AI-Driven Multimodal Data Fusion for Early Cancer

Literature Review

1. Introduction

Cancer is one of the leading causes of mortality worldwide, and early diagnosis significantly improves clinical outcomes. Advances in artificial intelligence (AI) have enabled precise prediction of tumor behaviour through large-scale clinical and molecular data analysis. However, most traditional diagnostic approaches rely on a single data type, which can overlook important relationships between patient characteristics and tumor genomic drivers.

This project—**AI-Driven Multimodal Data Fusion for Early Cancer Diagnosis**—focuses on integrating **clinical variables** and **genomic mutation profiles** to predict **short-term survival (within 12 months)** in patients diagnosed with clear-cell renal cell carcinoma (ccRCC). The dataset is derived from the **TCGA-KIRC** (The Cancer Genome Atlas – Kidney Renal Clear Cell Carcinoma) cohort.

By building a multimodal, imaging-free deep learning framework, this study aims to uncover cross-modal correlations between patient demographics, tumor staging, and genomic alterations to support early-diagnosis tools for precision oncology.

2. Literature Review

2.1 Artificial Intelligence in Cancer Diagnosis

AI techniques have significantly advanced cancer diagnosis by enabling the analysis of complex, heterogeneous biomedical data. Traditional machine learning models such as Logistic Regression, Random Forest, and XGBoost are widely used for classification and survival prediction tasks. More recent deep learning approaches allow automated representation learning and can model non-linear dependencies across clinical and molecular data. These AI-driven approaches consistently outperform traditional statistical methods in prediction accuracy and scalability.

2.2 Multimodal Learning in Oncology

Multimodal learning integrates multiple complementary data sources to form a comprehensive understanding of tumor biology. In oncology, combining clinical and genomic information has been shown to enhance prognostic accuracy by capturing correlations between patient characteristics and molecular alterations. Studies leveraging TCGA datasets demonstrate that incorporating mutation profiles into clinical

prediction pipelines improves risk stratification and survival prediction in cancers such as ccRCC.

2.3 Data Fusion Techniques

Data fusion aims to combine heterogeneous modalities into a unified predictive framework. Common fusion strategies include:

- **Early fusion**: concatenating multiple feature sets before training.
- **Late fusion**: combining outputs of modality-specific models.
- **Intermediate fusion**: merging learned embeddings within a deep neural architecture.

Selecting an effective fusion method depends on data quality, dimensionality, and modality balance. Deep learning–based intermediate fusion is increasingly used due to its flexibility and ability to capture cross-modal relationships.

2.4 Clinical and Genomic Predictors of ccRCC Survival

Clinical variables such as age, tumor stage, and grade are well-established predictors of ccRCC survival. Genomic alterations—especially mutations in **VHL, PBRM1, and TTN**—have been associated with tumor aggressiveness, metabolic changes, and prognosis. Integrating these two data domains may improve survival prediction by linking molecular pathways to clinical progression.

2.5 Gaps in Existing Research

While multimodal AI models have demonstrated strong performance, several gaps remain:

- Most studies integrate only **two** modalities (e.g., clinical + genomics).
- Limited standardization in multimodal pipeline design.
- Data heterogeneity and missing genomic profiles limit reproducibility.
- Interpretability of deep fusion models remains challenging.

This project addresses these gaps by developing an imaging-free, dual-modality multimodal fusion framework using standardized TCGA clinical and genomic data.

2.6 Summary of Findings

The literature consistently shows that combining clinical and genomic data enhances cancer diagnosis and survival prediction. While AI-based multimodal methods outperform single-modality approaches, challenges persist regarding model generalizability and interpretability. This study contributes an imaging-free, multimodal framework for short-term survival prediction in ccRCC.

3. Methodology

3.1 Dataset Description

The dataset consists of **TCGA-KIRC** patients with two data modalities:

- **Clinical features:** age at diagnosis, gender, tumor grade, AJCC stage, cancer history.
- **Genomic features:** binary mutation status of key ccRCC-associated genes (VHL, PBRM1, TTN).

Each patient is identified using a unique TCGA barcode (e.g., TCGA-BP-4992). Only patients with complete clinical and genomic records were included.

3.2 Preprocessing

- Imputation of missing numerical values using **mean** and categorical values using **mode**.
- One-hot encoding for categorical variables (gender, race, stage).
- Standardization of continuous features using **StandardScaler**.
- Genomic mutation statuses converted to binary format (0 = wild-type, 1 = mutated).

3.3 Feature Engineering

- Random Forest and XGBoost models used to evaluate feature importance.
- Top clinical predictors (e.g., stage, grade, age) retained.
- Genomic predictors included based on biological relevance and statistical significance.

