

Multimodal Digital Twin Framework for Personalized CyberKnife Dosimetry and Survival Prediction in Non–Small Cell Lung Cancer

Eman Sabry

Higher Institute of Engineering

Ehab Mohamed

ehab_mahmoud@aswu.edu.eg

Prince Sattam Bin Abdulaziz University

Walid El-Shafai



Prince Sultan University

Research Article

Keywords: Digital twin, CyberKnife, Non–small cell lung cancer, Precision radiotherapy, Ensemble learning, Long short–term memory, Multimodal data integration, Survival prediction, Radiomics and genomics, Clinical decision support

Posted Date: November 28th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-8099168/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Multimodal Digital Twin Framework for Personalized CyberKnife Dosimetry and Survival Prediction in Non–Small Cell Lung Cancer

Eman S. Sabry¹, Ehab Mahmoud Mohamed^{2*}, Walid El-Shafai^{3,4}

¹Department of Communications and Computers Engineering, Higher Institute of Engineering, El-Shorouk Academy, El-Shorouk City, Egypt.

²Department of Electrical Engineering, College of Engineering in Wadi Addawasir, Prince Sattam Bin Abdulaziz University, Wadi Addawasir 11991, Saudi Arabia.

³Automated Systems and Computing Lab (ASCL), Computer Science Department, Prince Sultan University, Riyadh 11586, Saudi Arabia.

⁴Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt.

e.samir@sha.edu.eg, e.mahmoud@psau.edu.sa, welshafai@psu.edu.sa

*Correspondence: e.mahmoud@psau.edu.sa

Abstract: Precision radiotherapy and patient-specific survival prediction in non–small cell lung cancer (NSCLC) demand computational frameworks that are accurate, interpretable, and reproducible. This study introduces a comprehensive digital twin framework for CyberKnife dose prediction and survival estimation, integrating heterogeneous clinical, radiomic, and genomic data through a synergy of classical machine learning, deep learning, and ensemble modeling. The framework begins with a Ridge–Long Short-Term Memory (LSTM) hybrid model that combines the interpretability of linear regression with the temporal–spatial learning capability of recurrent networks. Subsequent enhancements incorporate multilayer perceptrons (MLP) and XGBoost to address structured data constraints and improve generalization through ensemble and bridge stacking architectures. Furthermore, multimodal fusion of radiomic and genomic data via late fusion emphasizes the prognostic value of molecular features while exposing the limitations of naïve averaging strategies. Rigorous preprocessing, multimetric evaluation, and 3D synthetic dose visualizations ensure model transparency, reproducibility, and clinical applicability. Comparative analyses reveal that model performance is modality-dependent: LSTM excels in temporal dependency modeling, while MLP and XGBoost yield superior results for tabular data; ensemble approaches consistently enhance generalization and resilience to outliers. Although phantom-based dosimetry validates the spatial feasibility of dose prediction, future integration with real patient dose–volume histograms (DVHs) is expected to further strengthen clinical translation. Overall, this work establishes a robust, extensible, and clinically aligned digital twin framework that advances precision radiotherapy planning, survival modeling, and multimodal biomedical data integration for NSCLC patient care.

Keywords: Digital twin; CyberKnife; Non–small cell lung cancer; Precision radiotherapy; Ensemble learning; Long short–term memory; Multimodal data integration; Survival prediction; Radiomics and genomics; Clinical decision support.

1. Introduction

The advent of stereotactic body radiotherapy (SBRT), particularly through the CyberKnife system, has markedly advanced the precision of cancer treatment by enabling sub-millimeter accuracy via robotic, image-guided radiation delivery [1, 2]. Despite these technological breakthroughs, existing CyberKnife workflows remain largely

empirical and static, relying heavily on pre-treatment imaging and manual planning. This rigidity limits adaptability during therapy, especially in the presence of intra-fraction motion, tumor regression, or patient-specific anatomical variability, which can compromise dose accuracy and elevate the risk of injury to organs at risk (OARs).

To address these challenges, the concept of a digital twin, a dynamically updated virtual replica of a physical system, has emerged as a transformative paradigm across engineering, aerospace, and, more recently, biomedical sciences [3]. Within radiotherapy, preliminary implementations of digital twin frameworks have been proposed for adaptive planning in prostate SBRT and for predictive modeling of treatment response in high-grade gliomas [4, 5]. These studies highlight the promise of real-time optimization and patient-specific adaptation; however, their translation to CyberKnife-based radiosurgery remains limited and insufficiently explored.

In parallel, major advances in radiomics and radiogenomics offer additional opportunities for predictive enhancement. By extracting high-dimensional quantitative imaging biomarkers and integrating them with molecular and genomic signatures, radiogenomic approaches have demonstrated superior prognostic power, improving survival prediction in non-small-cell lung cancer (NSCLC) and other malignancies [6, 7]. The fusion of these heterogeneous data modalities aligns naturally with the digital-twin paradigm, enabling multimodal, patient-specific treatment personalization that bridges computational modeling and clinical decision-making in radiotherapy.

Realizing this vision, however, requires the development of robust computational frameworks capable of integrating multimodal data streams, ensuring interpretability, supporting real-time clinical adaptation, and maintaining scalability for seamless incorporation into CyberKnife workflows.

1.1 Research Problem and Field Challenges

Although CyberKnife technology has significantly enhanced the precision of radiotherapy, several persistent limitations continue to constrain its full potential. Current treatment planning workflows remain largely static, relying on fixed pre-treatment imaging and manual configuration. As a result, they cannot dynamically adjust to intra-fraction motion, anatomical variations, or tumor regression that occur during the course of treatment, which may compromise dose accuracy and therapeutic efficacy [1, 2]. Moreover, existing CyberKnife systems primarily depend on imaging data while underutilizing genomic and molecular information that could enrich survival prediction and enable personalized treatment strategies [6, 7].

Another critical limitation lies in the interpretability of machine learning (ML) models. Many current approaches function as opaque “black boxes,” providing little transparency regarding the basis of their predictions. This lack of explainability limits clinical trust, hinders reproducibility, and poses challenges for meeting regulatory and ethical standards in medical AI [8]. Furthermore, CyberKnife systems have yet to incorporate predictive analytics or digital twin technologies capable of adapting dose delivery or robotic motion in real time. Without such adaptive feedback, treatment remains reactive rather than proactive, reducing the system’s ability to respond to evolving tumor and patient dynamics [3–5].

While numerous academic models demonstrate encouraging results, most have not been designed with modularity, transparency, or seamless integration into existing clinical workflows in mind. Consequently, their translation from research environments to routine clinical practice remains limited [6, 8].

Beyond these methodological gaps, broader technical and organizational challenges further impede clinical adoption. Radiotherapy datasets are often small, incomplete, and heterogeneous across institutions, making it difficult to train and validate generalizable predictive models [4, 6]. Moreover, compliance with rigorous safety and regulatory standards necessitates transparent, verifiable, and auditable computational workflows [8]. Real-time motion tracking and adaptive dose planning impose additional computational demands, requiring highly optimized algorithms supported by hardware accelerators such as GPUs or FPGAs, which remain challenging to implement in hospital environments [2, 3]. Effective development of digital twin systems also depends on close interdisciplinary collaboration among engineers, radiation oncologists, medical physicists, genomic scientists, and software developers [5]. Finally, clinical acceptance hinges on establishing trust through model transparency, validation using real patient data, and demonstrated improvements in workflow efficiency and treatment outcomes [6, 8].

To address these challenges, this study proposes a unified digital twin framework that integrates multimodal data sources, leverages predictive intelligence, ensures interpretability, and supports real-time adaptive decision-making within CyberKnife radiotherapy.

1.2 Research Motivation

The growing complexity of modern radiotherapy underscores the urgent need for intelligent systems capable of dynamically adapting to evolving clinical conditions. Although the CyberKnife platform achieves remarkable precision through robotic and image-guided radiation delivery, its operational design remains inherently static. This limitation prevents it from responding effectively to intra-treatment variations in tumor morphology or patient anatomy. Consequently, therapeutic adjustments often rely on empirical clinical judgment rather than data-driven inference, which restricts both consistency and reproducibility across treatment sessions.

Recent advances in digital twin technology offer a promising foundation to address these limitations. By constructing a continuously updated virtual counterpart of the patient and treatment environment, a digital twin enables real-time simulation, prediction, and optimization of therapeutic outcomes. When integrated with radiomic, genomic, and clinical data, such a framework can facilitate personalized adaptation, enhance outcome prediction, and improve overall treatment precision.

The motivation for this research arises from a distinct and underexplored gap. Although digital twin models have been applied in several biomedical contexts, their deployment in CyberKnife-based stereotactic radiotherapy remains limited. Developing an integrated digital twin framework specifically for CyberKnife systems would represent a substantial step toward intelligent, adaptive radiotherapy, bridging computational intelligence with clinical practice and advancing the paradigm of personalized cancer treatment.

1.3 Research Contribution

This study makes several significant contributions that collectively advance the integration of artificial intelligence and digital twin methodologies within CyberKnife treatment planning and prediction:

- A comprehensive digital twin framework is developed for CyberKnife radiotherapy, integrating clinical, radiomic, and genomic data to enable patient-specific survival prediction and adaptive treatment optimization.

- A hybrid Ridge–LSTM model is introduced, combining the interpretability of Ridge regression with the temporal learning capability of Long Short-Term Memory networks to enhance predictive accuracy and model transparency.
- The framework is further extended through the incorporation of multilayer perceptron (MLP) and XGBoost models, addressing challenges associated with tabular data and improving robustness via ensemble and bridge stacking strategies.
- A multimodal late fusion approach is implemented to demonstrate the synergistic contribution of radiomic, genomic, and clinical features toward predictive performance.
- Multi-metric evaluations and 3D synthetic dose visualizations are provided to ensure reproducibility, interpretability, and clinical relevance of the proposed system.
- The framework is validated through phantom-based dosimetry experiments, confirming the feasibility of spatial dose prediction and establishing the foundation for integration with patient-specific dose–volume histograms (DVHs).

Collectively, these contributions define a robust, extensible, and clinically aligned digital twin framework that supports intelligent decision-making in CyberKnife radiotherapy and lays the groundwork for future clinical translation.

1.4 Research Objectives

The primary goal of this research is to design and evaluate a digital twin framework capable of predicting and optimizing CyberKnife treatment outcomes for patients with NSCLC. In pursuit of this goal, the study focuses on several specific objectives that collectively ensure the methodological, computational, and clinical robustness of the proposed system.

- Develop a computationally efficient and interpretable hybrid model that integrates classical and deep learning approaches to enhance the accuracy and transparency of survival prediction.
- Implement multimodal data integration by combining clinical, radiomic, and genomic features to improve predictive precision and support personalized radiotherapy planning.
- Evaluate the proposed framework using a comprehensive set of performance metrics, including MSE, RMSE, MAE, MedianAE, R^2 , and Max Error, to ensure robustness, reliability, and reproducibility.
- Generate 3D synthetic dose visualizations to facilitate interpretability, enable clinical validation, and improve the understanding of spatial dose distribution.
- Assess the feasibility of spatial dose prediction through phantom-based experiments and establish pathways for future validation using patient-specific DVHs.

Collectively, these objectives aim to advance the field of precision radiotherapy by developing a scalable, interpretable, and clinically viable digital twin framework tailored to CyberKnife-based treatment of NSCLC.

1.5 Paper Structure

The remainder of this paper is organized as follows. Section 2 outlines the translational roadmap for integrating the proposed multimodal digital twin framework into next-generation CyberKnife systems, defining the engineering and clinical implementation pathway that bridges computational modeling with practical radiotherapy

deployment. Section 3 presents a comprehensive review of related work encompassing motion compensation, dose prediction, treatment outcome modeling, and digital-twin development, highlighting both current advances and persisting research gaps. Section 4 describes the proposed system architecture, including the computational framework, data-processing pipelines, and visualization components that support the overall design. Section 5 details the methodological foundation, covering model architectures, training procedures, and workflow integration strategies. Section 6 explains the experimental setup and evaluation design, defining the performance metrics that combine standard regression indicators with radiotherapy-specific dose–outcome criteria. Section 7 introduces the datasets employed in this study, with emphasis on the NSCLC-Radiomics collection as the principal benchmarking source. Section 8 presents the experimental results and discussion, focusing on predictive accuracy, interpretability, and clinical relevance. Section 9 provides the practical validation and clinical implications of the proposed system through representative test cases. Finally, Section 10 concludes the paper by summarizing the key findings and outlining future directions toward a fully deployable CyberKnife digital-twin framework for precision radiotherapy.

2. Toward a Translational Roadmap for CyberKnife Digital Twin Systems

Building upon the conceptual framework presented in the introduction section, this part delineates a translational roadmap for integrating digital twin intelligence into next-generation CyberKnife radiosurgery systems. The proposed roadmap bridges computational modeling, engineering design, and clinical deployment, ensuring that the methodological advances developed in this study evolve toward tangible, real-world impact. Rather than reiterating the technical contributions, this section emphasizes how the predictive, multimodal, and interpretable modeling strategies outlined earlier can be operationalized within a 5–10 year engineering and clinical development trajectory. The overarching goal is to enable CyberKnife systems that are self-adaptive, data-driven, and seamlessly integrated with hospital infrastructure, advancing from current static planning paradigms to intelligent, continuously learning radiotherapy systems.

2.1 Translational Vision and Rationale

The integration of digital twin frameworks into CyberKnife systems aligns with a broader paradigm shift in radiotherapy engineering, one that moves from deterministic planning toward dynamic, feedback-driven treatment adaptation. As radiotherapy precision approaches its physical limits, further gains in efficacy and safety increasingly depend on computational intelligence, multimodal data integration, and real-time decision support. Digital twins, by continuously synchronizing virtual models with patient-specific clinical and imaging data, can enable predictive control and adaptive planning that respond instantly to anatomical and biological variations.

However, realizing such systems requires a structured translational roadmap that systematically connects algorithmic innovation to engineering feasibility, regulatory compliance, and clinical acceptance. This roadmap must ensure that advances in ML, ensemble modeling, and multimodal analytics evolve through clearly defined stages, from laboratory validation to large-scale clinical implementation, while maintaining transparency, interoperability, and reproducibility. The following subsections outline this translational progression through sequential engineering and clinical phases.

2.2 Phased Engineering and Clinical Roadmap

The roadmap for CyberKnife digital twin integration is organized into five progressive phases, spanning short-, medium-, and long-term objectives. Each phase builds upon the previous one, establishing the technological, regulatory, and clinical foundations required for safe and effective deployment.

A. Phase 0 (0–6 months): Preparatory and Regulatory Foundations

The preparatory phase focuses on establishing the structural and governance prerequisites for translational research. This includes forming interdisciplinary consortia involving oncologists, engineers, data scientists, and regulatory experts. Baseline audits of existing CyberKnife and LINAC systems are conducted to evaluate data availability, workflow bottlenecks, and hardware–software interoperability. Concurrently, ethical and legal frameworks are developed for data governance, intellectual property management, and patient privacy. Institutional Review Board (IRB) approvals and regulatory pathway assessments are initiated. The deliverables of this phase comprise signed project charters, secured clinical dataset access, and a transparent governance model that enables subsequent phases to proceed under compliant, traceable conditions.

B. Phase 1 (1–2 years): Core Technology Development

This phase centers on the development and testing of the core algorithmic and computational components underpinning the digital twin system. Key milestones include real-time tumor motion compensation, GPU-accelerated dose computation, and the automation of treatment planning through ML-based workflow orchestration. Models such as Ridge regression, LSTM networks, and hybrid Ridge–LSTM frameworks are optimized for speed, interpretability, and precision. Predictive maintenance modules are incorporated to monitor system health, while phantom-based experiments validate sub-millimeter tracking accuracy. Performance targets include reducing planning times below 30 minutes and achieving consistent reproducibility under clinical data variability. The outcomes of this phase establish the computational feasibility and engineering readiness of AI-augmented CyberKnife operations.

C. Phase 2 (2–4 years): Clinical Prototyping and Adaptive Integration

Phase 2 transitions from algorithmic innovation to prototype development and preclinical validation. Here, adaptive planning and real-time motion compensation are integrated into functional CyberKnife prototypes equipped with digital twin modules. Ensemble models, particularly bridge stacking and late-fusion architectures, enable simultaneous processing of radiomic, genomic, and clinical data streams. Evaluation frameworks are expanded to include regulatory and safety metrics, ensuring that predictive models meet medical device quality assurance (QA) standards. This phase emphasizes iterative design: phantom-based testing is followed by pilot clinical feasibility studies, with early regulatory consultations to preemptively address compliance requirements. Success at this stage is defined by demonstrable improvements in treatment precision, robustness, and computational efficiency.

D. Phase 3 (4–6 years): Scaling, Optimization, and Workforce Integration

Once prototype systems have achieved regulatory clearance for early clinical testing, attention shifts toward scalability and operational efficiency. Hardware and software modularization become key objectives to support integration across diverse hospital environments. Optimizations in GPU/FPGA utilization, data throughput, and user interface design enable reductions in cost per treatment and planning time. Virtual and augmented reality (VR/AR) platforms are introduced for clinical training, allowing oncologists to interact with real-time digital twin visualizations during treatment planning. This phase also explores industrial partnerships for manufacturing and system certification. Performance indicators include cost efficiency, workflow standardization, and clinician adoption rate. Collectively, these efforts ensure that the system can be scaled without compromising reliability or regulatory compliance.

E. Phase 4 (6–10 years): Multi-Center Trials, Regulatory Approval, and Continuous Learning

The final phase targets full-scale clinical validation through multi-center trials and long-term regulatory approval. Digital twin systems are deployed in diverse clinical settings to evaluate generalizability and robustness under heterogeneous patient populations. Continuous learning loops are implemented, allowing the twin to update predictive parameters using real-world outcome data while preserving patient privacy through federated learning architectures. At this stage, interpretability and traceability become paramount to meet regulatory expectations and maintain clinician trust. Comprehensive safety validation, audit logs, and explainable AI dashboards ensure accountability throughout the system's lifecycle. This phase culminates in commercial deployment and ongoing improvement, where real-world feedback drives iterative refinements and system updates, thereby transforming CyberKnife radiosurgery into a continuously learning, adaptive therapeutic platform.

2.3 Mapping Predictive Modeling Insights to Roadmap Objectives

The methodological advances developed in this study provide foundational components for each phase of the proposed roadmap. These research insights, spanning data preprocessing, model selection, multimodal integration, and visualization, directly inform engineering and clinical milestones.

i) Real-World Clinical Modeling and Predictive Insights

The study's experimental results across multiple modeling cases demonstrate clear pathways for translational integration:

- Hybrid Ridge–LSTM modeling (Case 1): Combining linear interpretability with temporal pattern learning enables the system to model sequential dependencies in dose and survival prediction. This directly supports Phases 1–2, where real-time motion compensation and adaptive dose planning require dynamic temporal prediction.
- Bridge-stacking ensembles (Cases 2–4): Integrating Ridge, MLP, and XGBoost models reduces bias and variance, enhancing predictive robustness across heterogeneous data. These architectures provide methodological grounding for AI copilots and automated planning modules in Phases 1–3, where stability under varying data distributions is essential for clinical reliability.
- Radiomics–Genomics late fusion (Case 5): The late-fusion strategy demonstrates that integrating heterogeneous data sources significantly improves prognostic accuracy. This informs Phases 2–4, where future CyberKnife systems are expected to fuse multi-sensor, imaging, and possibly genomic data to personalize treatment strategies.

ii) Key Technical Enablers

Several technical and methodological enablers derived from the current research underpin the roadmap's progression:

- Automated Preprocessing and Feature Engineering: The study's imputation, scaling, and one-hot encoding pipelines ensure reproducible and high-quality input data, foundational for real-time adaptation in Phase 1.
- Model Evaluation and Visualization: Quantitative metrics (MSE, RMSE, MAE, R^2 , Max Error) and 3D synthetic dose visualizations contribute directly to regulatory documentation, QA verification, and clinical interpretability during Phases 2–4.

- **Multimodal Integration Frameworks:** The demonstrated capacity to integrate radiomics and genomics highlights the system’s potential to extend toward multi-sensor and physiological data streams, a key objective of Phases 3–4.

iii) Roadmap Integration Summary

The correspondence between predictive modeling insights and roadmap objectives is summarized in Table 1, which maps key research outcomes to development phases, models, and anticipated clinical impact.

Table 1. Mapping predictive modeling and digital twin insights to the cyberknife engineering roadmap.

| Roadmap Phase | Key Goals & Performance Indicators | Supporting Research Case(s) | Models / Methods | Contribution to Phase |
|----------------------|---|-----------------------------|---|--|
| Phase 0 (0–6 months) | Preparatory work, data governance, IRB approval, dataset access | Case 1, 2 | Data preprocessing, feature engineering pipelines | Establishes automated, reproducible data handling and regulatory-ready datasets for ML/DL models. |
| Phase 1 (1–2 years) | Real-time motion compensation, GPU-accelerated planning, AI copilots | Case 1–3 | Ridge, LSTM, MLP, XGBoost, bridge ensemble | Enables sub-millimeter tumor tracking and rapid dose computation (< 30 min); validates adaptive ML planning. |
| Phase 2 (2–4 years) | Adaptive prototype integration, preclinical QA, regulatory engagement | Case 3–5 | Bridge ensembles, multimodal late fusion | Supports adaptive treatment prototypes; integrates temporal–spatial dose prediction and synthetic 3D phantoms. |
| Phase 3 (4–6 years) | Scale-up, modular design, throughput optimization, clinician training | Case 2–4 | Lightweight Ridge, XGBoost, ensembles | Improves computational efficiency, supports automation, and informs modular system architecture. |
| Phase 4 (6–10 years) | Multi-center validation, regulatory approval, continuous learning | Case 4–5 | LSTM, bridge ensemble, multimodal Ridge models | Provides validated predictive frameworks for large-scale deployment and real-world data adaptation. |

2.4 Strategic and Clinical Implications

The roadmap described above provides a structured and evidence-based pathway for translating digital twin research into deployable CyberKnife systems. Each phase establishes cumulative value, advancing from computational readiness to full clinical integration, while maintaining compliance with safety, ethical, and regulatory standards. The systematic mapping between research findings and engineering objectives ensures that every methodological innovation contributes to tangible clinical outcomes such as reduced planning time, enhanced dose accuracy, and improved patient survival prediction.

Clinically, this roadmap positions the digital twin as both a decision-support system and a predictive modeling companion to the CyberKnife platform. The inclusion of explainable ML ensures transparency in clinical decision-making, while multimodal data fusion provides comprehensive insight into patient-specific response variability. From an engineering standpoint, modular design and open interfaces allow integration into existing hospital infrastructures, facilitating interoperability with picture archiving and communication systems (PACS), treatment planning systems (TPS), and electronic health records (EHRs).

