Multi-Modal Deep Learning Framework for Personalized Treatment Decision Support in Early-Stage Non-Small Cell Lung Cancer

Bowen Lou * and Shuxin Mo

Posted Date: 25 March 2026

doi: 10.20944/preprints202603.1988.v1

Keywords: NSCLC; deep learning; personalized medicine; decision support; multi-modal

 doi:10.20944/preprints202603.1988.v1

Article

Multi-Modal Deep Learning Framework for Personalized Treatment Decision Support in Early-Stage Non-Small Cell Lung Cancer

Bowen Lou * and Shuxin Mo

Kunming University of Science and Technology, China
* Correspondence: 202073158421@stu.kust.edu.cn

Abstract

Personalized treatment for early-stage non-small cell lung cancer (NSCLC), particularly in choosing between SBRT and surgery, is challenging due to complex, heterogeneous patient data. We introduce MM-Care, a novel deep learning framework for objective, interpretable, and personalized treatment decision support. MM-Care integrates patient-specific CT imaging, clinical indicators, and genomic data through a sophisticated multi-branch neural network. Its core innovations include multi-modal feature extraction, an adaptive Transformer-based fusion network for deep inter-modal interaction, and a dual-task prediction head for overall survival and local control across both interventions. An explainable decision report module, utilizing feature importance methods, enhances clinical trust. Evaluated on public and proprietary cohorts comprising thousands of patients, MM-Care consistently outperforms traditional models and deep learning baselines. Our experiments demonstrate superior prognostic performance for survival and local control. Ablation studies validate critical architectural contributions. Human evaluation with oncologists confirms high trust, utility, and interpretability, showing significant time savings and strong agreement with expert consensus. MM-Care also achieves high accuracy in aligning with retrospectively identified optimal treatment choices. These results highlight MM-Care's robust capability to provide precise, patient-specific prognostic predictions and optimal treatment recommendations, poised to significantly enhance personalized medicine in early-stage NSCLC.

Keywords: NSCLC; deep learning; personalized medicine; decision support; multi-modal

1. Introduction

Non-small cell lung cancer (NSCLC) remains one of the leading causes of cancer-related morbidity and mortality worldwide. For patients with early-stage NSCLC, stereotactic body radiation therapy (SBRT) and surgical resection are two primary curative treatment modalities. However, determining the optimal treatment strategy for individual patients to maximize survival benefits while minimizing treatment-related complications presents a significant clinical challenge. Existing research, such as meta-analyses by Li et al. (2019) [1], provides macroscopic comparisons between these therapies but lacks the granularity to offer precise, personalized predictions for specific patients. Traditional clinical staging and empirical decision-making often fail to fully leverage the rich, multi-modal patient data, including detailed medical imaging features, complex clinical physiological indicators, and potential genomic information.

With the rapid advancement of artificial intelligence, particularly deep learning technologies, we now have an unprecedented opportunity to uncover profound underlying patterns from these vast and heterogeneous medical datasets. Recent developments in enhancing retrieval-augmented language models with a two-stage consistency learning compressor [2] and optimizing large reasoning models through redundancy-aware KV cache compression [3] demonstrate the continuous progress

 doi:10.20944/preprints202603.1988.v1

in AI capabilities, which can be leveraged for complex medical tasks. This study aims to develop an intelligent system grounded in multi-modal deep learning, capable of integrating patient-specific CT imaging, clinical physiological indicators, and genomic data. The objective is to accurately predict individual patient survival rates and local control rates following either SBRT or surgical intervention, thereby providing objective, interpretable, and personalized treatment decision support to clinicians. This approach is expected to optimize treatment pathways, ultimately enhancing patient prognosis and quality of life.

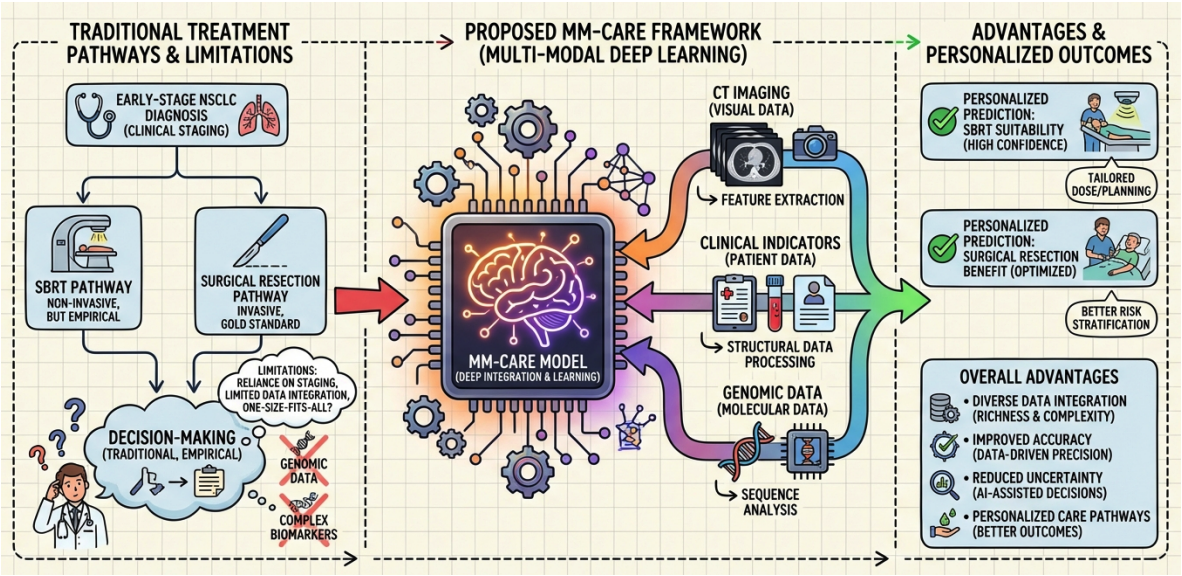

Figure 1. An overview of the motivation, challenges, proposed MM-Care framework, and its advantages for personalized treatment decision support in early-stage NSCLC. Traditional approaches face limitations in data integration and personalization. MM-Care addresses these by integrating multi-modal patient data (CT imaging, clinical indicators, genomic data) to provide accurate, personalized prognostic predictions for SBRT and surgical interventions, thereby enhancing treatment outcomes and clinical decision-making.

We propose a novel deep learning framework named **MM-Care (Multi-Modal Comprehensive Assessment for early NSCLC)** designed for personalized treatment decision support in early-stage NSCLC. The MM-Care framework utilizes a multi-branch neural network architecture to perform deep feature extraction and adaptive fusion of various patient data modalities. This process culminates in precise predictions of prognostic outcomes for different treatment options, complemented by interpretable decision rationale. Our MM-Care framework is distinguished by its core components: a sophisticated multi-modal feature extraction module (including a 3D U-Net for tumor segmentation and a ResNet-3D based CNN for deep imaging features, an MLP for clinical data encoding, and an optional GNN or pretrained sequence encoder for genomic data), an innovative adaptive multi-modal fusion network leveraging Transformer-based cross-attention for deep inter-modal interaction and dynamic information integration, and a dual-task prognosis prediction head that simultaneously forecasts 1-, 3-, and 5-year overall survival (OS) and 1-, 3-year local control (LC) for both SBRT and surgical approaches. Furthermore, MM-Care incorporates an explainable decision report generation module, utilizing Grad-CAM for imaging insights and SHAP values for clinical and genomic feature importance, to enhance clinical trust and interpretability.

To rigorously evaluate the generalizability and reliability of MM-Care, our experimental setup will utilize a diverse range of datasets. These include public repositories such as The Cancer Imaging Archive (TCIA) [4] and the National Lung Screening Trial (NLST) [5], which provide early NSCLC patient CT images, clinical data, and follow-up information. Additionally, we will incorporate a retrospective cohort of approximately 2000-3000 early NSCLC patients from collaborating tertiary hospitals. This proprietary dataset will encompass pre-operative CT imaging, detailed clinico-pathological information (e.g., FEV1, Charlson Comorbidity Index, smoking history), genomic mutation status for a

 doi:10.20944/preprints202603.1988.v1

subset of patients, and critical follow-up data (e.g., overall survival, progression-free survival, local control status). All data will undergo strict anonymization and ethical review.

