JOÃO RAFAEL
VIEIRA GRAÚDO

Machine Learning-Based Assessment of Mutational Profiles in Lung Carcinoma

Master's Dissertation Report in Biomedical Engineering

SUPERVISOR
Professor Miguel Guevara López, PhD
IPS-ESTS

CO-SUPERVISOR
Joana Albuquerque, MD
Hospital da Luz Setúbal e Lisboa

December 2025

JOÃO RAFAEL
VIEIRA GRAÚDO

Machine Learning-Based Assessment of Mutational Profiles in Lung Carcinoma

EXAMINATION BOARD

President: (PhD, Prof. Célio Pina, IPS-ESTS)

Supervisor: (PhD, Prof. Miguel Guevara López, IPS-ESTS)

Co-Supervisor: (MD, Joana Albuquerque, Hospital da Luz – Setúbal e Lisboa)

Examiner: (PhD, Prof. Ana Lima, Universidade do Minho)

December 2025

Ao meu pai,

Que partiu antes de eu começar este caminho, mas cuja presença me acompanhou em cada passo.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the Polytechnic Institute of Setúbal for providing me with the opportunity to undertake this academic journey and for contributing significantly to my growth, both academically and personally.

To my supervisor, Professor Miguel Guevara López, I extend my deepest appreciation for the guidance, constant support, and availability throughout this process. The opportunity to develop this work in the field of machine learning was an enriching challenge and an invaluable learning experience.

To my co-supervisor, Dr. Joana Albuquerque, I am grateful for the dedication, insightful feedback, and continuous support that greatly improved the quality and depth of this work.

To my family - my mother, my sisters, my nephew, and my girlfriend - I owe heartfelt thanks for their unconditional love, encouragement, and understanding, which were essential to the completion of this journey.

To my classmates and friends from the course, thank you for the collaboration, shared knowledge, and moments of companionship that made this path lighter and more enjoyable.

To my closest friends, I express my sincere appreciation for their constant motivation, friendship, and patience throughout this period.

Finally, I would like to express my gratitude to everyone who, directly or indirectly, contributed to the completion of this thesis.

To all, my heartfelt thanks.

ABSTRACT

This study aimed to develop machine learning models for automated prediction of overall survival and mutational status in advanced non-small cell lung cancer (NSCLC) patients with actionable molecular alterations. A retrospective cohort of 275 stage IV NSCLC patients from five hospitals in Southern Portugal (2016–2021) was analysed. Clinical, demographic, molecular, and therapeutic data were integrated into four supervised classification algorithms: Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and AdaBoost. Models were optimized using 10-fold stratified cross-validation with hyperparameter tuning via Grid Search. To address class imbalance, three complementary approaches were implemented: baseline modeling, Synthetic Minority Over-sampling Technique (SMOTE), and synthetic data augmentation using Conditional Tabular Generative Adversarial Network (CTGANSynthesizer). Performance was evaluated using accuracy, precision, recall, F1-score, and Area Under the Curve – Receiver Operating Characteristics (AUC-ROC) metrics.

Results demonstrated that ensemble-based methods (RF, XGBoost, and AdaBoost) substantially outperformed SVM, particularly when trained on balanced datasets. The third approach, incorporating both synthetic data generation and SMOTE oversampling, yielded the highest discriminatory performance, with AdaBoost achieving an AUC-ROC of 0.9217. Correlation analysis revealed that Eastern Cooperative Oncology Group Performance Status (r=0.220), bone metastases (r=0.179), and sex (r=0.159) were the strongest positive predictors of mortality, while Epidermal Growth Factor Receptor (EGFR) exon 19 deletions (r=-0.140) demonstrated the most favorable prognostic association. The most prevalent molecular alteration was Kirsten Rat Sarcoma Virus (KRAS) G12C (35.64%), followed by EGFR mutations (14.91%) and Anaplastic Lymphoma Kinase (ALK) rearrangements (7.27%), consistent with European epidemiological data.

This work demonstrates how machine learning tools can be valuable in predicting survival outcomes and personalizing treatments for patients with advanced NSCLC. Despite these advantages, there are still important challenges, such as the issue of data imbalance and the need to validate models in independent patient groups. Therefore, it is essential to maintain rigorous methodologies throughout the process. In the future, it will be important to test these models in different hospitals, integrate imaging data, and develop decision-support tools that are simple and transparent for healthcare professionals to use in their daily practice.

Keywords: Machine learning, Survival prediction, Lung cancer, Molecular alterations, Precision medicine

RESUMO

Este estudo teve como objetivo desenvolver modelos de aprendizagem automática para a predição automatizada da sobrevivência global e do estado mutacional em doentes com carcinoma do pulmão de não pequenas células (CPNPC) avançado com alterações moleculares acionáveis. Foi analisada uma coorte retrospetiva de 275 doentes com CPNPC em estádio IV provenientes de cinco hospitais do Sul de Portugal (2016–2021). Dados clínicos, demográficos, moleculares e terapêuticos foram integrados em quatro algoritmos de classificação supervisionada: Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost) e AdaBoost. Os modelos foram otimizados utilizando validação cruzada estratificada de 10 folds com afinação de hiperparâmetros através de Grid Search. Para abordar o desequilíbrio de classes, foram implementadas três abordagens complementares: modelação *baseline*, técnica *Synthetic Minority Over-sampling Technique* (SMOTE) e aumento de dados sintéticos utilizando *Conditional Tabular Generative Adversarial Network* (CTGANSynthesizer). O desempenho foi avaliado utilizando as métricas de exatidão, precisão, *recall*, *F1-score* e Área sob a Curva Característica Operacional do Recetor (AUC-ROC).

Os resultados demonstraram que os métodos baseados em ensembles (RF, XGBoost e AdaBoost) superaram substancialmente o SVM, particularmente quando treinados em conjuntos de dados balanceados. A terceira abordagem, incorporando tanto a geração de dados sintéticos como o oversampling com SMOTE, obteve o desempenho discriminativo mais elevado, com o AdaBoost a atingir uma AUC-ROC de 0,9217. A análise de correlação revelou que o Performance Status do *Eastern Cooperative Oncology Group Performance Status* (r=0,220), as metástases ósseas (r=0,179) e o sexo (r=0,159) foram os preditores positivos mais fortes de mortalidade, enquanto as deleções no exão 19 do Receptor do Fator de Crescimento Epidérmico (EGFR) (r=-0,140) demonstraram a associação prognóstica mais favorável. A alteração molecular mais prevalente foi a Kirsten Rat Sarcoma Virus (KRAS) G12C (35,64%), seguida das mutações EGFR (14,91%) e das rearranjos Anaplastic Lymphoma Kinase (ALK) (7,27%), consistente com dados epidemiológicos europeus.

Este trabalho demonstra como as ferramentas de aprendizagem automática podem ser valiosas na predição de resultados de sobrevivência e na personalização de tratamentos para doentes com CPNPC avançado. Apesar destas vantagens, persistem desafios importantes, como o problema do desequilíbrio de dados e a necessidade de validar os modelos em grupos independentes de doentes. Portanto, é essencial manter metodologias rigorosas ao longo de todo o processo. No futuro, será importante testar estes modelos em diferentes hospitais, integrar dados de imagem e desenvolver ferramentas de apoio à decisão simples e transparentes para uso quotidiano pelos profissionais de saúde.

Palavras-Chave: Aprendizagem automática, Previsão de sobrevivência, Cancro do Pulmão, Alterações moleculares, Medicina de precisão

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AJCC	American Joint Committee on Cancer
ALK	Anaplastic Lymphoma Kinase
AUC-ROC	Area Under the Curve – Receiver Operating Characteristics
BRAF	B-Rapidly Accelerated Fibrosarcoma
CNS	Central Nervous System
CTGAN	Conditional Tabular Generative Adversarial Network
ctDNA	Circulating Tumor DNA
CV	Cross-Validation
DRL	Deep Reinforcement Learning
ECOG-PS	Eastern Cooperative Oncology Group Performance Status
EGFR	Epidermal Growth Factor Receptor
EML4-ALK	Echinoderm Microtubule-Associated Protein-Like 4 - ALK
FDA	Food and Drug Administration
FN	False Negatives
FP	False Positives
GAN	Generative Adversarial Network
HER2	Human Epidermal Growth Factor Receptor 2
HT	Hyperparameter Tuning
IPO	Portuguese Institute of Oncology
KRAS	Kirsten Rat Sarcoma Virus
LDCT	Low-Dose Computed Tomography
LUAD	Lung Adenocarcinoma
ML	Machine Learning
MET	Mesenchymal-Epithelial Transition Factor
NELSON	Dutch-Belgian Lung Cancer Screening Trial
NGS	Next-Generation Sequencing
NTRK	Neurotrophic Receptor Tyrosine Kinase

NSCLC	Non-Small Cell Lung Cancer
PD-L1	Programmed Death-Ligand 1
PFS	Progression-Free Survival
RBF	Radial Basis Function
RET	Rearranged During Transfection
RF	Random Forest
ROS1	ROS Proto-Oncogene 1
SBRT	Stereotactic Body Radiotherapy
SCLC	Small-Cell Lung Carcinoma
SDV	Synthetic Data Vault
SHAP	Shapley Additive Explanations
SEM	Standard Error of the Mean
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TKI	Tyrosine Kinase Inhibitor
TN	True Negatives
TNM	Tumor, Node, Metastasis
TP	True Positives
XGBoost	eXtreme Gradient Boosting

CHAPTER 1

INTRODUCTION

Lung cancer represents one of the major causes of morbidity and mortality worldwide, having a significant impact on public health, patients' quality of life and healthcare system sustainability. The histological and molecular heterogeneity of these tumors, especially in non-small cell lung cancer (NSCLC), presents major challenges in diagnosis, treatment and prognostic stratification.

Currently, treatment personalization and the development of precision medicine strategies have become essential, requiring detailed molecular characterization of each tumor, and the identification of potentially actionable genetic alterations that guide therapeutic decisions.

Nevertheless, the routine implementation of molecular testing faces not only logistical and financial constraints, but also time-related limitations, which may delay clinical decision-making and compromise access to targeted therapies.

In this context, artificial intelligence and machine learning techniques emerge as promising tools to overcome these barriers. By integrating clinical, pathological, molecular and, in some cases, imaging data, they can identify hidden patterns associated with specific genetic alterations and tumor characteristics, supporting the early prediction of molecular profiles and enabling better patient selection to the most suitable treatment.

The present study combines approaches from biomedicine and data science, aiming to evaluate, in a national population of NSCLC patients, the utility of machine learning models for automatic prediction of relevant molecular alterations. This approach seeks to deepen understanding of the genetic-clinical determinants of lung cancer.

1.1. Motivation

Understanding the molecular profile of NSCLC and accurately identifying alterations with actionable therapeutic targets is fundamental to advancing in precision medicine. By tailoring treatment strategies to each tumour’s genetic profile, clinicians can maximize therapeutic efficacy, reduce unnecessary toxicity and costs, and ultimately improve patient outcomes.

Beyond its clinical relevance, the characterization of molecular alterations also contributes to the development of real-time registries that monitor the incidence and prevalence of specific mutations. This is particularly relevant in the context of population-based studies,

where the correlation between molecular profiles and clinicopathological characteristics can uncover important epidemiological trends, as well as therapeutic and preventive gaps. In the European population, and particularly in Portuguese population, such data remains limited and fragmented, underscoring the need for systematic approaches to support both research and clinical decision-making.

As molecular testing becomes increasingly central to treatment selection, there is a growing need for efficient, cost-effective, and scalable tools to support the diagnostic workflow. Machine learning techniques offer a powerful opportunity for automate the evaluation of mutational status, accelerate the diagnostic process and expand access to precision oncology, particularly in healthcare settings where molecular testing capacity and turnaround time are limited.

1.2. Objectives

The principal objective of this study is:

- Develop and validate machine learning models to predict overall survival outcomes in patients with advanced NSCLC harbouring actionable molecular alterations, integrating clinical, demographic, pathological, molecular, and therapeutic data.

As well as:

- Characterize the prevalence and epidemiology of targetable molecular alterations in the Portuguese population with advance NSCLC.
- Identify the clinical and pathological characteristics associated with different molecular subtypes.

1.3. Contributions

This dissertation provides the following principal contributions:

- **Integrated Clinical-Molecular Framework**: Analytical protocol integrating clinical, molecular and therapeutic data from 275 patients with Portuguese metastatic NSCLC, characterizing molecular epidemiology (KRAS 35.6%, EGFR 31.0%, ALK 7.3%) and clinical-prognostic correlations (ECOG-PS, bone metastases, sex as main predictors).
- **Machine Learning Optimization for Survival**: Demonstration that data synthesis (CTGANSynthesizer + SMOTE) is critical for balancing, achieving robust discrimination (AUC-ROC 0.9217, balanced sensitivity 80-89%) versus models on imbalanced data (AUC-ROC ≈0.5).
- **Exploratory Analysis of Mutational Prediction**: Evidence that isolated clinical data are inadequate for reliable prediction of mutational status, reaffirming need for direct molecular testing.

- **Methodological Contribution**: Reusable template for machine learning community in biomedicine facing extreme imbalance in clinical datasets.

1.4. Dissertation Structure

This dissertation is organized into five chapters, each contributing to the development and contextualization of the research problem:

- **Chapter 1 – Introduction**: This chapter, introduces the research topic, presents the motivation and relevance of the study, and outlines the main research objectives.
- **Chapter 2 – Literature Review**: Provides an overview of lung cancer subtypes and molecular characteristics, discusses conventional and automated methods for mutation detection, and reviews recent advances in the application of machine learning techniques in oncology.
- **Chapter 3 – Materials and Methods**: Describes the data sources used, the preprocessing steps applied, and the machine learning models implemented. This chapter also details the experimental design and evaluation metrics.
- **Chapter 4 – Results and Discussion**: Presents and interprets the results obtained from the application of the machine learning models, including performance metrics and comparisons across different configurations and data types.
- **Chapter 5 – Conclusion and Future Work**: Summarizes the main findings, reinforces the relevance of the work, and suggests directions for future research in this area.

CHAPTER 2

LITERATURE REVIEW

This chapter presents a review of the relevant literature, addressing the main subtypes of lung cancer, the most frequently mutated genes and recent approaches based on machine learning. Furthermore, it discusses key scientific contributions in the field.

2.1. Lung Cancer: Clinical and Molecular Background

Lung cancer remains one of the leading causes of cancer-related mortality worldwide and represents a major public health challenge. In 2022, GLOBOCAN[1] estimates indicated approximately 2.2 million new cases of lung cancer with 1.8 million related deaths, underscoring the continued global burden of this disease. With survival rates showing significant improvement but still challenging outcomes, it continues to be associated with variable prognosis depending on stage at diagnosis, molecular characteristics, and treatment accessibility [1], [2].

Histologically, lung cancer is broadly classified into small-cell lung carcinoma (SCLC) and NSCLC, with NSCLC accounting for approximately 80-85% of all diagnosed cases. NSCLC encompasses several distinct subtypes, most notably adenocarcinoma (representing 60-70% of NSCLC) and squamous cell carcinoma (25-30% of NSCLC), with large cell carcinoma comprising a smaller proportion. Recent epidemiological data indicate a sustained increase in the incidence of lung cancer, with a marked rise in adenocarcinoma cases among non-smokers and younger individuals [2].

The etiology of lung cancer remains multifactorial, with tobacco identified as the major risk factor, accounting for 80-90% of cases. Nevertheless, approximately 20% of lung cancer cases occur in never-smokers, highlighting the importance of other etiological factors including environmental exposures other than tobacco, occupational carcinogens and genetic predisposition. NSCLC exhibits substantial molecular and histological heterogeneity, even within identical histological subtypes, with distinct genomic landscapes that have profound implications for therapeutic approaches [3].

Among the most relevant molecular alterations in NSCLC include Kirsten Rat Sarcoma Virus (KRAS) mutations (30-40% of adenocarcinomas, the most frequently occurring mutation), Epidermal Growth Factor Receptor (EGFR) gene, identified in about 10-25% of lung adenocarcinomas, and rearrangements in the Anaplastic Lymphoma Kinase (ALK) gene, present

[1] https://gco.iarc.who.int/media/globocan/factsheets/cancers/15-trachea-bronchus-and-lung-fact-sheet.pdf

in 3-7% of cases. These driver alterations are more commonly observed in non-smokers, females, and patients with adenocarcinoma histology. Their detection is essential since they are associated with marked responses to targeted therapies, which have led to substantial improvements in survival compared to traditional cytotoxic chemotherapy [4].

Advances in understanding tumor biology and molecular mechanisms underlying lung carcinogenesis have revolutionized therapeutic strategies, particularly through the identification of actionable driver alterations and the development of precision medicine approaches. The introduction of tyrosine kinase inhibitors (TKIs) for patients with EGFR mutations or ALK rearrangements has transformed the natural history of advanced NSCLC, with median overall survival of 36 months for EGFR – FLAURA trial and 47 months for FLAURA2, for ALK can go up to >60 months [5]. Despite the availability of multiple treatment modalities, including surgical resection, chemotherapy, radiotherapy, targeted therapy, and immunotherapy; outcomes remain significantly influenced by stage at diagnosis, with approximately 75% of patients presenting with advanced disease [6].

In parallel, immunotherapy has emerged as a standard of care for patients with advanced NSCLC lacking actionable mutations, providing improved survival for a subset of patients, though the overall prognosis for stage IV disease remains poor, with five-year survival rates typically below 5% in the general population. Thus, the integration of molecular diagnostics and the rational use of novel therapies are now central to the management and prognostication of NSCLC, highlighting the importance of an individualized treatment approach [7].

2.2. Key Genetic Alterations in NSCLC

NSCLC encompasses a molecularly heterogeneous group of malignancies with distinct mutational landscapes that differ significantly between histological subtypes. Recent comprehensive genomic analyses have revealed that adenocarcinoma and squamous cell carcinoma should be treated as separate diseases due to their divergent molecular characteristics. However, emerging evidence indicates that the biological diversity within NSCLC subtypes extends well beyond histology, with molecular profiling now recognized as the key determinant of prognosis and therapeutic response [8].

Genome-wide studies have shown that genomic and transcriptomic differences between adenocarcinomas and squamous cell carcinomas remain profound, with specific patterns of gene amplification, mutation, and pathway activation. For instance, squamous cell carcinoma frequently harbors 3q amplification, particularly involving the 3q26-3q28 locus, which is rare in adenocarcinoma. Conversely, adenocarcinomas display higher rates of targetable driver mutations, such as EGFR, ALK, KRAS, and others.

More importantly, molecular characterization increasingly supersedes histological classification in guiding patient management. Recent advances in precision medicine demonstrate that molecular features - rather than histological subtype alone - dictate treatment options and expected outcomes. For example, within adenocarcinomas, a tumor harboring an

ALK rearrangement behaves distinctly from one with a KRAS mutation, both biologically and clinically. This paradigm shift underscores that, although histologic subtypes possess characteristic genetic and epigenetic traits, individual tumors must be stratified based on their unique molecular signatures, not just their microscopic appearance.

Thus, the field is moving towards an individualized, biomarker-centered approach to NSCLC, where personalized medicine based on actionable genomic alterations offers the potential for tailored therapy and improved prognosis, regardless of the broad histological classification [9], [10].

