Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P); 2583-7907 (E)
Vol.13 No.1, 2024, pp.45-58

www.crijournals.org
DOI: https://doi.org/10.70112/ajcst-2024.13.1.4266

Environmental and Genetic Interaction Models for Predicting Lung Cancer Risk Using Machine Learning: A Systematic Review and Meta-Analysis

Ernest E. Onuiri, Bright G. Akwaronwu and Kelechi C. Umeaka
Department of Computer Science, School of Computing, Babcock University, Nigeria
onuirie@babcock.edu.ng, akwaronwu0329@pg.babcock.edu.ng, umeaka0475@pg.babcock.edu.ng
(Received 8 March 2024; Revised 5 April 2024; Accepted 20 April 2024; Available online 25 April 2024)

***Abstract* - This systematic review and meta-analysis examined environmental and genetic interaction models for predicting lung cancer risk using machine learning techniques. The findings underscore the importance of considering both genetic and environmental factors to enhance predictive accuracy, with significant clinical implications. Among the models reviewed, the Rotation Forest model demonstrated superior performance, achieving an AUC of 0.993, reflecting excellent predictive capabilities. The mean AUC across all models was approximately 0.789, indicating moderate to good discrimination. These results hold promising implications for personalized medicine and clinical decision-making, potentially improving patient outcomes and reducing the global burden of lung cancer. The meta-analysis further highlighted strong performance, with an average TRI-performance (accuracy, precision, recall) of 88.1%, demonstrating robust predictive abilities. Integrating machine learning with multidimensional data deepens the understanding of the biological mechanisms underlying lung cancer and supports its application in precision oncology, paving the way for individualized interventions and improved clinical management.**
Keywords: **Lung Cancer Risk, Machine Learning, Genetic and Environmental Factors, Rotation Forest Model, Precision Oncology**

I. INTRODUCTION

The dynamics between nature and nurture are reflected in the biology of gene-environment (G×E) interaction [1]. These dynamics aid in addressing persistent issues in psychology regarding the roles of environment and heredity [2], [3]. Phenotypic diversity, produced by genetic mechanisms such as recombination and mutation, or by varying the expression of the same genome in response to environmental stimuli, allows organisms to adapt to a variety of environments [4], [5]. "Etiologic heterogeneity arises from variations in environmental and genetic factors across individuals, leading to diverse disease presentations within seemingly similar clinical groups" [5], [6]. To visually analyze nuclear alterations, light microscopy is used, which is an important tool in cancer detection. Nuclear structure is quantified using numerical parameters through computer-aided diagnostic tools, aiding in prognosis [7], [8]. Predicting cancer involves evaluating survivorship, recurrence, and susceptibility [9], [10]. Recurrence indicates the likelihood of cancer returning after therapy, survival assesses outcomes following diagnosis, and susceptibility predicts cancer risk [11], [12].

Lung cancer ranks as the leading cause of global cancer incidence and mortality, with 1.8 million documented deaths and approximately 2 million diagnoses each year [13], [14]. The biology of lung cancer is well-defined, offering valuable insights for comprehending its complexities [15]. This has enabled the development of individualized treatment plans based on biomarker profiling and histologic classification [16]. Despite declining smoking rates, a rising trend in smoking among women in developed nations has led to an increasing incidence of lung cancer in females [17], [18].

Correlations between radiologic, clinical, and pathologic findings highlight the growing occurrence of adenocarcinoma histology and advance our understanding of genetic profiling and personalized treatments [8], [12]. A medical diagnosis is essential for prognosis, which goes beyond diagnosis by emphasizing the importance of evaluating numerous variables [11], [19]. Lung cancer risk is influenced by various factors, the most common being lifestyle, environmental exposures, and occupational hazards. These factors have different impacts based on geography, gender, ethnicity, genetic predisposition, and their combined effects [17], [20]. The outcome of genetic-reactive chemistry varies among environmental hosts [1]. Identifying genes that affect a person's susceptibility to complex traits requires effectively managing environmental conditions. This is particularly true for lung cancer, where smoking and environmental factors play a significant role in its etiology [14], [18], [21].

As the application of machine learning (ML) grows in healthcare, effective regulations are needed to monitor and control the development of ML-based applications and other artificial intelligence (AI) software for early disease diagnosis and prediction across a range of medical conditions [18], [22]. The integration and analysis of large datasets related to lung cancer using ML models provide valuable tools in lung cancer research, offering opportunities for improved understanding and clinical application [23]. The potential of ML techniques has enhanced the accuracy of predictions related to survival, recurrence, and susceptibility to cancer, emphasizing the use of feature selection and classification methods, as well as the integration of multidimensional heterogeneous data for cancer inference [24], [25].

Data preprocessing is an essential first step in improving data quality, followed by feature extraction, model implementation, training, and parameter adjustment to ensure reliable options and accurate predictions [22], [26], [27]. Medical image analysis has significantly advanced through ML techniques, particularly in automated feature extraction [28], [29]. Effective medical diagnoses and classifications result from proper feature selection and model training, ensuring accurate predictions [30], [31]. This research study will examine the relationship between genetic variants and environmental factors in predicting lung cancer risk. The review will focus on studies employing ML techniques to develop lung cancer risk prediction models, examining the features, datasets, and methods used, as well as how the results may enhance lung cancer risk assessment and suggest potential clinical applications and future research directions.

A. Rationale

This systematic review was conducted to enhance our understanding of the predictive factors surrounding lung cancer risk, as well as models used to identify consistency and variability, and to evaluate their performance in terms of adaptability and clinical implications [32]. It aims to thoroughly examine the reliability of measurement instruments utilized in research, emphasizing their roles and features in ensuring accurate data collection [33], [34]. The review applies environmental and genetic interacting factors to develop models for predicting lung cancer risk using machine learning.

B. Objectives

This research aims to review existing articles on environmental and genetic factors influencing lung cancer risk prediction using machine learning techniques. It focuses on environmental and genetic factors, evaluates predictive model performance, and explores interactions between these factors to provide feasible directions for future practice. The study follows the PICOS framework [35], which includes:

1. *Population:* Analyze the efficiency of machine learning models in predicting lung cancer risk based on environmental and genetic factors.
2. *Intervention:* Identify limitations in research and propose recommendations for clinical applications and future studies in lung cancer risk prediction using machine learning.
3. *Comparison:* Investigate the interaction between environmental and genetic factors in the development of predictive models for lung cancer risk.
4. *Outcome:* Identify and summarize the environmental and genetic factors associated with the use of machine learning algorithms in lung cancer risk prediction models, and recommend additional applications for personalized and clinical use.
5. *Study Design:* Articles will be thoroughly screened, both manually and automatically, to ensure that only relevant documents are included in the review. The study applies the PRISMA 2020 guidelines [36]. Data sources for the analysis will include final articles extracted from Scopus, PubMed, and Google Scholar platforms.

II. METHODOLOGY

The method employed in this research study adheres to the PRISMA 2020 Statement [36]. Articles were retrieved from Scopus and PubMed on February 22, 2024, and from Google Scholar, along with random searches from various search engines, on February 16, 2024. Effective filtering techniques were applied to exclude articles unrelated to the research topics, including those not written in English. The search strategy incorporated terms related to “Environment*”, “Genetic”, “genetic*”, “gen* interact*”, “gene* variation”, “model* approach”, “Lung* cancer”, “Lung* carcinoma”, “Pulmonary cancer”, “Pulmonary carcinoma”, “respiratory carcinoma”, “bronchial carcinoma”, “lung tumor”, “respiratory malignancy”, “thoracic malignancy”, “pulmonary neoplasm”, “respiratory tumor”, “risk”, “predict*”, “estimation”, “analy*”, “forecast*”, “risk assessment”, “predictive modeling”, “data mining”, “machine learning”, and “Artificial Intelligence”.

