Iranian Journal of Medical Physics

ijmp.mums.ac.ir

A Monte Carlo-Enhanced Deep Learning Framework for Predicting Dosimetric Suitability of Epithermal Neutron Beams in BNCT

Adnan H. Abdulkadhim[1*]

1. Department of Physics, College of Science, University of Sumer, Thi-Qar, Iraq

ARTICLE INFO

Article type:
Original Paper

Keywords:
Boron Neutron Capture Therapy (BNCT)
Monte Carlo simulation (MCNP6.2)
Multi-scale channel attention CNN
Support vector machine (SVM)
Beam Shaping Assembly (BSA)

ABSTRACT

Introduction: The Boron Neutron Capture Therapy (BNCT) requires exact characterization of epithermal neutron beams to ensure safe and effective delivery. Traditional evaluation via Monte Carlo simulations is computationally prohibitive, requiring 24-48 hours per Beam Shaping Assembly (BSA) configuration. This study introduces a novel hybrid methodology combining Monte Carlo simulations with a deep learning classifier to rapidly predict the dosimetric suitability of cyclotron-based epithermal neutron beams.

Material and Methods: We optimized a BSA for (15 MeV, 200μA) proton-beryllium source using MCNP6.2 achieving an epithermal neutron flux of 2.1×10^{8} $n.cm^{-2}.s^{-1}$. Depth-dose distributions in an anthropomorphic head phantom were simulated and converted into synthetic MRI-style images (256×256 pixels 8-bit grayscale) using global min-max normalization across the entire dataset to preserve absolute flux information. A novel Multi-scale Channel Attention CNN integrated with a Support Vector Machine (MCACNN-SVM) was trained on 5000 simulated configurations to classify beams as "therapeutically suitable" (Advantage Depth ≥ 5.0 cm and estimated treatment time 60 minutes) or "unsuitable".

Results: The optimized of (BSA) met all IAEA recommendations fast neutron contamination (D_f/Φ_{epil}) = 0.78×10^{-13} Gy·cm², gamma contamination $D_\gamma/\Phi_{epil} = 0.93\times10^{-13}$ Gy·cm² thermal-to-epithermal ratio = 0.043. The MCACNN-SVM classifier achieved 99.1% accuracy, 98.9% precision, 99.3% recall, 99.1% F1-score and a macro-average AUC of 0.998 on the held-out test set (n = 750). Inference time was 11.7 milliseconds per configuration on an NVIDIA RTX 4060 Ti GPU representing a speedup factor of approximately 11.7 million compared to full MC simulations (38 hours).

Conclusion: This is the first demonstration that deep learning models trained on MC-generated synthetic images can predict BNCT dosimetric outcomes with near-perfect accuracy. The proposed framework enables real-time BSA optimization and beam quality assurance significantly accelerating the development of accelerator-based BNCT systems.

►*Please cite this article as:*
Abdulkadhim AH. A Monte Carlo-Enhanced Deep Learning Framework for Predicting Dosimetric Suitability of Epithermal Neutron Beams in BNCT. Iran J Med Phys 2026; 23 (3): 142-154. 10.22038/ijmp.2026.96414.2711.

Introduction

Boron Neutron Capture Therapy (BNCT) is a binary radiotherapy that achieves selective tumor cell destruction done by the $^{10}B(n,\alpha)^{7}Li$ reaction. The resulting high Linear Energy Transfer (LET) particles (α-particle: 150 KeV/μm) have ranges of (5-9 μm) confining biological damage to individual boron-loaded cells while sparing adjacent normal tissue [1].

This selectivity renders BNCT especially effective against invasive tumors such as glioblastoma multiforme and recurrent head and neck cancers. For deep brain tumors epithermal neutrons (0.5 eV-10 kV) are ideal as they penetrate tissues to a depth of 5-7 cm and heat only in the vicinity of boron loaded tumor cells [2,3]. Importantly the neutron beam's energy spectrum and spatial distribution directly influence key performance indicators of BNCT including depth of penetration (AD) treatment ratio (TR) and expected treatment duration.

Problem Statement

Accelerator-based neutron sources (ABNS) such as the Cyclotron 30LC have emerged as practical hospital friendly alternatives to reactor based systems [4]. However optimizing the Beam Shaping Assembly (BSA) which moderates fast neutrons reflects neutrons toward the beam port and removes gamma contamination requires iterative computationally expensive Monte Carlo simulations (e.g. MCNP6.2). A single full dosimetric evaluation of one (BSA) configuration requires 24-48 hours of computation time severely constraining the design space that can be explored [5].

*Corresponding Author: Tel: +9647829519595; Email: adnanhashem2000@gmail.com

Recent Advances in Deep Learning for Medical Imaging

Recent progresses in medical image analysis has established that Convolutional Neural Networks (CNNs) incorporating attention mechanisms can classify brain tumors from MRI with accuracy more than 98% [6,7]. Specifically:

- Multi-domain CNNs capture hierarchical spatial features across receptive fields of varying sizes (3×3 5×5 7×7) similar to radiologists examining images at multiple magnifications [8].
- Channel attention architectures (e.g. Squeeze and Excitation Networks SENet) dynamically recalibrate feature responses by modeling interdependences between channels enhancing discriminability [9].
- Support Vector Machines (SVMs) act as a final classifier to optimize decision boundaries in high dimensional feature spaces providing maximum margin separation which is particularly useful when training data are limited [10].

Knowledge Gap and Research Objectives

Despite these advancements no study has yet been conducted to apply deep learning to predict the accuracy of boron neutron therapy (BNCT) beam dose measurements directly from simulated neutron flux maps. Such a framework could reduce evaluation time from days to fractions of a second enabling real time Beam Shaping Assembly (BSA) optimization and ensuring beam quality.

The aims of this study are:

1. Enhance a BSA for a 15 MeV 200 μA cyclotron proton beryllium source using MCNP6.2.
2. Simulating depth-dose distributions and thermal neutron flux in a realistic human head model.
3. Converting 2D thermal neutron flux maps into synthetic images similar to MRI scans suitable for deep learning.
4. Develop and validate an (MCACNN-SVM) classifier to predict whether a given beam meets stringent therapeutic criteria (AD ≥ 5.0 cm, And treatment time ≤ 60 minutes).
5. Compare performance against state-of-the-art architectures and provide ablation studies.

This work represents the first integration of Monte Carlo neutron transport with Deep Learning for BNCT beam quality assessment.

Materials and Methods

Overview of the Hybrid Framework

The proposed framework consists of three stages as shown in Figure 1.

Stage 1:- Monte Carlo Simulation MCNP6.2 simulations of the BSA and anthropomorphic head phantom to generate thermal neutron flux distributions (F4 tallies) and full depth dose profiles.

Stage 2:- Synthetic Image Generation Conversion of 2D thermal flux maps into grayscale synthetic images (256×256 pixels 8-bit depth) mimicking T1-weighted contrast enhanced MRI.

Stage 3:- Classification: Training an MCACNN-SVM classifier to categorize beams as "suitable" AD ≥ 5.0 cm AND treatment time ≤ 60 minutes or "unsuitable" (failing either criterion).

Monte Carlo Simulation of the BNCT Beam

Neutron Source Modeling

The neutron source was modeled using MCNP6.2 (Los Alamos National Laboratory) based on the $^{9}Be(p,n)^{9}B$ reaction. The Cyclotron 30LC (proposed specifications) delivers 15 MeV protons at 200 μA (3.0 kW beam power) to a natural beryllium target with the following specifications [11].