2.2.1. Subtypes of Non-Small Cell Lung Carcinoma

- **Adenocarcinoma**: This represents the most prevalent NSCLC subtype, particularly among non-smokers and younger patients. It frequently harbors actionable driver mutations including EGFR (found in 10-15% of Western patients, up to 50% in Asian patients), ALK (3-7% of cases), ROS Proto-Oncogene 1 (ROS1) (1-2% of cases), and KRAS (25-30% of cases). Comprehensive molecular profiling studies, especially in never-smokers, detect driver mutations in up to 70-95% of patients, with an actionable mutation rate around 40-78% depending on population and testing strategy. These proportions are dynamic and likely to increase as next-generation sequencing expands the panel of testable mutations (e.g., Mesenchymal-Epithelial Transition Factor (MET), Rearranged During Transfection (RET), Neurotrophic Receptor Tyrosine Kinase (NTRK), B-Rapidly Accelerated Fibrosarcoma (BRAF)). The rise in molecular diagnostics is driving ongoing development of new targeted therapies, so the percentage of patients eligible for these approaches continues to grow, making the precise figure a moving target [11]. These molecular alterations make adenocarcinoma highly amenable to targeted therapeutic interventions [12].
- **Squamous Cell Carcinoma**: Strongly associated with tobacco exposure, this subtype demonstrates a complex mutational landscape with high tumor mutational burden. It is characterized by frequent alterations in TP53, PIK3CA, NFE2L2, and SOX2 pathways. Unlike adenocarcinoma, squamous cell carcinoma lacks well-defined targetable driver mutations, making it more challenging to treat with precision therapies [12].
- **Large Cell Carcinoma**: Representing the smallest proportion of NSCLC cases (2-3%) [13], this undifferentiated subtype often presents at advanced stages with poor clinical outcomes. Recent studies have identified unique biomarkers such as USP7 overexpression associated with poor prognosis [14].

2.2.2. Clinical Importance of Molecular Profiling

The integration of molecular profiling has transformed NSCLC management through precision oncology approaches. Current guidelines recommend testing for multiple biomarkers including EGFR, ALK, ROS1, Human Epidermal Growth Factor Receptor 2 (HER2), BRAF, KRAS, MET, RET, and NTRK alterations [4].

- **Targeted Therapies**: Detection of driver mutations enables application of specific targeted agents, with response rates that can exceed 70% in biomarker-positive populations, depending on the target mutation. ALK-rearranged NSCLC patients receiving next-generation ALK inhibitors (like alectinib or lorlatinib) also show markedly improved long-term survival, with 5-year survival estimates > 50% in some series [4].
  It is important to note that these survival rates vary significantly among different mutations and targeted treatments. For instance, KRAS G12C inhibitors, while promising, have shown more modest improvements in median progression-free survival and overall survival compared to EGFR or ALK inhibitors, and many other targetable alterations (e.g., ROS1, MET, RET) have emerging but less mature survival data. Hence, survival outcomes with targeted therapy are highly mutation and drug-specific, requiring precise molecular diagnosis and personalized treatment strategies [5].
- **Personalized Treatment Approaches**: Comprehensive genomic profiling combined with circulating tumor Deoxyribonucleic Acid (ctDNA) monitoring can support individualized treatment regimens, optimizing therapeutic efficacy while enabling early detection of resistance mechanisms. Recent advances include ultra-short circulating tumor Deoxyribonucleic Acid detection, which increases ctDNA targets by 1.6-fold, potentially improving sensitivity [15].

Despite these advancements, significant challenges remain in the clinical management of NSCLC. Late-stage diagnosis, tumor heterogeneity, and the emergence of resistance to targeted therapies underscore the complexity of this disease. Ongoing research is essential to further elucidate molecular mechanisms and to develop novel therapeutic strategies aimed at improving patient outcomes in NSCLC.

2.3. Lung Cancer Situation in Portugal

Lung cancer represents a significant public health challenge in Portugal, ranking as the fourth most prevalent cancer in the country, with approximately 6,000 new cases diagnosed annually. The mortality associated with this malignancy remains elevated, with lung cancer representing the leading cause of cancer-related death in men and the third leading cause in women, after breast cancer and colorectal cancer [16].

According to GLOBOCAN 2022 data, 4,253 new cases were registered in men and 1,092 new cases in women, reflecting a significant sex-based disproportion with a male-to-female ratio of 3.3:1, substantially higher than the global ratio of 2.1:1. Historical data from 2018 documented 5,284 new lung cancer cases with 4,671 related deaths, with projections indicating a 21.2% increase in new cases and 24.5% increase in deaths by 2040 [17], [18].

2.3.1. Histological Subtypes

According to recent epidemiological data from Portugal, NSCLC accounts for 80-90% of all lung cancer cases. The histological distribution highlights adenocarcinoma as the most prevalent

subtype (40.8%), followed by squamous cell carcinoma (22.7%), with large cell carcinoma comprising a smaller fraction of cases. Notably, there are significant sex-related differences: adenocarcinoma constitutes 57.0% of cases in women versus 35.8% in men, while squamous cell carcinoma is more common among men (26.8%) compared to women (9.3%). These disparities are attributed largely to differences in tobacco exposure, with men historically having a higher smoking burden. Furthermore, the incidence peak of lung cancer occurs later in women and continues to rise, reflecting the later onset of smoking habits in this group [16], [17].

2.3.2. Geographic Distribution

Epidemiological analysis of lung cancer in Portugal reveals markedly heterogeneous geographic patterns, with identification of high and low incidence clusters. The Azores demonstrate the highest incidence rate at 56.8 per 100,000 inhabitants, followed by the Lisbon region accounting for 29.1% of total national cases. The Northern region contributes 38.0% of total national cases. In contrast, the Central region demonstrates the lowest incidence rate at 25.6 per 100,000 inhabitants.

Spatial analysis identified four high-incidence lung cancer clusters, predominantly located in urban areas (Porto and Lisbon), and four low-incidence clusters in predominantly rural areas. These patterns suggest the influence of urban risk factors, historical industrialization, and differences in healthcare access [17].

Additionally, an ecological study conducted in Northern Portugal demonstrated a correlation between indoor radon exposure and lung cancer incidence rates in the region, highlighting radon as a significant risk factor for lung cancer in the Portuguese population. A multicentric case-control study including hospitals from Portugal and Spain showed that higher residential radon concentrations were associated with increased risk of small cell lung cancer, especially in heavy smokers, with a dose-response relationship. These findings emphasize the importance of radon as a major, preventable lung cancer risk factor in Portugal [19].

2.3.3. Stage at Diagnosis and Prognosis

According to clinical data from 2012-2016 from the Portuguese Institute of Oncology (IPO) Porto, the distribution by stage at diagnosis reveals:

- Locally advanced stage (IIIB): 23.6% of cases [18]
- Metastatic stage (IV): 58.4% of cases [18]

This high proportion of patients diagnosed with advanced disease (approximately 82%) in Portugal contrasts markedly with countries implementing lung cancer screening programs using low-dose computed tomography (LDCT). Randomized controlled trials such as the US National Lung Screening Trial (NLST) and the Dutch-Belgian Lung Cancer Screening Trial (NELSON) have demonstrated significant mortality reductions of approximately 20% and 24%, respectively, with LDCT screening compared to controls. These programs not only reduce lung

cancer mortality but also significantly increase the detection of early-stage disease, thereby decreasing the proportion of stage IV diagnoses by up to 40-50% [20].

Correspondingly, the prognosis of lung cancer patients improves with early detection, as reflected in stage-stratified survival data demonstrating median survival times extending beyond previously reported values in Portugal:

- Stage IIIB: Median survival around 16.7 months [20]
- Stage IV: Median survival about 9.8 months. [20]

Portugal currently lacks a nationwide lung cancer screening program, which substantially limits early detection and contributes to poor overall outcomes. Furthermore, a considerable proportion of patients (24.2%) with advanced NSCLC receive only supportive care due to comorbidities or performance status constraints, highlighting barriers related to healthcare access and socioeconomic factors [18].

2.4. Machine Learning Algorithms

Machine learning algorithms represent the fundamental building blocks of artificial intelligence systems, enabling computational models to extract patterns from data and make intelligent decisions without explicit programming. These algorithms are systematically categorized into supervised, unsupervised, semi-supervised, and reinforcement learning paradigms, each addressing distinct computational challenges and application domains [21]. Recent advances have demonstrated that the optimal algorithm selection depends critically on data characteristics, computational constraints, scalability requirements, and domain-specific performance metrics. Modern implementations increasingly emphasize the integration of multiple learning paradigms, with hybrid approaches showing superior performance across diverse applications [22].

Categories of Machine Learning Algorithms

- **Supervised Learning**: Encompasses algorithms including linear and logistic regression, decision trees, support vector machines, and advanced ensemble methods, trained on labeled datasets to predict outcomes or classify data instances. Recent developments emphasize the integration of Bayesian optimization for hyperparameter tuning and the incorporation of explainable Artificial Intelligence (AI) frameworks to enhance model interpretability [23].
- **Unsupervised Learning**: Includes advanced clustering techniques, dimensionality reduction methods such as Principal Component Analysis (PCA), and modern self-supervised learning approaches that identify latent patterns in unlabeled data without predefined target variables. Current research demonstrates significant advances in self-supervised learning, which has emerged as a critical subset capable of learning discriminative features from unlabeled data [24].

- **Semi-Supervised Learning**: It is situated between supervised and unsupervised learning that uses a combination of labeled and unlabeled data. The main goal of this approach is to leverage the large amount of information available in unlabeled data to create a better and more accurate model than what would be possible using only the few labeled data available [25].
- **Reinforcement Learning**: Focuses on sequential decision-making through trial-and-error interactions with dynamic environments, utilizing reward signals to optimize long-term behavioral strategies. Recent comprehensive surveys highlight the evolution from foundational tabular methods to sophisticated Deep Reinforcement Learning (DRL) techniques, with particular emphasis on scalability and sample efficiency [26].

2.4.1. Supervised Learning

Supervised learning involves training predictive models on datasets where each instance comprises input features paired with corresponding ground-truth labels. The fundamental objective is to learn generalizable mapping functions that accurately predict outputs for previously unseen data instances. This learning paradigm addresses two primary problem categories:

- **Classification**: Discrete categorical prediction tasks, including binary classification (spam detection) and multi-class problems (image recognition, medical diagnosis). Recent applications demonstrate exceptional performance in medical diagnostics, with accuracy rates exceeding 98% in specialized domains [27].
- **Regression**: Continuous value prediction problems, encompassing financial forecasting, real estate valuation, and scientific parameter estimation. Modern implementations achieve significant improvements through the integration of contextual open data and advanced optimization techniques [28].

Contemporary supervised learning evaluation emphasizes comprehensive performance assessment using cross-validation techniques, with metrics tailored to specific task requirements including precision, recall, F1-score, and domain-specific measures [27].

2.4.2. Support Vector Machines

Support Vector Machines were introduced by Cortes and Vapnik in 1995 as part of the broader theoretical framework of statistical learning theory proposed by Vapnik in 1998. The core principle of SVM is to separate classes by identifying the decision boundary that maximizes the margin between them. This large-margin principle is strongly linked to generalization performance, which explains why SVMs often achieve robust results on unseen data [29].

One of the key innovations of SVM is the use of kernel functions, which make it possible to model non-linear relationships by projecting data into higher-dimensional feature spaces without explicit transformation. Schölkopf and Smola (2002) describe in detail in their book "Learning with Kernels" the theoretical and practical foundations of SVMs and related kernel

methods. The most used kernels include linear, polynomial, and radial basis function (RBF) kernels, each suited for different problem structures [30].

SVM is particularly effective in high-dimensional spaces and in cases where the number of features exceeds the number of samples, such as in text classification and genomic studies. It is also robust to overfit when parameters are properly tuned. However, limitations include high computational cost on large datasets, sensitivity to kernel and parameter choices, and the absence of native probabilistic outputs, which often require calibration, for example, Platt scaling [29].

Applications of SVM include handwriting recognition, bioinformatics tasks such as protein classification, and document categorization. Despite newer algorithms gaining popularity, SVM remains a valuable method in domains with high-dimensional but relatively small datasets [29].

2.4.3. Random Forests

Random Forests were introduced by Breiman in 2001 as an extension of decision tree ensembles. The algorithm constructs multiple decision trees; each trained on a bootstrap sample of the dataset. At each node, only a random subset of features is considered for splitting, which increases diversity among the trees. The final prediction is made by averaging predictions (in regression) or by majority vote (in classification).

They offer numerous advantages. Are robust to overfitting, relatively easy to tune, and capable of handling both numerical and categorical data. They also provide internal performance estimates through out-of-bag errors and can produce measures of feature importance, which is useful for exploratory data analysis [31].

Despite these strengths, Random Forests can be computationally demanding when the number of trees is large, and their interpretability is more limited compared to simpler models. Additionally, traditional feature importance metrics can be biased toward variables with more categories or higher variance, as demonstrated by Strobl et al. (2007) [32].

Applications of Random Forests are broad, including disease prediction in healthcare, ecological modeling for species distribution, and financial applications such as credit risk assessment. Their combination of robustness and accuracy has made them a strong baseline model in many applied research contexts [31].

2.4.4. AdaBoost

Adaptive Boosting, or AdaBoost, was introduced by Freund and Schapire in 1997 as one of the first widely adopted boosting algorithms. Its principle is to combine multiple weak learners into a single strong learner by training them sequentially. Each new model focuses more on instances that previous models misclassified, thereby improving overall accuracy.

The theoretical appeal of AdaBoost comes from its close connection to margin theory, similar in spirit to SVM. Research has shown that AdaBoost tends to improve generalization by

increasing the margins of training examples. Its simplicity and relatively small number of hyperparameters make it easy to implement and effective across diverse tasks.

Nevertheless, AdaBoost has important weaknesses. Because it increases the weight of misclassified samples, it is highly sensitive to noise and outliers, which can harm generalization. It can also overfit when complex base learners are used [33].

Historically, AdaBoost played a transformative role in computer vision, particularly in the Viola-Jones face detection framework, which demonstrated its ability to perform real-time object detection with limited computational resources. Today, although it has been largely surpassed by gradient boosting methods, AdaBoost remains influential as the conceptual foundation of modern boosting algorithms, as highlighted by Schapire and Freund (2012) in their book "Boosting: Foundations and Algorithms" [34].

2.4.5. Extreme Gradient Boosting

Developed by Chen and Guestrin in 2016, is an efficient and scalable implementation of gradient boosting machines originally introduced by Friedman in 2001. Like AdaBoost, XGBoost builds ensembles sequentially, but instead of focusing solely on misclassified instances, it minimizes a specified loss function by fitting trees to the residual errors of previous models.

XGBoost introduced several innovations that contributed to its success. These include explicit regularization to prevent overfitting, shrinkage (a learning rate that scales the contribution of each tree), subsampling of features and data to reduce variance, and parallelized tree construction for scalability. Together, these features make XGBoost not only highly accurate but also computationally efficient on very large datasets [35].

The strengths of XGBoost are well documented. It consistently outperforms many competitors on structured/tabular data and has dominated predictive modeling competitions such as Kaggle. It can naturally handle missing values, allows flexible loss function specification, and scales effectively to millions of samples. Its weaknesses include high model complexity, a large hyperparameter space, and reduced interpretability compared to simpler models. However, modern interpretability frameworks such as Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) have made progress in addressing this limitation [36].

XGBoost has found applications in marketing (customer churn prediction), finance (credit risk modeling), medicine (disease prognosis), and industrial systems (fault detection). Its performance and flexibility have made it a standard in applied machine learning for tabular data [35].

2.4.6. Comparative Discussion

Each of the four models reviewed here represents a significant milestone in the evolution of supervised learning. SVMs emphasize theoretical rigor and perform especially well in high-dimensional but small datasets. AdaBoost introduced a simple but powerful framework for combining weak learners, though its sensitivity to noise limits its use in practice. Random Forests

provide a balance between accuracy, robustness, and ease of use, making them a strong default option for many tabular tasks. Finally, XGBoost represents the current state of the art in gradient boosting methods, achieving high predictive accuracy and scalability at the cost of interpretability.

From the perspective of bias and variance, SVMs and AdaBoost are effective at reducing bias but may struggle with variance in noisy datasets. Random Forests reduce variance by averaging many diverse trees, while XGBoost achieves both low bias and controlled variance through regularization and shrinkage, as described by Friedman (2001).

In terms of interpretability, SVMs with linear kernels and AdaBoost with simple stumps are relatively transparent, whereas Random Forests provide partial interpretability through feature importance measures. XGBoost is the least interpretable of the four, though modern interpretability frameworks like SHAP have made progress in addressing this limitation.

Finally, scalability varies significantly. While SVMs struggle with very large datasets, Random Forests and XGBoost scale much better, with XGBoost being particularly suited to large-scale industrial applications. AdaBoost is efficient for small and medium-sized datasets but less so for big data contexts.

2.5. Evaluation Metrics

The evaluation of machine learning models represents a cornerstone in the development of effective intelligent systems. Evaluation metrics provide a rigorous quantitative basis for assessing algorithmic performance, thereby enabling objective comparisons across diverse methodological approaches.

In this study, five metrics were selected - Precision, Accuracy, Recall, AUC-ROC, and F1-Score, as presented in Figure 2.1 - based on both the specific characteristics of the classification task under consideration and the evaluation requirements of the project. This methodological choice ensures a multidimensional assessment of model performance, capturing distinct facets of predictive quality and establishing a robust foundation for the interpretation of the empirical results.

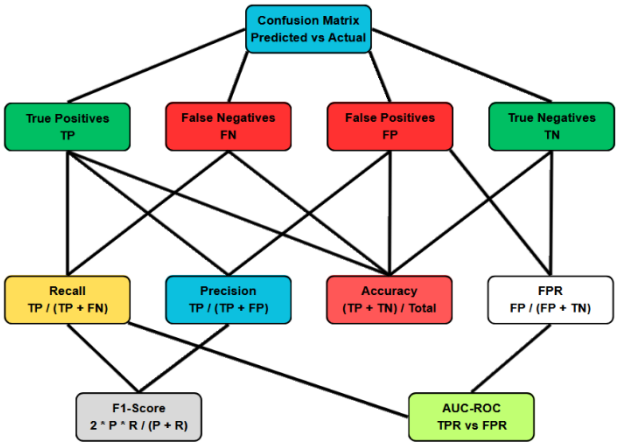

Figure 2.1 - Performance Metrics Framework Derived from Confusion Matrix.

The confusion matrix constitutes the mathematical foundation for deriving binary classification metrics utilized in this project. This two-dimensional matrix organizes model predictions into four fundamental categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The mathematical structure of this matrix enables precise quantification of different types of classification errors, providing a robust statistical basis for the implemented evaluation [37].

Precision represents the proportion of instances correctly classified as positive among all positive predictions made by the developed model. Mathematically, it is defined as shown in Equation 2.1:

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

This metric was selected due to its relevance in evaluating the model's ability to minimize false positives, a critical aspect in the specific context of the developed application. Precision addresses the fundamental question: "What is the probability that a positive prediction is correct?" being particularly valuable when costs associated with false positives are significantly high [38].

Accuracy quantifies the overall proportion of correct classifications relative to the total instances evaluated in the test set. Its mathematical formulation according to Equation 2.2 is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

Despite presenting limitations in imbalanced datasets, accuracy was included as a reference metric due to its intuitive interpretation and widespread use in literature, enabling direct comparisons with other studies [39].

