# Automatic detection and classification of treatment deviations in proton therapy from realistically simulated prompt gamma imaging data

Running title: “Automatic classification of PT-PGI data”

Julian Pietsch[a,b,*], Chirasak Khamfongkhruea[a,c], Jonathan Berthold[a,b], Guillaume Janssens[d], Kristin Stützer[a,b], Steffen Löck[a,e,f,#], Christian Richter[a,b,e,f,#]

[a] OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden – Rossendorf, Dresden, Germany

[b] Helmholtz-Zentrum Dresden - Rossendorf, Institute of Radiooncology – OncoRay, Dresden, Germany

[c] Princess Srisavangavadhana College of Medicine, Chulabhorn Royal Academy, Bangkok, Thailand

[d] Ion Beam Applications SA, Louvain-la-Neuve, Belgium

[e] German Cancer Research Center (DKFZ), Heidelberg and German Cancer Consortium (DKTK) partner site Dresden, Dresden, Germany

[f] Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany

[#] Shared last authorship

* Corresponding author:

Julian Pietsch, OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, PF 41, 01307 Dresden, Germany, Tel: +49 351 458 2141, julian.pietsch@oncoray.de

## Abstract

**Background:**

A clinical study regarding the potential of range verification in proton therapy by prompt gamma imaging (PGI) is carried out at our institution. Manual interpretation of the detected spot-wise range shift information is time-consuming, highly complex, and therefore not feasible in a broad routine application.

**Purpose:**

Here, we present an approach to automatically detect and classify treatment deviations in realistically simulated PGI data for head and neck cancer (HNC) treatments using convolutional neural networks (CNNs) and conventional machine learning (ML) approaches.

**Methods:**

For 12 HNC patients and one anthropomorphic head phantom (n=13), pencil beam scanning (PBS) treatment plans were generated and one field per plan was assumed to be monitored with the IBA slit camera. In total, 386 scenarios resembling different relevant or non-relevant treatment deviations were simulated on planning and control CTs and manually classified into 7 classes: non-relevant changes (NR) and relevant changes (RE) triggering treatment intervention due to range prediction errors (±RP), setup errors in beam direction (±SE), anatomical changes (AC), or a combination of such errors (CB). PBS spots with reliable PGI information were considered with their nominal Bragg peak position for the generation of two 3D spatial maps of 16×16×16 voxels containing PGI-determined range shift and proton number information. Three complexity levels of simulated PGI data were investigated: (I) optimal PGI data, (II) realistic PGI data with simulated Poisson noise based on the locally delivered proton number, and (III) realistic PGI data with an additional positioning uncertainty of the slit camera following an experimentally determined distribution.

For each complexity level, 3D-CNNs were trained on a data subset (n=9) using patient-wise leave-one-out cross-validation and tested on an independent test cohort (n=4). Both the binary task of detecting RE and the multi-class task of classifying the underlying error source were investigated. Similarly, four different conventional ML classifiers (logistic regression, multi-layer perceptron, random forest, support vector machine) were trained using five previously established handcrafted features extracted from the PGI data and used for performance comparison.

**Results:**

On the test data, the CNN ensemble achieved a binary accuracy of 0.95, 0.96, and 0.93 and a multi-class accuracy of 0.83, 0.81, and 0.76 for the complexity levels (I), (II), and (III), respectively. In the case of binary classification, the CNN ensemble detected treatment deviations in the most realistic scenario with a sensitivity of 0.95 and a specificity of 0.88. The best performing ML classifiers showed a similar test performance.

**Conclusions:**

This study demonstrates that CNNs can reliably detect relevant changes in realistically simulated PGI data and classify most of the underlying sources of treatment deviations. The CNNs extracted meaningful features from the PGI data with a performance comparable to ML classifiers trained on previously established handcrafted features. These results highlight the potential of a reliable, automatic interpretation of PGI data for treatment verification, which is highly desired for broad clinical application and a prerequisite for the inclusion of PGI in an automated feedback loop for online adaptive proton therapy.

**Keywords:** range verification, prompt gamma imaging, proton therapy, artificial intelligence, machine learning

## 1. Introduction

Treatment verification in proton therapy (PT) is highly desirable as it paves the way towards full exploitation of the favorable depth dose characteristics of protons and, thus, better sparing of healthy tissue[1,2]. Current state-of-the-art proton treatments use conservative treatment field arrangements and large margins around the tumor to mitigate the effect of proton range uncertainties[3], originating mainly from the conversion of the information in the computed tomography (CT) image to the stopping power ratio, deviations in patient positioning, and anatomical changes during treatment[4].

As clinical proton beams typically stop inside the patient, different strategies have evolved for treatment verification[2]. The detection of secondary particles and radiation like prompt gamma-rays, which are emitted as a result of the de-excitation of target nuclei, can be used to infer the proton penetration depth in near real-time without additional dose to the patient[5–7]. Various techniques exist[8], exploiting different properties of the emitted prompt gamma (PG) radiation, but only prompt gamma imaging (PGI) – by means of a collimated slit camera – has already shown first success in proof-of-principle pencil beam scanning (PBS) patient treatments[9,10]. However, manual interpretation of the spot-wise range shifts – extracted by a comparison between PGI reference profiles based on the patient anatomy in the planning CT images and the measured PGI profiles – requires a lot of time and manpower, which potentially hinders the broad application of PGI within the clinical workflow. Therefore, a system for the automatic evaluation of measured PGI data is necessary, which should be capable of differentiating relevant from non-relevant treatment deviations and potentially also classifying the underlying source for relevant deviations, e.g. into anatomical changes, setup errors, and errors of the stopping power prediction. Examples of possible treatment interventions in a future workflow could be the following: the detection of setup errors could potentially trigger new patient positioning using planar X-ray images, whereas the detection of anatomical changes would trigger new 3D image acquisition and possibly replanning. Real-time detection and classification of these errors would represent an important step in the direction of automatic treatment verification and online adaptive proton therapy[11].

In a previous study[12], we have already shown that automatic detection and classification of treatment deviations is possible using simulated data without any measurement effects from the PGI slit camera system of Ion Beam Applications (IBA, Louvain-la-Neuve, Belgium)[7,13,14]. Global features calculated from all PBS spots with reliable PGI information were used to heuristically construct a decision tree model to classify treatment deviations. This approach, however, discards potentially valuable information about the spatial relation of PGI spots and may not be applicable in a more realistic scenario.

In this manuscript, we propose convolutional neural networks (CNNs) for the automatic detection and classification of treatment deviations in realistically simulated PGI data of head and neck cancer patients. The use of CNNs eliminates the need to find task-specific features as CNNs can directly utilize the spot-wise PGI data as input for the classification task, thus incorporating the spatial structure of the measured range shifts of the whole treatment field. To quantify the suitability of CNNs, we compare their performance with conventional machine learning classifiers trained on handcrafted features that we established recently[12]. Aiming towards a future clinical translation, the data set under investigation contains different error scenarios that may occur in patient irradiations. In addition, we generalize our simulated data by considering Poisson noise of the PGI signal and the uncertainty of the positioning of the PGI slit camera with respect to the patient.

## 2. Materials and methods

### 2.1. Study design

We investigated the approach of using CNNs for the detection and classification of relevant treatment deviations based on optimal (I), noisy (II), and noisy PGI data with the uncertainty of camera positioning (III), as shown in Figure 1. For comparison, an approach using conventional machine learning (ML) classifiers, namely logistic regression (LR), multi-layer perceptron (MLP), random forest (RF), and support vector machine (SVM), based on five previously established handcrafted features[12] was used. For both methods, training was performed on an exploratory cohort including data sets from eight patients and one phantom (n=9) using patient-wise leave-one-out cross-validation (LOO-CV) and the resulting model ensembles were tested on an independent test cohort (n=4). Model performance was evaluated using multi-class accuracy as well as binary accuracy, sensitivity, and specificity for the case of multi-class (7 classes) and binary classification, respectively.