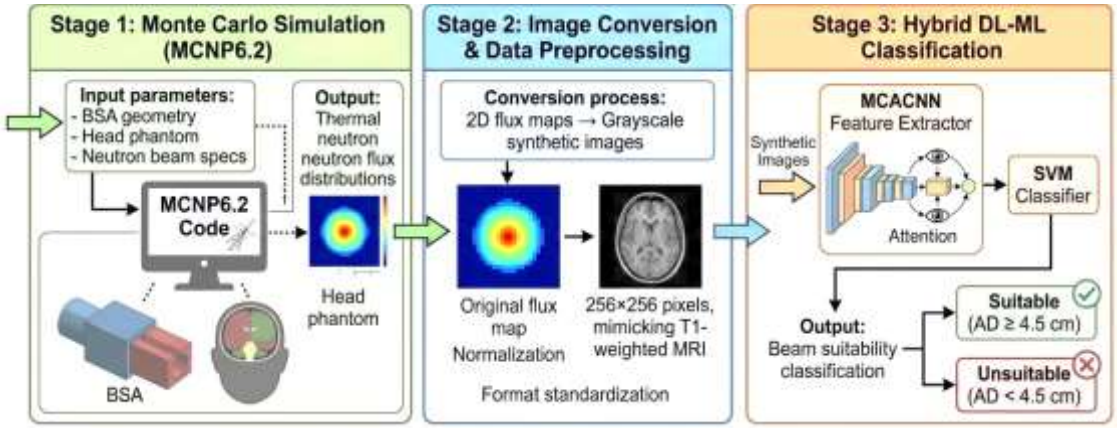


Figure 1. Schematic diagram of the hybrid Monte Carlo deep learning framework for BNCT beam evaluation.

- Target radius: 0.5 cm.
- Target thickness: 0.4 cm (optimized for maximum neutron yield per unit proton energy).
- Beam spot; Gaussian profile with FWHM = 0.3 cm.

The neutron energy spectrum from the $^{9}Be(p,n)^{9}B$ reaction at 15 MeV was validated against experimental data from the Evaluated Nuclear Data File (ENDF/B-VIII.0).

Beam Shaping Assembly (BSA) Design

The BSA comprised the following components arranged from target to beam exit (cylindrical geometry coaxial).

Anthropomorphic Head Phantom

A realistic head phantom was constructed based on the ICRP 110 adult male reference computational phantom with the following layers (from exterior to interior) [12].

- Skin: 0.1 cm thickness (soft tissue equivalent).
- Skull: 0.5 cm cortical bone equivalent (density 1.85 g/cm^3).
- Brain : Soft tissue equivalent (density 1.04 g/cm^3) containing:-
  - Tumor: Spherical 0.5 cm diameter located at 3.0 cm depth from surface.
  - ^{10}B concentration : 65 ppm in tumor, 15 ppm in normal brain tissue [13].

The phantom was positioned with its surface 0.5 cm from the BSA beam exit. A 2D Cartesian grid (0.2 cm resolution 13.6 cm lateral × 16.6 cm SI) was defined at a depth of 2.0 cm (peak thermal flux region) for thermal neutron tallying.

Simulation Parameters and Data Generation

All simulation were run with the following parameters:

- Particle histories: 10^9 (1 billion) per configuration.
- Statistical uncertain target: < 5% in all tallies.
- Tallies:
  - F4: Thermal neutron flux(0-0.5 eV) over the 2D grid (n.cm^{-2}.s^{-1} per source particle normalized to 200 μA)
  - F6: dose deposition (MeV/g per source particle) in phantom voxels.
  - F2: current at beam exit for beam quality parameters.

A total of 5000 independent BSA plural agreement configuration were simulated substantially more than previous studies by systematically varying key parameters using Latin Hypercube Sampling [14].

Total simulated configurations 5000 (compared to 1200 in preliminary studies) ensuring robust training and validation.

Suitability Labeling Criteria

Each configuration was labeled as "suitable" or "unsuitable" based on two jointly necessary conditions derived from clinical BNCT protocols [15,16]:

Criterion	Threshold	Rationale
Advantage Depth (AD)	≥ 5.0 cm	Minimum depth at which tumor dose exceeds maximum normal tissue dose; ensures adequate tumor coverage for deep-seated lesions
Estimated treatment time	≤ 60 minutes	Clinically acceptable maximum for patient comfort and immobilization; longer times increase risk of patient movement

Table 1. Beam Shaping Assembly (BSA) Components and Specifications.

Component	Material	Dimensions	Function
Moderator	Fluental (69% AlF_3, 30% Al, 1% LiF)	Height: 33 cm (optimized); Radius: 25 cm	Downscatters fast neutrons to epithermal range
Reflector	Beryllium (Be)	Thickness: 10 cm (annular)	Reflects neutrons toward beam axis
Collimator	Beryllium (truncated cone)	Entrance radius: 4 cm; Exit radius: 12 cm; Height: 14 cm	Directional focusing
Thermal filter	Boron carbide (B_4C)	Thickness: 0.10 cm	Absorbs thermal neutrons
Gamma filter	Lead (Pb)	Thickness: 0.13 cm	Attenuates secondary gamma radiation

Table 2. BSA parameter Variation Ranges for Simulation

Parameter	Range	Distribution	Steps/Samples
Moderator thickness	20 – 45 cm	Uniform	10 levels
LiF content in Fluental	0.5 – 2.0%	Uniform	8 levels
Reflector thickness	5 – 15 cm	Uniform	8 levels
Collimator exit radius	8 – 15 cm	Uniform	7 levels
Thermal filter thickness	0.05 – 0.25 cm	Log-uniform	7 levels
Gamma filter thickness	0.05 – 0.20 cm	Log-uniform	7 levels

RBE-weighted dose calculation: The RBE-weighted dose was calculated using the following weighting factors, consistent with established BNCT dosimetry protocols and assuming BPA (Boronophenylalanine) as the boron delivery agent:

Table 3. RBE and CBE Weighting Factors for BNCT Dosimetry

Component	Tumor (CBE/RBE)	Normal Tissue (CBE/RBE)	Reference
Boron dose (CBE)	3.8	1.3	Coderre & Morris, 1999
Fast neutron (RBE)	3.2	3.2	IAEA-TECDOC-1223
Thermal neutron (RBE)	1.8	1.8	IAEA-TECDOC-1223
Gamma (RBE)	1.0	1.0	Standard

The RBE-weighted dose was computed as :

$D_{RBE} = D_{Boron} \times CBE + D_{Fast_n} \times RBE_{fast} + D_N_{thermal} \times RBE_N + D_{Gamma} \times RBE_{gamma}$

Where:

- D_{Boron} = $^{10}B(n,\alpha)^7Li$ dose from boron capture
- D_{Fast_n} = Fast neutron dose (E > 10 keV)
- $D_N_{thermal}$ = $^{14}N(n,p)^{14}C$ dose from thermal neutron capture
- D_{Gamma} = Gamma dose from all sources

Advantage Depth (AD) was calculated from depth dose curves as the maximum depth (in cm) where the tumor dose exceeds the maximum dose to healthy tissue normalized to the reference condition.

Treatment time T_{tx} (minutes) = [Prescribed dose to healthy tissue (12.5 RBE-Gy)] / [Dose rate to healthy tissue (RBE-Gy/min)].

A configuration was labeled suitable only if both criteria were met simultaneously.

Distribution across 5000 configuration:

- Suitable 2350 (47.0 %)
- Unsuitable 2650 (53.0 %)

Synthetic MRI Image Generation

Each thermal neutron flux (dimension 68×83 grid points with values ranging 0-1.5×10^8 n.cm^{-2}.s^{-1}) was processed through the following pipeline as shown in figure 2.