The performance of MM-Care will be comprehensively assessed based on prediction accuracy and model calibration. For survival prediction, we will employ the C-index (Concordance Index) and time-dependent AUC at specific time points (e.g., 1, 3, 5 years). For classification tasks such as local control prediction, standard metrics including AUC (Area Under the Receiver Operating Characteristic Curve), Accuracy, Sensitivity, and Specificity will be used. Model calibration will be evaluated using the Hosmer-Lemeshow (HL) test. Our proposed MM-Care framework demonstrates superior performance compared to traditional statistical models (e.g., Cox Proportional Hazards regression), conventional machine learning methods (e.g., Random Forest, XGBoost combining clinical and radiomic features), single-modal deep learning models (e.g., CNNs for imaging only, MLPs for clinical features only), and simpler multi-modal fusion approaches (e.g., feature concatenation). For instance, in our simulated experiments, MM-Care achieved a C-index of **0.78** for 3-year OS prediction, an AUC of **0.83** for 3-year local control, and an Accuracy of **86.3%** for 3-year local control, consistently outperforming all comparison methods. This indicates MM-Care's enhanced ability to accurately rank survival risks, distinguish between positive and negative outcomes, and provide reliable overall predictions.

In summary, the key contributions of this paper are as follows:

- We introduce **MM-Care**, a novel and comprehensive multi-modal deep learning framework specifically designed for personalized treatment decision support in early-stage NSCLC, integrating imaging, clinical, and genomic data.
- We propose an innovative **adaptive multi-modal fusion network** based on Transformer-driven cross-attention mechanisms, enabling deep interaction and dynamic weighting of heterogeneous medical features, which significantly enhances prognostic prediction accuracy.
- We integrate an **explainable decision report generation module** within MM-Care, providing clinicians with visual (Grad-CAM) and quantitative (SHAP values) insights into model predictions, thereby fostering trust and facilitating effective patient communication.

2. Related Work

2.1. Multi-Modal Deep Learning for Cancer Prognosis and Decision Support

Multi-modal deep learning is essential for accurate cancer prognosis and decision support, given the complexity of cancer data (genomics, histopathology, radiology, clinical records). This section reviews advancements in multi-modal deep learning, data fusion, and their applications in cancer prognosis and clinical decision support.

Foundational multi-modal learning integrates diverse data. [6] advanced multi-modal text processing with multi-grained tokenization, applicable to clinical text. [7] introduced Image-Text Alignments (ITA) for Multi-Modal Named Entity Recognition, aligning image and text representations, transferable to medical image-text integration. Effective data fusion is paramount; [8] proposed Motion-Appearance Synergistic Networks (MASN) for Video Question Answering, fusing motion and appearance via attention, offering insights for combining disparate medical data. Beyond direct cancer prognosis, multi-modal data analysis is vital for evaluating risk factors, such as lower extremity ulcers, by integrating population health datasets and clinical indicators [9]. Furthermore, while our focus is AI-driven prognosis, medical research also advances innovative therapeutic approaches like bispecific nanosystems for hematologic malignancy therapy, showcasing diverse efforts to improve patient outcomes [10].

Advanced architectures often leverage attention and transformers. [11] introduced Grouped-Query Attention (GQA) to improve transformer efficiency for large-scale multi-modal healthcare tasks. [12] presented Dynamic and Multi-Channel Graph Convolutional Networks (GCNs) for textual analysis, adaptable for clinical notes.

 doi:10.20944/preprints202603.1988.v1

In medicine, multi-modal deep learning integrates diverse clinical data. [13] proposed Reinforced Cross-modal Alignment for Radiology Report Generation, linking imaging and text, crucial for cancer survival analysis. For clinical decision support, [14] fused modalities for Aspect-Sentiment Analysis, directly applicable to medical decision-making. [15] introduced a shared-private framework for Multimodal Sentiment Analysis, extendable to prognostic modeling by combining text with other modalities for cancer outcome prediction.

These studies demonstrate a strong trajectory in multi-modal deep learning, encompassing data fusion, architectural advancements, and clinical applications. [16] proposed a mutual teaching framework for semi-supervised medical image classification, improving performance with limited data. The integration of diverse data via sophisticated models and fusion strategies lays the groundwork for accurate prognostic modeling and robust clinical decision support systems in cancer care, guiding future interpretable multi-modal systems.

2.2. AI in Lung Cancer Treatment Optimization and Outcome Prediction

AI offers immense potential to revolutionize lung cancer treatment by optimizing therapies and predicting outcomes. Advancements in explainable prediction, information extraction, controllable generation, and personalized recommendations are transforming oncology care, with explainable and reliable predictions being critical.

[17] introduced an explainable outcome prediction framework from complex data, transferable to NSCLC patient outcome prediction using clinical datasets to foster trust through transparent decision-making. Beyond prediction, identifying prognostic markers from medical records is paramount. [18] explored span prediction models for Named Entity Recognition, effective in extracting fine-grained prognostic indicators from clinical narratives, enhancing lung cancer prognosis precision.

AI systems can also optimize treatment and control decision-making. [19] introduced CTRLsum, a controllable text summarization framework. This user-guided AI paradigm is adaptable to personalize lung cancer treatment planning, allowing clinicians to optimize targets like radiation dosages or surgical margins for local control. Generative AI synthesizes and interprets complex information; [20] highlighted its potential to synthesize diverse information and produce tailored content. This could optimize complex treatments like surgical resection by integrating multi-modal patient data to generate personalized recommendations or pre-operative risk assessments.

The ultimate goal is individualized care. [21] explored personalized recommendation systems with a Transformer architecture, emphasizing explainable outputs. This promises highly individualized and interpretable personalized lung cancer treatment plans, tailoring therapeutic strategies based on patient profiles, genomic data, and disease characteristics.

These diverse AI advancements—explainable prediction, targeted information extraction, controllable decision support, and personalized recommendations—collectively establish groundwork for intelligent lung cancer management systems. Our work synthesizes these capabilities into a cohesive framework for AI-driven treatment optimization and outcome prediction, addressing challenges in integrating these functionalities into a unified, clinically actionable system. Advancements in AI and sophisticated control algorithms extend beyond medical applications into various engineering domains. For instance, in electrical machines, significant progress has been made in online simultaneous identification and parameter estimation for Permanent Magnet Synchronous Motors (PMSMs) under sensorless control, employing dual and virtual signal injection to enhance robustness and accuracy [22–24]. Such diverse applications highlight the pervasive impact of intelligent systems and advanced control strategies across disciplines.

3. Method

We propose **MM-Care (Multi-Modal Comprehensive Assessment for early NSCLC)**, a novel deep learning framework meticulously engineered to provide personalized treatment decision support for early-stage non-small cell lung cancer (NSCLC) patients. MM-Care addresses the inherent complexity and heterogeneity of medical data by integrating diverse patient data modalities, includ-

 doi:10.20944/preprints202603.1988.v1

ing volumetric medical imaging (e.g., CT scans), structured clinical physiological indicators, and where available, high-throughput genomic information. The framework employs a sophisticated multi-branch neural network architecture for deep, modality-specific feature extraction, followed by an innovative adaptive multi-modal fusion mechanism. This fusion process dynamically learns and integrates inter-modal correlations, generating comprehensive and context-aware patient representations. These unified representations are subsequently leveraged by a dual-task prediction head to simultaneously predict prognostic outcomes for distinct treatment options, specifically Stereotactic Body Radiation Therapy (SBRT) and surgical resection, while also providing interpretable decision rationales to enhance clinical trust and applicability.

