SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I3P116

Volume 12 Issue 3, 161-170, March 2025

Review Article

Emerging Trends in Lung Cancer Detection: Multiomics and Deep Learning Synergy

Nrupal Sankpal[1], Sandeep Musale[2], Supriya Mangale[3]

[1,2,3]*Department of Electronics & Telecommunication Engineering, MKSSS's Cummins College of Engineering for Women, Maharashtra, India.*

[1]*Corresponding Author : nrupalsankpal@gmail.com*

Received: 10 January 2025 Revised: 11 February 2025 Accepted: 15 March 2025 Published: 29 March 2025

Abstract *- Lung cancer remains the leading cause of cancer-related deaths worldwide, with Lung Adenocarcinoma (LUAD) accounting for a significant portion of non-small cell lung cancer cases. Early detection and accurate prognosis prediction are critical for improving treatment outcomes. Recent advancements in multi-omics data integration, including genomics, transcriptomics, and histopathology, have shown promise in enhancing the prediction accuracy for LUAD. Lung cancer detection benefits significantly from machine learning models trained on multi-omics datasets, including gene expression, methylation, and mutations. Techniques such as Random Forest, SVM, and GLM have been employed to achieve robust prediction outcomes. The contribution of features was further analyzed using SHAP values. In particular, models such as LungDWM, which uses Generative Adversarial Networks (GANs) and attention-based feature encoders, have shown superior performance in diagnosing LUAD subtypes and predicting patient outcomes. This review explores the application of multiomics approaches, such as identifying key prognostic genes and developing machine learning models, to improve survival prediction and cancer staging while highlighting the current state of research, challenges, and future directions in using multi-omics for lung cancer detection and prognosis.*

Keywords *- Lung cancer, Machine Learning, Multiomics, Oncology.*

1. Introduction

The introduction should be succinct, with no subheadings. Limited figures may be included only if they are truly introductory and contain no new results. Lung cancer remains a significant global health concern, with Non-Small Cell Lung Cancer (NSCLC) constituting 85% of all lung cancer cases and Lung Adenocarcinoma (LUAD) being the predominant subtype, accounting for 40% of these cases. Despite advances in treatment, the prognosis for LUAD patients remains poor, with a five-year survival rate as low as 15%, primarily due to delayed diagnosis and tumor heterogeneity. Accurate survival prediction and risk stratification are critical for improving personalized treatment strategies. Multi-omics data, which integrates mRNA expression, miRNA, DNA methylation, and Copy Number Variations (CNV), has emerged as a promising approach to better understand LUAD mechanisms and identify reliable prognostic biomarkers. However, challenges such as data sparsity, high dimensionality, and class imbalance complicate effective feature selection and analysis. Machine learning and deep learning models solve these issues by improving predictive accuracy and identifying strong gene correlations. Furthermore, the tumor microenvironment, gut microbiome, and epigenomic modifications play crucial roles in cancer progression, necessitating integrative, explainable models for clinical applications. Addressing these challenges can advance early detection, enhance therapeutic outcomes, and provide actionable insights for LUAD prognosis. [3, 7, 8, 14-16]

Multi-omics data significantly enhances survival prediction in lung cancer by integrating diverse biological layers such as genomics, transcriptomics, proteomics, and metabolomics. This combination of data provides a comprehensive view of cancer biology, revealing insights into the molecular mechanisms that underlie cancer development and progression. By incorporating gene expression, DNA methylation, mutations, and protein alterations, multi-omics approaches offer a deeper understanding of tumor heterogeneity and allow for more accurate prognosis predictions, which are crucial for personalized treatment strategies. Studies show that multi-omics data can significantly outperform single-omics models in predicting survival risks for patients with Lung Adenocarcinoma (LUAD) [1, 16].

Moreover, multi-omics data plays a vital role in identifying molecular subtypes within lung cancer. These

subtypes exhibit distinct clinical behaviors, and by analyzing various data layers such as mRNA, miRNA, and copy number variations, researchers can identify specific patterns that correlate with patient outcomes. This integration enhances the ability to classify cancer subtypes more accurately, essential for selecting appropriate therapeutic interventions. Including DNA methylation data, for example, has been shown to improve survival predictions and support the identification of target groups for screening, ultimately leading to better risk stratification for lung cancer patients. [14, 16]

In addition to improving prognostic accuracy, multiomics data aids in understanding the complex interplay between genetic mutations and epigenetic regulation in lung cancer. Researchers gain new insights into tumour evolution and drug resistance mechanisms by examining how mutations interact with epigenetic factors such as DNA methylation and histone modifications. Furthermore, advances in single-cell profiling technologies have enhanced the ability to study tumor heterogeneity at a finer resolution, providing insights into cellular-level changes during cancer progression. These insights are crucial for developing personalized treatments tailored to the specific molecular profiles of individual patients [1, 11, 14].

Finally, integrating multi-omics data has profound implications for drug discovery and therapeutic strategies. By identifying molecular targets across various biological levels, multi-omics supports the development of precision therapies tailored to the specific characteristics of an individual's tumor. This approach not only aids in improving treatment efficacy but also plays a key role in overcoming challenges such as drug resistance. Furthermore, it accelerates the identification of biomarkers that could guide future therapeutic strategies, providing a foundation for precision medicine in lung cancer care. The holistic nature of multi-omics data transforms cancer research and clinical decision-making, offering the potential for more effective, personalized treatments [6].

