Annals of Biomedical Engineering
https://doi.org/10.1007/s10439-023-03354-3

S.I. : CONCUSSIONS II

Efficient Generation of Pretraining Samples for Developing a Deep Learning Brain Injury Model via Transfer Learning

Nan Lin[1] · Shaoju Wu[1] · Zheyang Wu[2] · Songbai Ji[1,3]

Received: 15 June 2023 / Accepted: 18 August 2023

Abstract

The large amount of training samples required to develop a deep learning brain injury model demands enormous computational resources. Here, we study how a transformer neural network (TNN) of high accuracy can be used to efficiently generate pretraining samples for a convolutional neural network (CNN) brain injury model to reduce computational cost. The samples use synthetic impacts emulating real-world events or augmented impacts generated from limited measured impacts. First, we verify that the TNN remains highly accurate for the two impact types ($N = 100$ each; R^2 of 0.948–0.967 with root mean squared error, RMSE, ~ 0.01, for voxelized peak strains). The TNN-estimated samples (1000–5000 for each data type) are then used to pretrain a CNN, which is further finetuned using directly simulated training samples (250–5000). An independent measured impact dataset considered of complete capture of impact event is used to assess estimation accuracy ($N = 191$). We find that pretraining can significantly improve CNN accuracy *via* transfer learning compared to a baseline CNN without pretraining. It is most effective when the finetuning dataset is relatively small (e.g., 2000–4000 pretraining synthetic or augmented samples improves success rate from 0.72 to 0.81 with 500 finetuning samples). When finetuning samples reach 3000 or more, no obvious improvement occurs from pretraining. These results support using the TNN to rapidly generate pretraining samples to facilitate a more efficient training strategy for future deep learning brain models, by limiting the number of costly direct simulations from an alternative baseline model. This study could contribute to a wider adoption of deep learning brain injury models for large-scale predictive modeling and ultimately, enhancing safety protocols and protective equipment.

Keywords Synthetic data · Transfer learning · Convolutional neural network · Transformer neural network · Traumatic brain injury

Glossary

Augmented impacts: impact kinematic profiles derived from real-world, measured impacts using component permutation, random rotational axis rotation and magnitude scaling.

Estimated responses: voxelized peak brain strains estimated from a deep learning model.

Finetuning samples or dataset: impacts and the corresponding (directly simulated) responses to finetune a deep learning model.

Measured impacts: impact kinematic profiles from real-world measurement (on-field or laboratory reconstruction).

Pretraining samples or dataset: impacts and the corresponding (estimated) responses used to pretrain a deep learning model so that to apply transfer learning.

Simulated responses: peak brain strains obtained from baseline finite element model simulation, further resampled

Associate Editor Stefan M. Duma oversaw the review of this article.

Resubmitted to the Annals of Biomedical Engineering (AMBE) August 16, 2023.

✉ Songbai Ji
sji@wpi.edu

[1] Department of Biomedical Engineering, Worcester Polytechnic Institute, 60 Prescott Street, Worcester, MA 01605, USA

[2] Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609, USA

[3] Department of Mechanical Engineering, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, USA

Published online: 29 August 2023

into a voxelized format (at an isotropic spatial resolution of 4 mm in this study).

Synthetic impacts: impact kinematics (specifically, rotational velocity profiles) generated based on feature distributions obtained from measured impacts.

Introduction

Finite element (FE) models of the human brain simulate the transient process of head impact based on physical laws. They provide spatially and temporally detailed mechanical responses such as strain, strain rate, and stress of the entire parenchyma. These responses are critical to the understanding of why and where brain injury occurs [25] but are difficult or infeasible to measure in a live brain under injury-relevant conditions [2]. Therefore, there is consensus that brain injury models play a critical role in the investigation of biomechanical mechanisms of traumatic brain injury (TBI). Over the past several decades, brain injury models have significantly evolved in sophistication [14, 24, 37]. A major limitation, nonetheless, is that these models are notoriously demanding in computational resources in terms of both runtime and hardware. They are impractical for routine applications.

