

AI-DRIVEN CLINICAL AND GENOMIC DATA FUSION FOR EARLY CANCER DIAGNOSIS

PROJECT PROPOSAL

Project Title: AI-Driven Clinical and Genomic Data Fusion for Early Cancer Diagnosis

Domain: Research

Project Overview:

This project focuses on developing an AI model that integrates clinical and genomic data from the MMIST-ccRCC dataset to predict early-stage cancer characteristics and survival outcomes. Multi-modal fusion leverages complementary data to improve predictive accuracy compared to single-modality models.

Objectives:

- Use **MMIST-ccRCC** CSV files (clinical+genomic_split.csv) to train and evaluate machine learning models.
- Implement **early fusion** (combining clinical and genomic features) and **late fusion** (combining model predictions).
- Evaluate model performance using **accuracy, AUC, precision, recall, and F1-score**.
- Provide interpretability insights into the contribution of each modality.

Significance:

Early detection of **clear cell renal cell carcinoma (ccRCC)** can significantly improve patient outcomes. By focusing on **clinical and genomic data**, this project investigates the effectiveness of multi-modal AI fusion without relying on imaging, aiming for models that are **cost-effective, reproducible, and scalable** across healthcare systems.

Expected Outcomes:

- Preprocessed, cleaned multi-modal dataset.
- AI models implement early and late fusion.
- Evaluation metrics demonstrating model performance.
- Visualizations showing feature importance and modality contributions to prediction.

Statement of Work (SOW)

Advisor: Rozhin Yasaei

- Provide guidance on dataset usage and modeling approaches.
- Review methodology and project deliverables.
- Approve project milestones and final submissions.

Team Responsibilities:

- Harika Sunkara: Dataset handling, AI model development.
- Swetha Kusampudi: Data preprocessing, feature engineering.
- Farzana Dudekula: Model evaluation, visualization, and reporting.

Date: October 1, 2025

TABLE OF CONTENTS:

S.NO	Contents	PAGE NO
1.	Executive Summary	3
2.	Literature review/Market research	3
3.	Research project deliverables	4
	3a What Analysis Is Being Run?	4
	3b What Accuracy Is Expected?	4
	3c What if the Analysis Doesn't Work?	4
	3d What if the Data Isn't Available?	4
4.	Project Timeline and Gannt Chart	4 & 5
5.	Ethics	6
6.	Approvals	8
7.	Appendix	8
	A. Advisor Engagement	8
	1. Project Team Responsibilities	8
	2. Faculty Advisor Responsibilities	9

	B. Ground Rules	9
--	-----------------	---

1. EXECUTIVE SUMMARY

This project proposes the development of an AI model for **early cancer diagnosis** by leveraging multimodal fusion techniques on the **MMIST-ccRCC dataset**. The dataset includes **clinical and genomic data**.

The primary objective is to evaluate **early fusion** (feature-level integration) and **late fusion** (prediction-level integration) approaches for predicting **clear cell renal cell carcinoma (ccRCC) survival and outcomes**. The expected outcome is to demonstrate that multimodal models provide **higher predictive accuracy** and **better interpretability** than single-modality models, thereby contributing to early detection strategies in healthcare.

2. LITERATURE REVIEW / MARKET RESEARCH

Relevant Studies Using MMIST-ccRCC

1. *MMIST-ccRCC: A Real World Medical Dataset for the Development of Multi-Modal Systems*
 - Authors: Tiago Mota et al.
 - Task: 12-month survival prediction
 - Approach: Early and late fusion, latent representation reconstruction
 - Finding: Multimodal fusion outperforms unimodal approaches.
2. *MIL vs. Aggregation: Evaluating Patient-Level Survival Prediction Strategies Using Graph-Based Learning*
 - Authors: M. Rita Verdelho et al.
 - Task: Patient-level survival prediction
 - Approach: MIL vs. aggregation using GNNs
 - Finding: MIL improved accuracy by selecting representative slides.
3. *A Multimodal Ensemble Approach for Clear Cell Renal Cell Carcinoma Treatment Outcome Prediction*
 - Authors: Meixu Chen et al.
 - Task: Treatment outcome prediction
 - Approach: Multimodal ensemble combining predictions from each modality
 - Finding: Ensemble multimodal approaches surpassed single-modality performance.