A. Eligibility Criteria

The eligibility criteria for this research study followed the PICOS framework guidelines [35], which are:

1. *Population:* Studies involving human populations, particularly those at risk of lung cancer or diagnosed with lung cancer.
2. *Intervention:* Studies investigating environmental and genetic factors that may influence lung cancer risk, and studies utilizing machine learning techniques to develop predictive models or algorithms for lung cancer risk prediction.
3. *Comparison:* Studies comparing the predictive performance of models incorporating both environmental and genetic factors versus models considering either factor alone. Studies evaluating different machine learning algorithms for lung cancer risk prediction.
4. *Outcome:* Studies that aim to predict the risk of lung cancer using machine learning techniques, and studies reporting performance metrics.
5. *Study Design:* Articles extracted from Scopus, PubMed, and Google Scholar platforms will be screened and included or excluded from the review based on the stipulated criteria to ensure the selection of high-quality and relevant articles. The analysis will focus on model performance and proposed clinical applications.

B. Inclusive Criteria

The inclusion criteria were applied based on the following:

1. Articles focusing on human environments, specifically populations at risk of lung cancer or diagnosed with lung cancer.
2. Investigations of environmental and genetic factors that may influence lung cancer risk.
3. Lung cancer risk prediction using machine learning techniques.

4. Observational and interventional studies that utilize machine learning methods for lung cancer risk prediction.
5. Research focusing on human lungs.
6. Studies published in English only.
7. Articles published between 2013 and 2024.
8. Articles available in full-text format.

C. Exclusive Criteria

The exclusion criteria were based on the following.

1. Studies focusing solely on non-machine learning methods for risk prediction.
2. Studies not related to lung cancer risk prediction.
3. Studies with inadequate reporting of methods and results.
4. Duplicated studies.
5. Studies not written in English.
6. Incomplete articles, newspapers, and conference papers.
7. Articles published before 2013.

These criteria ensure that relevant studies focusing on environmental and genetic interaction models for lung cancer risk prediction using machine learning are included, while excluding studies that do not meet the specific objectives of the review.

D. Information Sources

1. Scopus Querry - Thursday February 22, 2024 https://www.scopus.com/
2. PubMed - Thursday February 22, 2024 https://pubmed.ncbi.nlm.nih.gov/
3. Google Scholar - Friday February 16, 2024 https://scholar.google.com/

E. Search Strategy

The search strategy was designed to cover every aspect of the investigation related to the topic of the research: Environmental and Genetic Interaction Models for Lung Cancer Risk Prediction using Machine Learning. It encompasses fields such as Medicine, Biological and Chemical Sciences, Genetics, Computer and Information Technology, Environmental Science, and Decision Sciences. The strategy also includes research focusing on human environments, environmental and genetic factors that may influence lung cancer risk prediction using machine learning techniques, and articles published in English from 2013 to 2024.

F. Selection Process

Scopus: A total of 491 documents were successfully retrieved from the Scopus web platform using the following query string:
TITLE-ABS-KEY (((environment* OR genetic* OR “gene* interact*” OR “gene* variation”) AND (model* OR approach) AND (“Lung* cancer” OR “Lung* carcinoma” OR “Pulmonary cancer” OR “Pulmonary carcinoma” OR “respiratory carcinoma” OR “bronchial carcinoma” OR “lung tumor” OR “respiratory malignancy” OR “thoracic malignancy” OR “pulmonary neoplasm” OR “respiratory tumor”) AND (risk OR predict* OR estimation OR analy* OR forecast* OR “risk assessment” OR “predictive modeling” OR “data mining”) AND (“machine learning” OR “Artificial Intelligence”))) AND PUBYEAR > 2009 AND PUBYEAR < 2025 AND PUBYEAR > 2013 AND PUBYEAR < 2025 AND (LIMIT-TO (SRCTYPE , “j”)) AND (LIMIT-TO (DOCTYPE , “ar”)) AND (LIMIT-TO (SUBJAREA , “BIOC”) OR LIMIT-TO (SUBJAREA , “COMP”) OR LIMIT-TO (SUBJAREA , “MEDI”) OR LIMIT-TO (SUBJAREA , “IMMU”) OR LIMIT-TO (SUBJAREA , “ENVI”) OR LIMIT-TO (SUBJAREA , “NEUR”) OR LIMIT-TO (SUBJAREA , “DECI”)) AND (LIMIT-TO (LANGUAGE , “English”)) AND (LIMIT-TO (PUBSTAGE , “final”))

PubMed: A total of 275 documents were successfully extracted from the PubMed web platform using the following query string:
((“environment*”[All Fields] OR “genetic*”[All Fields] OR “gene interaction”[All Fields] OR “gene variation”[All Fields]) AND (“model*”[All Fields] OR “approach”[All Fields] OR “approach s”[All Fields] OR “approachability”[All Fields] OR “approachable”[All Fields] OR “approache”[All Fields] OR “approached”[All Fields] OR “approaches”[All Fields] OR “approaching”[All Fields] OR “approachs”[All Fields]) AND (“lung cancer”[All Fields] OR “lung carcinoma”[All Fields] OR “pulmonary cancer”[All Fields] OR “pulmonary carcinoma”[All Fields] OR “respiratory carcinoma”[All Fields] OR “bronchial carcinoma”[All Fields] OR “lung tumor”[All Fields] OR “respiratory malignancy”[All Fields] OR “thoracic malignancy”[All Fields] OR “pulmonary neoplasm”[All Fields] OR “respiratory tumor”[All Fields]) AND (“risk”[MeSH Terms] OR “risk”[All Fields] OR “predict*”[All Fields] OR “estimability”[All Fields] OR “estimable”[All Fields] OR “estimate”[All Fields] OR “estimated”[All Fields] OR “estimates”[All Fields] OR “estimating”[All Fields] OR “estimation”[All Fields] OR “estimations”[All Fields] OR “estimator”[All Fields] OR “estimator s”[All Fields] OR “estimators”[All Fields] OR “analy*”[All Fields] OR “forecast*”[All Fields] OR “risk assessment”[All Fields] OR “predictive modeling”[All Fields] OR “data mining”[All Fields]) AND (“machine learning”[All Fields] OR “Artificial Intelligence”[All Fields])) AND (y_10[Filter] AND fha[Filter] AND humans[Filter] AND english[Filter])

Google Scholar: The research title “Environmental and Genetic Interaction Models for Lung Cancer Risk Prediction Using Machine Learning” was used as a query string on the Google Scholar platform, and a total of 64 documents were retrieved.

G. Data Collection Process

Documents retrieved from Scopus were exported in “Comma-Separated Values” (CSV) and “Research Information Systems” (RIS) file formats, while documents from PubMed were saved in CSV and PubMed file formats. Both sets were successfully loaded into Mendeley Reference Manager. Documents sourced from Google Scholar were

directly saved to Mendeley Reference Manager using the Mendeley extension in the web browser. The final search was conducted on Thursday, February 22, 2024, and all references were exported to Hubmeta (https://hubmeta.com/) for duplication checking, title screening, and full-text screening. A total of 830 documents were submitted to Hubmeta for review. Following the application of both inclusion and exclusion criteria for selecting essential articles, 81 articles were retrieved from Hubmeta for manual screening and review.

H. Data Items

Article bibliographies were retrieved from Scopus and PubMed on February 22, 2024, using the query strings described in section 2.6 above, and from Google Scholar on February 16, 2024, using the research title as a query string. A total of 830 documents were retrieved from these three bibliographic databases as follows: Scopus (491), PubMed (275), and Google Scholar (64). The screening process was conducted to exclude files as follows: 215 duplicates were identified, and 54 documents were deemed ineligible by the Hubmeta screening tool. The second stage of screening involved title and abstract review, which revealed that 453 documents did not align with the objectives of the research. Additionally, the full text of 27 documents was unavailable, 2 documents were retracted, and the contents of 67 documents did not explicitly cover the basis of this research study.

