scientific reports

Check for updates

OPEN

Prediction of therapeutic intensity level from automatic multiclass segmentation of traumatic brain injury lesions on CT-scans

Clément Brossard, Jules Grèze, Jules-Arnaud de Busschère, Arnaud Attyé, Marion Richard, Florian Dhaussy Tornior, Clément Acquitter, Jean-François Payen, Emmanuel L. Barbier, Pierre Bouzat & Benjamin Lemasson✉

The prediction of the therapeutic intensity level (TIL) for severe traumatic brain injury (TBI) patients at the early phase of intensive care unit (ICU) remains challenging. Computed tomography images are still manually quantified and then underexploited. In this study, we develop an artificial intelligence-based tool to segment brain lesions on admission CT-scan and predict TIL within the first week in the ICU. A cohort of 29 head injured patients (87 CT-scans; Dataset1) was used to localize (using a structural atlas), segment (manually or automatically with or without transfer learning) 4 or 7 types of lesions and use these metrics to train classifiers, evaluated with AUC on a nested cross-validation, to predict requirements for TIL sum of 11 points or more during the 8 first days in ICU. The validation of the performances of both segmentation and classification tasks was done with Dice and accuracy scores on a sub-dataset of Dataset1 (internal validation) and an external dataset of 12 TBI patients (12 CT-scans; Dataset2). Automatic 4-class segmentation (without transfer learning) was not able to correctly predict the apparition of a day of extreme TIL (AUC = 60 ± 23%). In contrast, manual quantification of volumes of 7 lesions and their spatial location provided a significantly better prediction power (AUC = 89 ± 17%). Transfer learning significantly improved the automatic 4-class segmentation (DICE scores 0.63 vs 0.34) and trained more efficiently a 7-class convolutional neural network (DICE = 0.64). Both validations showed that segmentations based on transfer learning were able to predict extreme TIL with better or equivalent accuracy (83%) as those made with manual segmentations. Our automatic characterization (volume, type and spatial location) of initial brain lesions observed on CT-scan, publicly available on a dedicated computing platform, could predict requirements for high TIL during the first 8 days after severe TBI. Transfer learning strategies may improve the accuracy of CNN-based segmentation models.

Trial registrations Radiomic-TBI cohort; NCT04058379, first posted: 15 august 2019; Radioxy-TC cohort; Health Data Hub index F20220207212747, first posted: 7 February 2022.

Abbreviations

CT	Computed Tomography
AI	Artificial Intelligence
TBI	Traumatic brain injury
TILsum	Therapeutic intensity level summary
ICU	Intensive care unit
CNN	Convolutional neural network
AUC	Area under the curve

Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, Grenoble Institut Neurosciences (GIN), U1216, Eq. "Neuroimagerie Fonctionnelle et Perfusion Cérébrale", 38700 Grenoble, France. ✉email: benjamin.lemasson@univ-grenoble-alpes.fr

Traumatic Brain Injury (TBI) is a leading cause of death and disability in the world. Despite significant progress in their management, half of severe TBI patients will have long-term disabilities. One big issue is to have reliable tools to predict patient outcomes after TBI[1,2]. Models have been developed to predict outcome at 6 months post-trauma such as IMPACT[3] and CRASH[4], which include clinical and CT-scan data such as the presence of intracerebral hemorrhagic lesions, midline shift and compression of basal cisterns. However, analysis of CT imaging is qualitative and observer-dependent[5]. Such issue could be solved with the development of artificial intelligence (AI) applied to CT-scan imaging, providing CT-scan quantification[6–11] or automated delineation of traumatic brain lesions[12,13].

The BLAST-CT algorithm, developed to automatically delineate intraparenchymal hematoma (IPH), extra-axial hematoma (EAH), intraventricular hemorrhage (IVH) and perilesional oedema (Od) after severe TBI, is to our knowledge the most advanced segmentation tool of TBI lesions on CT-scans[14].

While predicting TBI patient outcome at 6 months using qualitative analysis of CT scan may be difficult, information contained in initial CT-scan could be used to predict short-term evolution such as the intensity of therapies required for each patient. In severe TBI patients, most therapies are directed to control intracranial pressure (ICP), a strong driver of outcome after severe TBI. Eight ICP-treatment modalities have been then collected to validate a daily scoring system, the TIL sum with a maximum score of 38 points[15]. A TIL sum of 11 points or more is considered as moderate-to-intense requirements for therapies. Although the presence of midline shift and compressed basal cisterns are usually considered as radiological signs of high ICP, nothing is said about the predictive value of CT-scan findings on TIL sum.

We hypothesized that an automated delineation of the most frequent traumatic brain lesions from initial CT-scan could predict a moderate-to-severe TIL sum assessed during the first week after admission to the intensive care unit (ICU).

In this study, we extracted metrics from brain CT-scans representing the volume, the type and the spatial location of injuries using automatic or manual segmentations. Two transfer learning strategies were used to re-train the BLAST-CT algorithm in order to improve the automatic segmentations. Finally, segmentation and classification models were validated using internal and external datasets of patients.

Methods

Data retrieval

Dataset 1

The first dataset contains 30 head injured patients admitted at the Universitary Hospital of Grenoble (CHUGA) between January 2020 and April 2021 (Radiomic-TBI cohort; NCT04058379). Inclusion was prospective, conditioned to patient agreement and an Abbreviated Injury Score (AIS) ≥ 3[16], corresponding to the presence of an injury visible in the CT-scan acquired the day of admission. The following clinical data were retrieved at the admission in the Intensive Care unit (ICU): Age, Glasgow Coma Scale (GCS), Mean Arterial Pressure (MAP), presence/absence of antiaggregants and Hemoglobin (Hb) rate. The data needed to compute the TILsum was retrieved daily during the 8 first days in the ICU. Finally, Marshall[17] and Rotterdam[18] scores were computed from the CT-scans.

Among the 30 patients of the cohort, one was excluded because of a primary admission outside of the CHUGA, as described on the flowchart on Fig. 1, and 84 CT-scans were finally acquired on these 29 patients. This dataset, characterized in Table 1, was split into a train, validation and test sub-datasets in a 60/20/20 proportion. For obvious independence reasons, all scans of a patient were in the same sub-dataset.