Gap Analysis

- Most studies explore imaging fusion; **limited exploration of clinical + genomic-only fusion**.
- Need for interpretable and lightweight models for clinical deployment.

Motivation

This project will build on these studies by benchmarking **early vs. late fusion**, providing interpretability insights, and publishing **reproducible pipelines** for multimodal healthcare AI.

3. RESEARCH PROJECT DELIVERABLES

3a. What Analysis Is Being Run?

- Data preprocessing (cleaning, missing values, normalization).
- Baseline models (Logistic Regression, Random Forest).
- Deep learning models for clinical and genomic modalities.
- Fusion approaches: **early fusion** - Combine clinical + genomic features and **late fusion** - Combine predictions from each modality.
- Evaluation with metrics: Accuracy, Precision, Recall, F1, AUC.

3b. What Accuracy Is Expected?

- Target: **70–85% accuracy, AUC \geq 0.75** depending on modality balance and sample size.

3c. What if the Analysis Doesn't Work?

- Report unimodal baselines (clinical-only and genomic-only).
- Experiment with alternative fusion (stacked ensemble, weighted averaging).

3d. What if the Data Isn't Available?

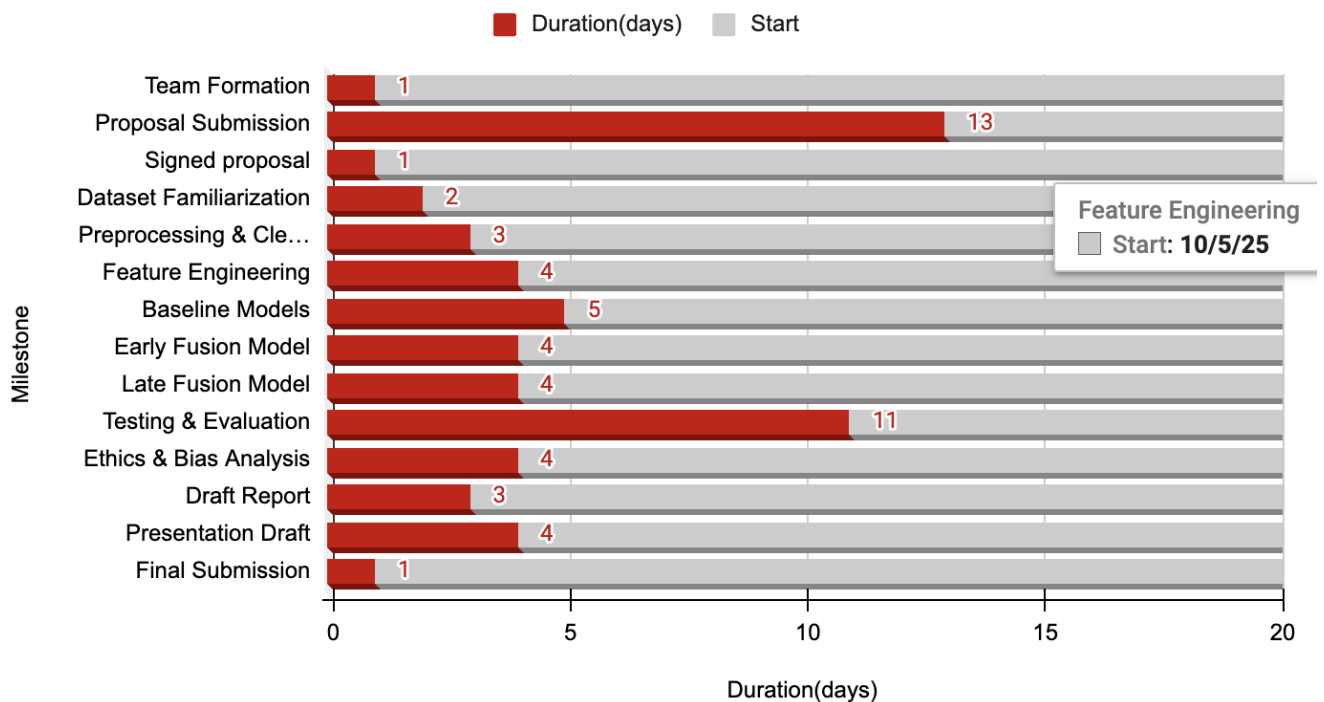
- If parts of the MMIST-ccRCC dataset are restricted:
 - Use **public clinical datasets** (e.g., TCGA for genomics).
 - Simulate multimodal structure using available clinical-genomic pairs.

