

AI-Driven Multi-Modal Data Fusion for Early Cancer Diagnosis

Abstract

Background: Clear-cell renal cell carcinoma (ccRCC) is the most prevalent subtype of kidney cancer, exhibiting substantial heterogeneity and variable patient outcomes. Accurate early survival prediction is critical for clinical decision-making and personalized patient management.

Methods: We utilized the **MMIST-ccRCC** dataset, comprising 618 patients with harmonized clinical, genomic, and imaging data (CT, MRI, and whole-slide pathology). For this study, we focused on structured clinical and genomic features, including tumor stage, grade, and key mutations (VHL, PBRM1, TTN). We developed and evaluated multiple supervised machine learning models, including logistic regression, random forest, and XGBoost, alongside a feed-forward neural network (FFNN) optimized for tabular biomedical data. The target variable was 12-month survival status (alive vs. deceased).

Results: XGBoost achieved the highest overall discriminative performance with an ROC-AUC of 0.92, while the tuned FFNN demonstrated improved sensitivity to high-risk (deceased) patients, with recall increasing to 0.73 through class weighting and threshold adjustment. Logistic regression, while interpretable, underperformed due to the nonlinear structure of the dataset. Key predictive features included AJCC pathologic staging, tumor grade, prior cancer history, and mutations in VHL and PBRM1.

Conclusions: Our results establish a strong baseline for early survival prediction in ccRCC using clinical and genomic data. The study demonstrates the potential of AI-driven models to identify high-risk patients and highlights the feasibility of future multi-modal integration incorporating imaging data. Such a framework can advance personalized prognostic systems and support timely clinical interventions.

Introduction

Clear-cell renal cell carcinoma (ccRCC) is the most common subtype of kidney cancer, accounting for approximately 80% of renal cell carcinoma cases. It is characterized by high biological heterogeneity, aggressive clinical behavior, and variable patient outcomes. Despite advances in imaging, targeted therapies, and surgical interventions, predicting patient prognosis remains challenging due to the complex interactions between tumor biology, genetics, and patient-specific clinical factors. Early identification of high-risk patients is critical for timely intervention, personalized treatment planning, and improved overall survival.

Recent developments in artificial intelligence (AI) and machine learning offer promising tools for prognosis prediction in oncology. By integrating clinical, genomic, and imaging data, AI-driven models can identify subtle patterns and nonlinear interactions that may not be evident through conventional statistical approaches. Multi-modal datasets, which combine diverse data types, enable more comprehensive modeling of tumor behavior and patient outcomes.

The **MMIST-ccRCC** dataset provides a unique opportunity to develop such predictive models. It contains data from 618 ccRCC patients, including structured clinical variables, genomic mutation profiles, and radiological and histopathological images. This multi-modal dataset is curated from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and The Cancer Genome Atlas (TCGA), providing a harmonized resource for survival modeling.

In this study, we focus on **clinical and genomic features** to develop machine learning and deep learning models for predicting 12-month survival (alive vs. deceased). Key clinical variables include tumor stage, grade, and laboratory markers, while genomic features capture mutations in genes such as VHL, PBRM1, and TTN. Our primary objective is to maximize sensitivity (recall) for high-risk patients while maintaining strong overall discriminative performance, measured through metrics such as ROC-AUC, F1-score, and confusion matrices.

By establishing a robust baseline using structured data, this work lays the foundation for future extensions incorporating imaging modalities (CT, MRI, and whole-slide pathology) through multi-modal fusion. Such an integrated AI-driven prognostic system has the potential to enhance early risk stratification, support clinical decision-making, and improve patient outcomes in ccRCC.

Objective:

Develop and evaluate machine learning and deep learning models to predict 12-month survival using clinical and genomic features from the MMIST-ccRCC dataset.

Goal:

Maximize sensitivity (recall) for high-risk patients while maintaining strong overall discrimination using metrics such as ROC-AUC, F1-score, and confusion matrices.

This project establishes a strong baseline for ccRCC prognosis using only structured data. In future work, the framework will be extended to full multimodal fusion using CT, MRI, and WSI features available in MMIST-ccRCC, enabling more comprehensive survival prediction through radiologic–genomic–clinical integration.