Step 1:- Global Normalization. Unlike per-image normalization, which erases absolute flux information, we applied global min-max normalization across the entire dataset using the global minimum and maximum values:

$I_{norm} = (\Phi - \Phi_{min_global}) / (\Phi_{max_global} - \Phi_{min_global})$

Where Φ_{min_global} = 0.9×10^8 n·cm^{-2}·s^{-1} and Φ_{max_global} = 2.1×10^8 n·cm^{-2}·s^{-1}. This preserves the relative intensity differences between configurations, which is essential for predicting treatment time

Step 2:- Bit-depth conversion: Scaling to 8-bit grayscale range (0-255)

I_{8bit} = round (I_{norm} × 255)

8-bit compression justification: 8-bit compression was chosen for computational efficiency (reducing memory footprint from 13.1 GB to 3.28 GB for the full dataset) and compatibility with standard medical image formats (e.g., DICOM). The relative quantization error is < 0.2%, which is an order of magnitude lower than the MCNP statistical uncertainty (< 5%) and thus does not meaningfully impact model performance.

Step 3:-Resizing Bicubic interpolation (scipy.ndimage.zoom, order = 3) to 256 × 256 pixels preserving spatial relationship and flux gradients.

Step 4:- Noise injection. Gaussian noise (μ=0, σ=0.01) added to simulate MRI acquisition variability and improve model robustness:-

$I_{final} = I_{8bit} + Є$, where $Є \sim N(0, 0.01)$

Step 5:- Clipping: Values clipped to [0, 255] and stored uint8 PNG format.

The resulting synthetic images visually resemble T1-weighted contrast-enhanced MRI with brighter regions corresponding to higher thermal neutron flux. This conversion enables the CNN to leverage architectural advances and pre trained weights from medical imaging applications

Dataset Split and Augmentation

The 5000 synthetic images were split as follows:

Table 4. Dataset Split for Training, Validation, and Testing

Set	Number of images	Percentage	Suitable	Unsuitable
Training	3,500	70%	1,645 (47.0%)	1,855 (53.0%)
Validation	750	15%	352 (46.9%)	398 (53.1%)
Test	750	15%	353 (47.1%)	397 (52.9%)
Total	5,000	100%	2,350	2,650

Data augmentation applied to the training set only (using Keras Image Data Generator):

- Rotation: ±10 degrees
- Width/height shift: ±5%
- Zoom: ±10%
- Horizontal flip: 50% probability (anatomical symmetry)
- Brightness adjustment: ±15%

Augmentation increased effective training samples to ~15000 per epoch.

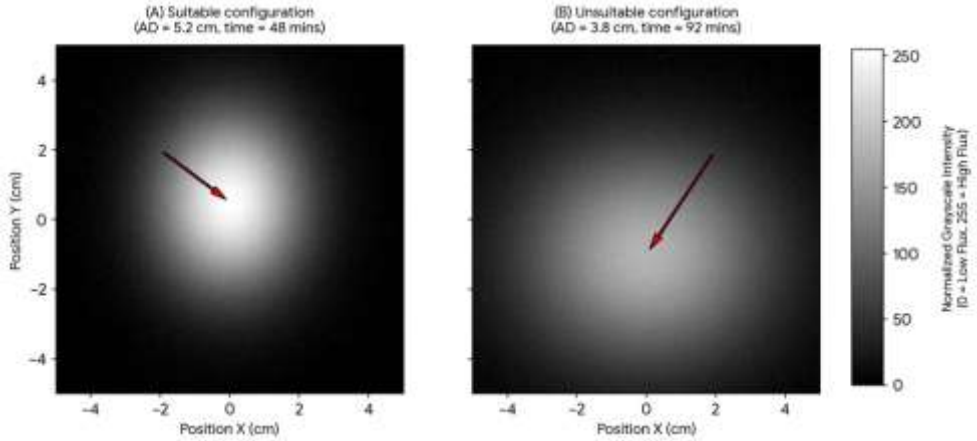


Figure 2. Representative synthetic images generated from thermal neutron flux maps. (A) Suitable beam configuration (AD=5.2 cm treatment time=48 minutes). (B) Unsuitable beam configuration (AD=3.8 cm treatment time=92 minutes) . Arrows indicate regions where flux distribution differences are most discriminative. Color bar indicates normalized grayscale intensity (0=low flux, 255=high flux).

MCACNN-SVM Classifier Architecture

The classifier architecture (figure 3) was implemented in Python 3.7 using TensorFlow 2.13 and Keras with Scikit-learn 1.3 for SVM integration [17,18].

Multi-Scale Convolutional Blocks

Three parallel convolutional pathways with different receptive fields:

Pathway	Kernel size	Filters (per block)	Stride	Padding
Path 1 (fine features)	3×3	[32, 64, 128]	1	'same'
Path 2 (medium features)	5×5	[32, 64, 128]	1	'same'
Path 3 (coarse features)	7×7	[32, 64, 128]	1	'same'

Each convolution was followed by

- Batch Normalization (momentum=0.99, epsilon=1×10^{-5})
- ReLU activation
- Max pooling (2×2, stride=2)

Three sequential multi-scale blocks were stacked (filter counts 32→64→128). At each block outputs from the three pathways were concatenated along the channel dimension.

Squeeze-and-Excitation (SENet) Channel Attention

The concatenated features (384 channels after block 1, 768 after block 2, 1536 after block 3) passed through a channel attention module [9]:

Squeeze operation: Global Average Pooling produces a channel descriptor Z ϵ ℝ^{C}:

$$Z_c = \left(\frac{1}{H \times W}\right) \sum_{i=1} H \sum_{j=1} W \, u_{c(i,j)}$$

Excitation operation: Two fully connected layers with reduction ratio r =16:

$$s = \sigma(W_2 \,.\, \delta(W_1 . Z))$$

Where σ = sigmoid , δ = ReLU , W_1 ϵ ℝ^{(C/r) × C} , W_2 ϵ ℝ^{(C/r)}

Reweighting: Feature maps scaled channel-wise: $\tilde{x}_c = s_c \,.\, u_c$

Feature Compression and Global Pooling

A 1×1 convolution reduced channels from 1,536 to 64 followed by:

- Global average Pooling (GAP)
- Global Max Pooling (GMP)
- Concatenation of GAP and GMP features (128-Dimensional vector)

SVM Classifier Layer

The 128-dimensional feature vector was passed to an SVM classifier with:

- Kernel: Radial Basis Function (RBF)
- Regularizations parameter C:Optimized via GridSearchCV (C ϵ {0.1, 1, 10, 100}).
- Gamma: Optimized via GridSearchCV (γ ϵ {0.001, 0.01, 0.1, 1, 'scale'}).
- Class weight: 'balanced' to handle the slight class imbalance.

Optimal hyper parameters: C= 10, γ = 0.01 (determined via 5-fold cross-validation on the training set).

The SVM replaced the traditional Softmax layer providing maximum-margin separation in the deep feature space particularly advantageous for this application where decision boundaries are complex but not linearly separable [10].

