

**NSCLC360**  
**LEVERAGING MULTIOMICS DATA FOR**  
**PERSONALIZED LUNG CANCER PROGNOSIS**  
**THROUGH INTEGRATED HEALTH PROFILES**

24-25J-211

Project Proposal Report

Arudchayan Pirabakaran

B.Sc. (Hons) in Information Technology Specializing in Data  
Science

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology Sri  
Lanka

August 2024



## **ACKNOWLEDGMENT**

I would like to express my deepest and most heartfelt appreciation to my supervisor, Mr. Samadhi Rathnayake, and my co-supervisor, Ms. Thisara Shyamalee, from the Faculty of Computing. Their unwavering support, expert guidance, and dedicated mentorship have been invaluable as I embark on this research project. Throughout this journey, their insightful feedback and constructive critiques have been instrumental in shaping the direction of my work, refining the focus of my research, and ultimately enhancing the overall quality and depth of my study. Their commitment to excellence and their encouragement have motivated me to push the boundaries of my knowledge and strive for the highest standards in my research.


In addition, I would like to acknowledge the significant contributions of my fellow team members. Their collaboration, shared insights, and mutual support have been a source of inspiration and have greatly enriched my understanding of the research topic. The collective effort and synergy of the team have not only broadened my perspective but have also fostered a spirit of camaraderie and intellectual curiosity that has propelled our research forward.

I am also deeply grateful to Dr. Nuradh Joseph, our external supervisor with extensive expertise in clinical oncology. His invaluable assistance and specialized knowledge in this area have been crucial as we initiated this research project. Dr. Joseph's guidance has provided us with a deeper understanding of the clinical aspects of our study, and his support has been instrumental in bridging the gap between theoretical research and practical application in the field of oncology.

## DECLARATION

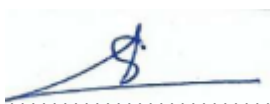
I declare that this is our own work and this proposal does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name	Student ID	Signature
P Arudchayan	IT21190698	

The supervisor/s should certify the proposal report with the following declaration.

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.



.....  
**Signature of the supervisor**

...08/23/2024....

**Date:**



.....  
**Signature of the co-supervisor**

...08/23/2024....

**Date:**

## ABSTRACT

This report presents a comprehensive approach for enhancing prognostic analysis in Non-Small Cell Lung Cancer (NSCLC) by integrating multi-omic data with Explainable Artificial Intelligence (XAI). The proposed solution, NSCLC360, consists of four primary components designed to address key challenges in NSCLC management. These components include: (1) Lung Cancer Image Analysis for early and accurate detection using advanced deep learning models; (2) An Explainable Quantitative Approach for Enhancing Prognostic Analysis in Non-Small Cell Lung Cancer (3) Side Effect Prediction for lung cancer treatments to improve patient quality of life through personalized risk assessments; and (4) Recurrence Prediction using multimodal data to develop personalized post-operative treatment plans.

NSCLC is a complex and often late-diagnosed cancer with significant challenges in personalized treatment and prognosis. Current methods lack comprehensive integration of diverse biological data and sufficient transparency in predictive models. This research aims to address these gaps by leveraging multi-omic data and applying XAI techniques to develop a more accurate and interpretable prognostic model. The expected outcomes include improved prognostic accuracy, better personalization of treatment, and enhanced trust in AI-driven recommendations, ultimately leading to more effective and data-driven cancer care.

The report emphasizes the implementation, functionality, and requirements of integrating XAI into the prognostic model, focusing on ensuring transparency and usability in clinical settings. The proposed system aims to provide a holistic and user-friendly approach to NSCLC management, leveraging readily available data and technologies.

**Keywords:** Multi-omics, Explainable AI, NSCLC, Prognostic Modeling, Machine Learning, Biomarker Identification, Treatment Efficacy, Recurrence Prediction, Data Integration, Clinical Decision Support

## TABLE OF CONTENTS

ACKNOWLEDGMENT .....	i
ABSTRACT .....	iii
LIST OF ABBREVIATIONS .....	vi
1. INTRODUCTION .....	1
1.1 Background & Literature Survey .....	3
1.2 Research Gap .....	4
2. OBJECTIVES.....	7
2.1 Main Objective.....	7
2.2 Specific Objectives.....	7
3. METHODOLOGY .....	8
3.1 Project Overview.....	8
3.2 Individual Component.....	9
4. RESEARCH & DEVELOPMENT OVERVIEW.....	16
4.1 Sources for Test Data and Analysis .....	16
4.2 Anticipated Benefits .....	16
4.3 Expected Research Outcomes .....	17
4.4 Research Constraints .....	17
4.5 Project Plan .....	18
CONCLUSION.....	21
REFERENCES .....	22

## **LIST OF FIGURES**

Figure 1:High level diagram of the proposed system.....	9
Figure 2:Model Training Process .....	11
Figure 3: WBS.....	19
Figure 4: Gantt Chart.....	20