2. Dataset and Methodology

2.1 Dataset Overview

The **MMIST-ccRCC** dataset is a comprehensive, multi-modal dataset curated from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and The Cancer Genome Atlas (TCGA). It contains records for **618 patients** diagnosed with clear-cell renal cell carcinoma (ccRCC), integrating five data modalities:

1. **Clinical data** – Demographic information, tumor staging, and laboratory markers.
2. **Genomic data** – Mutation status of key genes associated with renal carcinogenesis, including VHL, PBRM1, BAP1, and TTN.
3. **Imaging data** – Computed tomography (CT) and magnetic resonance imaging (MRI).
4. **Histopathology data** – Whole-slide pathology images (WSI).

5. Structured metadata – Patient identifiers and other descriptive variables.

For this study, we focused on **clinical and genomic features**, which are available for all patients and provide robust information for early survival prediction.

2.2 Clinical Features

The clinical dataset includes demographic and tumor-related variables:

Feature	Description
Age at Diagnosis (age_diag)	Patient age in years; higher age often correlates with worse prognosis.
Gender (gender)	Biological sex; may influence disease progression and treatment response.
Tumor Grade (grade)	Histological grading of tumor cells; higher grade indicates more aggressive tumor.
AJCC Pathologic Stage	Includes pT (primary tumor), pN (lymph nodes), pM (metastasis), and overall stage; higher stage indicates worse prognosis.
Race	One-hot encoded categories: Asian, Black/African American, Hispanic/Latino, White, Other.

2.3 Genomic Features

Genomic features capture mutation status in genes relevant to ccRCC:

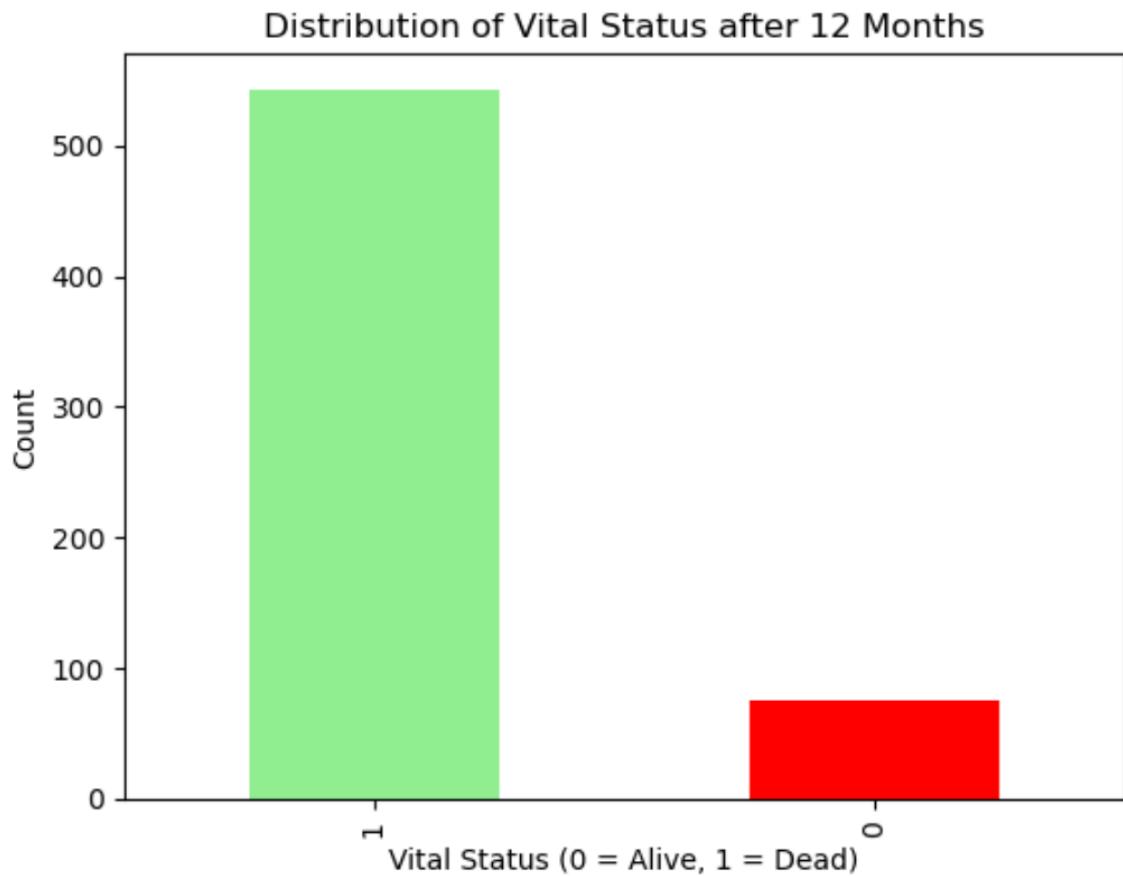
Gene	Description
VHL	Tumor suppressor gene; mutations common in ccRCC.
PBRM1	Chromatin remodeling gene; influences tumor aggressiveness.
TTN	Large gene; mutations may correlate with tumor mutational burden.