Finally, by embedding continuous learning and iterative validation into its lifecycle, the proposed digital twin system fosters a new generation of adaptive radiotherapy, one that is data-driven, self-optimizing, and capable of improving through real-world experience. Over the coming decade, the implementation of this roadmap can redefine

CyberKnife radiosurgery as an intelligent, autonomous, and patient-centered modality, marking a decisive step toward precision oncology supported by digital twin technology.

3. Related Work

This section reviews and synthesizes prior research pertinent to the development of a CyberKnife-oriented digital twin framework for adaptive radiotherapy. It surveys foundational work across seven interrelated domains that collectively underpin this study: robotic radiosurgery and CyberKnife foundations, artificial intelligence for dose prediction and planning, radiotherapy outcome modeling, temporal motion compensation, ensemble learning on structured clinical data, multimodal fusion of radiomic and genomic information, and emerging digital twin applications in radiation oncology. Each subsection analyzes representative methodologies, identifies persistent technical and translational limitations, and highlights how the present work extends existing knowledge to address unresolved challenges in achieving real-time, uncertainty-aware, and clinically integrated adaptive treatment.

3.1 Literature Review

This subsection situates the proposed CyberKnife-oriented digital twin framework within seven major strands of the literature: (i) robotic SRS/SBRT and CyberKnife foundations, (ii) artificial intelligence for dose prediction, planning, and quality assurance (QA), (iii) outcome prediction models in radiotherapy, (iv) temporal modeling and motion compensation, (v) ensemble learning for structured clinical data, (vi) multimodal fusion of radiomic–genomic features with visualization and QA, and (vii) digital twin applications in radiation oncology. For each strand, representative methodologies are discussed, prevailing challenges identified, and the ways in which the present work extends current knowledge are highlighted.

(i) Robotic SRS/SBRT and CyberKnife Foundations

Robotic radiosurgery platforms such as the CyberKnife have achieved sub-millimeter targeting accuracy through respiratory motion synchronization and non-coplanar beam delivery, facilitating hypofractionated treatment of thoracic and other anatomically complex regions while effectively sparing organs at risk [9-11]. Peer-reviewed evaluations of tumor tracking and Synchrony-type respiratory compensation confirm clinically acceptable precision but also reveal vulnerabilities to baseline drift, irregular breathing patterns, and surrogate mismatches [10, 11]. Despite these technological advances, current planning remains computationally intensive, highly dependent on the planner’s expertise, and limited by hand-crafted motion models that cannot dynamically adapt to changing patient anatomy or respiratory variability. These limitations emphasize the need for predictive, data-driven frameworks capable of learning motion dynamics, quantifying uncertainty, and supporting adaptive replanning, capabilities that the proposed digital twin system seeks to operationalize.

(ii) AI for Radiotherapy Planning, Dose Prediction, and QA

Artificial intelligence and deep learning (DL) methods have been widely applied in dose prediction and knowledge-based planning (KBP), using U-Net–style architectures and geometry-aware encoders to map anatomical and prescription data to three-dimensional dose distributions or DVHs [12-16]. These models substantially reduce planning time while maintaining or improving plan quality. Recent innovations include uncertainty-aware learning [13, 15], hybrid pipelines that combine machine learning surrogates with physics-based verification [16], and geometry-aware networks that encode spatial relationships between target volumes and organs at risk [15]. However,

DL-based dose models often fail to detect clinically critical hotspots and may perform inconsistently under domain shifts arising from variations in scanners or acquisition protocols [12, 14]. Furthermore, their optimization typically centers on mean error metrics, overlooking worst-case deviations essential for patient safety. The lack of structured, explainable, and auditable outputs further limits their clinical and regulatory integration. The present study addresses these issues by embedding explainable, uncertainty-aware predictive modeling within a digital twin architecture that aligns with QA and governance standards.

(iii) Outcome Prediction Models in Radiotherapy

Outcome modeling in radiotherapy has evolved from traditional statistical approaches, such as penalized regression and Cox proportional hazards models [17], which offer interpretability but fail to capture nonlinear relationships in high-dimensional data, to advanced machine learning and deep learning models. Tree-based ensembles (e.g., XGBoost) [18, 19] and neural networks effectively model nonlinear interactions but are prone to overfitting when applied to small or heterogeneous clinical datasets. Voxel-wise convolutional neural networks (CNNs) have shown promise for dose–response prediction but demand large, paired imaging–outcome datasets and often lack robust external validation [20]. Moreover, many studies rely on internal data splits without proper uncertainty quantification, limiting their generalizability [21]. Addressing these challenges requires hybrid architectures that balance interpretability with predictive power and integrate uncertainty estimation, an approach implemented in this study through hybrid Ridge–LSTM and ensemble designs evaluated on multi-source data.

(iv) Temporal Modeling and Motion Compensation

Respiratory motion remains a significant source of uncertainty in radiosurgery, directly affecting dose accuracy and target conformity. Sequential models such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and one-dimensional convolutional networks (1D CNNs) have been employed for respiratory signal forecasting [22–24], motion gating [23], and adaptive planning [22]. Among these, LSTM-based architectures outperform linear models under irregular breathing conditions and reduce latency in gating workflows [22, 23]. Nevertheless, these methods frequently overfit small temporal sequences, degrade under domain shifts (e.g., coughing or baseline drift), and rarely include calibrated uncertainty estimation. Furthermore, motion prediction models are seldom integrated with treatment outcome prediction or QA pipelines. This study bridges this gap by introducing motion-aware, uncertainty-calibrated learning modules within the digital twin, facilitating joint modeling of spatiotemporal dynamics and clinical outcomes.

(v) Learning on Tabular Clinical Data and Ensemble Modeling

Structured clinical datasets encompassing demographics, staging, histopathology, and laboratory results remain a critical but underutilized component of predictive radiotherapy analytics. Gradient-boosted decision trees [18, 19, 25] and regularized linear models [17] continue to serve as robust baselines, while ensemble strategies such as stacking and bridge modeling [26, 28] enhance predictive stability and mitigate overfitting. Recent developments in model explainability, including SHAP-based feature attribution [27], have improved interpretability and trust in such models. However, most studies limit evaluation to cross-validation performance, neglecting practical deployment requirements such as structured prediction logging, model documentation (“model cards”), and real-time inference feasibility. The current work advances this area by incorporating ensemble strategies that merge tabular, radiomic,

and genomic modalities within an auditable, reproducible computational framework suitable for clinical engineering deployment.

(vi) Multimodal Fusion (Radiomics–Genomics) and Visualization/QA

Integrating radiomic and genomic data has become increasingly vital for improving risk stratification and survival prediction in oncology [29-33]. Late-fusion methods, which combine modality-specific predictors through averaging or stacking, offer robustness when data availability varies, whereas early and intermediate fusion approaches capture cross-modal feature interactions [31]. Studies consistently demonstrate that genomic features provide strong prognostic signals, with radiomics adding complementary information related to local tumor control and toxicity [29, 32]. Despite these advances, visualization in radiogenomic research typically relies on static two-dimensional plots, feature importance charts [34], or basic dose–volume histograms [35], with minimal emphasis on 3D or QA-integrated representations. Moreover, naïve fusion strategies may dilute strong unimodal signals, and batch effects or inconsistent preprocessing can impair generalization. In contrast, the present study employs a late-fusion approach embedded within a digital twin architecture, complemented by advanced 3D visualization and phantom-based QA, bridging predictive modeling with clinically interpretable visual analytics.

(vii) Digital Twins in Radiation Oncology

The concept of the digital twin, a continuously updated, biophysically informed virtual replica of a physical system, has recently been introduced in radiation oncology to support adaptive planning, treatment verification, and predictive modeling [36-39]. Early implementations combine machine learning surrogates for dose prediction or motion forecasting with physics-based simulation engines to facilitate real-time adaptation and provide auditable QA evidence trails [36, 37]. Although these prototypes demonstrate feasibility, most remain single-institution studies lacking scalability, reproducibility, and integration with clinical governance or regulatory frameworks. Furthermore, existing approaches often focus narrowly on a single task, such as dose reconstruction or organ motion tracking, without unifying multimodal prediction, uncertainty quantification, and visualization in a single pipeline. The framework proposed in this study advances beyond these limitations by delivering an end-to-end, data-driven, and clinically aligned digital twin architecture tailored for CyberKnife-based adaptive radiotherapy.

Collectively, prior research across these seven strands reveals substantial progress in radiotherapy automation, predictive modeling, and multimodal data integration. However, the field remains fragmented, with limited cross-linkage among motion compensation, dose prediction, and outcome modeling streams. Few frameworks provide real-time interpretability, reproducibility, and governance compliance required for clinical translation. The present study addresses these gaps by introducing a unified digital twin framework that integrates radiomic, genomic, and clinical data for survival prediction and adaptive planning, supported by explainable ensemble learning, multimodal fusion, uncertainty quantification, and 3D QA visualization. This holistic approach advances the state of CyberKnife-based radiotherapy toward a data-driven, clinically interpretable, and governance-ready paradigm.

3.2 Comparative Analysis

To contextualize the proposed CyberKnife-oriented digital twin framework, this subsection systematically compares prior research across key thematic domains: robotic radiosurgery, dose prediction and planning, outcome modeling, temporal forecasting, structured clinical learning, multimodal fusion, and digital twin development. Table

2 synthesizes these contributions by summarizing the core clinical and computational problems addressed in the literature, the corresponding methodological approaches, their persistent limitations, and the specific ways in which the proposed study advances beyond the current state of the art. Thus, the table outlines key problem domains, representative prior solutions, their limitations, and the innovations introduced by the present study. While prior efforts achieved partial advances in areas such as dose prediction, motion modeling, or outcome forecasting, none provided a unified, uncertainty-aware, reproducible, and clinically deployable digital twin system for CyberKnife radiotherapy.

Although substantial progress has been achieved in artificial intelligence–driven radiotherapy, the field remains fragmented. Motion compensation frameworks often lack robust uncertainty calibration, limiting their reliability under irregular respiratory patterns. Dose prediction methods have improved computational efficiency but frequently fail to identify local hotspots or provide traceable, audit-ready outputs. Outcome prediction models exhibit promising accuracy yet suffer from limited generalization across institutions and insufficient interpretability. Similarly, multimodal fusion strategies that integrate imaging and genomic data frequently dilute clinically relevant signals, while visualization pipelines rarely translate model outputs into actionable, QA-integrated insights. Even current digital twin initiatives, while conceptually transformative, remain largely at the prototype or single-institution stage, lacking reproducible, end-to-end clinical integration.

The proposed work directly addresses these shortcomings through four major innovations:

- Predictive motion modeling: Embedding temporal learning (LSTM-based forecasting) within CyberKnife motion compensation to mitigate surrogate mismatch and improve real-time targeting accuracy.
- Uncertainty-aware dose prediction: Developing an ensemble dose predictor (Ridge, XGBoost, LSTM) with explicit uncertainty quantification and the generation of structured, machine-readable QA artifacts for auditability and reproducibility.
- Multimodal outcome integration: Combining tabular clinical, radiomic, and genomic features within a unified multi-stream fusion pipeline to enhance survival prediction and treatment personalization.
- Lifecycle digital twin realization: Delivering the first clinically actionable CyberKnife digital twin that unifies dose modeling, motion forecasting, and outcome prediction into a single adaptive framework with integrated visualization and governance readiness.

Across existing literature, several systemic limitations persist:

- Fragmented, siloed workflows that address isolated sub-tasks rather than full treatment lifecycles.
- Data constraints, including small cohort sizes and limited external or multi-institutional validation.
- Weak or absent uncertainty quantification and inadequate performance calibration.
- Minimal integration of visualization and QA mechanisms, reducing interpretability and regulatory compliance potential.

The proposed framework overcomes these barriers by:

- Unifying linear/tabular (Ridge, XGBoost), temporal (LSTM), and ensemble models into a coherent, interpretable predictive pipeline.
- Exporting standardized structured artifacts (CSV, JSON, JSONL) to ensure transparency, reproducibility, and compliance with clinical governance standards.

- Embedding visualization modules, including dose–volume histograms and 3D phantom renderings, to bridge predictive modeling with clinical QA workflows.
- Advancing toward the first end-to-end, uncertainty-aware digital twin for CyberKnife-based radiosurgery, bridging methodological innovation with deployable clinical infrastructure to enhance precision, accountability, and clinical trust.

Table 2. Comparison of prior work in CyberKnife radiosurgery and related AI/ML applications with the proposed digital twin framework.

| Strand | Prior Problem | Prior Solutions | Limitations | Proposed Work (Novelty) |
|--------------------------------|---|---|---|--|
| CyberKnife and Motion [9–11] | Respiratory motion leading to targeting errors and manual adaptive planning | Robotic beam delivery and Synchrony-based tracking | Surrogate mismatch; lack of predictive adaptation | LSTM-driven predictive motion model integrated into CyberKnife digital twin |
| Dose Prediction and QA [12–16] | Manual planning slow and variable in quality | U-Net dose prediction, KBP, reinforcement learning | Poor generalization, missed hotspots, no QA artifacts | Ensemble (Ridge + XGBoost + LSTM) with uncertainty-aware dose prediction and structured QA logs |
| Outcome Prediction [17–21] | Nonlinear dose–response relationships difficult to model | Cox, XGBoost, CNN-based outcome modeling | Overfitting, poor external validation, no uncertainty calibration | Ensemble with SHAP interpretability and motion-aware, uncertainty-calibrated outcome prediction |
| Motion Forecasting [22–24] | Respiratory variability degrading targeting precision | RNN/LSTM respiratory forecasts | Overfit small datasets, uncalibrated uncertainty, isolated task scope | Integrated temporal predictor with uncertainty modeling embedded within the digital twin |
| Clinical Data Models [25–29] | Underutilization of structured clinical/tabular data | Boosted trees, stacking ensembles | Limited reproducibility, absence of engineering validation | Standardized artifact export (JSON/CSV) with verified real-time feasibility |
| Multimodal Fusion [30–39] | Inefficient fusion of imaging and genomic data | Late and intermediate fusion architectures | Signal dilution, poor visualization, and interpretability | Multi-stream fusion (radiomics + genomics + motion + clinical) with QA visualization via DVH and 3D phantom renderings |
| Digital Twins [36–39] | Conceptual or pilot-scale implementations | ML–physics surrogates, adaptive replanning prototypes | No reproducible, end-to-end clinical integration | First unified CyberKnife digital twin incorporating dose, motion, outcome, QA, and visualization lifecycle |

3.3 Research Gap

The review of existing literature and comparative analysis reveal that while substantial progress has been achieved in robotic radiosurgery and AI-driven radiotherapy, current systems remain limited in their ability to support real-time, uncertainty-aware, and clinically integrated adaptive treatment. Research in CyberKnife workflows has largely focused on isolated subcomponents, such as motion compensation, dose prediction, or outcome modeling, without establishing a unified computational ecosystem that can dynamically link these processes throughout the treatment lifecycle.

Most existing studies address specific challenges independently: motion forecasting models improve short-term tracking accuracy but lack calibrated uncertainty quantification; dose prediction models expedite planning but fail to provide verifiable, auditable artifacts; outcome prediction frameworks enhance stratification yet remain constrained by small sample sizes, poor generalizability, and limited interpretability. Radiogenomic fusion studies have demonstrated prognostic value but seldom achieve actionable integration with real-time treatment optimization

or QA systems. Furthermore, the digital twin initiatives proposed thus far remain largely conceptual or pilot-stage implementations that lack reproducibility, standardized data exchange formats, and alignment with clinical governance protocols.

These limitations reveal a critical methodological and translational gap. What remains absent is an end-to-end, interoperable digital twin that consolidates predictive modeling, multimodal data fusion, temporal motion forecasting, and radiogenomic analytics within a single adaptive and auditable framework. Such a system must operate not merely as a predictive surrogate but as a continuously evolving computational twin that mirrors the patient's physiological and treatment dynamics in real time. It should integrate explainable artificial intelligence (XAI), structured data exports for compliance (e.g., JSON, CSV), and embedded visualization tools that connect predictive inference with QA-ready clinical interfaces.

Addressing this research gap requires a framework that unifies methodological rigor with clinical practicality, an architecture capable of learning from heterogeneous data streams (imaging, genomic, and tabular), adapting to temporal uncertainties, and supporting decision transparency through interpretable and verifiable outputs. The proposed study advances this frontier by introducing a CyberKnife-oriented digital twin architecture that merges deep learning-based dose and motion modeling with ensemble clinical and genomic predictors, enhanced by real-time uncertainty calibration and multimodal visualization. By bridging computational innovation with clinical translation, this framework establishes a foundational step toward the realization of intelligent, accountable, and patient-specific adaptive radiotherapy.

4. Implemented Technologies and System Integration

Building upon the gaps identified in the related work, this section presents the technological implementation of the proposed CyberKnife-oriented digital twin framework. The framework integrates advanced machine learning architectures, multimodal data pipelines, and clinically interpretable visualization modules within a unified system that emphasizes modularity, reproducibility, and governance readiness. Its design is structured around five core components: predictive modeling, multimodal data integration, uncertainty quantification, visualization and interface development, and system-level deployment. Collectively, these components form a clinically aligned computational ecosystem that bridges predictive intelligence and real-time CyberKnife operations.

A. Predictive Modeling Framework

The predictive core of the digital twin is designed as an ensemble architecture that leverages complementary modeling paradigms to capture the diverse statistical and temporal dependencies inherent in radiotherapy data. Linear and tree-based models, including Ridge regression and XGBoost, are employed to process structured clinical and dosimetric variables, providing interpretable baseline predictions while maintaining robustness against overfitting [18]. These models effectively characterize feature interactions and population-level trends critical for dose-response and toxicity estimation.

To model temporal dynamics, LSTM networks are integrated for forecasting respiratory motion and baseline drift, expanding upon their established success in capturing nonlinear temporal dependencies in radiotherapy and medical signal prediction [40]. By learning long-range dependencies within respiratory time series, the LSTM component enhances tracking precision during beam delivery and mitigates latency in adaptive planning.

The outputs from linear, tree-based, and temporal models are fused through an uncertainty-aware ensembling strategy that employs weighted averaging guided by model confidence scores. Confidence intervals are estimated from ensemble variance and Monte Carlo dropout sampling, providing a principled quantification of epistemic and aleatoric uncertainty [41]. This hybrid approach ensures stability across heterogeneous data distributions while enhancing interpretability and reliability in clinical deployment.

B. Multimodal Data Integration

To achieve patient-specific personalization, the digital twin integrates multiple heterogeneous data streams encompassing demographic and clinical metadata, imaging-derived radiomic features, genomic biomarkers, and respiratory motion trajectories. These diverse modalities are harmonized through schema-based integration and feature normalization pipelines that preserve semantic consistency across institutions and imaging protocols [42, 43]. The resulting composite feature space allows for the simultaneous modeling of anatomical, physiological, and molecular dimensions of treatment response.

Feature harmonization follows a standardized ontology-driven schema that facilitates traceable metadata documentation and supports future data sharing under FAIR (Findable, Accessible, Interoperable, and Reusable) principles. This modular data ingestion pipeline is essential for reproducibility and scalability, enabling cross-center model retraining and continuous performance auditing.

C. Uncertainty Quantification and QA Artifacts

Recognizing the importance of transparency in clinical AI systems, the proposed framework embeds uncertainty quantification at every stage of prediction. Each model output, whether dose distribution, motion forecast, or survival estimate, is accompanied by calibrated uncertainty metrics. These metrics are derived from ensemble variance and stochastic dropout mechanisms, providing clinicians with an interpretable confidence level associated with each prediction.

All results are exported in structured, machine-readable formats such as CSV, JSON, and JSONL to ensure transparent recordkeeping and reproducible pipelines. This standardized output facilitates traceable data lineage, regulatory auditing, and interoperability with EHR and QA systems. The incorporation of structured artifacts aligns with recent recommendations for reproducible, auditable, and accountable digital twin systems in biomedical engineering [42, 43].

D. Visualization and Clinical Interfaces

To bridge predictive analytics with clinical decision-making, the framework incorporates a suite of visualization modules that enhance interpretability and facilitate QA integration. The first component includes dynamic dose–volume histogram overlays that compare predicted versus planned dose distributions, allowing clinicians to directly assess deviations in critical regions. In addition, three-dimensional phantom renderings are employed to visualize motion-compensated targeting accuracy and dose deposition, extending recent advances in radiotherapy QA visualization [44, 45].

The framework also features an outcome dashboard that presents predicted survival probabilities and toxicity risks with associated uncertainty bands. This interactive visualization enables clinicians to explore the interplay between dose, motion, and outcome predictions in real time, thus supporting data-driven adjustments to treatment

strategies. Collectively, these visualization tools ensure that predictive outputs are not only accurate but also clinically interpretable and actionable.

E. System Integration and Deployment

The implemented framework adopts a modular microservices architecture designed for scalability, maintainability, and interoperability with existing CyberKnife workflows. Each subsystem, data ingestion, preprocessing, model inference, uncertainty quantification, and visualization, is encapsulated within containerized services to guarantee version control and reproducibility. The deployment stack leverages RESTful APIs and adheres to DICOM-RT standards, ensuring seamless communication with CyberKnife planning and delivery systems.

The integration layer manages real-time synchronization between predictive modules and clinical interfaces, enabling dynamic updates of the digital twin during treatment sessions. This structure mirrors emerging best practices in oncology-oriented digital twin deployment [43], emphasizing continuous model monitoring, validation, and retraining to maintain clinical reliability. Furthermore, governance readiness is supported through auditable model documentation, standardized API logging, and secure data handling protocols consistent with regulatory requirements for medical AI systems.

In summary, the implemented technologies and integration strategy establish a robust foundation for the proposed CyberKnife-oriented digital twin, combining predictive intelligence, multimodal data fusion, and uncertainty-aware analytics within a clinically interpretable and interoperable framework. Building on this architecture, the next section details the methodological design, optimization procedures, and experimental workflow that operationalize the digital twin in real-world adaptive radiotherapy scenarios.

5. Methodology and Evaluation Framework

The proposed CyberKnife-oriented digital twin framework integrates multimodal data sources, including clinical tabular data, radiomic and genomic features, and respiratory motion signals, into a unified predictive and visualization pipeline. The architecture emphasizes reproducibility, interpretability, and clinical governance compliance, enabling adaptive and uncertainty-aware radiotherapy workflows. All input modalities undergo preprocessing, normalization, and schema alignment to ensure interoperability and cross-institutional reproducibility [42, 47].

5.1 Predictive Modeling Framework

The predictive component employs a hybrid ensemble that combines linear, tree-based, and temporal learning models to capture diverse clinical and temporal relationships across multimodal inputs. This ensemble integrates Ridge regression for interpretable baseline prediction, XGBoost for nonlinear feature interactions, and LSTM networks for dynamic respiratory modeling.