Machine Learning (ML) and Deep Learning (DL) techniques have revolutionized the prediction of cancer treatment outcomes, particularly in lung cancer. By leveraging high-dimensional, complex datasets, ML algorithms can extract and select features that identify key biomarkers influencing disease progression. Algorithms such as random forests, Support Vector Machines (SVM), and Bayesian networks are widely used to build predictive models that analyze genetic, epigenetic, and clinical data. These models help clinicians make more accurate predictions regarding patient prognosis, facilitating personalized treatment plans. Additionally, ML models like random survival forests improve survival analysis accuracy by incorporating multi-omics data such as genomics, transcriptomics, and proteomics, enhancing the prediction of survival outcomes in cancer patients. Deep learning further enhances cancer treatment prediction by modeling complex relationships in omics data. Convolutional Neural Networks (CNNs) and autoencoders are powerful DL techniques that extract features from high-dimensional genomic data, offering valuable insights into cancer development and treatment response. Deep learning's ability to learn from large datasets makes it highly effective in identifying prognostic biomarkers, classifying cancer subtypes, and predicting treatment outcomes. Moreover, autoencoders and Deep Neural Networks (DNNs) are used for feature reduction and data fusion, improving accuracy in survival prediction by capturing intricate patterns in multi-omics data. Machine learning applications extend beyond survival prediction and treatment outcomes to lung cancer's early detection and staging. By analyzing gene expression data, clinical symptoms, and histopathological images, ML models can classify patients according to cancer progression and predict appropriate treatment options. Techniques such as SVMs and neural networks have successfully distinguished between early and late-stage cancers, thereby improving diagnostic capabilities. Integrating multiomics data—including microbiome and metabolic information—further strengthens the diagnostic process, providing a more comprehensive understanding of cancer. This approach is particularly beneficial in advancing personalized medicine, where treatment options are tailored to an individual's molecular profile [1, 4-6, 10, 11].

In addition to ML and DL, Cox regression plays a crucial role in cancer prognosis, particularly for survival analysis. This statistical technique helps identify the relationship between various risk factors and patient survival by modeling the hazard function. In lung cancer, Cox regression is valuable for identifying independent prognostic factors, such as genetic mutations, which influence cancer progression and treatment outcomes. By integrating multi-omics data, Cox regression provides deeper insights into how biomarkers and clinical variables contribute to survival predictions. Its ability to assess gene signatures and survival risk factors enhances the precision of prognosis and aids in developing personalized treatment strategies for cancer patients [1, 7, 9, 16].

Integrating multi-omics data, including gene expression, DNA methylation, mutations, and copy number variations, has proven to be a robust approach for predicting lung cancer survival, identifying biomarkers, and improving diagnostic accuracy. Multi-omics studies reveal that combining various data types enables better classification of high-risk and low-risk patient groups and improves cancer staging outcomes. Machine learning algorithms, such as random forest, support vector machines, and deep learning, are widely employed to handle high-dimensional omics data, demonstrating improved predictive capabilities compared to traditional statistical methods. Gene signatures and metabolic

biomarkers have emerged as key indicators for early diagnosis and prognostic evaluations, highlighting the importance of biological pathways and molecular subtypes in understanding tumor progression. Furthermore, research has introduced novel hallmarks of cancer, such as disrupted differentiation, phenotypic plasticity, and epigenetic reprogramming, as contributors to tumorigenesis [5-7, 10]. Advanced computational approaches have addressed challenges such as missing values, dataset imbalance, and heterogeneity, ensuring more reliable and interpretable results. Models developed using multi-omics data have been validated across multiple independent datasets, solidifying their effectiveness in clinical applications. Techniques leveraging histopathological imaging, microbial transcriptome analysis, and plasma metabolites further complement multi-omics studies, enabling improved survival predictions and early detection strategies. Additionally, functional analyses have identified critical biological processes and key prognostic genes that influence cancer progression and patient outcomes. These findings underscore the transformative potential of multi-omics and machine learning in advancing personalized medicine, improving risk stratification, and enhancing therapeutic strategies for lung cancer patients [4, 11].

Lung cancer prognosis and research face multifaceted challenges stemming from delayed detection, limited screening programs, and the complex nature of tumor biology. Late-stage diagnosis often hinders effective intervention and accurate survival predictions. The heterogeneity of Non-Small-Cell Lung Cancer (NSCLC) further complicates prognostic modeling due to the variability in tumor characteristics and patient-specific factors. Additionally, data-related challenges, such as missing values, high dimensionality, and noise, impact gene expression analysis and feature selection reliability. The lack of standardized datasets, insufficient clinical variables, and small sample sizes also limit the development and validation of robust predictive models across diverse populations [1, 2, 5, 8, 11-13].

