Leaf-Specific Classification of Multi-Leaf Collimator Positioning Errors in Volumetric Modulated Arc Therapy Using a Convolutional Neural Network

Ju Yeol Shin , Chang Heon Choi , Jung-In Kim , Jong Min Park , Wonjoong Cheon *,† , So-Yeon Park *,†

Posted Date: 21 May 2026

doi: 10.20944/preprints202605.1394.v1

Keywords: multi-leaf collimator; quality assurance; deep learning; convolutional neural network; fluence map; volumetric modulated arc therapy; MLC error classification

 doi:10.20944/preprints202605.1394.v1

Article

Leaf-Specific Classification of Multi-Leaf Collimator Positioning Errors in Volumetric Modulated Arc Therapy Using a Convolutional Neural Network

Ju Yeol Shin [1,2], Chang Heon Choi [1,2], Jung-In Kim [1,2], Jong Min Park [2], Wonjoong Cheon [3,†,*] and So Yeon Park [4,5,†,*]

[1] Paprica Lab. Co., Ltd, Seoul 03123, Republic of Korea
[2] Department of Radiation Oncology, Seoul National University Hospital, Seoul 03080, Republic of Korea
[3] Department of Radiation Oncology, Seoul St. Mary's Hospital, College of Medicine, the Catholic University of Korea, Seoul 06591, Republic of Korea
[4] Department of Radiation Oncology, Veterans Health Service Medical Center, Seoul 05368, Republic of Korea
[5] Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul 03080, Republic of Korea
* Correspondence: wjcheon@catholic.ac.kr (W.C.); vsoyounv@gmail.com (S.Y.P.)
† These authors have contributed equally to this work.

Abstract

Background/Objectives: Multi-leaf collimator (MLC) positioning accuracy critically affects delivered dose fidelity in volumetric modulated arc therapy (VMAT), yet conventional gamma-based quality assurance (QA) provides only plan-level pass/fail outcomes without leaf-specific error localization. This study developed and validated a convolutional neural network (CNN) framework that classifies the magnitude and direction of individual MLC leaf positioning errors directly from fluence map data. **Methods**: Three patient cohorts were analyzed: 20 prostate cancer patients for model development under an 8:1:1 train/validation/test split, and 20 additional prostate and 10 head and neck (H&N) patients reserved for external validation. For the inner MLC leaves 21–40, systematic offsets from −5 mm to +5 mm in 1.0 mm increments were independently applied to the two leaf banks, yielding 121 error combinations per leaf. A CNN was trained as a 121-class classifier on two-channel inputs pairing the reference and error-induced fluence map regions, and was compared against three tree-based baselines using five-fold cross-validation. **Results**: The CNN achieved 97.21% accuracy on the internal test set and 96.54 ± 0.43% accuracy across the five cross-validation folds, significantly outperforming all three baseline models. Over 94% of predictions in both leaf banks fell within 1.0 mm of the true offset, consistent with the AAPM TG-142 MLC positioning tolerance. External validation yielded 96.19% accuracy on the additional prostate cohort and 93.72% on the H&N cohort, supporting both reproducibility within the same treatment site and generalizability across anatomically distinct sites. **Conclusions**: The proposed CNN framework enables leaf-specific identification of MLC positioning errors in both magnitude and direction, moving patient-specific QA beyond plan-level pass/fail assessment toward actionable error localization.

Keywords: multi-leaf collimator; quality assurance; deep learning; convolutional neural network; fluence map; volumetric modulated arc therapy; MLC error classification

1. Introduction

With advancements in radiation therapy technology, techniques such as Volumetric Modulated Arc Therapy (VMAT) and Intensity-Modulated Radiation Therapy (IMRT) have been developed and widely adopted. VMAT, in particular, is a well-established radiation therapy modality that

 doi:10.20944/preprints202605.1394.v1

modulates the multi-leaf collimator (MLC), dose rate, and gantry speed to deliver high doses to the tumor while sparing surrounding healthy tissues [1,2]. However, the dynamic and simultaneous modulation of these parameters inherently introduces greater delivery uncertainty compared to conventional techniques.

Of the mechanical factors governing VMAT delivery — including gantry rotation, dose rate, and MLC positioning — errors in MLC leaf positions have been widely recognized as having a particularly significant impact on delivered dose accuracy, since the MLC directly collimates the radiation beam to define the shape and intensity of the radiation field. Even submillimeter deviations can lead to substantial discrepancies between planned and delivered doses, compromising treatment efficacy and increasing the risk of adverse effects on healthy tissues. Such errors are especially critical in VMAT, where MLC leaves must continuously reposition in synchronization with gantry rotation and dose rate modulation, leaving minimal tolerance for positioning deviations [3–8]. To mitigate these uncertainties, rigorous quality assurance (QA) procedures have become indispensable, with patient-specific delivery QA serving as a critical step to verify that the measured dose distribution matches the treatment planning system (TPS) calculation [3,9].

For the evaluation of delivery QA results, gamma analysis is the most widely used method, evaluating agreement between measured and planned dose distributions by calculating dose differences and distance to agreement within defined criteria. However, gamma analysis alone is insufficient for comprehensive error localization: it lacks sensitivity to localized delivery errors, reduces complex spatial dose discrepancies to a single passing rate that is ultimately interpreted as a pass/fail outcome against a predefined threshold, and falls short of identifying which specific MLC leaves are misaligned or characterizing the magnitude of individual leaf deviations [8–10].

As complementary approaches, modulation indices and texture-based indices derived from modulating parameters and fluence maps have been developed to quantify treatment plan complexity and predict delivery accuracy [6,7,11,12]. Additionally, indices of achievement, which measure the alignment between planned and delivered fluences, have shown promise in improving QA reliability [13]. However, these approaches remain indirect — they predict the likelihood of delivery errors at the plan level but do not localize errors to individual MLC leaves or quantify the magnitude and direction of specific leaf deviations.