4. PROJECT TIMELINE & GANTT CHART

MILESTONE	Start	End	Duration(days)
Team Formation	9/15/25	9/15/25	1
Proposal Submission	9/19/25	10/1/25	13

Signed proposal	10/1/25	10/1/25	1
Dataset Familiarization	9/28/25	9/29/25	2
Preprocessing & Cleaning	10/2/25	10/4/25	3
Feature Engineering	10/5/25	10/8/25	4
Baseline Models	10/9/25	10/13/25	5
Early Fusion Model	10/14/25	10/17/25	4
Late Fusion Model	10/18/25	10/21/25	4
Testing & Evaluation	10/22/25	11/1/25	11
Ethics & Bias Analysis	11/2/25	11/5/25	4
Draft Report	11/6/25	11/8/25	3
Presentation Draft	11/9/25	11/12/25	4
Final Submission	11/15/25	11/15/25	1

Project Timeline & Gannt Chart



5. ETHICS

Ethics & Risk Assessment Chart

#	Question	Y/N/M	Explanation / Mitigation
1	Could a user sell drugs or other illegal items on your platform?	N	Not applicable – the platform is for medical AI research only.
2	Could a user engage in sex trafficking?	N	No social or commerce functionality.
3	Could a user sell class notes or cheat on homework using your platform?	N	Not designed for educational marketplaces.
4	Could a stalker use your project to find someone?	N	No location-tracking or personal identifiers are shared.
5	Could your app be used to spy on or track individuals?	M	Clinical/genomic data could indirectly reveal sensitive info. Mitigation: Encrypt all data, role-based access.
6	Could your app/software access the camera or microphone without consent?	N	No camera/mic access required.
7	Could someone be re-traumatized or have mental health impacted?	M	Cancer predictions may cause anxiety. Mitigation: Only deliver results through clinicians with proper support.
8	Could your algorithm promote traumatizing or upsetting content?	N	Not a content platform.
9	Would users be upset if their data was shared with others?	Y	Yes, medical/genomic data is sensitive. Mitigation: HIPAA/GDPR compliance, anonymization, explicit consent.
10	Could a data leak lead to identity theft?	Y	Genomic data is personally identifiable. Mitigation: Strong encryption, anonymization, minimal storage.
11	If hacked, could users lose their job, spouse, or family?	M	Insurance/employment discrimination possible. Mitigation: Secure storage, limited access, oversight.
12	Should there be an age limitation?	Y	Only adults or minors with guardian consent should be included.


