NSCLC360: LEVERAGING MULTIOMICS DATA FOR PERSONALIZED LUNG CANCER

PROGNOSIS THROUGH INTEGRATED HEALTH PROFILES

24-25J-211

Project Proposal Report

Irfan Nawaz Abdul Azeez

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**
Irfan N.A.A	IT21331022	

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.

……………………………….. …08/23/2024…

Signature of the supervisor **Date:**

……………………………….. …08/23/2024…

Signature of the co-supervisor **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) Predictive Modeling of NSCLC treatment outcomes by integrating genomic, transcriptomic, proteomic, and clinical data; (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

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
NSCLC	Non-Small Cell Lung Cancer
XAI	Explainable Artificial Intelligence
ML	Machine Learning
TCGA	The Cancer Genome Atlas
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations

1. INTRODUCTION

The prediction of recurrence of a disease based an individual's health status and disease characteristics, plays a crucial role in guiding treatment decisions and post treatment care. 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 and better predict the locality and time frame of the recurrence of cancer. This integrated approach offers the potential for a more accurate and personalized recurrence prediction of diseases like Non-Small Cell Lung Cancer (NSCLC), a subtype of lung cancer characterized by its late diagnosis and complex molecular landscape (Kent et al., 2020)[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 recurrence, 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 (Hulsen et al., 2019)[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 (Raufaste-Cazavieille et al., 2022)[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. **Predicting Outcomes for NSCLC Treatment Using Multimodal and Multiomics Data**: This component integrates genomic, transcriptomic, proteomic, and clinical data to enhance predictive modeling for personalized treatment outcomes. By addressing the molecular complexity of NSCLC, it aims to improve the choice of treatment and patient outcomes.
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 recurrence analysis in NSCLC. For example, Aryan Ghazipour et al have researched into using post radiation therapy ct images with RNN/CNN deep learning to predict the survival of patients from Stereotactic body radiation therapy. And Jaryd R. Christie et all has researched into Predicting recurrence risks in lung cancer patients using multimodal radiomics and random survival forests. And most importantly, Panyanat Aonpong et all has researched into Improved Genotype-Guided Deep Radiomics Signatures for Recurrence Prediction of Non-Small Cell Lung Cancer. Yet these studies lack true explainability and integration of multi modal data. They only consider radiomics and genomics. Disregarding proteomic and clinical data. As well as Panyanat Aonpong et all is using genomic data as a predictive element in recurrence prediction. It is not integrated into the feature pipeline, but is merely used as a validation dataset.

Research has also investigated the integration of multi-omic data to improve recurrence accuracy. A study by Smith et al. [3] demonstrated that combining genomic, transcriptomic, and proteomic data could enhance the prediction of treatment responses. Similarly, Lee et al. [4] 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.

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 by Zhang et al. [5] applied XAI techniques to cancer prognostic models, demonstrating improved trust and usability in clinical settings.

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 NSCLC:

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.

1. **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.
2. **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.
3. **Challenges in Translating Research into Clinical Practice**: Despite advancements in multi-omic data integration and XAI, translating these research findings into practical, user-friendly clinical tools remains a challenge.

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 populations and real-world clinical settings.

1.3 Research Problem

The primary research problem is the need for an effective and interpretable approach to predict the recurrence of NSCLC. Existing models often fail to integrate diverse data types and provide transparent insights into the decision-making process. This research aims to develop a quantitative approach that addresses these limitations by integrating multi-omic data, incorporating XAI techniques, and validating the model in real-world settings so that better post operative care can be provided to patients

2. OBJECTIVES

2.1 Main Objective

To develop an advanced quantitative approach for enhancing recurrence prediction in NSCLC by integrating multi-omic data and incorporating Explainable Artificial Intelligence (XAI) to provide transparent and interpretable insights for personalized post operative care.