Recent advances in deep learning have further enhanced patient-specific QA. Nyflot et al. [14], and Wootton et al. [15] applied CNNs and radiomic analysis to gamma images to detect systematic and random MLC errors in IMRT, while Kimura et al. [16] extended this approach to VMAT using cylindrical detector measurements with a multi-task CNN capable of classifying multiple error types simultaneously. Nakamura et al. [17] further demonstrated that deep learning applied to MLC modeling parameters — specifically transmission factor and dosimetric leaf gap — could distinguish error types from dose difference maps with high sensitivity [17]. Beyond dose-based inputs, trajectory log-file data have also been leveraged; Carlson et al. [18], Osman et al. [19], and Chuang et al. [20] used machine learning to predict MLC positional deviations from such log-file parameters, yet log-file readings reflect only the machine's recorded motor positions and may not reveal fluence-level discrepancies when motor encoders report nominal values. This limitation can be mitigated by fluence map–based approaches that reflect the actually delivered radiation distribution.

Despite these advances, existing approaches share a fundamental shortcoming in resolving leaf-specific delivery errors: dose- and gamma-based methods predominantly operate at the plan level through aggregate metrics or their pass/fail dichotomization, while log-file–based methods, though leaf-specific, capture only mechanical positions rather than the resulting delivered fluence [10]. Furthermore, fluence maps — which directly encode the cumulative radiation intensity shaped by MLC positions across all control points — remain largely unexplored as a primary input for leaf-specific error classification. To bridge this methodological gap, this study develops a CNN-based framework that directly analyzes fluence map data to classify MLC leaf positioning errors at the individual leaf level, simultaneously identifying both the magnitude and direction of deviations across a range of −5 mm to +5 mm as a 121-class classification problem. This level of leaf-specific error

 doi:10.20944/preprints202605.1394.v1

resolution has not been addressed in prior fluence-based QA studies. This leaf-specific, multi-class approach moves beyond plan-level pass/fail assessments toward actionable error localization. Its generalizability is further evaluated across anatomically distinct treatment sites to support broader clinical translation.

2. Methods

2.1. Study Design and Patient Cohorts

This retrospective study (approved by the Institutional Review Board of the Veterans Health Service Medical Center, approval number: 2022-01-002-005) analyzed VMAT treatment planning data from three independent patient cohorts (Table 1): a primary prostate cohort (n = 20) for model development under an 80%/10%/10% training/validation/test split, an additional prostate cohort (n = 20) for external validation within the same anatomical site, and a head and neck (H&N) cohort (n = 10) for cross-site generalizability assessment.

All treatment plans comprised two full arcs and were delivered using the Varian High Definition (HD) MLC system (Varian Medical Systems, Palo Alto, CA), which consists of 60 leaf pairs — each pair comprising two opposing leaves from Bank A and Bank B — with leaf widths of 5 mm for outer leaves (leaves 1–14 and 47–60) and 2.5 mm for inner leaves (leaves 15–46) at the isocenter. A summary of the study design is provided in Table 1.

The overall workflow of the proposed framework — encompassing data preparation, fluence map generation, MLC error simulation, model training, and performance evaluation — is summarized in Figure 1. Each methodological component is described in detail in the following sections.

Table 1. Overview of the study design and dataset composition.

Cohort	Patients	Purpose	Data Split	Samples per Leaf	Total Samples
Prostate (primary)	20	Model development	Training 80% / Validation 10% / Test 10%	2,420	48,400
Prostate (additional)	20	External validation (same site)	All external test	2,420	48,400
Head & Neck	10	External validation (cross-site)	All external test	1,210	24,200

Samples per leaf = 121 error combinations × number of patients. Total samples = samples per leaf × 20 inner leaves (leaves 21–40). Details of error simulation and leaf selection are provided in subsequent sections.

