

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

# 22BIO201: Intelligence of Biological Systems – 1

November – 2024

Project Report

**AI-Based Pathway Analysis for Understanding Disease Mechanisms**

**GROUP: 5**

**Group Members**

ADITYA K SAMANT - CB.SC.U4AIE23012

JAYAN SUBRAMANIAN – CB.SC.U4AIE23033

Date of Submission: 11/11/2024

Signature of the Project Supervisor:

|  |  |
| --- | --- |
| **TITLE** | **PAGE NUMBER** |
| 1. ABSTRACT | 3 |
| 2. INTRODUCTION | 4 |
| 3. RELATED WORKS | 7 |
| 4. METHODOLOGY | 9 |
| 5. EXPERIMENTS | 11 |
| 6. RESULTS | 14 |
| 7. CONCLUSION | 17 |
| 8. FUTURE WORKS | 18 |
| 9. REFRENCES | 19 |

# ABSTRACT

"Lung adenocarcinoma is a complex disease with multifaceted biological pathways that require a deep understanding to develop effective treatments. This project introduces an AI-powered tool, 'PathInsight,' designed to analyze and interpret the molecular mechanisms underpinning disease progression. Using a Graph Attention Network (GAT) model, the application processes and integrates data from genomics, transcriptomics, and proteomics to identify high-risk patterns and pathways associated with lung adenocarcinoma. A user-friendly web application interfaces with the model, offering risk predictions based on patient data and delivering actionable insights. Additionally, the app leverages the Maps API to provide users with tailored recommendations for healthcare providers, resources, and support locations nearby. This report details the model architecture, data processing pipeline, and technological framework, demonstrating how 'PathInsight' can enhance our understanding of lung adenocarcinoma mechanisms and support timely, data-driven interventions. By combining advanced analytics with localized support, this innovation aims to empower clinicians and patients in managing lung adenocarcinoma effectively."

# INTRODUCTION

## Background and Motivation:

## Lung adenocarcinoma is a leading cause of cancer-related deaths, and understanding its complex molecular mechanisms is crucial for effective diagnosis, treatment, and patient outcomes. Despite significant advancements in cancer research, the pathways underlying the disease remain poorly understood, making it challenging to personalize treatment plans. With the increasing volume of genomic, transcriptomic, and proteomic data available, there is a growing need for advanced AI tools that can analyze this data to uncover hidden patterns and relationships. Our project, "AI-Based Pathway Analysis for Understanding Disease Mechanisms," seeks to leverage AI to bridge this gap, providing deeper insights into lung adenocarcinoma.

## Problem Statement:

## Lung adenocarcinoma’s intricate disease mechanisms are influenced by genetic, transcriptomic, and proteomic factors that are difficult to analyze using traditional methods. Current approaches lack a comprehensive model that integrates these data sources to predict disease progression and personalize treatment. Additionally, clinicians and patients need easy access to actionable information about risk levels and available support resources. This project addresses these challenges by using a Graph Attention Network (GAT) to analyze multi-omics data and provide risk predictions, while also offering a web-based application that integrates these insights with localized healthcare recommendations.

## Project Objectives:

The primary objective of this project is to develop an AI-driven platform that provides actionable insights into lung adenocarcinoma mechanisms. The platform features include:

* **Advanced risk prediction** using a Graph Attention Network (GAT) to analyze genomics, transcriptomics, and proteomics data.
* **Personalized recommendations** for actionable next steps based on individual risk profiles.
* **Web application** that interfaces with the AI model to deliver results and utilizes the Maps API to suggest nearby healthcare providers and resources.

This tool aims to empower clinicians and patients with data-driven insights that enhance decision-making and support better health outcomes for lung adenocarcinoma patients.

**Technical Overview:**

Our application utilizes cutting-edge technologies in machine learning, AI, and web development to provide a comprehensive solution for lung adenocarcinoma pathway analysis and risk prediction:

* **Graph Attention Network (GAT):** This deep learning model is employed to analyze and integrate multi-omics data (genomics, transcriptomics, and proteomics) to identify key disease mechanisms and predict patient-specific risks for lung adenocarcinoma.
* **Web Application Interface:** The web-based platform connects users to the AI model, offering an intuitive interface for accessing personalized risk predictions and actionable insights. The application facilitates easy interaction with the model and visualizes risk data in a user-friendly manner.
* **Maps API:** Integrated into the app to suggest healthcare providers, treatment centers, and support resources based on user location. This feature provides a seamless experience for patients and clinicians to find localized options for care and support.