I. Study Risk of Bias Assessment

The methods for evaluating bias in each article involved assessing both the outcome and study levels. At the outcome level, the risk of bias was evaluated by examining the quality of reporting and methodology for each outcome of interest, such as lung cancer risk prediction using machine learning techniques. This assessment included evaluating factors such as the clarity and completeness of outcome reporting, the appropriateness of statistical analyses, and the potential for outcome misclassification or measurement bias. At the study level, the risk of bias was evaluated by considering various factors that could impact the validity of the study results. This assessment included examining the study design, participant selection methods, potential confounding variables, and sources of bias such as funding sources or conflicts of interest. The information gathered from assessing the risk of bias for each article was used in data integration to analyze and interpret the findings. Articles with a high risk of bias were given less weight in the overall analysis, while studies with a lower risk of bias were prioritized.

J. Data Flowchat

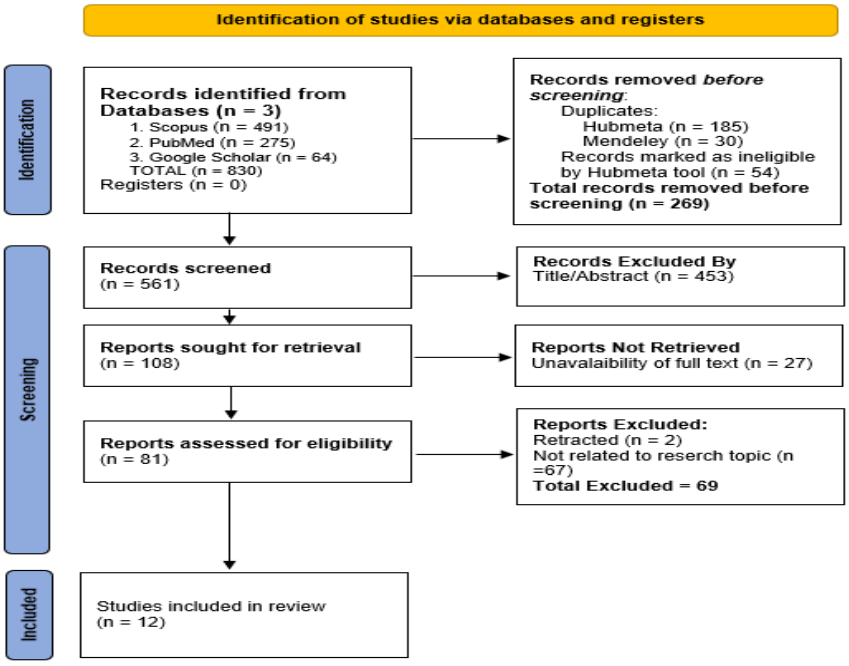

Fig. 1 2020 PRISMA Flowchart [36]

III. RESULTS

This research presents a summary of 12 studies that focus on various factors influencing prognosis, environmental and genetic factors, and predictive elements. The included studies analyzed how these factors contribute to specific outcomes, such as disease progression and prediction. By categorizing the studies based on these factors, the research aims to provide an in-depth analysis of the landscape in this area of study. The document highlights the importance of considering various factors, both internal and external, in understanding and predicting outcomes in different contexts. Table I below identifies key themes across the studies, contributing to a broader understanding of prognostic and predictive factors in lung cancer using machine learning.

TABLE I FINDINGS OF INCLUDED STUDIES

Pg = Prognostic features; Er = Environmental Factor Considered; Gn = Genetic Factor Considered; Pr = Predictive Elements

Sl. No.	Ref	Author(s) (Year)	Title	Machine Learning Methodology	Pg	Er	Gn	Pr	Measures	Validation	Findings	Results
1	[1]	Manuck *et al.,* (2014)	Gene-Environment (GxE) Interaction	Conceptual models related to gene-environment (G × E) interactions, such as the diathesis-stress model, vantage sensitivity, and differential susceptibility.	Yes	Yes	Yes	Yes	G×E interaction studies in psychology	Examination of conceptual models and rationales for G × E interactions in psychological research.	Controversy exists regarding the interpretation and significance of G × E interactions in psychology. Some researchers embrace the prospect of G × E interactions affecting behavior, while others criticize the weaknesses and inflated claims associated with G × E research.	The increase in G × E literature corresponds with decreasing expectations of swift advancements in pinpointing genes directly linked to psychological traits and disorders. G × E interaction is proposed as one of several hypotheses to elucidate the limited success in identifying genetic variants for behavioral phenotypes.
2	[7]	Kukreja S. *et al.,* (2023)	A Heuristic Machine Learning-Based Optimization Technique for Predicting Lung Cancer Patient Survival	Naive Bayes, SSA, Rapid Decision Tree Learner, and K-Nearest Neighbor algorithms were implemented and evaluated using the lung cancer dataset.	Yes	No	Yes	Yes	The mean absolute error (MAE) of the predictions, accuracy, recall, and precision.	The validation process includes assessing the accuracy, recall, and precision of the proposed technique's predictions.	The study utilized a dataset that included over 100 cases sourced from the Wisconsin Prognostic Lung Cancer subdirectory. By leveraging this dataset, the researchers discovered a significant volume of real-world data related to lung cancer cases, focusing on prognostic information. The dataset contained detailed information about the characteristics of cancer cell nuclei, such as radial distance, opacity, and location.	When predicting how long a lung cancer patient would survive within five years, the mean absolute error (MAE) of the predictions made by this technique is accurate within a month. It achieved an accuracy, recall, and precision of 98.78%, 98.4%, and 98.6%, respectively.
3	[18]	Dritsas *et al.,* (2022)	Lung Cancer Risk Prediction Using Machine Learning Models	Machine Learning Models	Yes	Yes	No	Yes	Accuracy, Precision, Recall, F-Measure, AUC	10-fold cross-validation	The Rotation Forest model achieved high performance with an AUC of 99.3%, and accuracy, precision, recall, and F-measure values of 97.1%.	The proposed model outperformed other models and demonstrated superior results compared to the reference models.
4	[37]	Wang N. *et al.,* (2023)	Development and Validation of Polyamine Metabolism-Associated Gene Signatures to Predict Prognosis and Immunotherapy Response in Lung Adenocarcinoma	Utilizing the Least Absolute Shrinkage and Selection Operator (LASSO) method to build a risk score model based on gene signatures associated with polyamine metabolism.	Yes	Yes	Yes	Yes	AUC, p-value	Validated the prognostic prediction model in an independent cohort. (GSE72094)	LUAD patients were divided into high- and low-risk groups according to the risk score methodology. It was found that there are two unique subgroups of LUAD patients (C1 and C2). A total of 291 differentially expressed genes (DEGs) were identified by comparing the two subgroups; these genes were primarily enriched in the cell cycle, nuclear division, and organelle fission.	To create the nomogram, three independent prognostic factors - PSMC6, SMOX, and SMS-were identified, all of which were found to be elevated in LUAD cells.
5	[38]	Y. Liu *et al.,* (2022)	Development and Validation of Machine Learning Models to	Logistic Regression (LR), Decision Tree (DT), Random	No	Yes	Yes	Yes	CAL, AUC, DCA, C-index and Brier score	validation cohort, 10-fold cross-validation	Identification of sixteen radiomics features selected for	The RF classifier outperformed the LR, DT, and SVM classifiers in