Multi-omics integration offers significant promise but comes with its own difficulties, including incomplete datasets, redundancy in features, and the inability to fully capture relationships among various omics layers. Existing models often lack generalizability, as they are trained on limited or single institution datasets. Traditional machine learning methods struggle to handle the complexity of high dimensional omics data, which often contain noise and imbalanced features. Further challenges arise from the need for experimental validation, limited understanding of the role of epigenetic and metabolic factors, and insufficient consideration of comorbidities in prognostic models. Addressing these issues requires improved data integration methods, larger and more diverse datasets, and the development of more reliable and interpretable prediction frameworks [5-7, 13, 14]. Multi-omics approaches integrating genomics, transcriptomics, epigenomics, and proteomics have emerged as promising tools to enhance our understanding of LUAD and improve survival predictions. These methods offer a more comprehensive view of tumor biology by capturing diverse molecular alterations. However, challenges such as high-dimensional data, sparsity, class imbalance, and the lack of standardization across studies hinder the translation of multi-omics insights into clinical applications. Additionally, while Machine Learning (ML) and Deep Learning (DL) techniques have shown potential in analyzing complex omics data, issues related to model interpretability, overfitting, and reproducibility remain unresolved.

This review aims to comprehensively evaluate recent advancements in multi-omics data integration for LUAD prognosis. We critically assess various feature selection methods, ML/DL models, and their contributions to survival prediction, highlighting their strengths and limitations. Furthermore, we explore the role of the tumor microenvironment, epigenetic modifications, and molecular interactions in shaping patient outcomes. This review aims to offer insights into future directions for developing more robust and clinically relevant prognostic models by addressing the challenges associated with multi-omics analysis.

2. Multiomic Data Types in Lung Cancer Detection

Multiomic data integrates various biological datasets that capture information at different molecular levels, providing a comprehensive view of cellular processes and disease mechanisms. Genomic data, such as somatic mutations obtained from Whole-Exome Sequencing (WES) and Copy Number Variations (CNV), highlight DNA alterations and mutational landscapes critical to cancer development. DNA methylation data, a key component of epigenomics, identifies changes in gene regulation through methylation profiling. Transcriptomic data, including gene expression levels derived from total RNA sequencing (RNA-seq) and microRNA expression from miRNA sequencing, provides insights into transcriptional and post-transcriptional processes. These datasets collectively uncover variations in genetic and epigenetic mechanisms that influence disease progression and therapeutic responses [1, 4, 6, 8, 9, 11, 14, 16]. Additionally, multiomic data encompasses proteomics, metabolomics, and chromatin accessibility profiling, offering deeper functional and regulatory insights. Proteomics examines the entire protein landscape, analyzing protein expression, modifications, and interactions, while metabolomics investigates metabolic profiles, capturing biochemical alterations in cellular activity. Chromatin accessibility and histone modification analyses contribute to understanding transcriptional regulation and epigenetic

changes. Integrating these diverse omic data types, including post-transcriptional modifications and microbiome data, allows for a more robust understanding of complex diseases like cancer. This holistic approach bridges the gaps between genetic alterations, gene expression, and cellular functionality, enabling improved predictive models and biomarker discovery for diagnosis and prognosis [8, 10, 11].

2.1. Integration and Fusion of Multiomic Data for a Comprehensive View

Integrating and fusing multiomic data combine diverse datasets, such as genomics, transcriptomics, proteomics, and metabolomics, to provide a holistic understanding of complex biological systems and diseases. This process enhances predictive model accuracy by leveraging complementary features from different omics layers, addressing the limitations of single omic analyses. Fusion strategies, including early fusion (combining raw data) and late fusion (integrating processed features), are employed to optimize data integration. Machine learning methods, such as Convolutional Neural Networks (CNNs), Autoencoders (AEs), and Graph Neural Networks (GNNs), facilitate the discovery of intricate correlations between datasets, improving outcomes like cancer classification, survival prediction, and treatment response accuracy. Challenges such as missing data, varying dataset comparability, and high dimensionality are mitigated using Generative Adversarial Networks (GANs) and dimensionality reduction techniques. Furthermore, integrating multiomic data with histopathological images or microbiome profiles has improved staging prediction and prognostic models, enhancing cancer diagnosis's clinical relevance and accuracy [1, 4, 6, 8, 11].

3. Machine Learning Algorithms for Lung Cancer Detection

Machine learning algorithms are crucial in leveraging multiomic data for lung cancer detection. This section reviews various machine learning algorithms commonly employed in lung cancer detection and their applications in the field. These algorithms enable the identification of patterns and relationships within multiomic data, facilitating accurate prediction and classification of lung cancer subtypes. The materials and methods section should contain sufficient detail so that all procedures can be repeated. It may be divided into headed subsections if several methods are described.

3.1. Supervised Learning Algorithms

Random Forest, a widely-used supervised learning model, is effective for classification and predictive modeling tasks. It works by constructing multiple decision trees using random subsets of the training dataset, combining their predictions to produce the final output via majority voting. This ensemble approach helps mitigate overfitting and improves the model's generalization ability on unseen data. For instance, in studies involving the TCGA dataset, Random Forest demonstrated superior performance in terms of the Area Under the Curve (AUC) when compared to other methods such as SVM, LDA, GLM, and PLS. Additionally, nested cross-validation and independent dataset validation were used to ensure the robustness and reliability of the model. Random Survival Forest (RSF), a variation of Random Forest, has also been employed for survival data analysis, improving prognosis prediction accuracy for LUAD patients by handling censored data effectively [1, 11, 15, 16].