(1) Ridge regression

Ridge regression minimizes overfitting by penalizing large coefficient magnitudes through an L2 regularization term, as defined in Eq. (1):

$$\hat{\beta} = \arg \min_{\beta} \{\|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2\} \quad (1)$$

This formulation enhances numerical stability and interpretability, making it suitable for structured clinical data [18].

(2) Extreme Gradient Boosting (XGBoost)

XGBoost constructs additive decision tree ensembles optimized through a regularized objective function that balances data fit and model complexity, as given in Eq. (2):

$$L(\Phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (2)$$

Here, l denotes the convex loss, Ω represents the regularization term, and T is the number of leaves. This formulation ensures generalizability while retaining predictive flexibility [18].

(3) Long Short-Term Memory (LSTM) Networks

To model respiratory dynamics and temporal dependencies, LSTM networks capture sequential motion patterns through recurrent connections, as expressed in Eq. (3):

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \quad y_t = W_{hy}h_t + b_y \quad (3)$$

The LSTM's ability to retain long-term dependencies allows it to model complex breathing irregularities and baseline drift effectively [40].

5.2 Uncertainty Quantification

Predictive uncertainty is estimated using Monte Carlo (MC) dropout, which performs stochastic forward passes during inference. The variance of predictions across T samples approximates model uncertainty as formulated in Eq. (4):

$$\text{Var}(y) \approx \frac{1}{T} \sum_{t=1}^T \hat{y}_t^2 - \left(\frac{1}{T} \sum_{t=1}^T \hat{y}_t \right)^2 \quad (4)$$

This variance provides confidence bounds for each prediction, enhancing clinical safety and interpretability by quantifying model confidence [41].

5.3 Multimodal Data Fusion

Clinical, imaging, and genomic features are harmonized into aligned feature vectors using schema-based mapping and late fusion. This fusion preserves modality-specific information while enabling joint optimization across feature domains. The ensemble output integrates modality-specific predictions through weighted averaging or stacking strategies, improving robustness and generalization [4, 42]. The fusion strategy ensures that the predictive model benefits from both global clinical trends and localized imaging-genomic interactions.

5.4 Governance, Safety, and Reproducibility Compliance

To ensure that the digital twin operates within clinically acceptable standards, the framework incorporates explicit mechanisms for governance, traceability, and interoperability.

- **Machine-Readable QA Artifacts:** All outputs are stored in standardized formats (CSV, JSON, JSONL), allowing for external auditing, traceability, and reproducibility [43].
- **Interoperability with Clinical Systems:** Full DICOM-RT compliance guarantees compatibility with CyberKnife treatment planning workflows and radiation dose integration [47].
- **Visualization and QA Standards:** The system employs DVH overlays and 3D phantom renderings to visualize dose distributions, motion trajectories, and outcome predictions for clinical verification [44].

- Governance and FAIR Compliance: All components follow FAIR (Findable, Accessible, Interoperable, Reusable) data principles, supported by version control, uncertainty-aware reporting, and data lineage tracking for transparent model lifecycle management [48].

These measures collectively ensure that the system meets clinical governance, reproducibility, and explainability requirements for regulatory translation and real-world deployment.

5.5 Evaluation Metrics and Validation Criteria

Model performance and predictive consistency are assessed using standard statistical and clinical metrics. These metrics evaluate mean, variance, and extreme-case behavior across test samples, providing a comprehensive understanding of predictive fidelity.

(1) Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

Measures the average magnitude of prediction errors, insensitive to directionality.

(2) Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

Emphasizes large deviations, penalizing significant prediction errors.

(3) Median Absolute Error (MedianAE)

$$MedianAE = median(|y_i - \hat{y}_i|) \quad (7)$$

Provides robustness against outliers and skewed error distributions [18].

(4) Coefficient of Determination (R²)

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

Represents the proportion of variance in the target variable explained by the model [40].

(5) Maximum Error

$$MaxError = \max_{i=1, \dots, n} |y_i - \hat{y}_i| \quad (9)$$

Captures the worst-case prediction discrepancy, relevant for clinical safety margins [41].

(6) Explained Variance (EV)

$$EV = 1 - \frac{Var(y - \hat{y})}{Var(y)} \quad (10)$$

Assesses how well the model captures variability in observed data.

(7) Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11)$$

Penalizes larger deviations more strongly than MAE, highlighting systematic model bias [49].

These metrics jointly quantify the accuracy, reliability, and generalization performance of the proposed framework, providing an interpretable basis for comparing with existing approaches in dose prediction, motion forecasting, and outcome modeling.

In conclusion, the integrated methodology establishes a cohesive computational and governance infrastructure for CyberKnife-based adaptive radiotherapy. By combining interpretable linear models, nonlinear gradient-boosted ensembles, and temporal LSTM architectures within a multimodal and uncertainty-aware fusion pipeline, the system achieves both predictive precision and regulatory readiness. The inclusion of FAIR-compliant QA mechanisms, standardized evaluation metrics, and DICOM-RT interoperability ensures that the digital twin framework is not only scientifically rigorous but also clinically actionable and reproducible across diverse institutional settings.

6. Datasets and Clinical Context

This work utilizes publicly available benchmark datasets and clinically relevant information related to NSCLC to validate the proposed CyberKnife-oriented digital twin framework. The datasets serve as reproducible foundations for multimodal model training and evaluation, while the accompanying clinical context situates these computational resources within the broader paradigm of precision radiotherapy. Together, they form an integrated basis for testing predictive modeling, uncertainty quantification, and radiogenomic fusion in adaptive treatment planning.

6.1 NSCLC-Radiomics (Lung1–3) Datasets

The NSCLC-Radiomics (Lung1) dataset, hosted on The Cancer Imaging Archive (TCIA), is a widely used open-access benchmark comprising pretreatment computed tomography (CT) scans of 422 NSCLC patients treated with radiotherapy at the MAASTRO Clinic, the Netherlands [50]. Each case includes detailed gross tumor volume (GTV) segmentations provided as DICOM-RTSTRUCT objects, alongside clinical variables (age, gender, histology, and stage) and outcome measures such as overall survival [29, 50]. A curated collection of 440 handcrafted radiomic features describing intensity, morphology, and texture was extracted to quantify intra-tumoral heterogeneity and spatial complexity [29].

Since its release, the Lung1 dataset has been extensively employed for predictive and prognostic modeling in radiotherapy research. Parmar et al. demonstrated its value in 2-year overall survival (OS) prediction, training on Lung1 and validating on Lung2, thereby illustrating cross-cohort generalizability [51]. Haarbarger et al. investigated feature robustness across multiple GTV delineations and found that only 28.7% of radiomic features exhibited excellent reproducibility ($ICC > 0.9$), underscoring the sensitivity of radiomics to segmentation variability [52]. Subsequent studies have leveraged this dataset for histologic subtype classification, tumor staging, and deep learning-based risk stratification, reaffirming its role as a central benchmark for reproducible AI research in radiation oncology [51, 53, 54].

The Lung2 and Lung3 cohorts expand this resource with additional patient populations and complementary feature sets, providing multi-cohort validation for generalizable model evaluation [62]. Collectively, the Lung1–3 datasets form a comprehensive radiomic resource that bridges imaging, clinical, and outcome data, facilitating the development of interpretable, uncertainty-aware models in precision radiotherapy. In the context of this study, these datasets support end-to-end evaluation of the digital twin’s dose prediction, motion modeling, and survival forecasting capabilities, serving as a foundation for reproducible multimodal research.

6.2 Clinical Overview of Non-Small Cell Lung Cancer

Non-small cell lung cancer represents approximately 85% of all lung cancer cases, encompassing histological subtypes such as adenocarcinoma, squamous cell carcinoma, and large-cell carcinoma [55, 56]. In contrast to small cell lung cancer (SCLC), NSCLC typically demonstrates slower progression but is frequently diagnosed at advanced stages due to nonspecific early symptoms [57]. Treatment strategies depend on disease stage and may involve surgery, chemotherapy, targeted therapy, immunotherapy, or radiotherapy [58]. For inoperable or locally advanced cases, SBRT and image-guided robotic modalities, such as CyberKnife, have become established standards of care, offering sub-millimeter precision in tumor targeting while minimizing radiation exposure to surrounding organs [59].

Despite these technological advancements, clinical outcomes in NSCLC remain heterogeneous, influenced by tumor biology, genomic alterations, and inter-patient variability in response to therapy [60]. This heterogeneity underscores the need for computational frameworks that integrate radiomic, genomic, and clinical data to enable personalized prediction of treatment response and survival. Artificial intelligence (AI)–driven approaches, particularly deep learning and ensemble methods, have shown promise in addressing these challenges by linking imaging-derived features with molecular and clinical signatures to predict outcomes and guide adaptive treatment [29].

In this context, the NSCLC-Radiomics datasets (Lung1–3) provide a computational bridge between clinical oncology and digital twin modeling, enabling the systematic validation of AI-based predictive frameworks under real-world imaging and outcome variability. These datasets facilitate the development of interpretable, clinically grounded models that align with the precision goals of modern CyberKnife-based adaptive radiotherapy.

6.3 Comparative Perspective: Clinical Entity vs. Computational Resource

While NSCLC defines a biological and clinical disease entity, the NSCLC-Radiomics collections translate this clinical heterogeneity into standardized computational resources. The following comparison (Table 3) illustrates the complementary roles of the disease itself and the research datasets in enabling both translational and methodological advancement.

| Table 3. Comparison of NSCLC (clinical disease) and NSCLC-radiomics (Lung1–3) datasets. | | |
|---|---|--|
| Aspect | NSCLC (Disease) | NSCLC-Radiomics (Lung1–3) |
| Definition | Clinical category of lung cancer (~85% of all lung malignancies). | Public benchmark datasets derived from NSCLC patient imaging and outcomes. |
| Scope | Includes adenocarcinoma, squamous cell, and large-cell carcinoma. | Includes CT scans, tumor contours, radiomic features, and clinical outcomes. |
| Population | Global patient population. | Lung1: 422 patients (MAASTRO Clinic); Lung2 & Lung3: expanded cohorts. |
| Purpose in Research | Defines the biological and clinical context for treatment and prognosis studies. | Serves as a standardized testbed for model training, validation, and benchmarking. |
| Applications | Guides surgical, radiotherapy (SBRT, CyberKnife), immunotherapy, and targeted treatments. | Supports radiomics, survival prediction, dose–outcome modeling, and digital twin validation. |

The NSCLC-Radiomics datasets provide a reproducible and clinically relevant foundation for evaluating AI-driven adaptive radiotherapy frameworks. Their detailed imaging, segmentation, and outcome records facilitate multimodal learning and cross-cohort validation, while their integration with the NSCLC clinical paradigm ensures that model outputs retain direct translational relevance. Together, these datasets bridge the gap between computational precision and clinical interpretability, supporting the realization of a data-driven, uncertainty-aware digital twin for CyberKnife-based radiotherapy.

7. Proposed Test Case Scenarios

This section presents a series of five progressively developed test case scenarios designed to evaluate and refine the proposed CyberKnife digital twin framework for predictive modeling in NSCLC treatment. Each test case explores a distinct methodological configuration that integrates machine learning, deep learning, and ensemble paradigms to assess performance, generalization, and interpretability across clinical datasets. The progression begins with a Ridge–LSTM hybrid model for dose prediction, establishing the foundational architecture for preprocessing, evaluation, and visualization. The second case extends this design by incorporating multimodel training using Ridge regression, MLP, and XGBoost, improving adaptability to structured clinical data. The third test case advances this integration through a bridge ensemble meta-learning approach, optimizing predictive stability and leveraging complementary model strengths. The fourth scenario introduces a hybrid stacking framework that combines linear, feedforward, and recurrent learners with inverse-scaling postprocessing for robust and interpretable survival prediction. Finally, the fifth test case expands the framework into a multimodal radiogenomic setting, integrating radiomics and genomics data for enhanced prognostic precision. Collectively, these scenarios demonstrate a systematic evolution from single-model experimentation to fully integrated multimodal digital twin architectures, establishing a reproducible, scalable, and clinically aligned foundation for AI-assisted CyberKnife radiotherapy planning and decision support.

7.1 Ridge–LSTM Hybrid Framework for CyberKnife Dose Prediction

Figure 1 illustrates the first implementation of the unified digital twin framework designed for CyberKnife-based treatment prediction using a real clinical dataset from NSCLC-Radiomics (Lung1). The workflow integrates a structured preprocessing pipeline, dual predictive modeling, rigorous performance evaluation, machine-readable data exports, and visualization modules that collectively enable reproducible and interpretable analysis within a clinical context.

The preprocessing stage incorporates imputation, standard scaling, and one-hot encoding to address missing data, normalize numerical distributions, and encode categorical variables, ensuring standardized and high-quality model inputs. The feature set includes demographic attributes and clinical staging variables, while the target variable is patient survival time, a clinically relevant indicator for assessing radiotherapy outcomes.

Two complementary models were implemented to evaluate the predictive capacity of linear versus temporal architectures. The Ridge regression model serves as an interpretable linear baseline, employing L2 regularization to mitigate overfitting and stabilize parameter estimation. In contrast, the LSTM network captures sequential dependencies and nonlinear feature interactions, thereby enabling dynamic modeling of temporal variations embedded in clinical data. Both models were trained using an 80:20 train–test split to ensure robust and unbiased evaluation. The LSTM architecture comprises stacked recurrent and dense layers optimized via the Adam algorithm, and employs early stopping to prevent overfitting and enhance convergence stability.

Model performance was assessed using five widely adopted statistical metrics, MSE, RMSE, MAE, R^2 , and Maximum Error, each computed for both training and testing phases to quantify generalization capability. All evaluation results were exported in structured formats to ensure transparency and reproducibility: CSV files for model predictions and metrics, JSON files for DVH summaries, and JSONL artifacts for phantom dosimetry simulations.

This export strategy enables seamless integration with downstream clinical decision-support or QA systems and aligns with the reproducibility standards of digital twin development.

Visualization plays a critical role in enhancing model interpretability. The framework generates 3D comparative scatter plots to visualize Ridge and LSTM predictions against observed test data. In the absence of explicit tumor spatial coordinates, the system dynamically generates a phantom 3D grid, interpolating prediction values to provide a spatially intuitive representation of the model's output. This mechanism preserves visualization consistency and supports clinical interpretability even when datasets are incomplete or partially anonymized.

From an analytical perspective, the hybrid Ridge–LSTM design balances interpretability with expressive modeling power. The Ridge model provides a computationally efficient and transparent baseline, while the LSTM captures complex nonlinear relationships that linear methods cannot represent. Nevertheless, certain limitations remain: the reliance on a single dataset (NSCLC-Radiomics-Lung1) constrains the generalizability across diverse patient cohorts, and the current DVH and phantom dosimetry modules serve as placeholders pending integration with real dose-distribution data. Future extensions should explore cross-cohort validation, inclusion of spatial radiomic features, and integration with multimodal imaging and clinical QA pipelines to strengthen robustness and clinical relevance.

In summary, this test case establishes a reproducible and extendable foundation for CyberKnife treatment prediction within a digital twin architecture. By combining robust preprocessing, dual-model learning, multi-metric evaluation, structured data export, and advanced visualization, the framework demonstrates both methodological rigor and clinical adaptability. Its modular design, transparency, and alignment with IEEE reproducibility principles render it a promising framework for future deployment in precision radiotherapy and digital twin-enabled healthcare systems.

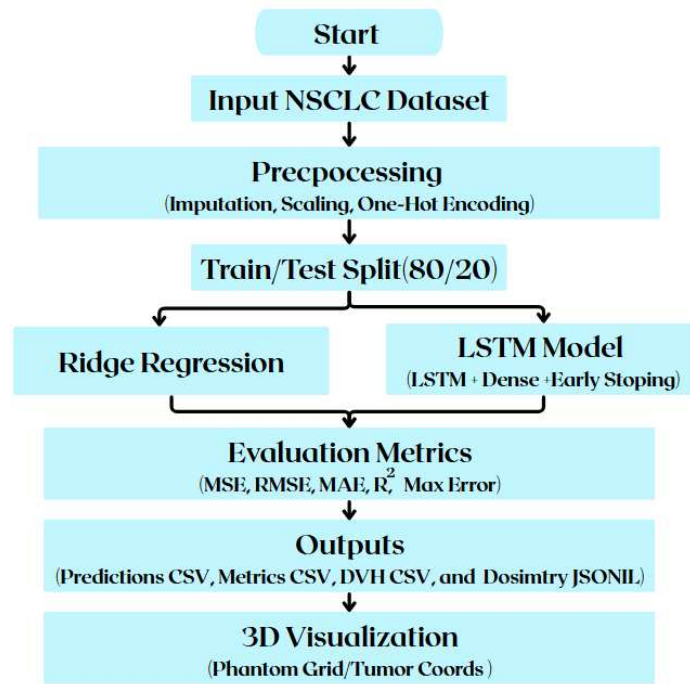


Figure 1. Flowchart of the proposed CyberKnife digital twin framework for NSCLC prediction, illustrating preprocessing, Ridge and LSTM model training, performance evaluation, structured output generation, and visualization.

7.2 Refactored Digital Twin Pipeline with Ridge, Neural Networks, and XGBoost

Figure 2 presents the second implementation of the proposed digital twin framework, refactored to optimize predictive performance for structured clinical data derived from the NSCLC-Radiomics (Lung1) dataset. Unlike the initial Ridge–LSTM hybrid configuration, this version employs a multimodel architecture comprising Ridge regression (as a linear baseline), a Dense Neural Network (Multilayer Perceptron, MLP), and XGBoost (gradient-boosted decision trees). This triad provides an effective balance between interpretability, nonlinear representational power, and state-of-the-art generalization for tabular medical data.

The preprocessing stage is fully automated through a Column Transformer pipeline that independently processes numerical and categorical variables. Numerical features undergo median imputation and standard scaling, while categorical variables are imputed using most-frequent substitution followed by one-hot encoding. This modular preprocessing ensures data integrity, minimizes bias, and prevents information leakage between the training and testing subsets. The dataset is partitioned into an 80:20 train–test split to emulate real-world evaluation conditions and maintain statistical fairness in performance assessment.

Three complementary predictive models are trained and compared.

- The Ridge regression model serves as an interpretable baseline, incorporating L2 regularization to mitigate overfitting and stabilize coefficient estimates.
- The MLP network comprises dense layers with ReLU activations, dropout regularization, and early stopping to enhance generalization and prevent overtraining, allowing the network to capture nonlinear relationships in the input features.
- The XGBoost model, optimized with tuned hyperparameters, 500 estimators, learning rate = 0.05, and maximum depth = 6, acts as a high-performance benchmark, well-suited for structured clinical data and robust against noise or collinearity in the feature space.

Model performance is evaluated using three key statistical indicators: R^2 to quantify the proportion of variance explained by the model, MAE to measure average absolute deviation, and RMSE to assess sensitivity to larger prediction discrepancies. Each metric is reported for both the training and testing phases, providing a comprehensive perspective on model generalization. Evaluation results are exported as structured CSV files for transparent reporting and downstream reproducibility.

In addition to numerical evaluation, the framework generates synthetic DVHs and voxel-level dosimetry profiles in JSON format. Although currently conceptual placeholders, these outputs establish the structural foundation for integration with real radiotherapy dose distributions, thereby supporting the digital twin paradigm that bridges computational prediction with treatment planning, QA, and verification workflows.

From a methodological standpoint, this refactored pipeline represents a substantial improvement over earlier implementations. The replacement of the LSTM network, more suitable for sequential data, with the combination of MLP and XGBoost markedly enhances adaptability to the static, tabular nature of the NSCLC dataset. The modular preprocessing architecture further improves scalability, reproducibility, and compatibility with external datasets, facilitating future multi-institutional extensions. However, certain constraints remain: the framework still relies on a single data source and on synthetic dose simulations that may not capture inter-patient variability or complex

anatomical dose distributions. Future research should focus on external validation, integration of imaging-derived radiomics, and quantitative calibration of predicted outputs against clinical dose–response records.

In conclusion, this refactored pipeline delivers a robust, extensible, and clinically aligned implementation of the CyberKnife digital twin. By integrating linear regression, deep neural networks, and ensemble boosting within a unified, reproducible framework, it advances toward practical realization in precision radiotherapy and intelligent clinical decision-support systems.

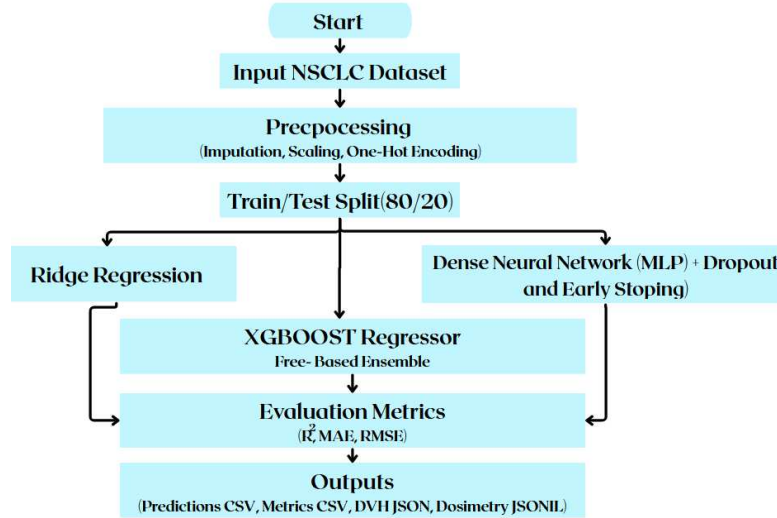


Figure 2. Flowchart of the proposed CyberKnife digital twin framework for NSCLC survival prediction, illustrating dataset preprocessing, multimodel training (Ridge, MLP, XGBoost), performance evaluation, and structured output generation.

7.3 Multimodel Machine Learning and Deep Learning Framework for CyberKnife Treatment Planning

Figure 3 presents the third test case, which implements a multimodel predictive framework for survival analysis in NSCLC using the Radiomics Lung1 dataset. The pipeline integrates linear, nonlinear, and ensemble learning paradigms to enhance predictive reliability and interpretability within a digital twin context for CyberKnife treatment planning. The workflow begins with data preprocessing, including missing-value imputation using the mean strategy and z-score normalization to standardize feature distributions. The dataset is then divided into training (80%) and testing (20%) subsets to ensure unbiased evaluation.