2.2 Specific Objectives

1. **Identifying Potential Recurrence features**: Discover and validate features by integrating multi-omic data (genomic, transcriptomic, proteomic) and ensuring transparency in the identification process using XAI techniques.
2. **Identifying Locality based features**: Analyze the interaction between features and recurrence localities and time period, incorporating XAI to provide clear explanations of the model's predictions.
3. **Creating a Recurrence prediction Model**: Develop and validate an advanced recurrence prediction model that integrates diverse data types (genomic, clinical, imaging) and incorporates XAI features to enhance post operative care and model interpretability.

3. METHODOLOGY

3.1 Project Overview

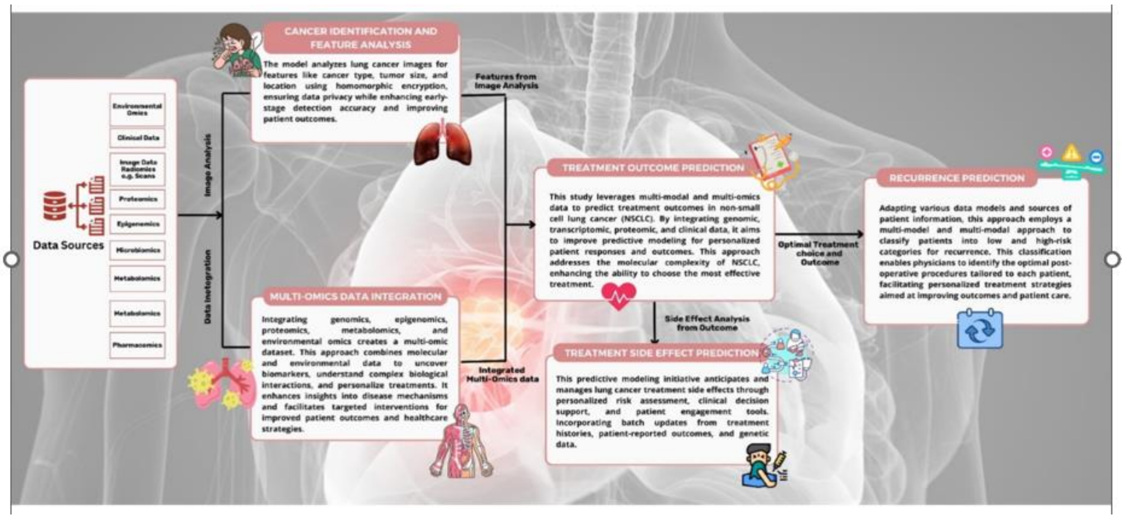
The methodology for this project, titled NSCLC360, is designed to address the limitations of current research by integrating multi-omic data with Explainable Artificial Intelligence (XAI) techniques. Current research often focuses on individual omics layers, such as genomics or proteomics, without integrating 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. **Predicting Outcomes for NSCLC Treatment Using Multimodal and Multiomics Data:** This component focuses on integrating genomic, transcriptomic, proteomic, and clinical data to enhance predictive modeling for personalized treatment outcomes. By addressing the molecular complexity of NSCLC, this component aims to refine treatment choices and improve patient outcomes, providing a comprehensive view of how various biological factors influence disease progression and response to treatment.
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.


Data Sources
Environmental Omics
Clinical Data
Image Data Radiomics e.g. Scans
Proteomics
Epigenomics
Microbiomics
Metabolomics
Metabolomics
Pharmacomics
Image Analysis
Data Inetegration
Features from Image Analysis
Integrated Multi-Omics data
Optimal Treatment choice and Outcome
Side Effect Analysis from Outcome

CANCER IDENTIFICATION AND FEATURE ANALYSIS

The model analyzes lung cancer images for features like cancer type, tumor size, and location using homomorphic encryption, ensuring data privacy while enhancing early-stage detection accuracy and improving patient outcomes.

MULTI-OMICS DATA INTEGRATION

Integrating genomics, epigenomics, proteomics, metabolomics, and environmental omics creates a multi-omic dataset. This approach combines molecular and environmental data to uncover biomarkers, understand complex biological interactions, and personalize treatments. It enhances insights into disease mechanisms and facilitates targeted interventions for improved patient outcomes and healthcare strategies.