			Predict Epidermal Growth Factor Receptor Mutation in Non-Small Cell Lung Cancer: A Multi-Center Retrospective Radiomics Study	Forest (RF), Support Vector Machine (SVM)							building the model for lung cancer diagnosis.	both training and validation cohorts, achieving higher AUC, calibration, and C-index, and a lower Brier score. The DeLong test showed no significant difference in AUC between the training and validation cohorts for the RF classifier but revealed significant differences for the LR, DT, and SVM classifiers.
6	[39]	H. Lee *et al.,* (2024)	Evaluating county-level lung cancer incidence from environmental radiation exposure, PM2.5, and other exposures using regression and machine learning models.	Tree-based machine learning (ML) models	No	Yes	No	Yes	MAPE, RMSE	fivefold cross-validation	The correlation between environmental radon levels and PM2.5 was found to be significant, suggesting a potential synergistic influence of both on health outcomes.	Poisson regression: MAPE = 6.29, RMSE = 12.70; Poisson random forest regression: MAPE = 1.22, RMSE = 8.01.
7	[40]	Q. Cai *et al.,* (2020)	Exploration of predictive and prognostic alternative splicing signatures in lung adenocarcinoma using machine learning methods.	Random forest-based classifiers, Cox regression model, random survival forest analysis, and forward selection model.	Yes	No	No	Yes	ROC, AUC	fivefold cross-validation.	Each ASE pair exhibited a complete inverse correlation (correlation coefficient = -1). The 12-ASE classifier effectively assessed lymph node metastasis (LNM) status in LUAD patients and identified crucial prognosis-related ASEs. A 16-ASE prognostic model was developed to predict overall survival in LUAD patients.	The 12-ASE-based classifier effectively assessed lymph node metastasis (LNM) in LUAD patients, with an AUC exceeding 0.7. The prognostic model demonstrated consistent performance over 1, 3, 5, and 10 years in both the training and internal test groups.
8	[41]	Rani *et al.,* (2023)	Exploring Machine Learning in Lung Cancer: Predictive Modeling, Gene Associations, and Challenges	Support Vector Machines, Random Forests, Deep Learning Architectures, Network-Based Methodologies	Yes	No	Yes	Yes	Prediction Methods, Gene Association Analyses, Biomarker Identification	Analysis includes diverse data types such as gene expression, genomic variants, and clinical data.	Extensive investigation into various machine learning algorithms, focusing on their practical applications in predictive modeling, biomarker identification, and drug discovery pathways.	Specific emphasis on the utilization of SVM, RF, logistic regression, CNNs, RNNs, and GCN seeks to provide a solid foundation for upcoming advancements in treatment, diagnosis, and prognosis.
9	[42]	Pati J. (2019)	Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach	Advanced machine learning techniques	Yes	Yes	Yes	Yes	CI, Accuracy, Precision, Recall, F-Measure,	Evaluation of gene expression data for lung cancer. Selection and prediction of the optimal subset of genes likely to cause lung cancer.	Identification of the strongest candidate genes with accurate prognostic value for determining susceptibility to lung cancer.	Strong emphasis is placed on utilizing gene expression data analysis and advanced machine learning techniques to identify key genes associated with lung cancer susceptibility, aiming to improve early

											prediction and understanding of the eco-genomics of cancer.
10	[43]	Okser *et al.,* (2013)	Genetic Variants and Their Interactions in Disease Risk Prediction: Machine Learning and Network Perspectives	Predictive modeling approaches	Yes	Yes	Yes	Yes	Mining genotype-phenotype relationships	Internal and external cross-validation	Acknowledgement of the role that interactions between genetic loci play in the development of complex phenotypic traits and human diseases. Highlighting the challenges in identifying genetic markers for disease risk prediction. Introduction of machine learning approaches to address the issue of missing heritability and enhance disease risk prediction models.	Emphasis on utilizing machine learning-based approaches to identify hidden interactions among genetic factors. Discussion of the challenges in implementing scalable algorithms for genetic feature selection and validating predictive models. Suggestion of incorporating additional biological information, such as physical protein interaction networks, to enhance the model construction process.
11	[44]	Wang *et al.,* (2022)	How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine Learning-Based Modeling and Benchmarking	Stepwise regression, variance inflation factor, and machine learning algorithms (linear regression, support vector regression, random forest, k-nearest neighbors, Cubist model tree).	Yes	Yes	No	Yes	RMSE, R-squared	5-fold cross-validation	Identification of risk factors impacting lung cancer incidence rates, including air pollution, tobacco use, socioeconomic status, employment status, marital status, and environmental factors.	The study identified the Cubist model tree, utilizing feature selection, as the best-performing model, with an RMSE of 3.310 and an R-squared of 0.960. Significant contributors to lung cancer incidence included smoking, employment percentage, and other factors. Random forest models were used to interpret the most significant contributing variables.
12	[45]	Li *et al.,* (2022)	Prediction of lung cancer risk in the Chinese population with genetic and environmental factors using extreme gradient boosting.	Extreme Gradient Boosting (XGBoost)	No	Yes	Yes	Yes	P-value, AUC	10-fold cross-validation	A prediction model has been developed for the early detection of lung cancer in the Chinese population.	The model incorporating SNPs and applying XGBoost significantly improved predictive performance, particularly for the squamous cell carcinoma (SCC) subtype.

ACC: Accuracy; PRE: Precision; REC: Recall; AUC: Area Under the Curve; PA: Performance Accuracy; N/A: Not Available; LASSO: Least Absolute Shrinkage and Selection Operator; LogReg: Logistic Regression; XGBoost: Extreme Gradient Boosting; Naive: Naive Bayes + SSA; LinReg+: Linear Regression; SVR: Support Vector Regression; K-NN: K-Nearest Neighbors; Cubist: Cubist Model; PSMC6: Proteasome 26S Subunit, ATPase 6; SMOX: Spermine Oxidase; SMS: Spermine Synthase; RotForest: Rotation Forest; LogR: Logistic Regression; DT: Decision Tree; RF, RF2: Random Forest; SVM: Support Vector Machines; MLP: Multilayer Perceptron; RandSub: Random Subspace; SMO: Sequential Minimal Optimization; CAL: Calibration Curve; DCA: Decision Curve Analysis; C-index: Concordance Index; MAPE: Mean Absolute Percentage Error; RMSE: Root Mean Square Error; ROC: Receiver Operating Characteristic.

Table I provides a summary of the included studies, focusing on various factors influencing prognosis, including environmental factors, genetic factors, and predictive elements. It categorizes the studies based on these factors and includes information such as the author(s) and year, title, machine learning methodology used, measures evaluated, validation methods, findings, and results. The table enhances the organization of information across different studies, providing a comprehensive overview of the research landscape in this specific area.

A. Findings (Table I)

The studies included in the document cover a wide range of factors influencing prognosis, including environmental factors, genetic factors, and risk predictive elements using machine learning models across various research fields.

1. Manuck [1] examined various conceptual models related to gene-environment interactions in psychology, highlighting controversy and differing opinions on the significance of these interactions. The study suggested that the concentration of ambient radon and PM2.5 may have a synergistic influence on health consequences [39].
2. The development of a risk score model based on gene signatures linked to polyamine metabolism aimed to predict lung adenocarcinoma (LUAD) patient prognosis and immunotherapy response [44]. This study identified distinct patient subgroups, correlated gene expression profiles, and evaluated responses to immunotherapy [44]. Additionally, the study explored the use of deep learning architectures, random forests, and support vector machines in biomarker identification, gene association studies, and predictive modeling for lung cancer, with a focus on combining multi-omics data for comprehensive understanding [41].
3. An improved use of advanced machine learning techniques to analyze and evaluate genetic datasets for early lung cancer prediction was demonstrated, focusing on identifying key genes associated with disease susceptibility [42]. Utilizing genetic interactions in disease risk prediction through machine learning techniques helps uncover undiscovered relationships between genomic loci, thereby increasing prediction accuracy and enhancing understanding of disease networks [43].
4. Wang [44] focused on polyamine metabolism-associated gene signatures, offering insights into personalized treatment strategies for lung adenocarcinoma patients. The development of a prognostic model for early lung cancer detection in the Chinese population showed significant improvement in predictive performance, particularly for the squamous cell carcinoma subtype, emphasizing the importance of genetic-environment interactions [45].
5. The association between lung cancer incidence rates and environmental risks was explored using various machine learning algorithms. The Cubist model tree provided the optimal model, exhibiting an RMSE of 3.31 and an R-squared of 0.96. A risk score model based on 14 gene signatures associated with polyamine metabolism demonstrated significant prognostic value for determining prognosis and responsiveness to immunotherapy in lung cancer patients [44].
6. The association between lung cancer incidence rates and environmental risks showed high statistical measures, indicating accurate lung cancer risk prediction through machine learning techniques [44].