To dramatically improve efficiency while retaining high accuracy, several deep learning brain injury models have been developed [6, 11, 26, 33–35, 40]. The basic idea is to train a neural network by iteratively adjusting its weighting parameters to learn the complex and nonlinear but smooth and continuous high-dimensional mapping relationship between head impact kinematics and the resulting brain responses. If a deep learning model achieves sufficient accuracy relative to the baseline counterpart, it can then be used as an efficient surrogate. This could enable the underlying baseline FE model for large-scale impact simulations critical for rapid concussion risk estimation [6, 11, 40], incorporation of the cumulative effects from many subconcussive head impacts [27], and iterative design and testing of protective headgears [10]. As a result, the brain modeling community has recommended further integration of deep learning models into future TBI biomechanics research and practice [14]. The potential for routine and large-scale model simulations could have profound implications for injury biomechanics and other related fields in general, given that applications of deep learning techniques may well be extended beyond TBI biomechanics as focused here.

Nevertheless, to achieve high accuracy, a deep learning brain model usually requires a large amount of training samples. This poses two major challenges. First, generating response samples from impact kinematics still relies on computationally rather costly FE model simulations. For example, an earlier convolutional neural network (CNN) using ~ 5700 impacts in contact sports required ~ 8 months nonstop simulations [11] and ~ 10 months for ~ 3200 impacts from automotive impacts [33], even with parallel processing. These studies have already taken advantage of the head geometrical symmetry property relative to the mid-sagittal plane to halve the parametric space [16]. Another study using a fully connected network required an estimated total runtime of ~ 26 months to generate ~ 2500 response samples (without considering parallel processing [40]).

In addition, although more training samples are anticipated to improve estimation accuracy, it is unclear what an "optimal" or minimum training data size is for a given desired accuracy. For example, with ~ 5700 training samples, a high success rate (SR, a measure that considers the accuracy in both magnitude and spatial distribution pattern) of up to 97.1% was achieved [11]. When the training sample size was substantially reduced to ~ 1400, the SR only dropped to 86.2% for a morphologically individualized CNN *via* transfer learning (i.e., ~ 75.4% reduction in sample size vs. ~ 11.2% decrease in SR) [22]. Transfer learning is to initiate the neural network weighting factors based on a converged network instead of using random values; thus, expedites training [5]. Identifying a minimum training sample size is important to limit the cost of direct model simulations.

The second challenge is that real-world measured head impacts are usually difficult to acquire, and only relatively small datasets have been reported (e.g., dozens for laboratory reconstructed impacts [30], a few hundred for on-field measurements [13, 34, 40] or automotive impacts [9, 33]). As a work-around, data augmentation has been used to increase training sample size [11, 34, 35]. Measured impacts from helmet testing [6, 10] or through dummy head model simulation [40] can reach thousands, but their kinematic profiles are relatively "simple" in "feature" space. They are mostly composed of single peak/single rotational axis. Thus, they may not be representative of much more complex real-world impacts on live humans, where multiple peaks and rotational axes during the impact temporal window seem more common [13].

Given the previously simulated large amount of head impacts now available, here we explore how best to design an efficient training strategy to develop a deep learning brain model. There are two aims of this study. First, we evaluate whether a recently developed transformer neural network (TNN) of high accuracy (e.g., coefficient of determination, $R^2 > 0.99$, for spatially detailed peak strains) can be utilized to efficiently generate brain response samples without relying on the costly direct simulation from the baseline FE model. With sufficient accuracy, the response samples can then be employed to pretrain a CNN model [33] before finetuning *via* transfer learning with additional directly simulated training samples. This is anticipated to

reduce training samples required compared to baseline training without transfer learning.

Second, we also explore whether it is feasible to generate synthetic impacts based on kinematic feature characteristics from limited measured impacts. A data-driven emulator based on principal component analysis (PCA) was previously developed to generate synthetic impacts [1]. However, it does not allow direct control of kinematic profile complexity, which, intuitively, may have implications for the accuracy of the surrogate deep learning model. If synthetic data would facilitate model training when limited real-world training samples are available, as similarly found in image recognition [21] and car detection [32] problems, they may alleviate some burden of having sufficient measured data that could be difficult to obtain in the real-world.

We expect findings from this study may quantify the effectiveness of pretraining for CNN model development in the context of TBI biomechanical modeling. While these findings may not directly benefit deep learning brain models that have already been developed, they could offer important insight into an economical training strategy for a *future* deep learning surrogate model aimed at efficiently replicating responses of an alternative or an upgraded baseline FE model. As fresh impact-response samples are necessary from direct and costly model simulations, a guideline for generating training samples and training strategies could be valuable, especially if application of deep learning is to be expanded on a large scale [14], or to be potentially extended to microscale axonal injury models [28, 41] as well. These efforts could contribute to a wider adoption of deep learning brain injury models for large-scale predictive modeling and ultimately, enhancing safety protocols and protective equipment.