TREATMENT OUTCOME PREDICTION

This study leverages multi-modal and multi-omics data to predict treatment outcomes in non-small cell lung cancer (NSCLC). By integrating genomic, transcriptomic, proteomic, and clinical data, it aims to improve predictive modeling for personalized patient responses and outcomes. This approach addresses the molecular complexity of NSCLC, enhancing the ability to choose the most effective treatment.

TREATMENT SIDE EFFECT PREDICTION

This predictive modeling initiative anticipates and manages lung cancer treatment side effects through personalized risk assessment, clinical decision support, and patient engagement tools. Incorporating batch updates from treatment histories, patient-reported outcomes, and genetic data.

RECURRENCE PREDICTION

Adapting various data models and sources of patient information, this approach employs a multi-model and multi-modal approach to classify patients into low and high-risk categories for recurrence. This classification enables physicians to identify the optimal post-operative procedures tailored to each patient, facilitating personalized treatment strategies aimed at improving outcomes and patient care.

Figure 1:High level diagram of the proposed system

3.2 Individual Component

3.2.1 Algorithm

Develop advanced algorithms to analyze multi-omic data and identify key features for cancer recurrence locality, recurrence time period and patient stratification. This includes:

- **Data Integration**: Combine genomic, transcriptomic, and proteomic data using data fusion techniques.

- **Model Development**: Employ machine learning algorithms to identify patterns and correlations between features and recurrence.
- **XAI Techniques**: Apply XAI methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to ensure that model predictions are interpretable and understandable.

3.2.2 XAI Integration

Incorporate XAI to enhance model transparency and interpretability:

- **Feature Importance Analysis**: Use XAI techniques to identify and visualize the most important features influencing model predictions.
- **Decision Explanations**: Provide clear, human-readable explanations for the model's predictions, helping clinicians understand the rationale behind recommendations.
- **Model Visualization**: Create visual tools to represent the model's decision boundaries and prediction logic, facilitating better understanding and trust in the model's outputs.

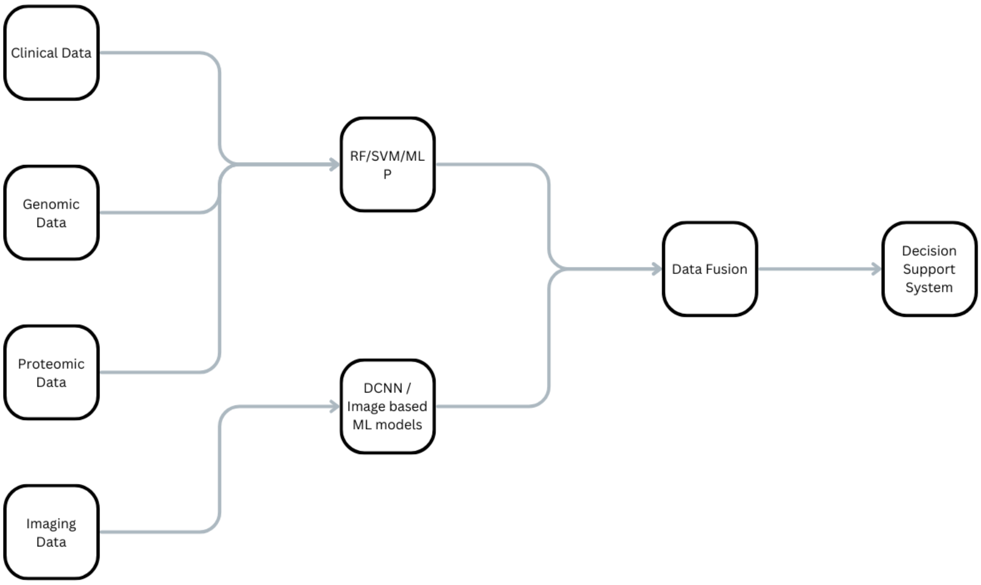


Figure 2:Overall component diagram

3.2.3 Commercialization

Explore pathways for commercializing the developed prognostic model:

- **Partnerships**: Collaborate with healthcare providers and technology companies to integrate the model into clinical practice.
- **Regulatory Approval**: Ensure the model meets regulatory standards for clinical use, including data privacy and security requirements.
- **User Training**: Provide training and support for clinicians to effectively utilize the model and XAI features in their practice.