## 2.2. Prompt gamma simulations

### 2.2.1 Patient data

The basis of this study are CT imaging data of 12 patients and of an anthropomorphic heterogeneous head phantom (n=13) with tissue equivalent materials (CIRS, Norfolk, USA) as well as PBS treatment plans already used in a previous study[12]. The phantom data set increases the size and variability of the training data and was included as we did not see any notable differences in extracted PGI range shift data between patients and phantom.

The imaging data consist of a dual-energy CT (DECT) scan for the phantom as well as the clinical single-energy planning CT (pCT) and 2 - 5 (median 4) control CT (cCT) scans from patients with loco-regionally advanced head and neck squamous cell carcinoma, previously treated with intensity-modulated photon therapy at our institution between January and July 2016[15]. The conversion from CT values to stopping power ratio (SPR) values was conducted using Hounsfield look-up tables (HLUTs). For the phantom data set, the HLUT was applied to a pseudo-monoenergetic data set derived from a weighted superposition of the DECT scans[16].

The clinical target volumes (CTVs) and organs-at-risk (OARs) were defined in all CT scans by an experienced radiation oncologist. PBS treatment plans for both, patients and the phantom, were generated in RayStation 5.99 (RaySearch Laboratories AB, Stockholm, Sweden) using three beams and a simultaneous integrated boost to achieve a prescribed dose of 70 Gy to the high-risk CTV and 57 Gy to the low-risk CTV in 33 fractions. A constant relative biological effectiveness (RBE) of 1.1 for proton beams was assumed. Robust optimization was used to account for range uncertainties of ±3.5% and setup uncertainties of 3 mm[17]. More detailed information about the planning constraints can be found in a previous publication[15]. For each plan, the treatment field that delivers the highest dose to the high-risk CTV is considered for potential PGI range monitoring by the IBA slit camera[7,13,14] within this study resulting in PGI data from 13 unique treatment fields.

### 2.2.2 Definition of ground truth classes

To investigate the capability of CNNs to detect and classify treatment deviations, three fundamental error scenarios were introduced to both the patient and phantom data:

- Setup errors (±0.5 mm, ±1 mm, ±3 mm) by shifting the pCT images in beam direction,
- Range prediction errors (±0.5%, ±1%, ±2%, ±3.5%) by globally changing the SPR, and
- Anatomical changes by overwriting the CT number in the phantom data set and by using cCT data with real anatomical changes in the patient data set, respectively.

All cases were reviewed for the demand of treatment adaptation[12]. Each case was thereby classified as a clinically relevant (RE) or non-relevant deviation (NR), respectively, which is described in detail in the following paragraph.

Scenarios with setup errors of ±3 mm, as well as range prediction errors of ±2% and ±3.5%, were defined as relevant treatment deviations and labeled as relevant setup errors (SE) and relevant range prediction errors (RP), respectively. These scenarios correspond to treatment deviations that we aim to detect in a possible application of PGI in a clinical online adaptive workflow. In contrast to previous work[12], the cases of range prediction errors of ±1% were classified as non-relevant, as they lead to a maximum proton range shift of about 1 mm for the investigated head and neck treatment fields. For the cases of anatomical changes, a comparison between the dose on the pCT and cCT images using dose-volume histogram (DVH) constraints was carried out. Consequently, patient cCT scans with no or only minor anatomical changes resulting in minor, clinically acceptable dosimetric deviations were labeled as NR, the others as relevant anatomical changes (AC).

Parts of the relevant cases were also used to create combinations of error scenarios (CB) consisting either of two or three of the fundamental classes (RP ±2%, SE ±3 mm, AC). These were added to resemble patient irradiation more closely, as a combination of errors is also not unlikely to occur over the course of fractionated treatment.

In summary, the ground truth classification yielded seven classes (±SE, ±RP, AC, CB, NR), out of which six are considered relevant deviations that would trigger treatment adaptation. For each of the 13 investigated treatment fields, 28 - 31 (median 30) cases were defined, resulting in 386 total cases with corresponding ground truth definitions. For binary classification, all six relevant error classes (±RP, ±SE, AC, CB) were combined to the RE class. Table1 gives an overview of the absolute frequency of ground truth classes during training and testing.

### 2.2.3 PGI simulation of reference and measurement-like profiles

PGI data were simulated assuming that the respective treatment fields were monitored with the IBA slit camera[7,13,14] currently used in a clinical study in our institution (DRKS00009224). Consisting of two rows of scintillation detectors behind a knife-edge slit collimator, this prototype scores PG-ray events in the energy region of 3 - 6 MeV. The response of the slit camera is modeled by the REGGUI software, which uses an analytical model based on Monte Carlo simulations[18].

The camera was virtually placed perpendicular to the beam direction projecting the PG emissions onto the detector elements resulting in a one-dimensional profile for each spot of the monitored PBS field. The distance of the collimator center from the detector and the beam central axis of the field was set to 16 cm and 20 cm, respectively. The resulting 10 cm field of view (FOV) at the central beam axis was focused on the high-risk CTV. For both, patients and the phantom, the pCT scan and the corresponding treatment plan were used to simulate reference prompt gamma profiles for each PBS spot of the monitored treatment field.

The REGGUI software calculates a range shift for each PBS spot of the treatment field by using a one-dimensional least-square matching of the simulated reference profile (based on the unmodified pCT) with an actual measured profile. In this simulation-only study, the measured profiles were replaced with simulated profiles of the corresponding relevant or non-relevant error scenarios.

To evaluate the suitability of the proposed models for more realistic measurement conditions, three different complexity levels were investigated:

- (I) Optimal PGI data:

  Independent of the monitor unit (MU) information in the treatment plan, each spot of the treatment field was assumed to have $10^9$ protons, like in a previous publication[12].
- (II) PGI data with realistic noise:

  Using the MU information in the treatment plan, the number of protons in each PBS spot was calculated and Poisson noise was simulated for each detector element according to the number of detected PG-rays. Consequently, high weighted spots were influenced less by noise and their range shift information was more reliable.
- (III) PGI data with realistic noise and an additional uncertainty of camera positioning:

  In addition to Poisson noise, a PGI camera positioning uncertainty in beam direction was added to imitate a realistic camera setup by random sampling from a normal distribution $N$(μ=0 mm, σ=0.5 mm). The mean μ and standard deviation σ of the distribution are derived from quality assurance measurements of the positioning reproducibility of the clinically used 2nd generation PGI camera system performed over a period of one year[14]. For each treatment field, the simulated global positioning error of the camera results in a shift of all measured PGI profiles compared to the corresponding reference profiles. Consequently, the positioning error translates directly to a change in the extracted range shift information for each spot of the treatment field.

## 2.3. Classification using CNNs

### 2.3.1 Preprocessing of PGI data

For all complexity levels, the PGI profiles were aggregated with those from their neighboring spots within the same energy layer by the application of a 7 mm Gaussian-shaped lateral convolution[10] to reduce statistical noise. Furthermore, the corresponding profiles were smoothed by a Gaussian kernel with full width at half maximum of 20 mm[19].

After shift determination, PBS spots with reliable PGI information were selected and converted to voxel grids used as an input for the CNN. This processing of the spot-wise PGI data consisted of three steps:

1. Definition of spots with reliable PGI information by applying four filters accounting for the location of the spots with respect to the camera FOV and the external contour, the amount of range mixing, and the length of the distal profile fall-off, as defined previously[12].
2. Additional elimination of spots with less than $0.25\times10^8$ protons after aggregation in the case of noisy PGI data simulations ((II) + (III)) as spots with too low counting statistics are prone to erroneous range shift retrieval. Spots having more protons than the used threshold showed an average PGI range shift precision smaller than 2 mm (cf. Supplement S1). The number of protons for all aggregated spots of the investigated treatment fields ranged from $0.02\times10^8$ – $3.56\times10^8$ (median $0.46\times10^8$).