Methods

Figure 1 illustrates the overall procedure of the study. Two separate impact datasets (synthetic based on extracted features vs. augmented based on limited measured data) are fed into the TNN to efficiently generate their respective voxelwise maximum principal strain (MPS) of the whole brain without FE model simulation (seconds vs. weeks or months). The impact kinematics and the corresponding voxelwise peak MPS constitute pretraining samples. They are then used to initially train a CNN. The resulting pretrained CNN is further refined using finetuning training samples that are simulated directly from the baseline FE model, and then resampled into voxelwise MPS. The resulting target CNN is finally used for performance evaluation on a separate testing dataset. The details of each step are elaborated in the following sections.

Synthetic Impacts

Feature Extraction

Assuming a rigid body skull, head impact kinematic profiles about the three anatomical axes at the head center of gravity fully describe the head excursion. Linear acceleration, alone, generates little brain strain due to brain's near incompressibility. The exception is the inferior-to-superior component that could generate strain in the brainstem [33] due to the mobility along the neural axis, but which may be compensated for *via* superposition [10]. Therefore, isolated head rotational kinematics are sufficient for brain strain estimation [3, 18]. This simplifies model input. Both magnitudes and temporal locations of local extrema (peaks and valleys) in a rotational velocity profile are important "features" [4, 29] because they dictate the physical process of head acceleration and deceleration. Head impact profiles typically have various numbers of peaks, which lead to varied strain time

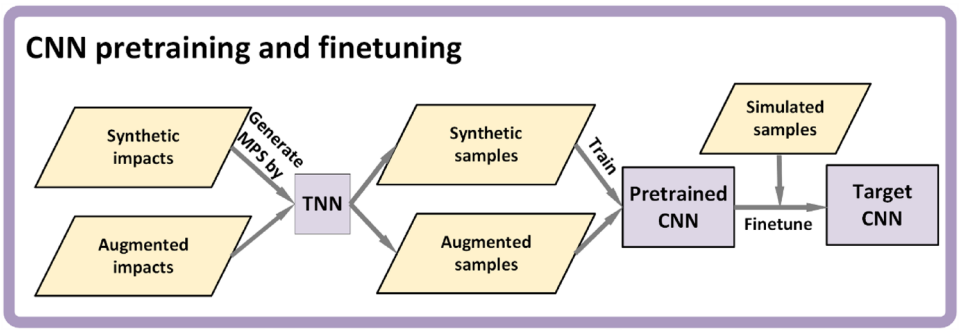

Fig. 1 Procedure for generating pretraining samples for a CNN deep learning brain model and finetuning. In this study, both pretraining and finetuning samples are based on the same baseline FE model for evaluation. For future applications, however, the finetuning dataset could be from an alternative baseline FE model that requires costly model simulations to generate fresh training samples.

history patterns [1, 15]. Based on these observations, therefore, we considered kinematic features of interest to include the number, magnitudes, and temporal locations of local extrema in the three rotational velocity profile components. Other kinematic features also exist, such as combinations of extrema magnitudes of accelerations, velocities, and other variants including integrations and differences [40]. They were not used here due to the loss of temporal information necessary to derive impact kinematic profiles.

The two sets of measured impacts were employed to extract features of interest ($N = 163$, from 53 reconstructed NFL impacts [30] and 110 mouthguard impacts [13]). The mouthguard impacts were mostly from 30 collegiate American football players, along with impacts from 2 professional box players and 1 mixed martial artist [13]. First, all notable peaks and valleys for each component of the rotational velocity profile were identified ("findpeaks.m" in Matlab). To avoid an excessive number of extrema, smaller prominences and depressions (< 10% of the maximal resultant velocity) were neglected. When two peaks or valleys were too close to each other (< 10 ms), only the one with the larger magnitude was collected, as they were mostly bumps, summits, or saddles in the same acceleration/deceleration phase. Given the importance of rotational velocity magnitude on brain strain, we ordered the resulting kinematic features based on the magnitude. Specifically, the maximum magnitudes in the three components were first arranged in a descending order, from which two ratios between a smaller value and its larger neighbor were calculated, along with their temporal locations relative to the maximum peak. Figure 2 illustrates the procedure for feature extraction. Specific anatomical axes were not considered in analyzing features.