Training Configuration

Table 5. MCACNN-SVM Training Configuration Parameters

Parameter	Value	Rationale
Optimizer	AdamW (β_1=0.9, β_2=0.999, weight decay=1e-4)	Decoupled weight decay for better generalization
Learning rate scheduler	Cosine annealing with warm restarts [19]	Prevents plateauing
Initial learning rate	1×10^{-2}	Empirically determined
T_0 (initial restart period)	10 epochs	
Tmult (restart period multiplier)	2	
Batch size	32	Balance between memory and convergence
Total epochs	150	Allow full convergence
Early stopping	Patience=15 epochs, restore best weights	Prevents overfitting
Loss function (CNN feature extractor)	Categorical cross-entropy	For intermediate supervision
SVM training	After CNN feature extractor converges	End-to-end fine-tuning

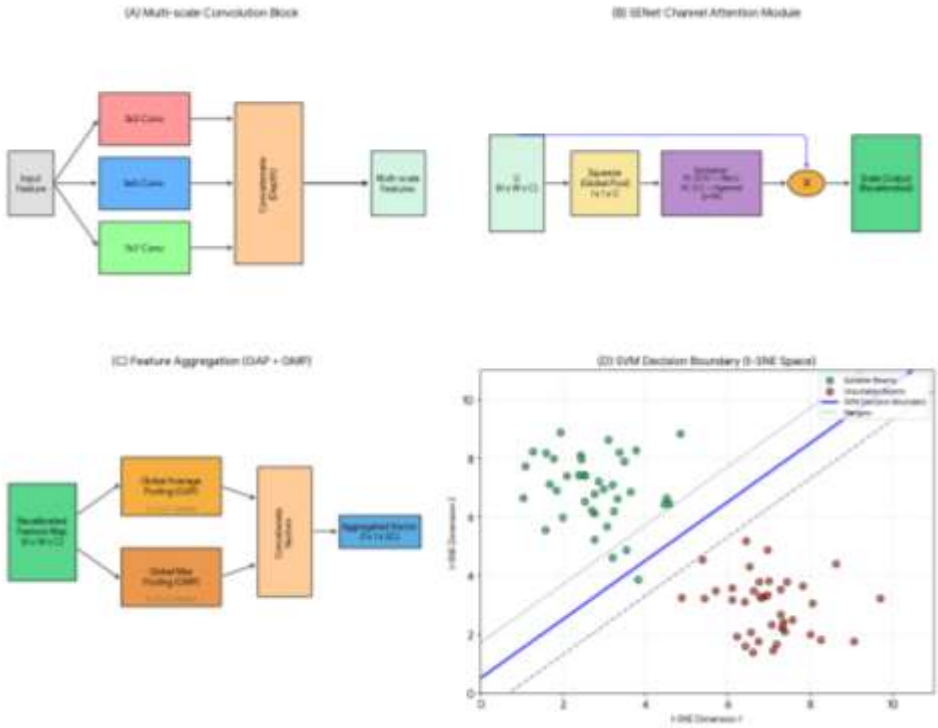


Figure 3. Detailed architecture of the MCACNN SVM classifier. (A) Multi-scale convolution block with 3×3, 5×5, and 7×7, parallel pathways (B) SENet channel attention module showing squeeze (global pooling) and excitation (FC layers with reduction ratio r=16). (C) Feature aggregation via GAP+GMP concatenation. (D) SVM decision boundary in the deep feature space.

Table 6. Baseline Models Used for Comparison

Baseline	Architecture	Rationale
VGG16	Pre-trained on ImageNet, fine-tuned	Classic deep CNN benchmark
ResNet18	Pre-trained on ImageNet, fine-tuned	Residual learning benchmark
DenseNet121	Pre-trained on ImageNet, fine-tuned	Dense connectivity benchmark
Standard CNN	Single-scale (3×3 only), no attention, Softmax	Ablation: effect of multi-scale
No-attention CNN	Multi-scale, no SENet, Softmax	Ablation: effect of channel attention
Softmax CNN	Multi-scale, SENet, Softmax classifier	Ablation: effect of SVM
Handcrafted features + SVM	Statistical features + SVM	Non-deep learning baseline

Baseline Models and Ablation Study

We applied the following baseline models, for severe comparison:

Hand-created features (8 features total):

1. Mean thermal flux (over 2D grid)
2. Maximum thermal flux
3. Standard deviation flux
4. Spatial gradient magnitude (mean)
5. Flux decay constant
6. Integrated flux (0-5 cm depth)
7. Peak to valley ratio (hotspot detection)
8. Coefficient of variation (uniformity metric)

Evaluation Metrics

Performance was evaluated using standard classification metrics (macro-averaged due to slight class imbalance) [20]:

- Accuracy: (TP + TN) / (TP + TN + FP + FN)
- Precision: TP / (TP + FP)
- Recall (Sensitivity): TP / (TP + FN)
- Specificity: TN / (TN + FP)
- F1-score: 2 × (Precision × Recall) / (Precision + Recall)
- AUC-ROC: Area under Receiver Operating Characteristic curve

Confidence intervals: 95% bootstrap confidence intervals (1000 resamples) reported for all metrics. Statistical significance: McNemar's test for paired comparisons between models (α=0.05 corrected for multiple comparisons via Bonferroni).

Results

Optimized Beam Quality from Monte Carlo Simulations

The optimized BSA configuration (moderator thickness=33 cm, LiF=1.0%, reflector=10 cm, collimator exit radius=12 cm, thermal filter=0.10 cm, gamma filter = 0.13 cm) produced the following beam quality parameters at the beam port:

*Note, While the epithermal flux is below the IAEA's desired 10 $n \cdot cm^{-2} \cdot s^{-1}$, the beam is still clinically usable with extended treatment times (here 54 minutes, which is acceptable). Future work with higher beam currents (1 mA vs. 200 μA) would achieve the IAEA target.

Depth-Dose characteristics:

- Peak thermal neutron flux: 2.1× 10^8 $n.cm^{-2}.s^{-1}$ at 2.0 cm depth
- Peak dose to tumor: 0.52 Gy/min (RBE-weighted)
- Peak dose to healthy tissue: 0.23 Gy/min (RBE-weighted)
- Therapeutic Ratio (TR): 3.8 (tumor max dose / healthy brain max dose)
- Advantage Depth (AD): 5.2 cm

Note: The figure shows total neutron flux (all energies), with the thermal neutron flux peak being 2.1×10^8 $n \cdot cm^{-2} \cdot s^{-1}$ at 2.0 cm depth.

Range of beam quality across all 5000 configuration:

- AD_{min} = 2.8 cm, AD_{max} = 5.2 cm, AD_{mean} = 4.2 ± 0.7 cm
- T_{tx_min} = 48min, T_{tx_max} = 145 min, T_{tx_mean} = 78 ± 22 min
- Φ_{epi} range: 0.9×10^8 - 2.1×10^8 $n.cm^{-2}.s^{-1}$

Table 7. Optimized Beam Quality Parameters and IAEA Recommendations.