 doi:10.20944/preprints202605.1394.v1

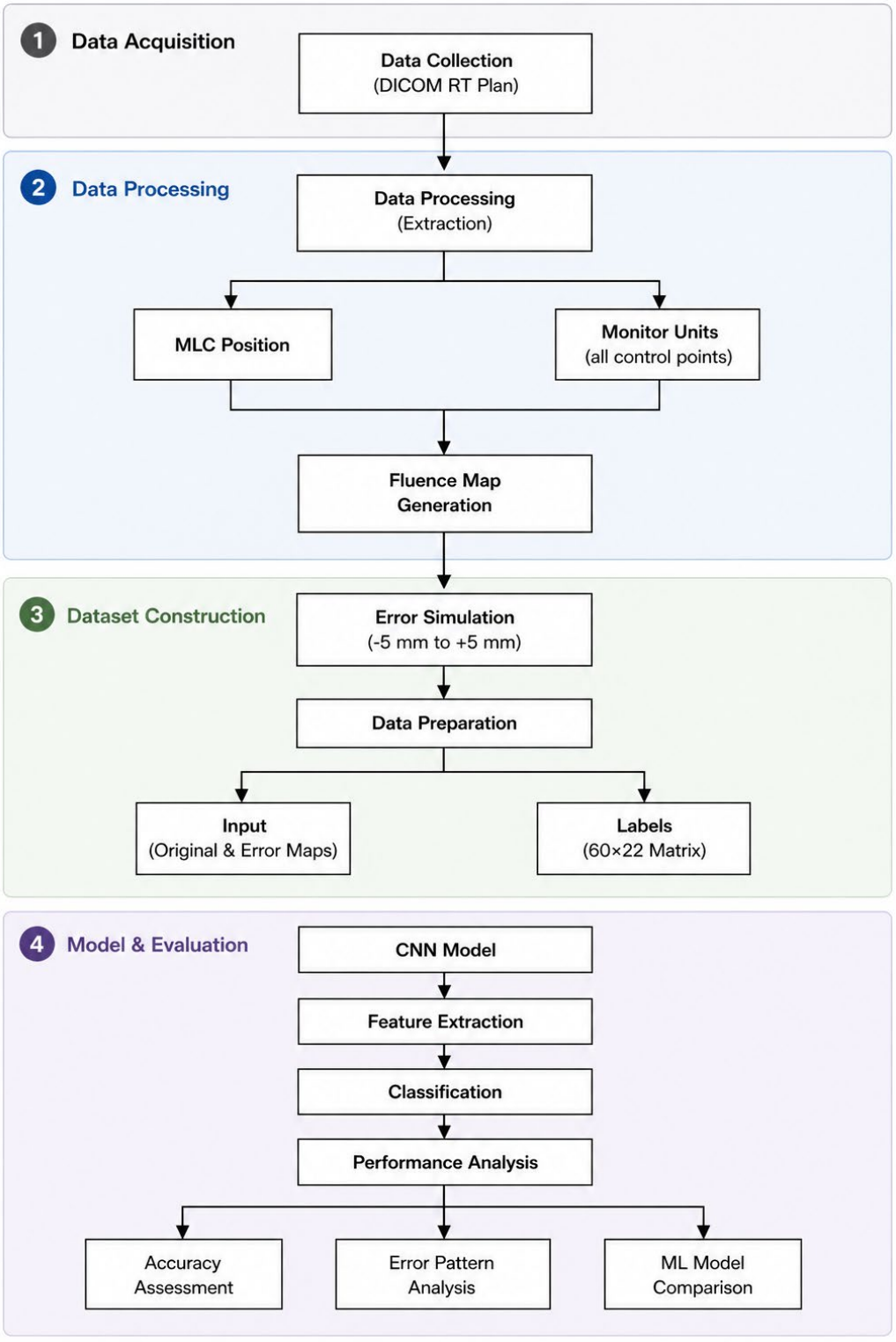

Figure 1. Overview of the Data Processing and Deep Learning Model Workflow for MLC Error Detection.

2.2. Fluence Map Generation and ROI Extraction

For each treatment plan, fluence maps were generated from the Digital Imaging and Communications in Medicine (DICOM) RT plan file through a three-step process: (1) extraction of MLC positions and monitor units (MUs) at all control points; (2) processing of leaf positions from both MLC banks; and (3) accumulation of MLC-shaped beam contributions weighted by the corresponding MUs to produce the composite fluence map. The resulting maps represented the

 doi:10.20944/preprints202605.1394.v1

cumulative radiation intensity distribution delivered over the entire arc, at a spatial resolution of approximately 1.0 mm per pixel.

From each fluence map, reconstructed at the isocenter plane, a leaf-specific region of interest (ROI) was extracted to serve as input for the deep learning model. Analysis was restricted to inner MLC leaves 21–40, which shape the central treatment field in central-target VMAT plans and represent the most clinically relevant region for positioning error detection. To represent each leaf consistently, an ROI of three vertical pixels by 400 horizontal pixels was defined. At the isocenter level, the inner leaves have a projected width of 2.5 mm; the three-pixel vertical extent was chosen to fully encompass this leaf width while providing a uniform input size across all leaves, and the 400-pixel lateral extent covered the full range of leaf travel within the treatment field.

For model input, the ROI from the reference fluence map and the ROI from the corresponding error-induced fluence map were stacked to form a two-channel tensor of shape 3 × 400 × 2. This paired representation enabled the model to learn from the direct comparison between reference and error-induced fluence distributions rather than from a single fluence map in isolation. Figure 2 illustrates the fluence map generation and ROI extraction process, showing the reference and error-induced composite fluence maps together with the extracted ROI for MLC leaf #30.

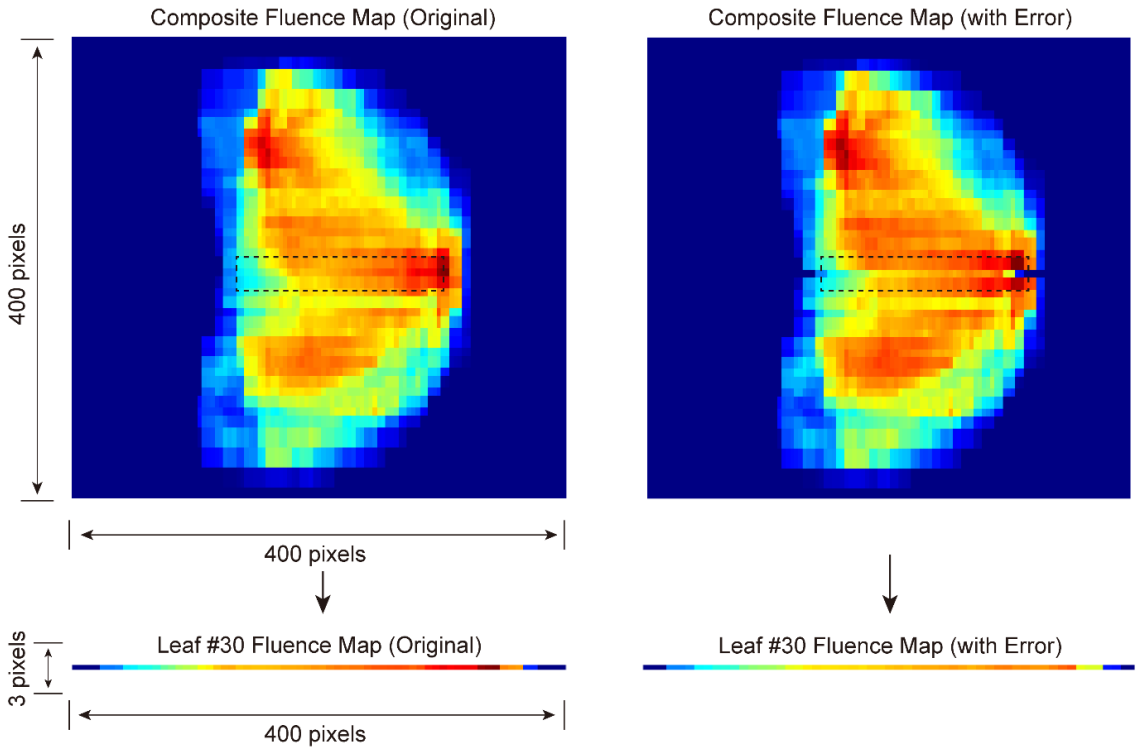

Figure 2. Fluence map generation and leaf-specific ROI extraction. (Upper) Original and error-induced composite fluence maps. (Lower) Extracted ROIs for MLC leaf #30, consisting of three vertical pixels and 400 horizontal pixels. These segments are paired and stacked to form a model input of size 3 × 400 × 2.