## LIST OF ABBREVIATIONS

<b>Abbreviation</b>	<b>Full Form</b>
<b>NSCLC</b>	Non-Small Cell Lung Cancer
<b>XAI</b>	Explainable Artificial Intelligence
<b>ML</b>	Machine Learning
<b>TCGA</b>	The Cancer Genome Atlas
<b>SHAP</b>	SHapley Additive exPlanations
<b>LIME</b>	Local Interpretable Model-agnostic Explanations
<b>SEER</b>	Surveillance, Epidemiology, and End Results
<b>PLCO</b>	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial



# 1. INTRODUCTION

Prognosis, the prediction of disease outcomes based on an individual's health status and disease characteristics, plays a crucial role in guiding treatment decisions. By leveraging multiomics data, which includes various biological layers such as genomics, proteomics, and metabolomics, researchers can uncover the intricate molecular diversity and heterogeneity present within tumors. This integrated approach offers the potential for a more accurate and personalized prognosis, diagnosis, and treatment of diseases like Non-Small Cell Lung Cancer (NSCLC), a subtype of lung cancer characterized by its late diagnosis and complex molecular landscape [1].

However, one of the significant challenges in applying advanced ML models, particularly in healthcare, is the lack of transparency or interpretability of these models, often referred to as the "black box" problem. This has led to the emergence of Explainable AI (XAI), which aims to make ML models more transparent, interpretable, and trustworthy. In the context of NSCLC prognosis, integrating XAI methods can help clinicians understand how different multiomics data contribute to the model's predictions, thereby improving trust and adoption of these technologies in clinical settings [2].

Despite the promise of multiomics and XAI, current research has primarily focused on individual omics layers, such as genomics or proteomics, rather than integrating them. This has limited the ability to fully understand the interplay of various factors affecting disease progression and treatment outcomes. A comprehensive, multimodal approach that incorporates data from multiple omics layers, along with XAI techniques, is needed to provide a holistic and interpretable view of NSCLC and improve precision medicine strategies [3]. This research aims to fill this gap by integrating multiomics data with XAI to develop predictive models for NSCLC prognosis, treatment outcomes, and recurrence, thereby advancing personalized and explainable medicine in lung cancer care.

### **Nature of the Solution: NSCLC360**

To address the complexities of NSCLC management, the proposed solution, "NSCLC360," is structured around four key components:

1. **Lung Cancer Image Analysis:** This component focuses on developing a platform for early and accurate cancer detection using advanced deep learning models. By analyzing medical imaging data, it can classify lung cancer types, identify tumor locations and sizes, and ensure patient data privacy through homomorphic encryption.
2. **An Explainable Quantitative Approach for Enhancing Prognostic Analysis in Non-Small Cell Lung Cancer:** This component focuses on identifying factors that influence outcomes and treatment decisions in NSCLC. By analyzing clinical, genetic, and environmental data, it aims to create a prognostic model that is both accurate and explainable. This approach enhances the ability to predict patient outcomes and tailor treatments, ultimately improving survival rates and quality of care.
3. **Predict Side Effects of Lung Cancer Treatments:** This predictive modeling component anticipates and manages the side effects of lung cancer treatments. It utilizes personalized risk assessments and real-time data to inform adaptive treatment strategies, improving patient quality of life through multidisciplinary efforts.
4. **Recurrence Prediction for NSCLC Using Multimodal and Multiomics Data:** By adapting various data models, this component provides a classification system that identifies patients at low or high risk of recurrence. This allows for personalized post-operative treatment plans, optimizing long-term patient care.

NSCLC360 thus aims to deliver a comprehensive and personalized approach to NSCLC management by leveraging the power of multiomics data and machine learning.

## 1.1 Background & Literature Survey

Non-Small Cell Lung Cancer (NSCLC) is the most common form of lung cancer, accounting for approximately 85% of all lung cancer cases. Prognostic analysis in NSCLC is crucial for personalized treatment planning and improving patient outcomes. Traditional approaches, such as clinical staging and genomic profiling, have significantly advanced our understanding of NSCLC. However, these methods often lack the ability to integrate and interpret complex multi-omic data, leading to suboptimal prognostic accuracy.

Several studies have explored various methods to enhance prognostic analysis in NSCLC. For instance, genomic profiling tools like FoundationOne CDx [4] have provided detailed insights into genetic mutations but are limited by their narrow focus on genomic data without integrating clinical or imaging data. Additionally, IBM Watson for Oncology [5] uses artificial intelligence to analyze large volumes of medical literature and patient data, providing treatment recommendations. However, the lack of transparency in AI decision-making processes can hinder trust and clinical adoption.