Recall measures the model's ability to correctly identify all positive instances present in the test dataset. Its mathematical definition represented in Equation 2.3 is:

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

This metric was considered essential for the project due to the importance of minimizing false negatives in the context of the developed application, where non-detection of positive cases may carry significant consequences [39].

AUC represents the area under the ROC (Receiver Operating Characteristic) curve, offering an integral assessment of the developed model's performance, independent of the chosen classification threshold. This metric was selected for its ability to provide an aggregate measure of performance across different decision thresholds [40].

AUC value interpretation in the context of this project follows established conventions [41]:

- AUC = 0.5: Random performance
- AUC ∈ [0.7, 0.8]: Acceptable performance
- AUC ∈ [0.8, 0.9]: Satisfactory performance
- AUC > 0.9: Excellent performance

The F1 Score constitutes the harmonic mean between precision and recall, providing a unified metric that balances both components in evaluating the implemented model. Its mathematical formulation, in Equation 2.4, is [42]:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.4}$$

This metric was included for its relevance in evaluating models in contexts where both precision and recall are equally important, providing a balanced measure of overall performance [43].

Cross-validation (CV) addresses a critical challenge in predictive modeling: evaluating model performance when test data is unavailable or limited. The technique operates by dividing the dataset into multiple subsets, iteratively using different portions for training and validation to obtain a comprehensive assessment of model accuracy. This approach serves three primary purposes: estimating prediction errors, tuning hyperparameters, and selecting the most suitable model among competing alternatives [44].

The fundamental goal of cross-validation is to provide an unbiased estimate of a model's generalization performance - that is, its ability to make accurate predictions on unseen data. By repeatedly training and testing the model on different data partitions, cross-validation reduces the variance associated with a single train-test split and provides more reliable performance estimates [45].

K-fold cross-validation represents the most widely adopted cross-validation strategy. The methodology involves partitioning the dataset into k equally sized subsets or "folds". The model training process then proceeds iteratively: in each of the k iterations, one-fold serves as the validation set while the remaining k-1 folds constitute the training set. This process continues until every fold has been used exactly once as the validation set.

The final performance metric is computed as the average of the k individual validation scores, providing a robust estimate of model accuracy. Common choices for k include 5 or 10, though the optimal value depends on dataset characteristics and computational constraints. When k equals the number of samples in the dataset, the method is termed leave-one-out cross-validation (LOOCV), where each individual observation serves as the validation set once.

The k-fold approach offers several advantages over simple holdout validation. First, it maximizes data utilization by ensuring that every observation is used for both training and

validation. Second, it provides a more stable and less variable estimate of model performance compared to a single train-test split. Third, it helps detect overfitting by evaluating the model on multiple independent validation sets [44].

2.6. Related Work

The intersection of genomic data analysis and machine learning has revolutionized survival prediction in lung carcinoma patients, generating substantial research focused on identifying prognostic biomarkers and developing predictive models. This comprehensive review examines recent advances in predicting both survival status (alive/dead) and survival time using genomic features and supervised machine learning approaches.

Random Forest has emerged as one of the most successful algorithms for lung cancer survival prediction. Yang et al. developed prediction models for recurrence and survivability in lung adenocarcinoma (LUAD) and squamous cell carcinoma using The Cancer Genome Atlas (TCGA) data, integrating genomic, clinical, and demographic data with copy number variation and mutation information from 15 selected genes. Their RF-based approach demonstrated superior performance in personalizing treatment decisions [46].

Chen et al. conducted a large cohort study identifying a novel prognosis prediction model for LUAD through machine learning strategies, where they initially used three distinct algorithms (sigFeature, random forest, and univariate Cox regression) to evaluate each gene's prognostic relevance. After 100,000 iterations of model construction, they successfully built a 16-gene-based prediction model capable of classifying LUAD patients into high-risk and low-risk groups [47].

A comprehensive meta-analysis by researchers comparing machine learning models for lung cancer survival prognostication showed that machine learning models achieved a C-statistic of 0.78 compared to 0.70 for traditional logistic regression models, demonstrating superior discriminative ability [48].

In 2023, Kang et al. proposed a Generative Adversarial Network (GAN) based method to generate synthetic data, using a divide-and-conquer strategy to preserve logical relationships between variables. It was validated on three clinical datasets (NSCLC, breast cancer, and diabetes). The presented method showed consistent improvements in classification model performance, particularly in AUC-ROC, while reducing limitations such as data scarcity and class imbalance, thus facilitating the application of ML in clinical contexts [49].

CHAPTER 3

MATERIALS AND METHODS

Chapter 3 outlines the methodological foundation of this work. It details the characteristics of the dataset, the design of the framework developed to address the research goals, the preprocessing procedures applied to the data, and the machine learning models implemented throughout the study.

A raw dataset comprising genetic mutation profiles of lung carcinoma patients, along with their corresponding clinical outcomes in terms of survival, was compiled for this study. To improve the robustness of the analysis, in collaboration with the clinician, the most discriminative clinical, molecular, and pathological features were selected, discarding the less informative ones. Subsequently, a set of machine learning algorithms was employed to model the relationship between the mutational landscape and patient outcomes. Hyperparameter tuning (HT) was conducted to optimize the performance of the algorithms and ensure the most suitable configuration for predictive accuracy. The trained models were then applied to unseen mutation data to estimate survival probability and classify patients into alive or death groups. Finally, the predictive results were systematically examined to extract meaningful insights into the association between specific genetic alterations and survival outcomes in lung carcinoma. The proposed pipeline is represented in Figure 3.1.

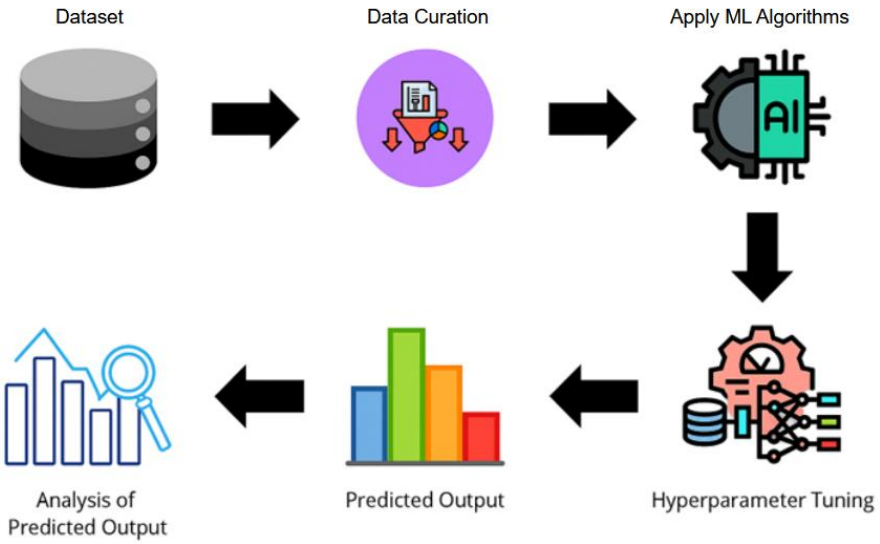

Figure 3.1 - Integrated Machine Learning Workflow for Prognostic Assessment.

3.1. Dataset

Data on the molecular profile of metastatic NSCLC patients were retrospectively collected from 5 hospitals in the South of Portugal, including patients diagnosed between January 2016 and July 2021 with a targetable alteration detected in next-generation sequencing (NGS) or equivalent method. Patients with incomplete clinical records or without histological confirmation were excluded from the analysis.

In addition to the identification of key target mutations (EGFR, ALK, KRAS, BRAF, among others), a broad set of demographic and clinical variables was recorded. These encompassed age at diagnosis, sex, Eastern Cooperative Oncology Group Performance Status (ECOG-PS), smoking history, tumor histology, clinical stage, metastatic sites, and PD-L1 expression, as well as treatment regimens administered and corresponding outcomes.

To guarantee the reliability and integrity of the dataset, a comprehensive data curation procedure was implemented. This process involved systematic data cleaning, identification and management of missing values, detection of inconsistencies and potential entry errors, feature selection and cross-validation against the original medical record.

The dataset included these variables:

a. Patient Demographics

   - Age At Diagnosis: Age at the time of lung cancer diagnosis;
   - Sex: Patient Sex;
   - ECOG-PS: Eastern Cooperative Oncologic Group - Performance Status;
   - Smoking history.

b. Disease Characteristics

   - Date Of Diagnosis: Date of initial diagnosis;
   - Histology: Histological subtype (adenocarcinoma, squamous cell carcinoma, or other rare subtypes);
   - Stage: Clinical stage (according to the American Joint Committee on Cancer – Tumor Node Metastasis (AJCC TNM) classification valid at the time of diagnosis);
   - Burden of Metastatic Disease: Site of metastases (lung, central nervous system, adrenal glands, bone, liver, others);

- PD-L1 Expression: PD-L1 expression assessed by immunohistochemistry (0%, ≥1%, or >50%).

c. Primary Tumor Treatment Characteristics

- Type of treatment: (surgery; surgery plus (neo)adjuvant chemotherapy; chemoradiotherapy; chemoradiotherapy plus immunotherapy; radiotherapy; surgery plus adjuvant osimertinib; other);

- Treatment Discontinuation: Reason for treatment discontinuation (completion of planned therapy, disease progression, toxicity, secondary malignancy, death, other).

d. Central Nervous System (CNS) Disease Treatment

- Type of treatment: (surgery, stereotactic body radiotherapy [SBRT], whole-brain radiotherapy, systemic therapy, other);

- DateCNSProg: Date of CNS progression.

e. First-Line Treatment in Metastatic Disease

- TypeTreatment_1L: Therapeutic regimen used (chemotherapy, immunotherapy, targeted therapy, radiotherapy, surgery, Palliative Care);

- Start and end dates of treatment;

- Treatment response: Treatment response (complete response, partial response, stable disease, or disease progression);

- Discontinuation: Reason for discontinuation (completion of planned therapy, progression, toxicity, secondary malignancy, or death);

- Date of progression.

f. Subsequent Lines of Treatment (Second- and Third-Line Metastatic Therapy)

- TherapyReg: Therapeutic regimen (chemotherapy, immunotherapy, targeted therapy, other);

- Start and end dates of treatment;

- Treatment response: Treatment response (complete response, partial response, stable disease, or disease progression);

- Discontinuation: Reason for discontinuation (progression, toxicity, secondary malignancy, or death);
- Date of progression;
- Date of last follow-up;
- Status: Survival status (alive or deceased);
- Date of death/last follow-up.

3.2. Design and Implemented Strategies

The methodological design of this study was structured around an integrated ML pipeline, conceived to predict overall survival and, in an exploratory context, molecular profile in patients with stage IV NSCLC harbouring molecular alterations. The development of this pipeline involved four fundamental strategic dimensions [50]:

- Definition of a data preprocessing and feature representation strategy;
- Structured selection and optimization of supervised learning algorithms;
- Systematic design of mechanisms to address class imbalance;
- Establishment of an evaluation framework grounded in multiple performance metrics appropriate for clinical applications.

3.2.1. Modeling Framework

The overall framework, represented in Figure 3.2., was conceived as a binary supervised classification problem in which each patient is represented by a vector of clinical, pathological, molecular, and therapeutic features, and the target variable corresponds to survival status (alive vs. deceased) or, exploratorily, to the predominant actionable mutation category. This framework was deliberately designed to reflect clinical decision-making: rather than focusing on a single data domain, the models integrate heterogeneous information that is realistically available in oncology practice.

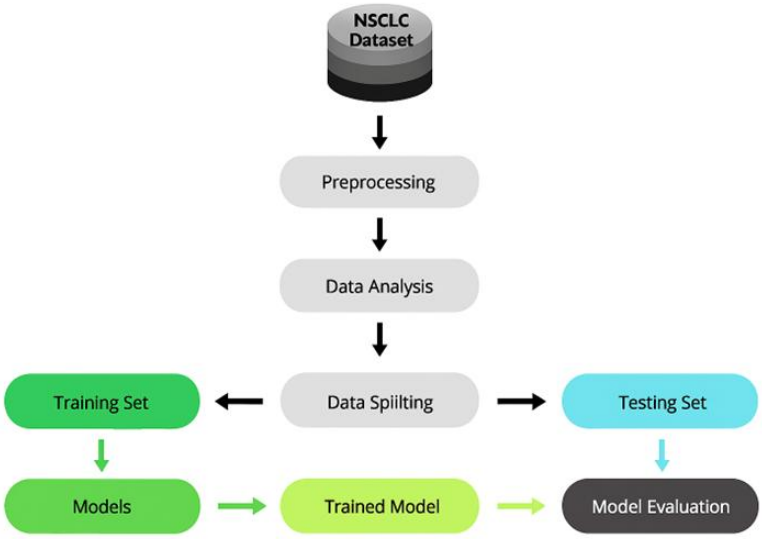

Figure 3.2 - Framework of Data Preprocessing, Analysis, and Model Evaluation

The dataset was divided into training and test subsets using a stratified 80/20 split, ensuring preservation of the original class distribution. Within the training subset, models were developed, optimized, and internally validated using stratified 10-fold cross-validation. The independent test subset was reserved for final performance evaluation, thus providing an unbiased estimate of generalization to unseen patients.

3.2.2. Data Preprocessing and Feature Representation

The preprocessing strategy was defined beforehand to ensure methodological consistency across all models. Continuous variables (e.g., age at diagnosis) and ordinal clinical scales (e.g., ECOG-PS) were standardized using z-score normalization (StandardScaler), ensuring that each feature presented zero mean and unit variance. This step was considered vital given the heterogeneity of variable scales, particularly for algorithms sensitive to feature scaling such as SVM.

Categorical variables (e.g., histology, smoking history, treatment type) were transformed into numerical representations compatible with the learning algorithms. Nominal variables were encoded as integers (LabelEncoder). Variables with extensive missing data or limited clinical relevance were excluded during a preliminary feature selection phase, while key prognostic factors identified in the literature (e.g., burden of metastatic disease, and PD-L1 expression) were retained to ensure that the models remained aligned with current clinical knowledge.

This preprocessing pipeline was implemented in a modular manner so that the same sequence of transformations could be consistently applied to both training and test data, therefore preventing information leakage and ensuring full reproducibility of the analyses.

3.2.3. Model Selection

Four supervised classification algorithms were selected to model survival and molecular profile: SVM, RF, XGBoost, and AdaBoost. These models were chosen for their complementary properties and extensive validation in structured clinical datasets.

- SVM was included as a large-margin classifier particularly suited to high-dimensional spaces and relatively small sample sizes;
- Random Forest represents a robust, non-parametric ensemble method capable of modeling complex non-linear interactions;
- XGBoost was selected as a gradient boosting implementation, recognized for its high predictive performance on tabular data and ability to handle heterogeneous feature types;
- AdaBoost was incorporated as a boosting algorithm with a distinct learning dynamic, focusing sequentially on misclassified instances and thus providing a useful point of comparison with both Random Forest and XGBoost.

Hyperparameter optimization for each model was performed using grid search embedded in a stratified 10-fold cross-validation scheme, with F1-score defined as the primary optimization metric. This procedure ensured systematic exploration of the hyperparameter space and reduced the risk that identified differences between algorithms would be driven by sub-optimal configurations rather than by intrinsic model capabilities.

3.2.4. Strategies to Adress Class Imbalance

An important consideration in this work was the pronounced class imbalance observed in the survival outcome, with a clear predominance of deceased patients (n=232) over survivors (n=43), as well as in several mutational categories. To address this issue, three complementary strategies were implemented and compared:

In the first approach (baseline), the models were trained on the original, imbalanced dataset, serving as a reference that reflects the natural distribution of outcomes but is known to bias learning towards the majority class.

The SMOTE technique was applied in the second approach, exclusively to the training folds, generating synthetic instances of the minority class through interpolation between nearest neighbours. This strategy aimed to improve the models' ability to detect the minority class while preserving the overall structure of the feature space.

Finally, in the third approach, SMOTE was combined with data augmentation using Conditional Tabular Generative Adversarial Network (CTGANSynthesizer). For each class, a CTGAN-based synthesizer was trained to learn the conditional probability distribution, considering both numerical and categorical variables and enforcing clinically admissible value ranges. Synthetic samples were then generated to equalize the size of all classes, thereby producing fully balanced training datasets. This combined strategy was designed to test whether

modelling a balanced and augmented distribution of patients could improve the discriminative performance of the algorithms without compromising clinical validity.

By comparing these three approaches across all algorithms, the study evaluates the impact of different class imbalance mitigation strategies on survival prediction and molecular profile classification.

3.2.5. Performance Evaluation

Model performance was evaluated using a set of complementary metrics derived from the confusion matrix: accuracy, precision, recall, and F1-score, together with AUC-ROC. Given the asymmetric clinical consequences of incorrectly classifying the minority class, particular emphasis was placed on recall and F1-score, which more adequately capture the balance between sensitivity and precision in imbalanced settings. ROC curves and confusion matrices were additionally inspected to provide more granular understanding of error patterns and trade-offs between sensitivity and specificity.

3.3. Preprocessing

The initial stage of dataset preprocessing involved several systematic steps to ensure data integrity, consistency, and preparation for analysis. The original dataset, stored in Comma-Separated Values (CSV) format, was imported into a Pandas DataFrame. Initially it contained 69 columns and 529 rows encompassing clinical, demographic, and molecular data from lung carcinoma patients. Each variable was examined to determine its data type (integer, floating-point, or object) and level of data integrity.

To streamline the dataset, several columns that stored numeric information as object types, including age at diagnosis, smoking history, ECOG-PS, clinical stage (ESTADIOCAT), histology, and number of metastatic sites, were converted to floating-point values. Any formatting inconsistencies or conversion errors were handled through coercion of invalid entries into missing values (NaN), thus ensuring consistency and reliability for subsequent analyses.

Patterns with data missing were carefully reviewed, and the dataset was improved by retaining only variables with adequate data coverage (more than 340 valid entries per column). Columns with sparse data were removed to improve analytical rigor. Clinically significant date variables were standardized to a uniform datetime format, allowing the derivation of new time-based metrics, including overall survival and follow-up time, which were added as additional variables.

Rows with critical missing values were excluded, resulting in a clean dataset ready for modeling and analysis.

Following this initial phase of data cleaning and structuring, attention turned to exploratory variable analysis and preparatory transformations. The intermediate dataset comprised 48

columns and 275 patient records, encompassing demographic, clinical, molecular, and therapeutic information.

Finally, exploratory data analysis was performed using frequency counts for key variables, including gender, smoking history, ECOG-PS, histology subtype, number of genetic alterations, PD-L1 expression, number and location of metastatic disease, and survival status.

3.4. Classification Models and Technical Implementation

This research employed supervised classification algorithms implemented in Python (Anaconda Inc.), leveraging the Scikit-Learn library and the imbalanced-learn package to address dataset imbalances through SMOTE augmentation. The project implementation was based on a set of libraries: Pandas for data manipulation, NumPy for numerical operations, Scikit-learn for algorithm implementation, imbalanced-learn (imblearn) for SMOTE, Synthetic Data Vault (SDV) for CTGANSynthesizer, and Matplotlib for visualization.