Within each profile component, the identified extrema magnitudes were similarly ordered to calculate ratios between two consecutive (i.e., adjacent) neighbors. In addition, the extrema temporal locations relative to the ordered neighbor (of a larger magnitude) were also recorded. The resulting magnitude ratios (ranging from 0 to 1) and relative temporal locations (ranging from − 100 ms to 100 ms) were measured to construct their corresponding statistical distributions. Quantifying these features based on ordered peak magnitudes allowed an effective control of kinematic profile shapes to focus more on larger velocity extrema when generating synthetic data. Velocity profile shapes are known to be important to the induced brain strains [3, 43]. Figure 3 reports the statistical distributions of the features for the top three extrema and those of the number of extrema points.

Generating Synthetic Rotational Velocity Profiles

The ordered peak magnitude ratios and their corresponding relative temporal locations were used to generate synthetic rotational velocity profiles for each component independently. First, the number of local extrema was randomly generated following its distribution. Next, the temporal location of the largest peak with a normalized magnitude of 1.0 was randomly generated within a range of 30–70 ms, as the TNN estimates strains starting from 30 ms. The upper limit ensured sufficient time for brain strains to reach peak values within the given 100 ms time window [15, 23].

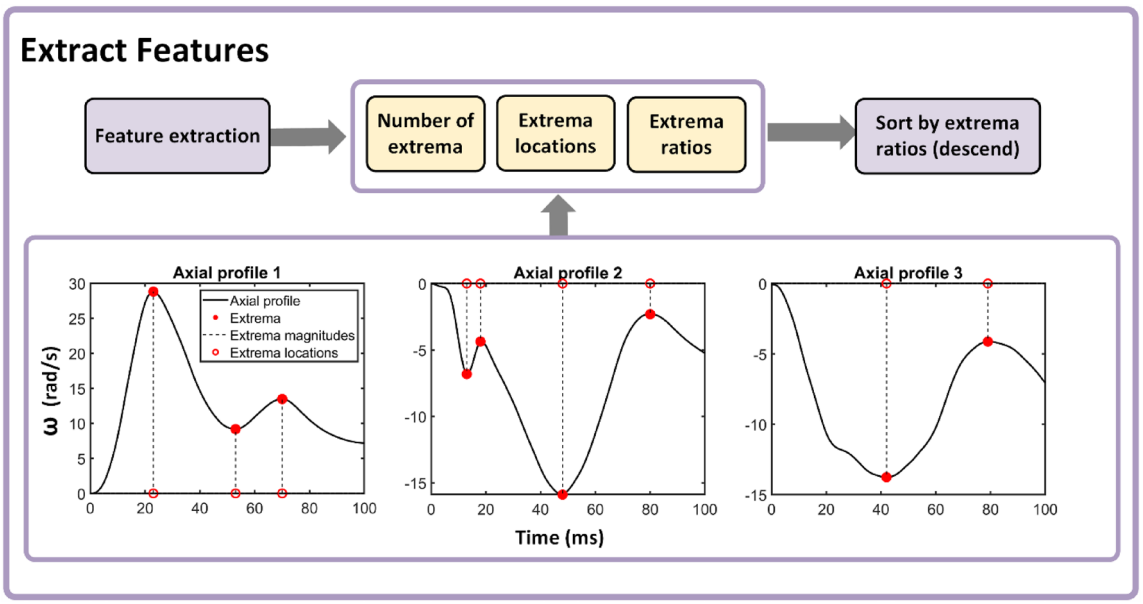

Fig. 2 Procedure for feature extraction of the three rotational velocity profile components, by first identifying peaks and valleys, from which to obtain their temporal locations and magnitude ratios. The features are sorted in a descending order based on magnitude ratios.