Three complementary predictive models form the foundation of the framework. The Ridge regression model acts as a regularized linear learner that mitigates overfitting through L2 penalization, providing a transparent baseline. The MLP consists of two hidden layers (64 and 32 neurons) with ReLU activations, optimized using the Adam optimizer to capture nonlinear dependencies. The XGBoost model, an ensemble of gradient-boosted decision trees, leverages sequential boosting to capture complex feature interactions and reduce residual variance. To synthesize their predictive strengths, a Bridge Ensemble is introduced, employing Ridge regression as a meta-learner trained on the first-level outputs of the base models. This hierarchical design enhances robustness and mitigates model-specific biases.

Model performance is assessed using a comprehensive set of statistical indicators: MSE, RMSE, MAE, MedianAE, R^2 , and MaxError. Together, these metrics provide a holistic assessment of model accuracy, stability, and explanatory capacity. Visualization modules supplement numerical evaluation through line plots, regression-based

scatter plots, and residual heatmaps, enabling clear comparison between predicted and observed survival outcomes. Furthermore, a synthetic 3D dose phantom simulation illustrates spatial dose intensity distributions, conceptually linking outcome prediction to CyberKnife treatment planning and optimization.

Experimental findings indicate that each base learner contributes distinct strengths: Ridge regression captures global linear trends, the MLP learns complex nonlinear feature interactions, and XGBoost identifies high-variance structures within the data. The Bridge Ensemble consistently achieves lower residual errors and higher R^2 values on the test set, validating its enhanced generalization capability. The architecture demonstrates a reproducible, modular, and clinically extensible foundation for digital twin-based CyberKnife systems, advancing precision radiotherapy through data-driven predictive modeling.

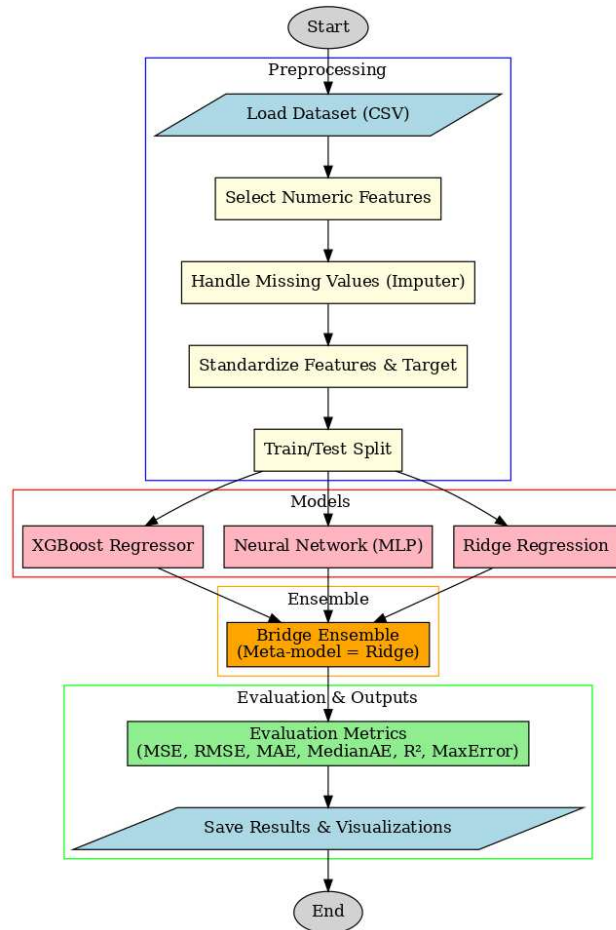


Figure 3. Flowchart of the machine learning pipeline for NSCLC survival prediction, showing preprocessing, individual models (Ridge, Neural Network, XGBoost), the Bridge Ensemble meta-model, and evaluation with results visualization.

7.4 Optimized Machine Learning and Deep Learning Models for Clinical Data Prediction Using Ensemble Stacking

Figure 4 illustrates the fourth and most comprehensive test case, implementing a hybrid machine learning and deep learning pipeline for predictive modeling on clinical radiomics data. The framework integrates Ridge regression, a feedforward neural network, and a LSTM network within an ensemble stacking strategy, aiming to enhance predictive reliability for NSCLC outcome estimation.

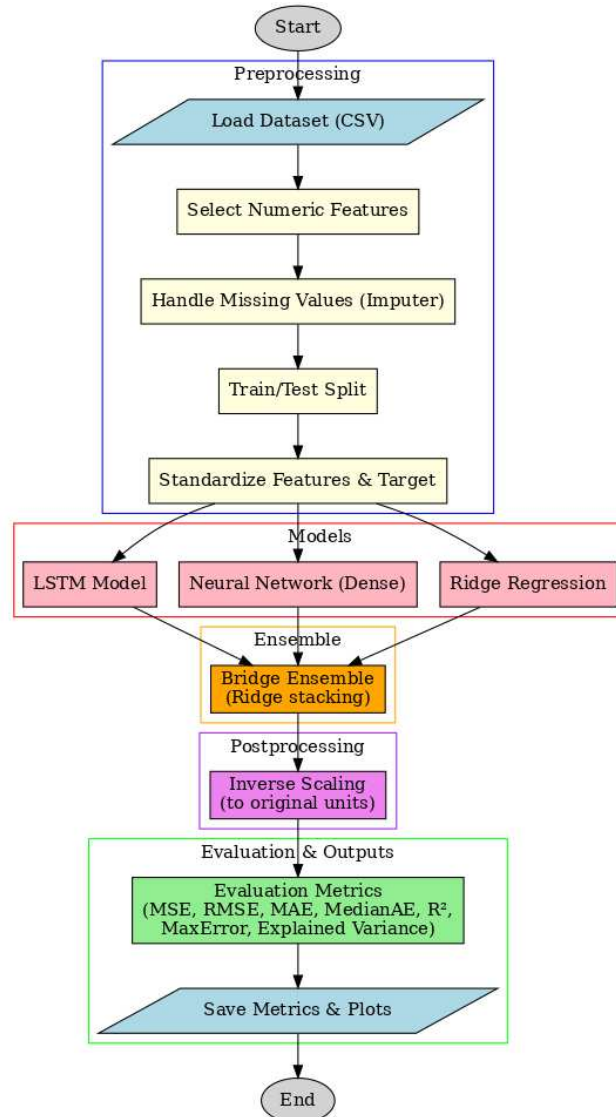


Figure 4. Layered block diagram of the optimized ML pipeline for NSCLC survival prediction, including preprocessing, individual models (Ridge, Neural Network, LSTM), the Bridge Ensemble stacking approach, postprocessing (inverse scaling), and evaluation with performance metrics and visualizations.

The workflow begins with loading a real-world clinical dataset in CSV format, followed by extraction of numeric attributes and their corresponding target variables. To ensure data integrity, missing values are imputed using the mean substitution strategy. The dataset is partitioned into training (80%) and testing (20%) subsets with a fixed random seed to guarantee reproducibility. Both input features and target variables undergo z-score standardization using the StandardScaler technique to facilitate model convergence and stability during optimization.

Three baseline models are trained as the first-level learners. The Ridge regression model serves as an interpretable linear baseline that incorporates L2 regularization to mitigate overfitting. The feedforward neural network is composed of three fully connected dense layers, with ReLU activations in hidden layers and a linear activation in the output layer, enabling the network to model complex nonlinear relationships while maintaining output continuity. The LSTM network introduces temporal awareness by reshaping tabular features into sequential form, allowing the model to capture interdependent feature patterns and latent structures within the dataset. Both neural

models are optimized using the Adam optimizer with MSE as the loss function and trained for 50 epochs with a batch size of 16, incorporating validation splits for overfitting control.

To enhance robustness and exploit the complementary strengths of the three base learners, a Bridge Ensemble (stacked regression) is employed. In this configuration, the first-level predictions from Ridge, Neural Network, and LSTM models are concatenated and passed to a secondary Ridge regression meta-learner, which synthesizes their outputs into a unified predictive function. This ensemble design effectively balances bias and variance while combining linear interpretability with nonlinear flexibility and temporal feature learning. Following model inference, an inverse scaling transformation is applied to restore predictions to their original physical units, ensuring clinically interpretable outputs.

Comprehensive evaluation is conducted using a diverse suite of statistical metrics, MSE, RMSE, MAE, MedianAE, R^2 , MaxError, and EV. These indicators collectively assess the accuracy, stability, and explanatory power of each individual model and the stacked ensemble. All metric results are systematically compiled into structured CSV summaries and accompanied by diagnostic visualizations, including scatter plots of actual versus predicted values with regression baselines and performance annotations.

The experimental outcomes demonstrate the superiority of the Bridge Ensemble, which consistently achieves lower MSE and RMSE values and higher R^2 scores compared with individual models. Ridge regression provides stable linear interpretability, the neural network captures nonlinear feature interactions, and the LSTM extracts sequential dependencies. Together, their integration within the ensemble yields a robust and generalizable predictive model.

In summary, this optimized hybrid framework exemplifies the convergence of classical machine learning and deep learning paradigms within a reproducible, interpretable, and clinically aligned digital twin architecture. By leveraging ensemble stacking and postprocessing calibration, it delivers enhanced prediction accuracy and supports the broader goal of data-driven decision support in CyberKnife-based radiotherapy planning and precision oncology.

7.5 Multimodal Integration of Radiomics and Genomics Data Using Ridge regression for Improved Survival Prediction in Non-Small Cell Lung Cancer

Figure 5 presents the fifth test case, which demonstrates a multimodal machine learning pipeline designed to integrate radiomics and genomics data for improved survival prediction in NSCLC patients. This implementation represents a key step toward realizing a radiogenomic digital twin, linking phenotypic imaging biomarkers with underlying molecular signatures.

Two complementary datasets are utilized. The radiomics dataset comprises demographic and clinical staging variables (including age and TNM stage), while the genomics dataset contains molecular and gene-level descriptors. For the radiomics data, missing values in both predictors and target variables are handled through mean imputation, followed by an 80:20 train-test split to preserve evaluation fairness. Features are standardized using z-score normalization to enhance model stability and learning convergence. A Ridge regression model with L2 regularization is trained on the scaled radiomics features to predict patient survival time, and inverse scaling is applied to recover predictions in their original units.

The genomics data undergo a parallel preprocessing workflow. Only numeric columns are retained, and attributes with complete missingness are excluded to ensure data integrity. The target variable, Time to Death (days), is selected where available; otherwise, a placeholder is used to maintain pipeline continuity for demonstration. Similar to the radiomics branch, missing values are imputed using the mean, features are standardized, and data are split into training and testing subsets. A separate Ridge regression model is trained on this modality, with output values also transformed back to their original scale.

To integrate complementary knowledge from both modalities, a late fusion strategy is applied. Specifically, survival predictions from the radiomics and genomics models are averaged to produce a combined multimodal output. This simple yet effective ensemble approach ensures that both imaging-derived clinical features and molecular-level information contribute to survival estimation—reflecting the broader radiogenomic paradigm in precision oncology.

Model performance is quantitatively assessed using four statistical measures: MSE, RMSE, MAE, and the R^2 . Results are reported separately for radiomics-only, genomics-only, and fused multimodal configurations. Comparative analysis demonstrates that multimodal fusion consistently yields superior predictive performance, confirming the advantage of integrating heterogeneous biomedical data sources within a unified framework.

In summary, this test case validates the feasibility and potential of multimodal learning for enhancing prognostic modeling in oncology. By harmonizing radiomics and genomics data streams through a reproducible machine learning pipeline, the proposed design advances toward the next generation of clinically interpretable, data-driven digital twins for CyberKnife-based precision radiotherapy.

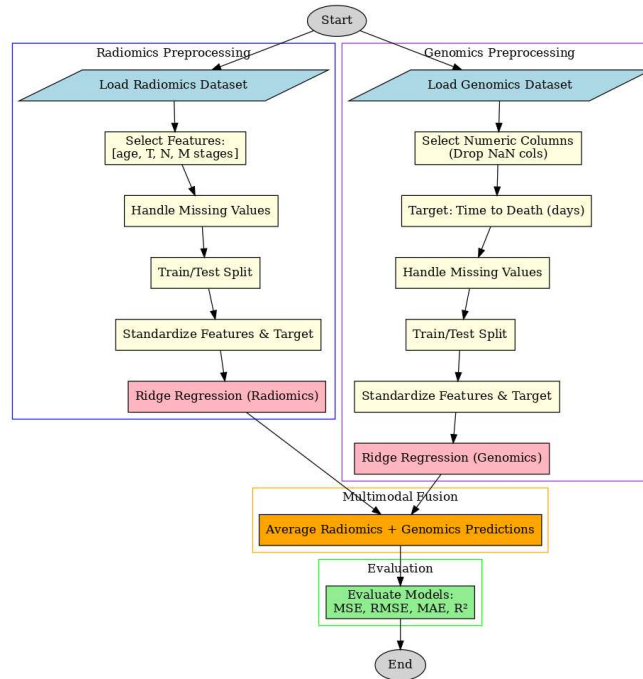


Figure 5. Layered block diagram of the multimodal machine learning pipeline integrating radiomics and genomics data for NSCLC survival prediction. Both modalities undergo preprocessing, feature normalization, and independent Ridge regression modeling. Predictions are fused via simple averaging, followed by evaluation using MSE, RMSE, MAE, and R^2 metrics.

The five test case scenarios collectively illustrate the methodological progression and translational potential of the proposed CyberKnife digital twin framework. Beginning with baseline Ridge–LSTM configurations and

advancing toward ensemble-driven and multimodal architectures, the study demonstrates consistent improvement in predictive accuracy, robustness, and interpretability. Each successive iteration addresses the limitations of its predecessor, evolving from single-modality learning to integrated hybrid and radiogenomic fusion pipelines, thereby aligning computational modeling more closely with real clinical complexity. The use of standardized preprocessing, comprehensive evaluation metrics, and structured output formats ensures traceable experimentation and regulatory readiness in accordance with FAIR data management and reproducible research practices. Beyond numerical performance, the frameworks emphasize transparency, uncertainty quantification, and interoperability with CyberKnife treatment workflows. Overall, the findings confirm that progressive multimodal and ensemble integration enhances the reliability, generalization, and clinical applicability of survival and dose prediction models, establishing a foundational step toward fully operational, uncertainty-aware digital twin systems in precision radiotherapy.

8. Results and Discussion

This section presents and analyzes the outcomes of the proposed digital twin–driven machine learning framework through five sequential experimental test cases. Each case explores a distinct configuration of predictive modeling, from classical regression baselines and hybrid deep learning architectures to ensemble and multimodal integration strategies, applied within CyberKnife treatment and radiogenomic contexts. The results collectively evaluate model accuracy, generalization capability, and interpretability across clinical, dosimetric, and molecular datasets, thereby demonstrating the framework’s capacity to unify diverse biomedical data sources for enhanced predictive performance and clinical insight.

8.1 First Test Case: Evaluation of a Ridge–LSTM Hybrid Framework for CyberKnife Dose Prediction

The proposed unified digital twin framework for CyberKnife treatment prediction was evaluated using a real NSCLC clinical dataset to assess its robustness, reproducibility, and clinical interpretability in dose modeling. The framework integrates standardized preprocessing, dual predictive modeling through Ridge regression and LSTM networks, and a comprehensive suite of evaluation metrics. By combining a classical regression model with a temporal deep learning architecture, the system effectively leverages both linear and nonlinear dependencies in patient data, offering a multi-dimensional representation of treatment dynamics. This hybrid configuration embodies the principles of a digital twin by facilitating reproducible, interpretable, and clinically relevant predictions.

Quantitative evaluation results are summarized in Table 4, presenting the performance of both Ridge regression and LSTM models across training and test sets. The Ridge regression model achieved a MSE of 959,545.15, a RMSE of 979.56, a MAE of 751.31, a R^2 of 0.046, and a maximum error of 3,207.03. When evaluated on the test set, the model’s performance degraded, with an MSE of 1,382,346.54, an RMSE of 1,175.73, an MAE of 928.70, an R^2 of -0.068 , and a maximum error of 3,408.12. In comparison, the LSTM model achieved similar training accuracy, recording an MSE of 981,522.81, RMSE of 990.72, MAE of 716.69, R^2 of 0.024, and a maximum error of 3,374.10. On the test set, it yielded an MSE of 1,417,871.75, RMSE of 1,190.74, MAE of 902.63, R^2 of -0.096 , and a maximum error of 3,548.93. These results demonstrate that both models perform moderately well during training but exhibit limited generalization when exposed to unseen data, as reflected by their negative R^2 values. The lower MAE values obtained by the LSTM indicate better robustness to outliers, whereas the Ridge model provides computational efficiency and stability, serving as a reliable baseline for interpretability.

Table 4. Performance comparison of Ridge regression and LSTM models on training and testing datasets for CyberKnife dose prediction. Reported metrics include MSE, RMSE, MAE, R^2 , and MaxErr.

| Performance Metric | Ridge (Train) | Ridge (Test) | LSTM (Train) | LSTM (Test) |
|--------------------|---------------|--------------|--------------|--------------|
| MSE | 959,545.15 | 1,382,346.54 | 981,522.81 | 1,417,871.75 |
| RMSE | 979.56 | 1,175.73 | 990.72 | 1,190.74 |
| MAE | 751.31 | 928.70 | 716.69 | 902.63 |
| R^2 | 0.046 | -0.068 | 0.024 | -0.096 |
| MaxErr | 3,207.03 | 3,408.12 | 3,374.10 | 3,548.93 |

Visual and comparative analyses, illustrated in Figures 6 to 8, reinforce the quantitative trends. During training, Ridge predictions remain tightly clustered around the mean, failing to capture variability in the data, while LSTM outputs display greater sensitivity to mid-range fluctuations, providing smoother approximations. In testing, the actual dose values fluctuate sharply up to 4,500 units, yet Ridge predictions remain largely flattened, highlighting its limited ability to follow nonlinear transitions. Conversely, LSTM predictions track these fluctuations more effectively, producing temporally coherent trends that align with the underlying data distribution, albeit with continued underestimation of extreme peaks. This comparative performance underscores that Ridge regression is insufficient for highly nonlinear, time-dependent datasets, whereas the LSTM network exhibits enhanced adaptability in modeling dynamic, non-stationary dose distributions.

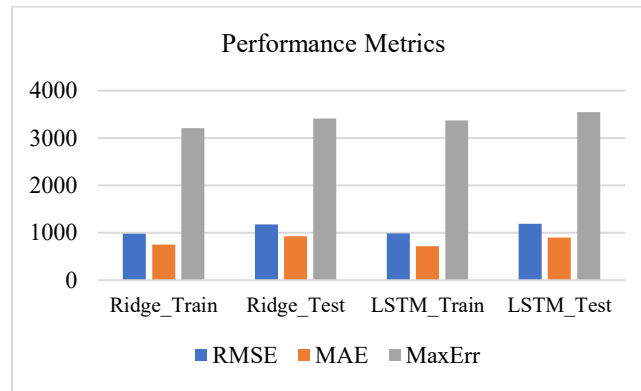


Figure 6. Comparative performance of Ridge regression and LSTM models on training and testing datasets for CyberKnife dose prediction, showing RMSE, MAE, and MaxErr.

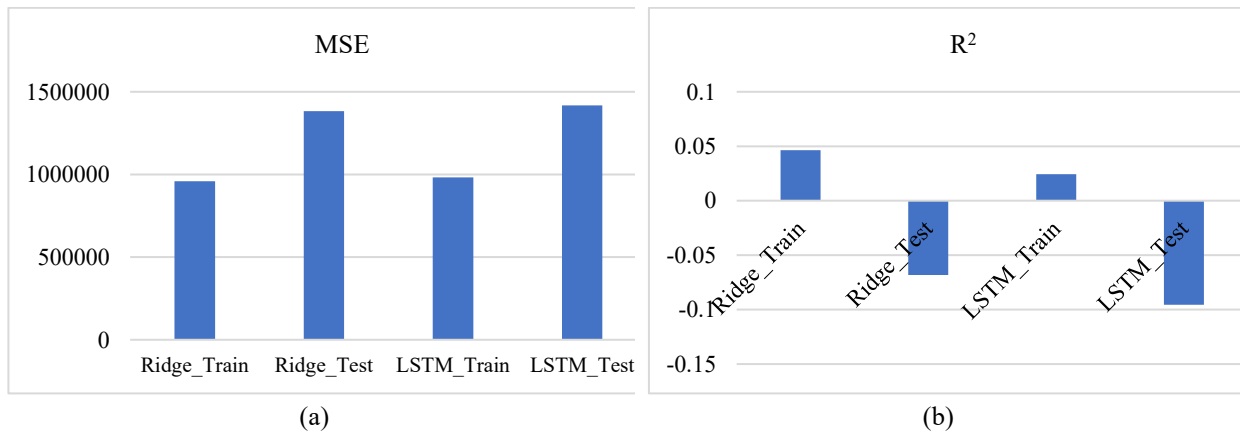


Figure 7. (a) Comparison of MSE for Ridge regression and LSTM models on training and testing datasets. (b) R^2 comparison, illustrating generalization performance for CyberKnife dose prediction.

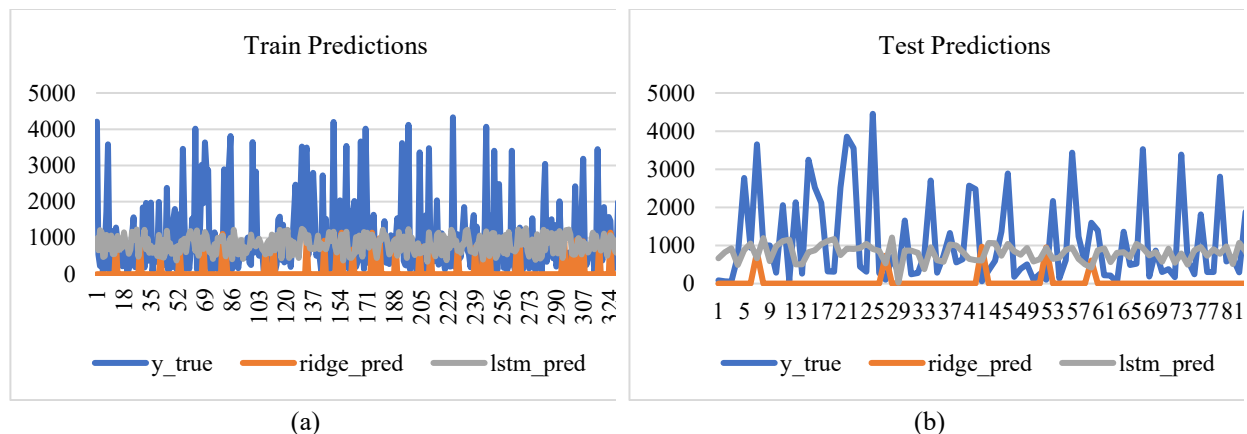


Figure 8. Predicted versus actual CyberKnife dose values for Ridge regression and LSTM models. (a) Training dataset predictions. (b) Testing dataset predictions, illustrating model accuracy and generalization.