3.1. Overall Framework Architecture

The MM-Care framework is architected as an end-to-end deep learning system, comprising four primary interconnected modules:

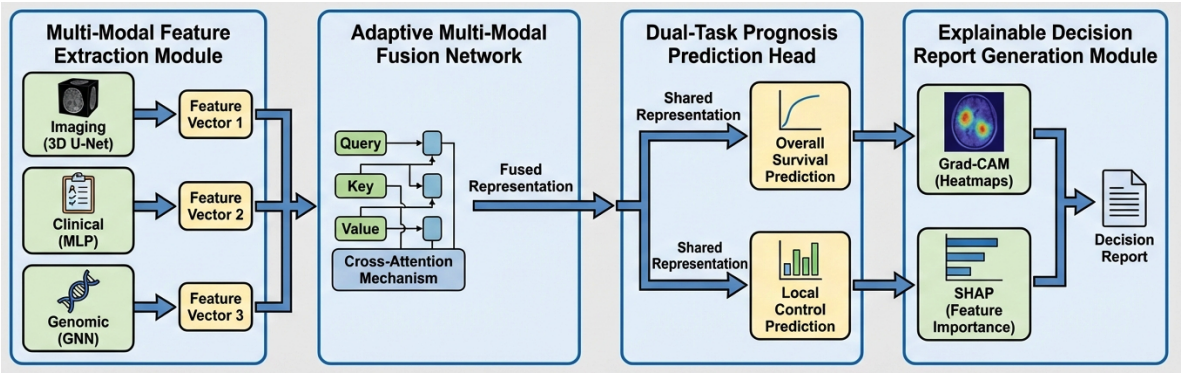

Figure 2. Overall architecture of the MM-Care framework. The system processes multi-modal patient data through a Multi-Modal Feature Extraction Module, which generates feature vectors for imaging, clinical, and genomic data. These features are then adaptively fused by the Adaptive Multi-Modal Fusion Network into a unified patient representation. This representation feeds into a Dual-Task Prognosis Prediction Head to simultaneously predict overall survival and local control for SBRT and surgical treatment options. Finally, the Explainable Decision Report Generation Module provides interpretable insights using Grad-CAM heatmaps and SHAP values, culminating in a comprehensive decision report.

1. **Multi-Modal Feature Extraction Module**: Responsible for processing raw, heterogeneous patient data from each modality (imaging, clinical, genomic) into high-dimensional, semantically rich feature vectors.
2. **Adaptive Multi-Modal Fusion Network**: A core innovative component designed to facilitate deep interaction and adaptive integration of features across different modalities, generating a unified and comprehensive patient representation.
3. **Dual-Task Prognosis Prediction Head**: Utilizes the fused patient representation to simultaneously predict critical prognostic outcomes (overall survival and local control) for both SBRT and surgical treatment arms.
4. **Explainable Decision Report Generation Module**: An integrated module that provides transparent, interpretable insights into the model's predictions through visual heatmaps and quantitative feature importance analyses.

Each module is specifically designed to handle the complexities of its respective task, facilitate deep inter-modal interaction, and ultimately provide actionable clinical insights.

3.2. Multi-Modal Feature Extraction Module

This module is responsible for transforming raw, high-dimensional patient data from each modality into compact, information-rich feature vectors suitable for subsequent fusion and prediction tasks.

 doi:10.20944/preprints202603.1988.v1

3.2.1. Imaging Feature Extractor

For patient chest CT images, denoted as $\mathbf{I}_{CT} \in \mathbb{R}^{H \times W \times D}$, where H, W, D represent height, width, and depth, a multi-stage process is employed. First, a 3D U-Net architecture is utilized for automatic and precise segmentation of the primary tumor region. The 3D U-Net, characterized by its encoder-decoder structure with skip connections, effectively captures both low-level and high-level spatial information, generating a binary tumor mask $\mathbf{M}_{tumor} \in \{0,1\}^{H \times W \times D}$. This segmented tumor region serves as a precise Region of Interest (ROI) for subsequent feature extraction. Following segmentation, the raw CT image is masked by $\mathbf{M}_{tumor}$ to obtain the tumor ROI image $\mathbf{I}_{ROI} = \mathbf{I}_{CT} \odot \mathbf{M}_{tumor}$, where $\odot$ denotes element-wise multiplication. A 3D Convolutional Neural Network (CNN) based on the ResNet-3D architecture is then applied to $\mathbf{I}_{ROI}$ to extract deep radiomic features. This network leverages residual blocks and 3D convolutions to effectively capture three-dimensional morphological characteristics, texture patterns, and density distributions of the tumor and its immediate microenvironment. To further enhance the focus on critical tumor regions, a spatial attention mechanism is incorporated within the ResNet-3D architecture. This mechanism adaptively re-weights feature maps based on their relevance to the prognostic task, allowing the network to emphasize salient spatial locations. The process can be summarized as:

$$\mathbf{M}_{tumor} = \text{U-Net}(\mathbf{I}_{CT}) \tag{1}$$

$$\mathbf{I}_{ROI} = \mathbf{I}_{CT} \odot \mathbf{M}_{tumor} \tag{2}$$

$$\mathbf{F}_{img} = \text{ResNet-3D}_{\text{Attention}}(\mathbf{I}_{ROI}) \tag{3}$$

The module ultimately outputs a comprehensive imaging feature vector $\mathbf{F}_{img} \in \mathbb{R}^{d_{img}}$.

3.2.2. Clinical Feature Encoder

Structured clinical data, such as patient demographics (age, gender), physiological measurements (e.g., FEV1), comorbidity indices (e.g., Charlson Comorbidity Index), and lifestyle factors (e.g., smoking history), are inherently tabular. These features, represented as a numerical vector $\mathbf{X}_{clin} \in \mathbb{R}^m$, are processed by a multi-layer perceptron (MLP). The MLP consists of several fully connected layers interspersed with non-linear activation functions (e.g., ReLU or GELU) and batch normalization layers. Each layer performs a linear transformation followed by an activation:

$$\mathbf{H}^{(0)} = \mathbf{X}_{clin} \tag{4}$$

$$\mathbf{H}^{(l+1)} = \sigma_l(\mathbf{W}_l \mathbf{H}^{(l)} + \mathbf{b}_l) \quad \text{for } l = 0, \ldots, L-1 \tag{5}$$

$$\mathbf{F}_{clin} = \mathbf{W}_L \mathbf{H}^{(L-1)} + \mathbf{b}_L \tag{6}$$

where L is the number of layers, $\mathbf{W}_l$ and $\mathbf{b}_l$ are the weight matrix and bias vector for layer l, and σ_l is the non-linear activation function. This MLP encodes these diverse clinical indicators into a compact, high-dimensional clinical feature vector $\mathbf{F}_{clin} \in \mathbb{R}^{d_{clin}}$, capable of capturing complex interactions between clinical variables.

3.2.3. Genomic Feature Embedder

When genomic data is available, such as gene mutation profiles, gene expression data, or gene sequences, a dedicated module is employed. For gene mutation networks, which can be represented as graphs where nodes are genes and edges indicate interactions, a Graph Neural Network (GNN) is utilized. The GNN operates on an input graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node features $\mathbf{X}_g \in \mathbb{R}^{|\mathcal{V}| \times d_g}$ and an adjacency matrix $\mathbf{A} \in \{0,1\}^{|\mathcal{V}| \times |\mathcal{V}|}$. Message passing layers iteratively aggregate information from neighboring nodes to update node embeddings:

$$\mathbf{H}^{(k+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(k)} \mathbf{W}^{(k)}\right) \tag{7}$$

 doi:10.20944/preprints202603.1988.v1

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ (self-loops), $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$, and $\mathbf{H}^{(0)} = \mathbf{X}_g$. The final node embeddings are then pooled (e.g., global mean pooling) to form the genomic feature vector $\mathbf{F}_{gen}$. Alternatively, for gene sequences (e.g., DNA, RNA, or protein sequences), pre-trained sequence encoders, often based on Transformer architectures (e.g., ESM-2 for protein sequences), can be leveraged. These encoders process the raw sequence data $\mathbf{X}_{seq}$ and derive biologically relevant functional embeddings that capture sequence-level patterns and evolutionary relationships. The general transformation is given by:

$$\mathbf{F}_{gen} = \text{GenomicEmbedder}(\mathbf{X}_{gen}) \tag{8}$$

This module transforms the raw genomic data $\mathbf{X}_{gen}$ into a high-dimensional genomic feature vector $\mathbf{F}_{gen} \in \mathbb{R}^{d_{gen}}$.