Research has also investigated the integration of multi-omic data to improve prognostic accuracy. A study demonstrated that combining genomic, transcriptomic, and proteomic data could enhance the prediction of treatment responses [6]. Similarly, another study highlighted the potential of integrating imaging data with genomic information to provide a more comprehensive prognostic assessment. However, these approaches often suffer from challenges related to data integration and model interpretability.[7]

The concept of Explainable Artificial Intelligence (XAI) has emerged as a solution to address these challenges. XAI aims to make AI models more transparent and understandable to clinicians by providing clear explanations of the decision-making process. A recent study applied XAI techniques to cancer prognostic models, demonstrating improved trust and usability in clinical settings.[8]

Despite these advancements, there remain significant gaps in integrating multi-omic data, ensuring model interpretability, and translating research findings into practical

clinical tools. This research aims to address these gaps by developing a quantitative approach that incorporates XAI to enhance prognostic analysis in NSCLC.

## 1.2 Research Gap

The existing literature reveals several limitations and gaps in the current approaches to prognostic analysis in Non-Small Cell Lung Cancer (NSCLC):

1. **Limited Integration of Multi-Omic Data:** While genomic profiling tools like FoundationOne CDx provide detailed genetic information, they do not integrate other relevant data types, such as clinical and imaging data. This lack of integration limits the ability to provide a comprehensive prognostic assessment.
2. **Lack of Model Interpretability:** AI-based tools, such as IBM Watson for Oncology, offer advanced analytical capabilities but often lack transparency in their decision-making processes. This opacity can hinder clinicians' ability to trust and effectively utilize the recommendations provided by these systems.
3. **Inadequate Validation Across Diverse Populations:** Many studies and tools have been validated in specific populations or controlled environments, leading to limited generalizability. There is a need for validation in diverse patient populations to ensure the robustness and applicability of prognostic models.
4. **Underrepresentation of Rare Genetic Variants:** Most existing models focus on common genetic mutations, potentially overlooking the prognostic significance of rare genetic variants. These rare variants could play a crucial role in individual patient outcomes, and their inclusion in prognostic models could enhance personalized treatment strategies.
5. **Suboptimal Integration of Emerging Biomarkers:** New biomarkers, including those derived from liquid biopsies and immune profiling, are rapidly emerging as important factors in NSCLC prognosis. However, these biomarkers are not yet widely integrated into existing models, potentially missing opportunities to improve prognostic accuracy and patient stratification.
6. **Simplistic Consideration of Biomarkers:** Existing prognostic models often focus solely on the presence or absence of specific biomarkers, without

considering the levels of these biomarkers or the potential interactions between different biomarkers. This approach can oversimplify the complexity of NSCLC, as the levels and combinations of multiple biomarkers may provide a more nuanced and accurate prognostic assessment.

7. **Insufficient Focus on Treatment Response Prediction:** Current prognostic models often emphasize overall survival or disease progression, with less attention given to predicting individual responses to specific treatments. A more precise prediction of treatment efficacy could guide clinicians in selecting the most appropriate therapeutic options for each patient.

To address these gaps, this research will focus on:

- Developing a comprehensive prognostic model that integrates multi-omic data.
- Incorporating XAI techniques to enhance model transparency and usability.
- Validating the model across diverse patient cohorts.
- Analyzing biomarker levels and their interactions for a more refined prognostic assessment.
- Including rare genetic variants to enhance personalized treatment strategies.
- Integrating emerging biomarkers to improve prognostic accuracy and patient stratification.
- Enhancing treatment response prediction to optimize therapeutic decision-making.

### **1.3 Research Problem**

The primary research problem is the need for a more effective, interpretable, and comprehensive approach to prognostic analysis in Non-Small Cell Lung Cancer (NSCLC). Existing methods often fall short in several key areas: they tend to overlook the integration of diverse data types, such as multi-omic, clinical, and imaging data; they fail to provide transparent insights into the decision-making process, which can erode trust and usability among clinicians; and they do not adequately account for the complexity of biomarker interactions or the dynamic nature of disease progression over time. Additionally, current models often neglect the inclusion of rare genetic variants and emerging biomarkers, which could be critical for personalized treatment strategies.

This research aims to address these limitations by developing an explainable, quantitative approach that not only integrates multi-omic data but also incorporates Explainable AI (XAI) techniques to enhance model transparency. The model will be rigorously validated in diverse, real-world clinical settings, ensuring its robustness and applicability. Furthermore, this approach will focus on identifying factors that significantly affect patient outcomes and treatment choices, ultimately leading to more personalized and effective therapeutic strategies.

## 2. OBJECTIVES

### 2.1 Main Objective

To develop an advanced quantitative approach for enhancing prognostic analysis in NSCLC by integrating multi-omic data and incorporating Explainable Artificial Intelligence (XAI) to provide transparent and interpretable insights for personalized treatment planning.