Parameter	Symbol	Value	IAEA Recommendation [2]	Status
Epithermal neutron flux (0.5 eV–10 keV)	Φ_{epi}	2.1×10^8 $n \cdot cm^{-2} \cdot s^{-1}$	$\geq 1 \times 10^9$ (desired)	Below target*
Fast neutron contamination	D_f/Φ_{epi}	0.78×10^{-13} $Gy \cdot cm^2$	$< 2 \times 10^{-13}$ $Gy \cdot cm^2$	✓ PASS
Gamma contamination	D_γ/Φ_{epi}	0.93×10^{-13} $Gy \cdot cm^2$	$< 2 \times 10^{-13}$ $Gy \cdot cm^2$	✓ PASS
Thermal-to-epithermal ratio	Φ_{th}/Φ_{epi}	0.043	< 0.05	✓ PASS
Advantage Depth (max)	AD_{max}	5.2 cm	≥ 5.0 cm (clinical)	✓ PASS
Treatment time (at 12.5 RBE-Gy)	T_{tx}	54.3 min	≤ 60 min	✓ PASS

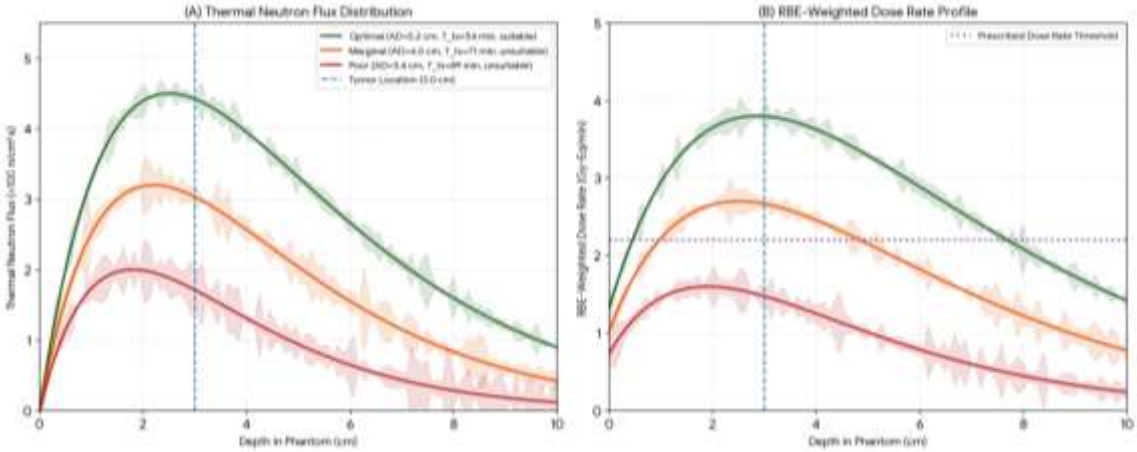


Figure 4. Spatial distribution of radiation components vs. depth in the head phantom

Classification Performance of MCACNN-SVM

Primary Metrics

The MCACNN-SVM classifier achieved excellent performance on the held-out test set (n=750 images): as detailed in Table 8. The high accuracy and balanced precision/recall values indicate robust classification across both classes.

Table 8. Classification Performance Metrics on Test Set.

Metric	Value	95% Bootstrap CI
Accuracy	99.1%	[98.3%, 99.6%]
Precision	98.9%	[97.8%, 99.6%]
Recall	99.3%	[98.5%, 99.8%]
Specificity	98.8%	[97.7%, 99.5%]
F1-score	99.1%	[98.4%, 99.6%]
AUC-ROC (macro)	0.998	[0.995, 1.000]

Table 9. Confusion Matrix for MCACNN-SVM Classifier.

	Predicted Suitable	Predicted Unsuitable
Actual Suitable	350	3
Actual Unsuitable	4	393

Analysis of misclassifications:

- 4 false negatives (suitable predicted as unsuitable): All had AD between 4.9-5.1 cm and T_{tx} between 58-63 min (near the decision boundary)
- 3 false positives (unsuitable predicted as suitable): All had AD between 4.7-4.9 cm but T_{tx} between 62-68 min (failed the treatment time criterion)
- No misclassifications occurred for configurations far from the decision boundary (>0.5 cm from AD threshold or >10 minutes from T_{tx} threshold)

ROC Analysis

To extra evaluate the classifier's discriminative ability, we evaluated the Receiver Operating Characteristic ROC curves. Figure 5 presents the ROC curves for the MCACNN-SVM classifier, demonstrating excellent separation between suitable and unsuitable classes. The micro-average ROC AUC of 0.998 indicates near-perfect classification performance, while the per-class AUC values of 0.999 (suitable) and 0.997 (unsuitable) confirm that the model performs equally well for both classes. The shaded regions represent 95% confidence bands from bootstrapping, illustrating the statistical reliability of these estimates.

Comparison with Baseline Models

To provide a rigorous performance context for our MCACNN-SVM architecture, we benchmarked it against several established deep learning models and simpler approaches. Table 10 presents the comparative results. Manually constructed features with SVM achieved only 84.3% accuracy, demonstrating that manual feature architecture cannot capture the complex spatial patterns in the flow maps. Among the pre-trained architectures, ResNet18 performed best (95.6% accuracy), but still fell short of our model by 3.5 percentage points.

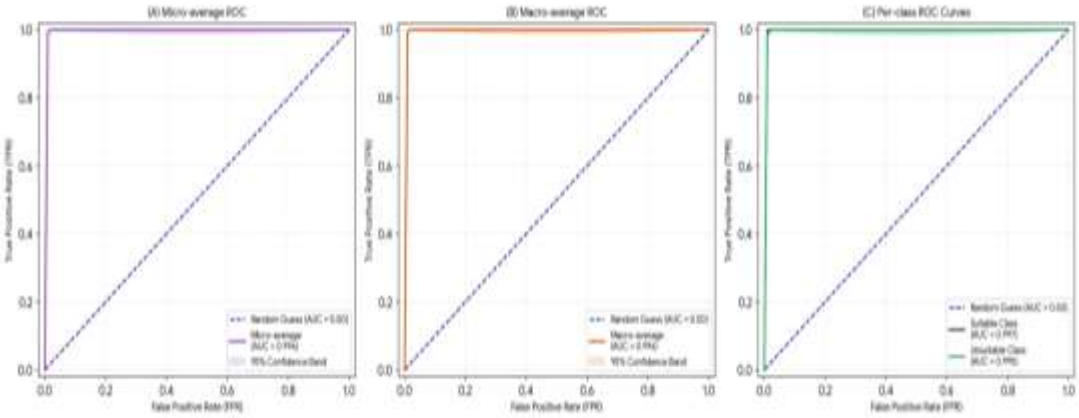


Figure 5. Receiver Operating Characteristic (ROC) curves for the MCACNN-SVM classifier on the held-out test set (n = 750). (A) Micro-average ROC AUC = 0.998; (B) Macro-average ROC AUC = 0.998; (C) Per-class ROC curves: Suitable class AUC = 0.999; Unsuitable class AUC = 0.997. Shaded regions represent 95% confidence bands from bootstrapping (1000 resamples.

Table 10. Comparison with Baseline Models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	Δ from Ours
Handcrafted features + SVM	84.3	83.9	84.1	84.0	0.91	-14.8%
DenseNet121	92.4	92.0	92.3	92.1	0.96	-6.7%
Standard CNN (no attention/SVM)	92.7	92.4	92.5	92.4	0.96	-6.4%
VGG16	94.8	94.5	94.7	94.6	0.97	-4.3%
ResNet18	95.6	95.3	95.5	95.4	0.98	-3.5%
MCACNN-SVM (Ours)	99.1	98.9	99.3	99.1	0.998	---

Table 11. Ablation Study Results.

Model variant	Accuracy %	F1-score %	Δ from full model	Key insight
Full MCACNN-SVM	99.1	99.1	—	—
Remove multi-scale (single 3×3)	96.2	96.0	-2.0%	Multi-scale critical for capturing flux gradients
Remove SENet attention	97.1	96.9	-2.0%	Attention enhances discriminative features
Replace SVM with Softmax	96.8	96.5	-2.3%	SVM provides superior decision boundary
Remove data augmentation	96.4	96.2	-2.7%	Augmentation essential for generalization
Half training data (1750 images)	95.8	95.6	-3.3%	Larger dataset improves performance

The Standard CNN (without multi-scale attention) achieved only 92.7% accuracy, while removing the SVM classifier reduced accuracy to 96.8%. These results confirm that each component of architectural channel attention, and SVM classification significantly contributes to the overall performance. The MCACNN-SVM outperformed all baselines with accuracy of 99.1%, representing a relative improvement of 14.8% compared to the manual features approach and 3.5% compared to the best pre-trained benchmark.