2.3. MLC Error Simulation and Label Encoding

Error simulation was applied to the inner MLC leaves (leaves 21–40), which shape the central treatment field in central-target VMAT plans and represent the most clinically relevant region for positioning error detection. Systematic offsets ranging from −5 mm to +5 mm in 1.0 mm increments were independently applied to Bank A and Bank B, yielding 11 discrete error states per bank and 121 unique error combinations (11 × 11) per leaf. This design ensured uniform representation of all possible error states across the 121-class classification task, with each combination treated as a distinct class so that both the magnitude and direction of deviations in the two banks could be identified simultaneously.

 doi:10.20944/preprints202605.1394.v1

For each sample, the ground truth was encoded as a categorical label corresponding to one of the 121 error combinations, which directly mapped to the 121-unit softmax output of the CNN, enabling end-to-end multi-class classification. An illustrative example is shown in Figure 3: when a +3 mm offset is applied to Bank A and a −1.0 mm offset to Bank B, the corresponding combination is encoded as the class index assigned to this pairing within the 121-class scheme.

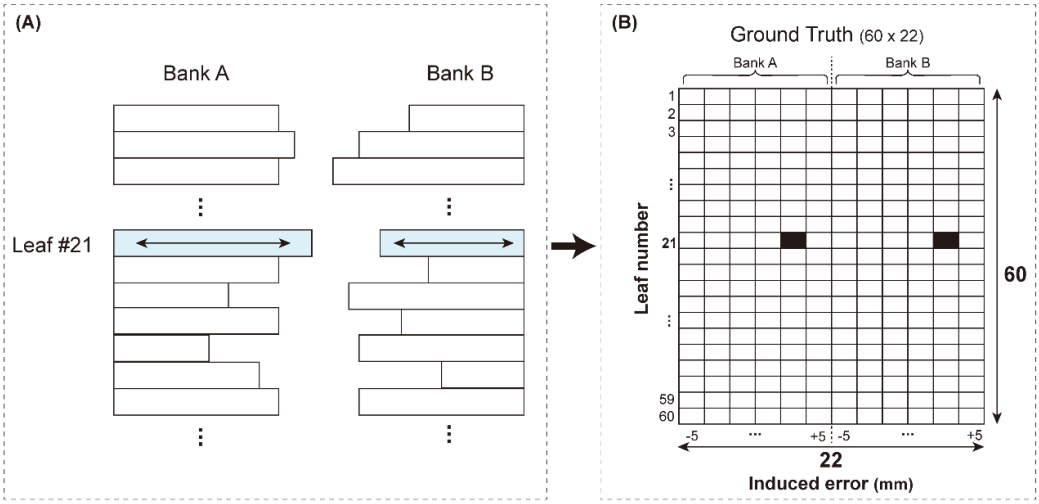

Figure 3. MLC error simulation and 121-class label encoding. (A) Systematic offsets from −5 mm to +5 mm in 1.0 mm increments are applied independently to Bank A and Bank B, producing 121 unique error combinations (11 × 11) per leaf. (B) The resulting error combination is encoded into a corresponding location within a two-dimensional label matrix, forming a 121-class representation (11 × 11). The vertical axis of the matrix represents the 60 MLC leaves, while the horizontal axis represents the 22 possible error combinations generated from independent offsets applied to Bank A and Bank B. The illustrated example shows Bank A shifted by +3 mm and Bank B by −1.0 mm, mapped to a specific class index in the encoding scheme. The grid is schematic and not drawn to the actual scale of 60 × 22.

2.4. Deep Learning Model Architecture and Optimization

A CNN architecture was developed to classify MLC positioning errors from the paired fluence map ROI defined in the preceding sections. The network received an input tensor of shape 3 × 400 × 2, where the two channels corresponded to the reference and error-induced fluence map segments, enabling the model to learn from the direct comparison between the two states rather than from a single fluence distribution. The architecture is illustrated in Figure 4.

The feature extraction backbone consisted of six convolutional layers organized into three sequential blocks, each block comprising two convolutional layers with a kernel size of 1 × 3 followed by a max pooling layer with pool size 1 × 4 to progressively reduce the lateral dimension. The number of filters increased with depth and was determined through hyperparameter optimization, with block 1 containing 16 and 32 filters, block 2 containing 64 and 64 filters, and block 3 containing 128 and 128 filters. All convolutional layers used ReLU activation and "same" padding to preserve spatial dimensions, and batch normalization and dropout were applied after each convolutional layer to stabilize training and reduce overfitting. The output of the final block was flattened and passed through a dense layer with 512 units and ReLU activation, followed by the final output layer of 121 units with softmax activation, corresponding to the 121 error combinations defined in the MLC Error Simulation and Label Encoding section.

 doi:10.20944/preprints202605.1394.v1

Model hyperparameters were optimized using Keras Tuner with a RandomSearch strategy. The number of convolutional blocks was fixed, while the number of convolutional layers within each block (ranging from 1 to 3) and the learning rate (1×10^{-2}, 1×10^{-3}, and 1×10^{-4}) were explored as hyperparameters during optimization over 15 trials. The architecture described above represents the optimized configuration selected through this search.

The model was trained with the Adam optimizer using categorical cross-entropy as the loss function and a batch size of 8, for a maximum of 20 epochs. To mitigate overfitting, early stopping was applied with a patience of 10 epochs based on validation loss. All experiments were implemented in TensorFlow 2.10.0 and executed on an NVIDIA GeForce RTX 4090 GPU with CUDA Toolkit 11.8.