Dataset 2

The second dataset contains 12 patients suffering a severe non penetrating TBI (GCS ≤ 8) admitted at the Universitary Hospital of Grenoble between August 2018 and April 2021 (RadioxyTC cohort; Health Data Hub index F20220207212747). TILsum score during the 5 first days in the ICU was retrieved and TILsum until the 8th day was estimated from clinical reports, in order to estimate the overall maximum TILsum score. This dataset was used to perform an external validation. Characterization of this dataset can be found in the Supplementary Table S4.

Outcome

The TILsum score is computed daily from the list of interventions and treatments undergone by the patient in the ICU and is an integer between 0 and 38. Details about its computation can be found in the supplementary material S1. One can define a day of extreme management if the TILsum reaches 11 or more[19,20]. In this work, we tried to detect patients that underwent an extreme management day during the 8 first days in the ICU (group TILsum_High) from the others (group TILsum_Low).

Preprocessing

All 84 CT-scans were extracted from the Hospital Storage System in DICOM and then converted to NIfTI format thanks to the MP3 software[21]. The brain was extracted with a MATLAB-based Skull removal algorithm[22]. Then, all images from a patient were rigidly co registered to his first CT-scan obtained at admission using the FLIRT algorithm from the FSL toolbox[23] and finally resampled at $1mm^3$.

Segmentations

On Dataset 1, we segmented TBI lesions on the CT-scans in 6 different ways. First, we applied the DeepMedic-based[24] Convolutional Neural Network (CNN) called BLAST-CT[14], which aims to automatically segment 4 lesions typical of TBI : IPH, EAH, Od and IVH leading to the *BLAST-CT segmentation*. Then, this segmentation

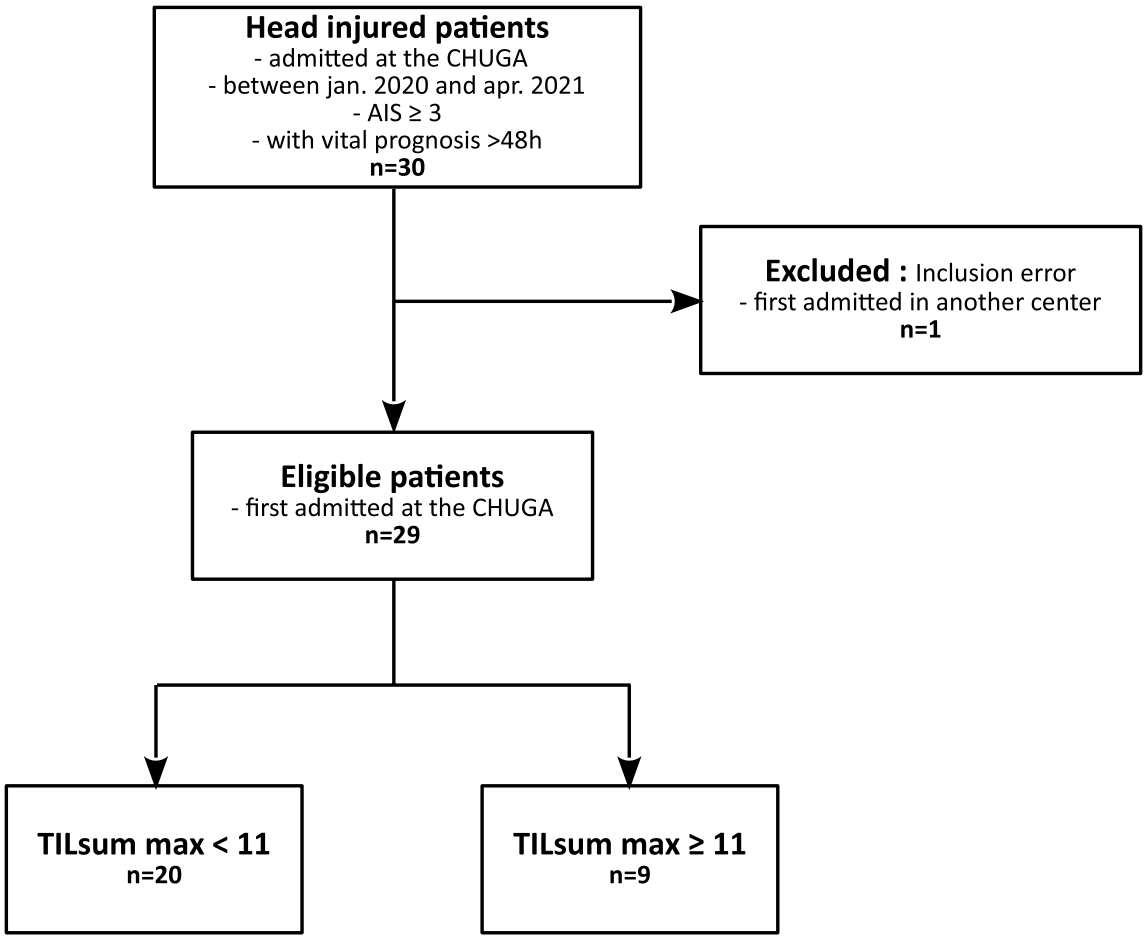

Figure 1. Flowchart of inclusions in Dataset1.

	TILsum_Low	TILsum_High
Nb of patients	20	9
Age (years)	48.3 ± 23.2 [18–79]	42.1 ± 18.6 [22–73]
Sex	13M/7F	8M/1F
Weight (kg)	70.8 ± 13.6 [51–100]	76.7 ± 6.1 [68–86]
Height (cm)	175.3 ± 9.8 [160–197]	175.5 ± 5.8 [169–185]
Glasgow Coma Score	9.3 ± 4.3 [3–15]	6.9 ± 4.0 [3–14]
Mean Arterial Pressure (mmHg)	87.8 ± 17.9 [41–115]	92.7 ± 14.6 [67–106]
Hemoglobin (g/l)	125 ± 27.7 [44–155]	129 ± 20.6 [95–164]
Presence Antiaggregants	4/20	0/9
Marshall score	2.6 ± 1.3 [2–6]	4.9 ± 1.1 [3–6]
Rotterdam score	3.0 ± 0.8 [2–5]	4.4 ± 1.2 [2–6]
TILsum maximum during the 8 first days in ICU	6.3 ± 2.6 [1–10]	18.9 ± 4.5 [11–28]