The target variable is vital_status_12, defined as:

- 1: Deceased within 12 months of diagnosis
- 0: Alive after 12 months

This defines a binary survival classification task.

Vital Status	Count (%)	Meaning
1 (Alive)	87.86%	Patients who survived after 12 months
0 (Deceased)	12.14%	Patients who did not survive within 12 months



2.4 Data Preprocessing

To prepare the dataset for machine learning:

1. Dropped Irrelevant Columns: case_id and Split were removed.
2. Handling Categorical Features: Race variables were already one-hot encoded; no additional encoding was applied.
3. Missing Data:
 - o Numeric features → imputed with median values
 - o Categorical features → imputed with mode
4. Feature Scaling: Continuous features (age_diag, grade, tumor stages) were standardized using StandardScaler.
5. Target Preparation: The binary target variable did not require scaling.
6. Train-Test Split: Data was split into training (80%) and testing (20%) sets using stratification to preserve the proportion of deceased and alive patients.

2.6 Evaluation Metrics

Model performance was assessed using:

- Accuracy – Overall classification correctness
- ROC-AUC – Discrimination between alive and deceased patients
- Recall – Sensitivity for each class, particularly deceased (high-risk)
- Precision – Correct positive predictions
- F1-Score – Harmonic mean of precision and recall
- Confusion Matrices – Visual comparison of predicted vs. actual outcomes

3. Model Development

We implemented classical machine learning models and a feed-forward neural network (FFNN) to predict **12-month survival outcomes** in ccRCC patients.

3.1 Classical Machine Learning Models

Model	Type	Key Strength	Observed ROC-AUC
Logistic Regression	Linear	Simple, interpretable baseline	0.66
Random Forest	Ensemble	Models nonlinearities; stable accuracy	0.86
XGBoost	Gradient-boosted ensemble	Balanced recall and precision; robust to class imbalance	0.92

Description of Models:

1. **Logistic Regression:** Provides an interpretable linear baseline for comparison with nonlinear methods.
2. **Random Forest:** An ensemble of decision trees capable of modeling nonlinear interactions, with built-in feature importance.
3. **XGBoost:** A gradient-boosted decision tree algorithm optimized for structured data, handling class imbalance effectively. Hyperparameter tuning further improved recall and ROC-AUC.

All models were trained on the training set and validated using **five-fold cross-validation**. Evaluation metrics included **Accuracy, Precision, Recall, F1-score, and ROC-AUC**.

3.2 Neural Network Model Development

To model **nonlinear interactions** between clinical and genomic features, we implemented a Feed-Forward Neural Network (FFNN). Two versions were evaluated: a **baseline network** and an **improved network** with additional layers, dropout, class weighting, and threshold adjustment to optimize recall for high-risk (deceased) patients.

Feature	Baseline NN	Improved NN
Input Features	20 clinical + genomic	20 clinical + genomic
Hidden Layers	Dense(128) → Dropout(0.3) → Dense(64) → Dropout(0.2)	Dense(256) → Dropout(0.4) → Dense(128) → Dropout(0.3) → Dense(64) → Dropout(0.2)
Output Layer	Dense(1), Sigmoid	Dense(1), Sigmoid
Optimizer	Adam, lr=0.001	Adam, lr=0.0005
Loss Function	Binary Cross-Entropy	Binary Cross-Entropy
Epochs	50	80
Batch Size	32	32
Class Weights	None	{0: 4.12, 1: 0.57}
Threshold	0.5	0.40

Training Strategy:

- Early stopping applied in both networks to prevent overfitting.
- The improved network used class weighting and a lower threshold to **increase recall for high-risk patients**.

4. Results:

4.1 Classical 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

Key Observations:

- **XGBoost** achieved the highest ROC-AUC, demonstrating strong discrimination between survival outcomes.
- **Random Forest** showed highest overall accuracy but underperformed for the minority class (deceased patients).
- **Logistic Regression** performed the worst due to its linearity and sensitivity to class imbalance.