3.4.1. Pipeline Architecture and Data Flow

Each model follows a structured pipeline architecture that chains transformations sequentially: raw data pass through the StandardScaler (normalization), then through SMOTE, applied only to training data, and finally reach the classifier. This design ensures that all transformations are applied consistently and reproducibly, using ImbPipeline to correctly integrate SMOTE into the cross-validation workflow, thereby preventing information leakage.

The dataset was first divided into training and test subsets using a stratified 80/20 (train/test) split to preserve class proportions. To address the imbalance in the target classes, SMOTE was applied exclusively to the training subset, generating synthetic samples for the minority class and thus improving model training.

Model performance was evaluated on the test set using multiple metrics: accuracy, precision, recall, F1-score, and AUC-ROC. Confusion matrices were also plotted to visually assess classification errors and true positive rates.

3.4.2. Hyperparameter Optimization via Grid Search

To optimize hyperparameters, both Grid Search and cross-validation were employed with a stratified 10-fold scheme, optimizing the F1-score to balance the precision-recall trade-off. Parameter grids included variations in the regularization and kernel parameters for SVM, number of estimators and tree depth for RF and XGBoost and learning rates and estimators for AdaBoost.

In the SVM model, it was used three parameters for HT, as presented in Table 3.1.

Table 3.1 - Hyperparameter optimization parameters (Grid Search) for the Support Vector Machine algorithm

Parameter	Description	Values Used
C	Controls the trade-off between maximizing the margin and minimizing classification errors	0.1, 1, 10, 100
gamma	Defines the influence range of each training point	‘scale’, ‘auto’, 0.1, 0.01, 0.001
kernel	Defines the function mapping data to higher dimensions	‘rbf’, ‘poly’

Table 3.2 presents the five parameters used for HT in the RF model.

Table 3.2 - Hyperparameter optimization parameters (Grid Search) for the Random Forest algorithm

Parameter	Description	Values Used
n_estimators	Number of trees in the forest	50, 100, 200, 300
max_depth	Maximum depth each tree can grow	None, 5, 10, 15, 20
min_samples_split	Minimum samples required to split a node	2, 5, 10
min_samples_leaf	Minimum samples required in a leaf node	1, 2, 4
max_features	Number of features considered for splits	‘sqrt’, ‘log2’

The XGBoost model employed three parameters for HT, as shown in Table 3.3.

Table 3.3 - Hyperparameter optimization parameters (Grid Search) for the XGBoost algorithm

Parameter	Description	Values Used
n_estimators	Number of boosting rounds (trees) to build	100, 200, 300
max_depth	Maximum depth of each tree	3, 5, 7, 9
learning_rate	Controls the contribution of each tree	0.01, 0.05, 0.1, 0.2
min_child_weight	Minimum sum of instance weights in a child node	1, 3, 5
subsample	Fraction of training data used for each tree	0.6, 0.8, 1.0

Two parameters for HT were utilized in the AdaBoost model, as demonstrated in Table 3.4.

Table 3.4 - Hyperparameter optimization parameters (Grid Search) for the AdaBoost algorithm

Parameter	Description	Values Used
n_estimators	Number of boosting iterations	50, 100, 200, 300
learning_rate	Controls the contribution of each weak learner	0.5, 1.0, 1.5
algorithm	Boosting strategy	‘SAMME’, ‘SAMME.R’

3.4.3. Synthetic Data Generation for Class Balancing

For the exploratory analysis of mutational status prediction, a categorical variable named "MutationalStatus" was created in the dataset. In the original dataset, all mutations were coded in binary format (0/1) in individual columns (EGFR, KRAS, ALK, BRAF, ROS1, NTRK, MET, HER2, RET, and others). However, multiclass classification requires a single categorical variable as target, consolidating information dispersed across multiple columns.

To address this requirement, was implemented the following strategy. For each patient, the primary mutation was identified following a clinical priority hierarchy:

1. EGFR ; 2. KRAS ; 3. ALK ; 4. Other (BRAF, ROS1, NTRK, MET, HER2, RET)

This hierarchy reflects current clinical practice, where therapeutic selection prioritizes the most clinically relevant mutation in cases of multiple genetic alterations.

Application of this procedure to N=275 patients with valid mutational status data produced the following distribution:

- EGFR: 105 patients (38.4%);
- KRAS: 90 patients (32.8%);
- ALK: 19 patients (6.9%);
- Other: 60 patients (21.9%).

One patient presented no recorded mutation and was therefore excluded from this analysis.

All remaining mutations not explicitly included in the hierarchy (BRAF, ROS1, NTRK, MET, HER2, RET, and rare variants) were consolidated in the "Other" category, resulting in a categorical variable suitable for multiclass modeling. This approach consolidates dispersed binary information into a single target variable, enabling multiclass classification. However, it implies loss of information regarding co-mutations.

The exploratory mutational status prediction analysis employed identical methodology to the primary survival analysis, differing only in the target variable definition. Specifically, same feature

set, preprocessing, class imbalance correction strategies (two-stage approach), cross-validation (10-fold stratified), classification algorithms (SVM, RF, XGBoost, AdaBoost), hyperparameter tuning (Grid Search with identical parameter spaces), and evaluation metrics.

Class imbalance requires data augmentation strategies. Synthetic samples were generated from both the minority and majority classes to achieve equal numbers of instances for both. CTGAN was employed, specifically through the CTGANSynthesizer from the SDV library. CTGANSynthesizer was selected because it learns the joint distribution of features while preserving correlations between variables; automatically handles mixed data types (continuous and categorical); and imposes data integrity constraints ensuring that values remain within specified ranges.

3.4.3.1. Procedure Used for Creation of Synthetic Data

The first step involved stratifying the dataset by the target variable. For the exploratory analysis, the EGFR class contained N=105 patients, the KRAS class contained N=90 patients, the ALK class contained N=19 patients, and the Other class contained N=60 patients. Each class was processed independently, enabling selective augmentation of minority classes.

In the second step, for each class, metadata was automatically detected from the data through a procedure that identifies data types (integer, floating-point, categorical) for each column, statistical properties (minimum, maximum, mean, standard deviation) for numerical variables, and unique values with their frequencies for categorical variables.

In step three, the synthesizer was initialized with the following hyperparameters and constraints: enforce_min_max_values set to True to ensure that generated values remain within the ranges of each variable; enforce_rounding set to True to ensure that if original values are integers without decimal places, synthetic data maintains this precision; epochs set to 300 representing the number of training iterations; and verbose set to True to monitor data convergence. The enforce_min_max_values=True constraint is critical for medical data, avoiding generation of biologically implausible values (for example, age greater than 120 years) where out-of-range values have no clinical interpretation.

Subsequently, the CTGANSynthesizer was fitted to each class's data through a procedure that trains the neural network to learn the conditional probability distribution of each class, with training converging when discriminator and generator losses stabilize.

Following training, synthetic samples were generated with specific quantities to achieve multiclass balance. Specifically, 45 synthetic samples were generated for EGFR (105 original + 45 synthetic = 150 total), 60 synthetic samples for KRAS (90 original + 60 synthetic = 150 total), 131 synthetic samples for ALK (19 original + 131 synthetic = 150 total), and 90 synthetic samples for Other (60 original + 90 synthetic = 150 total). Generated samples had the target class label re-appended.

Original and synthetic data were concatenated into a single augmented matrix.

Finally, validation procedures were performed to ensure quality of synthetic data. In the range verification validation, for each numerical feature, the ranges of synthetic values were compared to original values by checking whether the minimum and maximum of synthetic data remained within [minimum_original, maximum_original]. In the descriptive statistics comparison validation, descriptive statistics (mean, median, standard deviation, minimum, maximum, quartiles) of synthetic data were compared to original data.

For the primary analysis of survival outcome prediction, an equivalent methodology to that described for the exploratory analysis was applied, differing only in the target variable definition. The extreme imbalance of the minority class (only 43 alive patients out of 275) required data augmentation strategies. Synthetic samples were generated from both classes (deceased and alive) to achieve equal numbers of instances.

43 synthetic samples were generated for class 1 (Deceased: 232 original + 43 synthetic = 275 total) and 232 synthetic samples were generated for class 0 (Alive: 43 original + 232 synthetic = 275 total), totalling 550 data instances.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the results of statistical analysis and predictive classification models developed to predict survival outcomes in patients with NSCLC with actionable molecular alteration using machine learning algorithms. The statistical analysis involved a comprehensive exploration of relationships between survival outcomes (alive vs. deceased) and a range of clinical, demographic, and pathological variables. This analysis applied descriptive statistics, correlation analysis, and classification modeling techniques to elucidate patterns and associations within the data that determine NSCLC prognosis.

For developing predictive models of survival, four machine learning algorithms were utilized: Support Vector Machines, Random Forest, XGBoost, and AdaBoost. The algorithms were trained on a subset of data and subsequently evaluated for their predictive capacity on independent test samples, utilizing 10-fold cross-validation. Performance of the models was evaluated using multiple classification metrics appropriate for clinical prognosis prediction.

The results of the developed classification models are presented in detail in this chapter, including: (i) sample characterization and outcome distribution; (ii) comparative performance of predictive models through confusion matrices and ROC curves; (iii) quantification of discriminative capacity of each algorithm; and (iv) identification of significant prognostic factors associated with survival in NSCLC.

4.1. Statistical Analysis

In the final analysis, shown on Figure 4.1, 275 patients with advanced NSCLC and a targetable molecular alteration were included, predominantly male with a history of smoking (n=186, 67.6%). This reflect the traditional epidemiology and risk factors associated with NSCLC, although, in recent years an increase in incidence has been among women and non-smokers.

Variables	Total Dataset (n = 275)
Age at Diagnosis	65.5 ± 11.35
Sex	
Male	165 (60 %)
Female	110 (40 %)
Smoking History	
Yes	186 (67.64 %)
No	89 (32.36 %)
PS ECOG	
0	75 (27.27 %)
1	131 (47.64 %)
2	55 (20 %)
3	12 (4.36 %)
4	2 (0.73 %)
Histology	
Adenocarcinoma	253 (92 %)
Squamous Cell Carcinoma	10 (3.64 %)
Adenosquamous	4 (1.45 %)
Other	8 (2.91 %)
Number of Genes with Alterations	
1	220 (80 %)
2	45 (16.36 %)
3	8 (2.91 %)
4	1 (0.36 %)
9	1 (0.36 %)
PDL1 Expression	
< 1%	105 (38.18 %)
1% - 49%	57 (20.73 %)
≥ 50%	68 (24.73 %)
Unknown	45 (16.36 %)

Figure 4.1 - Demographic and clinicopathological distribution of the study cohort (n=275 patients with stage IV NSCLC)

Performance Status was assessed according to the ECOG scale. Most patients present preserved or mildly impaired functional status (n=206, 74.9%), with 75 patients (27.3%) classified as ECOG-PS of - 0 and 131 patients (47.6%) as ECOG-PS of - 1. A further 55 patients (20.0%) exhibited ECOG-PS 2, while 12 (4.4%) and 2 (0.7%) patients were classified as ECOG-PS 3 and ECOG-PS 4, respectively. These findings indicate that most patients maintained a good general condition at the time of diagnosis, which may represent a favorable prognostic factor for treatment response and overall outcomes.

Histologically, there was a clear predominance of adenocarcinoma, identified in 253 patients (92.0%), followed by squamous cell carcinoma in 10 patients (3.6%), other histological types in 8 patients (2.9%), and large cell or unspecified carcinoma in 4 patients (1.5%). These are consistent with the epidemiological profile of NSCLC, in which adenocarcinoma is the most frequent histological subtype overall, particularly among patient harboring targetable molecular alterations.

Regarding molecular status, most patients present a single gene alteration, (n=220, 80.0%), suggesting a dominant oncogenic driver in most of the cases. In the other patients, 45 (16.36%) show alterations in two genes and 8 (2.91%) in three genes, one (0.36%) in 4 genes, and another (0.36%) in 9 mutated genes.

PD-L1 expression has a heterogeneous distribution, with 105 patients (38.18%) with PD-L1 <1%, 57 patients (20.73%) with expression between 1-49%, 68 patients (24.73%) with PD-L1 ≥50%, and 45 patients (16.36%) with unknown PD-L1 expression. This stratification assumes particular importance for guiding first-line treatment, especially in patients without target mutations or less common target mutations, where immunotherapy with anti-PD1/PDL1 and anti-CTL4 can be done in combination with chemotherapy or in monotherapy if PD-L1 ≥5.

The burden of metastatic disease in this cohort reveals a spectrum of disease dissemination (Figure 4.2): 36.73% of patients presented with metastatic involvement confined to a single site, while 33.45% had metastases in two sites. A smaller but significant proportion exhibited more extensive spread, with 17.09% having three metastatic sites, 8.36% with four sites, and 3.64% presenting metastases in five sites. Only one patient (0.36%) demonstrated metastatic disease in six distinct sites, and a single case had no documented distant metastases, potentially representing locally advanced unresectable disease. All the data are represented in Figure 4.2.

Variables	Total Dataset (n = 275)
Burden of Metastatic Disease	
0	1 (0.36 %)
1	101 (36.73 %)
2	92 (33.45 %)
3	47 (17.09 %)
4	23 (8.36 %)
5	10 (3.64 %)
6	1 (0.36 %)
MTxLymphNodes	71 (25.82 %)
MTxCentralNervousSystem	78 (28.36 %)
MTxSupraRenal	41 (14.91 %)
MTxLiver	41 (14.91 %)
MTxBone	114 (41.45 %)
MTxPleural	86 (31.27%)
MTxLung	105 (38.18 %)

Figure 4.2 - Distribution of metastases and metastatic burden in the cohort.

The distribution of metastatic sites are in line with well-described hematogenous dissemination patterns typical of NSCLC. Bone is the most frequently affected site, involved in 41.45% of patients, followed by pulmonary metastases (38.18%) and pleural involvement (31.27%). Central nervous system (CNS) metastases occurred in 28.36% of patients, while

distant lymph node metastases and metastases to the liver and adrenal glands affected approximately 25.82% and 14.91% of patients, respectively. This pattern holds important prognostic and therapeutic implications.

Emerging evidence suggests that metastatic patterns may differ according to molecular tumor profiles, with certain driver mutations conferring specific organotropism. For example, tumors harboring EGFR mutations or ALK rearrangements more frequently metastasize to the brain and bone, while KRAS-mutated tumors show a predilection for other metastatic sites. Incorporating molecular data into the analysis of metastatic burden not only refines prognostication but also potentially guides the application of site-directed therapies and surveillance strategies.

Detailed analysis of specific molecular alterations (Figure 4.3) shows that KRAS mutation constitute the most frequent oncogenic driver in this cohort, present in 98 patients (35.64%). This prevalence aligns well with published literature reporting KRAS mutation rates ranging from approximately 25% to 35% in European NSCLC populations. ALK rearrangements were identified in 20 patients (7.27%), consistent with the generally reported prevalence of 3-7% worldwide. BRAF mutations (4.36%), MET alterations (4.73%), HER2 amplifications or mutations (4.73%), RET fusions (1.45%), and ROS1 rearrangements (1.82%) were within expected ranges. No NTRK fusions were detected, as expected given their rarity (<1%). Additionally, 56 patients (20.36%) harbored other molecular alterations, reflecting the characteristic heterogeneity of NSCLC [51].

Variable	Total dataset (n=275)
KRAS G12C	98 (35.64 %)
EGFR	
EGFR_del_ex19	37 (13.45 %)
EGFR_L858R_ex21	41 (14.91 %)
EGFR_T790M	1 (0.36 %)
EGFR_insex20	4 (1.45 %)
EGFR_ex21	5 (1.82 %)
EGFR_ex20	7 (2.55 %)
EGFR_ex19	1 (0.36 %)
EGFR_ex18	9 (3.27 %)
EGFR_other	7 (2.55 %)
EML4-ALK	20 (7.27 %)
MET_skippingex14	13 (4.73 %)
HER2	13 (4.73 %)
BRAF_V600E	12 (4.36 %)
ROS	5 (1.82 %)
RET	4 (1.45 %)
NTRK	0 (0%)
Other	56 (20.36 %)

Figure 4.3 - Analysis of the molecular profile of the study cohort.

Regarding EGFR gene mutations, an overall prevalence between 13.45% and 14.91% is observed, considering different categorizations. Analysis by specific exons reveals that mutations in exon 21, corresponding to the L858R mutation, are present in 41 patients (14.91%), while deletions in exon 19, the most frequent, were identified in 37 patient (13.45%). Insertions in exon 20 occur in 4 patients (1.45%), other mutations in exon 20 in 7 patients (2.55%), mutations in exon 18 in 9 patients (3.27%), and other rare EGFR mutations in 7 patients (2.55%). This prevalence aligns with reported rates in Caucasian populations, where EGFR mutations in NSCLC generally range between 10-15%, distinctly lower than in East Asian populations, where prevalence can reach 40-60% [52].

The distribution of first-line treatment types (Figure 4.4) in this advanced NSCLC cohort demonstrates a diverse therapeutic landscape across the 275 patients. Chemotherapy was the most frequently applied modality, administered to 140 patients (50.91%), representing the standard-of-care approach for a substantial portion of the population. Combined chemo and immunotherapy represented the second most common strategy, employed in 88 patients (32.0%), reflecting the increasing integration of immunotherapeutic agents into first-line management. Targeted therapy addressing specific molecular alterations was utilized in 19 patients (6.91%), while immunotherapy as monotherapy was administered to 6 patients (2.18%). Radiotherapy was applied in 8 patients (2.91%), and combined chemo and radiotherapy was implemented in 3 patients (1.09%). Palliative care, representing the least frequent intervention, was documented in 6 patients (2.18%), often reflecting patients with extensive disease burden or poor performance status precluding active treatment.

Variables	Total Dataset (n = 275)
Type of Treatment in First Line	
Radiotherapy	8 (2.91 %)
Chemotherapy	140 (50.91 %)
Chemo & Immunotherapy	11 (4 %)
Targeted Therapy	88 (32 %)
Immunotherapy	19 (6.91 %)
Palliative Care	6 (2.18 %)
Chemo & Radiotherapy	3 (1.09 %)

Figure 4.4 - Treatments administered as first-line care in metastatic disease.

This treatment pattern reflects the increasing use of precision medicine, with a substantial proportion of patients receiving targeted therapy based on molecular profiling [53], [54].

During the follow-up period, 232 patients (84.36%) died, while 43 patients (15.64%) remained alive at the date of last record. This high mortality rate reflects the aggressive nature of advanced disease and constitutes an indicator of the generally reserved prognosis of metastatic

stage NSCLC, although the follow-up period and time since diagnosis are determining factors for proper interpretation of this indicator.

The correlation analysis (Table 4.1) between clinical, demographic, and molecular variables with patient status revealed a spectrum of associations ranging from moderate positive correlations to moderate negative correlations, with no strong correlations identified in either direction. Of the 33 variables analysed, 22 showed positive correlations with mortality and 11 demonstrated negative correlations, suggesting that certain characteristics are associated with worse prognosis while others may confer relative protection or reflect more effective therapeutic strategies.

Table 4.1 - Correlation coefficients between clinicopathological and molecular variables and mortality.