3. Conversion of the spot-wise information, i.e. the range shift as well as proton number information, to $16\times16\times16$ voxel grids.
   - The in-plane spot coordinates (horizontal and vertical deflection in beam direction) and the mean proton range inside the patient (taking range mixing into account) were extracted from the REGGUI software for each spot. For the cases with range prediction errors, this coordinate system ensures that spots which have picked up comparable range errors travelling through tissue are sorted into the same depth bin of the voxel grid. As a result, cases with range prediction errors show a distinctive signature across all patient voxel grids.
   - The voxel grid size was defined based on the training data and the voxel size was kept constant through all patients ($1.46\times1.46\times0.63\ \text{cm}^3$, where the last value corresponds to the direction of the central beam axis).
   - The range shift information of all spots belonging to the same voxel was averaged using the number of protons of each spot for weighting. Additionally, the number of protons was summed over all spots per voxel and normalized to the range of (0, 1] for each treatment field.

For each scenario, this resulted in two $16\times16\times16$ voxel grids (range shift and proton number) that were used as two separate input channels for the CNN. Voxels without PBS spots contained a value of zero for both channels. For all used data sets, a percentage of 4.7% - 16.1% (median 9.0%) and 4.1% - 13.5% (median 7.9%) of voxels contained spot information for optimal and noisy data, respectively. An example case for the resulting PGI voxel grids as well as the underlying PGI profiles is shown in Figure 2 for all complexity levels.

### 2.3.2 Classification task and evaluation metrics

The CNNs were trained for binary as well as multi-class labels independently. For the binary evaluation, the ability of the model to differentiate between RE and NR was evaluated using the metrics sensitivity (SENS) and specificity (SPEC), defined as follows:

$$\mathrm{SENS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (1)$$

$$\mathrm{SPEC} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (2)$$

where true positive (TP) refers to the number of correctly predicted relevant error scenarios and true negative (TN) refers to the number of correctly predicted non-relevant deviations. False positive (FP) corresponds to non-relevant deviations that were incorrectly identified as relevant error scenarios and false negative (FN) corresponds to relevant error scenarios that were incorrectly identified as non-relevant deviations.

Additionally, the accuracy (ACC), both binary and multi-class, describing the fraction of correct classifications, was calculated:

$$\mathrm{ACC} = \frac{1}{n}\sum_{i=1}^{n}(Y_i = Z_i) \quad (3)$$

where $n$ is the number of cases, $Y_i$ is the prediction of the model for case $i$, and $Z_i$ represents the corresponding ground truth class.

### 2.3.3 CNN architecture

The CNN network architecture is loosely based on the LeNet architecture[20], but uses three-dimensional convolutional operations to fully utilize the spatial information of the PGI data input. The architecture consists of ten layers in total and comprises a convolutional and a fully connected part, cf. Figure 3. For the binary classification task, the same network is used but the last layer consists only of two hidden units corresponding to the two classes: relevant and non-relevant deviations. For both classification tasks, the predicted class corresponds to the label with the highest probability after the last layer.

The architecture was implemented using TensorFlow 2.1.0 and Python 3.7.7 and training was performed on an NVIDIA Tesla V100 with 32 GB and CUDA 10.1. Training for the multi-class classification task took on average 10 minutes, 3 hours, and 12 hours for the different complexity levels (I), (II), and (III), respectively. The trained CNN ensemble provided sub-second predictions on individual data sets using a normal desktop PC.

### 2.3.4 CNN training and testing

For each classification task, binary or multi-class, and each complexity level (I) – (III), the training of the CNNs was carried out in the following way: training was done from scratch using patient-wise leave-out-one cross-validation (LOO-CV) on eight patient data sets as well as the phantom data set (n=9). In this approach, each data set was used as a validation set once, which resulted in nine trained models.

The ensemble of nine trained models was then tested on an independent dataset of four patients (n=4). For each case, the predictions of the nine models were averaged and the class with the highest probability was used as the final prediction of the CNN ensemble.

To quantify the degree of certainty of the final model ensemble, the 95% confidence interval (CI) was calculated for each of the used metrics using 2000 bootstrap samples on the test data. For each bootstrap sample, the entirety of the test data was sampled with replacement and the corresponding metric was calculated on this sample. The final 95% CI was then defined by the 2.5th and 97.5th percentile of the resulting sample estimates for each respective metric.

For the investigated complexity levels, the data used for training and testing are summarized in the following

- (I): 266 labeled cases for training and 120 for testing.
- (II): for each case in (I), Poisson noise was simulated 20 times for each profile of the treatment field.

- (III): for each case in (II), 100 different randomly drawn camera positioning values were used for validation and for testing.

For each case in (I), there were a total of 20 cases with noisy shift values (II) and a total of 2000 cases with noisy shift values and added camera positioning uncertainty (III).

Training was performed by minimizing the cross-entropy loss for 500 iterations using the Adam optimizer[21] with a learning rate of $10^{-3}$, L2 regularization of $10^{-3}$, and a batch size of 64. For each cross-validation split, the training iteration where the current model achieved the highest accuracy – binary or multi-class depending on the classification task – on the held-out validation data was selected. Class imbalance was addressed by using class weighting in the loss function inversely proportional to the class frequencies in the used training set. For complexity levels (II) and (III), a batch size of 512 was used.

To limit overfitting of the model during training, three-dimensional spatial augmentations were implemented using the batchgenerators library[22] for all complexity levels (I) – (III). These augmentations included random rotation of the input data around the central beam axis between [-18°, +18°] and randomly zooming in and out of the input data by a maximum of 10%. For (III), a random camera positioning is drawn each time a training example is used and added to all range shift voxels with PGI information prior to the spatial augmentations. This positioning affects all spots equally and, therefore, directly translates into the averaged shift value in the 16×16×16 voxel grid.

### 2.4. Classification using conventional ML models

Four conventional ML models, namely logistic regression (LR), multi-layer perceptron (MLP), random forest (RF), and support vector machine (SVM), were trained on the same data and in the same way as the CNNs. Instead of the creation of two 16×16×16 voxel grids, five features were calculated for each case as described in a previous publication[12] after the first two filtering steps described in Section 2.3.1. No standardization was performed on these features.

The ML models were implemented using scikit-learn 0.22.1 in Python 3.7.7. For each model, the corresponding hyperparameters were optimized for 400 iterations using the TPESampler in optuna 2.3.0[23]. More detailed information can be found in Supplement S2.

For complexity level (III), the 100 camera positioning uncertainties used for the validation of the CNN models were introduced to the training data as no data augmentations during the training process can be applied to these conventional ML models. During testing, the same positioning uncertainties were used as for the testing of the CNNs.

## 3. Results

### 3.1. Binary classification

For the binary classification task of distinguishing between relevant and non-relevant deviations, the resulting CNN-based confusion matrices for all complexity levels (I) – (III) obtained during LOO-CV as well as for applying the trained CNN ensemble to the test data are shown in

Figure 4.

The CNN ensemble achieved binary test accuracies of 0.95, 0.96, and 0.93 for the respective complexity levels (I) – (III). In the most realistic scenario (III), the CNN ensemble reached a specificity of 0.88 and a sensitivity of 0.95 on the test data. All other evaluation metrics on the test data can be found in Supplement Table S2.

### 3.2. Multi-class classification

The ensemble of CNNs reached multi-class test accuracies of 0.83, 0.81, and 0.76 for the complexity levels (I), (II), and (III), respectively (

Figure 5). For complexity levels (II) and (III), all test cases of AC were misclassified. All other evaluation metrics of the test data can be found in Supplement Table S3.