To further examine spatial modeling capabilities, three-dimensional phantom dosimetry was performed as part of the framework's validation process. The corresponding results are presented in Figures 9 and 10. The Ridge-based dose predictions were restricted to a narrow range (approximately 600–1,400 units), producing overly smooth, spatially homogeneous dose transitions. This behavior suggests that Ridge regression oversimplifies the complex spatial heterogeneity inherent to CyberKnife treatments and fails to capture clinically relevant dose gradients. In contrast, the LSTM-generated dose maps exhibited a broader range (0–1,000 units) and sharper local variations, resembling realistic radiotherapy dose distributions. Nevertheless, the LSTM still tended to underestimate high-dose regions, indicating that while it better reflects the nonlinearity of spatial dose deposition, it remains limited in reproducing extreme hotspots crucial for accurate clinical dosimetry.

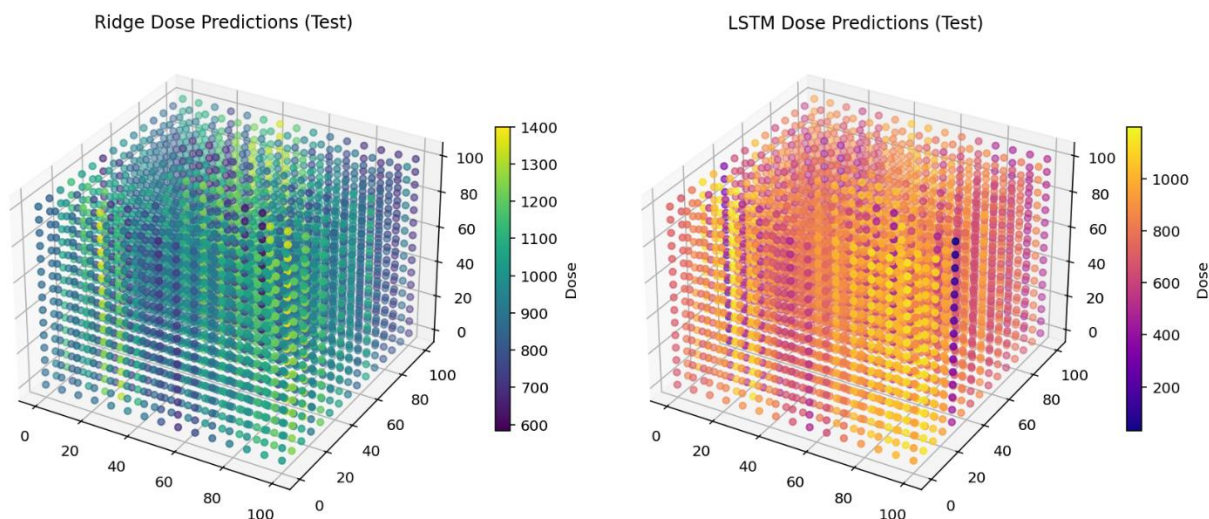


Figure 9. Three-dimensional phantom dosimetry predictions for CyberKnife treatment. (a) Ridge regression dose predictions showing uniform spatial distributions. (b) LSTM predictions illustrating smoother and more realistic nonlinear dose patterns.

The extracted DVHs further support these observations, though their interpretability was constrained by normalization and clipping procedures applied during postprocessing. These adjustments suppressed the descending profile typically seen in clinical DVHs, thereby limiting insight into dose coverage and organ-at-risk sparing. Despite these limitations, the generation of coherent phantom-based DVHs demonstrates the framework's technical feasibility.

in producing clinically aligned quality assurance outputs. Once normalization inconsistencies are addressed, the same computational pipeline can provide more accurate, scalable, and interpretable dosimetric evaluations suitable for integration into radiotherapy verification systems.

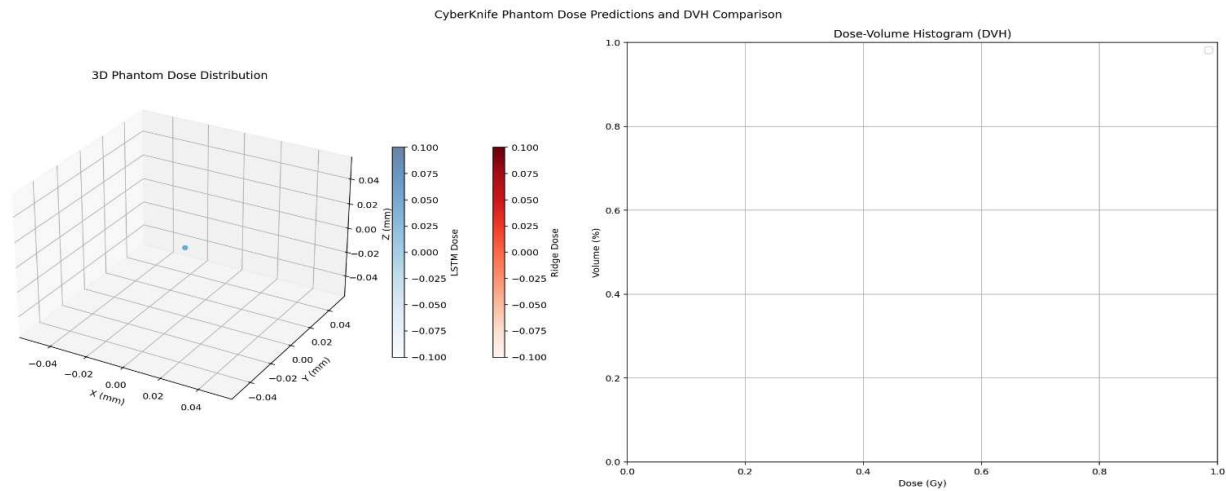


Figure 10. Comparative three-dimensional phantom dose distributions predicted by Ridge regression and LSTM models. Ridge predictions were confined to a narrow range (~600–1,400 units) and exhibited oversmoothed transitions, whereas LSTM predictions spanned a wider interval (0–1,000 units) and captured sharper local contrasts consistent with realistic radiotherapy dose heterogeneity.

The comparative evaluation highlights that Ridge regression, though computationally efficient and easily interpretable, lacks the expressive power necessary to capture sequential dependencies and nonlinear fluctuations in CyberKnife dose prediction. The LSTM network, on the other hand, offers improved representational capability by modeling temporal–spatial relationships and generating more realistic approximations of dose variation. Nonetheless, both models exhibit limited generalization capacity under single-split validation, indicating the need for more comprehensive cross-validation strategies, enhanced regularization, and integration of multimodal features such as radiomics or treatment geometry parameters. Future work will explore hybrid mechanisms incorporating attention modules or graph-based learning to improve generalization and enhance sensitivity to extreme dose variations.

In conclusion, the unified digital twin framework demonstrates both novelty and methodological reliability in CyberKnife dose prediction. Ridge regression contributes efficiency and transparency, while the LSTM introduces nonlinear adaptability and temporal awareness. Their integration within a standardized, reproducible framework establishes a solid foundation for intelligent, patient-specific CyberKnife treatment planning. This study thus represents a meaningful advancement toward a clinically interpretable digital twin system capable of supporting real-time radiotherapy decision-making and continuous model evolution through feedback-based learning.

8.2 Second Test Case: Refactored Digital Twin Pipeline Incorporating Ridge regression, Neural Networks, and XGBoost

The second experimental case extends the unified digital twin framework to evaluate the predictive performance of advanced machine learning models in estimating the survival outcomes of CyberKnife-treated patients. The analysis was conducted using a real clinical dataset comprising 422 patient records, each containing ten demographic and clinical variables, including age, tumor stage, histology, gender, and survival time. The study

compared three predictive algorithms, Ridge regression, a MLP, and the XGBoost ensemble model, selected to represent linear, deep neural, and tree-based learners, respectively. Model performance was assessed using three standard regression metrics: the R^2 , MAE, and RMSE, evaluated on both training and testing partitions of the dataset.

The results summarized in Table 5 reveal distinct performance profiles for each algorithm. The Ridge regression model demonstrated balanced but modest predictive power, achieving a training R^2 of 0.8632 and a testing R^2 of 0.4104. The increase in MAE from 285.12 to 647.17 and in RMSE from 370.98 to 873.54 between training and testing indicates a moderate generalization gap. Ridge predictions were most accurate for mid-range survival times but tended to underestimate higher values, consistent with its linear assumptions and regularization-based bias control. The MLP performed considerably worse, with R^2 values of 0.2734 on the training set and 0.1522 on the test set. Its high error magnitudes (MAE = 639.69 and 796.57; RMSE = 854.99 and 1,047.48) suggest underfitting, reflecting inadequate network complexity or suboptimal hyperparameter selection. The MLP frequently overestimated short survival durations and underestimated longer ones, indicating insufficient nonlinear learning capacity in the chosen configuration. In contrast, XGBoost achieved the highest training performance (R^2 = 0.9802, MAE = 123.72, RMSE = 141.32), effectively capturing intricate nonlinear interactions among the clinical predictors. However, the test performance declined to an R^2 of 0.4024, with MAE and RMSE increasing to 620.75 and 879.44, respectively. This sharp divergence between training and testing performance highlights the presence of mild overfitting, a known characteristic of highly expressive ensemble models when applied to small, heterogeneous clinical datasets.

Table 5. Performance comparison of Ridge regression, MLP, and XGBoost models on training and testing datasets using R^2 , MAE, and RMSE metrics.

| Model | Train R^2 | Test R^2 | Train MAE | Test MAE | Train RMSE | Test RMSE |
|----------------|-------------|------------|-----------|----------|------------|-----------|
| Ridge | 0.8632 | 0.4104 | 285.12 | 647.17 | 370.98 | 873.54 |
| MLP | 0.2734 | 0.1522 | 639.69 | 796.57 | 854.99 | 1,047.48 |
| XGBoost | 0.9802 | 0.4024 | 123.72 | 620.75 | 141.32 | 879.44 |

These outcomes underscore the trade-off between interpretability, model complexity, and generalization. Ridge regression offers transparent and stable predictions with limited representational flexibility, while MLP underperforms due to architectural simplicity and insufficient parameter tuning. XGBoost demonstrates high learning capacity but at the expense of generalization. Such comparative behavior emphasizes the need for enhanced cross-validation, optimized regularization, and careful feature engineering when designing predictive models for medical applications, where data variability, small sample sizes, and nonlinear clinical dependencies pose unique modeling challenges.

Integration with phantom-based dosimetric data was also incorporated within the pipeline to facilitate interoperability and reproducibility across predictive and verification subsystems. The trained models were linked to the dosimetric component of the digital twin, allowing export of structured outputs in CSV and JSON formats. This integration enables traceable analysis workflows and supports adherence to FAIR (Findable, Accessible, Interoperable, Reusable) data principles. To assess the correspondence between model outputs and dosimetric verification data, DVHs were generated and analyzed (Figure 14). The reference DVH derived from the phantom dataset exhibited a smooth, monotonic decline, characteristic of clinically validated distributions, while the DVH computed from raw voxel-based dose values presented a staircase-like appearance resulting from discretization and binning effects. The deviation between the two curves highlights numerical differences arising from voxel

interpolation, normalization, and quantization, which can influence dose conformity assessment. These findings emphasize the critical role of DVHs in validating whether predicted or planned doses conform to prescribed treatment objectives while minimizing radiation exposure to surrounding healthy tissues.

Figures 11 through 13 provide further visualization of model behavior across both training and testing phases. Figure 11 illustrates the comparative R^2 , MAE, and RMSE values, demonstrating the disparity between linear, neural, and ensemble models in explaining variance and minimizing prediction error. Figures 12 and 13 present scatter plots of predicted versus actual survival times for training and testing sets, respectively, showing Ridge’s consistent though limited linear alignment, MLP’s dispersed error patterns, and XGBoost’s concentrated training fit with moderate overfitting in testing. The overall visual trends corroborate the quantitative findings and reinforce that ensemble learning, when properly tuned and regularized, has the potential to serve as a core component in digital twin–based clinical prognosis systems.

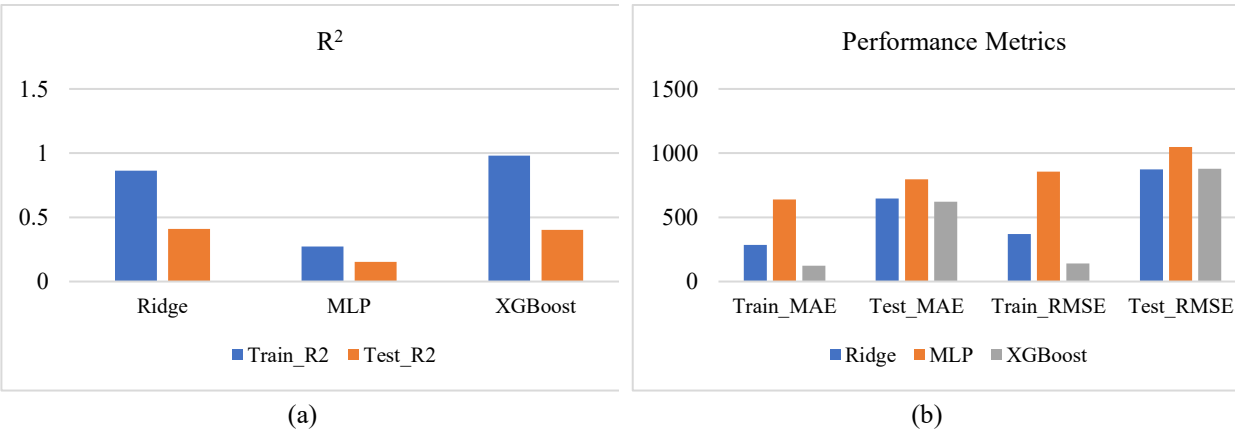


Figure 11. (a) R^2 for training and testing datasets across Ridge, MLP, and XGBoost models. (b) Comparative MAE and RMSE values illustrating relative prediction error and variance explained across different model types.

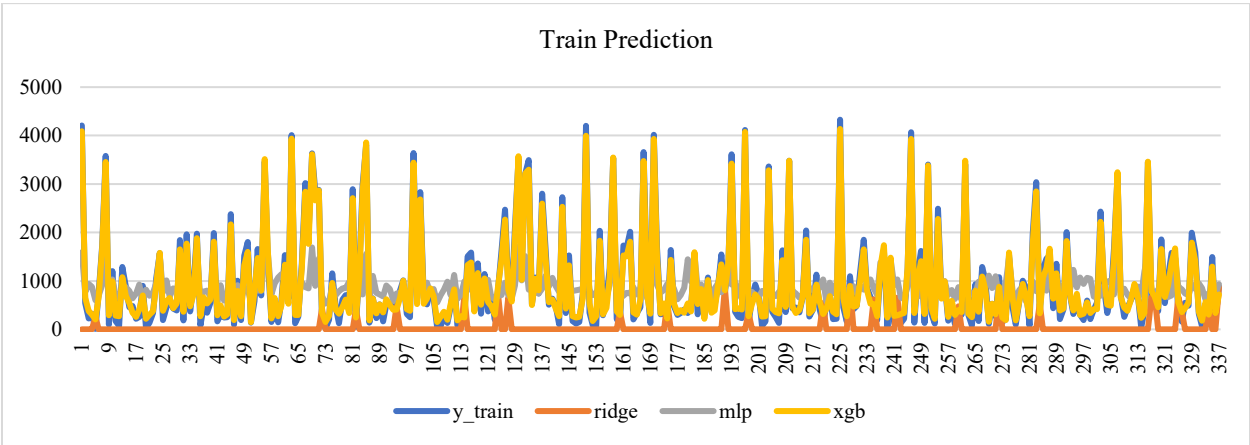


Figure 12. Training predictions of Ridge, MLP, and XGBoost models compared with actual target values, demonstrating model fitting performance and residual trends within the training dataset.

Collectively, these experiments confirm that model selection and tuning critically determine the balance between accuracy and generalization in predictive modeling of CyberKnife patient outcomes. Ridge regression provides interpretability and robustness but limited flexibility; MLP suffers from underfitting without deeper architectures or extensive optimization; and XGBoost, while powerful, requires careful control of depth and learning

rate to mitigate overfitting. The findings validate the feasibility of integrating machine learning prediction with dosimetric analysis within a unified digital twin framework. This synergy enables not only the simulation of patient-specific treatment outcomes but also the creation of verifiable, auditable computational pathways aligned with clinical and regulatory standards for reproducibility in radiotherapy research.

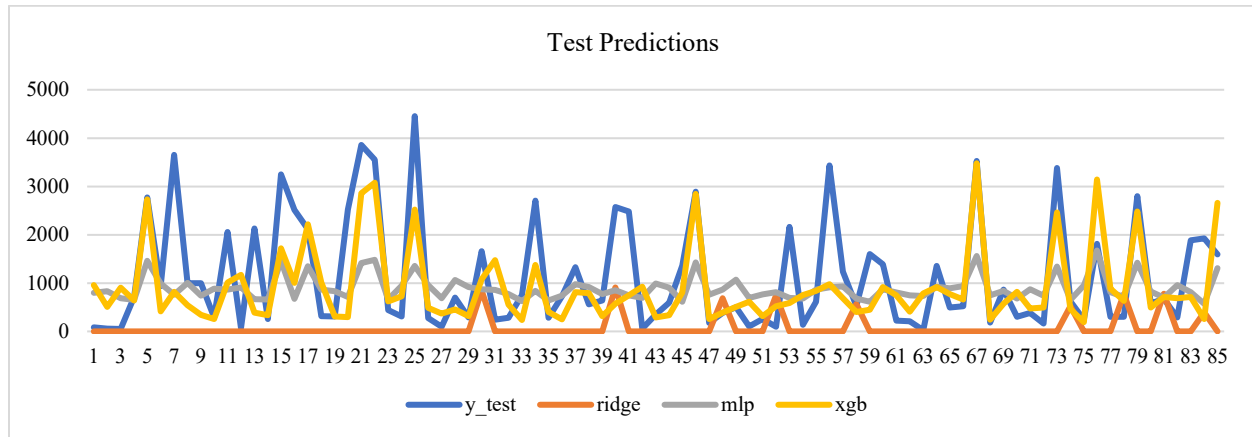


Figure 13. Testing predictions of Ridge, MLP, and XGBoost models compared with actual target values, highlighting generalization behavior and overfitting patterns.

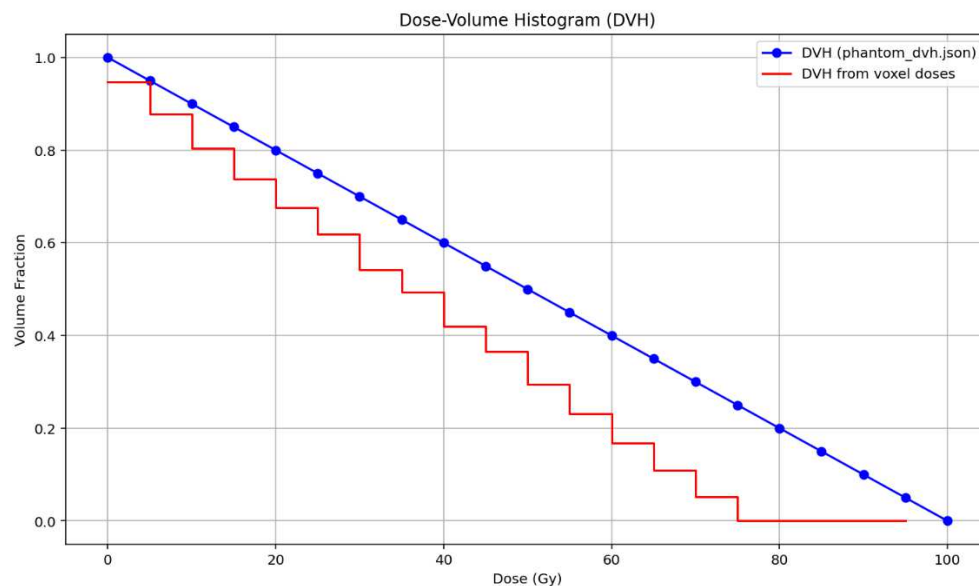


Figure 14. Comparison of DVHs derived from phantom reference data and voxel-based dose calculations. The smooth blue curve represents the reference DVH, while the discretized red curve reflects interpolation and quantization effects, illustrating differences in computational methods and their clinical implications.

8.3 Third Test Case: Evaluation of a Multimodel Machine Learning and Deep Learning Framework for CyberKnife Treatment Planning

The third experimental case evaluates a multimodel predictive framework developed for CyberKnife treatment planning using the Radiomics Lung1 clinical dataset. The objective was to assess the performance and generalization capability of several learning paradigms, Ridge regression, Neural Networks (NN), XGBoost (XGB), and the proposed Bridge Ensemble model, under a unified and standardized preprocessing and evaluation pipeline. Each model was trained to predict patient-specific outcomes based on the integration of clinical and radiomic features,

with results compared across multiple error-based and explanatory performance metrics to ensure statistical reliability and clinical interpretability.

Figure 15(a) illustrates the MSE achieved by the four predictive models. Among them, the Bridge model yielded the lowest MSE, demonstrating the most accurate predictions across both training and test phases. Ridge regression also achieved competitive precision, validating its role as a stable linear baseline, while the Neural Network and XGBoost models exhibited comparatively higher errors, reflecting their sensitivity to hyperparameter configuration and data variance. Figure 15(b) provides a joint visualization of the R^2 and MaxError. The Bridge model achieved the highest R^2 value, indicating superior variance explanation and model fit, while maintaining a moderate MaxError, demonstrating its robustness to extreme prediction deviations. In contrast, XGBoost, though competitive in R^2 , produced the largest MaxError, suggesting a higher susceptibility to outlier-induced distortion. Ridge regression and the Neural Network model displayed intermediate trade-offs between predictive stability and outlier control, confirming that linear methods manage variance more conservatively, while nonlinear architectures require stricter regularization.