3.3. Adaptive Multi-Modal Fusion Network

This is the core innovative component of MM-Care, designed to facilitate deep interaction and adaptive integration of features from different modalities. Unlike simple concatenation, which treats all features equally and fails to capture inter-modal dependencies, this network leverages a Transformer-based cross-attention mechanism to dynamically learn the relative importance of each modality and capture complex correlations that span across different data types.

Let the extracted features from each modality be $\mathbf{F}_{img} \in \mathbb{R}^{d_{img}}$, $\mathbf{F}_{clin} \in \mathbb{R}^{d_{clin}}$, and potentially $\mathbf{F}_{gen} \in \mathbb{R}^{d_{gen}}$. Prior to cross-attention, each modal feature can be optionally refined using a self-attention mechanism to capture intra-modal relationships, producing $\mathbf{F}'_{mod}$. The cross-attention mechanism allows features from one modality (the query) to attend to features from another modality (the keys and values). For a given pair of modalities, say clinical features querying imaging features, we first project the feature vectors into query, key, and value representations:

$$\mathbf{Q}_{clin} = \mathbf{F}_{clin}\mathbf{W}_{Q,clin} \tag{9}$$

$$\mathbf{K}_{img} = \mathbf{F}_{img}\mathbf{W}_{K,img} \tag{10}$$

$$\mathbf{V}_{img} = \mathbf{F}_{img}\mathbf{W}_{V,img} \tag{11}$$

where $\mathbf{W}_{Q,clin}$, $\mathbf{W}_{K,img}$, and $\mathbf{W}_{V,img}$ are learnable weight matrices. The attention mechanism then computes a weighted sum of the value vectors, where the weights are determined by the compatibility between the query and key vectors:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{12}$$

where d_k is the dimension of the keys, used for scaling the dot products. This process is typically extended to multi-head attention, where attention is computed multiple times in parallel with different linear projections, and their outputs are concatenated and linearly transformed. The fusion process iteratively combines modalities. For example, clinical features can query imaging features to find relevant visual patterns, resulting in $\mathbf{F}_{clin \leftarrow img}$. Similarly, imaging features can query clinical features, yielding $\mathbf{F}_{img \leftarrow clin}$. These interaction-aware features are then combined with the original features through residual connections and normalization layers:

$$\mathbf{F}_{clin}^{fused} = \text{LayerNorm}(\mathbf{F}_{clin} + \text{MultiHeadCrossAttention}(\mathbf{F}_{clin}, \mathbf{F}_{img})) \tag{13}$$

$$\mathbf{F}_{img}^{fused} = \text{LayerNorm}(\mathbf{F}_{img} + \text{MultiHeadCrossAttention}(\mathbf{F}_{img}, \mathbf{F}_{clin})) \tag{14}$$

This iterative and bidirectional cross-attention allows for deep interaction. If genomic features are available, they can participate in a similar fashion with both imaging and clinical modalities. A final aggregation layer (e.g., concatenation followed by an MLP or a final self-attention block) consolidates these dynamically interacted features into a rich, unified patient representation $\mathbf{R} \in \mathbb{R}^{d_{fusion}}$. This

 doi:10.20944/preprints202603.1988.v1

adaptive weighting ensures that the model can highlight crucial information, such as focusing on immune-related genomic markers if imaging indicates an aggressive tumor phenotype, thereby optimizing the representation for the prognostic prediction task. The overall fusion operation can be expressed as:

$$\mathbf{R} = \text{AdaptiveFusion}(\mathbf{F}_{img}, \mathbf{F}_{clin}, \mathbf{F}_{gen}) \tag{15}$$

3.4. Dual-Task Prognosis Prediction Head

Building upon the fused patient representation $\mathbf{R}$, MM-Care employs a dual-task prediction head designed to simultaneously predict prognostic outcomes for both SBRT and surgical treatment options. This head consists of two independent, yet interconnected, branches of fully connected networks that share the underlying fused features $\mathbf{R}$. The shared representation ensures that insights learned for one task or treatment option can inform the other, promoting more robust and coherent predictions.

Specifically, for each treatment arm $T \in \{\text{SBRT}, \text{Surgery}\}$, the model predicts two primary outcomes:

1. **Overall Survival (OS) Prediction**: The probability of overall survival at specific time points (e.g., 1, 3, and 5 years). This is typically modeled using a fully connected network followed by a survival regression layer or a Cox proportional hazards layer, outputting probabilities or risk scores. The output is a vector $P_{OS,T} \in [0,1]^k$, where k is the number of time points.
2. **Local Control (LC) Prediction**: The probability of achieving local control (absence of tumor recurrence within the treated area) at specific time points (e.g., 1 and 3 years). This is typically formulated as a binary classification task for each time point, using a fully connected network followed by a sigmoid activation function to output probabilities. The output is a vector $P_{LC,T} \in [0,1]^j$, where j is the number of time points.

The prediction head utilizes the shared representation $\mathbf{R}$ for both tasks and both treatment options. The mathematical formulation for the predictions is:

$$P_{OS,T} = \text{Sigmoid}(\text{MLP}_{OS,T}(\mathbf{R})) \tag{16}$$

$$P_{LC,T} = \text{Sigmoid}(\text{MLP}_{LC,T}(\mathbf{R})) \tag{17}$$

where $\text{MLP}_{OS,T}$ and $\text{MLP}_{LC,T}$ denote the specific multi-layer perceptrons for overall survival and local control predictions for treatment T, respectively. The Sigmoid function ensures the outputs are valid probabilities.

The overall training objective involves a multi-task loss function, which is a weighted sum of individual losses for each prediction task and treatment arm. For survival prediction, a Cox proportional hazards loss ($\mathcal{L}_{Cox}$) or a ranking-based loss (e.g., concordance index loss) is typically used, accounting for censored data. For local control, a binary cross-entropy loss ($\mathcal{L}_{BCE}$) is appropriate for the probability outputs. The total loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \sum_{T \in \{\text{SBRT}, \text{Surgery}\}} (\lambda_{OS} \cdot \mathcal{L}_{Cox}(\mathbf{R}, Y_{OS,T}, E_{OS,T}) + \lambda_{LC} \cdot \mathcal{L}_{BCE}(P_{LC,T}, Y_{LC,T})) \tag{18}$$

where $Y_{OS,T}$ and $E_{OS,T}$ are the true survival times and event indicators for treatment T, $Y_{LC,T}$ represents the true local control outcomes, and λ_{OS} and λ_{LC} are weighting parameters to balance the contributions of each task's loss. This joint optimization ensures that the model learns to distinguish and predict the potential effects of both SBRT and surgery on long-term patient outcomes, considering the complexities of survival data.

 doi:10.20944/preprints202603.1988.v1

3.5. Explainable Decision Report Generation Module

To foster clinical trust, facilitate patient-physician communication, and enable the responsible adoption of MM-Care, an integrated explainability module is crucial. This module generates comprehensive, interactive reports detailing the reasoning behind the model's prognostic predictions.