This combination of advanced AI modeling, web technologies, and location-based services empowers users with a tool that not only predicts risk but also supports informed decision-making and better healthcare accessibility.

**Contributions:**

Each team member contributed to different aspects of the project, and their individual efforts are summarized below:

**Jayan Subramanian**:

* Developed the frontend of the web application using React.js, styled with Tailwind CSS and ShadCN components, providing an intuitive user interface for data input and results visualization.
* Implemented backend services with FastAPI to facilitate seamless communication between the frontend and the model hosted in Google Colab.
* Conducted data preprocessing on raw genomic data from the GDC, including normalization, feature extraction, and structuring, ensuring compatibility with the Graph Neural Network (GAT) model.
* Utilized Python libraries for efficient data handling and managed API endpoints for real-time data processing and prediction requests.
* Integrated Maps API to enhance user experience by displaying relevant and nearby facilities based on user needs and requirements

**Aditya Samant**:

* Designed and implemented the Graph Neural Network (GAT) model architecture in Google Colab, focusing on pathway analysis for disease mechanism understanding.
* Trained the GNN model on synthetic and real genomic data, refining it to accurately predict disease pathways and molecular interactions.
* Applied forward and backward propagation techniques to optimize model performance, fine-tuning weights and minimizing the loss function.
* Also built a small scale and more easily explainable model to remove all the black box perceptions about the Graph Neural Network (GAT).

# RELATED WORKS

## AI-Driven Pathway Analysis and Risk Prediction Applications:

With the increasing complexity of cancer research, several AI-based platforms have emerged to analyze molecular data and provide insights into disease mechanisms. These applications focus on leveraging multi-omics data, such as genomics and proteomics, to understand cancer progression. Examples include tools like OncoKB and CancerMine, which provide insights into cancer-specific mutations and therapeutic options. However, these tools often focus on general cancer research and lack integration with personalized risk prediction and localized healthcare recommendations, particularly for specific cancer types like lung adenocarcinoma.

Additionally, AI-based platforms in oncology are frequently used for diagnostic purposes, such as detecting mutations or analyzing imaging data. However, there is limited integration of pathway analysis with actionable, localized recommendations, making it harder for clinicians and patients to directly translate insights into practical care decisions.

## Gaps and Issues in Existing Solutions:

Despite progress in AI-based cancer research tools, there are several gaps in current solutions:

* **Lack of Integration Across Multi-Omics Data:** Many existing platforms focus on isolated omics data types (genomics, transcriptomics, or proteomics) without integrating them into a comprehensive analysis. This fragmented approach limits the ability to fully understand the complex interactions driving cancer progression, especially in lung adenocarcinoma.
* **Absence of Personalized Risk Prediction:** While several tools can analyze molecular data, few offer personalized risk predictions that tailor insights to individual patients' data, limiting the clinical utility of these platforms.
* **Limited Accessibility and Actionable Insights:** While some applications offer insights into cancer mechanisms, they often lack localized healthcare recommendations or actionable next steps for patients and clinicians. Additionally, existing platforms are not designed to guide users to nearby healthcare facilities or support services.

In Summary, while there are numerous AI-driven tools and mobile health applications that focus on analyzing cancer-related data, they often fall short in providing personalized, actionable insights that integrate multi-omics data and localized support. Our project addresses these gaps by using a Graph Attention Network (GAT) for comprehensive pathway analysis, delivering personalized lung adenocarcinoma risk predictions, and integrating a web application with Maps API for localized healthcare recommendations. This platform is designed to empower clinicians and patients with deeper, actionable insights, improving

decision-making and access to care.