### 2.2 Specific Objectives

1. **Identifying Potential Prognostic Biomarkers:** Discover and validate novel biomarkers by integrating multi-omic data (genomic, transcriptomic, proteomic) and ensuring transparency in the identification process using XAI techniques.
2. **Identifying Treatment Efficacy:** Analyze the interaction between biomarkers and treatments, and evaluate treatment responses over time, incorporating XAI to provide clear explanations of the model's predictions.
3. **Creating a Prognostic Model:** Develop and validate an advanced prognostic model that integrates diverse data types (genomic, clinical, imaging) and incorporates XAI features to enhance clinical decision-making and model interpretability.
4. **Ensuring Model Explainability and Transparency:**  
Guarantee that the prognostic model's decisions are clear and understandable by employing XAI techniques to make its outputs and decision-making processes transparent.
5. **Integrating into Clinical Decision Support Systems:**  
Seamlessly incorporate the prognostic model into clinical workflows by designing a user-friendly interface and integrating it with existing decision support systems to aid in informed decision-making.

### 3. METHODOLOGY

#### 3.1 Project Overview

The methodology for this project, titled NSCLC360, is designed to address the limitations of current research by integrating multi-omic data with Explainable Artificial Intelligence (XAI) techniques. Current research often focuses on individual omics layers, such as genomics or proteomics, without integrating them to understand the interplay of various factors affecting disease progression and treatment outcomes. NSCLC360 aims to provide a holistic and interpretable approach to NSCLC management by incorporating data from multiple omics layers along with XAI. The project is structured around four key components:

1. **Lung Cancer Image Analysis:** This component involves developing an advanced platform for early and accurate detection of lung cancer using deep learning models. By analyzing medical imaging data, the system will classify lung cancer types, determine tumor locations and sizes, and ensure data privacy through homomorphic encryption. This early detection is crucial for timely and effective intervention.
2. **An Explainable Quantitative Approach for Enhancing Prognostic Analysis in Non-Small Cell Lung Cancer:** This component focuses on identifying factors that influence outcomes and treatment decisions in NSCLC. By analyzing clinical, genetic, and environmental data, it aims to create a prognostic model that is both accurate and explainable. This approach enhances the ability to predict patient outcomes and tailor treatments, ultimately improving survival rates and quality of care.
3. **Predicting Side Effects of Lung Cancer Treatments:** This predictive modeling component will anticipate and manage the side effects of lung cancer treatments. By utilizing personalized risk assessments and real-time data, the system will inform adaptive treatment strategies, thereby improving patient quality of life and enabling more targeted management of treatment-related side effects.



#### 4. Recurrence Prediction for NSCLC Using Multimodal and Multiomics Data:

This component will develop a classification system to identify patients at low or high risk of disease recurrence. By integrating diverse data models, it will support the creation of personalized post-operative treatment plans, optimizing long-term patient care and monitoring.

By integrating these components, NSCLC360 aims to deliver a comprehensive, data-driven approach to NSCLC management, enhancing prognostic accuracy, treatment personalization, and model transparency. This methodology is designed to bridge current research gaps and advance precision medicine strategies in lung cancer care.

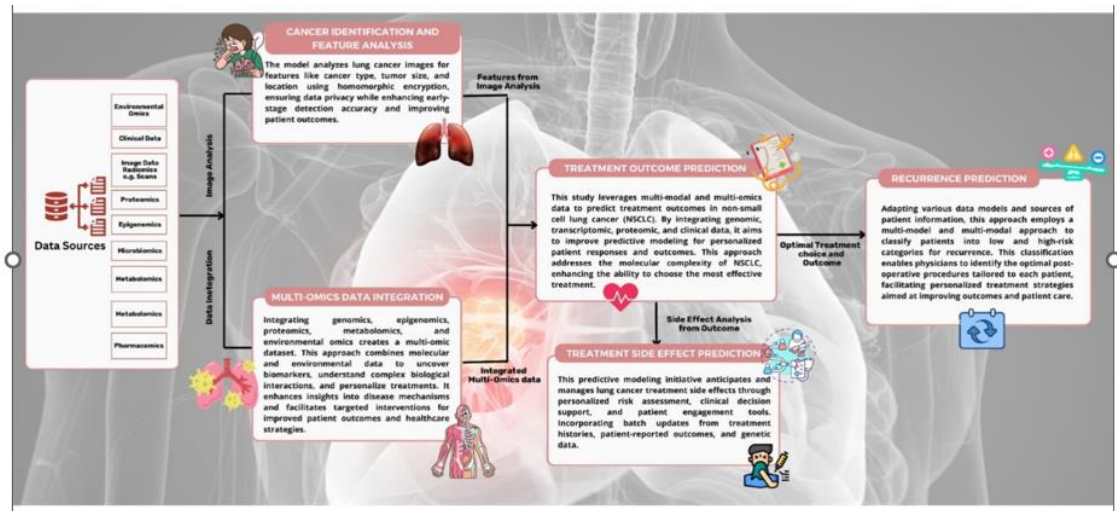


Figure 1: High level diagram of the proposed system

### 3.2 Individual Component

#### An Explainable Quantitative Approach for Enhancing Prognostic Analysis in Non-Small Cell Lung Cancer

##### 3.2.1 Overview

Non-Small Cell Lung Cancer (NSCLC) remains one of the most prevalent and challenging forms of lung cancer, characterized by its complexity and variability in patient outcomes. The prognosis for NSCLC patients is influenced by a range of factors, including clinical characteristics, genetic profiles, and environmental exposures. Current prognostic models often fall short in their ability to integrate these diverse factors comprehensively while remaining interpretable for clinical use.