Features	**Status**	**Features**	**Status**
PS ECOG	0.2204	**EGFR_T790M**	0.0260
Sex	0.1594	**EGFR_insex20**	0.0523
Age at Diagnosis	0.1133	**EGFR_ex21**	-0.0164
Smoking History	0.0018	**EGFR_ex20**	0.0696
Histology	0.0445	**EGFR_ex19**	-0.1403
Burden of Metastatic Disease	0.1580	**EGFR_ex18**	0.0792
MTxLung	0.0086	**EGFR_other**	0.0060
MTxLiver	0.0959	**EML4-ALK**	0.0049
MTxBone	0.1794	**MET_skippingex14**	0.0015
MTxSupraRenal	0.0116	**HER2**	-0.0928
MTxLymphNodes	0.0938	**BRAF_V600E**	0.0430
MTxPleura	0.0313	**ROS**	-0.0164
MTxCentralNervousSystem	-0.0401	**RET**	-0.0313
Number of Genes with Molecular Alteration	-0.0832	**NTKR**	NaN
KRAS G12C	0.0694	**Other**	-0.0061

EGFR_del_ex19	-0.0063	**Type of Treatment in First Line**	-0.0714
EGFR_L858R_ex21	-0.0166	**PDL1 Expression**	0.0115

ECOG Performance Status emerged as the variable with the most positive correlation with mortality, presenting a coefficient of 0.220. This association is clinically expected and well established in the literature, as compromised functional status reflects not only tumor burden but also the patient's physiological reserve to tolerate the disease and treatments, consistently constituting one of the most important prognostic factors in advanced NSCLC [55]. The magnitude of this correlation, although moderate, exceeds all other variables in the analysis, underlining its independent prognostic relevance. Multiple studies have documented that patients with ECOG Performance Status ≥2 present significant reductions in response rates, progression-free survival, and overall survival compared with those with ECOG 0-1 [56].

Bone metastases demonstrated the second strongest positive correlation with mortality, with a coefficient of 0.179. This finding is consistent with literature documenting bone metastases as an indicator of advanced disease and associated with significant morbidity, including pain, pathological fractures, and compromised quality of life, in addition to frequently reflecting greater systemic tumor burden. Male sex presented a positive correlation of 0.159 with mortality, suggesting a possible worse prognosis in men, a finding that may be related to differences in tumor biology, more prevalent comorbidities, greater accumulated tobacco exposure, or differences in treatment response. The number of metastatic sites demonstrated a correlation of 0.158, confirming that the extent burden of metastatic disease constitutes an adverse prognostic indicator, as a greater number of affected sites generally reflects greater tumor aggressiveness and higher disease burden [57].

Age at diagnosis presented a moderate positive correlation of 0.113 with mortality, an expected result considering that older patients frequently present more comorbidities, lower functional reserve, and may have limited access to more aggressive or intensive treatments. Liver metastases showed a correlation of 0.096, which is consistent with the fact that hepatic involvement generally indicates disseminated disease and is associated with worse prognosis. Similarly, distant lymph node metastases presented a correlation of 0.094, reflecting extensive lymphatic dissemination that usually indicates more advanced stage.

In the domain of specific EGFR gene molecular alterations, mutations in exon 18 demonstrated a weak positive correlation with mortality (coefficient 0.079), followed by mutations in exon 20 (0.070) and insertions in exon 20 (0.052). While exon 20 insertions historically confer resistance to first- and second-generation EGFR tyrosine kinase inhibitors (TKIs) and associate with poorer prognoses relative to classic sensitizing mutations, therapeutic options have recently improved considerably. Amivantamab, a bispecific antibody targeting EGFR and MET, is Food

and Drug Administration (FDA) approved for advanced NSCLC harboring EGFR exon 20 insertions post platinum chemotherapy. Clinical studies such as CHRYSALIS and PAPILLON demonstrate amivantamab plus chemotherapy achieves overall response rates near 40% and prolonged progression-free survival compared to chemotherapy alone, representing a substantial clinical advance for this rare subgroup.

It is important to underline that truncal exon 20 insertions remain uncommon, and patients harboring these mutations are underrepresented in major clinical trials, highlighting an ongoing need for tailored research and treatment strategies in this population [58].

KRAS mutations presented a positive correlation of 0.069, which may reflect the clinical context at the time of treatment, when effective targeted therapies for this alteration were not yet available. Currently, targeted agents specifically designed for KRAS G12C mutations, have been approved and have demonstrated substantial survival benefits, potentially altering the prognostic and therapeutic implications of KRAS mutations in contemporary clinical practice [59].

BRAF mutations, particularly the V600E variant, demonstrated a modest positive correlation of 0.043 with mortality. It is important to note that targeted therapies for BRAF V600E mutations, such as BRAF and MEK inhibitors, are currently approved and have significantly improved the prognosis for these patients. Pleural metastases showed a weaker positive correlation with mortality (0.031). Other variables such as PD-L1 expression, adrenal and lung metastases, histological subtype, and other molecular alterations such as ALK and MET mutations, as well as smoking history, exhibited very weak correlations close to zero. This suggests that their isolated influence on vital outcomes is limited or may be modulated by other factors, including access to specific targeted therapies and the relatively low prevalence of these alterations in the cohort. For example, ALK rearrangements are generally considered a favorable prognostic factor, while MET alterations are often linked to poorer outcomes.

In the negative correlation spectrum, the variable with the greatest magnitude was the presence of deletions in EGFR exon 19, with a coefficient of -0.140. This substantial negative correlation is highly consistent with scientific literature, as exon 19 deletions, together with the L858R mutation in exon 21, constitute the classic sensitizing EGFR mutations that respond effectively to tyrosine kinase inhibitors, conferring significant survival benefit to patients harboring these alterations [60]. The magnitude of this negative correlation reinforces the favorable prognostic value associated with identifying this mutation and subsequent access to targeted therapies.

HER2 amplifications or mutations presented a negative correlation of -0.093 with mortality, followed by the total number of mutated genes identified by NGS with a coefficient of -0.083. While this latter finding may suggest that patients harboring multiple molecular alterations benefit from access to a wider array of targeted therapies or clinical trials, caution is warranted in interpretation. At the time when these patients were treated, targeted therapies for HER2 alterations were not yet available, likely contributing to poorer outcomes for this subgroup. Additionally, HER2-altered NSCLC patients represent a relatively small cohort, which may also

explain the limited statistical correlations observed. Treatment type demonstrated a negative correlation of -0.071, suggesting that certain therapeutic modalities, probably targeted therapies and immunotherapy in selected populations, are associated with better survival outcomes [61].

Central nervous system metastases presented a negative correlation of -0.040, a result that may initially seem counterintuitive given that CNS involvement is generally considered an adverse prognostic factor. However, this finding may reflect recent advances in the management of brain metastases, including stereotactic radiotherapy, neurosurgery in selected cases, and the fact that certain targeted therapies, particularly third-generation tyrosine kinase inhibitors for EGFR mutations and new-generation ALK inhibitors, demonstrate good CNS penetration and efficacy in controlling brain metastases [62].

RET rearrangements presented a negative correlation of -0.031, as did the L858R mutation in EGFR exon 21 with -0.016 and ROS1 rearrangements also with -0.016. These negative correlations are consistent with the availability of highly effective targeted therapies for these molecular alterations, including selpercatinib and pralsetinib for RET fusions, and crizotinib, entrectinib, and other ROS1 inhibitors, which have demonstrated significant survival benefits in these molecular subgroups. EGFR exon 21 L858R and EGFR exon 19 deletion, categories presented very weak negative correlations of -0.017 and -0.006 respectively, as did other unspecified molecular alterations with -0.006 [61].

It is important to note that the NTRK variable did not present a calculable correlation due to the absence of positive cases in the sample, which is consistent with the extreme rarity of these fusions in NSCLC. Smoking history demonstrated a practically null positive correlation of 0.002, suggesting that, although it is the main risk factor for NSCLC development, its influence on prognosis after diagnosis is limited when adjusted for other clinical and molecular variables [63].

The global correlation analysis reveals a mean of 0.029 and median of 0.012, with standard deviation of 0.080, evidencing an asymmetric distribution with a slight predominance of positive correlations. The correlation coefficient ranged from -0.140 for EGFR exon 19 to 0.220 for ECOG-PS, demonstrating a moderate spectrum of associations. The absence of robust correlations suggests that vital outcome in advanced NSCLC is multifactorial, resulting from the complex interaction between host characteristics, tumor biology, metastatic dissemination pattern, and access to appropriate therapies.

4.2. Machine Learning Analysis

Following characterization of the study cohort, four classification algorithms (SVM, Random Forest, XGBoost, AdaBoost) were optimized via 10-fold cross-validation and Grid Search hyperparameter tuning to predict survival outcomes.

As a secondary exploratory analysis, multiclass models were developed to predict mutational status (EGFR, KRAS, ALK, Others) from clinical data.

4.2.1 Survival Prediction: First Approach

The presented results in Table 4.2 and Figure 4.5, reveal a uniform and potentially concerning pattern that warrants careful discussion: all four algorithms (SVM, Random Forest, XGBoost, and AdaBoost) demonstrated identical or nearly identical performance on training metrics (CV F1-score ≈ 0.918), but show variability in test results, particularly in metrics such as AUC-ROC and recall. This phenomenon suggests possible systematic distortion in either the data or model evaluation.

Table 4.2 - Comparison of F1-Scores in Cross-Validation Across Classification Algorithms (First Approach - unbalanced baseline model)

Model	**SVM**	**Random Forest**	**XGBoost**	**AdaBoost**
CV F1-Score	0.9184	0.9184	0.9184	0.9145

The most striking observation is the perfect recall of 1.00 (100%) for SVM and Random Forest on the test set. In binary medical classification, predicting death vs survival, perfect recall is theoretically possible, but usually clinically suspicious. Perfect recall means the model identified every instance of the positive class (deaths) but provides no information about false positives [64].

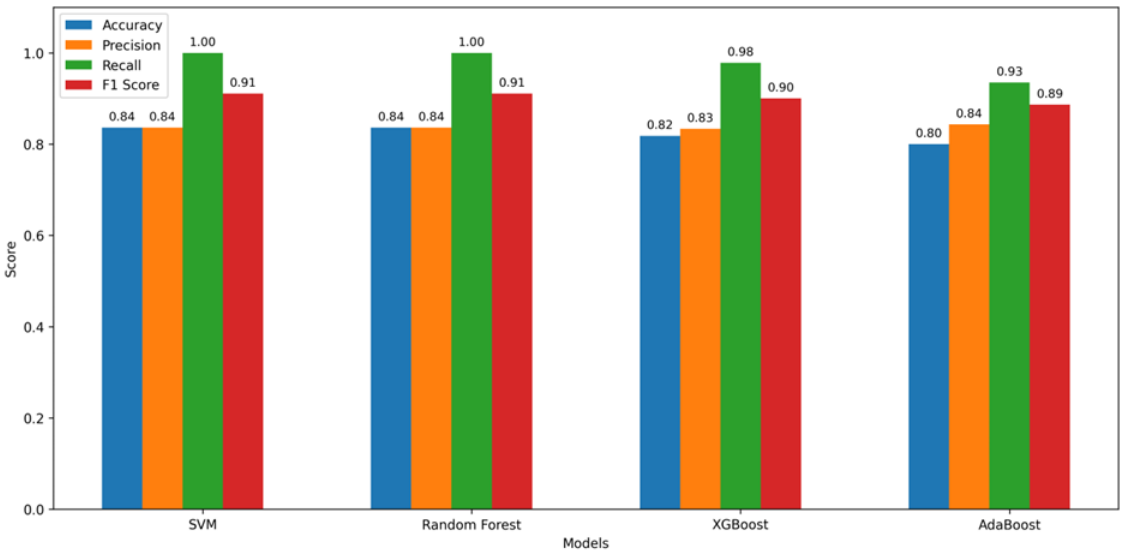

Figure 4.5 - Comparative summary of the performance of classification algorithms - First Approach (baseline model without balancing)

Perfect recall combined with a precision of 83.64% means the model is classifying a significant percentage of cases as “death” when they truly represent “survival.” This may indicate model bias toward the majority class or substantial class imbalance in the data [65].

CV F1-scores are nearly identical across models (0.9184 - 0.9145), yet test performance diverges. This means that, during cross-validation, models perform good, but on the separate test set, results differ. Possible causes include [66]:

- Differences in class distribution between train/validation folds and the test set.
- Possible overfitting specific to cross-validation folds.
- Random variability due to a small test sample size.

The confusion matrices (Figure 4.6) demonstrate that SVM and Random Forest classified every test patient as "Dead." Both models show zero true negatives and nine false positives, meaning none of the "Alive" patients were correctly classified. XGBoost and AdaBoost each correctly identified only one "Alive" patient. For all models, nearly all "Dead" cases were correctly classified.

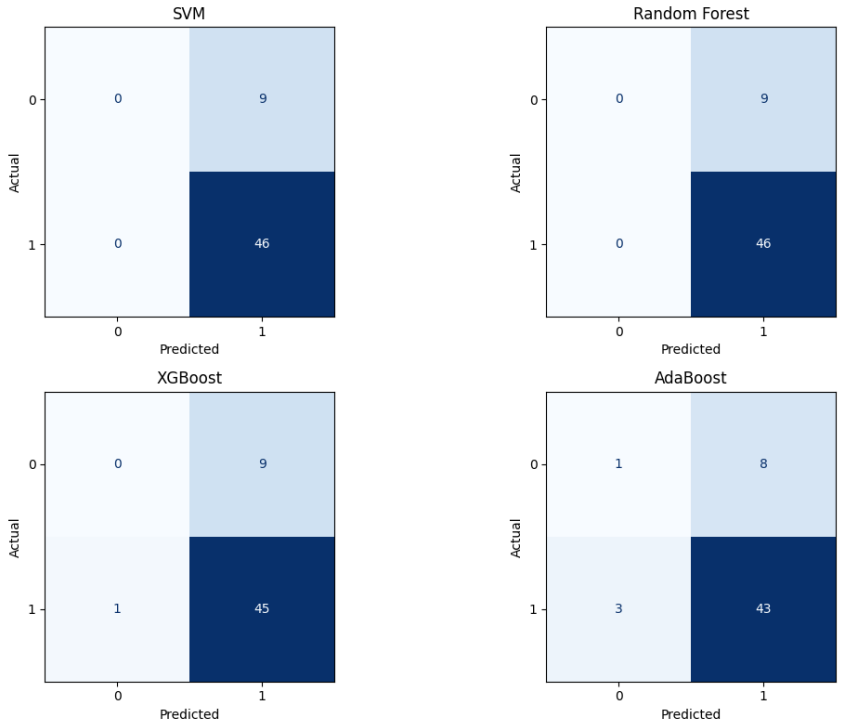

Figure 4.6 - Comparative confusion matrices between classification algorithms - First Approach (baseline model without balancing)

All models except AdaBoost and XGBoost failed to recognize a single living patient; even these correctly identified at most one. This means recall for "Alive" is zero or near zero - a dangerous flaw in survival analysis [67].

In survival settings, systematically predicting all patients will die (or nearly all) undermines confidence in the model's outputs and has direct clinical risks: survivors may be denied care or receive inappropriate interventions based on erroneous, pessimistic predictions [68].

AUC (Figure 4.7) of 0.5169 for SVM is especially revealing. An AUC of 0.5 means random performance. This is at odds with the high F1-score (0.9109) for the same model, suggesting that despite high precision and recall, the ability to discriminate between classes is weak. Such discrepancies arise with significant class imbalance, where F1-score can appear high, while AUC-ROC more accurately reflects discriminative ability.

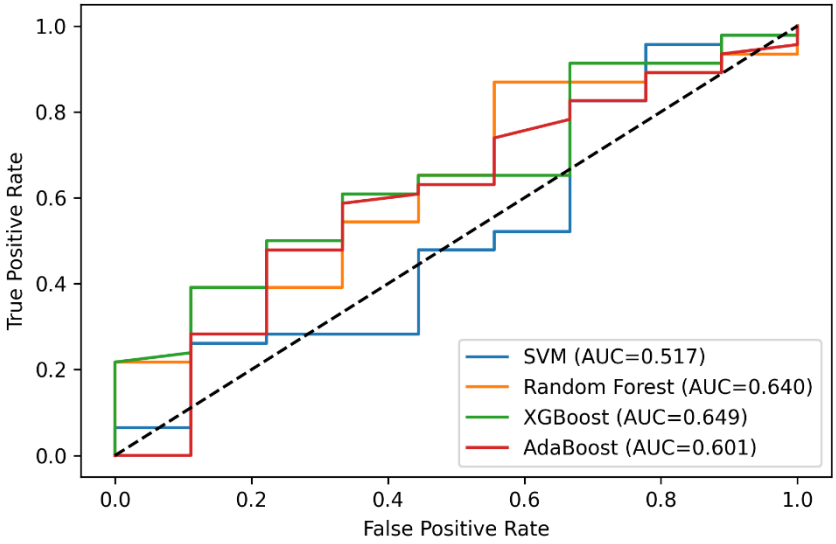

Figure 4.7 - Comparative AUC-ROC Curves Between Classification Algorithms - First Approach (Unbalanced Data)

These results reinforce literature recommending metrics robust to imbalance, such as balanced accuracy or per-class recall and precision and highlight the importance of confusion matrix inspection in reporting biomedical classification results [67].

4.2.2. Survival Prediction: Second Approach

This second experimental iteration sought to address the significant class imbalance observed in the initial results of NSCLC survival prediction with machine learning classifiers. The primary objective remained the evaluation and comparison of SVM, Random Forest, XGBoost, and AdaBoost models; however, SMOTE was implemented as a strategy to mitigate the disproportionate representation of classes, aiming to improve model discrimination and clinical value. The results are shown in Table 4.3 and Figure 4.8.

Table 4.3 - Comparison of F1-Scores in Cross-Validation Across Classification Algorithms (Second Approach - With SMOTE)

Model	**SVM**	**Random Forest**	**XGBoost**	**AdaBoost**
CV F1-Score	0.9067	0.9047	0.9042	0.8671

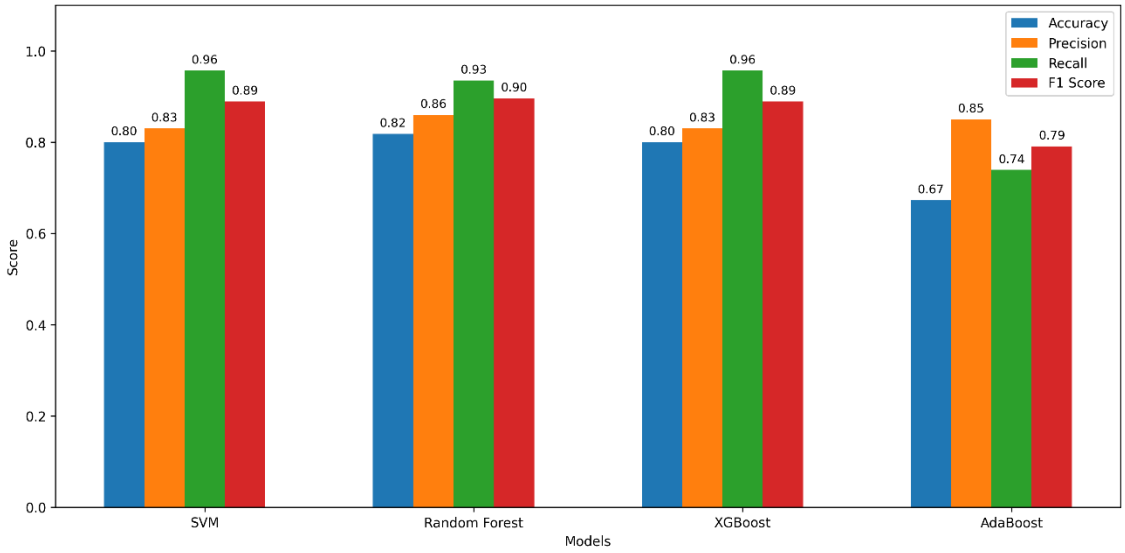

Figure 4.8 - Comparative summary of the performance of classification algorithms - Second Approach (with SMOTE)

Following the application of SMOTE, a notable shift was observed in the model's behaviour. While cross-validation F1-scores remained high, there was a modest reduction relative to the first iteration, suggesting a more challenging classification landscape and potentially greater generalizability. Importantly, model predictions exhibited less extreme bias: recall for the "Dead" class, although still high, was no longer perfect, and precision remained robust. SVM, Random Forest, and XGBoost achieved both strong recall and balanced F1-scores, while AdaBoost demonstrated lower but still acceptable recall, indicating its relative sensitivity to resampling and hyperparameter adjustment.