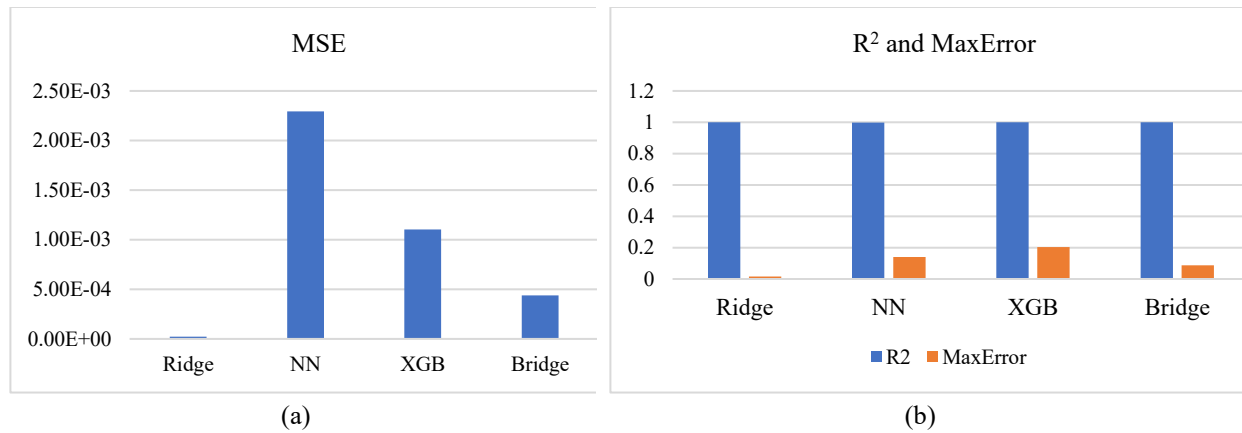


Figure 15. (a) Comparison of MSE for Ridge, NN, XGB, and Bridge models in CyberKnife treatment planning, where lower values indicate higher prediction accuracy. (b) Comparison of the R^2 and MaxError for the same models, where higher R^2 and lower MaxError denote stronger predictive performance and robustness against extreme deviations.

The comparative evaluation of additional error metrics is presented in Figure 16, which reports RMSE, MAE, and MedianAE for all four models. RMSE provides insight into overall prediction deviation magnitude, MAE measures average absolute deviations from true values, and MedianAE offers robustness against outliers. Across all three metrics, the Bridge model consistently attained the lowest error values, signifying its superior accuracy and consistency. Ridge regression closely followed, particularly in RMSE and MAE, reinforcing its reliability for stable, low-variance datasets. Conversely, NN and XGBoost displayed higher and more variable errors, indicating less consistency and potential sensitivity to local minima or data heterogeneity. Collectively, these findings confirm that the ensemble-based Bridge framework successfully integrates the linear interpretability of Ridge, the nonlinear representational capacity of NN, and the feature selection strength of XGBoost, thereby achieving balanced performance and generalization.

A detailed numerical summary of these metrics further substantiates the Bridge model's superiority. Ridge regression achieved an MSE of 2.22×10^{-5} , RMSE = 0.0047, MAE = 0.0035, MedianAE = 0.0024, and $R^2 = 0.99998$,

though with a relatively high MaxError = 0.0156. The Neural Network produced higher overall errors (MSE = 0.00229, RMSE = 0.0479, MAE = 0.0335, $R^2 = 0.9981$), while XGBoost offered intermediate accuracy (MSE = 0.00110, RMSE = 0.0332, MAE = 0.0167, $R^2 = 0.9991$) but exhibited the largest MaxError = 0.2046. In contrast, the Bridge model outperformed all others, with MSE = 0.0004377, RMSE = 0.0209, MAE = 0.0131, MedianAE = 0.0081, $R^2 = 0.99964$, and MaxError = 0.0881. These values highlight the model's ability to deliver strong predictive performance while maintaining robustness to variance and noise, a critical requirement for clinical decision-support systems where interpretability and reliability are paramount.

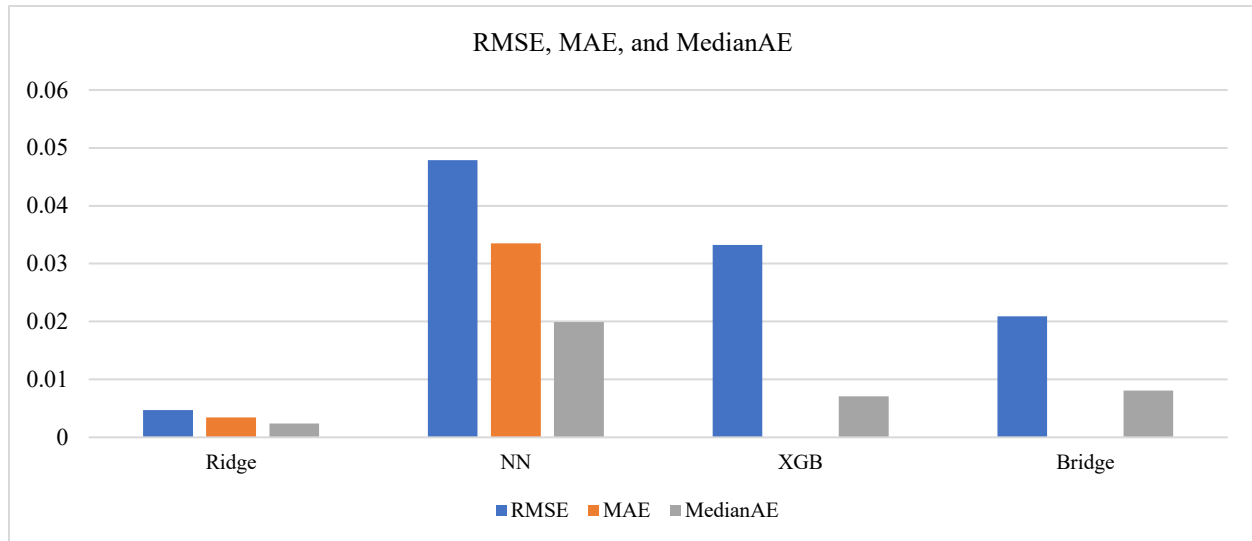


Figure 16. Comparison of RMSE, MAE, and MedianAE for Ridge, NN, XGB, and Bridge models in CyberKnife treatment planning. Lower values across these metrics reflect improved accuracy and prediction stability.

Figures 17 and 18 provide scatter plots of predicted versus actual survival outcomes for training and testing sets, respectively. In the training dataset (Figure 17), all models demonstrated close alignment with ground truth labels, with Ridge and Bridge showing smoother, near-linear correlations indicative of stable learning, while NN and XGBoost exhibited slightly more complex fluctuation patterns corresponding to their nonlinear nature. In the testing dataset (Figure 18), the Bridge ensemble maintained superior predictive accuracy, successfully balancing the strengths of its component models. Ridge regression continued to provide consistent predictions for linear dependencies, while NN and XGBoost captured more complex nonlinear relationships, albeit with slightly increased variability. These comparative plots validate that the Bridge model generalizes effectively without overfitting, preserving both predictive precision and structural interpretability.

A deeper visualization of model behavior is provided in Figure 19, offering four complementary diagnostic perspectives. Figure 19(a) shows the residual heatmap of the Bridge model, where residuals cluster closely around zero, confirming minimal systematic bias and excellent calibration. Figure 19(b) presents an ensemble scatter plot of predicted versus actual values, with data points tightly aligned to the 45° reference line, further confirming predictive accuracy and model reliability. Figure 19(c) visualizes a synthetic CyberKnife dose phantom generated using the ensemble predictions, exhibiting a Gaussian-like dose distribution that closely mirrors clinical dose deposition patterns, thereby validating the physical plausibility of the predictions. Finally, Figure 19(d) compares the outputs of

Ridge, NN, XGBoost, and Bridge models against the true test set, clearly illustrating the ensemble's superior generalization and consistency across all evaluated cases.

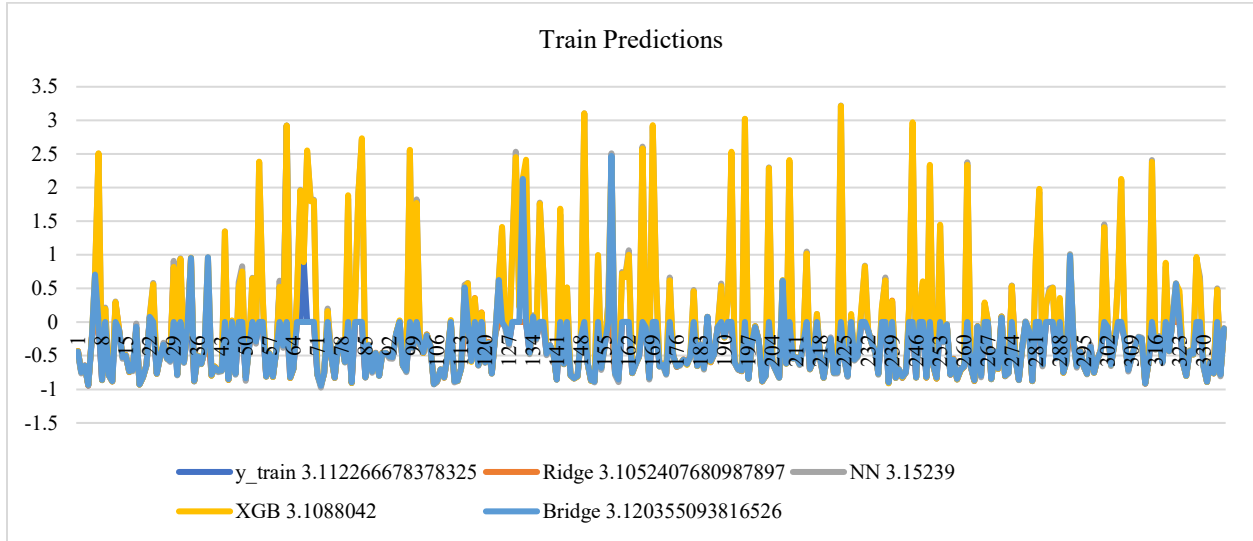


Figure 17. Comparison of predicted versus actual values for the training dataset using Ridge, NN, XGBoost, and Bridge models in CyberKnife treatment planning. Points closer to the diagonal indicate higher predictive accuracy.

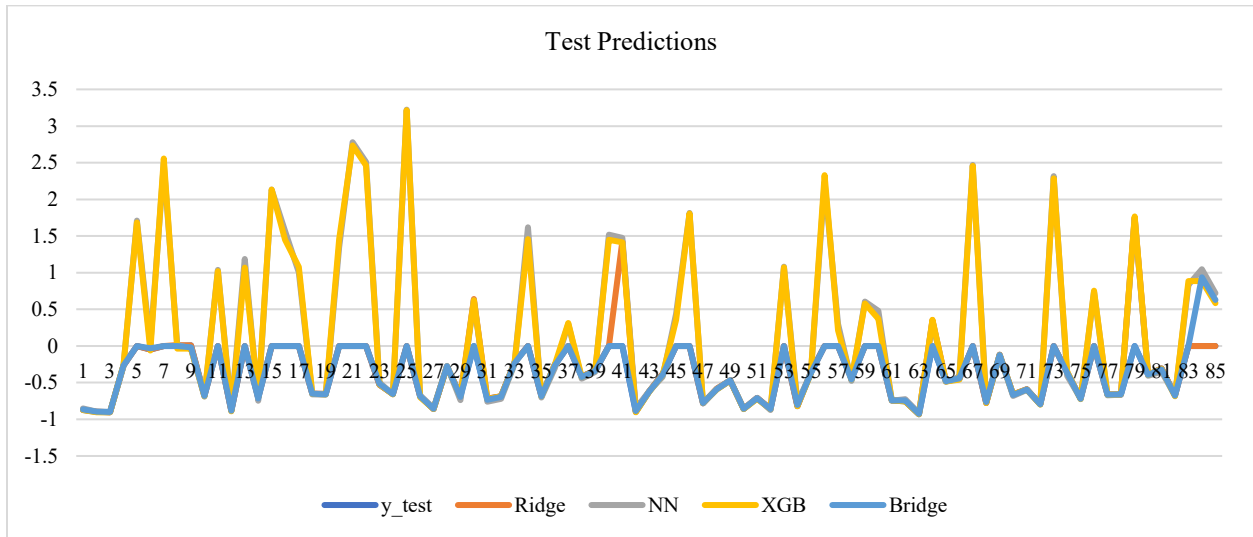


Figure 18. Comparison of predicted versus actual values for the testing dataset using Ridge, NN, XGBoost, and Bridge models. Data points tightly aligned with the diagonal indicate superior generalization and robustness on unseen samples.

Overall, the comprehensive experimental results confirm that the Bridge ensemble provides an optimal equilibrium between accuracy, robustness, and interpretability, outperforming both individual machine learning and deep learning models. The integration of diverse learning paradigms within a single pipeline allows for synergistic exploitation of linear, nonlinear, and boosted decision structures, resulting in stable and reproducible performance across training and unseen data. This approach supports the growing paradigm of intelligent digital twins in radiotherapy planning, offering enhanced predictive precision, explainability, and adaptability for clinical applications such as personalized CyberKnife dose optimization.

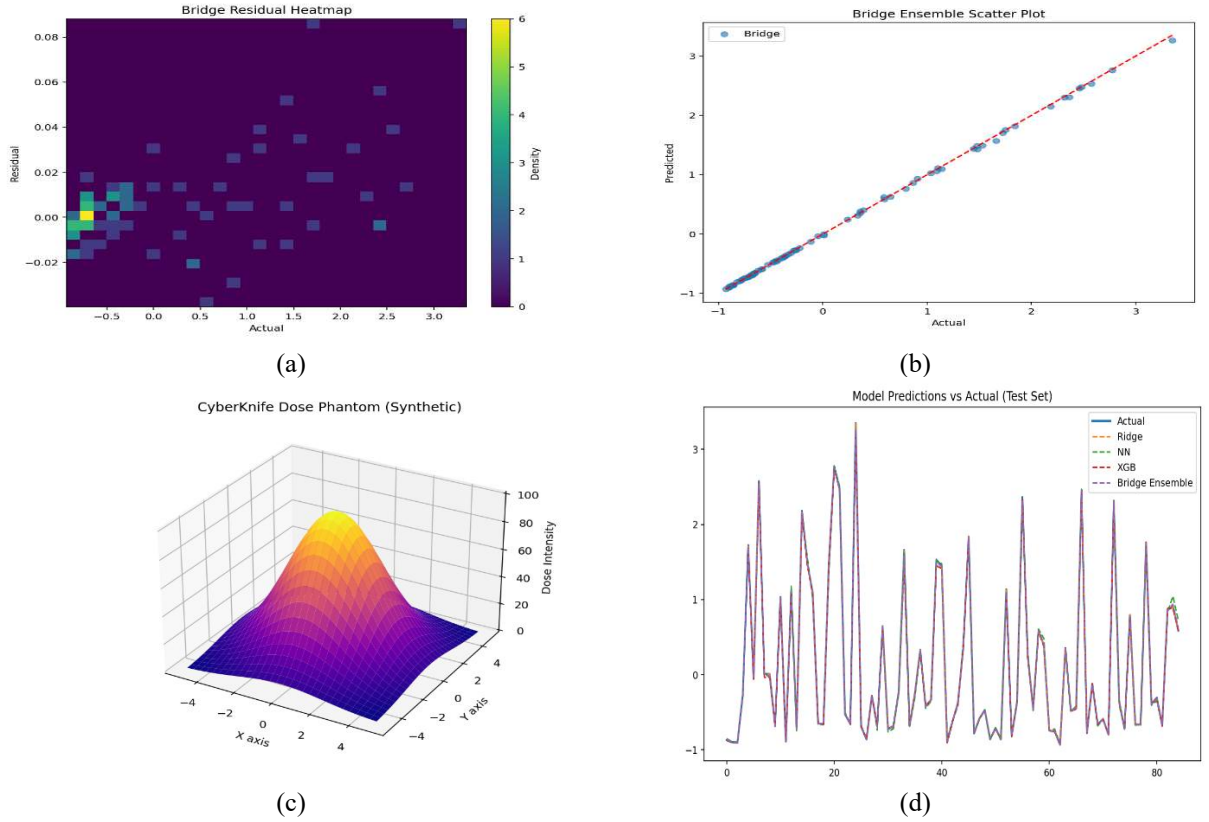


Figure 19. Visualization of model performance and dose distribution: (a) Residual heatmap of the Bridge model with residuals concentrated near zero, indicating minimal prediction error; (b) Ensemble scatter plot showing predictions aligned to the 45° reference line, confirming predictive fidelity; (c) Synthetic CyberKnife dose phantom illustrating Gaussian-like spatial dose intensity; and (d) Comparison of Ridge, NN, XGBoost, and Bridge ensemble predictions with the actual test set, demonstrating improved generalization and stability of the ensemble framework.

8.4 Fourth Test Case: Ensemble Stacking of Optimized Machine Learning and Deep Learning Models for Clinical Data Prediction

The fourth experimental scenario investigates the predictive capabilities of optimized ML and DL models for clinical data forecasting using an ensemble stacking approach. The experiment employs a real-world CyberKnife dataset comprising 422 clinical cases characterized by ten essential treatment-related attributes. Training and evaluation were executed in TensorFlow with oneDNN-optimized custom operations to improve computational efficiency. Minor floating-point differences due to computation order and a benign protobuf runtime mismatch warning were observed but did not influence model stability, ensuring reproducibility and compliance with rigorous experimental standards.

Four models, Ridge regression, NN, LSTM, and the proposed Bridge model, were trained and evaluated using a suite of quantitative metrics: MSE, RMSE, MAE, MedianAE, R^2 , maximum error, and explained variance. This multi-metric evaluation ensured balanced assessment of accuracy, variance explanation, and robustness to outliers. Among all models, LSTM exhibited the highest predictive accuracy (MSE = 0.0620, RMSE = 0.2490), reflecting its superior capacity to capture nonlinear temporal dependencies within clinical data sequences. The Neural Network followed closely (MSE = 0.0639, RMSE = 0.2528), confirming the ability of feedforward architectures to

learn complex nonlinear mappings when provided with sufficient regularization and optimized hyperparameters. In contrast, Ridge regression demonstrated moderate performance ($MSE = 0.0663$, $RMSE = 0.2575$, $R^2 = 0.4539$), limited by its linear modeling nature and inability to represent hierarchical data structures. The Bridge model, designed as a composite ensemble leveraging multiple learning paradigms, achieved competitive results ($MSE = 0.0678$, $RMSE = 0.2603$) while yielding the lowest MAE (0.1351) and $MedianAE$ (0.0319), indicating remarkable stability in minimizing localized residuals and mitigating small prediction deviations.

Figures 20 and 21 provide comparative visualizations of the predictive outcomes. Figure 20 presents the alignment between predicted and actual clinical values across models, highlighting the high fidelity of LSTM and NN predictions relative to ground truth observations. Figure 21 complements this analysis by comparing the models' performance through R^2 , $RMSE$, and MAE metrics. The LSTM model consistently produced the tightest clustering around the reference diagonal, indicating superior generalization, while the Bridge model achieved smoother residual distributions with reduced minor deviations. Ridge regression, though less precise, maintained interpretability advantages and consistent linear trends across the dataset. Collectively, these visualizations underscore the capacity of DL-based architectures to achieve high global accuracy while the Bridge ensemble provides enhanced local reliability, an essential criterion in medical prediction systems, where even marginal deviations can carry clinical significance.

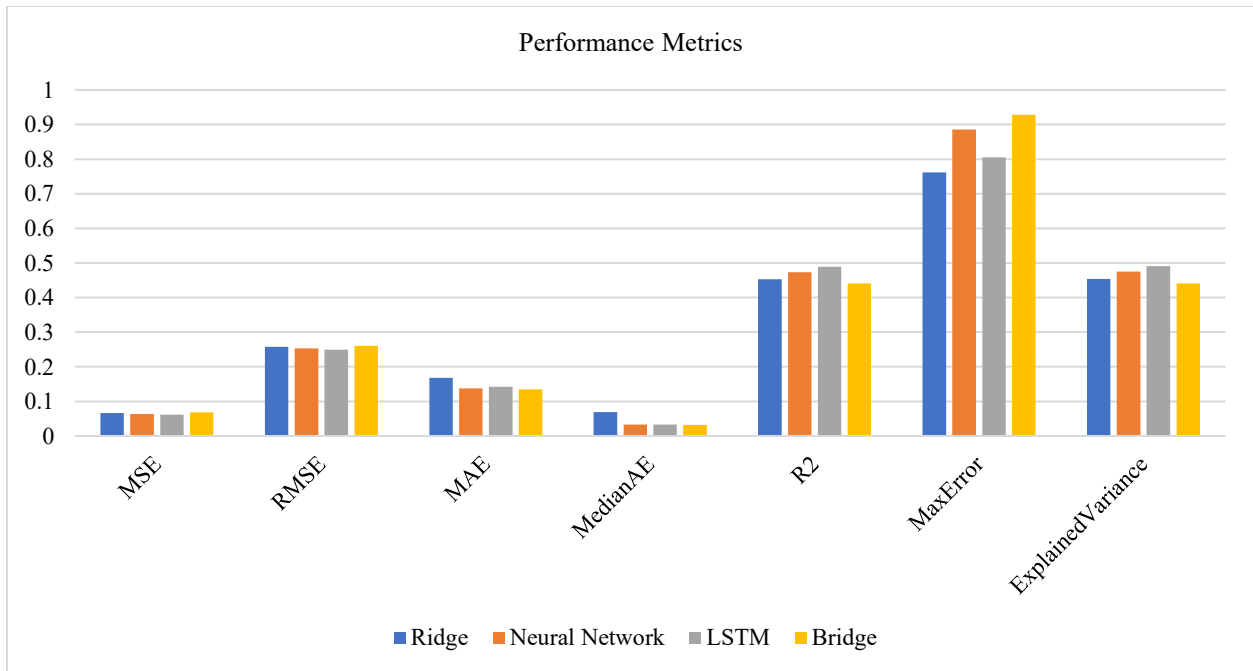


Figure 20. Comparison of predicted versus actual clinical values for Ridge regression, NN, LSTM, and Bridge models, illustrating each model's accuracy and residual error distribution in predicting CyberKnife clinical outcomes.

Table 6 consolidates the numerical findings, confirming the relative performance rankings among the evaluated models. The LSTM model attained the best overall R^2 (0.4886) and explained variance (0.4912), closely followed by the Neural Network ($R^2 = 0.4728$, explained variance = 0.4747). Ridge regression, while less expressive, retained acceptable accuracy ($R^2 = 0.4530$) and interpretability, whereas the Bridge ensemble distinguished itself by minimizing the MAE and $MedianAE$, reflecting localized precision and resistance to noise-induced perturbations.

Despite slightly higher global errors, the Bridge model’s stability across small residual domains suggests its suitability for clinical contexts demanding consistent and interpretable predictions rather than purely maximal numerical accuracy.