In comparative analyses, Random Forest often outperforms traditional machine learning algorithms like Logistic Regression (LR), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), particularly in high dimensional omic datasets. For cancer prediction tasks, supervised algorithms such as neural networks, SVM, and decision trees are evaluated based on performance metrics like accuracy, F1 macro, and F1 weighted scores. Random Forest's ability to handle complex, multifeatured data makes it particularly valuable in gene expression analysis, where traditional methods may struggle. Studies have also highlighted using fivefold cross-validation to validate model performance and address data sparsity and variability issues. Overall, the combination of Random Forest's interpretability and accuracy makes it a strong choice for supervised learning tasks in biomedical and cancer research domains [3, 5-7, 10].

3.2. Supervised Learning Algorithms

Unsupervised learning methods are pivotal in clustering and analyzing multi-omics data, particularly when labelled data is unavailable. These methods identify inherent patterns and relationships within datasets, making them ideal for exploratory data analysis and feature extraction. Techniques such as K-means clustering are widely employed for grouping multi-omics embeddings, enabling the discovery of biologically relevant clusters.

Autoencoders, a type of unsupervised artificial neural network, are commonly used to learn efficient representations of data by minimizing reconstruction errors. Variants like convolutional autoencoders enhance feature extraction capabilities and accelerate training, while advanced models such as efmmdVAE and lfAE have demonstrated superior clustering performance in multiomics analyses. Evaluation metrics like clustering indices ensure rigorous assessment of these methods, highlighting their ability to uncover hidden structures in complex datasets and provide meaningful insights into biological systems [9, 10, 16].

3.3. Deep Learning Models

Deep learning models are highly effective in predicting tumor types and subtypes, leveraging their ability to extract meaningful patterns from high-dimensional multi-omics

data. Techniques such as Mutation-Attention (MuAt) learn representations of somatic alterations, though local mutations often struggle to capture complex structural variants. Models like Autoencoders, Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), and Generative Adversarial Networks (GANs) are used to classify, cluster, and interpret biological data. Early and late fusion approaches combine diverse data sources, improving the accuracy and interpretability of results. Performance metrics such as accuracy, F1 score, and silhouette index are utilized for evaluation, with models like Autoencoders achieving significant survival differentiation and outperforming traditional single-omics methods. Despite challenges such as computational costs and the need for large datasets, deep learning methods, particularly those incorporating attention mechanisms and generative adversarial learning, show promise in enhancing precision and robustness in cancer diagnosis and prognosis [1, 2, 7, 11, 16].

4. Feature Selection and Dimensionality Reduction

Feature selection and dimensionality reduction are fundamental steps in processing multi-omics data for lung cancer detection, as they help manage the complexity of high-dimensional datasets. Feature selection focuses on identifying the most informative features contributing to predictive accuracy while discarding irrelevant or redundant variables. Techniques like Recursive Feature Elimination (RFE) systematically remove less significant features. LASSO regression imposes penalties to shrink coefficients of less relevant predictors to zero, ensuring a compact and meaningful feature set. Other methods, such as mutual information-based selection and genetic algorithms, capture non-linear dependencies, providing a comprehensive approach to isolating critical biomarkers in lung cancer. These methods not only improve computational efficiency but also enhance model interpretability, offering valuable insights into the biological mechanisms underlying lung cancer progression [1, 3, 14].

Dimensionality reduction, on the other hand, aims to transform high-dimensional data into a lower-dimensional space while preserving essential structures and patterns. Techniques like Principal Component Analysis (PCA) [13] reduce dimensionality by identifying orthogonal components that capture maximum variance in the data. Advanced methods like t-Distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are particularly effective for visualizing complex datasets by preserving local and global data relationships. In the realm of deep learning, autoencoders leverage neural networks to learn compressed representations of data. These techniques address the challenges of the curse of dimensionality, reducing the risk of overfitting and improving model generalization. By enabling the integration of multiple omics data types, such as genomic, transcriptomic, and proteomic datasets, feature selection and dimensionality reduction techniques empower researchers to uncover novel biomarkers and achieve more accurate lung cancer classification and subtype identification.

5. Comparative Analysis of different Machine Learning Models

Performance evaluation metrics assess the quality and effectiveness of machine learning models, classifiers, or algorithms in solving specific tasks, such as classification, regression, clustering, or recommendation. The choice of evaluation metric depends on the nature of the problem you are trying to solve.

In recent studies of multi-omics-based feature extraction and selection for predicting lung cancer survival, various machine-learning approaches have been employed to ensure robust evaluation and prediction accuracy. Jaksik et al. [1] utilized several machine learning methods, including Random Forest, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Generalized Linear Model (GLM), and Partial Least Squares (PLS), with nested cross-validation for a more reliable assessment. Feature importance was further analyzed using SHAP values, which provided insight into the contribution of each feature. They employed TCGA and CPTAC-3 datasets, encompassing 267 and 96 cases, respectively, and integrated multiple data types such as gene expression, methylation, mutations, and Copy Number Variations (CNVs). Notably, their study demonstrated an impressive AUC of 0.839 for TCGA and 0.815 for CPTAC-3 using Random Forest, where gene set aggregation showed the best feature extraction performance. The study highlighted key predictive features like gene expression, methylation, and mutations, supported by feature selection methods like Boruta, Lasso, and non-negative matrix factorization. Despite the promising results, challenges such as integrating clinical data, addressing batch effects, and overcoming limitations due to small sample sizes and structural variant data were noted. These issues stress the importance of enhancing feature extraction techniques and incorporating diverse datasets to improve prediction accuracy and better stratify patients for more precise survival outcomes.