Table 1. Characterization of the Radiomic-TBI cohort (mean ± STD [Min–Max]). TILsum_High gathers the patients who have undergone at least one day with a TILsum equal or higher than 11 during their 8 first days in ICU. TILsum_Low gathers the other patients, without a day of extreme TIL.

was manually corrected by JAdB and AA, respectively anesthesiologist and neuroradiologist with 2 and 10 years of experience, using ITK-SNAP software[25], to obtain the *4-class manual segmentation*. This manual segmentation was then refined by splitting EAH between subdural hemorrhage (SDH), epidural hemorrhage (EDH) and subarachnoïd hemorrhage (SAH), and by distinguishing petechiae (Pe) from IPH, leading to the *7-class manual segmentation*. Finally, we used two transfer learning techniques to refine the automatic segmentations of BLAST-CT: a Fine-Tuning approach[26] in order to obtain an automatic 4-class segmentation named *CNN2* and a Transfer Learning approach[27] which lead to two 7-class automatic segmentations named *CNN3* and *CNN4*, depending on

Name of the model	Number of lesions segmented	Weights initialization	Trained on our data?	Training method
CNN1: BLAST-CT	4	BLAST-CT weights	No	Not applicable
CNN2	4	BLAST-CT weights	Yes	Fine tuning
CNN3	7	Random weights	Yes	Classical training
CNN4	7	CNN2 weights	Yes	Transfer learning

Table 2. Description of the 4 CNN using to automatically segment TBI lesions on CT-scans.

their initialization. The principal characteristics of the 4 different automatic segmentations evaluated are summarized on Table 2. On Dataset 2, manual segmentations at 4 and 7 classes were drawn by FDH (anesthesiologist, 1 year of experience), JAdB, and AA.

Features extraction

A structural atlas was retrieved and adapted from the FSL toolbox[23] and a CT-scan template was downloaded from[28]. The atlas and template were co-registered, first linearly thanks to the FLIRT algorithm[23] and then elastically using ANTS[29], to all CT-scans acquired at admission. All voxels inside the brain were thereby ascribed to one of the 11 areas of the atlas: Frontal (FL), Parietal (PL), Occipital (OL) and Temporal Lobes, but also Caudate, Cerebellum, Insula, Putamen, Thalamus, the rest of the brain, mainly composed of the ventricles, and the extra-cerebral space.

For the 3 segmentation approaches (BLAST-CT/4-class manual/7-class manual), we extracted, as illustrated on Fig. 2, the volume of each injury (respectively 4/4/7) in each area of the atlas (11), leading to respectively 44/44/77 metrics. For each of these segmentations, we combined these metrics to perform 7 experiments, each with a different set of metrics, detailed on Table 3.

Classification

For each of the 7 experiments and each of the 3 segmentations (BLAST-CT/4-classes manual/7-class manual), we trained a classifier to predict whether patients belong to the TILsum_Low or the TILsum_high group. Our classifier was designed using the PhotonAI toolbox[30], which proposes machine learning pipelines, containing data preprocessing, data augmentation, feature selection, hyperparameter optimization and model evaluations. Due to the small number of patients in our dataset, we used a state-of-the-art method: a nested cross-validation[31,32], on the train and validation sub-datasets to assure statistical robustness in the tuning and evaluation of our classifiers. This procedure is detailed on Supplementary Figs. S1 and S2.

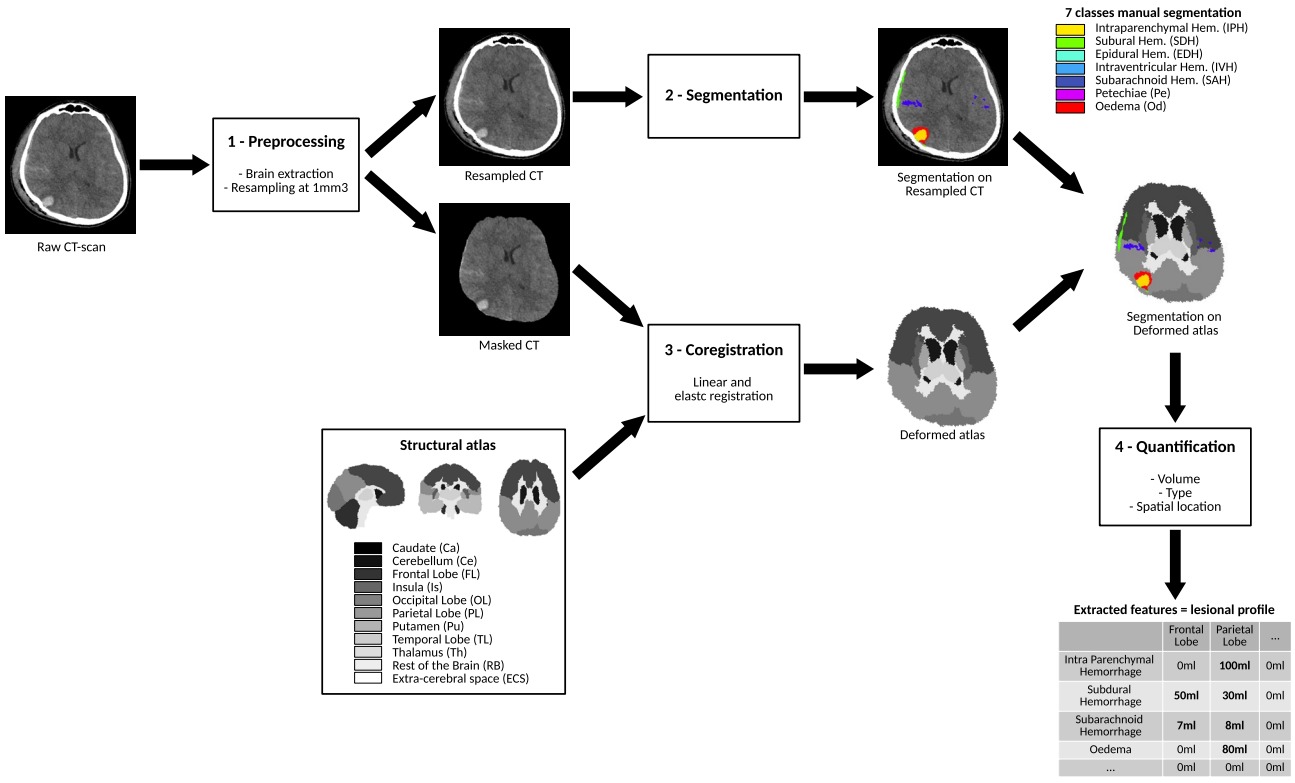

Figure 2. Overview of the lesion volume quantification by lesion type and spatial location.