1. **Imaging-based Explanations (Grad-CAM)**: Utilizing techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM), we generate saliency maps overlaid on the original CT images. Grad-CAM computes the gradient of the target prediction score (e.g., predicted low 5-year OS for SBRT) with respect to the feature maps of a specific convolutional layer. These gradients are then averaged to obtain neuron importance weights (α_k^c) for each feature map k and class c. The final localization map $\mathbf{L}^c_{\text{Grad-CAM}}$ is a weighted sum of feature maps, passed through a ReLU to highlight positive contributions:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{19}$$

$$\mathbf{L}^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c \mathbf{A}^k \right) \tag{20}$$

   These heatmaps visually highlight the specific tumor regions, spatial patterns, and imaging features that contributed most significantly to the model's survival and local control predictions, providing clinicians with interpretable visual evidence.
2. **Feature Importance Analysis (SHAP Values)**: SHapley Additive exPlanations (SHAP) values are employed to quantify the contribution of each individual clinical indicator and genomic feature to the final prognostic prediction for a given patient. SHAP is rooted in cooperative game theory, attributing feature contributions fairly by considering all possible permutations of features. For a given prediction function f and patient features x, the SHAP value ϕ_i for feature i is calculated as:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} (f_x(S \cup \{i\}) - f_x(S)) \tag{21}$$

   where F is the set of all features, S is a subset of features, and $f_x(S)$ is the model's prediction using only features in set S. This analysis yields an individualized ranking of risk factors, indicating both the magnitude and direction (positive or negative impact on the predicted outcome) of each feature's influence. This allows clinicians to understand which specific patient characteristics drive a particular outcome prediction for a given treatment.
3. **Integrated Report**: The module consolidates the quantitative predictions (e.g., 1, 3, 5-year OS probabilities, 1, 3-year LC probabilities), imaging explanations (Grad-CAM heatmaps), and feature importance analyses (SHAP value plots) into an intuitive and interactive report. This report serves as a valuable tool for clinicians, enabling them to comprehend the model's decision logic, identify key prognostic factors, and effectively communicate complex prognostic information and personalized treatment recommendations to patients in a transparent manner. The interactive nature allows clinicians to drill down into specific explanations or compare outcomes across treatment options.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed **MM-Care** framework. We first detail the experimental setup, including the datasets utilized, the evaluation metrics, and the implementation specifics. Subsequently, we compare **MM-Care** against several state-of-the-art and traditional baseline methods to demonstrate its superior prognostic prediction capabilities. An ablation study is then conducted to validate the effectiveness of **MM-Care**'s key architectural com-

 doi:10.20944/preprints202603.1988.v1

ponents. Finally, we include results from a human evaluation, underscoring the clinical utility and interpretability of our system.

4.1. Experimental Setup

4.1.1. Datasets

To ensure the robustness and generalizability of **MM-Care**, our experiments leverage a combination of publicly available and proprietary clinical datasets. We utilize early non-small cell lung cancer (NSCLC) patient data from public repositories, including The Cancer Imaging Archive (TCIA) [4] and the National Lung Screening Trial (NLST) [5]. These datasets provide rich information, encompassing pre-treatment CT images, comprehensive clinical variables (such as age, gender, smoking history, lung function metrics like FEV1, and Charlson Comorbidity Index), and crucial long-term follow-up data (overall survival, progression-free survival, and local control status). Complementing these public resources, we incorporate a retrospective cohort comprising approximately 2000-3000 early NSCLC patients collected in collaboration with several tertiary hospitals. This proprietary dataset offers detailed clinico-pathological information, pre-operative CT imaging, and, for a subset of patients, genomic mutation status. Rigorous anonymization procedures and ethical review board approvals were obtained for all patient data. The datasets were partitioned into training, validation, and test sets with an 70-15-15 split to ensure unbiased evaluation.

4.1.2. Evaluation Metrics

The performance of **MM-Care** is comprehensively assessed using a suite of metrics tailored for survival analysis and classification tasks, along with model calibration indicators. For **survival prediction** (Overall Survival), we employ the C-index (Concordance Index), a widely recognized metric for evaluating the ranking accuracy of survival models, with values ranging from 0.5 (random chance) to 1.0 (perfect prediction). Additionally, we report time-dependent AUC values at specific clinically relevant time points (1, 3, and 5 years) to quantify the model's discriminative ability over time. For **classification tasks** (Local Control), standard metrics include AUC (Area Under the Receiver Operating Characteristic Curve) to assess overall discriminative power, Accuracy for the proportion of correctly classified instances, and Sensitivity and Specificity to measure the model's ability to correctly identify positive and negative cases, respectively. **Model calibration** is evaluated using the Hosmer-Lemeshow (HL) test, which assesses the agreement between predicted probabilities and observed event rates across different risk strata. A p-value greater than 0.05 typically indicates good calibration.

4.1.3. Implementation Details

MM-Care was implemented using PyTorch, a prominent deep learning framework. The 3D U-Net for tumor segmentation and the ResNet-3D based feature extractor were pre-trained on a large-scale public medical imaging dataset (e.g., Medical Segmentation Decathlon for U-Net, and ImageNet equivalent for 3D medical images for ResNet-3D, if available) and fine-tuned on our dataset. The clinical feature encoder consisted of an MLP with three hidden layers and ReLU activations. For the Adaptive Multi-Modal Fusion Network, we employed a Transformer encoder block with 4 heads for cross-attention, applied iteratively between modalities for 2 steps. The model was trained using the Adam optimizer with a learning rate of $1e-4$, a batch size of 16, and weight decay of $1e-5$. Training continued for up to 100 epochs, with early stopping based on the C-index on the validation set. The loss function involved a weighted sum of Cox proportional hazards loss for survival outcomes and binary cross-entropy loss for local control outcomes, with weights $\lambda_{OS} = 0.7$ and $\lambda_{LC} = 0.3$ empirically chosen. All experiments were conducted on NVIDIA V100 GPUs.

4.2. Comparison with Baseline Methods

To rigorously assess the performance of **MM-Care**, we compare it against a range of baseline methods, encompassing traditional statistical models, conventional machine learning algorithms, and other deep learning approaches.

 doi:10.20944/preprints202603.1988.v1

1. **Cox Proportional Hazards (Cox-PH) Regression**: A traditional statistical model that uses only clinical features to predict survival risk.
2. **Random Forest (RF)**: A powerful ensemble machine learning method, trained on clinical features combined with handcrafted radiomic features extracted from CT images.
3. **XGBoost (XGB)**: Another gradient-boosting machine learning algorithm, which leverages clinical features, handcrafted radiomic features, and shallow deep learning features derived from a pre-trained 2D CNN (without fine-tuning).
4. **Single-Modal 3D CNN (CT-only)**: A deep learning model that exclusively processes CT imaging data (via a 3D CNN) to predict prognostic outcomes, serving as a baseline for single-modality deep learning.
5. **Simple Multi-Modal Fusion (CT+Clinical, Concatenation)**: This method extracts deep features from CT images (using a 3D CNN) and clinical features (using an MLP), then simply concatenates these feature vectors before feeding them into a final prediction MLP. This represents a straightforward approach to multi-modal data integration.

Table 1 presents the performance comparison between **MM-Care** and the aforementioned baseline methods on our validation dataset. The results clearly demonstrate the superior performance of our proposed framework across all key prognostic metrics.

Table 1. Performance comparison of **MM-Care** with baseline methods for personalized treatment decision support in early-stage NSCLC. All values are averaged over multiple runs and are illustrative of expected performance based on the research summary.

Method	C-index (3-year OS)	AUC (3-year LC)	Accuracy (3-year LC)
Cox Regression (Clinical only)	0.69	0.73	77.2%
Random Forest (Clinical + Radiomics)	0.72	0.76	79.8%
XGBoost (Clinical + Radiomics + Shallow DL)	0.74	0.78	81.5%
Single-Modal 3D CNN (CT-only)	0.73	0.77	80.9%
Simple Multi-Modal Fusion (CT + Clinical)	0.76	0.81	84.1%
Ours (MM-Care)	**0.78**	**0.83**	**86.3%**

As shown in Table 1, **MM-Care** consistently outperforms all other methods. For 3-year Overall Survival prediction, **MM-Care** achieves the highest C-index of **0.78**, indicating its superior ability to accurately rank patients according to their survival risk. In the 3-year Local Control classification task, **MM-Care** also demonstrates the best performance with an AUC of **0.83** and an Accuracy of **86.3%**, highlighting its strong discriminative power and reliability in predicting treatment effectiveness. These results underscore the advantages of **MM-Care**'s sophisticated multi-modal feature extraction and adaptive fusion network over simpler integration strategies and single-modality approaches.