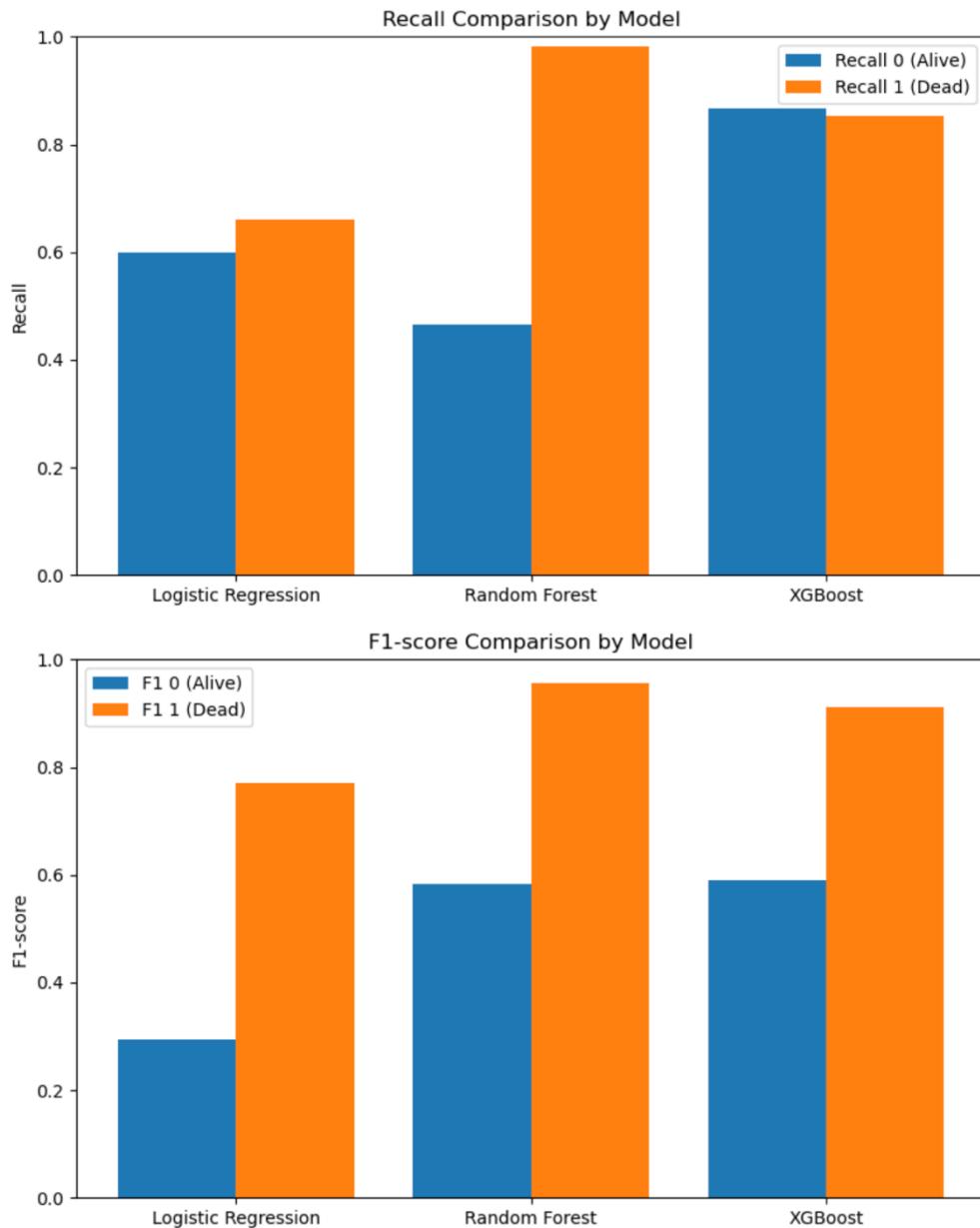
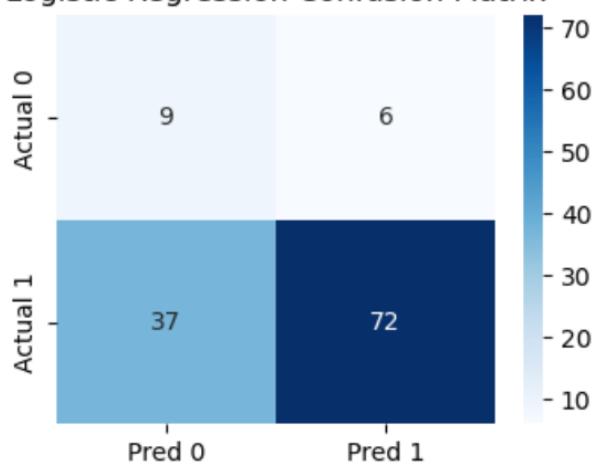
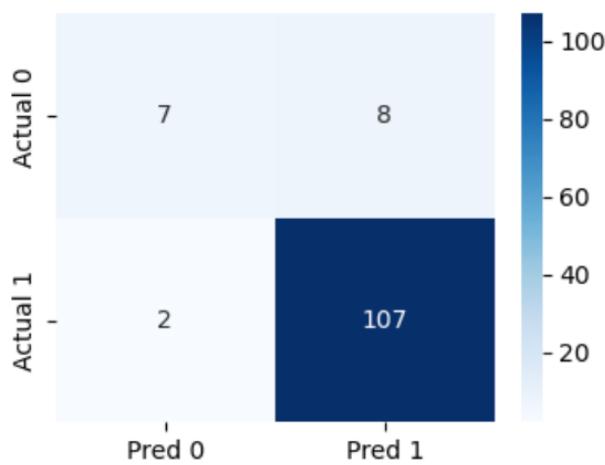