# METHODOLOGY

# Our project, "AI-Based Pathway Analysis for Understanding Disease Mechanisms," is structured with an integrated framework of frontend, backend, and data sources to deliver personalized insights for lung adenocarcinoma. The methodology combines advanced technologies in data processing, machine learning, and web development, resulting in an interactive platform for disease pathway analysis and risk prediction. Below is a step-by-step breakdown of the components and data flow as illustrated:

# Frontend Interface (React.js with ShadCN Library and Tailwind CSS):

### The user-facing frontend is built using React.js, styled with the ShadCN library and Tailwind CSS for efficient, responsive design. Users access the platform through this interface, providing initial data inputs on their health profile, which forms the basis for downstream analysis. Tailwind CSS ensures a smooth, consistent look and feel, while ShadCN enables reusable components for an optimized user experience.

### Backend Infrastructure (FastAPI):

### FastAPI serves as the core of the backend, handling data input from the frontend and routing it to the machine learning model. This API framework is designed for high-performance interaction, managing requests for prediction generation and data retrieval. FastAPI coordinates communication between the React.js frontend and the GAT model, ensuring efficient data flow.

### Graph Neural Network Model (Google Colab):

### The application’s disease mechanism analysis relies on a Graph Neural Network (GAT) model developed in Google Colab. This model utilizes multi-omics data (genomics, transcriptomics, and proteomics) to assess patient risk for lung adenocarcinoma. The GAT is trained on multi-dimensional datasets to analyze pathway interconnections, yielding prediction results based on disease risk scores.

### Data Processing and Libraries (Python):

Python libraries support data preprocessing and analysis before model training. These libraries process raw data, converting it into a format suitable for the GAT model, and perform essential steps like normalization, feature selection, and transformation. By streamlining the data preparation, this component enhances model accuracy and efficiency.

### Data Source (GDC - Genomic Data Commons):

#### The multi-omics data required for pathway analysis is sourced from the Genomic Data Commons (GDC). This rich data repository provides structured genomic, transcriptomic, and proteomic data for cancer research, offering insights into disease mechanisms and progression. Data from GDC is processed and fed into the GAT model, enabling comprehensive, personalized pathway analysis.

The combination of these components allows for efficient, AI-driven pathway analysis, leveraging real-world data to deliver risk predictions and actionable insights in a user-friendly, web-based platform.

**Flow diagram**

A diagram of a computer

Description automatically generated

# EXPERIMENTS

The project leverages machine learning and graph neural networks (GATs) to analyze biological pathways and understand disease mechanisms at a molecular level. Here, we outline the data sources, preprocessing steps, model architectures, and the roles of various components within the application.

### Dataset

#### Genomic Data (GDC):

* **Source:** The Genomic Data Commons (GDC) provides high-quality genomic data crucial for our disease pathway analysis. We collected relevant datasets from GDC, specifically targeting genes and molecular interactions associated with cancer pathways and disease progression.
* **Preprocessing:** To prepare the data for model training, we processed the raw genomic information, normalizing it and converting it into a format suitable for input into our GAT model. This step ensures consistent data handling and accurate downstream analysis by structuring data for compatibility with graph-based representations.

#### Training Data for GAT:

* **Data Augmentation and Labeling:** Since biological data often lacks structured labels, we generated synthetic data to supplement the training process. This dataset simulates potential relationships among genes and biological pathways, providing a structured foundation for training our GAT model to detect and predict molecular interactions in actual patient data.

#### Knowledge Base for Disease Pathways:

* **Data Compilation:** We constructed a knowledge base that integrates information from biomedical literature, existing pathway databases, and scholarly articles on disease mechanisms. This curated knowledge base includes molecular interactions, disease-gene associations, and pathway-specific knowledge essential for pathway analysis and understanding disease mechanisms.

### Explanation and Application of Models:

#### Graph Neural Network (GAT) Model - Disease Pathway Prediction:

* **Overview:** The core of the project is a GAT model, implemented in Google Colab, which predicts potential disease pathways based on molecular data. By modeling interactions between genes and identifying patterns in the pathway data, the GAT helps uncover previously unrecognized disease mechanisms.
* **Model Architecture:** The GAT model comprises layers designed for graph-based data, allowing it to capture both node features and edge relationships within the biological pathway. The architecture includes multiple graph convolutional layers that aggregate information across nodes, enabling effective learning of complex interactions in genomic data.
* **Training:** The model was trained on a synthetic dataset with known molecular interactions and pathways, then fine-tuned with real genomic data from GDC. By iterating through forward and backward propagation, the model learns the relationships between genes within disease pathways, allowing for accurate pathway predictions.
* **Application:** The GAT model outputs predictions on the likelihood of specific genes being involved in disease-related pathways. These predictions help researchers and clinicians identify key molecular players in disease mechanisms, potentially guiding further experimental research.