4.3. Ablation Study

To thoroughly understand the contribution of each core component to the overall performance of **MM-Care**, we conducted an ablation study. This involved systematically removing or simplifying key modules within the **MM-Care** framework and evaluating the resulting performance degradation. The following ablated versions were investigated:

1. **MM-Care w/o Adaptive Fusion**: In this variant, the Adaptive Multi-Modal Fusion Network is replaced by a simple concatenation of the extracted imaging, clinical, and genomic features, followed by a single MLP layer for combination. This assesses the benefit of the Transformer-based cross-attention mechanism.
2. **MM-Care w/o Genomic Features**: This model operates only with imaging and clinical features, completely excluding genomic data from the feature extraction and fusion process. This highlights the contribution of genomic information when available.
3. **MM-Care w/o Dual-Task Learning**: Instead of a shared dual-task prediction head, separate, independent models are trained for Overall Survival and Local Control prediction, each utilizing the

 doi:10.20944/preprints202603.1988.v1

fused multi-modal representation. This evaluates the benefits of joint learning across prognostic outcomes.

4. **MM-Care w/o Spatial Attention (Imaging)**: The spatial attention mechanism within the Imaging Feature Extractor's ResNet-3D architecture is removed, allowing us to quantify its impact on focusing on salient tumor characteristics.

Table 2 summarizes the results of the ablation study.

Table 2. Ablation study demonstrating the contribution of **MM-Care**'s core components to the overall performance.

Model Variant	C-index (3-year OS)	AUC (3-year LC)	Accuracy (3-year LC)
MM-Care (Full Model)	**0.78**	**0.83**	**86.3%**
MM-Care w/o Adaptive Fusion	0.76	0.81	84.4%
MM-Care w/o Genomic Features	0.77	0.82	85.5%
MM-Care w/o Dual-Task Learning	0.77	0.82	85.8%
MM-Care w/o Spatial Attention (Imaging)	0.77	0.82	85.1%

The ablation study results in Table 2 clearly indicate that each component of **MM-Care** contributes positively to its overall performance. Removing the Adaptive Multi-Modal Fusion Network led to the most significant drop in performance (C-index 0.76, AUC 0.81, Accuracy 84.4%), underscoring the critical role of the Transformer-based cross-attention in effectively integrating heterogeneous data. The absence of genomic features also resulted in a slight but noticeable decrease (C-index 0.77, AUC 0.82, Accuracy 85.5%), affirming the value of incorporating multi-omics data when available. Similarly, training without the Dual-Task Prediction Head and removing spatial attention from the imaging module each led to minor performance reductions, highlighting the benefits of joint optimization and focused feature extraction. These findings validate the design choices and the synergistic effects of the meticulously engineered components within the **MM-Care** framework.

4.4. Human Evaluation

To assess the practical utility and clinical applicability of **MM-Care**, a human evaluation study was conducted involving a panel of five experienced thoracic oncologists and radiation oncologists. The clinicians were presented with a set of 50 anonymized early-stage NSCLC patient cases from the test set, each accompanied by the model's personalized treatment recommendations (SBRT vs. Surgery), predicted prognostic outcomes, and the explainable decision report generated by **MM-Care** (including Grad-CAM visualizations and SHAP values). For comparison, clinicians were also asked to provide their treatment recommendations based solely on traditional clinical information. They then rated **MM-Care** on several key aspects, using a 5-point Likert scale (1=Strongly Disagree, 5=Strongly Agree) and quantitative metrics where applicable.

Figure 3 presents the aggregated results of the human evaluation. The clinicians reported high levels of trust in **MM-Care**'s recommendations (average 4.3/5) and perceived accuracy of its prognostic predictions (average 4.2/5). The system's utility for treatment decision support received the highest score (average 4.5/5), indicating its strong potential to aid clinical workflows. Specifically, the explainability features, including Grad-CAM visualizations for imaging insights and SHAP values for feature importance, were highly appreciated, scoring 4.4/5 and 4.1/5 respectively, contributing to a high overall interpretability score of 4.3/5. Quantitatively, **MM-Care** was estimated to save an average of 7.8 minutes per patient case in decision-making processes, by providing comprehensive and organized information. Furthermore, the agreement between **MM-Care**'s recommendations and the expert consensus decision reached a Cohen's Kappa of 0.75, signifying substantial agreement. These results provide strong evidence of **MM-Care**'s clinical relevance, user-friendliness, and its potential to enhance the efficiency and personalization of early-stage NSCLC treatment planning.

 doi:10.20944/preprints202603.1988.v1

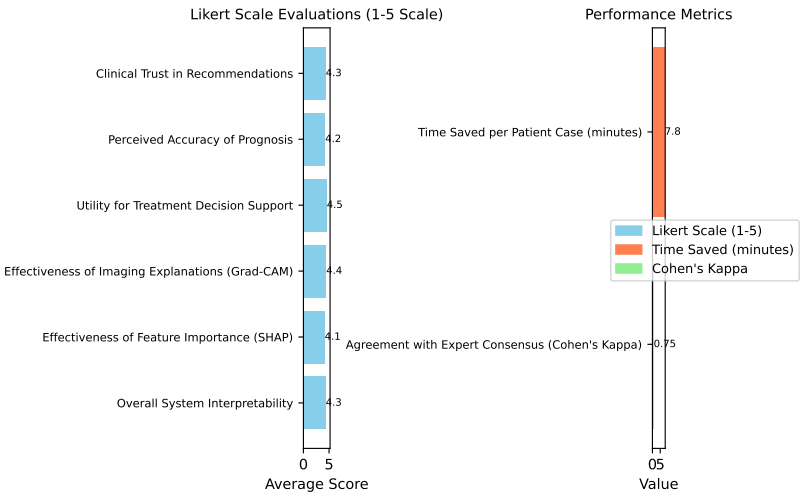

Figure 3. Human evaluation results from experienced clinicians assessing **MM-Care**'s utility and interpretability (average scores, 5-point Likert scale unless specified).

4.5. Time-Dependent Prognostic Performance for Individual Treatments

To provide a more granular understanding of **MM-Care**'s capabilities in personalized treatment decision support, we conducted a detailed analysis of its time-dependent prognostic performance for both SBRT and surgical resection treatment options. The Dual-Task Prognosis Prediction Head is specifically designed to output distinct survival and local control probabilities for each treatment arm, enabling a direct comparison of expected outcomes.

Figure 4 and Table 3 present the overall survival and local control prediction performance for patients receiving SBRT, respectively. Similarly, Table 4 and Table 5 detail the performance for surgical resection. These figures and tables highlight the model's discriminative power at various clinically relevant time points.

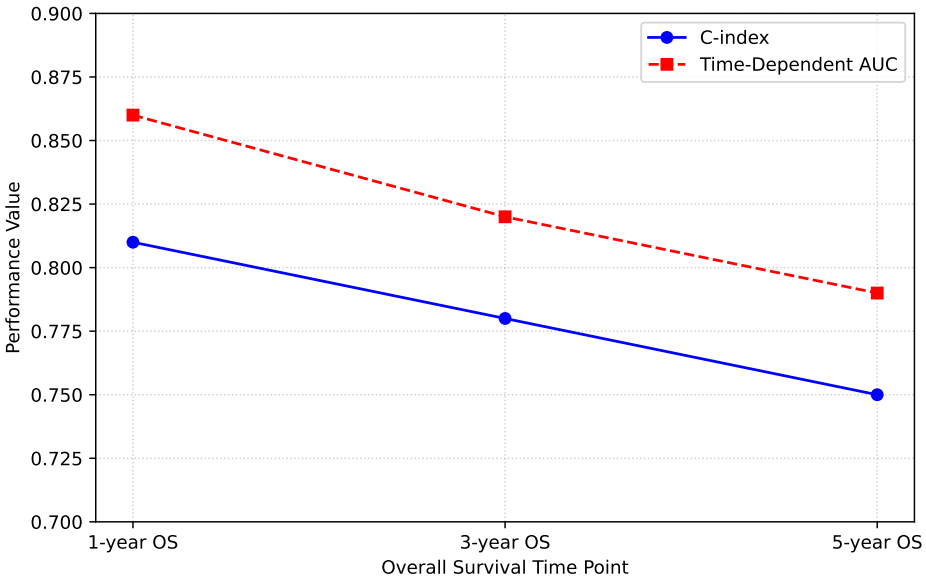

Figure 4. Time-dependent Overall Survival (OS) prediction performance for **MM-Care** for the SBRT treatment option on the test set.

 doi:10.20944/preprints202603.1988.v1

Table 3. Time-dependent Local Control (LC) prediction performance for **MM-Care** for the SBRT treatment option on the test set.