Figure 1: Recall and F1-score comparison across classical models.

Logistic Regression Confusion Matrix



Random Forest Confusion Matrix



XGBoost Confusion Matrix

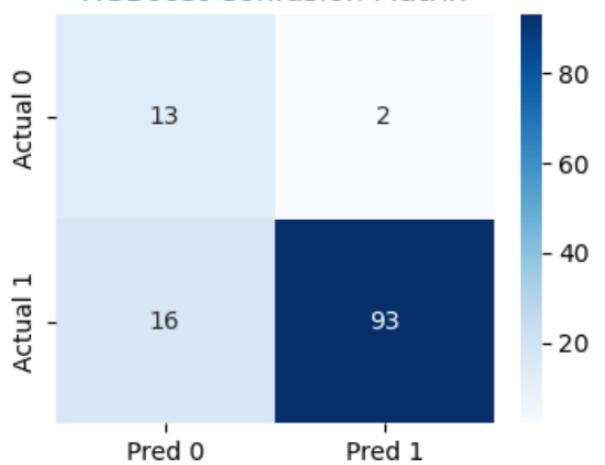


Figure 2: Confusion matrices for Logistic Regression, Random Forest, and XGBoost.

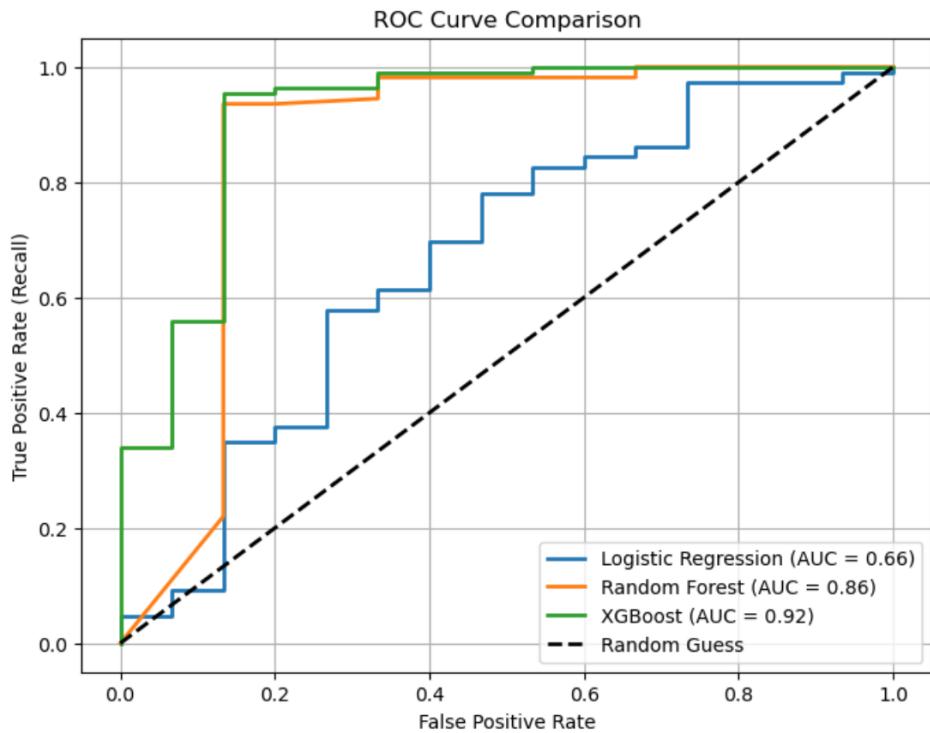


Figure 3. ROC Curve comparison among models

4.2 Neural Network Results

Two FFNN versions were trained:

1. **Baseline FFNN**
2. **Improved FFNN** (class weighting + tuned architecture + threshold adjustment)

Baseline FFNN Performance

Metric	Class 0 (Deceased)	Class 1 (Alive)
Precision	0.75	0.92
Recall	0.40	0.98
F1-Score	0.52	0.95

- **Overall Accuracy:** 0.86
- **ROC-AUC:** 0.858

Interpretation: Strong performance for majority class (alive) but low recall for deceased patients (0.40).