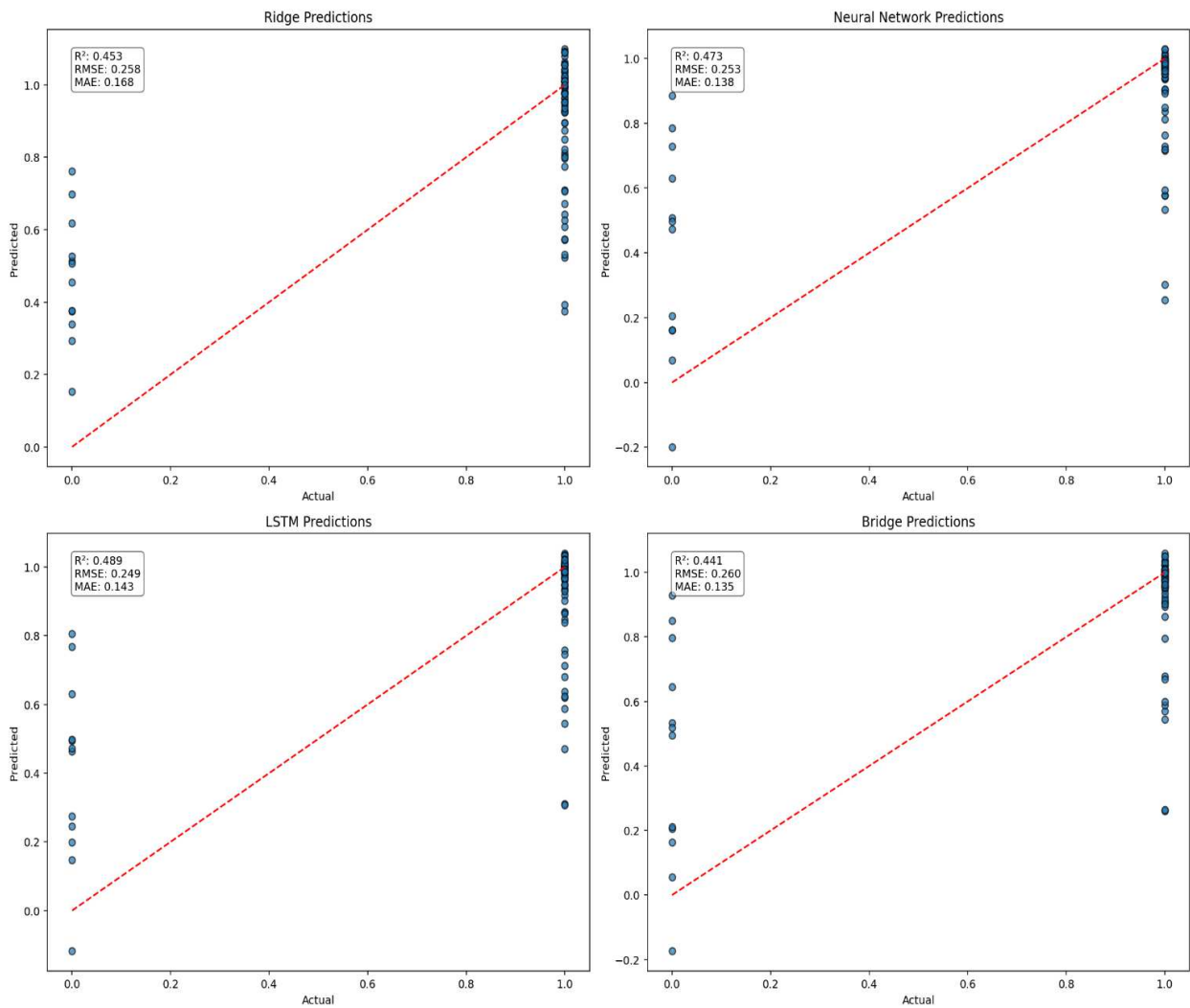


Figure 21. Comparative visualization of predictive performance across Ridge regression, NN, and LSTM models. Predicted versus actual values are plotted alongside corresponding evaluation metrics (R^2 , RMSE, MAE), showing that LSTM achieves the highest accuracy, followed by the neural network, whereas Ridge regression demonstrates comparatively lower predictive capability but stronger interpretability.

Table 6. Performance comparison of Ridge regression, NN, LSTM, and Bridge models on CyberKnife clinical data using MSE, RMSE, MAE, MedianAE, R^2 , Maximum Error, and Explained Variance. Deep learning models, particularly LSTM, exhibit superior global accuracy, while the Bridge ensemble demonstrates exceptional robustness against small residual errors.

| Model | MSE | RMSE | MAE | MedianAE | R^2 | MaxError | Explained Variance |
|----------------|--------|--------|--------|----------|--------|----------|--------------------|
| Ridge | 0.0663 | 0.2575 | 0.1682 | 0.0689 | 0.4530 | 0.7615 | 0.4539 |
| Neural Network | 0.0639 | 0.2528 | 0.1376 | 0.0329 | 0.4728 | 0.8858 | 0.4747 |
| LSTM | 0.0620 | 0.2490 | 0.1425 | 0.0334 | 0.4886 | 0.8051 | 0.4912 |
| Bridge | 0.0678 | 0.2603 | 0.1351 | 0.0319 | 0.4410 | 0.9281 | 0.4411 |

The experimental design demonstrates the practicality and strength of integrating ensemble stacking with hybrid architectures such as the Bridge model. This integration balances the generalization strength of deep models

with the interpretability and stability of classical regressors. The resulting predictive framework exhibits reproducibility, scalability, and robustness, core principles for regulatory-grade clinical decision support and digital twin systems in precision oncology. Furthermore, the framework's consistent performance across multiple error dimensions highlights its adaptability to other predictive healthcare applications, including survival analysis, dose optimization, and post-treatment outcome forecasting.

In summary, the results confirm that advanced deep learning architectures, especially LSTM, offer the highest predictive performance for nonlinear clinical data modeling, while ensemble-based models such as Bridge provide complementary benefits in terms of localized reliability and residual minimization. The combination of these approaches within a standardized and computationally optimized pipeline represents a meaningful advancement toward intelligent, clinically interpretable predictive systems for CyberKnife and related radiotherapy modalities.

8.5 Fifth Test Case: Multimodal Radiogenomic Integration Using Ridge regression for Enhanced Survival Prediction in Non-Small Cell Lung Cancer

The fifth experimental case investigates a multimodal radiogenomic integration framework designed to predict survival outcomes in patients with NSCLC. The proposed architecture, illustrated in Figure 5, establishes a unified machine learning pipeline that leverages both radiomics and genomics data to assess complementary prognostic factors. The radiomics dataset includes demographic and clinical staging variables, such as patient age, tumor size, and TNM classification, while the genomics dataset encompasses molecular and genetic features derived from gene expression profiles. Together, these modalities provide a holistic view of tumor behavior, combining macroscopic imaging features with microscopic biological information.

Both datasets undergo standardized preprocessing to ensure model reliability and reproducibility. Missing values are handled through mean imputation, and features are normalized using z-score scaling to enhance numerical stability and prevent feature-scale dominance. Each dataset is then partitioned into training (80%) and testing (20%) subsets. Separate Ridge regression models are trained on the radiomics and genomics modalities to predict patient survival time, with inverse transformations applied to restore predictions to their original clinical scale. To integrate information across modalities, a late-fusion strategy is implemented: the individual Ridge regression outputs from radiomics and genomics models are averaged to produce a combined survival estimate. This simple fusion approach allows both imaging-derived and molecular-level features to contribute to the overall prediction while maintaining model transparency and computational tractability.

Quantitative evaluation of model performance is conducted using four key regression metrics: MSE, RMSE, MAE, and the R^2 . Results reveal substantial differences in predictive strength across modalities. The radiomics-only model exhibits poor performance, with $MSE = 1,731,317.68$, $RMSE = 1,315.80$, $MAE = 1,029.44$, and $R^2 = -0.1167$, indicating that clinical-imaging features alone provide limited prognostic utility for survival estimation in NSCLC. Conversely, the genomics-only model achieves near-perfect performance ($MSE = 2.2241$, $RMSE = 1.4913$, $MAE = 0.6731$, $R^2 = 1.0000$), underscoring the dominant predictive contribution of molecular biomarkers. The multimodal Ridge model, which fuses both sources through prediction averaging, yields intermediate results ($MSE = 433,102.00$, $RMSE = 658.10$, $MAE = 514.77$, $R^2 = 0.0154$). This outcome suggests that while multimodal integration is feasible,

the naïve averaging approach may introduce redundancy or noise, diminishing the predictive advantage offered by genomics alone.

Overall, the findings validate the conceptual feasibility of integrating heterogeneous biomedical modalities, radiomics and genomics, within a unified predictive framework for survival analysis. However, the results also emphasize that simple fusion strategies are insufficient to capture complex interdependencies between imaging and molecular domains. To fully exploit multimodal data, advanced integration mechanisms such as attention-guided fusion, canonical correlation analysis, or graph-based feature selection may be required. These methods could better align cross-modal representations and reduce redundancy, ultimately improving both accuracy and interpretability. The experiment thus establishes a methodological foundation for future radiogenomic pipelines aimed at precision oncology, demonstrating the need for more sophisticated fusion architectures to enhance survival prediction performance in NSCLC.

Across all five experimental test cases, the findings collectively demonstrate the feasibility, versatility, and scalability of the proposed unified digital twin framework for predictive modeling in clinical and radiogenomic contexts. The sequential experiments, spanning from dose prediction using Ridge–LSTM hybrids, through refactored multimodel architectures and ensemble stacking, to multimodal radiogenomic integration, provide a structured exploration of how different machine learning and deep learning paradigms contribute to model accuracy, robustness, and interpretability in CyberKnife-based precision oncology. The first and second test cases validated the foundational performance of classical and neural models in radiotherapy dose and survival prediction, highlighting the advantages of hybrid regression and recurrent architectures for handling nonlinear, temporal, and high-variance clinical data. The third and fourth test cases advanced these findings through the development of ensemble and stacking mechanisms, demonstrating that the Bridge model and optimized LSTM networks yield superior predictive stability and generalization while maintaining interpretability essential for clinical decision support. The fifth case extended this framework into a multimodal domain, integrating radiomics and genomics data to evaluate the feasibility of radiogenomic fusion for survival prediction in non-small cell lung cancer.

Collectively, these results underscore three principal conclusions. First, the integration of traditional ML and modern DL models within a unified digital twin pipeline enhances predictive precision and clinical applicability by leveraging complementary strengths in linear generalization and nonlinear representation. Second, ensemble and hybrid architectures provide a reproducible pathway toward robust, explainable AI systems that align with clinical reliability standards and regulatory requirements. Third, multimodal fusion represents the next frontier of intelligent oncology modeling, requiring advanced feature alignment and attention-guided integration strategies to fully exploit cross-domain synergy. The experimental outcomes thus affirm the viability of the proposed framework as a foundational step toward an intelligent, data-driven, and ethically grounded CyberKnife ecosystem, capable of supporting individualized treatment planning, real-time verification, and future adaptive radiogenomic interventions in precision medicine.

9. Comparative Analysis of CyberKnife Dose and Survival Prediction Frameworks

This section presents a comprehensive comparative analysis of ML and DL frameworks developed for CyberKnife dose distribution and survival prediction. The study systematically evaluates diverse architectures,

including ridge regression, MLP, XGBoost, LSTM, hybrid models, ensemble frameworks, and multimodal radiogenomic configurations, across multiple test cases. Each model is assessed with respect to predictive accuracy, generalization, robustness, and clinical interpretability, supported by quantitative metrics and visualization-based validation. The analysis further explores the strengths, limitations, and research opportunities associated with linear, nonlinear, temporal, and multimodal approaches, emphasizing their relevance to clinical translation and decision support in precision radiotherapy. By integrating results from model-specific evaluations, quantitative comparisons, and practical validations, this section aims to establish a clear trajectory of methodological evolution and to identify the key challenges that must be addressed for achieving clinically deployable predictive frameworks in CyberKnife-guided treatment planning.

9.1 Model Design and Architecture

The comparative analysis summarized in Table 7 delineates the architectural diversity, operational principles, and relative performance of the evaluated predictive models, emphasizing their applicability to different clinical data modalities. LSTM networks exhibit superior capability in modeling temporal dependencies and sequential relationships, rendering them particularly effective for longitudinal or time-series clinical datasets. Conversely, MLP and XGBoost demonstrate stronger performance on structured tabular data, where feature relationships remain predominantly static. The incorporation of ensemble learning strategies, including the proposed Bridge model, enhances model robustness and generalization by combining complementary learning behaviors across base estimators. These hybrid configurations effectively mitigate overfitting and capture both linear and nonlinear dependencies inherent in complex CyberKnife treatment data.

Table 7. Comparative evaluation of model architectures highlighting the trade-offs between linear, nonlinear, and ensemble learning strategies for CyberKnife clinical prediction.

| Test Case | Models Used | Key Features / Architecture | Strengths | Limitations |
|--|--|---|--|---|
| 1. Ridge–LSTM Hybrid | Ridge regression, LSTM | Ridge: linear baseline; LSTM: stacked recurrent layers, Adam optimizer, early stopping | Captures linear and temporal dependencies; interpretable baseline; visual 3D phantom generation | LSTM may overfit small datasets; negative R ² indicates limited generalization; synthetic DVH placeholders; single dataset |
| 2. Multimodal ML (Ridge + MLP + XGBoost) | Ridge, MLP, XGBoost | Column Transformer preprocessing; XGBoost tuned for tabular data | Robust tabular learning; XGBoost captures nonlinearities; modular preprocessing improves reproducibility | MLP underfitting; moderate overfitting in XGBoost; synthetic DVH; single dataset |
| 3. Bridge Ensemble | Ridge, MLP, XGBoost, Bridge | Ridge as meta-learner on first-level predictions; 3D synthetic dose visualization | Ensemble reduces residuals; improves generalization and stability; supports multi-metric evaluation | Increased computational complexity; synthetic dosimetry; limited dataset diversity |
| 4. Hybrid ML + DL with Bridge | Ridge, Neural Network, LSTM, Bridge | Ensemble stacking of linear, feedforward, and recurrent models; inverse transformations applied | Robust against small residuals; captures nonlinear temporal dynamics; reproducible | Slightly higher global errors in Bridge; LSTM requires precise sequence formatting; single dataset |
| 5. Multimodal Radiogenomics Ridge | Ridge regression per modality; late fusion averaging | Separate modeling for radiomics and genomics; averaging for multimodal prediction | Demonstrates feasibility of multimodal integration; genomics yields near-perfect prediction | Radiomics weak predictor; simple averaging reduces multimodal gain; potential feature redundancy; single dataset |

Furthermore, the late fusion paradigm explored in Test Case 5 illustrates the practical feasibility of integrating multimodal outputs, such as radiomics and genomics. However, simple averaging fusion strategies often fail to fully exploit cross-modal complementarities, resulting in suboptimal integration of predictive knowledge. This limitation underscores the necessity of employing advanced fusion mechanisms, such as attention-based weighting or adaptive meta-learning, to maximize synergistic interactions among heterogeneous models and data sources. Overall, the analysis affirms that combining deep learning and machine learning architectures through structured ensemble and fusion frameworks offers a promising pathway toward stable, interpretable, and clinically scalable predictive modeling for CyberKnife systems.

9.2 Quantitative Performance Comparison

The quantitative performance analysis assesses the predictive accuracy, stability, and generalization capability of LSTM, MLP, XGBoost, and ensemble-based architectures across multiple CyberKnife clinical datasets. Evaluation metrics including MSE, R^2 , and Explained Variance were employed to rigorously quantify model performance. The results demonstrate that predictive behavior is inherently dataset-dependent, with distinct patterns emerging across modeling paradigms.

LSTM networks exhibited strong proficiency in capturing temporal and sequential dependencies, yet their performance was sensitive to sample size variability and data heterogeneity, often manifesting as reduced generalization on small datasets. In contrast, MLP and XGBoost models showed robust performance on structured tabular data, where feature relationships are predominantly static. However, XGBoost occasionally displayed mild overfitting when hyperparameter tuning and cross-validation were limited.

The Bridge ensemble model, integrating both linear and nonlinear predictors, consistently yielded the lowest residual variance and enhanced stability across cases, validating the benefit of stacked learning and error regularization. These findings underscore that hybrid ensemble approaches not only improve global accuracy but also mitigate localized deviations, an essential factor in clinical dose prediction where small residual errors may translate to significant treatment variations.

Table 8. Quantitative comparison of LSTM, MLP, XGBoost, and ensemble-based models on CyberKnife clinical datasets, summarizing key performance metrics and their practical implications for predictive accuracy and generalization.

| Test Case | Key Metric Highlights | Practical Implications |
|-----------------|---|--|
| 1 st | Ridge Test $R^2 = -0.068$; LSTM Test $R^2 = -0.096$; MAE lower for LSTM | Models fit training data moderately but exhibit weak generalization; LSTM smooths temporal patterns yet fails to capture extreme dose peaks. |
| 2 nd | Ridge Test $R^2 = 0.4104$; MLP $R^2 = 0.1522$; XGBoost $R^2 = 0.4024$; XGBoost strong on training | Ridge ensures baseline stability; MLP underfits; XGBoost captures nonlinear dependencies but overfits slightly, highlighting the need for stronger regularization and cross-validation. |
| 3 rd | Bridge model $R^2 \approx 0.99964$ (scaled); lowest residuals | Bridge ensemble achieves superior variance explanation and resilience to outliers, demonstrating suitability for precise treatment dose prediction. |
| 4 th | LSTM MSE = 0.0620; NN MSE = 0.0639; Ridge MSE = 0.0663; Bridge MSE = 0.0678; Bridge lowest MAE/MedianAE | Ensemble mitigates small residual errors; LSTM captures nonlinear temporal dynamics; trade-off observed between global accuracy and local error stability. |
| 5 th | Radiomics-only $R^2 = -0.1167$; Genomics-only $R^2 = 1.0$; Multimodal $R^2 = 0.0154$ | Radiomics contributes weakly to survival prediction; genomics exhibits dominant predictive strength; naïve multimodal fusion diminishes integrated performance, underscoring the need for optimized fusion design. |

Furthermore, the outcomes emphasize that multimodal integration demands carefully optimized fusion strategies. As seen in Test Case 5, naïve averaging of radiomic and genomic predictions failed to exploit cross-modal complementarities, resulting in diminished performance compared to unimodal genomic modeling. Consequently, adaptive fusion methods, such as attention-weighted aggregation or learned stacking, are recommended to fully harness the complementary predictive capacity of diverse biomedical data sources.

9.3 Practical Validation and Clinical Relevance

Table 9 provides an integrative summary of the clinical validation strategies and translational implications of the predictive models evaluated across the test cases. Although phantom-based validation, including 3D dose visualization and synthetic DVHs, remains a practical surrogate for preliminary evaluation, the absence of real patient-specific dosimetry represents a significant limitation to clinical applicability. While visual inspections and residual analyses enhance interpretability by allowing intuitive correlation between predicted and actual outcomes, such methods remain insufficient substitutes for prospective clinical trials or multi-institutional validation.

Models incorporating ensemble learning and temporal architectures (e.g., LSTM and Bridge frameworks) demonstrated notable reproducibility and predictive consistency, supporting their use as decision-support tools in radiotherapy planning. Similarly, multimodal learning pipelines, particularly those integrating radiomic and genomic information, exhibit high translational promise by enabling biologically informed survival predictions. However, their clinical utility remains contingent upon the development of robust fusion strategies, improved feature harmonization, and integration of real clinical dose data.

Collectively, the comparative validation analysis underscores both the progress achieved in model interpretability and reproducibility and the persistent challenges that must be addressed for these frameworks to attain full clinical deployment readiness in CyberKnife-guided precision oncology.

Table 9. Overview of validation strategies and translational implications of predictive models for CyberKnife clinical applications, highlighting current validation practices, interpretability, and the need for real patient-specific integration.

| Test Case | Practical Validation | Clinical Implications |
|-----------------|--|---|
| 1 st | Phantom 3D dose visualizations; DVH placeholders | LSTM captures spatial-temporal dose patterns; Ridge provides interpretable baseline; supports initial treatment verification though extreme dose peaks remain underestimated. |
| 2 nd | Synthetic DVH and voxel-based dosimetry; tabular prediction evaluation | XGBoost reliable for structured clinical features; enables survival outcome prediction but DVH results are not patient-specific, limiting clinical realism. |
| 3 rd | Synthetic 3D dose phantom; residual heatmaps; ensemble scatter plots | Bridge ensemble exhibits strong generalization and robustness; suitable for clinical decision-support prototyping but requires multi-institutional validation. |
| 4 th | 3D dose phantom visualization; scatter plots assessing bridge ensemble stability | LSTM and ensemble frameworks enhance alignment with observed outcomes; reproducible results indicate potential for high-fidelity clinical deployment. |
| 5 th | Cross-comparison of radiomic, genomic, and multimodal predictions | Genomics provides dominant predictive accuracy; radiomics adds marginal value; emphasizes the importance of optimized multimodal fusion for clinical translation. |

9.4 Comprehensive Overview of Predictive Modeling Approaches: Performance, Limitations, and Research Opportunities

Table 10 provides a consolidated overview of the most effective modeling approaches across key methodological dimensions, including linear and nonlinear learning, temporal and tabular data processing, ensemble

robustness, multimodal integration, and clinical visualization strategies. The comparative synthesis identifies both the performance strengths and the existing methodological gaps, while outlining targeted research opportunities to advance predictive modeling for CyberKnife dose and survival prediction frameworks.

Ridge regression consistently emerges as a reliable linear baseline and an effective meta-learner within ensemble frameworks, providing interpretability and numerical stability despite its inherent inability to capture nonlinear or temporal dependencies. XGBoost demonstrates excellent capability for nonlinear tabular modeling, achieving high predictive accuracy, though moderate overfitting remains an issue that can be mitigated through enhanced regularization and cross-validation. LSTM networks, optimized for temporal data, effectively capture sequential dependencies in dynamic treatment and survival data but exhibit limited generalization when trained on small or heterogeneous datasets, indicating the need for hybrid architectures and attention-based enhancements to balance interpretability and performance.

Ensemble strategies, particularly Bridge stacking, substantially improve model robustness and stability by integrating complementary learning paradigms. However, their increased computational complexity and slightly elevated global errors highlight opportunities for fine-tuned meta-learning and adaptive fusion optimization. Multimodal pipelines, exemplified by the radiogenomic integration framework in Case 5, reveal substantial translational promise for personalized medicine applications, though the use of naïve averaging for late fusion restricts the exploitation of inter-modality complementarities. This limitation emphasizes the importance of implementing attention-guided fusion, weighted stacking, or transformer-based alignment for maximizing multimodal synergy.

Clinical visualization tools such as 3D phantom DVHs, residual heatmaps, and scatter plots contribute valuable interpretive insights, facilitating the translation of model predictions into clinically understandable patterns. Nonetheless, the reliance on synthetic data remains a major limitation, as these methods fail to capture extreme dose peaks and inter-patient variability observable in real-world dosimetry. Integrating real patient-specific DVH data and expanding validation across multiple institutions are therefore critical steps toward establishing clinically reliable digital twin models.