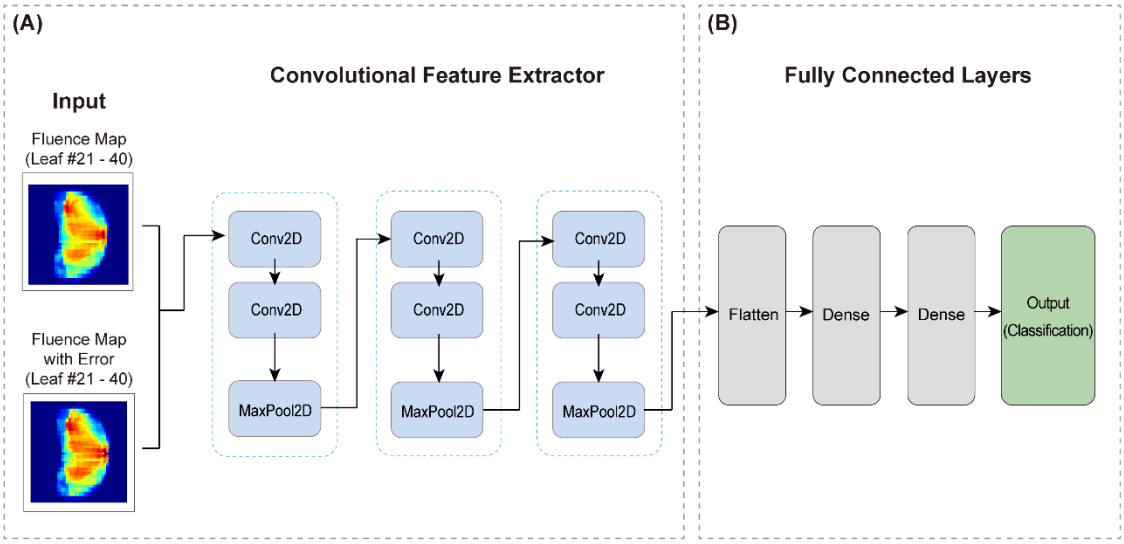

Figure 4. CNN architecture for MLC error detection and classification. (A) Feature extraction using convolutional and pooling layers. (B) Classification using fully connected layers.

2.5. Implementation of Traditional Machine Learning Models for Comparison

Three tree-based machine learning models — Random Forest, XGBoost, and CatBoost — were implemented as baselines for comparison with the CNN. All three are widely used classifiers for tabular data and were selected to provide a representative benchmark from traditional machine learning methods. To ensure a fair comparison, the same two-channel fluence map ROI used as CNN input was flattened into a 2,400-dimensional feature vector ($3 \times 400 \times 2$) and provided to each baseline model.

The Random Forest classifier was configured with 500 trees and a maximum depth of 10, with bootstrap aggregation and feature randomization applied to reduce overfitting. XGBoost was trained with 500 boosting iterations and a learning rate of 0.1, and CatBoost was implemented with parameters matched to those of XGBoost for consistency, leveraging its ordered boosting approach to reduce target leakage. Hyperparameters for all three models were selected through grid search. For the two gradient boosting models, L1 and L2 regularization on leaf weights were applied to stabilize training, and early stopping based on validation performance was employed across all three models to prevent overfitting.

2.6. Performance Evaluation

Model performance was evaluated using four standard classification metrics: accuracy, precision, recall, and F1-score. Given the 121-class nature of the task, precision, recall, and F1-score were computed using a one-vs-rest strategy and then macro-averaged across all 121 classes, ensuring equal weighting of each class regardless of its frequency in the dataset. In this formulation, true

 doi:10.20944/preprints202605.1394.v1

positives (TP) denote correctly identified error combinations, true negatives (TN) denote correctly rejected non-target classes, false positives (FP) denote incorrectly predicted error combinations, and false negatives (FN) denote missed error combinations.

The four metrics were defined as follows:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 - score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

To further characterize classification reliability beyond aggregate metrics, a prediction mismatch analysis was performed. For each misclassified sample, the absolute deviation between the predicted and true offset values was computed independently for Bank A and Bank B across the full range of −5 mm to +5 mm. This analysis quantified not only whether misclassifications occurred, but also how far the predicted error magnitudes deviated from the ground truth, providing a clinically interpretable measure of model reliability.

3. Statistical Methods

To compare CNN performance against each of the three baseline machine learning models (Random Forest, XGBoost, and CatBoost), paired t-tests were performed on the fold-wise accuracies obtained from the five-fold cross-validation procedure. To control the family-wise error rate across the three pairwise comparisons, the Holm–Bonferroni correction was applied, with statistical significance defined as a corrected p-value below 0.05. Effect sizes were quantified using Cohen's d to assess the practical magnitude of the observed performance differences alongside statistical significance. All analyses were conducted in Python 3.8.18 using SciPy 1.10.1 and scikit-learn 1.3.0.

4. Results

4.1. Model Training Performance

On the internal test set of the primary prostate cohort, the CNN model achieved a test accuracy of 97.21%, with precision of 97.14%, recall of 97.06%, and F1-score of 97.04% (Table 2). The balanced performance across these four metrics indicates that the model did not exhibit systematic bias toward any particular error class.

Table 20. epochs, the model reached a training accuracy of 97.10% and a validation accuracy of 96.81%, with close alignment between training and validation curves for both accuracy and loss (Figure 5), suggesting that the model generalized well without overfitting.

Table 2. Model performance metrics on the internal test set (primary prostate cohort).

Metric	Value (%)
Accuracy	97.21
Precision	97.14
Recall	97.06
F1-score	97.04

 doi:10.20944/preprints202605.1394.v1

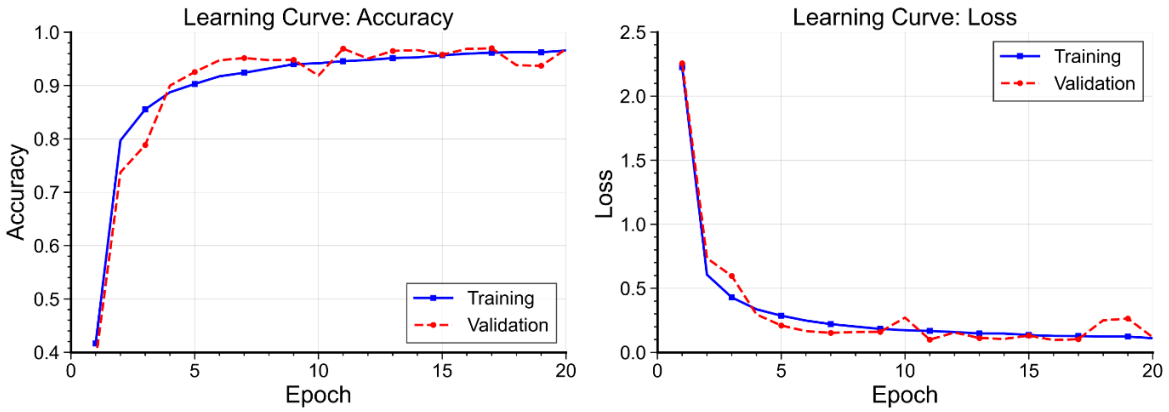

Figure 5. Model learning curves over 20 epochs. (Left) Accuracy and (Right) loss curves for training and validation sets. The close alignment between training and validation curves indicates robust model generalization without overfitting.