### 3.3. Comparison with ML models

For the conventional ML models (LR, MLP, RF, SVM) trained on PGI feature data, maximum binary test accuracies of 0.97 (MLP), 0.96 (SVM), and 0.93 (SVM) were achieved for the complexity levels (I), (II), and (III), respectively. The corresponding confusion matrices for the validation data and test data are shown in Supplement S3 and Supplement S4, respectively. For the multi-class task, test accuracies of 0.86 (SVM), 0.81 (SVM), and 0.80 (MLP) were achieved for the complexity levels (I), (II), and (III), respectively. The corresponding confusion matrices for the validation data and test data are shown in Supplement S5 and Supplement S6, respectively.

A comparison between the test performances for the CNN and conventional ML models is shown in

Figure 6 for both classification tasks and all complexity levels. For the case of binary classification, the CNNs show no big decline in performance with the addition of measurement effects and perform comparably to the best performing ML classifier for all complexity levels. For the multi-class classification, the achieved accuracies are generally lower than those of the binary classification and both approaches show comparable performances for the first two complexity levels (except for the LR model). For the most realistic scenario, three out of four ML classifiers slightly outperform the CNN approach.

## 4. Discussion

An essential requirement for the integration of PGI in a clinical proton therapy workflow for treatment verification is the automatic evaluation of the measured data. In this manuscript, we assess the use of CNNs for the automatic detection and classification of common treatment deviations in proton therapy using realistically simulated PGI data from head and neck cancer patients and compare their performance with conventional ML algorithms trained on handcrafted features. Two measurement effects with notable impact on the extracted PGI range shifts, Poisson noise and the PGI camera positioning uncertainty, were included in the simulations to investigate their influence on the performance of the different classifiers.

Both the CNN as well as the conventional ML approach yield very comparable results for the binary and the multi-class classification task even after the addition of the investigated measurement effects. In the most realistic scenario, the CNNs were able to detect relevant treatment deviations with a sensitivity of 0.95 and specificity of 0.88 in the independent test data. For the more complex task of classifying the underlying treatment deviation, the CNNs classified most sources of deviation correctly, reaching an accuracy of 0.76 in the most realistic scenario. The best performing ML model even reached an accuracy of 0.80 for this case.

In the field of treatment verification, ML has already been used to convert simulated distributions of PG-rays[24], positron emitters[25–27], and acoustic signals[28,29] to dose distributions. The mentioned publications use simulations on phantom data and mostly only investigate high-weighted PBS spots. In this study, however, real patient CT data and PBS plans with realistic proton numbers per spot were used. Furthermore, measurement effects were incorporated into the simulations which were carried out using a software package already used for an ongoing clinical study of an existing PGI camera prototype. As a result, this is, to our knowledge, the most realistic simulation study utilizing ML methods for treatment verification. Moreover, our approach differs from the one of the mentioned publications as we do not aim to solve the task of establishing the relationship between the distribution of secondary irradiation effects (PG-rays, positron emitters, or acoustic waves) and dose. For the case of PG-rays, we judge this spot-wise dose prediction for all voxels along the beam path to be very challenging especially for low-weighted PBS spots in the presence of a realistic noise level and other measurement uncertainties[10,13,30]. With our approach, we try to mitigate the influence of noise using spot aggregation as well as averaging of PGI range shifts in each voxel and focus on the simpler task of utilizing patterns in the PGI data of the whole treatment field to differentiate between treatment-relevant and non-relevant changes (yes-no question). In a future clinical application, this could serve as an intervention mechanism to trigger volumetric imaging and eventually plan adaptation. In this case, the conversion of PGI information to dose would be obsolete.

We compared the performance of CNNs with conventional ML models trained on five established handcrafted features previously extracted from optimal PGI data[12]. The creation of these features was based on the used data set and ground truth classes (without the addition of the combination error class) and can be seen as the performance benchmark on optimal PGI data. In this study, the test performance of classifiers trained on these features only slightly decreased from 0.86 to 0.81 and 0.80 when adding Poisson noise and the uncertainty of camera positioning, respectively. This shows that the underlying handcrafted features were robust against both investigated measurement effects.

Compared to the feature-based approach, we utilized CNNs, which can be considered as more sophisticated classifiers that take the spatial information embedded in the PGI data into account. This approach yielded comparable test performances compared to the best performing baseline ML models for all complexity levels. Only for the most realistic multi-class classification task, three out of four ML models slightly outperformed the CNN approach (accuracy: 0.78-0.80 vs. 0.76). A possible reason for that could be that all hyperparameters of the ML models were optimized for each complexity case separately while the same CNN architecture including most of the hyperparameters (e.g. number of feature channels, strength of regularization) was used for all complexity levels. The main motivation behind this was to show the versatility of the used CNN architecture which we assume to be a prerequisite for the translational application of a similar CNN to clinical PGI data. While the CNNs already achieved competitive results for all complexity levels, their performance might further be improved by e.g. increasing the training data and fine-tuning the respective hyperparameters. Future investigations have to show whether CNN or ML approaches are better suited for our use case. In summary, it could be shown that even with the limited training data size, the CNNs were able to extract meaningful features from the PGI data, which were suitable for the given classification task.

A subset of the scenarios in this study, such as isolated range prediction and setup errors, were included to allow for a comparison with an earlier publication[12]. As the handcrafted features were tailored to these ground truth classes, a re-evaluation of the established feature set, e.g. by potentially adding or removing features, would be required for the detection and classification of more complex treatment deviations. Additional measurement effects of the PGI acquisition, such as neutron background and detector response, which were not considered in this study, could also force such an adaption of features. The use of CNNs, on the other hand, would eliminate the need to design task-specific handcrafted features by extracting the most important features directly from the input data given that enough training data are available. While CNNs seem to be a more straightforward approach for the application to more complex data, the easier interpretation could be an advantage of the feature-based classification approaches. Therefore, both approaches will be evaluated in future studies, e.g. applied to clinical PGI data.

Looking into the future of a possible clinical implementation of such an automatic detection system, the main aim would be the sole detection of relevant treatment deviations, while the classification of the underlying cause would be of secondary priority. As a high percentage of treatment sessions are expected to be delivered without relevant deviations from the treatment plan, a critical aim is to keep the absolute number of false positive alarms as low as possible. For the most realistic scenario (III), CNNs trained on binary data reached a high sensitivity of 0.95 compared to the specificity of 0.88 on the test data. Therefore, we investigated an approach to increase model specificity on the test data described in detail in Supplement S7. Instead of selecting the model that maximizes the binary accuracy on the validation data, we used a modified metric consisting of a weighted mean of sensitivity and specificity during the model selection process at the training stage. When weighting specificity three times as high as the sensitivity, the specificity of error detection on the validation data could be raised from 0.92 to 0.98 for the most realistic scenario. For the test data, the specificity was raised from 0.88 to 0.90 while lowering the sensitivity from 0.95 to

0.92. This proves that it is principally possible to optimize the network depending on the desired balance of specificity and sensitivity even though with limited benefit in this case.

Regarding the error source classification performance on the test data, the CNN ensembles could reliably detect (with a detection rate ≥ 81% for all complexity levels) almost all ground truth classes, even underrepresented classes (cf. Table 1) like relevant setup errors (SE) and range prediction errors (RPE). Only the cases of relevant anatomical changes (AC) were mostly predicted to be cases of combination errors (CB) for both the CNN as well as the ML approach. This is mainly caused by very similar PGI characteristics found in both classes. As the CB class is overrepresented in the training data, using the mean prediction of all models as the final prediction has a bias towards this class. The ML approach was not expected to work well for the CB class, as the underlying features were designed without the CB class. Interestingly, the spatial information, available to the CNN approach, did not help with this task. From a clinical perspective, this misclassification is not a severe limitation as both classes are relevant treatment deviations.