This research seeks to address this gap by developing an explainable quantitative approach to enhance prognostic analysis in NSCLC. Our objective is to create a robust and transparent model that accurately predicts patient outcomes and informs treatment decisions. By leveraging a combination of clinical, genetic, and environmental data, we aim to build a model that not only provides reliable prognostic insights but also offers clear explanations of its predictions.

The significance of this approach lies in its potential to improve patient stratification, personalize treatment plans, and ultimately enhance survival rates and quality of care. Through detailed analysis and integration of multifaceted data, our goal is to advance the field of cancer prognosis and support clinicians in making more informed decisions for their patients.

### **3.2.2 Explainable Artificial Intelligence (XAI)**

The integration of Explainable Artificial Intelligence (XAI) is pivotal in enhancing the transparency and interpretability of our prognostic model for Non-Small Cell Lung Cancer (NSCLC). This incorporation addresses the critical need for models that are not only accurate but also comprehensible to clinicians. The XAI framework employed in this research encompasses several key methodologies:

- **Feature Importance Analysis:** We apply XAI techniques to assess and visualize the significance of individual features in influencing model predictions. This analysis elucidates which clinical, genetic, and environmental variables exert the greatest impact on the model's outcomes, thereby providing a deeper understanding of the factors driving prognosis.
- **Decision Explanations:** To facilitate clinical applicability, we provide detailed, human-readable explanations for the model's predictions. This aspect of XAI ensures that the rationale behind the model's recommendations is transparent,

thereby enabling healthcare practitioners to interpret and act upon the insights with greater confidence and clarity.

- **Model Visualization:** We employ advanced visualization techniques to illustrate the model's decision boundaries and the underlying prediction logic. These visual tools are designed to enhance the interpretability of the model, allowing for a clearer representation of how data inputs are processed and how predictions are derived.

The integration of XAI in this prognostic model not only improves its transparency but also supports the clinical decision-making process by making the model's operations and outputs more accessible and understandable. This approach aims to foster greater trust in the model and facilitate its practical application in personalized patient care, ultimately contributing to more informed treatment strategies for NSCLC.

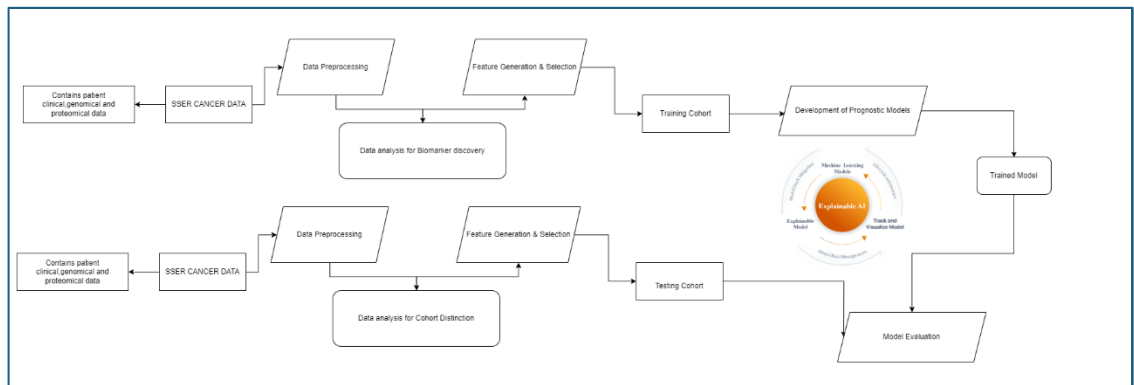


Figure 2: Model Training Process

### 3.3 Commercialization

To facilitate the transition of the "NSCLC360" clinical decision support system from research to clinical practice, three primary commercialization models are proposed:

**Direct Licensing to Healthcare Providers:** This model involves licensing the NSCLC360 system directly to hospitals, oncology centers, and clinical practices.

By embedding the system within existing healthcare infrastructures, this approach allows for the integration of advanced decision support tools into clinical workflows, accompanied by ongoing technical support and system updates.

**Strategic Partnerships with Biotechnology Firms:** Forming strategic alliances with biotechnology companies that specialize in cancer diagnostics and therapeutics presents a viable commercialization pathway. Such partnerships can capitalize on the firms' established market presence, distribution channels, and complementary technologies, thereby enhancing the reach and adoption of NSCLC360.