4.2. Cross-Validation Assessment

Five-fold cross-validation confirmed model stability across different data partitions. The average cross-validation accuracy was 96.54 ± 0.43%, and the average loss was 0.1083 (Figure 6). The narrow standard deviation of 0.43 percentage points across all five folds indicates that the reported performance was not sensitive to the specific choice of training–validation split, supporting the reliability of the internal test results.

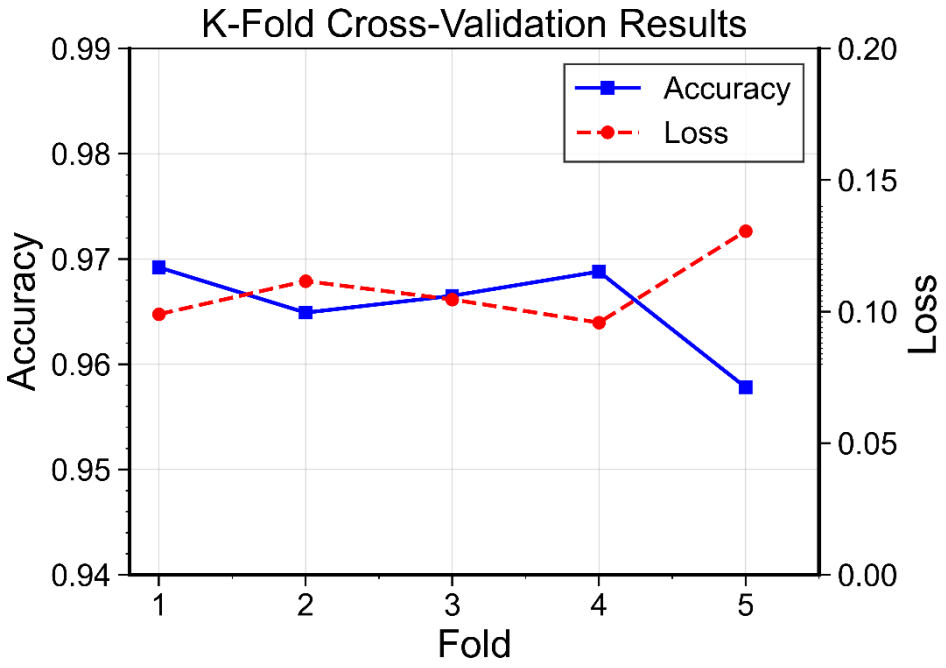

Figure 6. Five-fold cross-validation results for accuracy and loss.

4.3. Comparative Analysis with Traditional Machine Learning

The CNN was compared against three tree-based baseline models — Random Forest, XGBoost, and CatBoost — trained and evaluated on identical data splits. The results revealed a substantial generalization gap in the baseline models (Table 3). XGBoost and CatBoost achieved near-perfect training accuracies of 99.92% and 99.60%, respectively, but their test accuracies dropped to 69.86% and 63.95%. Random Forest showed the most pronounced generalization failure, with training

 doi:10.20944/preprints202605.1394.v1

accuracy of 94.29% falling to 22.00% on the test set. In contrast, the CNN maintained nearly identical accuracies on training (97.10%) and test (97.21%) sets, with a gap of only 0.11 percentage points.

Statistical comparison further confirmed the superiority of the CNN. Paired t-tests on fold-wise accuracies from the five-fold cross-validation yielded highly significant differences between the CNN and each baseline (CNN vs XGBoost: $p = 2.3 \times 10^{-7}$; CNN vs Random Forest: $p = 9.6 \times 10^{-9}$; CNN vs CatBoost: $p = 1.4 \times 10^{-7}$), with all corrected p-values remaining below 0.001 after Holm–Bonferroni correction (Table 4). Cohen's d values of 44.8, 99.6, and 43.0 for the three comparisons were classified as extremely large effects.

Table 3. Training and test accuracies of the CNN and baseline machine learning models.

Model Type	Training Accuracy (%)	Test Accuracy (%)
CNN	97.10	97.21
XGBoost	99.92	69.86
Random Forest	94.29	22.00
CatBoost	99.60	63.95

Table 4. Statistical comparison of the CNN and baseline machine learning models.

Comparison	p-value (paired t-test)	Corrected p-value (Holm-Bonferroni)	Effect Size (Cohen's d)	Interpretation
CNN vs XGBoost	2.3×10^{-7}	< 0.001	44.8	Extremely large effect
CNN vs Random Forest	9.6×10^{-9}	< 0.001	99.6	Extremely large effect
CNN vs CatBoost	1.4×10^{-7}	< 0.001	43.0	Extremely large effect

4.4. Model Reliability Assessment

To characterize the precision of individual predictions beyond aggregate classification metrics, a deviation analysis was performed on the internal test set. Of the 4,840 test samples, 145 (3.00%) were misclassified. For these misclassified samples, the mean deviation magnitudes remained small — 0.52 mm for Bank A and 0.72 mm for Bank B — although the maximum deviation reached 5 mm in both banks (Table 5).

The distribution of deviation magnitudes further illustrates the model's precision (Table 6). For Bank A, 95.9% of predictions fell within a 1.0 mm deviation of the true offset (55.1% exact matches and 40.7% within 1.0 mm), while 94.5% of Bank B predictions fell within the same threshold (35.9% exact matches and 58.6% within 1.0 mm). These results are consistent with the 1.0 mm MLC positioning tolerance recommended by AAPM TG-142 [21].

Table 5. Overall model reliability on the internal test set.

Metric	Value
Total samples	4,840
Misclassified samples	145
Misclassification rate	3.00 %
Mean Bank A deviation magnitude	0.52 mm
Mean Bank B deviation magnitude	0.72 mm
Maximum Bank A deviation magnitude	5 mm
Maximum Bank B deviation magnitude	5 mm

Table 6. Distribution of deviation magnitudes by MLC bank.

 doi:10.20944/preprints202605.1394.v1

Deviation magnitude	Bank A (%)	Bank B (%)
0 mm	55.1	35.9
1.0 mm	40.7	58.6
2 mm	2.1	4.1
3 mm	1.4	0.7
4 mm	0.0	0.0
5 mm	0.7	0.7

4.5. External Validation

To assess the generalizability of the CNN beyond the primary prostate cohort, the trained model was evaluated on two independent external datasets: an additional prostate cohort of 20 patients and a head and neck (H&N) cohort of 10 patients. Both datasets were generated using the same error simulation and sampling procedure described in Methods, and neither was used in any stage of model development.