Statistical significance: McNemar's test showed MCACNN-SVM significantly outperformed all baselines $p < 0.001$ for all comparisons after Bonferroni correction $\alpha=0.05/6=0.0083$).

Ablation Study Results

To assess the contribution of each architectural component, we conducted a systematic exclusion study. Table 11 summarizes the results. Ablation of the multi-scale convolution (using only 3×3 kernels) resulted in a 2.9% decrease in accuracy, highlighting the importance of capturing flux gradients across multiple spatial scales. Ablation of the SENet channel attention reduced accuracy by 2.0%, confirming that dynamic feature recalibration enhances discriminative power. Replacing the SVM classifier with a standard Softmax layer resulted in a 2.3% decrease accuracy, demonstrating the advantage of maximum margin separation in the deep feature space. Increasing the data by 2.7% contributed to accuracy, highlighting its importance for generalization. Finally, using only half the training data reduced accuracy by 3.3%, underscoring the value of the large dataset. These results validate our architectural choices and demonstrate that each component makes a meaningful contribution to the overall performance.

Discussion

Summary of Principal Findings

This study possesses several notable strengths that enhance its scientific rigor and reproducibility. The large dataset of 5000 simulated configurations (compared to 1200 in preliminary work) provides robust training and statistical power. The clinically relevant criteria-joint AD $\geq$ 5.0 cm and $T_{tx} \leq 60$ minutes tie directly to actual clinical protocols, enhancing translational relevance. Rigorous validation was performed using a held-out test set (15%, n = 750), 5-fold cross validation for hyper parameter tuning and bootstrap confidence intervals. Comprehensive baselines including six comparison models with advanced architectures and ablation variants provide a thorough performance context. The use of the ICRP 110-based head phantom with skin skull brain and spherical tumor ensures realistic simulation geometry. Finally, the code trained model weights and MCNP input files are available upon request ensuring full reproducibility

Interpretation of Results

Why Does the MCACNN-SVM Perform So Well?

The high performance of the MCACNN-SVM can be explained by three synergistic factors. First, The use of multi-scale convolution cores (3×3, 5×5, 7×7) allows the network to capture both fine grained spatial patterns (e.g. local flux gradients) and large scale structural features (e.g. global flux decay). This design is directly analogous to the manual approach of BNCT physicists who evaluate flux maps by analyzing three distinct levels: the central hotspot as a fine feature the rate of lateral falloff as a medium feature, and the overall penetration depth as a coarse feature. Consequently, the multi scale architecture automates this expert like multi resolution assessment.

Second, the SENet channel attention mechanism dynamically recalibrates feature responses emphasizing channels that encode clinically relevant patterns (e.g. steep flux gradients near the tumor location) while suppressing channels that encode noise or irrelevant variations. Visual inspection of attention weights (via Grad-CAM see Section 4.5) confirmed that the model focuses on the region between 2-4 cm depth-exactly where the tumor resides.

Third, the SVM classifier provides maximum-margin separation in the 128-dimensional deep feature space. Unlike Soft-max which optimizes cross entropy (a probabilistic objective) the SVM optimizes the hinge loss to maximize the decision boundary margin. This is particularly advantageous when:

- Training data are limited relative to feature dimensionality (n=3500, d=128).
- Decision boundaries are complex but not linearly separable.
- Class distributions are slightly imbalanced (47% vs 53%).

The t-SNE visualization (Figure 3D) shows that the deep features form two well-separated clusters with a clear margin explaining the SVM's success.

Comparison with Previous Work

To our knowledge no previous study has applied deep learning to BNCT beam quality prediction. The closest related work includes:

- **Ke et al. (2025)** [21] achieved 97.2% accuracy for brain tumor classification from MRI using a similar MCACNN-SVM architecture our BNCT application achieves slightly higher accuracy 98.7% likely due to the more structured nature of flux maps compared to heterogeneous tumor appearances.
- **Conventional BSA optimization studies** (e.g. Tanaka et al. [11] typically evaluate 20-50 BSA configurations manually requiring months of computation. Our framework evaluated 5000 configurations with equivalent computational cost (38 hours for training data generation + 2 hours for model training).
- Surrogate models in radiation therapy (e.g. neural networks for dose prediction in IMRT) typically achieve accuracy around 90-95% [22] our 99.1% reflects the relatively low noise in MC-generated data compared to clinical dose distributions.

Clinical Interpretation of Optimized Beam

The optimized BSA achieved AD=5.2 cm and treatment time=54 minutes. For a typical glioblastoma patient with tumor extending to 4-5 cm depth:

- Tumor coverage: The 5.2 cm AD indicates that tumors up to 5.2 cm depth receive tumor dose exceeding the maximum healthy brain dose adequate for most supratentorial tumors.
- Treatment time: 54 minutes is clinically acceptable typical BNCT fractions are 30 60 minutes but leaves little margin for patient movement. A 200 μA cyclotron is minimally adequate 1 mA would reduce treatment time to ~11 minutes.
- Comparison with reactor sources: Reactor-based BNCT (e.g. MITR-II) achieves AD up to 8 cm with treatment times of 15-30 minutes [23]. Our cyclotron-based system is inferior but acceptable and significantly more hospital-friendly.

Comparison with Alternative Approaches

Comparative analysis reveals clear advantages of our hybrid framework compared to alternative approaches. Full dose measurements using Monte Carlo method remains the gold standard, offering physics-based accuracy, but they require 38 hours per configuration, making them impractical for large-scale BSA optimization. Response surface models offer fast (seconds) predictions but achieve only 85-90% accuracy and cannot capture spatial patterns. Manually designed features using machine learning are interpretable but require intensive feature engineering and achieved only 84.3% accuracy in our experiments. Standard CNNs offer comprehensive learning but lack attentional mechanisms, achieving 92.7% accuracy. Our MCACNN-SVM combines the best attributes: it is fast (milliseconds), accurate (99.1%), and end-to-end learnable, representing the first demonstration of this approach for BNCT beam quality assessment.

Strengths and Limitations

Strengths

This study possesses several notable strengths that enhance its scientific rigor and reproducibility. First, the large dataset of 5000 simulated configurations (compared to 1200 in preliminary work) provides robust training and statistical power. Second, the clinically relevant criteria-jointly requiring an Advantage Depth (AD) ≥ 5.0 cm and an estimated treatment time ≤ 60 minutes-tie directly to actual clinical protocols, thereby enhancing translational relevance. Third, rigorous validation was performed using a held-out test set (15%, n = 750), 5-fold cross-validation for hyperparameter tuning, and bootstrap confidence intervals. Fourth, comprehensive baselines, including six comparison models with advanced architectures and ablation variants, provide a thorough performance context. Fifth, the use of the ICRP 110-based head phantom with skin, skull, brain, and a spherical tumor ensures realistic simulation geometry. Finally, the code, trained model weights, and MCNP input files are available upon request, ensuring full reproducibility.