**Subscription-Based SaaS Model:** The NSCLC360 system can be offered as a Software-as-a-Service (SaaS) solution, where access is provided through a subscription fee. This model supports continuous system updates and technical support while establishing a recurring revenue stream. The SaaS model also facilitates scalability and flexibility in deployment, catering to a diverse range of healthcare settings.

Each of these models presents unique advantages and opportunities for scaling the NSCLC360 system, ultimately contributing to its successful integration into clinical practice and improving decision-making processes in the management of Non-Small Cell Lung Cancer.

### **3.4 Functional Requirements**

The NSCLC360 clinical decision support system must fulfill the following functional requirements to effectively support the management of Non-Small Cell Lung Cancer (NSCLC):

#### Lung Cancer Image Analysis:

**Image Processing:** The system shall support the ingestion and analysis of medical imaging data in various formats, including but not limited to CT scans and MRI.

**Tumor Detection and Classification:** The system must employ advanced deep learning algorithms to accurately detect and classify tumor types, identify tumor locations, and determine tumor sizes.

**Data Privacy:** The system shall utilize homomorphic encryption to ensure the confidentiality and privacy of patient data throughout the analysis process.

#### Prognostic Analysis:

**Data Integration:** The system must integrate clinical, genetic, and environmental data to facilitate comprehensive prognostic analysis.

**Outcome Prediction:** The system shall provide accurate predictions of patient outcomes based on the integrated data.

**Explainability:** The system must offer clear, interpretable explanations of the prognostic predictions to support clinical decision-making.

#### Side Effect Prediction:

**Risk Assessment:** The system must utilize predictive modeling techniques to estimate the likelihood of side effects associated with lung cancer treatments.

**Adaptive Treatment Strategies:** The system should generate recommendations for modifying treatment plans to mitigate anticipated side effects based on real-time data.

#### Recurrence Prediction:

**Multimodal Data Analysis:** The system must analyze multimodal and multiomics data to assess the risk of cancer recurrence.

**Risk Classification:** The system shall classify patients into risk categories (low or high) for recurrence, informing post-operative treatment plans.

**Personalized Treatment Plans:** The system must develop tailored post-operative care strategies based on recurrence risk assessments.

### 3.5 Non-Functional Requirements

The NSCLC360 system must also satisfy the following non-functional requirements to ensure its effectiveness, reliability, and usability:

#### Performance:

**Response Time:** The system shall process and analyze medical images and related data within an acceptable time frame to support timely clinical decision-making.

**Scalability:** The system must be scalable to accommodate increasing data volumes and user demands without performance degradation.

#### Reliability:

**Availability:** The system must ensure high availability and minimal downtime to support continuous clinical operations.

**Fault Tolerance:** The system should incorporate fault tolerance mechanisms to handle and recover from errors or failures without data loss.

#### Usability:

**User Interface:** The system must feature an intuitive and user-friendly interface to facilitate ease of use by clinicians.

**Training and Support:** Comprehensive training and user support must be provided to ensure effective system utilization.

#### Security:

**Data Protection:** The system must implement stringent security measures, including encryption and access controls, to protect patient data.

**Regulatory Compliance:** The system must adhere to relevant regulatory standards and data protection laws, such as GDPR and HIPAA.

#### Maintainability:

**Modular Architecture:** The system should be designed with a modular architecture to facilitate updates and ongoing maintenance.

Documentation: Detailed documentation must be provided, including functional descriptions, user manuals, and technical support materials.

Interoperability:

Integration: The system must be compatible with existing healthcare systems and adhere to relevant standards (e.g., HL7, FHIR) to enable seamless data exchange and integration.

## 4. RESEARCH & DEVELOPMENT OVERVIEW

### 4.1 Sources for Test Data and Analysis

For the creation and validation of **NSCLC360**, the following sources will be utilized:

- **Clinical Trial Datasets:** Data from both ongoing and completed clinical trials will be leveraged to assess the model's performance. Notable datasets include the SEER database and the PLCO database. These sources will provide a comprehensive basis for evaluating the model's accuracy and reliability.
- **Genomic Databases:** The Cancer Genome Atlas (TCGA) will be employed to integrate genomic data into the model. TCGA offers extensive genomic information crucial for enhancing the model's prognostic capabilities and understanding of NSCLC at a molecular level.
- **Patient Records:** Anonymized patient records will be analyzed to assess the model's applicability in practical scenarios. This analysis will provide insights into how well the model performs with real-world data and its effectiveness in clinical settings.

The analytical methods utilized will include advanced statistical analysis, machine learning algorithms, and Explainable Artificial Intelligence (XAI) techniques. These methodologies will ensure a comprehensive evaluation and validation of the model, facilitating a robust assessment of its performance.

### 4.2 Anticipated Benefits

The implementation of the **NSCLC360** system is expected to yield several key benefits:

- **Enhanced Prognostic Accuracy:** By integrating multi-omic data and employing XAI techniques, the system is anticipated to significantly improve the accuracy of prognostic assessments, offering more precise predictions regarding patient outcomes.