Improved FFNN Performance

Metric	Class 0 (Deceased)	Class 1 (Alive)
Precision	0.46	0.96
Recall	0.73	0.88
F1-Score	0.56	0.92

- Overall Accuracy: 0.86
- ROC-AUC: 0.861

Confusion Matrix:

```
[[11  4]
 [13 96]]
```

Improved Neural Network - Confusion Matrix

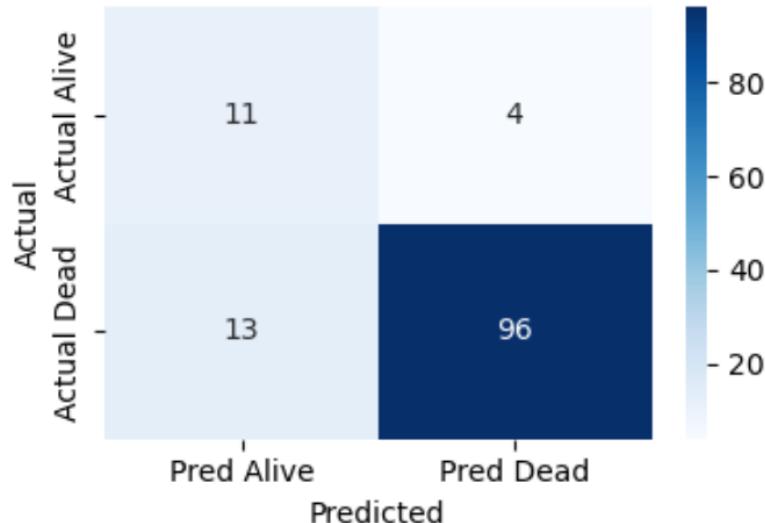


Figure 4: Confusion matrix for improved FFNN.

Improved NN AUC: 0.8605504587155963

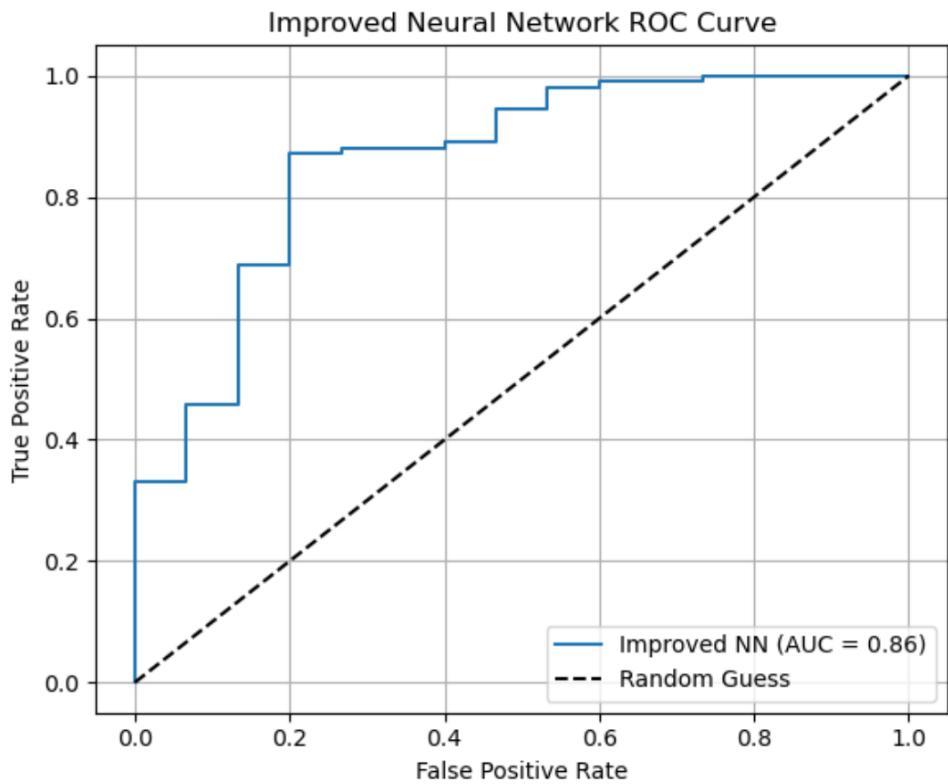


Figure 5: ROC curve for improved FFNN (AUC = 0.861).