B. Findings (Table II)

The findings highlight the complexity of factors influencing prognosis and the effectiveness of using machine learning in predictive modeling. The research underscores the importance of integrating genetic, environmental, and prognostic elements for a comprehensive understanding and prediction of outcomes in lung cancer. Machine learning methods have been widely employed to explore predictive and prognostic elements in lung cancer. In this research, the Rotation Forest model achieved an impressive AUC of 0.993, with high accuracy, precision, recall, and confidence interval scores, demonstrating the effectiveness of machine learning. One notable study on lung adenocarcinoma utilized LASSO Cox regression to identify a 16-ASE signature, achieving good results with an average confidence interval of 0.789 and an AUC exceeding 0.7 in fivefold cross-validation.

The study conducted within the Chinese population employed logistic regression and Extreme Gradient Boosting (XGBoost) to predict lung cancer risk, incorporating 61 SNPs and achieving an average ρ-value of <0.05 with an AUC of >0.7. Additionally, the Naive Bayes and SSA models demonstrated high accuracy (>0.98) in predicting lung cancer patient survival within five years by utilizing Naive Bayes and SSA biomarker genes. Studies employing diverse algorithms like LASSO Cox regression and XGBoost enhance predictive accuracy and survival prognosis. Environmental risk analysis and gene signature-based models further underscore the interdisciplinary approach, promising tailored interventions and improved patient

outcomes in lung cancer care. These findings collectively emphasize the importance of utilizing machine learning techniques to refine predictive and prognostic models for lung cancer, integrating genetic, environmental, and biomarker data to advance patient outcomes and treatment strategies.

TABLE II OUTCOME OF FINDINGS

Sl. No.	Ref.	Year	Model	Biomarkers	Datasets	Variables	ρ-value	Confidence Interval	PA/AUC
1	[40]	2020	LASSO Cox regression	16-ASE signature	TCGA database, 572 samples of LUADAS dataset, 502 samples of LUAD-AS dataset	ASEs, clinical variables	0.054	0.766-0.812 (Average: 0.789)	AUC > 0.7 (fivefold cross-validation)
2	[45]	2022	Logistic Regression	61 Single nucleotide polymorphisms (SNPs)	Chinese population Genotype Dataset: (974 lung cancer patients and 1004 healthy people)	Age, gender, smoking intensity, smoking duration, family history	0.0253	0.718-0.765 (95%) (Average: 0.7415)	AUC 0.742
			Extreme Gradient Boosting (XGBoost)				<0.001	0.737-0.782 (95%) (Average: 0.7595)	AUC: 0.759
3	[7]	2023	Naive Bayes + SSA	Biomarker genes	Wisconsin Prognostic Lung Cancer	Survival time with lung cancer	High	N/A	ACC: 98.78 PRE: 98.6 REC: 98.4
			Random Forest						ACC: 0.92.8 PRE: 88.2 REC: 93.4
4	[44]	2022	Linear regression, Support vector regression, K-nearest neighbor, Cubist model Random Forest model	CO, NO2, SO2, O3, PM10, VEHICLES, SMOKERS, NO2, EMPLOYED, FACTORIES	Taiwan	20 Risk Factors	N/A	1.68 - 4.62 (Average: 3.15) 95% (CI)	96%
5	[37]	2023	Risk Score	59 polyamines metabolism genes	TCGA; 59 polyamines metabolism genes	Polyamines metabolism genes	< 0.05	N/A	AUC values PSMC6: 0.822 SMOX: 0.802 SMS: 0.818
6	[18]	2022	Rotation Forest	Metabolic markers	309 Public dataset	Lung cancer Risk factors	N/A	N/A	AUC: 99.3% ACC: 97.1% PRE: 97.1% REC: 97.1%
7	[38]	2022	Random Forest	Radiomics features from CT images	346 patients from four centers	16 core radiomics features	<0. 05	N/A	AUCs LR: 0.658 DT: 0.567 RF: 0.880 SVM: 0.765
8	[42]	2018	Multilayer Perceptron,	FABP4, Platelet/endothelial cell adhesion molecule, Four and a half LIM domains1, Amine oxidase, copper containing 3, C-type lectin domain family 3, member B, Selenoprotein P, plasma, 1	7129 genes	gene expression data for the Lung cancer Via Kent Ridge Bio-Medical Dataset Repository	N/A	± 1.96	ACC: 86.67 PRE: 0.8714 REC: 0.8315
			Random Subspace,					± 1.96	ACC: 68.33 PRE: 0.6458 REC: 0.6025
			SMO					± 1.96	ACC: 91.67 PRE: 0.9125 REC: 0.9029

TABLE III STATISTICAL SUMMARY OF OUTCOME

Sl. No.	Models	Year	ρ-value	Confidence Interval	Performance			AUC
					Accuracy	Precision	Recall	
1	LASSO	2020	0.054	0.766-0.812	-	-	-	> 0.70
2	LogReg	2022	0.0253	0.718-0.765	-	-	-	0.742
	XGBoost	2022	<0.001	0.737-0.782	-	-	-	0.759
3	Naive	2023	-	-	0.988	0.986	0.984	-
	RF2	2023	-	-	0.928	0.882	0.934	-
4	LinReg+	2022	-	1.68 - 4.62	-	-	-	0.960
5	PSMC6	2023	< 0.05	-	-	-	-	0.822
	SMOX	2023						0.802
	SMS	2023						0.818
6	RotForest	2022	-	-	0.971	0.971	0.971	0.993
7	LogR	2022	<0. 05	-	-	-	-	0.658
	DT	2022						0.567
	RF	2022						0.880
	SVM	2022						0.765
8	Mpercept	2018	-	-1.96 - 1.96	0.867	0.871	0.832	-
	RandSub	2018	-	-1.96 - 1.96	0.683	0.646	0.603	-
	SMO	2018	-	-1.96 - 1.96	0.917	0.913	0.903	-

The accuracy, precision, and recall values across the models reflect their ability to make correct predictions and minimize false positives and false negatives, with average mean values of 0.892, 0.878, and 0.871, respectively. The models recorded >0.870 (>87%) in the TRI-performance rating and an overall rate of 0.881 (88.1%). Models such as Naive Bayes showed optimal results with regard to accuracy, precision, and recall. Models with lower ρ-values and narrower confidence intervals are more consistent and reliable in their predictions. The confidence intervals and ρ-values provide insights into the statistical significance and reliability of the predictions made by each model.

C. Model Performance Metrics

The models exhibited a wide range of performance metrics, including accuracy, precision, recall, confidence intervals, AUC values, and ρ-values. Models such as Rotation Forest, XGBoost, and Multilayer Perceptron demonstrated high AUC values, indicating their strong predictive capabilities in distinguishing between positive and negative classes.