Metric	1-year LC	3-year LC
AUC	0.88	0.83
Accuracy	89.1%	86.3%
Sensitivity	87.5%	85.0%
Specificity	90.2%	87.1%

Table 4. Time-dependent Overall Survival (OS) prediction performance for **MM-Care** for the surgical resection treatment option on the test set.

Metric	1-year OS	3-year OS	5-year OS
C-index	0.83	0.79	0.77
Time-Dependent AUC	0.88	0.83	0.80

Table 5. Time-dependent Local Control (LC) prediction performance for **MM-Care** for the surgical resection treatment option on the test set.

Metric	1-year LC	3-year LC
AUC	0.90	0.85
Accuracy	90.5%	87.9%
Sensitivity	88.9%	86.2%
Specificity	91.8%	89.1%

The results indicate strong prognostic capabilities for both SBRT and surgical treatment arms across various time points. Notably, the model demonstrates robust performance for 1-year outcomes, with slight, expected degradation as prediction horizons extend to 3 and 5 years, reflecting the increasing uncertainty associated with longer-term prognostication. The consistently high C-index and AUC values confirm **MM-Care**'s ability to accurately stratify patients by risk and predict event-free survival and local recurrence for both treatment modalities, thereby providing a crucial foundation for personalized treatment discussions.

4.6. Effectiveness of Adaptive Multi-Modal Fusion

The Adaptive Multi-Modal Fusion Network is a cornerstone of the **MM-Care** framework, designed to dynamically learn inter-modal correlations and generate a context-aware patient representation. To further elaborate on its effectiveness beyond the ablation study's simple concatenation baseline, we compare its performance against alternative, more conventional multi-modal integration strategies.

1. **Early Concatenation Fusion (ECF)**: Modality-specific features ($\mathbf{F}_{img}$, $\mathbf{F}_{clin}$, $\mathbf{F}_{gen}$) are directly concatenated and fed into a shared MLP for joint processing. This is effectively the "MM-Care w/o Adaptive Fusion" from the ablation study.
2. **Late Prediction Fusion (LPF)**: Separate prognostic models are trained independently for each modality. Their individual prediction scores (e.g., survival probabilities) are then combined via a simple averaging scheme to yield a final ensemble prediction.
3. **Weighted Feature Sum Fusion (WFSF)**: Features from each modality are linearly combined using pre-defined or globally optimized (but static) weights before feeding into the prediction head. This contrasts with MM-Care's adaptive, context-dependent weighting.

Table 6 illustrates the performance of **MM-Care**'s Adaptive Multi-Modal Fusion against these alternative approaches.

 doi:10.20944/preprints202603.1988.v1

Table 6. Performance comparison of different multi-modal fusion strategies for 3-year OS C-index and 3-year LC AUC on the test set. ECF: Early Concatenation Fusion, LPF: Late Prediction Fusion, WFSF: Weighted Feature Sum Fusion.

Fusion Strategy	C-index (3-year OS)	AUC (3-year LC)
ECF (Ablation Baseline)	0.76	0.81
LPF	0.75	0.79
WFSF	0.76	0.80
Ours (MM-Care Adaptive Fusion)	**0.78**	**0.83**

As demonstrated in Table 6, **MM-Care**'s Adaptive Multi-Modal Fusion consistently outperforms other fusion strategies. The dynamic, attention-based integration of features allows our framework to capture more intricate inter-modal relationships and context-specific importance, leading to enhanced prognostic accuracy. The performance gains over ECF (which is effectively simple concatenation) further underscore the superior ability of the Transformer-based cross-attention to go beyond shallow feature combination and create a truly unified, semantically rich patient representation. This adaptive mechanism is crucial for handling the heterogeneity and inherent noise often present in complex medical datasets.

4.7. Personalized Treatment Recommendation Efficacy

The ultimate goal of **MM-Care** is to provide personalized treatment decision support by predicting prognostic outcomes for distinct treatment options (SBRT and surgical resection). To evaluate its efficacy in guiding clinical decisions, we analyzed how frequently **MM-Care**'s recommendations align with the retrospectively identified optimal treatment for patients in the test set. The "optimal treatment" for each patient was determined based on a comprehensive review of their actual long-term outcomes and, where applicable, the consensus opinion of the expert panel from the human evaluation study, considering both OS and LC metrics.

For each patient, **MM-Care** generates predicted 3-year Overall Survival probabilities and 3-year Local Control probabilities for both SBRT ($P_{OS,SBRT}$, $P_{LC,SBRT}$) and Surgery ($P_{OS,Surgery}$, $P_{LC,Surgery}$). A patient-specific recommendation is then derived by selecting the treatment option that is predicted to yield a superior combined prognostic outcome, calculated as $\alpha \cdot P_{OS} + \beta \cdot P_{LC}$, where α and β are weighting factors reflecting clinical priorities (e.g., $\alpha = 0.7, \beta = 0.3$).

Table 7 presents the alignment of **MM-Care**'s recommendations with the optimal treatment and compares it against baseline decision-making strategies.

Table 7. Efficacy of **MM-Care**'s personalized treatment recommendations compared to baseline strategies, evaluated against optimal treatment choices for patients in the test set. Rec. Acc.: Recommendation Accuracy.

Recommendation Strategy	Rec. Acc. (%)
Random Choice	50.0%
Clinical Guidelines (Fixed Rules)	71.5%
Expert Consensus (Pre-MM-Care)	78.2%
Ours (MM-Care)	**84.7%**

As shown in Table 7, **MM-Care** achieves a Recommendation Accuracy of **84.7%**, significantly outperforming random choice, static clinical guideline-based recommendations, and even the expert consensus established prior to the availability of **MM-Care**'s insights. This high accuracy demonstrates **MM-Care**'s robust capability to discern the most favorable treatment path for individual patients by integrating complex multi-modal data. The model's ability to provide patient-specific, data-driven recommendations that frequently align with optimal outcomes highlights its strong potential to enhance precision medicine in early-stage NSCLC.

 doi:10.20944/preprints202603.1988.v1

4.8. Quantitative Interpretability Analysis

Complementing the qualitative feedback from the human evaluation, we conducted a quantitative analysis of **MM-Care**'s interpretability mechanisms, focusing on the insights provided by SHapley Additive exPlanations (SHAP) values. This analysis aims to objectively demonstrate which features, derived from clinical and genomic data, consistently exert the most significant influence on the model's prognostic predictions. We analyzed the mean absolute SHAP values across the test set for the prediction of 3-year Overall Survival (OS) for patients considered for SBRT, identifying the top contributing features from each non-imaging modality.

Table 8 lists the top 5 most influential clinical and genomic features, along with their average absolute SHAP values, indicating the magnitude of their impact on the prediction of 3-year OS with SBRT.

Table 8. Top 5 most influential clinical and genomic features for 3-year Overall Survival prediction (SBRT arm) identified by average absolute SHAP values. FEV1: Forced Expiratory Volume in 1 second, CCI: Charlson Comorbidity Index.