Experiment number	Metrics	Number of metrics
Exp 1	Age, GCS, MAP, Hb, antiaggregants	5
Exp 2	Exp1 set + Marshall + Rotterdam scores	7
Exp 3	Global volume of lesion in the brain	1
Exp 4	Volume of lesion per type of injury (e.g., 300 mm^3 of Intraparenchymal Hemorrhage, …)	4 for 4-class segmentations, 7 for 7-class segmentation
Exp 5	Volume of lesion per spatial location (e.g., 300 mm^3 of lesion in the Frontal Lobe, …)	11
Exp 6	Exp4 set concatenated with the Exp5 set	15 for 4-class segmentations, 18 for 7-class segmentation
Exp 7	Volume of lesion per type of injury and spatial location (e.g., 300 mm^3 of intraparenchymal Hemorrhage in the Frontal Lobe, …)	44 for 4-class segmentations, 77 for 7-class segmentation

Table 3. Nature of input metrics for the 7 experiments.

Classification model selection

To compare our classification models, optimized and trained with different sets of metrics and different segmentations, we considered their global Area Under the Curve (AUC), of the Receiver-Operating Characteristic (ROC) curve, computed by the PhotonAI toolbox, and summarizing the sensibility and specificity of a binary classification model. As a direct implication of the classification procedure, each experiment resulted in 60 AUC values. To evaluate our models, we eventually considered the mean and standard deviation of these 60 AUC values. Significance of the difference of AUC distributions was assessed with non-parametric two-sided Mann–Whitney tests. After the training, we evaluated the importance of each feature of the best model, using the Mean Decrease impurity, in order to identify the key metrics that influence the model prediction.

Validations

We performed 2 validations. First we used the test sub-dataset of Dataset1 (6 patients) to evaluate our segmentations and classification models on the same type of data as the ones used for the training. Then, we used Dataset2 (12 patients) to evaluate our algorithms on data from the same center but from another study. We evaluated the accuracy of segmentation and classification models on these 2 validation datasets as described below:

Segmentation evaluation

We evaluated the segmentations by computing the DICE score[12], with the toolbox[33], between the 4 automatic segmentations and the related manual segmentation, on each lesion but also on the overall segmentation, obtained by merging all the classes into a unique class representing the lesional tissues (All). We then separately compared the two 4-class CNN and the two 7-class ones. Significance of the difference of Dice scores distribution was assessed with non-parametric two-side Wilcoxon tests.

Classification evaluation

We retrieved the best classification model from the nested cross validation using 4-class segmentations and 7-class segmentations, leading to 2 classification models: "*4-class Classification model*" and "*7-class Classification model*", both aimed at predicting our TILsum-based outcome. We then extracted the Volume/Type/Spatial location from the 6 different segmentations available (2 manuals and 4 automatics) and applied the related classification model. Finally, we measured and compared the accuracy of classification for each segmentation. Others evaluation metrics were measured and included in the supplementary material S1.

Ethics approval and consent to participate

The study Radiomic-TBI involving human participants was reviewed and approved by the French institution Comité de protection des personnes (Radiomic-TBI cohort; NCT04058379, first posted: 15 august 2019). Informed consent was obtained from all subjects and/or their legal guardian(s). The study RadioxyTC was also allowed by the French Direction de la Recherche Clinique et de l'Innovation and registered on the Health Data Hub (Radioxy-TC cohort; Health Data Hub index F20220207212747, first posted: 7 February 2022). Patients were individually informed, but no written informed consent was required, although patients had the opportunity to decline their participation in the study. These studies were carried out in accordance with the french regulation.

Results

Cohort characterization

The cohort characterization can be found on Table 1, for both groups TILsum_High and TILsum_Low used for the classification task. One can observe the unbalanced distribution of men and women and of antiaggregants in the two groups. As expected, the imaging scores (Marshall and Rotterdam) are lower in the TILsum_Low group than the TILsum_High one, whereas GCS are higher. The characterization of the second split of Dataset1 (train, validation and test sub-datasets) is provided in the Supplementary Table S5.

Classification model TILsum prediction

The mean and standard deviation of the AUC on the outer folds of the nested cross validation are shown on Table 4. The results show that clinical metrics fail to predict our TILsum based outcome. Adding manually estimated imaging scores improves the prediction power to 66 ± 24%. The global volume of injury does not

	Exp 1: clinical	Exp 2: clinical + imaging scores	Exp 3: volume total lesion	Exp 4: volume per type	Exp 5: volume per spatial location	Exp 6: concatenation volume per type and volume per spatial location	Exp 7: volume per type and spatial location
4 classes BLAST-CT segmentation	54 ± 24	66 ± 24	59 ± 24	61 ± − 27	65 ± 25	64 ± 25	60 ± 23
4 classes manual segmentation			63 ± 22	69 ± 24	71 ± 25	68 ± 26	74 ± 26
7 classes manual segmentation				75 ± 27		73 ± 24	**89 ± 17**

Table 4. AUC (Mean ± STD) on the outer folds of the models trained for 3 different segmentations (1 automatic and 2 manual) and 7 metrics sets. Results from the nested cross-validation procedure on data from the Train and Validation sub-datasets of Dataset1 (23 patients). Best result in bold.

show good predictions but splitting it by type or spatial location improves the prediction. The best model using 4-class segmentation is obtained for Exp7 and the 4-class manual segmentation (AUC = 74 ± 26%, Bias corrected and accelerated two-sided bootstrap 99-confidence interval [65, 83]). For the best model resulting from this nested cross-validation, called "*4-class Classification model*", the most important metrics, regarding the Mean Decrease impurity, are the volume of EAH in FL (importance of 38%) and in OL (31%). The overall best result is obtained for Exp7 and the 7-class manual segmentation (AUC = 89 ± 17%, Bias corrected and accelerated two-sided bootstrap 99-confidence interval [83, 94]). For this best model, called "*7-class Classification model*", the two most important metrics used for achieving the prediction are the volume of SDH in PL (importance of 47%) and in FL (33%). The second best model, obtained for type metrics (Exp4) and the 7-class manual segmentation (AUC = 75% ± 27%), relies the most on the volume of SDH (importance of 53%), SAH (26%), and IVH (20%). These results need to be seen through the prism of the small sample size.