Wang et al. [2] introduced LungDWM, a deep learning-based model for lung cancer subtype diagnosis using weakly-paired multi-omics data. Their method integrates attention-based encoders for feature extraction, Generative Adversarial Networks (GANs) for imputing missing omics data, and a fusion strategy for combining features to classify cancer subtypes. This approach achieved remarkable performance metrics, including an accuracy of 0.942, AUROC of 0.961, F1-score of 0.937, and AUPRC of 0.958. These results highlight the model's robustness, particularly under

conditions with missing data, and demonstrate its superiority over existing methods in lung cancer diagnosis. Additionally, LungDWM offers excellent interpretability, which can aid in identifying therapeutic sites, making it a powerful tool for precision medicine. Furthermore, the model's ability to handle incomplete data and maintain high diagnostic accuracy positions it as a promising approach for clinical applications where missing data is common.

Dessie et al. [3] proposed the PRPML model, which utilizes a nine-gene signature to predict the risk of Lung Adenocarcinoma (LUAD) patients, demonstrating robust prognostic capability. Through differential analysis and Adaptive LASSO for feature selection, the model was validated on three datasets-TCGA, GSE30219, and CMUH-covering a total of 865 samples. It achieved strong AUC scores of 0.812 and 0.863, effectively stratifying patients into high- and low-risk groups. Functional analysis revealed that high-risk groups were associated with critical biological pathways linked to poor survival outcomes. This work underscores the potential of the PRPML model for clinical applications, suggesting that further validation through alternative feature selection methods and functional experiments could improve its predictive accuracy and enhance its clinical utility.

Zhang et al. [4] developed a seven-gene signature for predicting the survival of Lung Adenocarcinoma (LUAD) patients by integrating multi-omics data. The approach employed a Random Survival Forest to filter prognostic genes, GISTIC 2.0 for identifying Copy Number Variations (CNV), and Cox regression for survival analysis. Validated across multiple datasets, including 516 SNP 6.0 samples, 576 RNA-Seq samples, and 226 GSE31210 samples, the signature demonstrated superior AUC performance over clinical features. Notably, MAGEL2, SMIM4, BCKDHB, and GANC were identified as novel prognostic markers, and gene set enrichment analysis highlights key biological pathways. The study stresses the need for experimental validation and the inclusion of additional clinical features to refine predictive models and further investigate the biological roles of these markers.

Anil Kumar et al. [5] applied Support Vector Machines (SVM) with SMOTE to classify lung cancer from text datasets, achieving an impressive 98.8% accuracy across five cancer datasets. The study employed neural networks with backpropagation and utilized 10-fold cross-validation for model evaluation. Key preprocessing techniques included addressing missing values using KNN, transforming data into binary format, and normalizing attributes for prediction. Random Forest was incorporated to further improve model performance, enhancing the accuracy and validation metrics such as f-measure and specificity. The authors suggest integrating real-time data, improving preprocessing steps, expanding the datasets to include more diverse demographics, and conducting longitudinal studies in clinical settings for practical validation and application of the model.

The study by Li et al. [6] employed a random forest algorithm to predict lung cancer stages, achieving an accuracy of 0.809. The analysis integrated microbial and transcriptomic data from 189 lung cancer patients and 1524 cases from prior studies. Through differential analysis, 291 upregulated and 128 down-regulated genes were identified, and significant pathways were revealed through GO and KEGG enrichment. Key microbial genera, such as Ureaplasma (more prevalent in early stages), and genes like REG4, CALCA, PHOX2B (downregulated), along with FOXI1, CYP1A1, LGI1, DLK1 (upregulated) were highlighted. The study emphasizes combining metabolomics with microbiome analysis to explore microbial metabolic regulation, advocating for the use of advanced technologies to enhance precision medicine in lung cancer treatment.

Liu and Wu [7] developed a deep neural network model for lung cancer prediction that integrates KL divergence for gene selection and focal loss as the loss function. This approach demonstrated remarkable performance, achieving an AUC of 0.99 on the validation set. The model was trained using RNAseq data from TCGA and ICGC datasets, which included 533 lung cancer and 59 normal samples, along with 488 lung cancer and 55 normal samples, respectively. By identifying 194 lung cancer-related genes, the model effectively addressed issues of imbalanced data and high-dimensional gene expression, enhancing both feature selection and overall model accuracy. Implemented in TensorFlow and validated through Kfold cross-validation, this model outperformed traditional algorithms such as SVM, LR, KNN, and RF in both accuracy and training speed. The study suggests future exploration into advanced feature selection techniques, integrating multi-omics data, improving interpretability for clinical applications, and reducing computational costs to enhance the model's broader applicability.