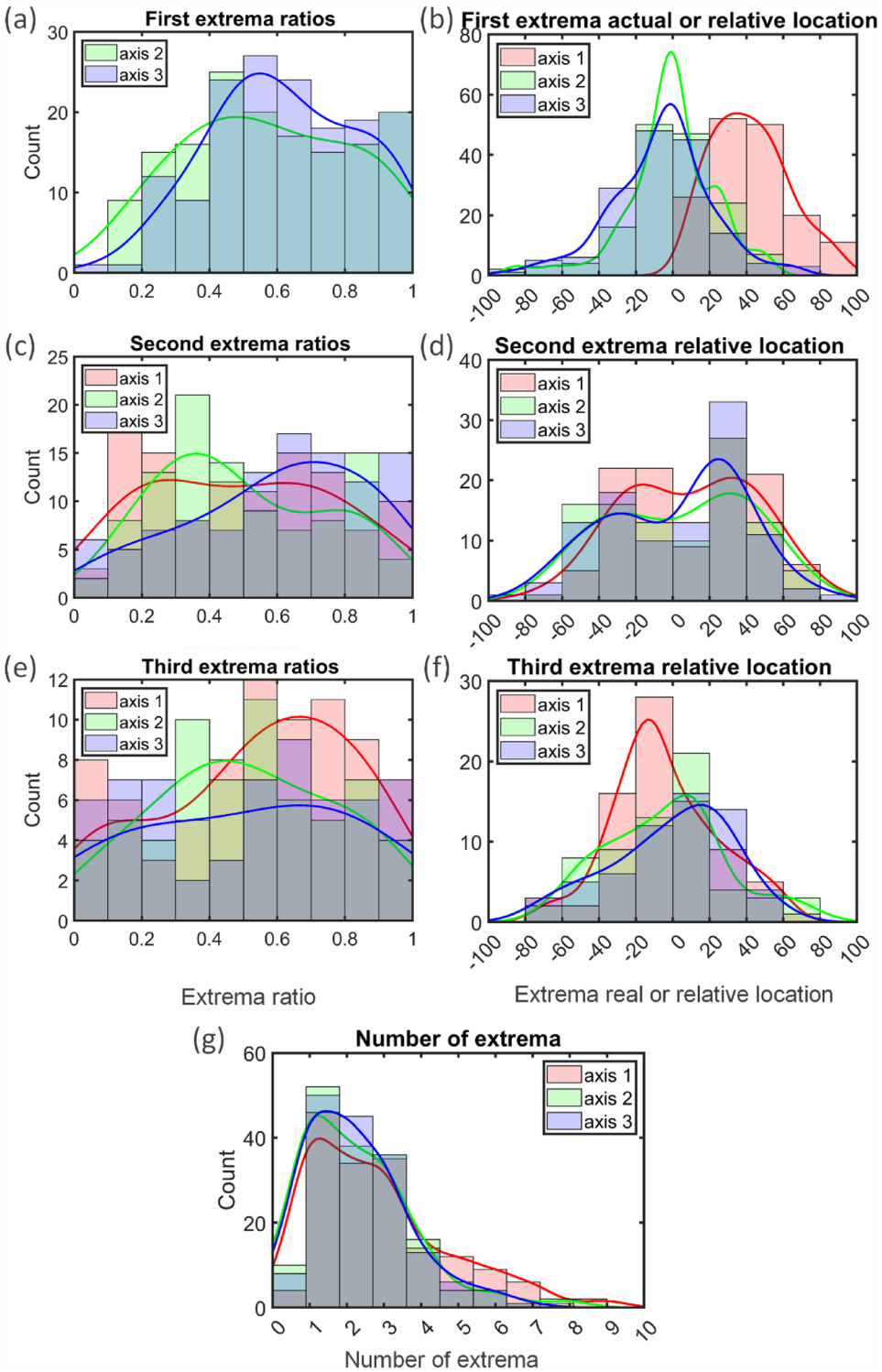

Fig. 3 Statistical distributions for (**a**) the ratios of the maximum extrema magnitudes of the second and third component relative to the those of the first and second along the three axes, respectively; (**b**) recorded temporal location of the maximum extrema for the first component, as well as and relative temporal locations of the maximum extrema for the second and third components; (**c**) ratios of the second largest extrema magnitudes relative to the maximum along each axes; (**d**) temporal locations of the second largest extrema relative to the maximum extrema along each axes; (**e**) relative ratios between the third and the second largest extrema along each axes; (**f**) relative temporal location for the third largest extrema relative to the second; and (**g**) number of extrema along each axis

Starting from this largest peak, smaller peaks were iteratively generated following the corresponding distributions for magnitude ratios and temporal locations (Figure 3). A scaling factor of –1.0 was applied to the magnitude, when necessary, to create a valley between two consecutive peaks (and vice versa). The resulting peaks and valleys served as "key points" for the rotational velocity profile, and they were up-sampled *via* "spline" interpolation to create a temporally continuous curve at 1000 Hz.