Regarding the segmentations, the 7-class manual segmentation shows better results than the 4-class manual segmentation, which is itself better than BLAST-CT on all experiments. On the Exp7, according to the bilateral Mann–Whitney test, 7 classes manual segmentation performed significantly better than the 2 others segmentations ($p < 0.01$ compared to 4-class manual segmentation, $p < 0.001$ compared to BLAST-CT segmentation) and the 4 classes manual segmentation performed significantly better than the BLAST-CT segmentation ($p < 0.01$).

Validations

Segmentation evaluation

Internal validation. The comparison of Dice scores between 4-class automatic segmentations (CNN1 and CNN2) and the 4-class manual segmentation is displayed on Fig. 3 and detailed on Supplementary Table S6, for each lesion type and for the overall lesion, as well as an illustration of the resulting segmentations.

On every lesion, CNN2 (average DICE score on overall lesions = 0.63), showed statistically significantly better results than CNN1 (BLAST-CT) (0.34). The gain is particularly large on Od and IVH lesions, where CNN1 performs poorly.

The comparison of Dice scores between 7-class automatic segmentations (CNN3 and CNN4) and the 7-class manual segmentation are displayed on Fig. 3 and detailed on Supplementary Table S7, for each lesion type and for the overall lesion, as well as an illustration of the resulting segmentations.

On the overall lesion, CNN4 (average DICE score on overall lesions = 0.64), showed statistically significantly better results than CNN3 (0.55). On the EDH lesion, results are unexpected, as CNN3 is better than CNN4, but not significatively, probably due to the small sample size of images containing this lesion.

External validation. Comparison of the DICE scores computed on Dataset2 (12 patients) between segmentations resulting from CNN1 to CNN4 and manual segmentations are shown on Fig. 4 and detailed on Supplementary Tables S8 and S9.

Classification evaluation

Prediction accuracy of the 2 best classification models ("*4-class Classification model*" and "*7-class Classification model*", see section "Classification model TILsum prediction") on the test sub-dataset of Dataset1 (Internal validation—6 patients) and on the Dataset2 (External validation—12 patients) for 6 segmentations (BLAST-CT, our 3 automatic segmentations, and the 2 manual segmentations) are shown on Table 5.

Best accuracies were obtained for 7-class segmentations. CNN1 was only able to correctly classify 50% and 67% respectively internally and externally. The two transfer learning automatic segmentations were able to reach the same accuracy as manual segmentations internally, and with the 7-class segmentation, transfer learning automatic segmentation outperformed the external prediction made with the manual segmentation (10/12 vs 8/12).

Discussion

In this study, we quantified the ability of volume, spatial location and type of brain lesions observed on admission CT-scans to predict the therapeutic intensity level of TBI patients within the first week in ICU. Volumes of 7 different lesions on 11 structural zones were able to predict this outcome with a mean AUC of 89% and a standard deviation of 17%. Although the small sample size limits the conclusions of this work, the most influential