#### Data Processing and API Integration (FastAPI and Python Libraries):

* **Functionality:** FastAPI serves as the bridge between the frontend and the GAT model, managing user data input and prediction outputs. Python libraries facilitate data preprocessing, normalization, and feature extraction, ensuring that raw genomic data from GDC is clean and ready for the GAT.
* **Application:** When a user submits genomic data, it is processed through Python libraries before being sent to the GAT model for analysis. After the GAT generates pathway predictions, FastAPI delivers these results back to the user, providing real-time insights into possible disease mechanisms.

#### Frontend Interface - User Interaction with Pathway Analysis:

* **User Interface:** The frontend, built using React.js and styled with Tailwind CSS and the ShadCN library, provides users with an intuitive way to input data, review pathway predictions, and visualize results. The login/signup feature allows users to create accounts and track analysis history.
* **Application:** After users log in, they can submit patient data, which is then processed and analyzed by the GAT. The frontend then presents the pathway predictions in a user-friendly format, making it easy for users to interpret the model’s findings and understand potential disease mechanisms.

1. **System Architecture:**

* **Frontend (React.js, Tailwind CSS, ShadCN Library):** This layer enables user interaction, taking user input and displaying pathway analysis results in a visually accessible format.
* **Backend (FastAPI and Google Colab GAT):** FastAPI manages data transfer between the frontend and backend, while Google Colab hosts the GAT model, which performs pathway prediction based on genomic data.
* **Data Sources (GDC):** Genomic data from GDC serves as the foundational dataset for training and prediction, while Python libraries facilitate data preprocessing.

# RESULTS

A screenshot of a computer

Description automatically generated

Figure – Home Page

A computer screen shot of a computer

Description automatically generated

Figure – File Input

A screenshot of a computer

Description automatically generated

Figure - Risk Prediction

A black screen with white text

Description automatically generated

Figure - Recommendations Home Page

A screen shot of a computer

Description automatically generated

Figure - Recommendations Page 1A screenshot of a computer screen

Description automatically generated

Figure - Recommendations Page 2

A screenshot of a computer

Description automatically generated

Figure - Recommendations Page 3

A screenshot of a computer

Description automatically generated

Figure - Recommendations Page 4

A screenshot of a computer

Description automatically generated

Figure - Recommendations Page 5

A screenshot of a computer

Description automatically generated

Figure - Recommendations Page 6

# CONCLUSION

In this project, titled "AI-Based Pathway Analysis for Understanding Disease Mechanisms," we developed a web application focused on aiding researchers and clinicians in analyzing disease mechanisms for lung adenocarcinoma. This platform integrates a Graph Neural Network (GAT) for pathway analysis using genomics, transcriptomics, and proteomics data, combined with a Google Maps API for location-based suggestions, providing users with comprehensive insights on disease risk and actionable steps.

We began by identifying the need for a system that could predict disease risk with a high degree of accuracy while offering accessible information on nearby resources and suggested actions. To meet these needs, the application combines cutting-edge technologies across frontend and backend systems, utilizing a React.js interface with ShadCN and Tailwind CSS for a seamless user experience. The backend, powered by FastAPI, facilitates efficient data flow between the frontend and the GAT model, which runs on Google Colab to analyze complex omics data sourced from the Genomic Data Commons (GDC). Python libraries manage data processing and feature selection, ensuring that the input data is optimized for the model.

Through testing, we validated the functionality of each module:

* **The GAT Model** demonstrated its ability to analyze omics data accurately, providing high-quality predictions for lung adenocarcinoma risk based on pathway analysis.
* **The Web Interface** proved user-friendly, enabling users to access insights easily while allowing seamless data input and prediction retrieval.
* **The Integration with Google Maps API** provided valuable location-based recommendations, enabling users to access resources near their vicinity, thereby enhancing the model’s practical applicability.