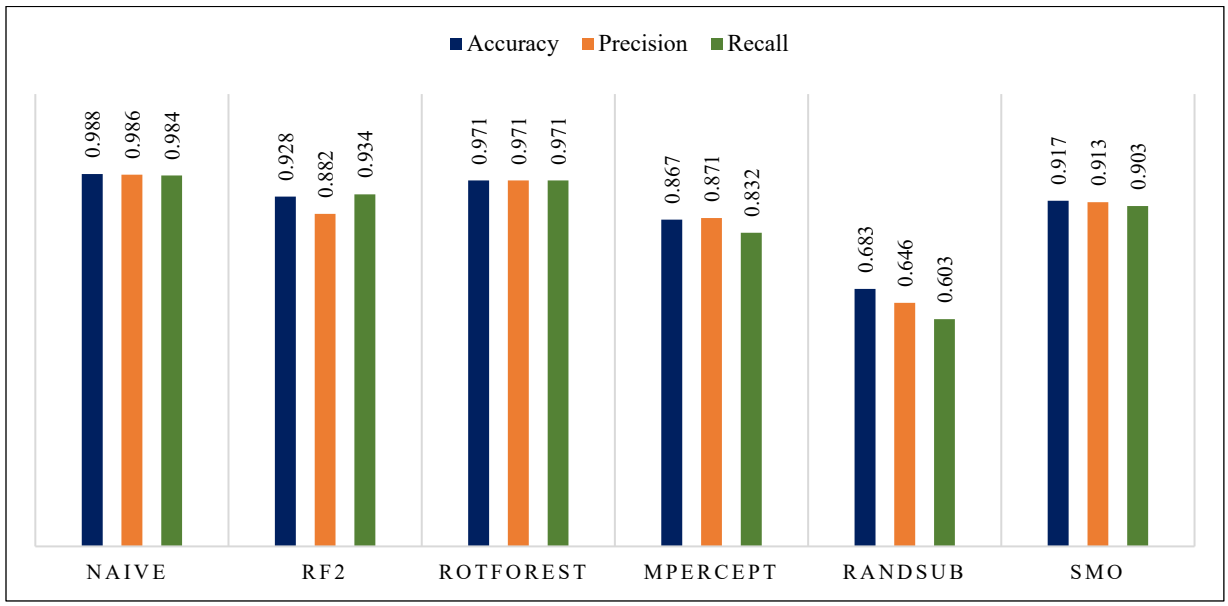

Fig. 2 Performance Accuracy

D. Mean AUC Performance

The twelve AUC datasets are summed up to give an average AUC value, as seen in the calculation below:

Sum of AUC values = (0.70 + 0.742 + 0.759 + 0.960 + 0.822 + 0.802 + 0.818 + 0.993 + 0.658 + 0.567 + 0.880 + 0.765) = 9.466

Number of AUC values (n) = 12

$$\text{Mean AUC} = \frac{\text{Sum of AUC values}}{\text{Number of AUC values}} = \frac{9.466}{12} \approx 0.789$$

The mean AUC of approximately 0.789 suggests that the predictive models evaluated in the research exhibit a moderate to good level of discrimination ability in distinguishing between classes.

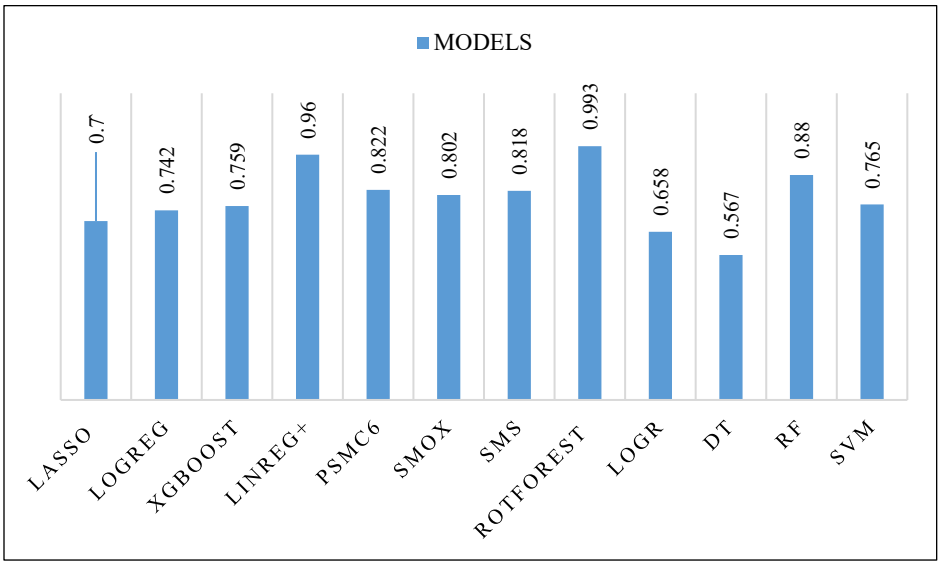

Fig. 3 Area under the curve statistics

This indicates that the models performed reasonably well in terms of predictive accuracy, although individual model performance may vary. Further analysis could explore factors influencing model performance and identify potential areas for improvement.

E. AUC Variability Check using Standard Deviation

(0.700 - 0.789)^2 ≈ 0.007921
(0.742 - 0.789)^2 ≈ 0.002209
(0.759 - 0.789)^2 ≈ 0.000900
(0.960 - 0.789)^2 ≈ 0.029241
(0.822 - 0.789)^2 ≈ 0.001089
(0.802 - 0.789)^2 ≈ 0.000169
(0.818 - 0.789)^2 ≈ 0.000841
(0.993 - 0.789)^2 ≈ 0.041616
(0.658 - 0.789)^2 ≈ 0.017161
(0.567 - 0.789)^2 ≈ 0.049284
(0.880 - 0.789)^2 ≈ 0.008281
(0.765 - 0.789)^2 ≈ 0.000576

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{(X - \bar{X})^2}{n}}$$

Sum of the squared differences ≈ 0.159288
n = 12

$$\text{Variance} \approx \frac{0.159288}{12} \approx 0.013274$$

$$\text{Standard Deviation} \approx \sqrt{0.013274} \approx 0.1152$$

The standard deviation of the AUC values is approximately 0.1152.

The standard deviation of approximately 0.1152 indicates the variability of the AUC values around the mean AUC of 0.789. A smaller standard deviation suggests that the AUC values are closer to the mean, indicating less variability among model performances. Conversely, a larger standard deviation implies greater variability among the AUC values.

In this case, a standard deviation of 0.1152 indicates moderate variability in the AUC values, suggesting that while the mean AUC is relatively stable, there are some differences in the performance of the predictive models. Further analysis is recommended to investigate factors contributing to this variability and explore strategies to improve model consistency.

F. Confidence Intervals and ρ-values

The results from Table III indicate that machine learning models are crucial for predicting lung cancer risk and are highly effective when considering genetic and environmental factors. The high AUC values, accuracy rates, and precision scores of certain models underscore their potential to enhance personalized treatment strategies and clinical decision-making in lung cancer care.

TABLE IV OUTCOME OF P-VALUES

Models	2020	2022	2023
LASSO	0.054		
LogReg		0.0253	
XGBoost		<0.001	
PSMC6			< 0.05
SMOX			< 0.05
SMS			< 0.05
LogR		< 0.05	
DT		< 0.05	
RF		< 0.05	
SVM		< 0.05	

The systematic analysis conducted on various models yielded p-values indicating their statistical significance. Among the models evaluated, LogReg, XGBoost, PSMC6, LogR, DT, RF, SVM, and LogR demonstrated significant evidence to reject the null hypothesis, with p-values less than 0.05. This suggests that these models significantly differ from the null hypothesis and can be effective in their respective tasks. Conversely, the LASSO model did not exhibit statistically significant evidence to reject the null hypothesis, with p-values exceeding 0.05. Consequently, while the LASSO model may still have utility, there is insufficient statistical evidence to conclude that it significantly outperformed the null hypothesis, warranting further investigation or consideration of alternatives.

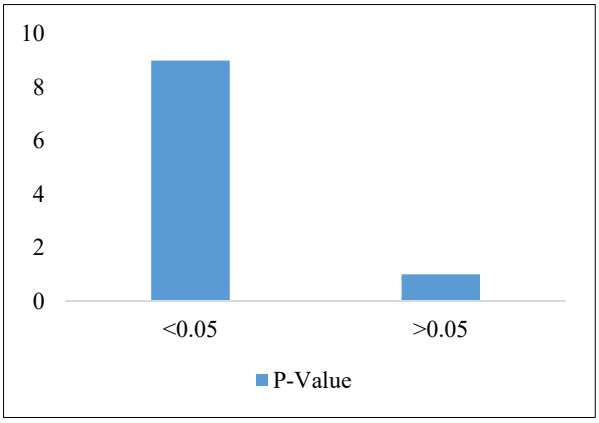

Fig. 3 ρ-Value Scale

Based on the research findings shown in Fig. 3, 9 out of the 10 models exhibit statistically significant results ($p < 0.05$), indicating that they likely have a meaningful impact on predicting lung cancer risk based on environmental and genetic interactions. This suggests promising avenues for further exploration and potential application in clinical settings. However, the LASSO model, with p-values greater than 0.05, may warrant closer examination. Although it did not achieve statistical significance in this study, it does not necessarily imply it is without merit. Further investigation is recommended to determine why it did not perform as expected and whether additional data collection could enhance its predictive accuracy.