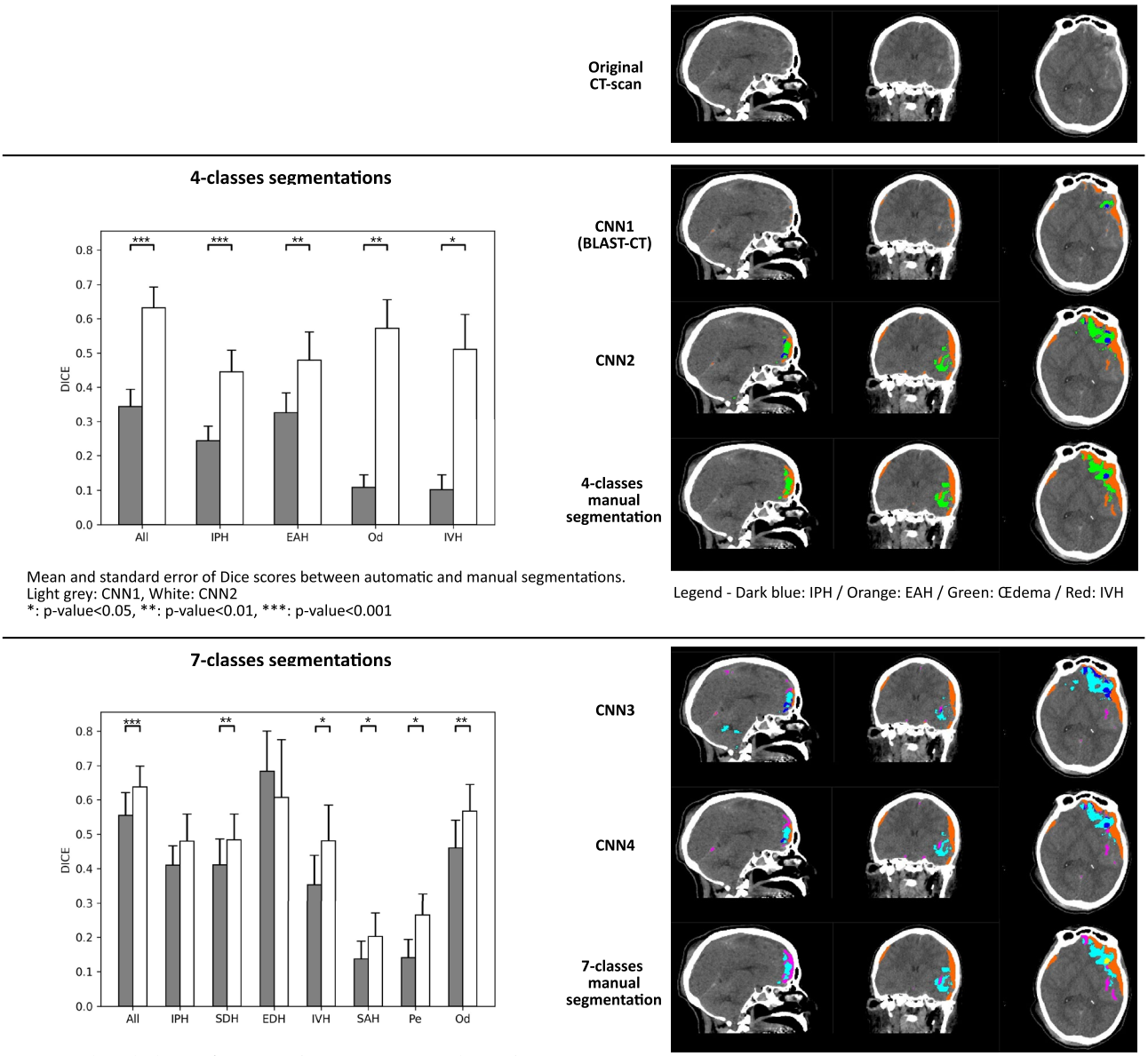

Figure 3. Barplots of the DICE scores (mean and standard error) computed on each lesion and on overall lesions between automatic and manual segmentations on the test sub-dataset of Dataset1 (6 patients—17 CT-scans). Upper part shows the comparison of CNN1 and CNN2, as well as an illustration of the resulting lesions. Lower part shows the comparison of CNN3 and CNN4, as well as an illustration of the resulting lesions. Significance was assessed with non-parametric two-side Wilcoxon tests. *$p < 0.05$.

metrics are the volume of lesions located in brain lobes, and especially the volume of subdural hemorrhage. This result is coherent with the medical experience about the large impact of SDH, often located in the frontal and parietal lobes, on medical care[34], and with the study pre-published by Rosnati et al. in which 6-months mortality of TBI patients has been predicted from frontal EAH lesions[11]. This recent study was conducted on more than 600 patients but only considered the 4-class BLAST-CT segmentation, when our study, conducted on way less patients, predicted a short-term outcome and exploited 7-class segmentations to highlight the influence of SDH.

We also highlighted the low predictive power of BLAST-CT to automatically predict our TILsum-based outcome. This lack of prediction could be explained by the poor segmentation obtained using BLAST-CT on our brain CT-scans. Indeed, although BLAST-CT was developed on large multicentric datasets (n = 839) the DICE obtained using BLAST-CT on our patients were lower than those published by Monteiro et al.[14]. In order to improve this automatic segmentation, we used 2 transfer learning approaches. First, we showed on the test sub-dataset of Dataset1 that the fine tuning of a deep learning algorithm on a small local dataset (n = 67 for training, consisting in the merge of train and validation sub-datasets) leads to a significantly increased segmentation

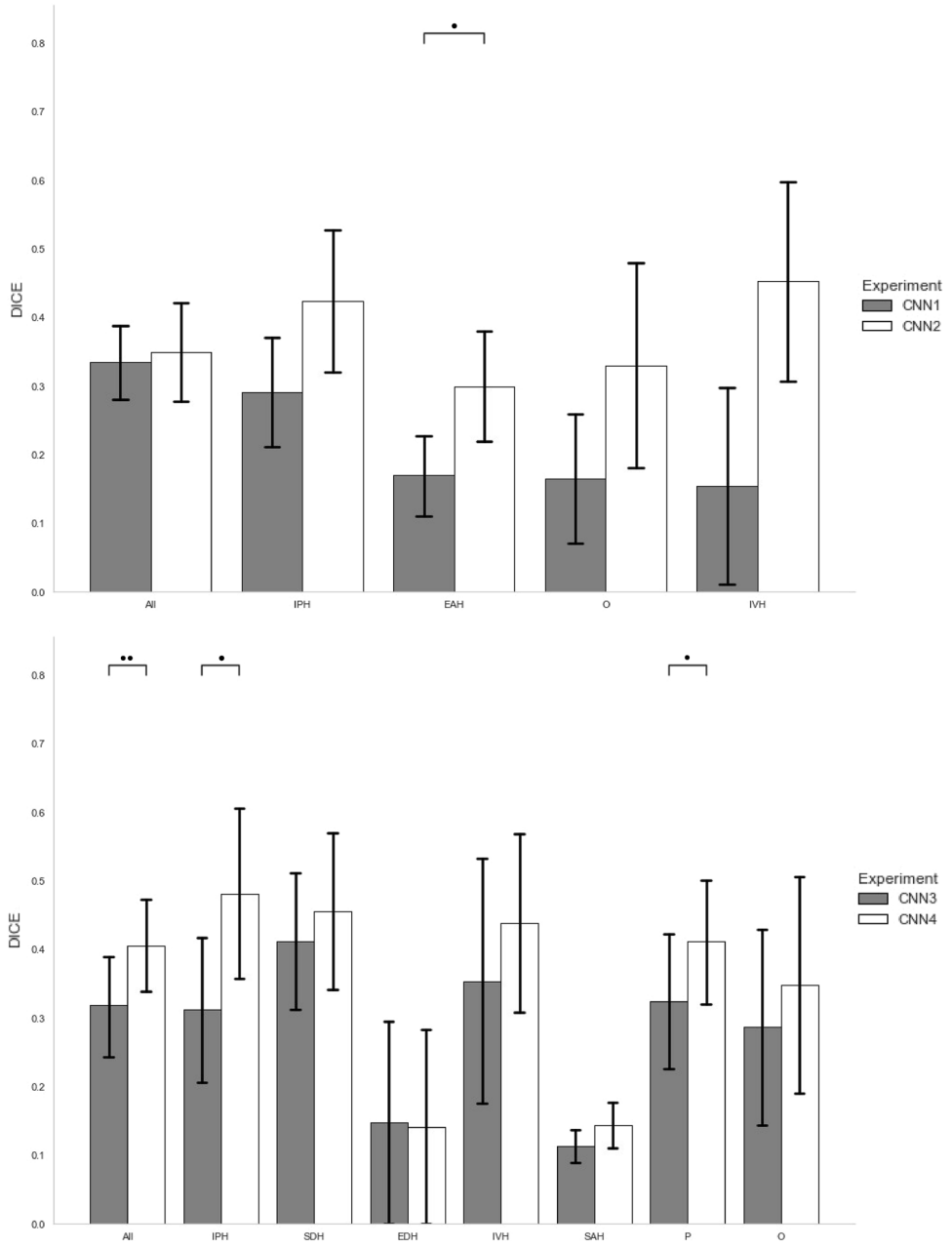

Figure 4. Barplots of the DICE scores (mean and standard error) computed on each lesion and on overall lesions between automatic and manual segmentations on Dataset2 (12 patients—12 CT-scans). Upper part shows the comparison of CNN1 (grey) and CNN2 (white). Lower part shows the comparison of CNN3 (grey) and CNN4 (white). Significance was assessed with non-parametric two-side Wilcoxon tests.*$p < 0.05$, **p-value < 0.01.