Overall, this project successfully demonstrates an innovative approach to disease mechanism analysis and risk prediction, combining powerful AI with user-focused design to bridge the gap between data analytics and actionable health insights.

For access to our project files, please visit:

https://github.com/VelvetThunder1/22BIO201-Group-5

# FUTURE WORKS

While the current implementation of "AI-Based Pathway Analysis for Understanding Disease Mechanisms" achieves its primary objectives, several enhancements could broaden its utility and impact:

1. **Expanded Omics Data Integration**: Incorporating additional omics data types, such as metabolomics or epigenomics, could provide a more comprehensive understanding of disease mechanisms and enable better predictions for complex conditions like lung adenocarcinoma.
2. **Improved Model Performance**: Utilizing more advanced neural architectures or fine-tuning the existing Graph Attention Network could improve the accuracy of pathway analysis, especially for rare mutations or complex interactions within biological pathways.
3. **Integration with Electronic Health Records (EHR)**: Adding EHR integration could allow the app to provide even more personalized insights by aligning patient-specific data with GAT-based predictions, which would be particularly valuable for clinicians.
4. **Real-Time Health Data Monitoring**: Incorporating real-time health tracking (e.g., patient vitals or lab results) could enable adaptive risk predictions that adjust based on real-time data, enhancing the application’s precision and relevance.
5. **Collaborative Research Features**: Enabling secure data sharing with healthcare providers or researchers could promote collaboration, allowing them to leverage the tool’s predictions to support clinical decision-making and research studies.
6. **Multi-Language and Localization Support**: Expanding language options and region-specific customization would make the platform accessible to a global user base, supporting international research and clinical applications.
7. **User Feedback and Model Refinement**: Implementing a feedback mechanism would allow users to contribute insights on predictions and recommendations, creating an iterative improvement cycle to refine model performance and accuracy.

These potential enhancements would allow the application to evolve into a more robust and clinically valuable tool, supporting researchers and clinicians in understanding and managing disease mechanisms with greater precision.

**REFERENCES**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.no.** | **Author (s)** | **Title** | **Observation** |
| **1.** | Lv Jin, Xiao-Yu Zuo, Wei-Yang Su, Xiao-Lei Zhao, Man-Qiong Yuan, Li-Zhen Han, Xiang Zhao, Ye-Da Chen, Shao-Qi Rao | Pathway-based Analysis Tools for Complex Diseases: A Review (2014) | Modern pathway-based analysis provides a technique, which allows a comprehensive understanding of the molecular mechanisms underlying complex diseases. |
| **2.** | Sreya Vadapalli, Habiba Abdelhalim, Saman Zeeshan, Zeeshan Ahmed | Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine (2022) | Review and compare scientific objectives, methodologies, datasets, data sources, ethics and gaps of AI/ML approaches used in genomics and precision medicine. |
| **3.** | Elisa Cirillo, Lawrence D. Parnell, Chris T. Evelo | A Review of Pathway-Based Analysis Tools That Visualize Genetic Variants (2017) | Pathway analysis is a powerful method for data analysis in genomics, most often applied to gene expression analysis. |
| **4.** | Clara W. T. Koh, Justin S. G. Ooi, Eugenia Ziying Ong & Kuan Rong Chan | STAGEs: A web-based tool that integrates data visualization and pathway enrichment analysis for gene expression studies (2023) | STAGEs (Static and Temporal Analysis of Gene Expression studies) that provides an interactive visualization of omics analysis outputs. |
| **5.** | Matthew Reyna, David Haan, Marta Paczkowska, Lieven Verbeke, Miguel Vazquez, Abdullah Kahraman, Sergio Pulido-Tamayo, Jonathan Barenboim, Lina Wadi, Priyanka Dhingra, Raunak Shrestha, Gad Getz, Michael Lawrence, Jakob Skou Pedersen, Mark Rubin, David Wheeler, Søren Brunak, Jose Izarzugaza, Ekta Khurana, Kathleen Marchal, Christian Mering, S. Cenk Sahinalp, Alfonso Valencia, Jüri Reimand, Joshua Stuart | Pathway and network analysis of more than 2500 whole cancer genomes (2020) | Expansion and future scope of the project, while a monumental task, is still quite achievable. |