3.2.4 Functional Requirements

- **Accurate Feature Identification**: The system must accurately identify and validate features.
- **Transparent Recurrence time period / locality prediction**: The system should provide interpretable insights into recurrence locality, time period and feature interactions.
- **User-Friendly Decision Support System**: Develop a decision support tool that integrates XAI features and is easy for clinicians to use.

3.2.5 Non-Functional Requirements

- **Scalability**: Ensure the model and XAI features can scale to handle large volumes of data and diverse patient populations.
- **Data Security and Privacy**: Implement robust measures to protect patient data and ensure compliance with data privacy regulations.
- **System Reliability and Performance**: The system should be reliable and perform efficiently in real-world clinical settings.

4. RESEARCH & DEVELOPMENT OVERVIEW

4.1 Sources for Test Data and Analysis

Data sources will include:

- **Clinical Trial Datasets**: Access data from ongoing and completed clinical trials to validate the model.
- **Genomic Databases**: Use databases such as The Cancer Genome Atlas (TCGA) for genomic data.
- **Patient Records**: Analyze anonymized patient records to assess model performance and applicability.

Analytical methods will include statistical analysis, machine learning, and XAI techniques to ensure comprehensive evaluation and validation.

4.2 Anticipated Benefits

- **Enhanced Recurrence Accuracy**: Improve the accuracy of recurrence assessments by integrating multi-omic data and XAI techniques.
- **Personalized Post-Treatment Options**: Provide more personalized and data-driven post-treatment care.

- **Increased Model Transparency**: Enhance trust and usability of the model through clear and interpretable AI outputs.

4.3 Scope and Specified Deliverables/Expected Research Outcome

Deliverables include:

- **Validated Features**: A comprehensive list of novel features validated through multi-omic data integration.
- **Cancer Locality predcition**: Detailed analysis of cancer locality and feature interaction with interpretable results.
- **Recurrence prediction Model**: A functional recurrence prediction model with XAI features, ready for clinical integration and use.

4.4 Research Constraints

- **Data Privacy Issues**: Ensuring compliance with data protection regulations while handling patient data.
- **Complexity of XAI Integration**: Integrating XAI techniques into the model may introduce additional complexity.
- **High Implementation Costs**: Potential costs associated with developing, validating, and commercializing the model.

4.5 Project Plan

The project plan will include:

- **Timeline:** Detailed schedule outlining key milestones and deliverables.
- **Milestones:** Specific goals and deadlines for each phase of the project.
- **Resource Allocation**: Budget and resources required for each component of the project, including personnel, software, and data acquisition.