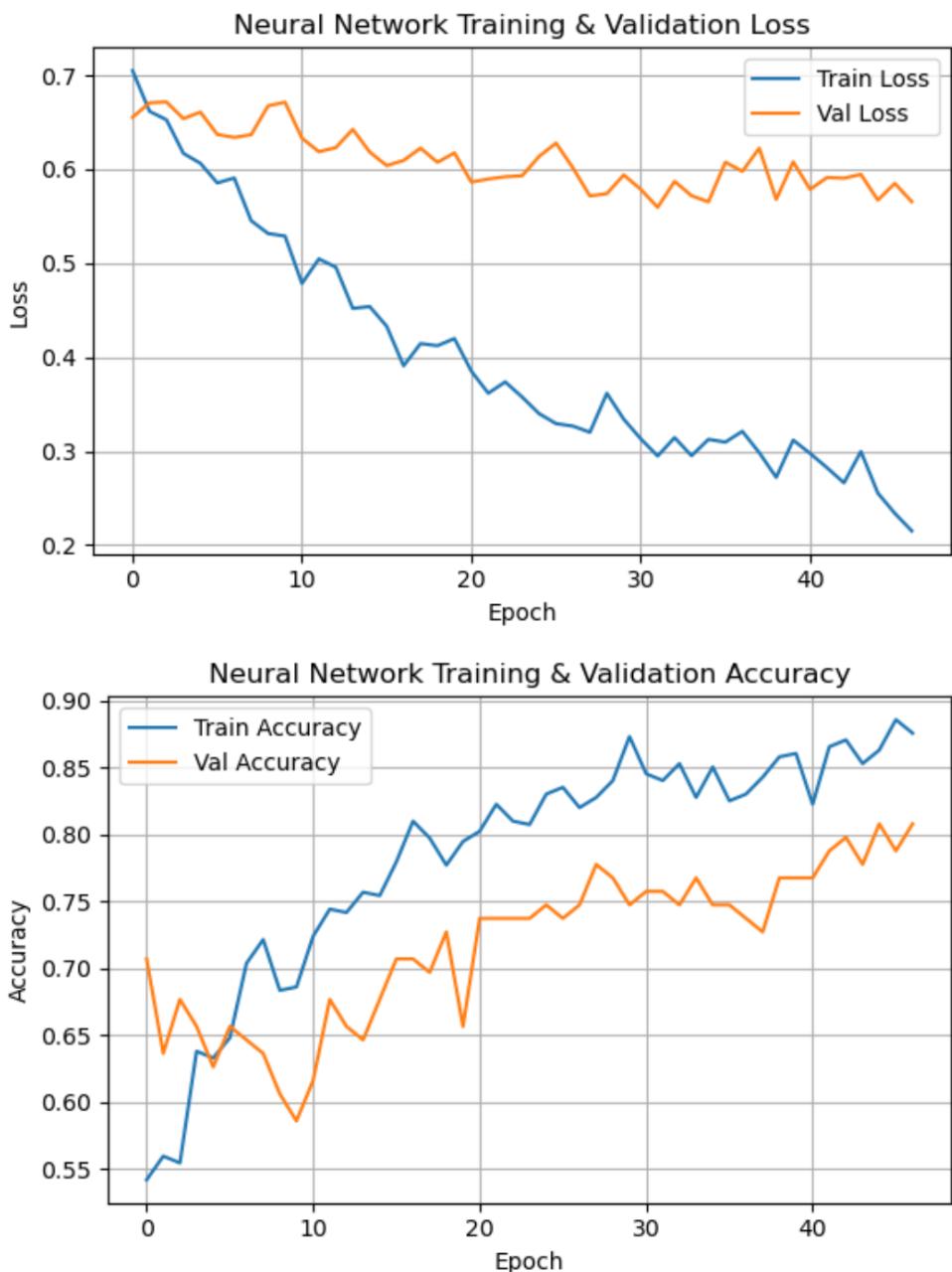


Figure 6: Training and validation loss/accuracy curves showing convergence and generalization.