- **Personalized Treatment Options:** The system will provide data-driven, personalized treatment recommendations, thereby enabling more tailored therapeutic strategies based on individual patient profiles.
- **Increased Model Transparency:** The incorporation of XAI techniques will enhance the interpretability of the model's outputs, fostering greater trust and usability among clinicians and supporting the practical application of the system in clinical environments.

### 4.3 Expected Research Outcomes

The research is expected to deliver the following outcomes:

- **Validated Biomarkers:** The identification and validation of novel prognostic biomarkers through multi-omic data integration. This outcome will contribute valuable insights into the biological underpinnings of NSCLC.
- **Treatment Efficacy Analysis:** A detailed analysis of treatment responses and interactions with biomarkers, providing interpretable results that inform therapeutic efficacy.
- **Prognostic Model:** A fully functional prognostic model incorporating XAI features, ready for integration into clinical practice. This model will support clinical decision-making by providing actionable insights into patient prognosis and treatment options.

### 4.4 Research Constraints

Several constraints may impact the research, including:

- **Data Privacy Issues:** Ensuring compliance with data protection regulations, such as GDPR and HIPAA, is crucial when handling and analyzing patient data. Adherence to these regulations will be strictly maintained to protect patient confidentiality.

- **Complexity of XAI Integration:** The integration of XAI techniques may introduce additional complexity into the model. This complexity could affect the development process and the overall performance of the system.
- **High Implementation Costs:** The development, validation, and commercialization of the model are associated with significant costs. These include expenses related to technology development, clinical trials, and regulatory compliance. Therefore the focus will be utilizing second hand data to create a prognostic model which if promising then proceeds to actual trials

## 4.5 Project Plan

The project plan will be structured as follows:

- **Timeline:** A comprehensive schedule outlining key milestones and deliverables will be established. This timeline will detail the phases of the project, from initial development through to final validation and implementation.
- **Milestones:** Specific goals and deadlines will be set for each phase of the project. Milestones will include data collection, model development, validation, and preparation for clinical integration.
- **Resource Allocation:** A detailed budget and resource plan will be developed, encompassing personnel, software, data acquisition, and other project-related expenditures. This plan will ensure that resources are effectively allocated to meet project objectives.