On the additional prostate cohort, which yielded 48,400 samples (2,420 per leaf × 20 inner leaves), the model achieved an accuracy of 96.19%, with precision of 96.17%, recall of 96.79%, and F1-score of 96.11% (Table 7). These results were nearly identical to those obtained on the internal test set (97.21% accuracy), with a difference of approximately 1 percentage point, indicating that model performance was reproducible across independent patient cohorts within the same treatment site.

On the H&N cohort, which yielded 24,200 samples (1,210 per leaf × 20 inner leaves), the model achieved an accuracy of 93.72%, with precision of 94.34%, recall of 93.71%, and F1-score of 93.78% (Table 8). Although the accuracy decreased by approximately 3.5 percentage points compared to the internal prostate test set — an expected reduction given the anatomical and geometric differences between the two treatment sites — the model retained high classification performance, supporting its applicability across anatomically distinct sites.

Table 7. Model performance on the additional prostate cohort.

Metric	Value (%)
Accuracy	96.19
Precision	96.17
Recall	96.79
F1-score	96.11

Table 8. Model performance on the head and neck cohort.

Metric	Value (%)
Accuracy	93.72
Precision	94.34
Recall	93.71
F1-score	93.78

5. Discussion

This study developed a CNN-based framework for leaf-specific classification of MLC positioning errors directly from fluence map data. On the internal test set of the primary prostate cohort, the model achieved an accuracy of 97.21% across the 121-class problem, with more than 94% of predictions falling within 1.0 mm of the true offset in both MLC banks — consistent with the 1.0 mm positioning tolerance recommended by AAPM TG-142. Performance was maintained on an independent prostate cohort (96.19% accuracy) and on a head and neck cohort (93.72% accuracy), supporting both the reproducibility of the model within the same treatment site and its generalizability across anatomically distinct sites.

 doi:10.20944/preprints202605.1394.v1

Previous studies have explored deep learning for MLC error detection, yet most have operated at the plan level rather than identifying individual leaf deviations. Nyflot et al. [14], and Wootton et al. [15] applied CNNs and radiomic analysis to gamma images to detect systematic and random MLC errors in IMRT, but their framework relied on gamma-based inputs and produced aggregate classifications rather than leaf-level localization. Kimura et al. [16] extended deep learning to VMAT using cylindrical detector measurements with a multi-task CNN, yet their approach also aggregated predictions at the plan level. Nakamura et al. focused on MLC modeling parameters — leaf transmission and dosimetric leaf gap — derived from dose difference maps, demonstrating high sensitivity for classifying modeling-related errors but not individual positioning deviations [17]. In contrast, the present study directly analyzes fluence map data and classifies both the magnitude and direction of positioning errors at the individual leaf level, addressing a gap that these prior works have not specifically targeted.

The methodological novelty of this work lies in three aspects. First, the problem was reformulated as a 121-class classification task, enabling simultaneous identification of both the magnitude and direction of leaf deviations — a formulation that, to our knowledge, has not been addressed in prior fluence-based QA studies. Second, the two-channel input representation, pairing the reference and error-induced fluence maps, allowed the model to learn from the direct comparison between the two states rather than from a single fluence distribution in isolation. This paired representation likely contributes to the model's ability to detect subtle delivery discrepancies that machine log data — which reflect only the recorded motor positions — may not capture, even when mechanical readings appear nominal. Third, the CNN consistently outperformed the three tree-based baselines (Random Forest, XGBoost, CatBoost), with corrected p-values below 0.001 and Cohen's d values of 43.0–99.6 for the three comparisons. The numerical magnitude of these effect sizes should be interpreted in the context of the very low fold-wise variance observed in the five-fold cross-validation (SD = 0.43 percentage points). Such tight variance, combined with the substantial accuracy gap, produced effect sizes that arithmetically exceed conventional reference ranges. While this does not imply that the accuracy difference is unprecedentedly large in a clinical sense, it does underscore that the performance advantage of the CNN was highly consistent across folds rather than driven by a few favorable splits.

The capability to identify the magnitude and direction of individual leaf deviations has direct implications for patient-specific QA. Current gamma-based workflows reduce complex spatial discrepancies to a single passing rate and a pass/fail judgment, providing little guidance on which leaves to investigate when a plan fails QA. In contrast, a leaf-level error map produced by the proposed framework could help physicists rapidly localize the source of delivery discrepancies, prioritize corrective actions, and — in principle — inform targeted re-calibration of specific MLC leaves. The precision within 1.0 mm observed in over 94% of predictions aligns with the AAPM TG-142 MLC positioning tolerance [21], suggesting that the method operates at a spatial resolution relevant to clinical action thresholds. Although the approach is currently limited to post-delivery analysis of fluence maps, its leaf-level granularity positions it as a complementary tool to existing QA methods rather than a replacement for them.

Several limitations should be acknowledged. First, fluence maps were generated from DICOM RT plans rather than derived from physical measurements; while this enabled systematic generation of all 121 error states under controlled conditions — a balance difficult to achieve from limited clinical datasets — validation with measured fluence data (e.g., EPID-based measurements) remains necessary before clinical translation. Second, analysis was restricted to the inner MLC leaves (leaves 21–40), which are the most active region in central-target VMAT plans; extending the framework to outer leaves is methodologically feasible but was not evaluated in this study. Third, each training sample contained a deviation applied to a single leaf, and the model was not explicitly trained on simultaneous multi-leaf error patterns; although leaf-wise scanning could, in principle, detect multiple simultaneous deviations by processing each leaf independently, this capability was not empirically verified. Fourth, the study focused solely on MLC positioning errors without considering

 doi:10.20944/preprints202605.1394.v1

other delivery uncertainties such as gantry angle, dose rate, and other mechanical parameters, which may interact with MLC errors in practice. Finally, the 1.0 mm granularity of the simulated error states, while aligned with the TG-142 tolerance, does not resolve sub-millimeter deviations that may still carry dosimetric significance in high-precision treatments.