accuracy. This result can change the classical paradigm in which the objective of segmentation studies is to train an algorithm on a large multicentric dataset to learn and overcome the intersite variability. It might then be possible to easily fine-tune with a few images an already trained algorithm in order to learn the specificities of a study. The second approach, aimed at automatically segmenting 7 lesions from the 4-class segmentation algorithm by transfer learning, showed good results on highly represented lesions but was less accurate on poorly represented ones (such as petechiaes or EDH), a classical behavior in machine learning.

Finally, in order to link our segmentation work to a clinically relevant issue, we validated our results by using the improved segmentations to predict our clinical outcome on the test sub-dataset of Dataset 1. We showed that our improved segmentations predict the TILsum based criteria as the manual ones do. Segmentation and classification were then validated on a new external dataset (Dataset2) leading, as obtained on the internal validation, to better results with transfer learning approaches, and a prediction accuracy of 83% with automatic segmentation (10/12), better than the one obtained with manual segmentation (8/12), which is counterintuivite. This former

Classification model	Applied segmentation	Internal validation accuracy	External validation accuracy
4-class classification model	CNN1 (BLAST-CT)	50% (3/6)	67% (8/12)
	CNN2	67% (4/6)	67% (8/12)
	Manual4	67% (4/6)	67% (8/12)
7-class classification model	CNN3	83% (5/6)	75% (9/12)
	CNN4	83% (5/6)	83% (10/12)
	Manual7	83% (5/6)	67% (8/12)

Table 5. Accuracies of the classification on internal and external validation datasets, for 6 segmentations (4 automatic, 2 manual).

result could be explained by the unperfect manual segmentation, as illustrated on the Supplementary Fig. S3. In this case, segmenting using deep learning would undoubtedly have produced a more accurate segmentation.

Compared to recent automated hemorrhage segmentation literature, most of the studies do not discriminate SDH from others hemorrhage. While Yao et al. only focused on hematoma volume estimation, Monteiro et al. merged SDH with SAH and EDH. To our knowledge, the study of Farzaneh et al. is the only study to segment SDH, which is crucial to discriminate against short term evolution, as shown by our classification study. Farzaneh et al. study reached a Dice score of more than 0.75 by combining deep learning and classical image processing methods, outperforming our external validation SDH Dice score of 0.46. While differences in patients' inclusion and statistical evaluation methods might explain part of this score, it is probable that non deep learning post processing might improve deep learning segmentation by making them closer to neuroradiologists segmentations.

To our knowledge, we developed the first automatic tool to predict the intensity level of medical care from CT-scans of brain-injured patients, linking image processing to clinical care. In order to share our CT-scan quantification tool described on Fig. 2 and named CT-TIQUA v1.4, we encapsulated our best segmentation model, atlas registration, and volumes extraction in a docker container and integrated it on the computing platform VIP[36], that enables anyone to execute any pipeline on dedicated computing resources from the web-interface: https://vip.creatis.insa-lyon.fr/. This tool is the first to provide a 7-class segmentation of TBI injuries as well as registered atlas. Its universal utilization might allow to easily try it on another study or task.

Of course, this study has some limitations. First, our datasets are small, leading to unstable classification performances and all these results must be validated on larger and multicentric cohorts before any further use in clinical practice. Secondly, since this is the first study to predict the therapeutic intensity level, we cannot compare ourselves to the literature and evaluate the quality of our CT-scans quantification. To overcome this limitation, one will soon evaluate the prediction of the 6-months mortality to be able to compare our classification results with the ones of the most similar study conducted by Rosnati et al.[11].

To conclude, we believe that the automatic quantification of CT-scans to predict short-term outcome of TBI patients has the potential to bring reproducible and reliable information that can help improve clinical care. One must multiply the research studies in this way but also investigate the lesions evolution on repeated CT-scans, that might contain crucial information currently unused.

Data availability

The datasets analysed during the current study are available from the corresponding author on reasonable request.

Received: 16 May 2023; Accepted: 7 November 2023
Published online: 17 November 2023

References

1. Peeters, W. *et al.* Epidemiology of traumatic brain injury in Europe. *Acta Neurochir. (Wien).* **157**(10), 1683–1696 (2015).
2. Maas, A. I. R. *et al.* Traumatic brain injury: Integrated approaches to improve prevention, clinical care, and research. *Lancet Neurol.* **16**(12), 987–1048 (2017).
3. Hukkelhoven, C. W. P. M. *et al.* Predicting outcome after traumatic brain injury: Development and validation of a prognostic score based on admission characteristics. *J. Neurotrauma.* **22**(10), 1025–1039 (2005).
4. MRC CRASH Trial Collaborators, Perel, P., Arango, M., Clayton, T., Edwards, P., Komolafe, E., *et al.* Predicting outcome after traumatic brain injury: Practical prognostic models based on large cohort of international patients. *BMJ.* **336**(7641), 425–429 (2008).
5. Chun, K. A. *et al.* Interobserver variability in the assessment of CT imaging features of traumatic brain injury. *J. Neurotrauma* **27**(2), 325–330 (2010).
6. Kim, H. *et al.* Quantitative analysis of computed tomography images and early detection of cerebral edema for pediatric traumatic brain injury patients: Retrospective study. *BMC Med.* **16**, 1 (2014).
7. Yao, H. Machine learning and image processing for clinical outcome prediction: Applications in medical data from patients with traumatic brain injury, ulcerative colitis, and heart failure [Internet] [Thesis]. 2021 [cité 3 mars 2022]. Disponible sur: http://deepblue.lib.umich.edu/handle/2027.42/171316.
8. Chen, W., Belle, A., Cockrell, C., Ward, K. R. & Najarian, K. Automated midline shift and intracranial pressure estimation based on brain CT images. *J. Vis. Exp.* **74**, 1 (2013).
9. Chilamkurthy, S. *et al.* Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study. *The Lancet* **392**(10162), 2388–2396 (2018).
10. Rosa, E. D. l., Sima, D. M., Vyvere, T. V., Kirschke, J. S., & Menze, B. A Radiomics approach to traumatic brain injury prediction in CT scans. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. p. 732–5 (2019).