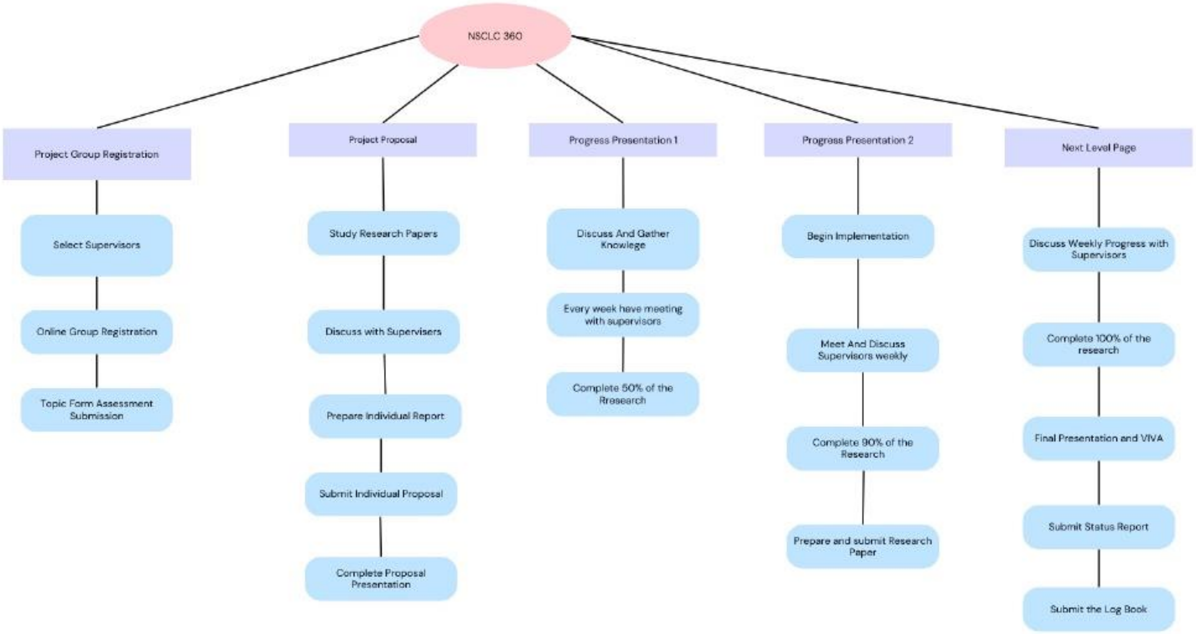

NSCLC 360
Project Group Registration
Select Supervisors
Online Group Registration
Topic Form Assessment Submission
Project Proposal
Study Research Papers
Discuss with Supervisers
Prepare Individual Report
Submit Individual Proposal
Complete Proposal Presentation
Progress Presentation 1
Discuss And Gather Knowlege
Every week have meeting with supervisors
Complete 50% of the Rresearch
Progress Presentation 2
Begin Implementation
Meet And Discuss Supervisors weekly
Complete 90% of the Research
Prepare and submit Research Paper
Next Level Page
Discuss Weekly Progress with Supervisors
Complete 100% of the research
Final Presentation and VIVA
Submit Status Report
Submit the Log Book