13	Could someone use your product to commit elder abuse?	N	No direct link.
14	If data was breached, could it be used for blackmail?	Y	Health/genomic info could be exploited. Mitigation: Strong security and ethical oversight.
15	Does your project imply bias against groups (race, gender, religion)?	M	Possible if dataset lacks diversity. Mitigation: Check datasets for representation, fairness audits.
16	Could your project be used to commit hate crimes?	N	Not applicable.
17	Does your algorithm focus on something unethical?	N	Purpose is medical diagnosis for patient benefit.
18	Does your app contain stereotypes (race/gender)?	M	Not inherent, but must test dataset for bias.
19	Could users scam others through your platform?	N	No user marketplace.
20	Is the algorithm biased toward one group?	M	Mitigation: Evaluate model fairness and ensure balanced representation.
21	Are users aware of how data will be used?	Y	Informed consent required for all data usage.
22	Could results be misinterpreted (e.g., by extremist groups)?	M	Mitigation: Results communicated only by clinicians; provide clear documentation.
23	Could use/purchase of your data help dangerous groups or regimes?	M	Possible misuse by insurers/governments. Mitigation: Limit data sharing agreements; ethics board approval.
24	Could your software cause injury in VR?	N	No VR features.
25	Are participants aware that their data will be collected?	Y	Informed consent obtained before participation.
26	Does your app contain addictive design elements?	N	Not relevant – no gamification.
27	Does your survey include compulsion or large incentives?	N	Participation is voluntary.
28	Could research outcomes harm individuals or entities?	M	Misuse of predictions may cause harm. Mitigation: Clinician oversight, ethics review

			board approval, secure handling of sensitive predictions.
--	--	--	---

6. APPROVALS

The signatures of the people below indicate an understanding of the purpose and content of this document. By signing, you approve the proposed project outlined in this Statement of Work (SOW), the division of work, the Ground Rules, and agree that next steps may be taken to create a Product Specification and proceed with the project.

Approver Table

Approver Name	Title	Signature	Date
Harika Sunkara	Team Member / AI Model Development	<i>Harika Sunkara</i>	10/01/2025
Swetha Kusampudi	Team Member / Data Preprocessing & Feature Engineering	<i>Swetha Kusampudi</i>	10/01/2025
Farzana Dudekula	Team Member / Model Evaluation & Visualization	<i>Farzana Dudekula</i>	10/01/2025
Rozhin Yasaei	Faculty Advisor / Mentor		10/08/2025
Nitika Sharma	Instructor		

7. APPENDIX

A. Advisor Engagement

1) Project Team Responsibilities

- The Project Manager will set up and facilitate **weekly calls/meetings with Rozhin Yasaei (Faculty Advisor)**.
- Team members will provide **weekly status updates**, including upcoming deliverables, critical issues, and adjustments to the Project Plan.

- Documents will be submitted to Rozhin Yasaei **at least 3 days before the due date** for review and signature.
- Design files and AI model outputs will be shared in a mutually agreed format.
- Modification requests from the Faculty Advisor will be reviewed and addressed within **1 week**.

2) Faculty Advisor Responsibilities

- Rozhin Yasaei will provide expertise to help the team advance their skills in **AI, data preprocessing, and multimodal fusion techniques**.
- Participate in **weekly or bi-weekly calls** to review project status, upcoming deliverables, priorities, issues, and progress.
- Review and approve/reject documents with adequate time for the team to meet deadlines.
- Provide feedback on design decisions, model evaluation, and visualizations.
- Resolve requested project plan modifications within **1 week**.
- Attend **iShowcase** and grade the finalized project using a skill-based rubric.

B. Ground Rules

As a team, we agree to the following principles to complete this project successfully:

1. **Stay focused on objectives and goals** – Every meeting starts with clear objectives; off-topic issues will be noted in a “sidebar” for later discussion.
2. **Listen actively** – Consider all viewpoints before contributing your own.
3. **Ensure all voices are heard** – No one dominates discussions; quieter team members are encouraged to share input.
4. **Respect differences of opinion** – Identify agreements before disagreements; discuss differing ideas constructively.
5. **Look for strengths in new ideas** – Assess the value of each suggestion when planning next steps.
6. **Focus on future solutions** – Learn from past experience, but direct discussions toward actionable solutions.
7. **Agree on action items and next steps** – Each meeting ends with assigned tasks and deadlines.
8. **Maintain accountability** – Each member is responsible for their assignments; failure to contribute affects the team as a whole.
9. **Communicate proactively** – Update the team on progress, issues, or blockers in a timely manner.