IV. DISCUSSION

The extensive utilization of machine learning methods in lung cancer research - focusing on predictive and prognostic elements and applying various models, including Rotation Forest, LASSO Cox regression, logistic regression, XGBoost, Naive Bayes, and SSA - exhibits high accuracy, precision, recall, and AUC values, showcasing their effectiveness in leveraging genetic and environmental factors for predictive modeling. The systematic analysis reveals significant findings, with most models demonstrating statistically significant results ($p < 0.05$) in predicting lung cancer risk based on environmental and genetic interactions, indicating promising potential applications in clinical settings. However, models that did not achieve statistical significance warrant further examination for performance improvement. Additional research is needed to enhance model performance, identify factors influencing variability, and improve predictive accuracy for personalized treatment strategies.

A. General Interpretation of Results

The studies aim to provide valuable insights into various aspects of research and management in the area of lung cancer. From predicting lung cancer incidence rates based on environmental exposures to developing prognostic models for overall survival and identifying candidate genes implicated in lung cancer development, the results offer significant contributions to the field. For example, machine learning models outperform traditional regression models in capturing the mutual influence between environmental exposures and health outcomes.

B. Limitations of Evidence

Regarding the limitations of evidence, the studies acknowledged potential issues such as sample size, data quality, biases in data selection, and the general relevance of findings [37]. These limitations may affect the validity and reliability of the results, emphasizing the need for further empirical analysis using larger and more diverse population samples to ensure the accuracy of the overall results and conclusions drawn from the research.

C. Limitations of Review Processes

The review processes discussed in the document have high coverage, which helps to minimize errors in the review process. However, limitations in comprehensiveness and critical appraisal of the included studies may exist. These limitations could impact the credibility of the review findings and the reliability of the conclusions drawn from the reviewed evidence.

D. Implications for Practice, Policy, and Future Research

The results of the studies suggest significant implications for public health practice, policy interventions, and future

research in the field of lung cancer. Recommendations include considering multiple environmental exposures in assessing lung cancer risk, developing targeted interventions based on machine learning predictions, and exploring gene-environment interactions for personalized treatment strategies. These implications highlight the importance of translating research findings into clinical applications, advancing policy initiatives to support research development, and enhancing patient outcomes through evidence-based interventions.

V. CONCLUSION

Lung cancer risk prediction and treatment using machine learning techniques have made significant strides. The specific objective across these studies is to build a comprehensive understanding of the origins of lung cancer, identify prognostic biomarkers, and develop predictive models for improved patient outcomes. Findings highlight the multifactorial nature of lung cancer, emphasizing the association between environmental exposures, genetic factors, and molecular pathways. Various machine learning models, including LASSO, Logistic Regression, XGBoost, Naive Bayes, and Rotation Forest, were explored. Key findings from the statistical summary and performance accuracy chart include AUC values with an average score of 0.789, a tri-performance rating (accuracy, precision, recall) of 0.881, confidence intervals, and ρ-values. Notably, the Rotation Forest model achieved the highest AUC of 0.993, indicating excellent predictive capabilities. These predictive models demonstrate strong performance in risk stratification and prognostic prediction, with implications for personalized medicine and clinical decision-making. The integration of machine learning approaches with multidimensional data has greatly enhanced our understanding of lung carcinoma and clinical management. This integration paves the way for precision oncology strategies and tailored interventions to improve patient outcomes and reduce the global burden of lung cancer.

REFERENCES

[1] S. B. Manuck and J. M. McCaffery, "Gene-Environment Interaction," *Annu. Rev. Psychol.*, vol. 65, no. 1, pp. 41-70, Jan. 2014, doi: 10.1146/annurev-psych-010213-115100.

[2] G. E. McClearn, "Nature and nurture: Interaction and coaction," *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, vol. 124B, no. 1, pp. 124-130, Jan. 2004, doi: 10.1002/ajmg.b.20044.

[3] C.-H. Yang, Y.-D. Lin, C.-Y. Yen, L.-Y. Chuang, and H.-W. Chang, "A Systematic Gene-Gene and Gene-Environment Interaction Analysis of DNA Repair Genes XRCC1, XRCC2, XRCC3, XRCC4, and Oral Cancer Risk," *Omi. A J. Integr. Biol.*, vol. 19, no. 4, pp. 238-247, Apr. 2015, doi: 10.1089/omi.2014.0121.

[4] G. Vogt, "Environmental Adaptation of Genetically Uniform Organisms with the Help of Epigenetic Mechanisms—An Insightful Perspective on Ecoepigenetics," *Epigenomes*, vol. 7, no. 1, Mar. 2023, doi: 10.3390/epigenomes7010001.

[5] L. M. Hernandez, D. G. Blazer, and A. S. Institute of Medicine (U.S.), *Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate*. National Academies Press, 2006.

[6] G. Sirugo, S. M. Williams, and S. A. Tishkoff, "The Missing Diversity in Human Genetic Studies," *Cell*, vol. 177, no. 1, pp. 26-31, Mar. 2019, doi: 10.1016/j.cell.2019.02.048.

[7] S. Kukreja, M. Sabharwal, M. A. Shah, and D. S. Gill, "A Heuristic Machine Learning-Based Optimization Technique to Predict Lung Cancer Patient Survival," *Comput. Intell. Neurosci.*, vol. 2023, p. 4506488, 2023, doi: 10.1155/2023/4506488.

[8] N. M. Carleton, G. Lee, A. Madabhushi, and R. W. Veltri, "Advances in the Computational and Molecular Understanding of the Prostate Cancer Cell Nucleus," *J. Cell. Biochem.*, vol. 119, no. 9, pp. 7127-7142, Sep. 2018, doi: 10.1002/jcb.27156.

[9] L. Pan et al., "Artificial Intelligence Empowered Digital Health Technologies in Cancer Survivorship Care: A Scoping Review," *Asia-Pacific J. Oncol. Nurs.*, vol. 9, no. 12, p. 100127, Dec. 2022, doi: 10.1016/j.apjon.2022.100127.

[10] A. Choudhary, A. Anand, A. Singh, and P. R., "Machine Learning-Based Ensemble Approach in Prediction of Lung Cancer Predisposition Using XRCC1 Gene Polymorphism," *J. of*, vol. 2023, pp. 1-10, Taylor & Francis, 2023, doi: 10.1080/07391102.2023.2242492.

[11] J. A. Cruz and D. S. W., "Applications of Machine Learning in Cancer Prediction and Prognosis," *Journals SAGEPUB*, [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/117693510600200030.

[12] D. Soldato et al., "The Future of Breast Cancer Research in the Survivorship Field," *Oncol. Ther.*, vol. 11, no. 2, pp. 199-229, Jun. 2023, doi: 10.1007/s40487-023-00225-8.

[13] K. C. Thandra, A. Barsouk, K. Saginala, J. S. Aluru, and A. Barsouk, "Epidemiology of Lung Cancer," *Termedia Publishing House Ltd.*, 2021, doi: 10.5114/wo.2021.103829.

[14] K. Chaitanya Thandra, A. Barsouk, K. Saginala, J. Sukumar Aluru, and A. Barsouk, "Epidemiology of Lung Cancer," *Współczesna Onkol.*, vol. 25, no. 1, pp. 45-52, 2021, doi: 10.5114/wo.2021.103829.