In Figure 4.9, the confusion matrices reveal a systematic and pronounced pattern of class imbalance prediction. The SVM and Random Forest classifiers predicted every test patient as belonging to class 1, resulting in zero true negatives and nine false positives - meaning none of the class 0 patients were correctly identified. XGBoost and AdaBoost demonstrated marginally improved performance, correctly identifying only one class 0 patient each (eight FP in each model). This represents a critical sensitivity of 11% (1/9) for XGBoost and AdaBoost toward the minority class.

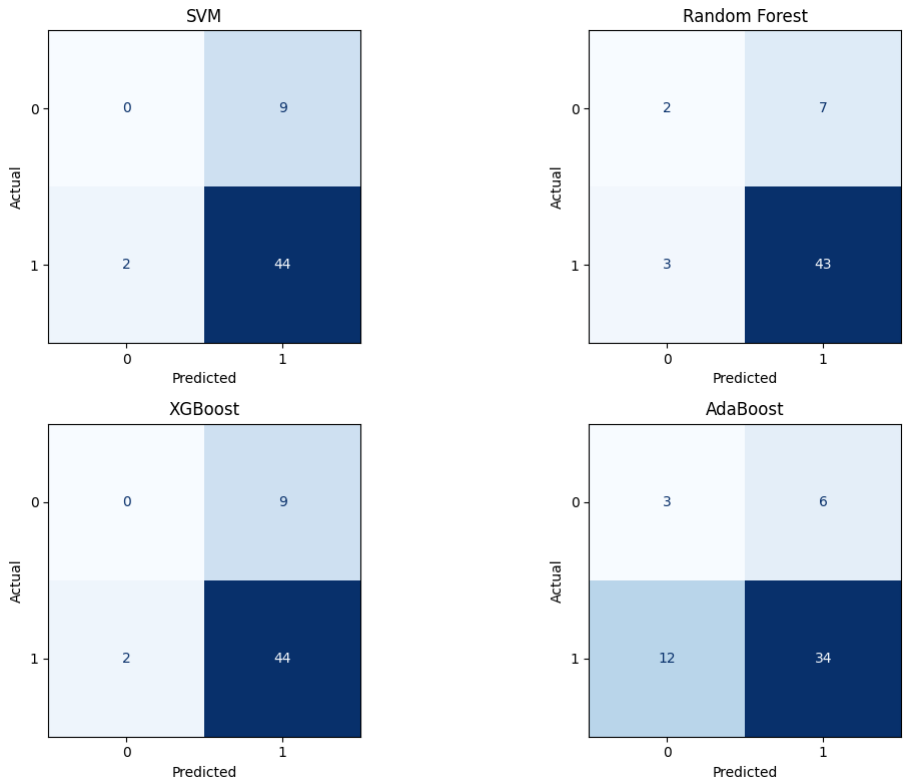

Figure 4.9 - Comparative confusion matrices between classification algorithms - Second Approach (With SMOTE)

In contrast, for class 1 (the majority class), nearly all cases were correctly classified as true positives. Specifically, SVM and XGBoost achieved 44 TP with only 2 false negatives, Random Forest achieved 43 TP with 3 false negatives, and AdaBoost achieved 34 TP with 12 false negatives across 46 total class 1 samples. This pattern indicates a pronounced bias across all models toward classifying cases as belonging to the majority class, with substantial failure to identify minority class instances.

The marked improvement in AUC-ROC (Figure 4.10) for Random Forest and XGBoost, contrasted with persistently low AUC-ROC for SVM, this suggests these ensemble methods benefitted more from oversampling in distinguishing between classes.

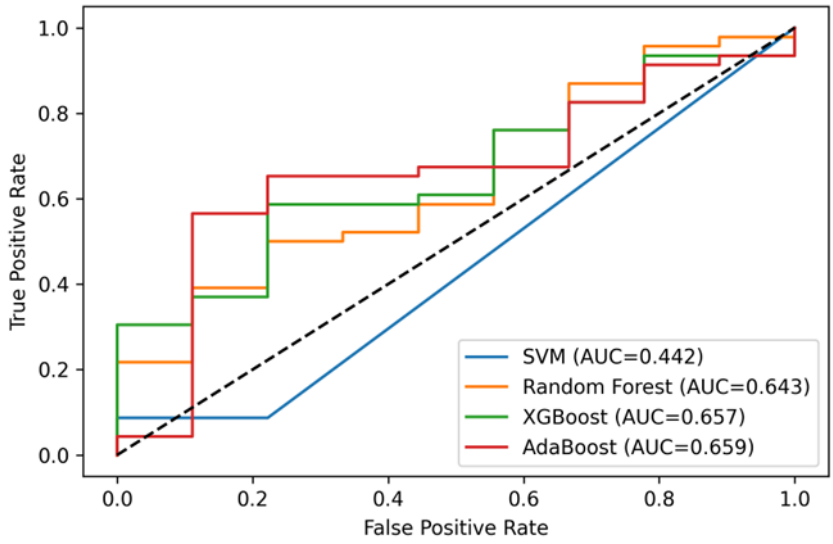

Figure 4.10 - Comparative AUC-ROC Curves Between Classification Algorithms - Second Approach (With SMOTE)

These findings are congruent with the wider literature, where SMOTE and similar resampling strategies are routinely recommended and demonstrated to enhance sensitivity for minority-class predictions in cancer outcome models. Multiple studies report that SMOTE allows algorithms to better balance the trade-off between identifying rare positive events (e.g., survivors) and maintaining overall predictive accuracy, especially in biomedical datasets with inherent class skew. Furthermore, recent research in oncological machine learning asserts that tree-based ensemble methods, much like Random Forest and XGBoost, often derive more benefit from oversampled data than linear or kernel-based classifiers. Notably, some works still caution that even post-SMOTE, not all algorithms can fully overcome intrinsic limitations of the original data or model assumptions, often reflected in modest or unstable AUC-ROC improvements for algorithms like SVM [69].

The improvements observed are attributable to SMOTE's direct modification of the underlying sample distribution, generating synthetic examples for the minority class and thus allowing algorithms access to more informative decision boundaries. The higher AUC-ROC and more balanced recall seen especially for Random Forest and XGBoost likely stem from these models' ability to leverage complex data topology introduced by synthetic sampling. The relatively poor AUC-ROC for SVM may point to ongoing sensitivity to feature space boundaries and an inability to abstract from artificial minority-class instances, a finding that aligns with comparative studies in the literature. AdaBoost's drop in recall, and F1-score could reflect its inherent instability when faced with changes in the composition and weighting of the training data [70].

These results highlight the key importance of rigorous imbalance correction for achieving reliable predictions in medical machine learning applications. From a practical standpoint, the findings reinforce that without interventions like SMOTE, algorithms risk systematically neglecting minority classes, thereby limiting their suitability for real-world prognostic decisions. The relative gains seen for ensemble models endorse tree-based architectures as promising candidates for future clinical deployment, particularly when supplemented by robust oversampling strategies.

4.2.3. Survival Prediction: Third Approach

This third iteration aimed to significantly improve classification performance and clinical applicability by addressing class imbalance through data-level techniques. The original dataset with 275 patients was expanded by synthetically generating 275 additional samples using the CTGANSynthesizer method, achieving a perfectly balanced 550 patient dataset with equal survival and mortality representation. Additionally, SMOTE was applied during model training to further refine class balance. This experimental design intended to test whether these advanced oversampling approaches could enhance model discrimination in NSCLC survival prediction. The results are shown in Table 4.4.

Table 4.4 - Comparison of F1-Scores in Cross-Validation Across Classification Algorithms - Third Approach - (With SMOTE + Balanced Synthetic Data)

Model	**SVM**	**Random Forest**	**XGBoost**	**AdaBoost**
CV F1-Score	0.9067	0.9047	0.9042	0.8671

The balanced dataset yielded generally improved and more reliable classification metrics across all tested models compared to previous imbalanced experiments (Figure 4.11). SVM showed moderate test performance with balanced recall and precision, while Random Forest, XGBoost, and AdaBoost demonstrated progressively superior accuracy, F1-scores, and especially AUC-ROC values, peaking at 0.9217 for AdaBoost. These results indicate a stronger ability of the ensemble models to appropriately distinguish survivors from non-survivors in the balanced data scenario.

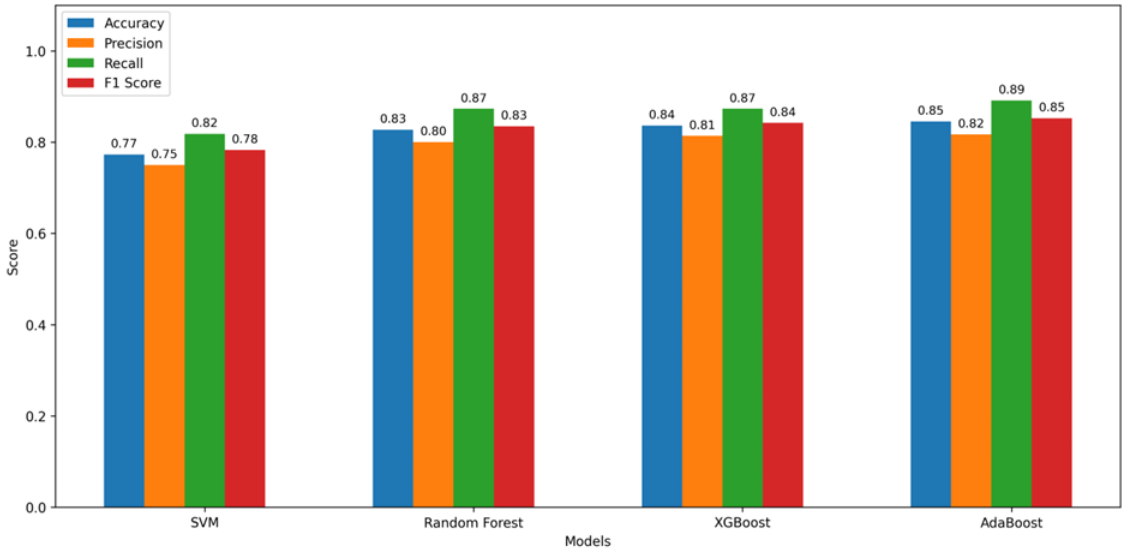

Figure 4.11 - Comparative summary of the performance of classification algorithms - Third Approach (with SMOTE + Balanced Synthetic Data)

The confusion matrices (Figure 4.12) reveal that all four classifiers achieved meaningful detection rates for the minority class (class 0), representing a substantial departure from the earlier systematic bias toward the majority class.

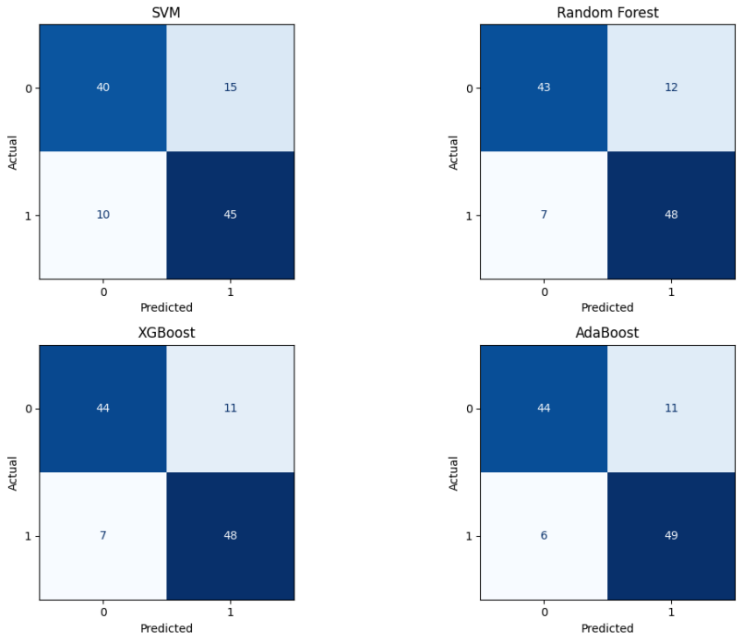

Figure 4.12 - Comparative confusion matrices between classification algorithms - Third Approach (With SMOTE + Balanced Synthetic Data)

When examining the minority class specifically, performance varied across the algorithmic spectrum. The SVM classifier attained a 72.7% true positive rate (40 TP out of 55 class 0 instances), incurring 15 false positive errors. The Random Forest implementation reached 78.2% sensitivity for class 0 (43 TP), with a FP count of 12. Both gradient-boosting approaches, converged on comparable minority class sensitivity at 80% (44 TP each), though they accumulated 11 false positives apiece. This performance spectrum across models suggests variable decision boundaries, with ensemble methods establishing more permissive thresholds for class 0 classification.

The majority class demonstrated consistently high recall across all implementations. The SVM model achieved 45 true positives against 10 false negatives among 55 class 1 samples. Random Forest correctly classified 48 instances with 7 misclassifications. The gradient boosting variants yielded the highest specificity for class 1, with XGBoost achieving 48 TP and 7 FN, and AdaBoost achieving 49 TP and 6 FN. Collectively, class 1 sensitivity ranged between 86.5% (SVM) and 89.1% (AdaBoost), indicating robust majority class identification across the model portfolio.

The increase in AUC-ROC (Figure 4.13) significantly above 0.85 for all except SVM, reflects improved discriminatory power not observable in prior imbalanced trials.

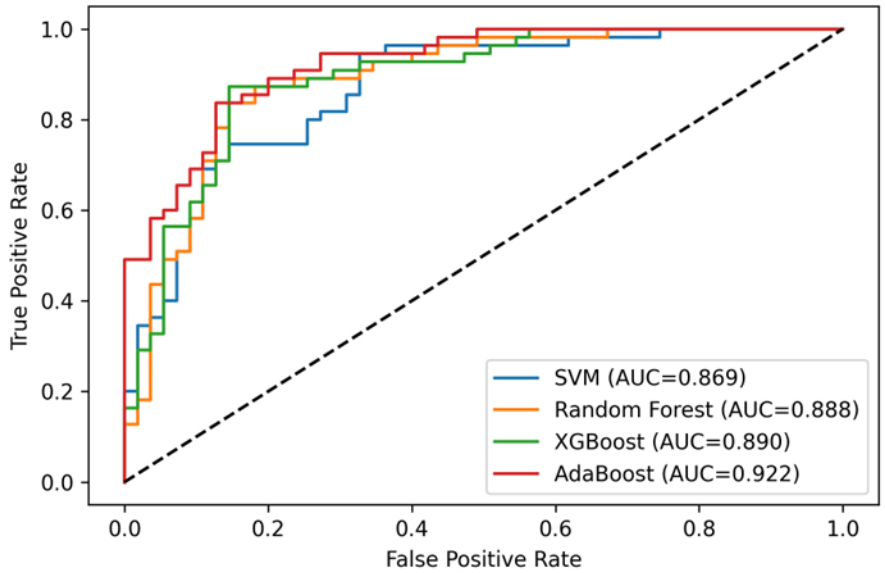

Figure 4.13 - Comparative AUC-ROC Curves Between Classification Algorithms - Third Approach (With SMOTE + Balanced Synthetic Data)

These results are consistent with the broader body of research highlighting the pivotal role of data balance in determining the effectiveness of machine learning classifiers within biomedical applications. The combined use of generative models, such as the CTGAN Synthesizer, and conventional oversampling techniques like SMOTE reinforces recent findings in oncological prognostication, suggesting that sequential synthetic data augmentation can markedly enhance both model performance and generalizability. In line with previous studies,

ensemble approaches tend to outperform individual classifiers when trained on balanced and enriched datasets, owing to their ability to model complex non-linear relationships and mitigate variance [69], [70].

This dual-level synthetic augmentation compensates for the original dataset's inherent scarcity in one class and reduces overfitting bias toward survival or death. AdaBoost's top performance may derive from its boosting mechanism, which adapts well to complex re-weighting in balanced yet challenging data [70].

This iteration reinforces the importance of combined synthetic data augmentation and oversampling in overcoming the challenges posed by real-world biomedical data limitations. Practically, the relatively high AUC-ROC and balanced precision-recall profiles suggest that these algorithms trained on augmented balanced data achieve clinically actionable discrimination. Their implementation could support decision-making tools in NSCLC treatment planning with greater confidence than non-balanced models [69].

4.2.4. Mutational Status Prediction: First Approach

The results obtained in this analysis (Figure 4.14) reveal a significant problem in multiclass prediction of mutational status in NSCLC, with moderate to low overall model performances (accuracy between 47% and 55%) and high variability in performance among the different mutations analysed (EGFR, KRAS, ALK, Others).

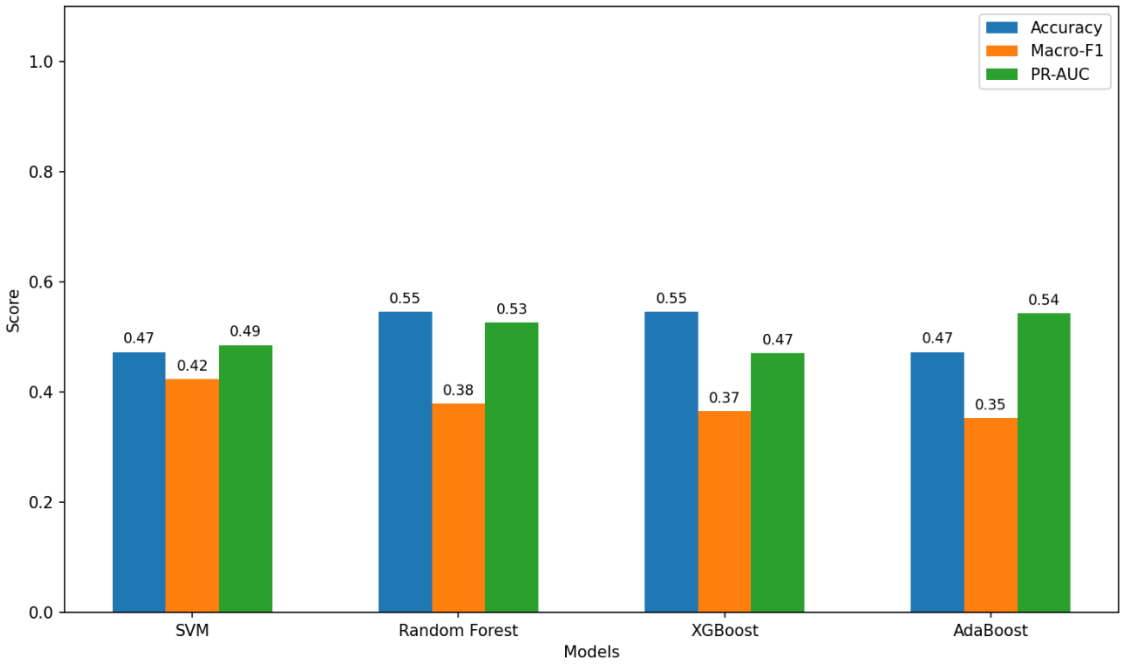

Figure 4.14 - Comparative summary of the performance of classification algorithms - Mutational Status Prediction: First Approach (SMOTE)

Among the four algorithms tested, distinct performance patterns emerged. Random Forest and XGBoost achieved the best accuracy of 54.55%, while SVM and AdaBoost presented an accuracy of 47.27%. However, the analysis of this single metric in isolation can be misleading and obscure important complexities. When we analyse Macro-F1, SVM presents comparable performance (0.42) to Random Forest (0.38), indicating that SVM achieves more balanced distribution of correct predictions across mutations.