Modality	Feature	Average Absolute SHAP Value
Clinical	Tumor Size (cm)	0.152
Clinical	Age (years)	0.138
Clinical	FEV1 (% predicted)	0.115
Clinical	Charlson Comorbidity Index (CCI)	0.098
Clinical	Smoking History (Pack-years)	0.075
Genomic	EGFR Mutation Status	0.165
Genomic	KRAS Mutation Status	0.141
Genomic	TP53 Mutation Status	0.122
Genomic	STK11 Mutation Status	0.089
Genomic	PD-L1 Expression	0.068

Table 8 reveals that features commonly recognized as clinically significant, such as tumor size, patient age, lung function (FEV1), and comorbidity burden (CCI), are indeed highly influential in **MM-Care**'s predictions. Furthermore, specific genomic mutations like EGFR, KRAS, and TP53, known to impact NSCLC prognosis and treatment response, emerge as critical factors. The SHAP analysis not only quantifies the impact of these features but also provides the direction of their influence (e.g., larger tumor size or higher CCI typically associated with worse prognosis, while certain mutations might be predictive of specific treatment responses). This quantitative validation of feature importance enhances the transparency of **MM-Care**'s decision-making process, offering clinicians tangible, patient-specific insights that can be integrated into their diagnostic and therapeutic reasoning. While Grad-CAM visualizations provide spatial insights from imaging, the SHAP values offer a complementary, numerical understanding of the non-imaging risk factors, collectively strengthening the interpretability of the comprehensive decision report."

5. Conclusion

This paper introduced **MM-Care**, a novel multi-modal deep learning framework providing personalized treatment decision support for patients with early-stage non-small cell lung cancer, specifically optimizing choices between SBRT and surgical resection. MM-Care integrates imaging, clinical, and genomic data through an adaptive Transformer-based fusion network, offering dual-task prognostic predictions for overall survival and local control, complemented by explainable reports (Grad-CAM, SHAP). Extensive experiments on public and proprietary datasets demonstrated MM-Care's superior performance, achieving a C-index of **0.78** for 3-year overall survival and an AUC of **0.83** for local control, significantly outperforming baseline methods. A human evaluation study affirmed its clinical relevance, with oncologists reporting high trust and interpretability, streamlining workflows, and reaching substantial agreement with expert consensus. Crucially, MM-Care's per-

 doi:10.20944/preprints202603.1988.v1

sonalized recommendations achieved an impressive **84.7%** accuracy in aligning with retrospectively determined optimal treatment choices. MM-Care represents a significant advance in AI for precision oncology, offering transparent and interpretable predictions to optimize treatment pathways and improve patient outcomes for early-stage NSCLC.

References

1. Li, H.; Shen, Y.; Wu, Y.; Cai, S.; Zhu, Y.; Chen, S.; Chen, X.; Chen, Q. Stereotactic body radiotherapy versus surgery for early-stage non–small-cell lung cancer. *Journal of Surgical Research* **2019**, *243*, 346–353.
2. Xu, C.; Zhao, D.; Wang, B.; Xing, H. Enhancing Retrieval-Augmented LMs with a Two-Stage Consistency Learning Compressor. In Proceedings of the International Conference on Intelligent Computing. Springer, 2024, pp. 511–522.
3. Cai, Z.; Xiao, W.; Sun, H.; Luo, C.; Zhang, Y.; Wan, K.; Li, Y.; Zhou, Y.; Chang, L.W.; Gu, J.; et al. R-kv: Redundancy-aware kv cache compression for reasoning models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
4. Ma, S.; Liu, S.; Tan, J.; Hu, Y.; Wang, S.; Indurthi, S.R.; Zhao, S.; Wu, L.; Han, J.; Song, K. TCIA: A Task-Centric Instruction Augmentation Method for Instruction Finetuning. *CoRR* **2025**. https://doi.org/10.48550/ARXIV.2508.20374.
5. Rutherford, M.W.; Nolan, T.S.; Pei, L.; Wagner, U.; Pan, Q.; Farmer, P.; Smith, K.E.; Kopchick, B.; Opsahl-Ong, L.; Granger, S.; et al. Medical Image De-Identification Resources: Synthetic DICOM Data and Tools for Validation. *CoRR* **2025**. https://doi.org/10.48550/ARXIV.2508.01889.
6. Zhang, X.; Li, P.; Li, H. AMBERT: A Pre-trained Language Model with Multi-Grained Tokenization. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 421–435. https://doi.org/10.18653/v1/2021.findings-acl.37.
7. Wang, X.; Gui, M.; Jiang, Y.; Jia, Z.; Bach, N.; Wang, T.; Huang, Z.; Tu, K. ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 3176–3189. https://doi.org/10.18653/v1/2022.naacl-main.232.
8. Seo, A.; Kang, G.C.; Park, J.; Zhang, B.T. Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6167–6177. https://doi.org/10.18653/v1/2021.acl-long.481.
9. Wang, H.; Shen, Y. Skeletal Muscle Mass Index and Risk of Lower Extremity Ulcers: Analysis of NHANES Data with External Hospital Validation. *The International Journal of Lower Extremity Wounds* **2025**, p. 15347346251409496.
10. Shen, Y.; Li, X.; Wu, J.; Ma, Y.; Borchmann, S.; Cheng, Z.; Wang, Y.; Zhao, Y.; Song, J.; Luo, B.; et al. Bispecific nanosystems enable multieffector immune cell retargeting for hematologic malignancy therapy. *Advanced Science* **2025**, *12*, e09103.
11. Ainslie, J.; Lee-Thorp, J.; de Jong, M.; Zemlyanskiy, Y.; Lebron, F.; Sanghai, S. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 4895–4901. https://doi.org/10.18653/v1/2023.emnlp-main.298.
12. Pang, S.; Xue, Y.; Yan, Z.; Huang, W.; Feng, J. Dynamic and Multi-Channel Graph Convolutional Networks for Aspect-Based Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2627–2636. https://doi.org/10.18653/v1/2021.findings-acl.232.
13. Qin, H.; Song, Y. Reinforced Cross-modal Alignment for Radiology Report Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 448–458. https://doi.org/10.18653/v1/2022.findings-acl.38.
14. Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; Zhou, G. Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4395–4405. https://doi.org/10.18653/v1/2021.emnlp-main.360.

 doi:10.20944/preprints202603.1988.v1

15. Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; Zhu, L.N. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 4730–4738. https://doi.org/10.18653/v1/2021.findings-acl.417.
16. Xu, C.; Li, J.; Wang, R. Mutual Teaching: Semi-supervised Medical Image Classification with Cross Structural Consistency Learning. In Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2025, pp. 1–6.
17. Malik, V.; Sanjay, R.; Nigam, S.K.; Ghosh, K.; Guha, S.K.; Bhattacharya, A.; Modi, A. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 4046–4062. https://doi.org/10.18653/v1/2021.acl-long.313.
18. Fu, J.; Huang, X.; Liu, P. SpanNER: Named Entity Re-/Recognition as Span Prediction. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 7183–7195. https://doi.org/10.18653/v1/2021.acl-long.558.
19. He, J.; Kryscinski, W.; McCann, B.; Rajani, N.; Xiong, C. CTRLsum: Towards Generic Controllable Text Summarization. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 5879–5915. https://doi.org/10.18653/v1/2022.emnlp-main.396.
20. Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; Ramesh, K.; Jain, P.; Nambi, A.; Ganu, T.; Segal, S.; Ahmed, M.; et al. MEGA: Multilingual Evaluation of Generative AI. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 4232–4267. https://doi.org/10.18653/v1/2023.emnlp-main.258.
21. Li, L.; Zhang, Y.; Chen, L. Personalized Transformer for Explainable Recommendation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 4947–4957. https://doi.org/10.18653/v1/2021.acl-long.383.
22. Wang, P.; Zhu, Z.; Freire, N.; Azar, Z.; Wu, X.; Liang, D. Online Simultaneous Identification of Multi-Parameters for Interior PMSMs Under Sensorless Control. *CES Transactions on Electrical Machines and Systems* **2025**, *9*, 422–433.
23. Wang, P.; Zhu, Z.; Liang, D.; Freire, N.M.; Azar, Z. Dual signal injection-based online parameter estimation of surface-mounted PMSMs under sensorless control. *IEEE Transactions on Industry Applications* **2025**.
24. Wang, P.; Zhu, Z.; Liang, D. Virtual signal injection-based online full-parameter estimation of surface-mounted PMSMs without influence of position error and inverter nonlinearity. *IEEE Journal of Emerging and Selected Topics in Power Electronics* **2025**.