The CC2DT method, proposed by Rong et al. [9], combines Convolutional Neural Networks (CNN) for classification and Convolutional Autoencoders (CAE) for dimensionality reduction, offering a robust solution for lung cancer diagnosis using multi-omics datasets, including mRNA, miRNA, and DNA methylation. This method outperformed traditional models such as SVM, RF, LDA, ET, and MLP, achieving an accuracy of 0.824, an AUC of 0.749, and an F1 score of 0.855. Using 10-fold cross-validation and gradient descent effectively mitigated overfitting, improving performance on high-dimensional data. This approach shows significant potential for real-time diagnostics and precision medicine, with future opportunities to integrate additional omics data, explore alternative algorithms, and enhance clinical applicability.

The study by Xie et al. [10] utilizes advanced machine learning techniques, such as KNN, Naive Bayes, AdaBoost, SVM, Random Forest, and Neural Networks, trained with a 10-fold cross-validation approach, to develop a diagnostic model for early lung cancer detection. Plasma metabolites from 110 lung cancer patients and 43 healthy individuals were analyzed using targeted metabolomic studies with LC-MS/MS, measuring 61 metabolites. Six key metabolic biomarkers were identified, achieving an impressive AUC of 0.989, with 98.1% sensitivity and 100% specificity, in distinguishing stage I lung cancer from healthy individuals. The study classified patients into stages I, II, and III, identifying tumor types such as adenocarcinomas and squamous carcinomas. Integrating these biomarkers significantly enhanced diagnostic performance, with Naïve Bayes identified as the most suitable model for early tumor prediction. The study highlights the potential of blood-based screenings combined with machine learning to provide non-invasive and accurate lung cancer detection, thereby improving survival rates through early diagnosis. Future work suggests combining plasma biomarkers with CT screening and conducting confirmatory studies across diverse patient groups while considering factors like age and smoking history.

Chen et al. [11] employed advanced machine learning techniques to extract quantitative features from histopathological images and integrate them with multiomics data for predicting genetic aberrations and survival outcomes in LUAD patients. The study achieved high AUCs for genetic aberrations, including ALK (0.879), BRAF (0.847), and EGFR (0.855), while transcriptional subtype AUCs ranged from 0.861 to 0.897. Prognostic predictions for overall survival showed AUCs between 0.717 and 0.825, with the best multi-omics-integrated model achieving a 5-year AUC of 0.908. Kaplan-Meier analysis revealed distinct survival outcomes between high-risk and low-risk groups, emphasizing the prognostic power of these models. The study utilized data from The Cancer Genome Atlas (TCGA) and tissue microarrays, with digital images scanned using the Aperio AT2 scanner. This research demonstrates the potential of histopathological features in prognosis prediction, although further training on more diverse samples and addressing dataset biases is recommended to enhance its clinical applicability.

Luan et al. [14] developed a comprehensive survival risk model for lung adenocarcinoma by integrating multiomics data, including DNA methylation, RNA expression, microRNA profiles, and DNA copy number variations from 439 cases. The study applied the LASSO regression algorithm to identify 21 CpG sites as prognostic markers, highlighting their association with survival risks, particularly in chromosomal regions 17q24.3 (amplification) and 11p15.5 (deletion). Furthermore, Cox regression analysis and the iCluster algorithm revealed six molecular subtypes of lung adenocarcinoma, effectively distinguishing high-risk patients with lower survival rates. The model demonstrated strong prognostic capabilities, with AUCs surpassing 0.7 at 12, 36, and 60 months. Univariate Cox regression analysis also identified 29 mutant genes strongly correlated with survival, underlining the independent predictive power of methylation scores. Cross-validated time-dependent ROC curves confirmed the model's reliability, emphasizing the value of multi-omics data integration in uncovering critical prognostic markers. This research sets the stage for further clinical investigations into gene-survival associations in lung cancer.

Ma et al. [15] developed a lung adenocarcinoma survival risk model using a 16-gene signature identified through machine learning techniques. The model leveraged multi-omics data, including RNA expression, DNA methylation, and microRNA profiles, to create a predictive framework. By applying a preprocessing step to exclude probes with missing values, the LASSO regression algorithm identified critical biomarkers associated with survival. The study identified molecular subtypes through Cox regression and the iCluster algorithm, significantly differentiating high-risk patients from low-risk groups. The model demonstrated strong prognostic capabilities and high AUC values at various times. These findings emphasize the importance of integrating multi-omics data for personalized cancer prognosis and provide a valuable tool for clinical use in predicting lung adenocarcinoma survival outcomes.

A deep learning-based autoencoding model was developed for analyzing multi-omics data (mRNA, miRNA, DNA methylation, and CNV) to predict survival in Lung Adenocarcinoma (LUAD). This model incorporated feature selection techniques such as univariate Cox regression, Lasso regression, and K-means clustering. The survival prediction model, evaluated using random forest, achieved a C-index of 0.65 and a significant Log-rank P value of 4.08e-09. The analysis, validated with 399 LUAD samples from TCGA and four independent datasets from GEO and TCGA, identified genes linked to survival-related biological processes. Lee et al. [16] suggest that future improvements could involve integrating additional omics and clinical data alongside expanding the dataset for more accurate predictions across various cancer types and populations.