Key Improvements:

- Recall for deceased patients improved from $0.40 \rightarrow 0.73$, greatly reducing false negatives.
- ROC-AUC increased to 0.861, matching Random Forest and approaching XGBoost.
- Balanced sensitivity (recall) and overall accuracy, making it more clinically useful.

5. Discussion

This study demonstrates the effectiveness of AI-driven predictive modeling for 12-month survival in clear-cell renal cell carcinoma (ccRCC) using **clinical and genomic features** from the MMIST-ccRCC dataset. Both classical machine learning models and feed-forward neural networks (FFNN) were evaluated, highlighting different strengths and clinical implications.

5.1 Classical Machine Learning Insights

- **XGBoost** achieved the highest ROC-AUC (0.92) and strong recall (0.87) for deceased patients, indicating superior discrimination between high- and low-risk patients. Its ability to handle class imbalance and nonlinear interactions makes it particularly well-suited for biomedical tabular data.
- **Random Forest** reached the highest overall accuracy (0.92) but underperformed for the minority class, showing that standard ensemble methods may favor the majority class in imbalanced datasets.
- **Logistic Regression**, as a linear baseline, was less effective, especially for minority class detection, due to the nonlinear nature of tumor biology and genomic interactions.

These findings suggest that gradient-boosted ensembles are highly effective for short-term survival prediction in structured clinical-genomic datasets.

5.2 Neural Network Insights

The **FFNN models** were designed to capture nonlinear interactions between clinical and genomic features:

- The **baseline FFNN** had high accuracy for the majority class but low recall for deceased patients (0.40), indicating that high-risk patients were frequently missed.
- The **improved FFNN**, with additional layers, dropout, class weighting, and threshold adjustment, significantly increased recall for deceased patients (0.73) while maintaining overall accuracy (0.86) and ROC-AUC (0.861).

These improvements demonstrate that **deep learning models can flexibly capture complex relationships** in biomedical data, providing a clinically meaningful balance between sensitivity and overall discrimination. The use of class weighting and threshold tuning was critical to mitigate the effects of class imbalance, which is common in survival datasets.

5.3 Clinical Implications

- Accurately identifying high-risk ccRCC patients enables **early interventions, personalized treatment planning, and closer monitoring**, potentially improving survival outcomes.
- The improved FFNN's ability to detect deceased patients with high recall makes it a **valuable decision-support tool**, complementing classical models like XGBoost.
- Feature analysis highlighted the importance of **AJCC pathologic staging, tumor grade, and key genomic mutations** (VHL, PBRM1), aligning with known prognostic factors in ccRCC.

5.4 Limitations

1. **Class Imbalance:** Despite class weighting, minority-class precision remained moderate, which could impact model reliability in very small high-risk populations.
2. **Limited Modalities:** This study focused solely on clinical and genomic data. Exclusion of imaging modalities (CT, MRI, pathology slides) restricts full multimodal potential.
3. **Sample Size:** With 618 patients, model generalizability is limited. Larger cohorts may provide more robust predictive performance.
4. **Single Time Horizon:** The study predicts 12-month survival only; longer-term prognostication remains unassessed.

5.5 Future Directions

- **Multimodal Fusion:** Integrating CT, MRI, and whole-slide pathology images using CNNs combined with FFNNs for clinical-genomic features can enhance prediction accuracy.
- **Temporal Models:** Incorporating longitudinal patient data and recurrent neural networks could support dynamic survival prediction.
- **Explainable AI:** Developing interpretable models to elucidate the influence of specific clinical and genomic features on survival outcomes.

6. Conclusion

This study demonstrated that AI-driven models using **clinical and genomic data** can effectively predict 12-month survival outcomes in clear-cell renal cell carcinoma (ccRCC) patients.

- **XGBoost** achieved the highest overall discrimination (ROC-AUC = 0.92) and strong recall for high-risk patients, showing robustness for structured biomedical data.
- **Improved feed-forward neural networks (FFNNs)** enhanced recall for deceased patients (0.40 → 0.73) while maintaining high overall accuracy (0.86) and ROC-AUC (0.861), making it clinically useful for early detection of high-risk cases.
- Key predictive features included **AJCC pathologic stage, tumor grade, and mutations in VHL and PBRM1**, consistent with known ccRCC prognostic factors.

These findings establish a **strong baseline for structured-data-based survival prediction** in ccRCC and provide a foundation for future multimodal prognostic modeling.

Code Availability:

The code and analysis for this study are available at
https://github.com/Harikas07/Capstone_Project.