As an immediate next step, the model should be validated against physically measured fluence data, such as EPID-acquired maps under controlled MLC perturbations, to confirm that the classification performance observed with simulated errors translates to actual delivery conditions. Evaluating the framework on multi-institutional datasets with different MLC systems would further clarify whether the learned features are specific to the Varian HD configuration or transferable across platforms. On the modeling side, training on simultaneous multi-leaf error patterns and incorporating additional delivery parameters (e.g., gantry angle and dose rate variations) are needed to better reflect the complexity of real-world delivery deviations. In the longer term, coupling this approach with real-time EPID-based in vivo dosimetry could enable online error detection during treatment delivery.

6. Conclusions

This study demonstrates that a CNN trained on paired fluence map representations can classify MLC positioning errors at the individual leaf level, resolving both the magnitude and direction of deviations within a 1.0 mm range. By providing leaf-specific, actionable information rather than plan-level pass/fail judgments, the proposed framework offers a complementary direction for patient-specific QA and a foundation for future validation with measured delivery data.

Author Contributions: Conceptualization, SYP, WJC; Project administration, SYP; Resources, SYP; Supervision, CHC, JIK, JMP, SYP, WJC; Programming, JYS, WJC; Writing of original draft, JYS; Writing of the review and editing, JYS, SYP, WJC; Data analysis, JYS, SYP, WJC.

Funding: This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00253604) and by the Technology Innovation Program (RS-2025-02221011, Development of Medical-Specialized Multimodal Hyperscale Generative AI Technology for Global Integration) funded by the Ministry of Trade, Industry & Energy (MOTIE, South Korea).

Institutional Review Board Statement: The study received approval from the Institutional Review Board of Veterans Health Service Medical Center (No. 2022-01-002-005). The requirement for informed consent was waived due to the retrospective nature of the study and the use of anonymized treatment planning data.

Informed Consent Statement: Not applicable.

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

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

AI: Artificial intelligence; CNN: Convolutional neural network; DL: Deep learning; IMRT: Intensity-modulated radiation therapy; ML: Machine learning; MLC: Multi-leaf collimator;
MU: Monitor unit; QA: Quality assurance; ROI: Region of interest; TPS: Treatment planning system; VMAT: Volumetric modulated arc therapy

References

1. Otto, K., Volumetric modulated arc therapy: IMRT in a single gantry arc. Med Phys, 2008. **35**(1): p. 310–7.
2. Brahme, A., Optimization of stationary and moving beam radiation therapy techniques. Radiother Oncol, 1988. **12**(2): p. 129–40.

 doi:10.20944/preprints202605.1394.v1

3. Ezzell, G.A., et al., Guidance document on delivery, treatment planning, and clinical implementation of IMRT: report of the IMRT Subcommittee of the AAPM Radiation Therapy Committee. Med Phys, 2003. **30**(8): p. 2089–115.
4. Fredh, A., et al., Patient QA systems for rotational radiation therapy: a comparative experimental study with intentional errors. Med Phys, 2013. **40**(3): p. 031716.
5. Heo, T., et al., The effect of beam interruption during VMAT delivery on the delivered dose distribution. Phys Med, 2015. **31**(3): p. 297–300.
6. Park, J.M., S.Y. Park, and H. Kim, Modulation index for VMAT considering both mechanical and dose calculation uncertainties. Phys Med Biol, 2015. **60**(18): p. 7101–25.
7. Park, S.Y., et al., Evaluation of the plan delivery accuracy of intensity-modulated radiation therapy by texture analysis using fluence maps. Phys Med, 2019. **59**: p. 64–74.
8. Heilemann, G., B. Poppe, and W. Laub, On the sensitivity of common gamma-index evaluation methods to MLC misalignments in Rapidarc quality assurance. Med Phys, 2013. **40**(3): p. 031702.
9. Kim, J.I., et al., The sensitivity of gamma-index method to the positioning errors of high-definition MLC in patient-specific VMAT QA for SBRT. Radiat Oncol, 2014. **9**: p. 167.
10. Yan, G., et al., On the sensitivity of patient-specific IMRT QA to MLC positioning errors. J Appl Clin Med Phys, 2009. **10**(1): p. 120–128.
11. Park, J.M., et al., *Modulation indices for volumetric modulated arc therapy.* Phys Med Biol, 2014. **59**(23): p. 7315–40.
12. Park, S.Y., et al., Textural feature calculated from segmental fluences as a modulation index for VMAT. Phys Med, 2015. **31**(8): p. 981–990.
13. Kim, D.S., et al., To propose adding index of achievement (IOA) to IMRT QA process. Radiat Oncol, 2018. **13**(1): p. 112.
14. Nyflot, M.J., et al., Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks. Med Phys, 2019. **46**(2): p. 456–464.
15. Wootton, L.S., et al., Error Detection in Intensity-Modulated Radiation Therapy Quality Assurance Using Radiomic Analysis of Gamma Distributions. Int J Radiat Oncol Biol Phys, 2018. **102**(1): p. 219–228.
16. Kimura, Y., et al., Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy. Phys Med, 2020. **73**: p. 57–64.
17. Nakamura, S., et al., Deep learning-based detection and classification of multi-leaf collimator modeling errors in volumetric modulated radiation therapy. J Appl Clin Med Phys, 2023. **24**(12): p. e14136.
18. Carlson, J.N., et al., A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. Phys Med Biol, 2016. **61**(6): p. 2514–31.
19. Osman, A.F.I., N.M. Maalej, and K. Jayesh, Prediction of the individual multileaf collimator positional deviations during dynamic IMRT delivery priori with artificial neural network. Med Phys, 2020. **47**(4): p. 1421–1430.
20. Chuang, K.C., W. Giles, and J. Adamson, A tool for patient-specific prediction of delivery discrepancies in machine parameters using trajectory log files. Med Phys, 2021. **48**(3): p. 978–990.
21. Klein, E.E., et al., Task Group 142 report: quality assurance of medical accelerators. Med Phys, 2009. **36**(9): p. 4197–212.