As CNNs cannot deal with missing values in their input data, a PGI range shift of zero was assigned to voxels in the PGI voxel grid that did not contain any PBS spot. This can be considered as an artificial generation of data points with perfect agreement between the PGI profile of the reference and error scenario and, thus, a proton range as in the planned scenario. However, the addition of a second channel consisting of the normalized number of protons per voxel as weighting information for the network eliminates this shortcoming. Empty voxels with no shift information can principally be ignored by the network whereas voxels with more protons deposit more dose and are therefore of higher importance in the classification task.

An obvious limitation of our study is the limited data size of 13 different head and neck treatment fields. Using the augmentations of rotations and scaling, the CNNs got varying PGI inputs and were able to generalize well onto treatment fields in the test data set.

However, an extended number of monitored treatment fields is needed for further research and might be able to improve model performance.

Camera positioning uncertainty was only investigated in beam direction as the positioning uncertainty in the other directions is comparably small[14] and has a minor effect on the measured PGI profiles. Similarly, only setup errors in beam direction were investigated. Setup errors in the other directions only lead to detectable PGI range shifts if the path length of the protons in the corresponding PBS spot changes. While relevant errors induced by such setup errors are in principle detectable, they would show ambiguous signatures in the PGI data dependent on the investigated treatment field and patient anatomy, comparable to the AC class.

In this study, information about surrounding organs at risk (OARs) is not considered. In the CNN approach, it would be possible to use an additional weighting factor for voxels that are close to OARs, where a deviation of the planned proton range would be of higher importance.

## 5. Conclusion

An automated approach to detect and classify treatment deviations in proton therapy was established using CNNs and simulated PGI data of three different complexity levels. In the most realistic scenario, taking Poisson noise and PGI camera positioning into account, the CNN approach detected treatment deviations with a sensitivity of 0.95 and specificity of 0.88 on the test data and could classify most of the underlying error sources reliably. The CNNs extracted meaningful features from the PGI data leading to comparable performance with ML classifiers trained on previously established handcrafted features. In conclusion, our study highlights the potential of a reliable, automatic interpretation of PGI data, which is highly desired for broad clinical application and a prerequisite for including PGI in an automated feedback loop for online adaptive proton therapy.

## Conflicts of Interest

OncoRay, HZDR and IBA have a research agreement in place. The authors report no conflict of interest.

## References

1. Parodi K. Latest developments in in-vivo imaging for proton therapy. *Br J Radiol*. 2020;93(1107):20190787. doi:10.1259/bjr.20190787

2. Knopf AC, Lomax A. In vivo proton range verification: A review. *Phys Med Biol*. 2013;58(15):131-160. doi:10.1088/0031-9155/58/15/R131

3. Lomax AJ, Boehringer T, Coray A, et al. Intensity modulated proton therapy: A clinical example. *Med Phys*. 2001;28(3):317-324. doi:10.1118/1.1350587

4. Paganetti H. Range uncertainties in proton therapy and the role of Monte Carlo simulations. *Phys Med Biol*. 2012;57(11):R99-R117. doi:10.1088/0031-9155/57/11/R99

5. Stichelbaut F, Jongen Y. Verification of the proton beam position in the patient by the detection of prompt gamma-rays emission. In: *39th Meeting of the Particle Therapy Co-Operative Group (San Francisco)*. 2003.

6. Min C-H, Kim CH, Youn M-Y, Kim J-W. Prompt gamma measurements for locating the dose falloff region in the proton therapy. *Appl Phys Lett*. 2006;89(18):183517. doi:10.1063/1.2378561

7. Smeets J, Roellinghoff F, Prieels D, et al. Prompt gamma imaging with a slit camera for real-time range control in proton therapy. *Phys Med Biol*. 2012;57(11):3371-3405. doi:10.1088/0031-9155/57/11/3371

8. Krimmer J, Dauvergne D, Létang JM, Testa É. Prompt-gamma monitoring in hadrontherapy: A review. *Nucl Instruments Methods Phys Res Sect A Accel Spectrometers, Detect Assoc Equip*. 2018;878(May):58-73. doi:10.1016/j.nima.2017.07.063

9. Xie Y, Petzoldt J, Janssens G, et al. Prompt gamma imaging for the identification of regional proton range deviations due to anatomic change in a heterogeneous region. *Br J Radiol*. 2020;93(1116):20190619. doi:10.1259/bjr.20190619

10. Xie Y, Bentefour EH, Janssens G, et al. Prompt Gamma Imaging for In Vivo Range Verification of Pencil Beam Scanning Proton Therapy. *Int J Radiat Oncol*. 2017;99(1):210-218. doi:10.1016/j.ijrobp.2017.04.027

11. Paganetti H, Beltran C, Both S, et al. Roadmap: proton therapy physics and biology. *Phys Med Biol*. 2021;66(5):05RM01. doi:10.1088/1361-6560/abcd16

12. Khamfongkhruea C, Berthold J, Janssens G, et al. Classification of the source of treatment deviation in proton therapy using prompt-gamma imaging information. *Med*

*Phys*. 2020;47(10):5102-5111. doi:10.1002/mp.14393

13. Nenoff L, Priegnitz M, Janssens G, et al. Sensitivity of a prompt-gamma slit-camera to detect range shifts for proton treatment verification. *Radiother Oncol*. 2017;125(3):534-540. doi:10.1016/j.radonc.2017.10.013

14. Berthold J, Khamfongkhruea C, Petzoldt J, et al. First-In-Human Validation of CT-Based Proton Range Prediction Using Prompt Gamma Imaging in Prostate Cancer Treatments. *Int J Radiat Oncol*. 2021;111(4):1033-1043. doi:10.1016/j.ijrobp.2021.06.036

15. Cubillos-Mesías M, Troost EGC, Lohaus F, et al. Including anatomical variations in robust optimization for head and neck proton therapy can reduce the need of adaptation. *Radiother Oncol*. 2019;131:127-134. doi:10.1016/j.radonc.2018.12.008

16. Wohlfahrt P, Möhler C, Hietschold V, et al. Clinical Implementation of Dual-energy CT for Proton Treatment Planning on Pseudo-monoenergetic CT scans. *Int J Radiat Oncol*. 2017;97(2):427-434. doi:10.1016/j.ijrobp.2016.10.022

17. Liu W, Frank SJ, Li X, et al. Effectiveness of robust optimization in intensity-modulated proton therapy planning for head and neck cancers. *Med Phys*. 2013;40(5):051711. doi:10.1118/1.4801899

18. Sterpin E, Janssens G, Smeets J, et al. Analytical computation of prompt gamma ray emission and detection for proton range verification. *Phys Med Biol*. 2015;60(12):4915-4946. doi:10.1088/0031-9155/60/12/4915

19. Priegnitz M, Barczyk S, Nenoff L, et al. Towards clinical application: prompt gamma imaging of passively scattered proton fields with a knife-edge slit camera. *Phys Med Biol*. 2016;61(22):7881-7905. doi:10.1088/0031-9155/61/22/7881

20. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278-2324. doi:10.1109/5.726791

21. Kingma DP, Ba J. ADAM: A method for stochastic optimization. *Int Conf Learn Represent*. Published online 2014. http://arxiv.org/abs/1412.6980

22. Isensee F, Jäger P, David WJZ, et al. batchgenerators - a python framework for data augmentation. Published online 2020.

23. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. *arXiv*. Published online 2019:2623-2631.