6. Discussion

In the realm of lung cancer detection, particularly LUAD (lung adenocarcinoma), multiomic data has proven to be invaluable in improving prognostic prediction and survival outcomes. Studies have identified many genes, such as sixteen key genes, demonstrating the ability to enhance prognosis prediction for LUAD patients. These gene signatures, integrated with clinical features, offer higher accuracy in predicting survival and treatment responses than traditional clinical indexes. Notably, advanced models such as Random Forests, deep learning, and ensemble Machine

Learning (ML) techniques have shown remarkable performance in predicting outcomes. For example, the deep neural network model outperformed traditional classifiers, and feature selection strategies further enhanced model accuracy, convergence, and training efficiency. Integrating various omic layers, including genomics, transcriptomics, and epigenomics, into predictive models has provided a more nuanced understanding of lung cancer biology. Moreover, multiomic fusion approaches have significantly improved the accuracy of lung cancer staging, highlighting the growing importance of integrating diverse biological data types for effective prognosis [1, 2, 4, 7, 11, 14].

Despite the advances, challenges remain in the practical application of these multiomic models. High-dimensional data, such as gene expression data, is often complicated using machine learning techniques like SVMs, which struggle with the high-dimensionality of features. Data preprocessing and feature selection are essential for ensuring reliable analysis and reducing computational costs associated with deep learning models. Furthermore, while multiomic approaches have enhanced biomarker discovery, the complexity of integrating data from various sources (such as genetic mutations, methylation, and copy number variations) still presents technical hurdles.

Additionally, the clinical applicability of these models is an ongoing challenge, as translating multiomic discoveries into actionable clinical tools requires extensive validation. Identifying specific genes such as REG4, CALCA, and PHOX2B, which show differential expression in lung cancer, can potentially develop targeted therapies. However, the continued development of computational strategies and the application of these models in diverse cancer types, as suggested by future work on expanding methods like LungDWM, will be critical to overcoming these challenges and improving early detection and treatment strategies [1, 2, 4, 7, 11, 14].

The comparative analysis of recent studies underscores the growing impact of multi-omics integration and advanced machine learning techniques in lung cancer prognosis and classification. Traditional models, such as Random Forest and SVM, have demonstrated strong predictive capabilities, as seen in [1, 6], where AUC values ranged from 0.809 to 0.839. However, deep learning approaches, particularly those incorporating attention mechanisms [2], Generative Adversarial Networks (GANs), and convolutional autoencoders [9], have shown significant improvements, achieving AUCs as high as 0.99 [7] while effectively handling missing data.

Additionally, survival risk prediction models leveraging feature selection techniques such as LASSO and Cox regression [3, 14] have successfully stratified high-risk patients, reinforcing the importance of multi-omics biomarkers in prognosis. Despite these advances, challenges persist in integrating diverse data types, mitigating batch effects, and ensuring clinical applicability.

Studies such as [11, 10] emphasize the need for improved interpretability and validation across larger, more diverse cohorts to enhance model generalizability. Future research should focus on refining feature extraction techniques, incorporating real-time clinical data, and exploring hybrid models that combine the strengths of traditional and deep learning approaches for more robust and interpretable predictions in lung cancer prognosis and diagnosis.

7. Summary & Conclusion

Lung cancer, particularly Lung Adenocarcinoma (LUAD), remains a major cause of cancer-related mortality worldwide. Recent advancements in multi-omics analysis have significantly enhanced the ability to predict the prognosis and survival outcomes for LUAD patients. Identifying key prognostic genes, such as the sixteen-gene signature and thirteen novel LUAD-related genes, has shown promise in improving predictive accuracy.

Machine learning models, including the use of random forests and deep learning techniques, have outperformed traditional prognostic methods like Cox models. Additionally, the integration of multi-omics data, including genomics, transcriptomics, and histopathological features, has improved survival predictions and lung cancer staging accuracy. Innovations such as the LungDWM model, which utilizes Generative Adversarial Networks (GANs) and attention-based feature encoders, have further improved the diagnosis and stratification of lung cancer subtypes. These findings underscore the potential of multi-omics fusion in advancing lung cancer detection and prognosis [1, 11, 15].

In conclusion, integrating multi-omics data has opened new avenues for improving the detection and prediction of LUAD. Significant strides have been made in highly accurate predicting survival and cancer subtypes by developing robust models, including those based on machine learning and deep learning techniques. The identification of genetic and epigenetic biomarkers, alongside the use of histopathological features, provides a more comprehensive approach to understanding LUAD biology and patient prognosis.

However, while these models demonstrate high potential, further validation and refinement are needed to ensure their clinical applicability. Future studies should focus on expanding multi-omics data integration across diverse populations, addressing the computational challenges of high-dimensional data, and validating these models in clinical settings to enhance early detection and personalized treatment strategies [1, 2, 11, 15].