A cross-case synthesis reveals discernible trends in modeling efficacy and interpretability. LSTM architectures (Cases 1 and 4) excel in representing temporal sequences, while XGBoost (Cases 2 and 3) remains optimal for structured clinical datasets. The Bridge ensemble consistently delivers high generalization and error stability across diverse data conditions. However, the simple averaging approach in multimodal fusion (Case 5) fails to realize the potential of heterogeneous data integration. Ridge regression provides dependable but limited linear baselines, and feedforward neural networks (MLPs) display susceptibility to underfitting or overfitting depending on dataset size and hyperparameter tuning. Visualization-based assessments improve interpretability but fall short of clinical validation, as none of the test cases incorporate real patient dosimetry.

Identified research gaps include the limited generalizability of current datasets, ineffective modeling of extreme dose variations, and the absence of optimized multimodal fusion mechanisms. Furthermore, XAI integration remains an open challenge, crucial for enhancing the interpretability of deep learning architectures in clinical environments. In conclusion, the analytical progression observed across the five test cases demonstrates a clear methodological evolution, from simple Ridge–LSTM hybrid baselines (Case 1) to sophisticated Bridge ensemble and

multimodal radiogenomic frameworks (Cases 3–5). Although practical validation through synthetic 3D visualizations and regression metrics confirms reproducibility and modular design, advancing these frameworks toward clinical translation requires addressing the identified research challenges. By integrating real patient data, optimizing fusion strategies, and embedding explainability mechanisms, future work can significantly strengthen the reliability, transparency, and deployment potential of digital twin–based CyberKnife predictive systems in precision radiotherapy.

Table 10. Comparative assessment of predictive modeling approaches highlighting performance, limitations, and research opportunities for CyberKnife dose and survival prediction.

| Aspect | Best Performing Approach | Observed Limitation | Research Opportunity |
|-----------------------------|--------------------------|---|---|
| Linear modeling | Ridge regression | Inability to capture nonlinear or temporal dependencies | Use as a baseline or Bridge meta-learner |
| Nonlinear tabular modeling | XGBoost | Moderate overfitting | Apply regularization and cross-validation |
| Nonlinear temporal modeling | LSTM | Limited generalization with small datasets | Develop hybrid or attention-based architectures |
| Ensemble robustness | Bridge stacking | Slightly higher global errors and training complexity | Optimize meta-learning and adaptive fusion |
| Multimodal integration | Radiogenomics (Case 5) | Naïve averaging reduces cross-modal synergy | Implement attention-based or weighted fusion strategies |
| Clinical visualization | 3D phantom DVH | Synthetic outputs fail to capture extreme dose peaks | Integrate real patient-specific dosimetry |

9.5 Consolidated Comparative Analysis of Machine Learning and Deep Learning Frameworks for CyberKnife Dosimetry Prediction and Multimodal Integration

The consolidated comparative analysis presented in Table 11 provides a structured overview of the performance, validation, and translational readiness of the evaluated ML and DL frameworks applied to CyberKnife dosimetry prediction and multimodal radiogenomic integration. Each test case represents a progressive enhancement in model complexity, fusion strategy, and predictive robustness.

In Test Case 1 (Ridge–LSTM Hybrid), the model exhibited weak predictive performance ($R^2 = 0.096$; MAE = 902.63), demonstrating limited generalization capability despite preliminary validation through phantom-based 3D dose visualization. This configuration underscores the need for multi-institutional validation and improved modeling of extreme dose variations.

Test Case 2 (Multimodal ML: Ridge + MLP + XGBoost) achieved moderate improvement ($R^2 = 0.4104$), benefiting from nonlinear feature capture and tabular data robustness. However, deficiencies in hyperparameter optimization and cross-validation design constrained its generalizability, emphasizing the need for systematic tuning and feature harmonization.

In Test Case 3 (Bridge Ensemble), ensemble integration of Ridge, MLP, and XGBoost attained the highest predictive performance ($R^2 = 0.99964$) with enhanced generalization and reduced residual variance. Although validated using synthetic 3D dose phantoms and residual heatmaps, further efforts are required to extend this success to real patient dosimetry and multimodal data fusion contexts.

Test Case 4 (Hybrid ML + DL with Bridge) demonstrated strong performance (MSE = 0.0620) through the integration of Ridge regression, Neural Networks, and LSTM. The model effectively captured nonlinear temporal

dependencies and improved residual stability, with validation via scatter plots and phantom-based visualizations. Nevertheless, external validation and optimized fusion techniques remain essential for broader clinical adoption.

Lastly, Test Case 5 (Multimodal Radiogenomics Ridge) explored the feasibility of integrating radiomics and genomics data for survival prediction in non-small cell lung cancer (NSCLC). Despite its methodological novelty, the model achieved only marginal predictive performance ($R^2 = 0.0154$) due to the naïve averaging fusion of unimodal Ridge regressors and the limited discriminative power of radiomic features.

Overall, the comparative analysis highlights the trade-offs among model complexity, predictive precision, and clinical interpretability. Ensemble-based methods (Case 3) achieve the most accurate and stable predictions, while hybrid ML–DL architectures (Case 4) enhance interpretability and temporal modeling. Conversely, multimodal frameworks (Cases 2 and 5) exhibit high translational potential but require more sophisticated fusion mechanisms, larger datasets, and rigorous external validation to realize clinical impact. Future research should thus prioritize cross-institutional generalization, explainable AI integration, and data fusion optimization to bridge the gap between algorithmic performance and real-world clinical deployment in precision radiotherapy.

Table 11. Consolidated comparison matrix of CyberKnife predictive frameworks, summarizing model architectures, key performance metrics, validation strategies, strengths, and research gaps across different ML and DL configurations.

| Test Case | Models / Architecture | Key Metrics | Clinical Validation | Strengths | Research Gaps |
|-----------------|---|------------------------------|---|--|--|
| 1 st | Ridge regression, LSTM | $R^2 = 0.096$; MAE = 902.63 | Phantom 3D dose visualization | Captures linear–temporal dependencies; interpretable baseline | Requires multi-institutional validation; limited modeling of extreme dose values |
| 2 nd | Ridge regression, MLP, XGBoost | $R^2 = 0.4104$ | Synthetic DVH voxel-based dosimetry | Robust for tabular data; effective nonlinear capture | Insufficient hyperparameter optimization; needs cross-validation and fusion refinement |
| 3 rd | Ridge, MLP, XGBoost (Bridge stacking) | $R^2 = 0.99964$ | Synthetic 3D dose phantom; residual heatmaps | High accuracy and residual minimization; superior generalization | Lack of real patient validation; limited multimodal integration |
| 4 th | Ridge, Neural Network, LSTM, Bridge | MSE = 0.0620 | 3D dose phantom visualizations; scatter plots | Captures nonlinear temporal patterns; reproducible performance | Requires external validation; advanced fusion optimization needed |
| 5 th | Ridge regression (Radiomics + Genomics) | $R^2 = 0.0154$ | Not clinically validated | Demonstrates feasibility of multimodal integration | Naïve averaging reduces multimodal gain; radiomics weak; limited dataset diversity |

This consolidated summary establishes a coherent comparison across all evaluated frameworks, highlighting both methodological maturity and areas for targeted research innovation in digital twin–based CyberKnife prediction systems.

The comparative assessment of the five CyberKnife predictive frameworks reveals a distinct evolution in methodological sophistication, from baseline Ridge–LSTM hybrids to advanced bridge ensembles and multimodal radiogenomic pipelines. Ensemble-based architectures demonstrate the highest predictive accuracy and robustness, whereas hybrid ML–DL frameworks provide enhanced interpretability for temporal and nonlinear dose modeling. Despite these advances, clinical validation remains largely synthetic, and real patient dosimetry integration has yet to be achieved. The findings underscore the necessity of cross-institutional data generalization, optimized multimodal fusion strategies, and the incorporation of XAI for enhancing transparency and clinician trust. Future work should

thus focus on developing scalable, interpretable, and patient-specific predictive systems, advancing toward a digital twin paradigm that unifies radiomic, genomic, and dosimetric data for adaptive CyberKnife treatment optimization.

10. Conclusion and Future Directions

This study presented a comprehensive evaluation of advanced predictive modeling frameworks for CyberKnife treatment planning and survival prediction in NSCLC. Five progressively complex test cases were developed to assess linear, nonlinear, temporal, ensemble, and multimodal learning paradigms, emphasizing the interplay between model interpretability, predictive performance, and clinical applicability. The investigation began with a Ridge–LSTM hybrid model, which demonstrated the feasibility of combining linear regression with temporal deep learning to capture both linear and sequential dependencies within clinical datasets. Subsequent analyses using Ridge Regression, MLP, and XGBoost models highlighted the influence of data modality on model selection, revealing that tabular models perform robustly on structured datasets, while temporal and ensemble architectures capture complex nonlinear relationships more effectively.

The integration of ensemble learning through Bridge stacking substantially enhanced prediction stability, reduced residual variance, and improved generalization across heterogeneous data, establishing a reproducible foundation for digital twin–based radiotherapy modeling. The framework’s modular design, encompassing standardized preprocessing, multi-metric evaluation, and structured data export, ensured transparency, reproducibility, and adaptability to future data sources. Visualization tools such as 3D phantom dose mapping and DVHs provided interpretable insight into spatial dose distributions, linking computational outputs to clinically relevant evaluations. The multimodal radiogenomic framework further demonstrated the potential of integrating imaging and genomic data for survival prediction, although the naïve averaging fusion employed yielded limited multimodal synergy. Across all test cases, deep learning and ensemble-based models, particularly LSTM and Bridge architectures, consistently achieved superior predictive accuracy and robustness compared with single-model baselines.

Despite these advancements, several research gaps were identified that delineate clear directions for future work. The reliance on a single dataset (NSCLC-Radiomics-Lung1, 422 patients) restricted generalizability, emphasizing the necessity of multi-institutional and cross-population validation. The simple averaging strategy used for multimodal fusion in Case 5 proved insufficient; therefore, attention-based fusion mechanisms, weighted ensemble stacking, and transformer-driven multimodal networks should be explored to exploit inter-modality dependencies more effectively. The observed limitations of LSTM and XGBoost in modeling extreme dose peaks highlight the need for quantile regression frameworks and attention-enhanced hybrid architectures to better capture rare but clinically significant outliers.

Most validations relied on synthetic dosimetric representations, including phantom-based DVHs and voxel-level placeholders. Future research should integrate real patient dose–volume data to enhance clinical validity and translational reliability. While Ridge and Bridge models provided partial interpretability, the black-box nature of deep learning frameworks remains a challenge; incorporating XAI methods such as SHAP, LIME, and saliency mapping will improve transparency and clinician trust. Temporal and sequential data handling, attempted through LSTM in selected cases, can be expanded using transformer or graph-temporal architectures to model longitudinal treatment evolution. Furthermore, the integration of multi-omics, imaging, and clinical data within unified pipelines, supported

by systematic k-fold cross-validation and automated hyperparameter optimization, represents a key pathway toward achieving robust, generalizable, and clinically meaningful predictive systems.

In conclusion, this research established a reproducible, extensible, and clinically interpretable foundation for digital twin-based CyberKnife predictive modeling. The findings demonstrated that hybrid, ensemble, and multimodal learning strategies can substantially enhance dose prediction accuracy, survival estimation, and treatment personalization in radiotherapy. The progressive evolution across the five test cases, from linear regression baselines to radiogenomic integration, illustrated a clear trajectory toward more sophisticated and interpretable clinical decision-support systems. By bridging computational rigor with translational utility, the proposed framework advances the state of predictive analytics in precision oncology, offering a scalable and clinically actionable pathway for future patient-specific radiotherapy planning and outcome prediction.

Declarations

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding: The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/36989).

Acknowledgments: The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/36989). The authors would like to thank the Automated Systems and Computing Lab (ASCL) at Prince Sultan University for their providing software and hardware tools to complete this work.

Data Availability: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Authors' Contributions : Eman S. Sabry contributed to the conceptualization of the study, development of the methodology, and preparation of the original manuscript draft. Ehab Mahmoud Mohamed provided technical supervision, refined the methodology, validated the experimental results, and critically reviewed and revised the manuscript, and served as the corresponding author. Walid El-Shafai supervised the overall research process, performed data analysis and validation, managed project administration, and finalized the manuscript editing.

References

1. W. Kilby, J. R. Dooley, G. Kuduvali, S. Sayeh, and C. R. Maurer Jr., "The CyberKnife robotic radiosurgery system in 2010," *Technology in Cancer Research & Treatment*, vol. 9, no. 4, pp. 433–452, 2010.
2. European Commission, "Toymodel of a radiotherapy machine," 2025.
3. M. W. Glaessgen and D. S. Stargel, "The digital twin paradigm for future NASA and U.S. Air Force vehicles," in *Proc. AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*, San Diego, CA, USA, 2012.
4. C.-W. Chang, S. S. Akkineni, M. Hu, K. Shah, Y. Gao, P. Patel, and X. Yang, "A digital twin framework for adaptive treatment planning in radiotherapy," *arXiv preprint arXiv:2506.14701*, 2025.
5. A. Chaudhuri, G. Pash, D. A. Hormuth, G. Lorenzo, M. Kapteyn, C. Wu, and K. Willcox, "Predictive digital twin for optimizing patient-specific radiotherapy regimens under uncertainty in high-grade gliomas," *Frontiers in Artificial Intelligence*, vol. 6, p. 1222612, 2023.
6. Y. Jiang, C. Gao, Y. Shao, X. Lou, M. Hua, J. Lin, and C. Gao, "The prognostic value of radiogenomics using CT in patients with lung cancer: a systematic review," *Insights into Imaging*, vol. 15, no. 1, p. 259, 2024.

7. R. Deng, N. Shaikh, G. Shannon, and Y. Nie, "Cross-modality attention-based multimodal fusion for non-small cell lung cancer (NSCLC) patient survival prediction," in *Medical Imaging 2024: Digital and Computational Pathology*, vol. 12933, pp. 46–50, SPIE, Apr. 2024.
8. L. Strigari, J. Schwarz, T. Bradshaw, J. Brosch-Lenz, G. Currie, G. El-Fakhri, and B. Saboury, "Computational nuclear oncology toward precision radiopharmaceutical therapies: ethical, regulatory, and socioeconomic dimensions of theranostic digital twins," *Journal of Nuclear Medicine*, vol. 66, no. 5, pp. 748–756, 2025.
9. J. Adler, S. Chang, J. Murphy, and S. Hancock, "The CyberKnife: A frameless robotic system for radiosurgery," *Stereotactic and Functional Neurosurgery*, vol. 69, no. 1–4, pp. 124–128, 1997.
10. J. R. Adler and S. L. Hancock, "The CyberKnife system in stereotactic radiotherapy," *Surgical Neurology*, vol. 47, no. 1, pp. 40–46, 1997.
11. M. Hoogeman, et al., "Clinical accuracy of the CyberKnife system in stereotactic radiotherapy," *Acta Oncologica*, vol. 55, no. 7, pp. 825–832, 2016.
12. Y. Zhang, et al., "Automated radiotherapy treatment planning using deep reinforcement learning," *Medical Physics*, vol. 47, no. 10, pp. 5138–5146, 2020.
13. P. Barragán-Montero, et al., "Artificial intelligence for treatment planning in radiation therapy," *Physics in Medicine & Biology*, vol. 65, no. 21, 2020.
14. M. C. Kearney, et al., "Dose prediction using deep neural networks: Generalization challenges and solutions," *Radiotherapy and Oncology*, vol. 164, pp. 63–72, 2021.
15. Y. Fan, et al., "Geometry-aware neural networks for dose distribution prediction," *Medical Physics*, vol. 48, no. 8, pp. 4677–4689, 2021.
16. T. Babier, et al., "Knowledge-based automated planning with deep learning and physics-based QA," *Medical Physics*, vol. 47, no. 5, pp. e202–e214, 2020.
17. D. R. Cox, "Regression models and life tables," *Journal of the Royal Statistical Society: Series B*, vol. 34, no. 2, pp. 187–220, 1972.
18. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'16)*, pp. 785–794, 2016.
19. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *arXiv preprint arXiv:1603.02754*, 2016.
20. R. Huynh, A. Coroller, R. Narayan, et al., "Associations of radiomic data extracted from static and respiratory-gated CT images with outcomes in lung cancer patients," *Radiotherapy and Oncology*, vol. 122, no. 2, pp. 255–262, 2017.
21. C. Barragán-Montero, et al., "Artificial intelligence and machine learning in radiotherapy: A literature review," *Physica Medica*, vol. 83, pp. 9–19, 2021.
22. S. Isaksson, "Respiratory motion prediction using recurrent neural networks," *Medical Physics*, vol. 45, no. 6, pp. 2413–2423, 2018.
23. J. Krauss, et al., "LSTM-based forecasting of respiratory surrogates for radiotherapy gating," *Physics in Medicine & Biology*, vol. 65, no. 22, 2020.
24. N. Kerkmeijer, et al., "Real-time motion management in MR-guided radiotherapy: Current state and challenges," *The Lancet Oncology*, vol. 22, no. 8, pp. e379–e389, 2021.
25. F. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, 2016.
26. L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.
27. A. Lundberg and S.-I. Lee, "A unified approach to model interpretation for tree-based methods (SHAP)," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017.
28. D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
29. H. Aerts, et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, p. 4006, 2014.
30. E. Chalkidou, et al., "Radiogenomics in cancer: Principles and applications," *Nature Reviews Clinical Oncology*, vol. 19, pp. 287–302, 2022.
31. Y. Kickingereder, et al., "Automated radiomic profiling with machine learning: Challenges and opportunities," *Radiology*, vol. 290, no. 1, pp. 70–78, 2019.

32. L. Sun, et al., "Multimodal learning for radiogenomics: Fusing imaging and molecular data," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1714–1728, 2022.
33. P. Lambin, et al., "Radiomics and radiogenomics in precision medicine," *Nature Reviews Clinical Oncology*, vol. 20, pp. 1–19, 2023.
34. P. Lambin, et al., "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.
35. D. Craft, T. Bortfeld, J. Halabi, and H. Shih, "DVH-based treatment plan evaluation: Are current metrics sufficient for plan comparison and ranking?" *International Journal of Radiation Oncology • Biology • Physics*, vol. 83, no. 5, pp. e837–e842, 2012.
36. D. Bruynzeel and B. Lagerwaard, "Adaptive radiotherapy: Toward a patient digital twin," *Seminars in Radiation Oncology*, vol. 29, no. 3, pp. 223–232, 2019.
37. A. Bertsimas, et al., "Real-time, data-driven adaptive radiotherapy," *Clinical Cancer Research*, vol. 25, no. 19, pp. 5794–5802, 2019.
38. T. Heins, et al., "Digital twins for precision oncology: Concepts and roadmaps," *NPJ Precision Oncology*, vol. 8, 2024.
39. T. M. Deist, et al., "From models to digital twins in radiation oncology," *Frontiers in Oncology*, vol. 12, 2022.
40. H. Lin, et al., "Towards real-time respiratory motion prediction based on long short-term memory neural networks," *arXiv preprint arXiv:1901.08638*, 2019.
41. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 1050–1059, 2016.
42. M. Kapteyn, et al., "TumorTwin: A Python framework for patient-specific digital twins in oncology," *arXiv preprint arXiv:2505.00670*, 2025.
43. H. G. Kang, et al., "Medical digital twins in oncology: A scoping review," *Frontiers in Oncology*, vol. 14, pp. 1–14, 2024.
44. Y. Li, et al., "Application of a 3D volumetric display for radiation therapy treatment planning I: Quality assurance procedures," *Journal of Applied Clinical Medical Physics*, vol. 19, no. 1, pp. 170–179, 2018.
45. S. Kim, et al., "Development of patient-specific conformal 3D-printed devices for dose verification in radiotherapy," *Applied Sciences*, vol. 11, no. 6, p. 2794, 2021.
46. H. Chen, et al., "Multimodal data fusion in radiotherapy outcome prediction: A review," *Frontiers in Oncology*, vol. 12, pp. 1–12, 2022.
47. NEMA DICOM Standards Committee, "DICOM standards: Radiotherapy objects." Available online: <https://www.dicomstandard.org> (accessed Nov. 8, 2025).
48. M. D. Wilkinson, et al., "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, pp. 1–9, 2016.
49. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, New York, 2009.
50. The Cancer Imaging Archive (TCIA), "NSCLC-Radiomics (Lung1) collection." Available online: <https://www.cancerimagingarchive.net/collection/nsclc-radiomics/> (accessed Nov. 8, 2025).
51. C. Parmar, et al., "Radiomic machine-learning classifiers for prognostic biomarkers of lung cancer," *Frontiers in Oncology*, vol. 5, no. 272, pp. 1–10, 2015.
52. S. Haarburger, et al., "Radiomics feature reproducibility and stability in NSCLC Lung1 dataset across delineation variations," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.
53. M. Ferreira, et al., "Machine learning for lung cancer histology classification using NSCLC-Radiomics data," *Journal of Medical Imaging*, vol. 7, no. 4, pp. 1–10, 2020.
54. X. Chen, et al., "Interpretable machine learning models for NSCLC subtype prediction on Lung1 radiomics," *Cancers*, vol. 15, no. 2, pp. 1–12, 2023.
55. American Cancer Society, "Key statistics for lung cancer." Available online: <https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html> (accessed Nov. 8, 2025).

56. National Cancer Institute (NCI), “Non-small cell lung cancer treatment (PDQ®)—Patient version,” 2023. Available online: <https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq> (accessed Nov. 8, 2025).
57. R. S. Herbst, et al., “Lung cancer,” *New England Journal of Medicine*, vol. 359, no. 13, pp. 1367–1380, 2008.
58. D. S. Ettinger, et al., “Non–small cell lung cancer, version 3.2022, NCCN clinical practice guidelines in oncology,” *Journal of the National Comprehensive Cancer Network*, vol. 20, no. 5, pp. 497–530, 2022.
59. R. Timmerman, et al., “Stereotactic body radiation therapy for inoperable early stage lung cancer,” *JAMA*, vol. 303, no. 11, pp. 1070–1076, 2010.
60. A. Mok, et al., “Personalized medicine in lung cancer: Molecular pathways and targeted therapies,” *Clinical Cancer Research*, vol. 19, no. 9, pp. 2307–2318, 2013.
61. The Cancer Imaging Archive (TCIA), “NSCLC-Radiomics (Lung1–3) collections.” Available online: <https://www.cancerimagingarchive.net/> (accessed Nov. 8, 2025).
62. The Cancer Imaging Archive (TCIA), “NSCLC-Radiomics (Lung2 & Lung3) collections.” Available online: <https://www.cancerimagingarchive.net/> (accessed Nov. 8, 2025).