- Final feature vector constructed by concatenating clinical + genomic attributes.

3.4 Model Design

Four supervised learning models were implemented:

Classical ML Models

- **Logistic Regression** – interpretable baseline.
- **Random Forest Classifier** – capable of modeling nonlinear interactions.
- **XGBoost** – boosted ensemble with regularization.

Deep Learning Fusion Model (FFNN)

Layer	Description
Input	Clinical + genomic features
Hidden 1	128 neurons, ReLU, Dropout (0.3)
Hidden 2	64 neurons, ReLU
Output	Sigmoid output for binary classification

Training Settings:

- Optimizer: Adam (lr = 0.001)
- Loss: Binary Cross-Entropy
- Epochs: 50 (Early Stopping enabled)

3.5 Fusion Strategies

Two fusion approaches were evaluated:

- **Early fusion:** direct concatenation of clinical + genomic features.
- **Late fusion:** separate clinical and genomic encoders merged before final output layer.

Both strategies were validated using **5-fold cross-validation**.

4. Results and Discussion

4.1 Model Performance

Model	Accuracy	ROC-AUC	Recall (Deceased)	Recall (Alive)
Logistic Regression	0.65	0.66	0.60	0.66
Random Forest	0.92	0.86	0.47	0.98
XGBoost	0.85	0.92	0.87	0.85
FFNN (Fusion Model)	0.86	0.8606	0.73	0.88

The deep fusion model achieved strong performance by leveraging cross-modal interactions between clinical and genomic features. XGBoost achieved the highest ROC-AUC, while Random Forest exhibited excellent recall for the "Alive" class.

4.2 Interpretability

- Feature importance analyses revealed **tumor stage**, **VHL mutation**, and **age** as key predictors of 12-month survival.
- These results align with known biological associations in ccRCC progression.

4.3 Challenges

- Dataset size limited due to strict requirement for complete multimodal records.
- Potential bias from class imbalance.
- High variance in clinical variables across patients introduces modelling challenges.

5. Conclusion and Future Work

This study demonstrates the effectiveness of multimodal data fusion using clinical and genomic features for short-term survival prediction in clear-cell renal cell carcinoma (ccRCC). By leveraging TCGA-KIRC clinical variables and key genomic mutations, the proposed fusion framework achieved strong predictive performance and provided interpretable insights into factors influencing 12-month survival. These results establish a solid baseline for multimodal survival-prediction research in ccRCC and highlight the

critical role of integrating heterogeneous patient data for improved early-risk stratification.

Future Work

Although the current study focuses on clinical and genomic modalities, future work will expand this framework to incorporate **medical imaging from the full MMIST-ccRCC dataset**, including:

- **CT scans**
- **MRI sequences**
- **Whole-slide digital pathology images**

Integrating these imaging modalities with clinical and genomic features will enable a comprehensive radiogenomic-clinical modeling strategy. Advanced **multimodal deep learning architectures**—such as transformer-based fusion networks, cross-attention models and encoder–decoder pipelines—will be developed to learn joint representations across all data types.

This broader multimodal integration is expected to:

- Capture **richer and more complementary tumor representations**
- Improve **recall for high-risk (deceased within 12 months) patients**
- Enhance **overall ROC-AUC performance** through deeper cross-modal feature learning
- Increase the biological interpretability of model predictions via explainable AI techniques

Ultimately, this research lays the groundwork for developing an **AI-driven clinical decision-support tool** capable of assisting clinicians in personalized treatment planning and early survival-risk assessment for ccRCC patients. The incorporation of deep learning and multimodal data represents a promising next step toward improved accuracy, generalizability and clinical utility in precision oncology.

References

- Clark, K., Vendt, B., Smith, K., et al. (2013). The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. Journal of Digital Imaging, 26(6), 1045–1057.

- van Griethuysen, J. J. M., Fedorov, A., Parmar, C., et al. (2017). Computational radiomics system to decode the radiographic phenotype. Scientific Reports, 7(1), 4470.
- Uhlig, J., et al. (2024). Radiomics for renal tumor subtyping and prognosis: A systematic review. European Radiology, 34(3), 1212–1228.
- Wang, H., Zhang, L., et al. (2023). Multimodal deep learning for clear-cell renal cell carcinoma grading and survival prediction. Insights into Imaging, 14(2), 54–69.
- PyRadiomics Documentation. (2017). Open-source radiomics feature extraction platform.
- TCGA Research Network. (2023). The Cancer Genome Atlas (KIRC cohort): Clinical and genomic data portal.