24. Liu C, Huang H. A deep learning approach for converting prompt gamma images to proton dose distributions: A Monte Carlo simulation study. *Phys Medica*. 2020;69(November 2019):110-119. doi:10.1016/j.ejmp.2019.12.006

25. Liu C, Li Z, Hu W, Xing L, Peng H. Range and dose verification in proton therapy using proton-induced positron emitters and recurrent neural networks (RNNs). *Phys Med Biol*. 2019;64(17):175009. doi:10.1088/1361-6560/ab3564

26. Ma S, Hu Z, Ye K, Zhang X, Wang Y, Peng H. Feasibility study of patient-specific dose verification in proton therapy utilizing positron emission tomography (PET) and generative adversarial network (GAN). *Med Phys*. 2020;47(10):5194-5208. doi:10.1002/mp.14443

27. Hu Z, Li G, Zhang X, Ye K, Lu J, Peng H. A machine learning framework with

anatomical prior for online dose verification using positron emitters and PET in proton therapy. *Phys Med Biol*. 2020;65(18):185003. doi:10.1088/1361-6560/ab9707

28. Yao S, Hu Z, Zhang X, et al. Feasibility study of range verification based on proton-induced acoustic signals and recurrent neural network. *Phys Med Biol*. 2020;65(21):215017. doi:10.1088/1361-6560/abaa5e

29. Yao S, Hu Z, Xie Q, Yang Y, Peng H. Further investigation of 3D dose verification in proton therapy utilizing acoustic signal, wavelet decomposition and machine learning. *Biomed Phys Eng Express*. 2022;8(1):015008. doi:10.1088/2057-1976/ac396d

30. Perali I, Celani A, Bombelli L, et al. Prompt gamma imaging of proton pencil beams at clinical dose rate. *Phys Med Biol*. 2014;59(19):5849-5871. doi:10.1088/0031-9155/59/19/5849

## Figure captions:

**Figure 1: Schematic of the study design. For each complexity level (I) – (III) of simulated prompt gamma imaging (PGI) data, convolutional neural networks (CNNs) as well as conventional machine learning models (LR – logistic regression, MLP – multi-layer perceptron, RF – random forest, SVM – support vector machine) were trained on an exploratory cohort (n=9) and the performance of each approach was evaluated on an independent test cohort (n=4) using different metrics depending on the classification task.**

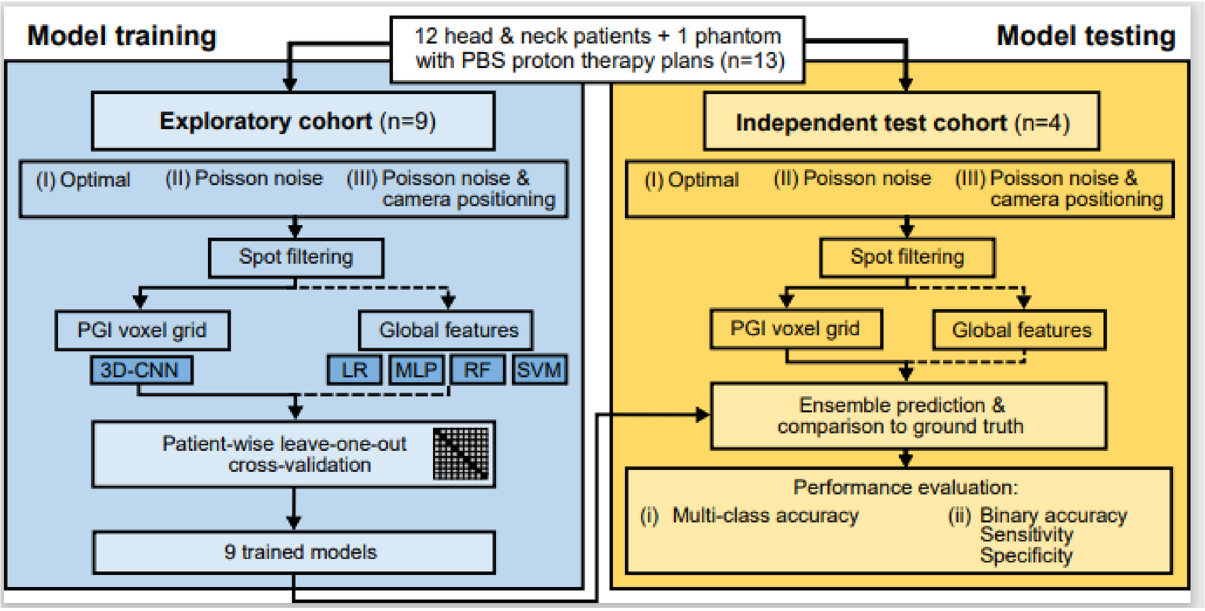


**Figure 2: Illustration of the three complexity levels (I) – (III) in terms of prompt gamma imaging (PGI) profiles (top) and 16×16×16 voxel grids (bottom) used as input for the convolutional neural networks (CNNs). The examples shown here are for the positive setup error class (+3 mm) and for the case of a camera positioning uncertainty of -1 mm (III).**

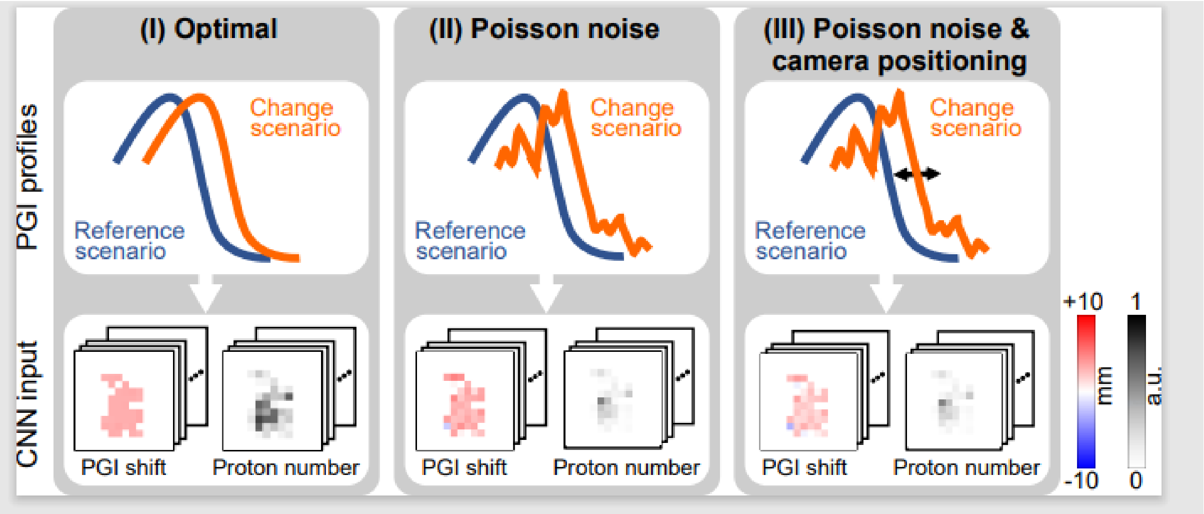
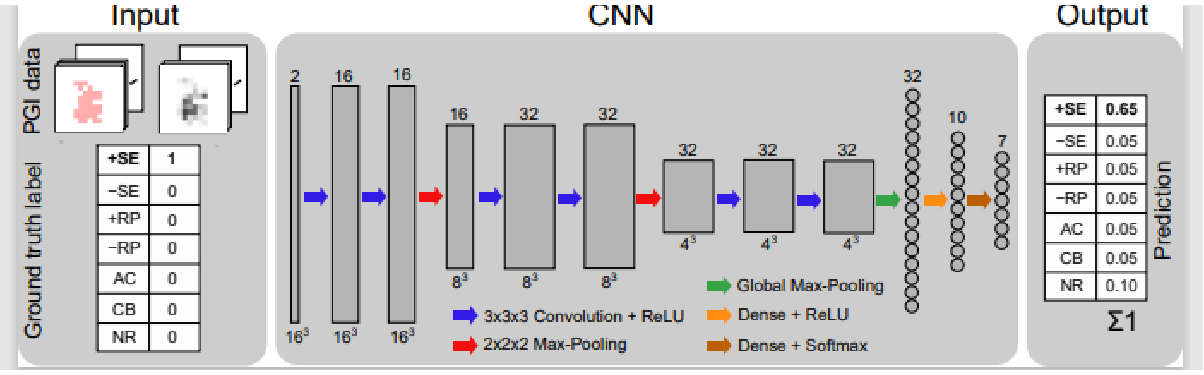