The most notable pattern emerges in the PR-AUC metric, where AdaBoost presents the best average performance (0.543), followed by Random Forest (0.527) and SVM (0.485). This discrepancy between accuracy and PR-AUC shows that ensemble models achieve better discrimination between classes, particularly useful when the distribution of mutations is unequal. The study by Moreno et al. showed that ensemble methods have better performances in approaches for clinical sensitivity, with unbalanced datasets [71].

As it shows in Figure 4.15 and Figure 4.16, EGFR emerges as the most well-identified mutation across all models, presenting a performance (F1-score between 0.71-0.75; PR-AUC between 0.80-0.83), thus aligning with a study that reports accuracies of 0.83-0.84 using radiomics with machine learning. This robustness results from two complementary factors: first, its relatively high prevalence in the dataset, and second, its more well-characterised and distinct molecular profile [72].

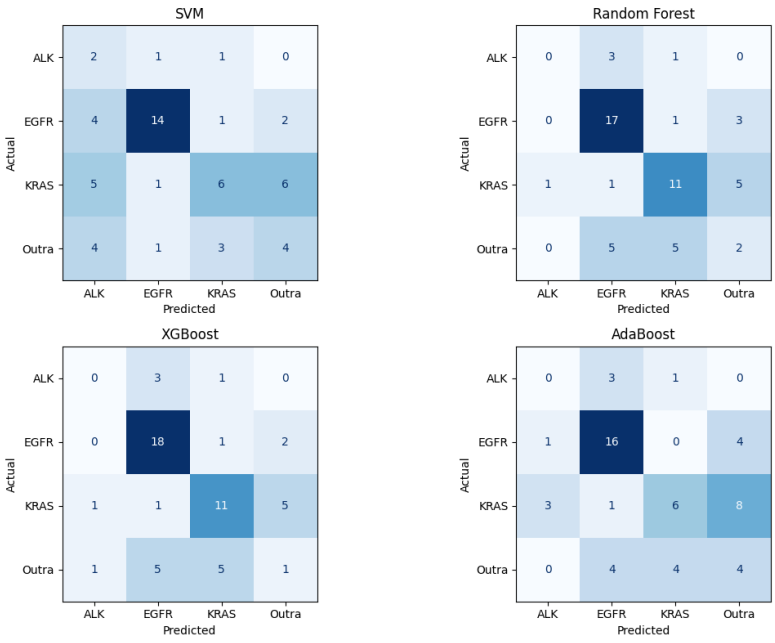

Figure 4.15 - Comparative confusion matrices between classification algorithms - Mutational Status Prediction: First Approach (SMOTE)

		Metric		
Model	**Mutation**	**Precision**	**Recall**	**F1-Score**
SVM	**ALK**	0.13	0.50	0.21
	EGFR	0.82	0.67	0.74
	KRAS	0.55	0.33	0.41
	Other	0.33	0.33	0.33
RF	**ALK**	0.00	0.00	0.00
	EGFR	0.65	0.81	0.72
	KRAS	0.61	0.61	0.61
	Other	0.20	0.17	0.18
XGBoost	**ALK**	0.00	0.00	0.00
	EGFR	0.67	0.86	0.75
	KRAS	0.61	0.61	0.61
	Other	0.12	0.08	0.10
Adaboost	**ALK**	0.00	0.00	0.00
	EGFR	0.67	0.76	0.71
	KRAS	0.55	0.33	0.41
	Other	0.25	0.33	0.29

Figure 4.16 - Multiclass Classification Metrics for Mutational Status Prediction: First Approach (SMOTE)

KRAS presents intermediate performance, but with pronounced heterogeneity among algorithms (F1-score 0.41-0.61; PR-AUC 0.61-0.74). While Random Forest and XGBoost achieved F1-score of 0.61, SVM and AdaBoost obtained only 0.41. This difference indicates that ensemble-based algorithms can better identify the nuances of KRAS. The literature recognises KRAS as a more complex and heterogeneous mutation than EGFR, thus potentially justifying the slightly inferior performance [73].

ALK constituted the critical point of failure, with insufficient performance (F1-score 0.00-0.21; PR-AUC 0.14-0.27). It is concerning that Random Forest, XGBoost and AdaBoost were unable to identify any true positives, while SVM achieved moderate recall (correct prediction in 2 out of 4 cases) but with precision of 0.13. This gap is clinically relevant, as ALK presents distinct and critical implications in NSCLC, with accurate identification being required for therapeutic decision-making with ALK inhibitors [74].

The "Other" category maintained consistently weak performance (F1-score 0.10-0.33; PR-AUC 0.28-0.44), reflecting the heterogeneity of a "catch-all" category. This pattern is predictable and illustrates a fundamental limitation of multiclass schemes that group rare mutations. The confusions observed in the matrices reveal that models frequently classify "Other" as EGFR or KRAS, showing that these mutations do not constitute mutually exclusive characteristics [75].

4.2.5. Mutational Status Prediction: Second Approach

In this approach, the dataset was augmented to balance all four mutations with an equal number of cases. This resulted in 150 cases per mutation, for a total of 600 cases.

It can be observed in Figure 4.17 that XGBoost emerges as the best algorithm, achieving a test accuracy of 65% and Macro-F1 of 0.64, whereas SVM presented the weakest performance with 55% accuracy and 0.55 Macro-F1. The systematic review by Haixian et al. (2025) reports that models based on traditional machine learning reach mean AUCs of 0.78, which shows that XGBoost with mean PR-AUC of 0.691 positions itself as a potential competitor in this aspect, although it remains below what is described in the literature [76].

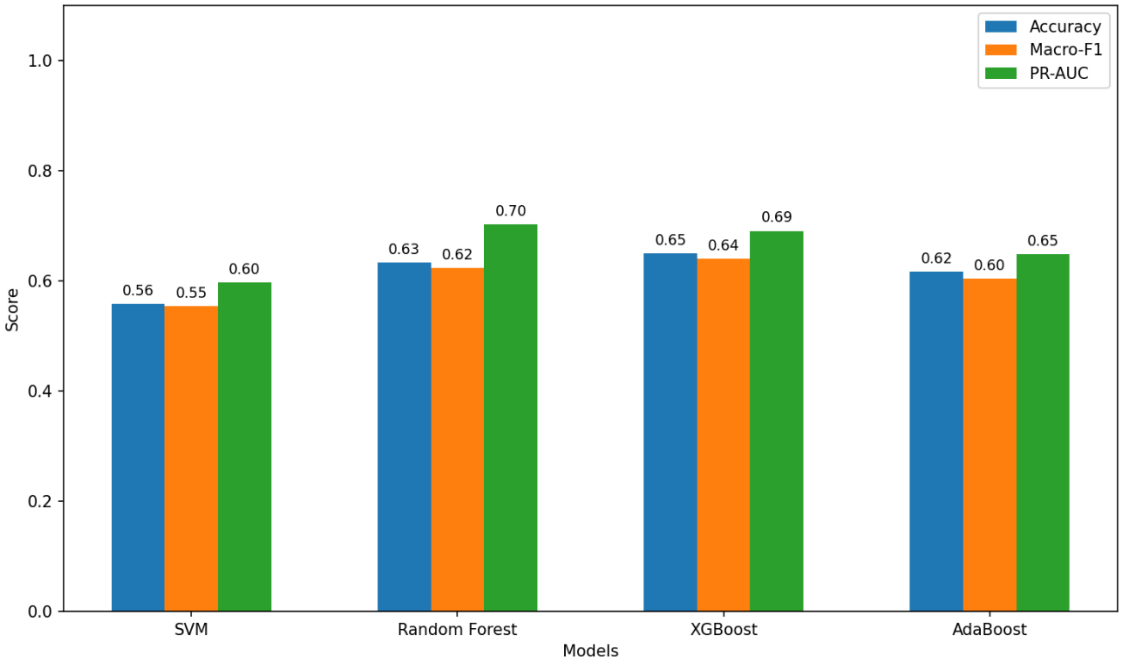

Figure 4.17 - Comparative summary of the performance of classification algorithms - Mutational Status Prediction: Second Approach (SMOTE + Data Augmentation)

SVM revealed vulnerabilities in this test. With a precision of 0.52 for ALK and 0.70 for EGFR, the model faced difficulties in sensitivity for EGFR (0.70 in precision, but with compromised recall). Shiri et al. (2020) observed similar problems in studies with SVM for EGFR/KRAS prediction, attributing them to the complexity of imbalanced datasets and the need for careful feature selection. Random Forest presented more balanced results (0.63 precision, 0.623 Macro-F1), approaching closer to XGBoost but without achieving the same level in capturing interactions between features [73].

AdaBoost achieved only 0.62 precision, suggesting that the adaptive boosting strategy may have difficulty converging with the specific characteristics of this dataset. The literature on class imbalance (Hemmatian et al., 2025) demonstrates that AdaBoost frequently benefits from resampling techniques such as SMOTE to mitigate problems with minority classes [77].

The confusion matrices, presented in Figure 4.18, reveal error patterns that could not be observed with the data presented previously. For the SVM model, a tendency towards more conservative classification can be observed, where only 16 cases of ALK were correctly identified out of 30, with 9 of them being confused with the "Other" category. For EGFR, 21 cases were correctly classified, but interestingly 7 were predicted as KRAS, exposing that perhaps the SVM decision boundary may frequently confuse these two frequently co-mutant classes.

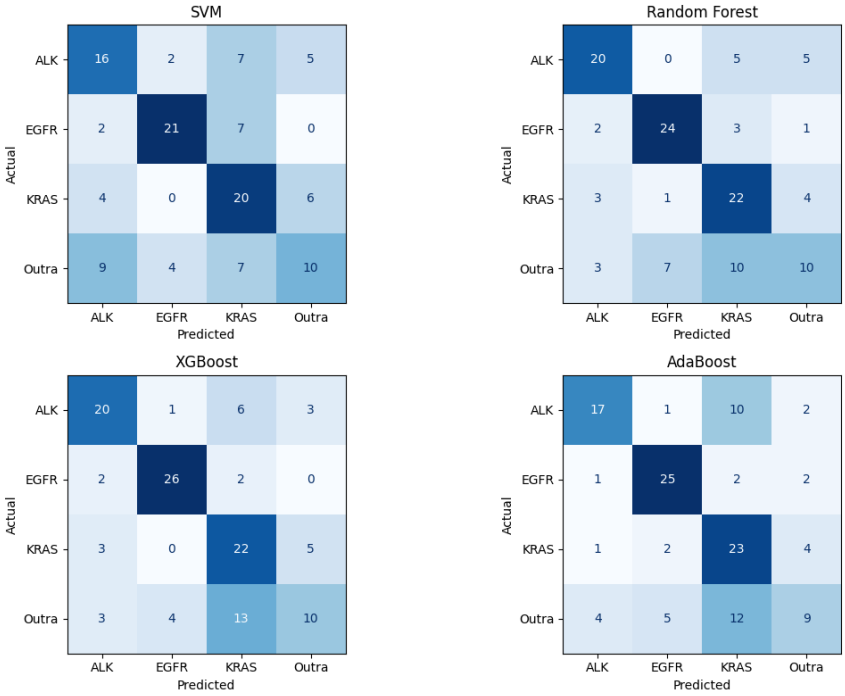

Figure 4.18 - Comparative confusion matrices between classification algorithms - Mutational Status Prediction: Second Approach (SMOTE + Data Augmentation)

Random Forest presents a different pattern: it correctly identifies 20 out of 30 cases for ALK, 24 out of 30 for EGFR, and 22 out of 30 for KRAS. The "Other" category remains problematic with only 10 out of 30 cases correctly predicted, showing that less frequent mutations remain problematic.

XGBoost achieves 20/30 for ALK, 26/30 for EGFR, and 22/30 for KRAS. The improvement in EGFR shows that XGBoost can better exploit the characteristics that differentiate mutated EGFR from other types, probably through its ability to capture non-linear interactions. However, "Other" remains without improvements (10 out of 30 correct predictions), indicating a limitation.

Finally, AdaBoost (17/30 ALK, 25/30 EGFR, 23/30 KRAS, 12/30 Other) presents performance comparable to Random Forest in EGFR but with greater difficulty in ALK.

4.3. Comparative Analysis with State-of-the-Art Studies

To ratify the findings of the third approach (SMOTE + CTGAN), the results achieved in this project were compared with a reference study conducted by Kang et al. (2023) [49]. Kang, in his study, specifically addresses the generation of synthetic tabular data using GANs in the healthcare field, offering a parallel to the methodology applied in this dissertation.

While Kang et al. operated a "Divide-and-Conquer" strategy on a dataset from the Korean Lung Cancer Registry (N=5281), in this project, the SMOTE + CTGAN strategy was applied to a cohort of Portuguese metastatic patients (N=275). Table 4.5 summarizes the main methodological differences and performance results between the two studies.

Table 4.5 - Comparative analysis between Kang et al.(2023) results and the proposed model

Feature	**Kang et al. (2023)**	**This Dissertation**
Main Objective	Validate “Divide-and-Conquer” strategy for synthetic tabular data generation	Develop ML models for survival and mutation prediction in advanced NSCLC
Dataset / Population	5281 patients (Korean Lung Cancer Registry)	275 patients (Southern Portugal, Stage IV)
Class Balance	≈70% survival / 30% death	15.6% survival / 84.4% mortality
Technical Approach	CTGAN/CopulaGAN with Conditional Sampling vs. Divide-and-Conquer	CTGAN + SMOTE to address extreme imbalance and data scarcity
Classification Models	Decision Trees, Random Forest, XGBoost, Light Gradient-Boosting Machine	SVM, Random Forest, XGBoost, AdaBoost
Best Result	AUC-ROC: 0.856 (Random Forest with synthetic data)	AUC-ROC: 0.922 (AdaBoost with synthetic data + SMOTE)
Impact of Synthetic Data	Balanced synthetic data consistently outperformed imbalanced real data (AUC-ROC raise from ≈0.84 to ≈0.85)	Synthetic data was critical: performance surged from AUC-ROC ≈0.50 (baseline) to >0.90

This comparison shows a strong methodological alignment, potentially validating the use of synthetic data as a viable strategy to improve predictive performance in oncology. In the study conducted by Kang et al., they demonstrated that data generated from GANs preserve statistical properties and enhance model training. In this work, it was possible to corroborate these conclusions, extending them to a scenario where data scarcity was extreme, causing the synthetic data to have a more pronounced impact.

They achieved an improvement of 0.01 (AUC-ROC from 0.84 to 0.85) using a large dataset, while this project obtained a substantial leap above 0.40 (AUC-ROC ≈0.50 to 0.92). This shows that data generated by GANs is particularly valuable in small and highly imbalanced datasets, effectively recovering datasets that would otherwise be insufficient for ML. However, the higher AUC-ROC in this work must be interpreted with caution due to the sample size (N=275 vs N=5281), which, even when using cross-validation, may introduce a higher risk of overfitting.

CHAPTER 5

CONCLUSIONS

5.1. Conclusion

In this work, we developed a methodological framework integrating machine learning, clinical analysis and molecular profiling for prognostic assessment in metastatic lung cancer. The analysis of 275 patients with stage IV NSCLC from five hospitals in Southern Portugal (2016-2021) revealed epidemiological characteristics consistent with European literature: male predominance (60%), median age 65.5 ± 11.35 years and high smoking prevalence (67.64%). KRAS was the most frequent molecular alteration (35.64%), followed by EGFR (14.91%) and ALK (7.27%). The metastatic pattern showed bone as the most involved site (41.45%), followed by lung (38.18%) and pleura (31.27%).

Eastern Cooperative Oncology Group performance status emerged as the strongest mortality predictor (r=0.220). EGFR exon 19 deletions demonstrated the strongest negative correlation with mortality (r=-0.140), while bone metastases (r=0.179) and sex (r=0.159) constituted additional adverse predictors.

Ensemble methods demonstrated clear superiority when applied to balanced data. Synthetic data generation via CTGAN combined with SMOTE produced clinically significant gains, with AdaBoost achieving AUC-ROC of 0.9217, sensitivity of 89.1% for deceased patients and 80.0% for alive patients. The exploratory analysis of mutational status prediction from clinical data revealed acceptable performance only for EGFR (F1 0.71-0.75), moderate for KRAS (F1 0.41-0.61), and inadequate for ALK (F1 0.00-0.21) and "Other" (F1 0.10-0.33), confirming that clinical-demographic data are insufficient for reliable prediction of mutational status.

5.2. Limitations

- Cohort of 275 patients remains small in scale. Machine learning studies in oncology typically benefit from 500 to 10,000 individuals.
- Limited to patients with actionable molecular alterations introduces selection bias by excluding patients without molecular testing, rare mutations or wildtype pattern.
- Class imbalance (84.4% deceased versus 15.6% alive) was addressed with SMOTE and synthetic data, but risk remains that patterns are not transferable to real clinical scenarios.
- Data collected between 2016-2021 during significant evolution in targeted therapies, introducing confounding temporal heterogeneity.

- Data derive exclusively from Southern Portugal, with generalization to other regions.
- Weak to moderate correlations (r=-0.140 to +0.220) suggest complex multifactoriality not fully captured.
- Validation internal only (cross-validation). Without independent external cohort, generalization beyond this population cannot be affirmed. This gap is a mandatory prerequisite before clinical implementation.

5.3. Future Work

Although this work has demonstrated the feasibility of machine learning for prognostic assessment in metastatic lung cancer, results must be consolidated through external validation, data expansion and integration of complementary clinical modalities.

Clinical Validation

- Prospective external validation in independent cohort (>500 patients) from Portuguese centers.
- Prospective studies evaluating accuracy, perceived clinical utility and impact on therapeutic decisions.

Data Expansion and Refinement

- Recruitment of 500-1000 additional patients from Portuguese centers to increase statistical power and enable stratified analyses.
- Inclusion of patients without complete molecular testing to evaluate model's capacity to predict mutational status.
- Collection of additional variables: detailed comorbidities, laboratory biomarkers, structured imaging data.