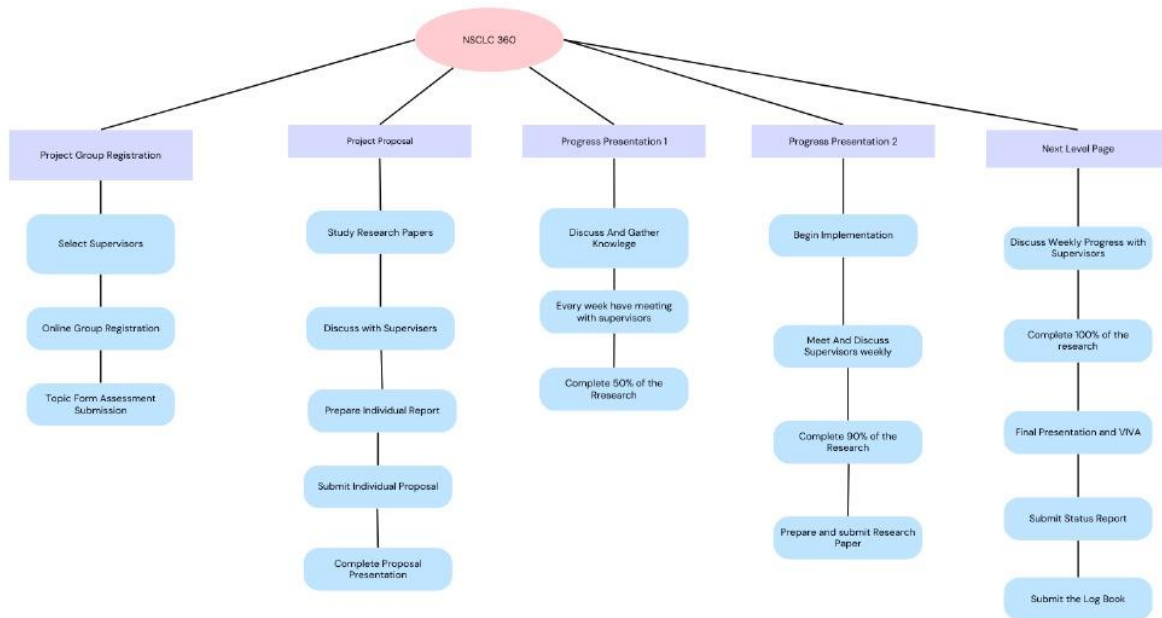


Figure 3: WBS

## NSCLC360 Gantt chart- 24-25J-211

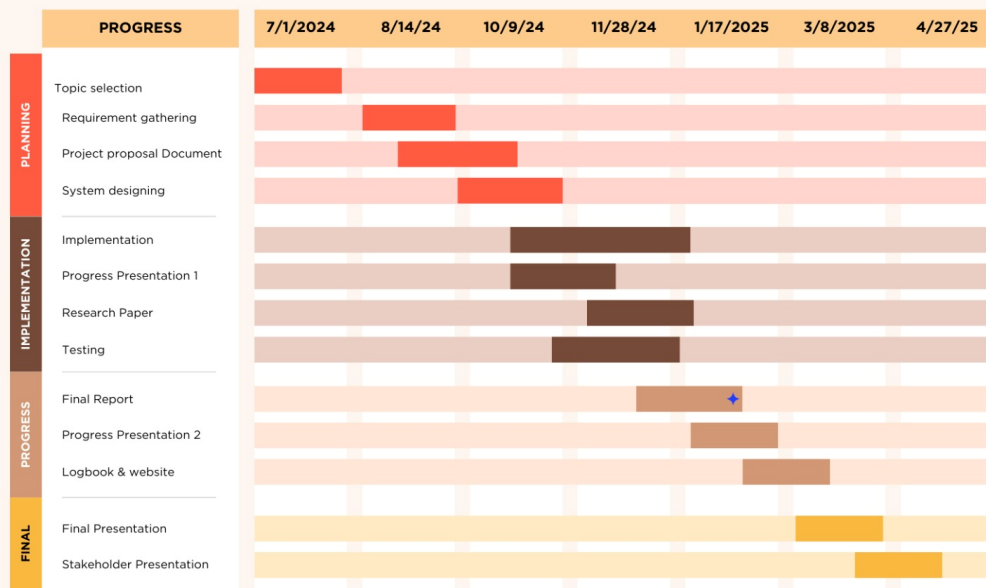


Figure 4: Gantt Chart

## **CONCLUSION**

This research aims to address critical gaps in NSCLC prognostic analysis by developing a comprehensive, interpretable approach that integrates multi-omic data and incorporates Explainable Artificial Intelligence (XAI). By enhancing prognostic accuracy, personalizing treatment recommendations, and improving model transparency, this study seeks to contribute to more effective and data-driven cancer care.

## REFERENCES

- [1] P. Kent, C. Cancelliere, E. Boyle, J. D. Cassidy, and A. Kongsted, “A conceptual framework for prognostic research,” *BMC Medical Research Methodology*, vol. 20, no. 1, Jun. 2020, doi: <https://doi.org/10.1186/s12874-020-01050-7>.
- [2] T. Hulsen *et al.*, “From Big Data to Precision Medicine,” *Frontiers in Medicine*, vol. 6, no. 34, Mar. 2019, doi: <https://doi.org/10.3389/fmed.2019.00034>.
- [3] V. Raufaste-Cazavieille, R. Santiago, and A. Droit, “Multi-omics analysis: Paving the path toward achieving precision medicine in cancer treatment and immuno-oncology,” *Frontiers in Molecular Biosciences*, vol. 9, Oct. 2022, doi: <https://doi.org/10.3389/fmolb.2022.962743>.
- [4] C. A. Milbury *et al.*, “Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors,” *PLOS ONE*, vol. 17, no. 3, p. e0264138, Mar. 2022, doi: <https://doi.org/10.1371/journal.pone.0264138>.
- [5] N. Zhou *et al.*, “Concordance Study Between IBM Watson for Oncology and Clinical Practice for Patients with Cancer in China,” *The Oncologist*, vol. 24, no. 6, pp. 812–819, Sep. 2018, doi: <https://doi.org/10.1634/theoncologist.2018-0255>.
- [6] Y. J. Heo, C. Hwa, G.-H. Lee, J.-M. Park, and J.-Y. An, “Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes,” *Molecules and Cells*, vol. 44, no. 7, pp. 433–443, Jul. 2021, doi: <https://doi.org/10.14348/molcells.2021.0042>.
- [7] M. Martínez-García and E. Hernández-Lemus, “Data Integration Challenges for Machine Learning in Precision Medicine,” *Frontiers in Medicine*, vol. 8, Jan. 2022, doi: <https://doi.org/10.3389/fmed.2021.784455>.
- [8] S. S. Makubhai, G. R. Pathak, and P. R. Chandre, “Predicting lung cancer risk using explainable artificial intelligence,” *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 2, pp. 1276–1285, Apr. 2024, doi: <https://doi.org/10.11591/eei.v13i2.6280>.
- [9] W. Che *et al.*, “How to use the Surveillance, Epidemiology, and End Results (SEER) data: research design and methodology,” *Military Medical Research*, vol. 10, no. 1, Oct. 2023, doi: <https://doi.org/10.1186/s40779-023-00488-2>.
- [10] C. S. Zhu *et al.*, “The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial and Its Associated Research Resource,” *JNCI Journal of the National Cancer Institute*, vol. 105, no. 22, pp. 1684–1693, Oct. 2013, doi: <https://doi.org/10.1093/jnci/djt281>.