**Figure 3: Architecture of the used convolutional neural network (CNN) for the multi-class classification task (middle). Numbers above and below each block designate the number of feature channels and the matrix size at the given state, respectively. Numbers above dense layers indicate the number of hidden units. On the left, an exemplary two-channel data input (top) and the corresponding ground truth label (bottom) are shown, while the resulting output prediction of the network is shown on the right.**



| | |
|---|---|
| +SE | 1 |
| −SE | 0 |
| +RP | 0 |
| −RP | 0 |
| AC | 0 |
| CB | 0 |
| NR | 0 |

| | |
|---|---|
| +SE | 0.65 |
| −SE | 0.05 |
| +RP | 0.05 |
| −RP | 0.05 |
| AC | 0.05 |
| CB | 0.05 |
| NR | 0.10 |

**Figure 4: Normalized binary confusion matrices of the convolutional neural networks (CNNs) both on the validation data (top row) and on the test data (bottom row) for the three complexity levels (I) – (III). Additionally, the corresponding accuracy (ACC) is displayed. CNNs were trained to differentiate between relevant (RE) and non-relevant (NR) treatment deviations.**

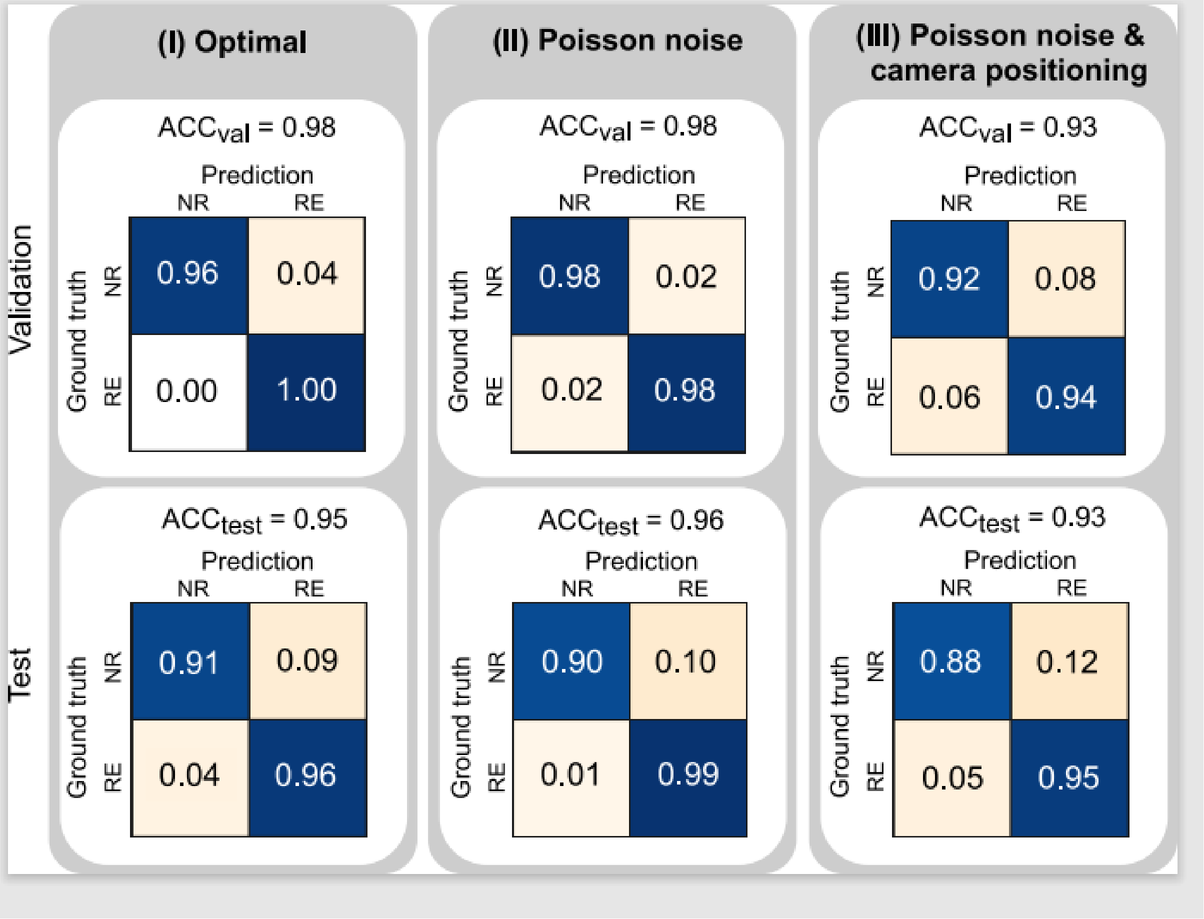
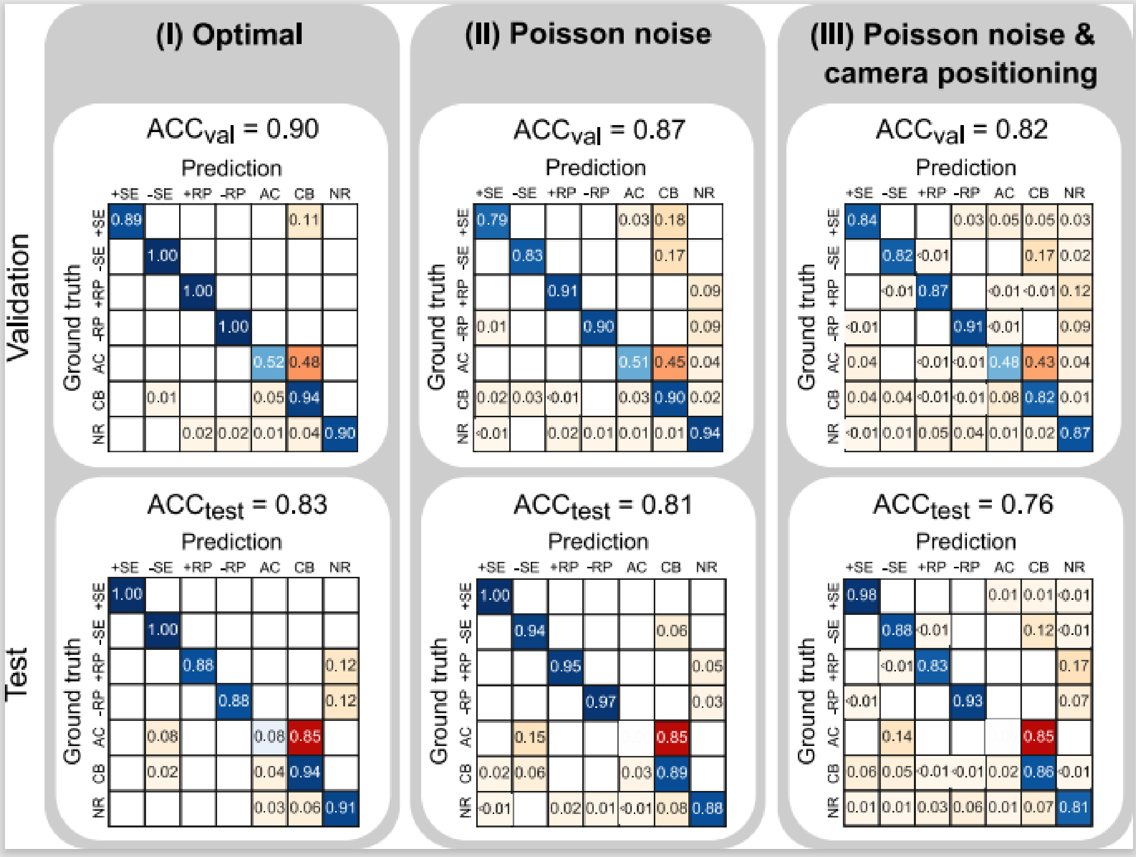