Figure 3: WBS

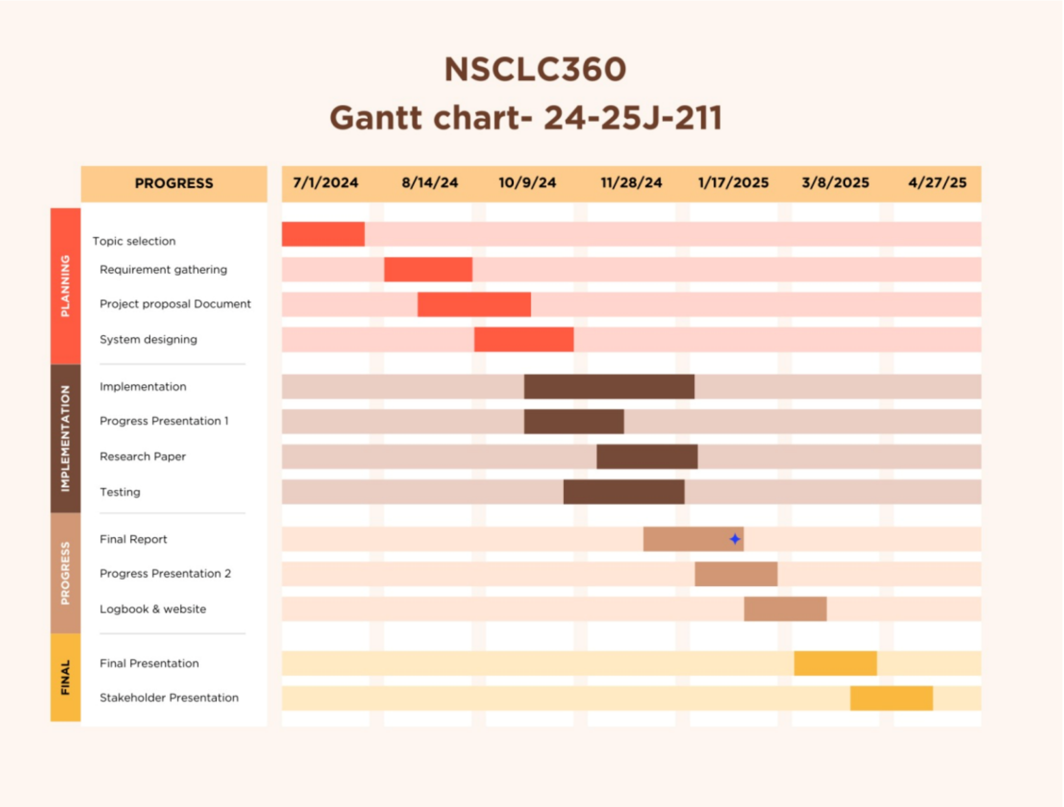


Figure 4: Gantt Chart

5. BUDGET AND BUDGET JUSTIFICATION

Provide a detailed budget including:

- **Data Acquisition**: Costs associated with obtaining and processing clinical and genomic data.
- **Software and Tools**: Expenses for software licenses, XAI tools, and computational resources.
- **Validation and Testing**: Costs related to validating the model and conducting clinical trials.

Justify each expense in terms of its contribution to achieving the research objectives and ensuring model interpretability.

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] Kent, P., Cancelliere, C., Boyle, E., Cassidy, J.D. and Kongsted, A. (2020). A conceptual framework for prognostic research. BMC Medical Research Methodology, 20(1). doi: https://doi.org/10.1186/s12874-020-01050-7

[2] Hulsen, T., Jamuar, S.S., Moody, A.R., Karnes, J.H., Varga, O., Hedensted, S., Spreafico, R., Hafler, D.A. and McKinney, E.F. (2019). From Big Data to Precision Medicine. Frontiers in Medicine, [online] 6(34). doi:https://doi.org/10.3389/fmed.2019.00034.

[3] Raufaste-Cazavieille, V., Santiago, R. and Droit, A. (2022). Multi-omics analysis: Paving the path toward achieving precision medicine in cancer treatment and immuno-oncology. Frontiers in Molecular Biosciences, 9. doi:https://doi.org/10.3389/fmolb.2022.962743.

[4] Christie JR, Daher O, Abdelrazek M, Romine PE, Malthaner RA, Qiabi M, Nayak R, Napel S, Nair VS, Mattonen SA. Predicting recurrence risks in lung cancer patients using multimodal radiomics and random survival forests. J Med Imaging (Bellingham). 2022 Nov;9(6):066001. doi: 10.1117/1.JMI.9.6.066001. Epub 2022 Nov 8. PMID: 36388142; PMCID: PMC9641263.

[5] Janik A, Torrente M, Costabello L, Calvo V, Walsh B, Camps C, Mohamed SK, Ortega AL, Nováček V, Massutí B, Minervini P, Campelo MRG, Del Barco E, Bosch-Barrera J, Menasalvas E, Timilsina M, Provencio M. Machine Learning-Assisted Recurrence Prediction for Patients With Early-Stage Non-Small-Cell Lung Cancer. JCO Clin Cancer Inform. 2023 Jul;7:e2200062. doi: 10.1200/CCI.22.00062. PMID: 37428988; PMCID: PMC10569772.

[6] Aonpong P, Iwamoto Y, Han XH, Lin L, Chen YW. Improved Genotype-Guided Deep Radiomics Signatures for Recurrence Prediction of Non-Small Cell Lung Cancer. Annual Int Conf IEEE Eng Med Biol Soc. 2021 Nov;2021:3561-3564. doi: 10.1109/EMBC46164.2021.9630703. PMID: 34892008.