[15] W. A. Cooper, D. C. L. Lam, S. A. O'Toole, and J. D. Minna, "Molecular Biology of Lung Cancer," *J. Thorac. Dis.*, vol. 5 Suppl 5, pp. S479-90, Oct. 2013, doi: 10.3978/j.issn.2072-1439.2013.08.03.

[16] M. Zheng, *Classification and Pathology of Lung Cancer*, W.B. Saunders, 2016, doi: 10.1016/j.soc.2016.02.003.

[17] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, *Global Epidemiology of Lung Cancer*, Ubiquity Press, 2019, doi: 10.5334/aogh.2419.

[18] E. Dritsas and M. Trigka, "Lung Cancer Risk Prediction with Machine Learning Models," *Big Data Cogn. Comput.*, vol. 6, no. 4, p. 139, Nov. 2022, doi: 10.3390/bdcc6040139.

[19] O. Ernest, O. Komolafe, S. O., and A. Oludele, "Ontology: A Case for Disease and Drug Knowledge Discovery," *Commun. Appl. Electron.*, vol. 5, no. 9, pp. 6-13, Sep. 2016, doi: 10.5120/cae2016652362.

[20] A. Shankar et al., "Environmental and Occupational Determinants of Lung Cancer," *Transl. Lung Cancer Res.*, vol. 8, no. Suppl 1, pp. S31-S49, May 2019, doi: 10.21037/tlcr.2019.03.05.

[21] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, "Global Epidemiology of Lung Cancer," *Ann. Glob. Heal.*, vol. 85, no. 1, Jan. 2019, doi: 10.5334/aogh.2419.

[22] N. Ghaffar Nia, E. Kaplanoglu, and A. Nasab, "Evaluation of Artificial Intelligence Techniques in Disease Diagnosis and Prediction," *Discov. Artif. Intell.*, vol. 3, no. 1, p. 5, Jan. 2023, doi: 10.1007/s44163-023-00049-5.

[23] Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, "Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis," *Genomics Proteomics Bioinformatics*, vol. 20, no. 5, pp. 850-866, 2022, doi: 10.1016/j.gpb.2022.11.003.

[24] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine Learning Applications in Cancer Prognosis and Prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8-17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[25] F. Alharbi and A. Vakanski, "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review," *Bioengineering*, vol. 10, no. 2, p. 173, Jan. 2023, doi: 10.3390/bioengineering10020173.

[26] A. A. Adegbenjo *et al.,* "Design and Analysis of an Automated IoT System for Data Flow Optimization in Higher Education Institutions," *J. Eur. des Systèmes Autom.*, vol. 56, no. 5, pp. 889-897, Oct. 2023, doi: 10.18280/jesa.560520.

[27] G. H. M. Sousa, R. A. Gomes, E. O. de Oliveira, and G. H. G. Trossini, "Machine Learning Methods Applied for the Prediction of Biological Activities of Triple Reuptake Inhibitors," *J. Biomol. Struct. Dyn.*, vol. 41, no. 20, pp. 10277-10286, Dec. 2023, doi: 10.1080/07391102.2022.2154269.

[28] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A Review on Deep Learning in Medical Image Analysis," *Int. J. Multimed. Inf. Retr.*, vol. 11, no. 1, pp. 19-38, 2022, doi: 10.1007/s13735-021-00218-1.

[29] S. Bindas and E. Onuiri, "A Deep Learning Approach to Speech Recognition for Detection of Mental Disorders," *Curr. TRENDS Inf. Commun. Technol. Res.*, vol. 2, no. 1, pp. 28-46, 2023, doi: 10.61867/pcub.v2i1a.042.

[30] E. E. Onuiri, O. Akande, O. B. Kalesanwo, T. Adigun, K. Rosanwo, and K. C. Umeaka, "A Systematic Review of Machine Learning Prediction Models for Colorectal Cancer Patient Survival Using Clinical Data and Gene Expression Profiles," *Rev. d'Intelligence Artif.*, vol. 37, no. 5, pp. 1273-1280, 2023, doi: 10.18280/ria.370520.

[31] Z. Sajjadnia, R. Khayami, and M. R. Moosavi, "Preprocessing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services," *Cancer Inform.*, vol. 19, p. 117693 512091795, Jan. 2020, doi: 10.1177/1176935120917955.

[32] H. Mohajan and H. K. Mohajan, "Two Criteria for Good Measurements in Research: Validity and Reliability," *Munich Personal RePEc Archive*, 2017.

[33] R. Caso *et al.,* "The Underlying Tumor Genomics of Predominant Histologic Subtypes in Lung Adenocarcinoma," *J. Thorac. Oncol.*, vol. 15, no. 12, pp. 1844-1856, Dec. 2020, doi: 10.1016/j.jtho.2020.08.005.

[34] Q. Li et al., "Combining Autophagy and Immune Characterizations to Predict Prognosis and Therapeutic Response in Lung Adenocarcinoma," *Front. Immunol.*, vol. 13, 2022, doi: 10.3389/fimmu.2022.944378.

[35] M. Amir-Behghadami and A. Janati, "Population, Intervention, Comparison, Outcomes and Study (PICOS) Design as a Framework to Formulate Eligibility Criteria in Systematic Reviews," *Emerg. Med. J.*, vol. 37, no. 6, pp. 387-387, Jun. 2020, doi: 10.1136/emermed-2020-209567.

[36] M. J. Page *et al.,* "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[37] N. Wang, M. Chai, L. Zhu, J. Liu, C. Yu, and X. Huang, "Development and Validation of Polyamines Metabolism-Associated Gene Signatures to Predict Prognosis and Immunotherapy Response in Lung Adenocarcinoma," *Front. Immunol.*, vol. 14, 2023, doi: 10.3389/fimmu.2023.1070953.

[38] Y. Liu *et al.,* "Development and Validation of Machine Learning Models to Predict Epidermal Growth Factor Receptor Mutation in Non-Small Cell Lung Cancer: A Multi-Center Retrospective Radiomics Study," *Cancer Control*, vol. 29, 2022, doi: 10.1177/10732748221092926.

[39] H. Lee *et al.,* "Evaluating County-Level Lung Cancer Incidence from Environmental Radiation Exposure, PM(2.5), and Other Exposures with Regression and Machine Learning Models," *Environ. Geochem. Health*, vol. 46, no. 3, p. 82, 2024, doi: 10.1007/s10653-023-01820-4.

[40] Q. Cai *et al.,* "Exploration of Predictive and Prognostic Alternative Splicing Signatures in Lung Adenocarcinoma Using Machine Learning Methods," *J. Transl. Med.*, vol. 18, no. 1, 2020, doi: 10.1186/s12967-020-02635-y.

[41] K. M. S. Rani and V. K. Prasad, "Exploring Machine Learning in Lung Cancer: Predictive Modelling, Gene Associations, and Challenges," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 6s, pp. 490-499, 2023, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85167990741&partnerID=40&md5=d96b5427eecc8ea6a8d04218bbf9290c.

[42] J. Pati, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach," *IEEE Access*, vol. 7, pp. 4232-4238, 2019, doi: 10.1109/ACCESS.2018.2886604.

[43] S. Okser, T. P.-B. Mining, "Genetic Variants and Their Interactions in Disease Risk Prediction - Machine Learning and Network Perspectives," *Biodata Mining*, vol. 6, no. 5, 2013, [Online]. Available: https://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-6-5.

[44] K.-M. Wang, K.-H. Chen, C. A. Hernanda, S.-H. Tseng, and K.-J. Wang, "How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine-Learning-Based Modeling and Benchmarking," *Int. J. Environ. Res. Public Health*, vol. 19, no. 14, 2022, doi: 10.3390/ijerph19148445.

[45] Y. Li et al., "Prediction of Lung Cancer Risk in Chinese Population with Genetic-Environment Factor Using Extreme Gradient Boosting," *Cancer Manag. Res.*, vol. 11, no. 23, pp. 4469-4478, 2022, doi: 10.1002/cam4.4800.