**Figure 5: Normalized multi-class confusion matrices of the convolutional neural networks (CNNs) both on the validation data (top row) and on the test data (bottom row) for the three complexity levels (I) – (III). Additionally, the corresponding accuracy (ACC) is displayed. The following classes of treatment deviations were utilized: relevant setup errors in beam direction (±SE), relevant range prediction errors (±RP), relevant anatomical changes (AC), or a combination of such errors (CB) as well as non-relevant changes (NR).**



(I) Optimal – Validation: $ACC_{val}$ = 0.90

| Ground truth \ Prediction | +SE | -SE | +RP | -RP | AC | CB | NR |
|---|---|---|---|---|---|---|---|
| +SE | 0.89 | | | | | 0.11 | |
| -SE | | 1.00 | | | | | |
| +RP | | | 1.00 | | | | |
| -RP | | | | 1.00 | | | |
| AC | | | | | 0.52 | 0.48 | |
| CB | | 0.01 | | | 0.05 | 0.94 | |
| NR | | | 0.02 | 0.02 | 0.01 | 0.04 | 0.90 |

(II) Poisson noise – Validation: $ACC_{val}$ = 0.87

| Ground truth \ Prediction | +SE | -SE | +RP | -RP | AC | CB | NR |
|---|---|---|---|---|---|---|---|
| +SE | 0.79 | | | | 0.03 | 0.18 | |
| -SE | | 0.83 | | | | 0.17 | |
| +RP | | | 0.91 | | | | 0.09 |
| -RP | 0.01 | | | 0.90 | | | 0.09 |
| AC | | | | | 0.51 | 0.45 | 0.04 |
| CB | 0.02 | 0.03 | <0.01 | | 0.03 | 0.90 | 0.02 |
| NR | <0.01 | | 0.02 | 0.01 | 0.01 | 0.01 | 0.94 |

(III) Poisson noise & camera positioning – Validation: $ACC_{val}$ = 0.82

| Ground truth \ Prediction | +SE | -SE | +RP | -RP | AC | CB | NR |
|---|---|---|---|---|---|---|---|
| +SE | 0.84 | | | 0.03 | 0.05 | 0.05 | 0.03 |
| -SE | | 0.82 | <0.01 | | | 0.17 | 0.02 |
| +RP | | <0.01 | 0.87 | | <0.01 | <0.01 | 0.12 |
| -RP | <0.01 | | | 0.91 | <0.01 | | 0.09 |
| AC | 0.04 | | <0.01 | <0.01 | 0.48 | 0.43 | 0.04 |
| CB | 0.04 | 0.04 | <0.01 | <0.01 | 0.08 | 0.82 | 0.01 |
| NR | <0.01 | 0.01 | 0.05 | 0.04 | 0.01 | 0.02 | 0.87 |

(I) Optimal – Test: $ACC_{test}$ = 0.83

| Ground truth \ Prediction | +SE | -SE | +RP | -RP | AC | CB | NR |
|---|---|---|---|---|---|---|---|
| +SE | 1.00 | | | | | | |
| -SE | | 1.00 | | | | | |
| +RP | | | 0.88 | | | | 0.12 |
| -RP | | | | 0.88 | | | 0.12 |
| AC | | 0.08 | | | 0.08 | 0.85 | |
| CB | | 0.02 | | | 0.04 | 0.94 | |
| NR | | | | | 0.03 | 0.06 | 0.91 |

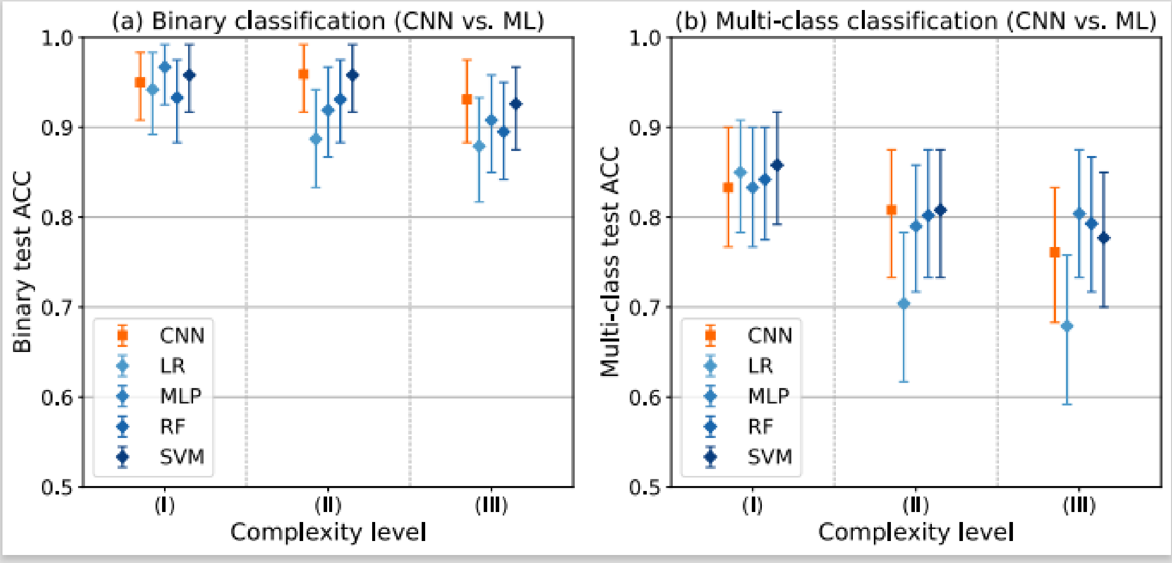
(II) Poisson noise – Test: $ACC_{test}$ = 0.81

| Ground truth \ Prediction | +SE | -SE | +RP | -RP | AC | CB | NR |
|---|---|---|---|---|---|---|---|
| +SE | 1.00 | | | | | | |
| -SE | | 0.94 | | | | 0.06 | |
| +RP | | | 0.95 | | | | 0.05 |
| -RP | | | | 0.97 | | | 0.03 |
| AC | | 0.15 | | | 0.00 | 0.85 | |
| CB | 0.02 | 0.06 | | | 0.03 | 0.89 | |
| NR | <0.01 | | 0.02 | 0.01 | <0.01 | 0.08 | 0.88 |

(III) Poisson noise & camera positioning – Test: $ACC_{test}$ = 0.76

| Ground truth \ Prediction | +SE | -SE | +RP | -RP | AC | CB | NR |
|---|---|---|---|---|---|---|---|
| +SE | 0.98 | | | | 0.01 | 0.01 | <0.01 |
| -SE | | 0.88 | <0.01 | | | 0.12 | <0.01 |
| +RP | | <0.01 | 0.83 | | | | 0.17 |
| -RP | <0.01 | | | 0.93 | | | 0.07 |
| AC | | 0.14 | | | 0.00 | 0.85 | |
| CB | 0.06 | 0.05 | <0.01 | <0.01 | 0.02 | 0.86 | <0.01 |
| NR | 0.01 | 0.01 | 0.03 | 0.06 | 0.01 | 0.07 | 0.81 |

Figure 6: Comparison of the test performances for the convolutional neural network (CNN) and the conventional machine learning (ML) based models (LR – logistic regression, MLP – multi-layer perceptron, RF – random forest, SVM – support vector machine) for both the binary (left) and multi-class (right) classification task. Error bars indicate the 95% confidence intervals of the test accuracy (ACC).


(a) Binary classification (CNN vs. ML)
Binary test ACC
CNN
LR
MLP
RF
SVM
(I)
(II)
(III)
Complexity level
(b) Multi-class classification (CNN vs. ML)
Multi-class test ACC
CNN
LR
MLP
RF
SVM
(I)
(II)
(III)
Complexity level