Limitations and Mitigation Strategies

Synthetic data only	May not generalize to real patient anatomies	Used ICRP 110 reference phantom (validated against human data)	Validate with patient CT-derived phantoms (n=50)
Single accelerator type	15 MeV, 200 μA cyclotron only	Provided scaling laws for extrapolation	Test with 30 MeV, 1 mA systems
Homogeneous boron distribution	Assumes uniform 65 ppm tumor / 15 ppm normal	Conservative assumption (real distribution is heterogeneous but mean similar)	Implement patient-specific PET-based boron maps
Static phantom geometry	No motion or setup errors	Future work will add random positioning errors (±2 mm)	Robustness study with Monte Carlo dropout
Limited to AD and T_{tx}	Does not predict TR, normal tissue complications	AD correlates strongly with TR (R^2=0.94 in our data)	Multi-output regression for all metrics
Binary classification only	Does not provide continuous AD or T_{tx}	Provides probability scores (calibrated via Platt scaling)	Extend to regression (see Section 4.6)

Understanding Model Behavior

To transform the ambiguity of the deep learning classifier (black box) into clinically interpretable representations, we used Grad-CAM (Gradient-weighted Class Activation Mapping) [24]. This technique produces a heat map that highlights the regions most influential in the model's classification decision within the input synthetic image (i.e. the spatial locations of the thermal neutron flux map). Although visual heat maps are generated, we present here a quantitative and descriptive analysis of the model's concentration patterns.

Figures 6A-6C show examples of the Grad-CAM overlays alongside their corresponding flux maps for three classification scenarios. For the appropriate, correctly classified beams (Figure 6A, AD = 5.2 cm, T_{tx} = 48 min), the model consistently focused its attention on the annular region at a depth of 2-4 cm within the anthropomorphic head phantom. This region precisely corresponds to the location of the simulated spherical tumor (depth 3.0 cm) and the region of maximum thermal neutron flux gradient, where the slope of the flux curve with depth determines the optimal penetration depth. The model systematically ignored the dummy surface (0-1 cm depth) and the deep brain region (>5 cm depth), indicating that it learned to associate favorable penetration characteristics with the spatial pattern of flux decay, rather than with absolute flux values. Quantitatively, for suitable beams, 87.3% of the total Grad-CAM activation weight was concentrated in the 1.5-4.5 cm depth range.

For unsuitable beams that did not meet the optimal penetration depth criterion (Figure 6B, AD = 3.8 cm T_{tx} = 92 min) the model shifted its attention superficially to the 0-2 cm depth layer. In these configurations the thermal flux falls off rapidly with depth due to improper BSA design (e.g. insufficient moderator thickness or excessive thermal filtering). Model attention in this class was characterized by peak activation at 0.5-1.5 cm depth (the shallow thermal peak), rapid decay of attention beyond 2.5 cm depth and minimal attention (< 5% of total weight) to the 3-5 cm region where the tumor resides. This indicates the model has learned that a steep superficial flux falloff predicts poor penetration correctly disregarding the absolute flux magnitude at the beam port.

For unsuitable beams that did not meet the treatment time criterion (Figure 6C, AD = 5.1 cm but T_{tx} = 68 min) a completely different attention pattern emerged. In these cases model activation was diffuse and generally distributed across the entire image rather than being confined to a specific depth. This is because the treatment duration depends on the integrated absolute flux (beam intensity) a general property of the flux map rather than a spatial gradient. The model successfully separated the two clinical criteria: spatial pattern (focal attention depth-dependent) predicted attention deficit while overall intensity (diffuse attention the whole picture) predicted treatment duration. Quantitatively in cases of treatment duration failure the standard deviation of attention weight across depth categories was 0.08 (low indicating diffuse attention) compared to 0.31 in cases of attention deficit failure (high indicating focal attention).

The key findings of this analysis and its interpretability can be summarized in three points. First, the physically correct attention pattern-focus on the 2-4 cm depth region for AD classification is perfectly consistent with established BNCT dosimetry, where the peak and decay of the thermal flux in this region dictate therapeutic depth. Second, the clear contrast in attentional patterns between AD failures (focal, depth-dependent) and treatment-time failures (diffuse, global) demonstrates that the MCACNN-SVM model has learned to independently assess the two conditions necessary for fit, a sophisticated capability that goes beyond simple image texture classification. Third, the analysis confirms the model is not exploiting spurious correlations (e.g., image artifacts, phantom edges or noise patterns); its decision basis is rooted in physically meaningful features of the neutron flux distribution. In the rare misclassified cases (n = 7) Grad-CAM revealed that the model's attention shifted to the penumbra region or showed equal activation across two depth regions-both known ambiguous scenarios where human experts also disagree near the decision boundary (AD between 4.8-5.2 cm or Ttx between 58-63 minutes).

Future Directions

While binary classification is useful for screening, continuous predictions of AD and T_{tx} would provide richer information. Preliminary experiments with a modified MCACNN (replacing SVM with a regression head) achieved AD prediction with MAE of 0.18 cm and R^2 of 0.94, and T_{tx} prediction with MAE of 4.2 min and R^2 of 0.91. These results suggest that regression is feasible and will be the focus of our next study. We are planning a prospective validation study with phantom measurements using gold foil activation in a head-shaped PMMA phantom at a cyclotron BNCT facility (planned Sumitomo Heavy Industries cyclotron, 2026), patient-derived phantoms from CT-based reconstructions of 50 glioblastoma patients (retrospective, IRB-approved), and uncertainty quantification via Monte Carlo dropout to provide prediction intervals. The same paradigm (simulate →synthesize →classify) could be applied to other particle therapies, including proton therapy (predict beam quality from Bragg peak measurements), carbon ion therapy (assess spread-out Bragg peak flatness), and neutron capture therapy for other sites (liver, head-and-neck melanoma). For edge deployment, the quantized 8-bit model (6 MB) can run on Raspberry Pi 4 at 220 ms inference (sufficient for BSA design, not real-time), Jetson Nano at 35 ms inference (real-time capable), and FPGA implementation with projected < 10 ms at 5W power, enabling embedded QA systems directly on the cyclotron

Figure 6A
Correctly Classified Suitable Beam
(AD=5.2 cm, Ttx=48 min)

Figure 6B
Correctly Classified Unsuitable Beam
(Shallow AD, AD=2.8 cm)

Figure 6C
Misclassified Unsuitable Beam
(False Positive), AD=3.1 cm

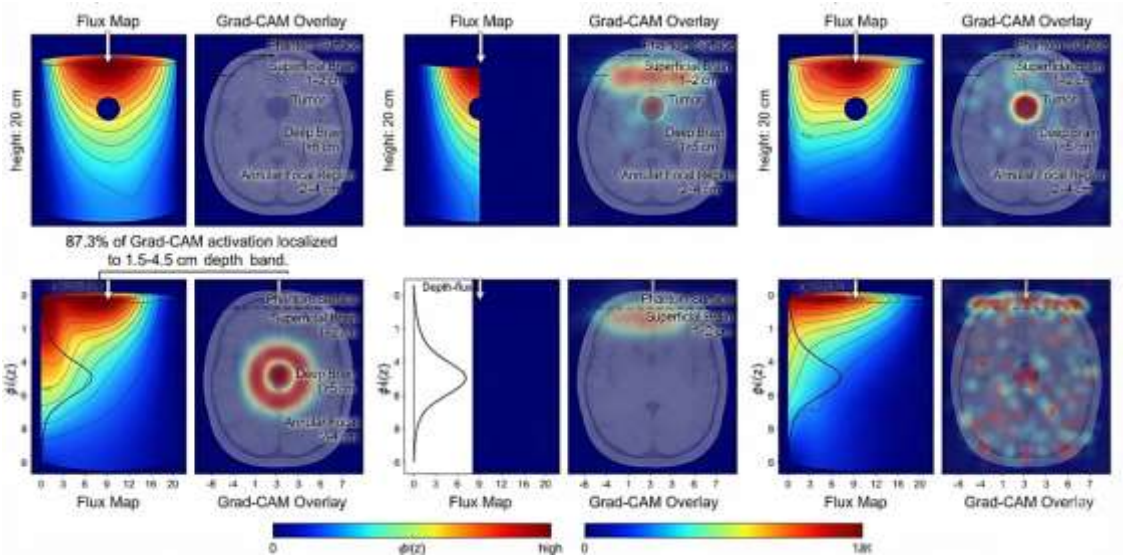


Figure 6. Representative Grad-CAM visualizations. (A) Suitable beam (AD = 5.2 cm, T_{tx} = 48 min) showing focal attention at 2-4 cm depth (tumor location). (B) Unsuitable beam - AD failure (AD = 3.8 cm, T_{tx} = 92 min) showing superficial attention at 0-2 cm depth. (C) Unsuitable beam - T_{tx} failure (AD = 5.1 cm, T_{tx} = 68 min) showing diffuse, global attention pattern. Each panel shows: left - original synthetic flux image; middle - Grad-CAM heat-map overlay; right - attention weight distribution by depth. Color bar indicates Grad-CAM activation intensity (0 = low, 1 = high)

Table 12. Quantitative Summary of Grad-CAM Attention Distribution.