11. Rosnati, M., Soreq, E., Monteiro, M., Li, L., Graham, N. S. N., & Zimmerman, K., *et al.* Automatic lesion analysis for increased efficiency in outcome prediction of traumatic brain injury [Internet]. arXiv; 2022 août [cité 19 août 2022]. Report No.: arXiv:2208.04114.
12. Brossard, C. *et al.* Contribution of CT-scan analysis by artificial intelligence to the clinical care of TBI patients. *Front. Neurol.* **12**, 1. https://doi.org/10.3389/fneur.2021.666875/full (2021).
13. V. V, Gudigar A, Raghavendra U, Hegde A, Menon GR, Molinari F, *et al.* Automated detection and screening of traumatic brain injury (TBI) using computed tomography images: A comprehensive review and future perspectives. *Int. J. Environ. Res. Public Health.* **18**(12), 6499 (2021).
14. Monteiro, M. *et al.* Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: An algorithm development and multicentre validation study. *Lancet Digit. Health* **2**(6), e314–e322 (2020).
15. Zuercher, P. *et al.* Reliability and validity of the therapy intensity level scale: Analysis of clinimetric properties of a novel approach to assess management of intracranial pressure in traumatic brain injury. *J. Neurotrauma.* **33**(19), 1768–1774 (2016).
16. Greenspan, L., McLellan, B. A. & Greig, H. Abbreviated injury scale and injury severity score: A scoring chart. *J. Trauma.* **25**(1), 60–64 (1985).
17. Marshall, L. F. *et al.* A new classification of head injury based on computerized tomography. *J. Neurosurg.* **75**, S14–S20 (1991).
18. Maas, A. I. R., Hukkelhoven, C. W. P. M., Marshall, L. F. & Steyerberg, E. W. Prediction of outcome in traumatic brain injury with computed tomographic characteristics: A comparison between the computed tomographic classification and combinations of computed tomographic predictors. *Neurosurgery.* **57**(6), 1173–1182 (2005).
19. Maas, A. I. R. *et al.* Standardizing data collection in traumatic brain injury. *J. Neurotrauma.* **28**(2), 177–187 (2011).
20. TBI-IMPACT. Therapy intensity level [Internet]. 2010. Disponible sur: http://www.tbi-impact.org/cde/mod_templates/T_TIL.9.1.pdf.
21. Brossard, C. *et al.* MP3: Medical software for processing multi-parametric images pipelines. *Front. Neuroinf.* **14**, 1. https://doi.org/10.3389/fninf.2020.594799/full (2020).
22. Najm, M. *et al.* Automated brain extraction from head CT and CTA images using convex optimization with shape propagation. *Comput. Methods Programs Biomed.* **176**, 1–8 (2019).
23. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *NeuroImage.* **62**(2), 782–790 (2012).
24. Kamnitsas, K. *et al.* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **36**, 61–78 (2017).
25. Yushkevich, P. A. *et al.* User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage.* **31**(3), 1116–1128 (2006).
26. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE Trans. Med. Imaging* **35**(5), 1299–1312 (2016).
27. Cheplygina, V., de Bruijne, M. & Pluim, J. P. W. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **54**, 280–296 (2019).
28. Rajashekar, D. *et al.* High-resolution T2-FLAIR and non-contrast CT brain atlas of the elderly. *Sci. Data.* **7**(1), 56 (2020).
29. Avants, B. B. *et al.* A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage* **54**(3), 2033–2044 (2011).
30. Leenings, R. *et al.* PHOTONAI—A Python API for rapid machine learning model development. *PLOS ONE.* **16**(7), e0254062 (2021).
31. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. ArXiv181112808 Cs Stat [Internet]. 10 nov 2020 [cité 18 févr 2022]; Disponible sur: http://arxiv.org/abs/1811.12808.
32. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**(1), 91 (2006).
33. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging.* **15**(1), 29 (2015).
34. Aromatario, M. *et al.* Traumatic epidural and subdural hematoma: epidemiology, outcome, and dating. *Medicina (Mex).* **57**(2), 125 (2021).
35. Farzaneh, N. *et al.* Automated segmentation and severity analysis of subdural hematoma for patients with traumatic brain injuries. *Diagnostics.* **10**(10), 773. https://doi.org/10.3390/diagnostics10100773 (2020).
36. Glatard, T. *et al.* A virtual imaging platform for multi-modality medical image simulation. *IEEE Trans Med. Imaging* **32**(1), 110–118 (2013).

Acknowledgements

The authors would like to thank the platform GRICAD of the University of Grenoble. Part of the results presented in this work were achieved using the CT-TIQUA v1.4 application through the Virtual Imaging Platform [36; Glatard et al.], which uses the resources provided by the biomed virtual organisation of the EGI infrastructure.

Author contributions

P.B., J.G. and B.L. designed the study. P.B., J.G. and M.R. supervised patient inclusions. J.A.dB., F.D.T. and A.A. manually segmented lesions on CT scans. Processing design was done by C.B., B.L., J.G. and C.A. C.B. and B.L. mainly conducted the processing and analyses. C.B. and B.L. wrote the first draft of the manuscript which was improved by J.G., P.B., J.F.P. and E.B. All authors read and approved the final manuscript. The patients/participants provided their written informed consent to publication.

Funding

This study was funded by the French Fondation des Gueules Cassées (grants number 17-2019, 15-2020 and 13-2021) and the Universitary Hospital of Grenoble. This research was supported by Fondation ARC pour la recherche sur le cancer (grant number SIGN'IT20181007790).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-46945-9.

Correspondence and requests for materials should be addressed to B.L.