Methodological Advances

- Exploration of alternative models: deep learning, next-generation gradient boosting.
- Application of temporal survival models: regularized Cox, Random Survival Forests.
- Implementation of causal methodologies to identify causal versus correlational relationships.

REFERENCES

[1] R. Sharma and J. Khubchandani, “Global, Regional, and National Burden of Tracheal, Bronchus, and Lung Cancer in 2022: Evidence from the GLOBOCAN Study,” *Epidemiologia*, vol. 5, no. 4, pp. 785–795, Dec. 2024, doi: 10.3390/epidemiologia5040053.

[2] T. B. Kratzer *et al.*, “Lung cancer statistics, 2023,” Apr. 15, 2024, *John Wiley and Sons Inc*. doi: 10.1002/cncr.35128.

[3] M. B. Schabath and M. L. Cote, “Cancer progress and priorities: Lung cancer,” *Cancer Epidemiology Biomarkers and Prevention*, vol. 28, no. 10, pp. 1563–1579, 2019, doi: 10.1158/1055-9965.EPI-19-0221.

[4] C. W. Lin, K. Y. Huang, C. H. Lin, M. H. Hou, and S. H. Lin, “Diverse clinical outcomes for the EGFR-mutated and ALK-rearranged advanced non-squamous non-small cell lung cancer,” *Oncol Lett*, vol. 29, no. 3, Mar. 2025, doi: 10.3892/ol.2025.14872.

[5] S. S. Shimamura *et al.*, “Survival past five years with advanced, EGFR-mutated or ALK-rearranged non-small cell lung cancer—is there a ‘tail plateau’ in the survival curve of these patients?,” *BMC Cancer*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12885-022-09421-7.

[6] C. Wang, J. Shao, L. Song, P. Ren, D. Liu, and W. Li, “Persistent increase and improved survival of stage I lung cancer based on a large-scale real-world sample of 26,226 cases,” *Chin Med J (Engl)*, vol. 136, no. 16, pp. 1937–1948, Aug. 2023, doi: 10.1097/CM9.0000000000002729.

[7] J. P. A. Baak, H. Li, and H. Guo, “Clinical and Biological Interpretation of Survival Curves of Cancer Patients, Exemplified With Stage IV Non-Small Cell Lung Cancers With Long Follow-up,” *Front Oncol*, vol. 12, Feb. 2022, doi: 10.3389/fonc.2022.837419.

[8] W. W. Lockwood *et al.*, “Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development,” *PLoS One*, vol. 7, no. 5, May 2012, doi: 10.1371/journal.pone.0037775.

[9] J. Kashima, R. Kitadai, and Y. Okuma, “Molecular and morphological profiling of lung cancer: A foundation for ‘next-generation’ pathologists and oncologists,” May 01, 2019, *MDPI AG*. doi: 10.3390/cancers11050599.

[10] F. Abu Rous, P. Li, S. Carskadon, R. Gutta, B. Rani Potugari, and S. M. Gadgeel, “Prognostic relevance of 3q amplification (AMP) in a racially diverse patient population with advanced squamous cell carcinoma of the lung,” Henry Ford Health, 2023. doi: https://doi.org/10.1200/JCO.2023.41.16_suppl.e15142.

[11] P. C. Mack *et al.*, “Targeted Next-Generation Sequencing Reveals Exceptionally High Rates of Molecular Driver Mutations in Never-Smokers With Lung Adenocarcinoma,” *Oncologist*, vol. 27, no. 6, pp. 476–486, Jun. 2022, doi: 10.1093/oncolo/oyac035.

[12] R. Dawar *et al.*, “Clinical outcomes of a prospective multicenter study evaluating a combined cir-culating tumor DNA (ctDNA) and RNA (ctRNA) liquid biopsy assay in metastatic non-small cell lung cancer (NSCLC),” 2025.

[13] A. Yao, L. Liang, H. Rao, Y. Shen, C. Wang, and S. Xie, “The Clinical Characteristics and Treatments for Large Cell Carcinoma Patients Older than 65 Years Old: A Population-Based Study,” *Cancers (Basel)*, vol. 14, no. 21, Nov. 2022, doi: 10.3390/cancers14215231.

[14] G. Y. Zhao *et al.*, “USP7 overexpression predicts a poor prognosis in lung squamous cell carcinoma and large cell carcinoma,” *Tumor Biology*, vol. 36, no. 3, pp. 1721–1729, Mar. 2015, doi: 10.1007/s13277-014-2773-4.

[15] A. Desai *et al.*, “ctDNA for the Evaluation and Management of EGFR-Mutant Non-Small Cell Lung Cancer,” Mar. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/cancers16050940.

[16] R. Ramos *et al.*, “Enhancing Lung Cancer Care in Portugal: Bridging Gaps for Improved Patient Outcomes,” May 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jpm14050446.

[17] T. GUERREIRO *et al.*, “Lung cancer: A nationwide study to characterize sex differences, incidence, and spatial patterns in portugal,” *In Vivo (Brooklyn)*, vol. 34, no. 5, pp. 2711–2719, Sep. 2020, doi: 10.21873/invivo.12092.

[18] M. Soares *et al.*, “Real-world treatment patterns and survival outcomes for advanced non-small cell lung cancer in the pre-immunotherapy era in Portugal: A retrospective analysis from the I-O Optimise initiative,” *BMC Pulm Med*, vol. 20, no. 1, Sep. 2020, doi: 10.1186/s12890-020-01270-z.

[19] B. Veloso, J. R. Nogueira, and M. F. Cardoso, “Lung cancer and indoor radon exposure in the north of Portugal - An ecological study,” *Cancer Epidemiol*, vol. 36, no. 1, Feb. 2012, doi: 10.1016/j.canep.2011.10.005.

[20] G. Bronte and C. Rolfo, “Semi-automated volumetric analysis in the NELSON trial for lung cancer screening: Is there room for diagnostic experience yet,” 2016, *AME Publishing Company*. doi: 10.21037/jtd.2016.11.36.

[21] S. H. Shetty, S. Shetty, C. Singh, and A. Rao, “Supervised Machine Learning: Algorithms and Applications,” in *Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools, and Applications*, 2022, pp. 1–16. doi: 10.1002/9781119821908.ch1.

[22] Z. Abdalhussain and A. S. Abdalrada, “Comprehensive Review of Machine Learning Approaches for Alzheimer’s Disease Diagnosis and Prognosis.” [Online]. Available: https://ijeaa.cultechpub.com/index.php/ijeaa

[23] N. V. D. S. S. V. Prasad Raju and P. N. Devi, “A Comparative Analysis of Machine Learning Algorithms for Big Data Applications in Predictive Analytics,” *International*

Journal of Scientific Research and Management (IJSRM), vol. 12, no. 10, pp. 1608–1630, Oct. 2024, doi: 10.18535/ijsrm/v12i10.ec09.

[24] J. Sharma, “Review of Machine Learning Algorithms and Their Applications,” *International Journal of Innovations in Science Engineering And Management*, pp. 113–119, Jun. 2024, doi: 10.69968/ijisem.2024v3i2113-119.

[25] J. E. van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Mach Learn*, vol. 109, no. 2, pp. 373–440, Feb. 2020, doi: 10.1007/s10994-019-05855-6.

[26] S. Selvarajan, H. Manoharan, A. O. Khadidos, and A. O. Khadidos, “Testing of Emerging Wireless Sensor Networks Using Radar Signals With Machine Learning Algorithms,” *IEEE Journal of Selected Areas in Sensors*, vol. 1, pp. 49–59, May 2024, doi: 10.1109/jsas.2024.3395578.

[27] I. D. Mienye and N. Jere, “Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction,” *Information (Switzerland)*, vol. 15, no. 7, Jul. 2024, doi: 10.3390/info15070394.

[28] F. Trindade Neves, M. Aparicio, and M. de Castro Neto, “The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities,” *Applied Sciences (Switzerland)*, vol. 14, no. 5, Mar. 2024, doi: 10.3390/app14052209.

[29] C. Cortes, “Support-Vector Networks,” 1995.

[30] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2001. doi: 10.7551/mitpress/4175.001.0001.

[31] L. Breiman, “Random Forests,” 2001.

[32] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable importance measures: Illustrations, sources and a solution,” *BMC Bioinformatics*, vol. 8, 2007, doi: 10.1186/1471-2105-8-25.

[33] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” 1997.

[34] R. E. . Schapire and Yoav. Freund, *Boosting : foundations and algorithms*. MIT Press, 2012.

[35] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[36] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Nov. 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

[37] I. Markoulidakis and G. Markoulidakis, “Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis,” *Technologies (Basel)*, vol. 12, no. 7, Jul. 2024, doi: 10.3390/technologies12070113.

[38] U. Anand, P. Vyshnavi, S. Sarvan, M. Gowtham, Nesarani. A, and Y. D. M. M, “Evaluation of Machine Learning Approaches for Predicting Cardiovascular Attacks,” in *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, 2024, pp. 468–474. doi: 10.1109/ICSCNA63714.2024.10864232.

[39] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O’Sullivan, “A review of model evaluation metrics for machine learning in genetics and genomics,” 2024, *Frontiers Media SA*. doi: 10.3389/fbinf.2024.1457619.

[40] L. Aissaoui Ferhi, M. Ben Amar, F. Choubani, and R. Bouallegue, “Enhancing diagnostic accuracy in symptom-based health checkers: a comprehensive machine learning approach with clinical vignettes and benchmarking,” *Front Artif Intell*, vol. 7, 2024, doi: 10.3389/frai.2024.1397388.

[41] G. Krivorotov, “Machine learning-based profit modeling for credit card underwriting - implications for credit risk: Machine learning-based profit modeling for credit card underwriting,” *J Bank Financ*, vol. 149, Apr. 2023, doi: 10.1016/j.jbankfin.2023.106785.

[42] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

[43] M. Abdelhamid and A. Desai, “Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification,” Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.19751

[44] J. Allgaier and R. Pryss, “Cross-Validation Visualized: A Narrative Guide to Advanced Methods,” *Mach Learn Knowl Extr*, vol. 6, no. 2, pp. 1378–1388, Jun. 2024, doi: 10.3390/make6020065.

[45] F. Emmert-Streib and M. Dehmer, “Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error,” Mar. 01, 2019, *MDPI*. doi: 10.3390/make1010032.

[46] Y. Yang, L. Xu, L. Sun, P. Zhang, and S. S. Farid, “Machine learning application in personalised lung cancer recurrence and survivability prediction,” *Comput Struct Biotechnol J*, vol. 20, pp. 1811–1820, Jan. 2022, doi: 10.1016/j.csbj.2022.03.035.

[47] Y. Li, D. Ge, J. Gu, F. Xu, Q. Zhu, and C. Lu, “A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies,” *BMC Cancer*, vol. 19, 2019, doi: 10.1186/S12885-019-6101-7.

[48] A. J. Didier, A. Nigro, Z. Noori, M. A. Omballi, S. M. Pappada, and D. M. Hamouda, “Application of machine learning for lung cancer survival prognostication—A systematic review and meta-analysis,” 2024, *Frontiers Media SA*. doi: 10.3389/frai.2024.1365777.

[49] H. Y. J. Kang, E. Batbaatar, D. W. Choi, K. S. Choi, M. Ko, and K. S. Ryu, “Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy,” *JMIR Med Inform*, vol. 11, no. 1, Jan. 2023, doi: 10.2196/47859.

[50] R. Chang, S. Qi, Y. Wu, Y. Yue, X. Zhang, and W. Qian, “Nomograms integrating CT radiomic and deep learning signatures to predict overall survival and progression-free survival in NSCLC patients treated with chemotherapy,” *Cancer Imaging*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s40644-023-00620-4.

[51] H. Zheng *et al.*, “Development of a Three-Dimensional Multi-Modal Perfusion-Thermal Electrode System for Complete Tumor Eradication,” *Cancers (Basel)*, vol. 14, no. 19, Oct. 2022, doi: 10.3390/cancers14194768.

[52] B. Melosky *et al.*, “Worldwide Prevalence of Epidermal Growth Factor Receptor Mutations in Non-Small Cell Lung Cancer: A Meta-Analysis,” Jan. 01, 2022, *Adis*. doi: 10.1007/s40291-021-00563-1.

[53] H. Jeon, S. Wang, J. Song, H. Gill, and H. Cheng, “Update 2025: Management of Non-Small-Cell Lung Cancer,” Dec. 01, 2025, *Springer*. doi: 10.1007/s00408-025-00801-x.

[54] T. Li, W. Ma, and E. Al-Obeidi, “Evolving Precision First-Line Systemic Treatment for Patients with Unresectable Non-Small Cell Lung Cancer,” Jul. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/cancers16132350.

[55] F. Huemer *et al.*, “Baseline absolute lymphocyte count and ECOG performance score are associated with survival in advanced non-small cell lung cancer undergoing PD-1/PD-L1 blockade,” *J Clin Med*, vol. 8, no. 7, Jul. 2019, doi: 10.3390/jcm8071014.

[56] K. Sehgal *et al.*, “Association of Performance Status with Survival in Patients with Advanced Non-Small Cell Lung Cancer Treated with Pembrolizumab Monotherapy,” *JAMA Netw Open*, vol. 4, no. 2, Feb. 2021, doi: 10.1001/jamanetworkopen.2020.37120.

[57] C. Chayangsu, J. Khorana, C. Charoentum, V. Sriuranpong, J. Patumanond, and A. Tantraworasin, “Exploring Prognostic Factors and Survival Outcomes in Advanced Non-Small Cell Lung Cancer Patients Undergoing First-Line Chemotherapy in Limited-Resource Settings,” *J Clin Med*, vol. 14, no. 2, Jan. 2025, doi: 10.3390/jcm14020335.

[58] D. H. Choi *et al.*, “Effectiveness and safety of amivantamab in EGFR exon 20 insertion (E20I) mutations in non-small cell lung cancer (NSCLC),” *Transl Lung Cancer Res*, vol. 12, no. 12, pp. 2448–2459, 2023, doi: 10.21037/tlcr-23-643.

[59] W.-Z. Zhong, Q. Zhou, and Y.-L. Wu, “The resistance mechanisms and treatment strategies for EGFR-mutant advanced non-small-cell lung cancer,” 2017. [Online]. Available: www.impactjournals.com/oncotarget/

[60] Z. Yao *et al.*, “Real-World Data on Prognostic Factors for Overall Survival in EGFR Mutation-Positive Advanced Non-Small Cell Lung Cancer Patients Treated with First-Line Gefitinib”, doi: 10.1634/theoncologist.2016.

[61] M. J. Vinolo-Gil, C. Herrera-Sánchez, F. J. Martin-Vega, R. Martín-Valero, G. Gonzalez-Medina, and V. Pérez-Cabezas, “Efficacy of tele-rehabilitation in patients with chronic obstructive pulmonary disease: a systematic review,” May 01, 2022, *Gobierno de Navarra*. doi: 10.23938/ASSN.0999.

[62] A. Desilets, M. Repetto, and A. Drilon, “ Repotrectinib: Redefining the therapeutic landscape for patients with ROS1 fusion-driven non-small cell lung cancer ,” *Clin Transl Med*, vol. 14, no. 10, Oct. 2024, doi: 10.1002/ctm2.70017.

[63] L. M. Forrest, D. C. McMillan, C. S. McArdle, W. J. Angerson, and D. J. Dunlop, “Comparison of an inflammation-based prognostic score (GPS) with performance status (ECOG) in patients receiving platinum-based chemotherapy for inoperable non-small-cell lung cancer,” *Br J Cancer*, vol. 90, no. 9, pp. 1704–1706, May 2004, doi: 10.1038/sj.bjc.6601789.

[64] J. H. Cabot and E. G. Ross, “Evaluating prediction model performance,” *Surgery*, vol. 174, no. 3, pp. 723–726, Sep. 2023, doi: 10.1016/j.surg.2023.05.023.

[65] Y. Liu, Y. Li, and D. Xie, “Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks,” *J Stat Comput Simul*, vol. 94, no. 1, pp. 183–203, 2024, doi: 10.1080/00949655.2023.2238235.

[66] B. J. Parker, S. Günter, and J. Bedo, “Stratification bias in low signal microarray studies,” *BMC Bioinformatics*, vol. 8, Sep. 2007, doi: 10.1186/1471-2105-8-326.

[67] S. Farhadpour, T. A. Warner, and A. E. Maxwell, “Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices,” *Remote Sens (Basel)*, vol. 16, no. 3, Feb. 2024, doi: 10.3390/rs16030533.

[68] S. J. J. Guesné, T. Hanser, S. Werner, S. Boobier, and S. Scott, “Mind your prevalence!,” *J Cheminform*, vol. 16, no. 1, Dec. 2024, doi: 10.1186/s13321-024-00837-w.

[69] Y. Yang and G. Mirzaei, “Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification,” *PLoS One*, vol. 19, no. 2 February, Feb. 2024, doi: 10.1371/journal.pone.0293607.

[70] A. K. Azlim Khan and N. H. Ahamed Hassain Malim, “Comparative Studies on Resampling Techniques in Machine Learning and Deep Learning Models for Drug-Target Interaction Prediction,” Feb. 01, 2023, *MDPI*. doi: 10.3390/molecules28041663.

[71] S. Moreno *et al.*, “A Radiogenomics Ensemble to Predict EGFR and KRAS Mutations in NSCLC,” *Tomography*, vol. 7, no. 2, pp. 154–168, Apr. 2021, doi: 10.3390/tomography7020014.

[72] N. Q. K. Le, Q. H. Kha, V. H. Nguyen, Y. C. Chen, S. J. Cheng, and C. Y. Chen, “Machine learning-based radiomics signatures for egfr and kras mutations prediction in non-small-cell lung cancer,” *Int J Mol Sci*, vol. 22, no. 17, Sep. 2021, doi: 10.3390/ijms22179254.

[73] I. Shiri *et al.*, “Next-Generation Radiogenomics Sequencing for Prediction of EGFR and KRAS Mutation Status in NSCLC Patients Using Multimodal Imaging and Machine Learning Algorithms,” *Mol Imaging Biol*, vol. 22, no. 4, pp. 1132–1148, Aug. 2020, doi: 10.1007/s11307-020-01487-8.

[74] X. Tan *et al.*, “Predicting EGFR mutation, ALK rearrangement, and uncommon EGFR mutation in NSCLC patients by driverless artificial intelligence: a cohort study,” *Respir Res*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12931-022-02053-2.

[75] T. Zhang *et al.*, “Simultaneous identification of egfr, kras, erbb2, and tp53 mutations in patients with non-small cell lung cancer by machine learning-derived three-dimensional radiomics,” *Cancers (Basel)*, vol. 13, no. 8, Apr. 2021, doi: 10.3390/cancers13081814.

[76] L. Haixian, P. Shu, L. Zhao, L. Chunfeng, and L. Lun, “Machine learning approaches for EGFR mutation status prediction in NSCLC: an updated systematic review,” 2025, *Frontiers Media SA*. doi: 10.3389/fonc.2025.1576461.

[77] J. Hemmatian, R. Hajizadeh, and F. Nazari, “Addressing imbalanced data classification with Cluster-Based Reduced Noise SMOTE,” *PLoS One*, vol. 20, no. 2 February, Feb. 2025, doi: 10.1371/journal.pone.0317396.