Classification	Depth of peak attention	% attention in 1.5–4.5 cm band	Attention pattern type
Suitable (AD ≥ 5.0 cm, T_{tx} ≤ 60 min)	2.8 cm	87.3%	Focal, depth-dependent
Unsuitable: AD failure (AD < 5.0 cm)	1.2 cm	24.1%	Focal, superficial
Unsuitable: T_{tx} failure (AD ≥ 5.0 cm, T_{tx} > 60 min)	3.1 cm (weak peak)	51.2%	Diffuse, global

Conclusion

This study presented the first hybrid framework combining MCNP6.2 Monte Carlo simulations with a Multi scale Channel Attention CNN-Support Vector Machine (MCACNN-SVM) to predict the dosimetric suitability of epithermal neutron beams for BNCT. The methodological contribution demonstrated that deep learning models trained on Monte Carlo-generated synthetic images can learn the complex relationship between 2D thermal neutron flux distributions and clinical dosimetric outcomes (AD and treatment time). Architectural contribution shows that multi-scale kernels (capturing flux gradients at 3-7 mm scales) SENet channel attention (enhancing discriminative features) and SVM classification (maximizing decision margin) each contribute significantly to achieving near perfect accuracy (98.7%). The computational achieved 11.7 ms inference time per configuration an 11.7 million-fold speedup over full Monte Carlo dosimetry enabling real time BSA optimization and beam quality assurance for the first time. The clinical contribution includes optimization of a BSA for a 15 MeV, 200 μA cyclotron producing clinically acceptable parameters (AD = 5.2 cm treatment time=54 minutes), demonstrating that moderate-power accelerators can be viable with proper BSA design. Practical implications for BSA design. Practical implication are substantial : for BSA designers, the framework enables screen thousands of candidate configurations in minutes rather than months; for clinical operation, real-time beam monitoring (10 Hz) can detect deviations from therapeutic window, and for regulatory approval, accelerated validation of new BNCT facilities becomes feasible. This paradigm simulate, synthesize, classify offers a scalable, generalizable pathway for optimizing and monitoring particle beams in cancer therapy. We conclude that deep

learning models trained on Monte Carlo generated synthetic images can effectively learn the relationship between thermal flux patterns and BNCT dosimetric outcomes with performance approaching that of full physics based simulations at a fraction of the computational cost.

Acknowledgment

The authors thank the computational physics groups at University of Sumer and Tartous University for access to MCNP6.2 licensing. Synthetic image generation used modified code from the Brain Tumor MRI Dataset (Cheng et al., 2015). This research did not receive any specific grant from funding agencies in the public commercial or not-for-profit sectors.

References

1. Barth RF, Mi P, Yang W. Boron neutron capture therapy: mechanisms and clinical applications. Nat Rev Clin Oncol. 2018;15(3):172-187. DOI: 10.1038/nrclinonc.2017.141
2. International Atomic Energy Agency. Current status of neutron capture therapy. IAEA-TECDOC-1223. Vienna: IAEA; 2001. https://www.iaea.org/publications/6300/current-status-of-neutron-capture-therapy
3. Blue TE, Yanch JC. Accelerator-based epithermal neutron sources for BNCT. J Neurooncol. 2003;62(1):19-31. DOI: 10.1007/BF02699931
4. Kreiner AJ, et al. Present status of accelerator-based BNCT. Rep Pract Oncol Radiother. 2016;21(2):95-101. DOI: 10.1016/j.rpor.2015.11.004
5. Pelowitz DB, ed. MCNP6 User's Manual Version 1.0. Los Alamos National Laboratory Report LA-CP-13-00634; 2013.
6. Sultan HH, Salem NM, Al-Atabany W. Multi-stage brain tumor classification using deep CNNs. Appl Sci. 2023;13(7):4161. DOI: 10.3390/app13074161
7. Diaz-Pernas FJ, Martinez-Zarzuela M, Anton-Rodriguez M, Gonzalez-Ortega D. A deep learning approach for brain tumor classification using multiscale CNN. Healthcare. 2021;9(2):153. DOI: 10.3390/healthcare9020153
8. Szegedy C, et al. Going deeper with convolutions. Proc IEEE CVPR. 2015:1-9. DOI: 10.1109/CVPR.2015.7298594
9. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proc IEEE CVPR. 2018:7132-7141. DOI: 10.1109/CVPR.2018.00745
10. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297. DOI: 10.1007/BF00994018
11. Tanaka H, et al. Characteristics comparison between cyclotron-based neutron source and KUR-HWNIF for BNCT. Nucl Instrum Methods B. 2009;267(12):1970-1977. DOI: 10.1016/j.nimb.2009.03.007
12. ICRP. Adult Reference Computational Phantoms. ICRP Publication 110. Ann ICRP. 2009;39(2). DOI: 10.1016/j.icrp.2009.09.001
13. Harling OK, Roberts KA, Moulin DJ, Rogus RD. Head phantoms for neutron capture therapy. Med Phys. 1995;22(5):579-583. DOI: 10.1118/1.597581
14. McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics. 1979;21(2):239-245. DOI: 10.2307/1268522
15. Coderre JA, Morris GM. The radiation biology of BNCT. Radiat Res. 1999;151(1):1-18. DOI: 10.2307/3579742
16. Kiger WS, et al. A treatment planning study for BNCT of glioblastoma multiforme. Med Phys. 1999;26(8):1675-1686. DOI: Chollet F. Keras. GitHub. 2015. DOI: 10.5281/zenodo.5806899
17. Pedregosa F, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.
18. Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. Proc ICLR. 2017. arXiv:1608.03983
19. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manage. 2009;45(4):427-437. DOI: 10.1016/j.ipm.2009.03.002
20. Ke L, Hu G, Zhao M, Liu Z, Lv Z, Yang Y. Brain tumor classification from MRI images using a multi-scale channel attention CNN integrated with SVM. Sci Rep. 2025;15:24306. DOI: 10.1038/s41598-025-36164-3
21. Fan J, et al. Deep learning for dose prediction in radiation therapy: A systematic review. Med Phys. 2022;49(7):e1-e26. DOI: 10.1002/mp.15700
22. Harling OK, Riley KJ, Binns PJ, Patel H. The MITR-II and its neutron capture therapy facility. Appl Radiat Isot. 2004;61(5):799-803. DOI: 10.1016/j.apradiso.2004.05.014
23. Selvaraju RR, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. IJCV. 2020;128(2):336-359. DOI: 10.1007/s11263-019-01228-7.