References

[1] Roman Jaksik et al., "Multiomics-Based Feature Extraction and Selection for the Prediction of Lung Cancer Survival," *International Journal of Molecular Sciences*, vol. 25, no. 7, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Xingze Wang et al., "Lung Cancer Subtype Diagnosis Using Weakly-Paired Multi-omics Data," *Bioinformatics*, vol. 38, no. 22, pp. 5092-5099, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Eskezeia Yihunie Dessie, Jan-Gowth Chang, and Ya-Sian Chang, "A Nine-Gene Signature Identification and Prognostic Risk Prediction for Patients with Lung Adenocarcinoma using Novel Machine Learning Approach," *Computers in Biology and Medicine* vol. 145, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Surong Zhang et al., "Identification of Seven-Gene Marker to Predict the Survival of Patients with Lung Adenocarcinoma using Integrated Multi-Omics Data Analysis," *Journal of Clinical Laboratory Analysis*, vol. 36, no. 2, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] C. Anil Kumar et al., "[Retracted] Lung Cancer Prediction from Text Datasets Using Machine Learning," *BioMed Research International*, vol. 2022, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Wei Li et al., "[Retracted] Lung Cancer Stage Prediction Using Multi-Omics Data," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Suli Liu, and Wu Yao, "Prediction of Lung Cancer using Gene Expression and Deep Learning with KL Divergence Gene Selection," *BMC Bioinformatics*, vol. 23, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Douglas Hanahan, "Hallmarks of Cancer: New Dimensions," *Cancer Discovery*, vol. 12, no. 1, pp. 31-46, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] ZHU Rong et al., "Diagnostic Classification of Lung Cancer Using Deep Transfer Learning Technology and Multi-Omics Data," *Chinese Journal of Electronics*, vol. 30, no. 5, pp. 843-852, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ying Xie et al., "Early Lung Cancer Diagnostic Biomarker Discovery by Machine Learning Methods," *Translational Oncology*, vol. 14, no. 1, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Linyan Chen et al., "Histopathological Images and Multi-omics Integration Predict Molecular Characteristics and Survival in Lung Adenocarcinoma," *Frontiers in Cell and Developmental Biology*, vol. 9, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Aditya Dubey, and Akhtar Rasool, "Efficient Technique of Microarray Missing Data Imputation using Clustering and Weighted Nearest Neighbour," *Scientific Reports*, vol. 11, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Xinshan Zhu et al., "An Efficient Ensemble Method for Missing Value Imputation in Microarray Gene Expression Data," *BMC Bioinformatics*, vol. 22, pp. 1-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mingyuan Luan et al., "Multi-omics Integrative Analysis and Survival Risk Model Construction of Non-Small Cell Lung Cancer Based on the Cancer Genome Atlas Datasets," *Oncology Letters*, vol. 20, no. 4, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Baoshan Ma et al., "Identification of a Sixteen-gene Prognostic Biomarker for Lung Adenocarcinoma using a Machine Learning Method," *Journal of Cancer*, vol. 11, no. 5, pp. 1288-1298, 2020. [Google Scholar] [Publisher Link]

[16] Tzong-Yi Lee et al., "Incorporating Deep Learning and Multi-omics Autoencoding for Analysis of Lung Adenocarcinoma Prognostication," *Computational Biology and Chemistry*, vol. 87, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Halil Ibrahim Kuru, Oznur Tastan, and A. Ercument Cicek, "MatchMaker: A Deep Learning Framework for Drug Synergy Prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 4, pp. 2334-2344, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Tongxin Wang et al., "MOGONET Integrates Multi-Omics Data using Graph Convolutional Networks Allowing Patient Classification and Biomarker Identification," *Nature Communications*, vol. 12, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Xiaohan Xing et al., "An Interpretable Multi-Level Enhanced Graph Attention Network for Disease Diagnosis with Gene Expression Data," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, USA, 556-561, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Muta Tah Hira et al., "Integrated Multi-omics Analysis of Ovarian Cancer using Variational Autoencoders," *Scientific Reports*, vol. 11, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Md. Mohaiminul Islam et al., "An Integrative Deep Learning Framework for Classifying Molecular Subtypes of Breast Cancer," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2185-2199, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Yuhua Fu et al., "A Gene Prioritization Method Based on a Swine Multi-Omics Knowledgebase and a Deep Learning Model. *Communications Biology*, vol. 3, no. 1, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Long-Yi Guo et al., "Deep Learning-Based Ovarian Cancer Subtypes Identification using Multi-Omics Data," *BioData Mining*, vol. 13, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Tianle Ma, and Aidong Zhang, "Integrate Multi-omics Data with Biological Interaction Networks using Multi-View Factorization Autoencoder (MAE)," *BMC Genomics*, vol. 20, pp. 1-11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Xiaoyu Zhang et al., "Integrated Multiomics Analysis using Variational Autoencoders: Application to Pan-Cancer Classification," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, pp. 765-769, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Jonathan Ronen, Sikander Hayat, and Altuna Akalin, "Evaluation of Colorectal Cancer Subtypes and Cell Lines using Deep Learning," *Life Science Alliance*, vol. 2, no. 6, pp. 1-16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[27] Kristina Preuer et al., "DeepSynergy: Predicting Anti-Cancer Drug Synergy with Deep Learning," *Bioinformatics*, vol. 34, no. 9, 1538-1546, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[28] Olivier B. Poirion, Kumardeep Chaudhary, and Lana X. Garmire, "Deep Learning Data Integration for Better Risk Stratification Models of Bladder Cancer," *AMIA Joint Summits on Translational Science Proceedings*, pp. 197-206, 2018. [CrossRef] [Google Scholar] [Publisher Link]