To combine the three component profiles, the two with smaller peak magnitudes were randomly scaled based on the distribution of peak magnitude ratios (Figure 3a). They were

further randomly shifted in time according to the established distributions of the relative temporal locations before randomly assigned to the x, y, and z axes. The three anatomical axes correspond to the posterior–anterior, right–left, and inferior–superior directions, respectively. Finally, the component magnitudes were uniformly scaled so that the peak resultant magnitude followed a uniform distribution in the range of 2–40 rad/s [11].

Intuitively, the number of local extrema reflects the complexity of the rotational velocity profile. Therefore, we empirically designated the resulting kinematic profile as "simple", if only one peak existed in the resultant rotational velocity profile, or "complex", otherwise.

Augmented Impact Dataset

For comparison, we also used the same earlier data augmentation scheme [11, 36] to generate a separate pretraining dataset. Briefly, the augmentation first permutes the x, y, and z components of the measured rotational velocity profile and then randomly rotates the rotational axis about the head center of gravity. The three anatomical axes correspond to head posterior-to-anterior, right-to-left, and inferior-to-superior direction, respectively. Finally, the resulting rotational velocity components are randomly and synchronously scaled so that the peak resultant magnitude follows a uniform distribution within the range of 2–40 rad/s. Similar to the synthetic impacts, the augmented impacts were also divided into "simple" vs. "complex" subsets, if the resultant velocity profile had one or more peaks, respectively.

TNN to Generate Pretraining Samples

The recently developed TNN served as a data generator in this study [35]. This neural network architecture is widely used in natural language processing tasks, which often has a superior performance due to its self-attention mechanism [12]. In this study, the TNN was retrained using the same earlier training samples [35] to predict relative brain-skull displacement for 70 ms duration (from 31st to 100th time frames; vs. 60 ms duration in the earlier study, from 31st to 90th; displacement and strain values for the first 30 ms are usually quite small to be of any interest) at a spatial isotropic resolution of 4 mm. Voxelized temporal MPS values at centroids were then obtained for each time frame.

To use the TNN for estimation, the same earlier preprocessing was employed for a given impact so that the dominant peak of rotational velocity occurred at a fixed temporal location of 50 ms (e.g., Figure 4). The rotational velocity profile was then concatenated with its corresponding acceleration profile (generated *via* forward differentiation; further scaled to 1% to maintain a comparable data range) before serving as TNN input. The TNN-estimated voxelwise relative brain-skull displacement from 31st to 100th ms were then used to calculate voxelwise MPS at every time point [17]. The peak MPS over the estimation time window of 70 ms were finally obtained to compare with the simulated counterparts resampled at the same voxel centroids.

CNN Pretraining and Finetuning

We adopted the previous CNN architecture [11] for training and evaluation. To be consistent with the voxelwise TNN output, the CNN output size was adjusted accordingly ($N=20036$ voxels for the brain parenchyma [35]). The CNN was first trained using the TNN-estimated pretraining samples from either the synthetic or augmented dataset. The network weights of the resulting pretrained CNN was further adjusted using finetuning training samples.

To investigate how the pretraining and finetuning sample sizes influenced CNN estimation accuracy, the two datasets were systematically varied in size. The pretraining dataset of sizes 1000 to 5000 (at a step size of 1000; $N=5$) with either synthetic or augmented impacts ($N=2$) were randomly selected from their respective pool of data. For each resulting pretrained CNN, finetuning dataset of sizes 250, 500, and then 1000 to 5000 (at a step size of 1000; $N=7$) were also randomly selected. To further investigate whether pretraining profile complexity affects CNN estimation accuracy, each pretraining dataset was also divided into "simple" vs. "complex" subsets for analysis ($N=2$). They led to a total of 70 ($=5\times7\times2$) pretraining/finetuning combinations. For each combination, 5 independent trials were executed, from which an averaged performance was obtained. In total, 350 ($=70\times5$) CNN models were pretrained and then finetuned for assessment. Finally, a baseline training without pretraining was conducted for each finetuning dataset ($N=7$) for additional comparison.

For all training tasks, mean squared error (MSE) between estimated peak MPS and those from direct simulation was used as the loss function. To determine network training hyperparameters, samples were divided into 90% for training and 10% for validation, for both pretraining and finetuning samples. For pretraining models, the batch size and learning rate were observed to be identical to the previous study (256 and 0.001, respectively [11]). The number of epochs was determined based on validation loss during cross-validation with an early stopping criterion [34] to avoid overfitting.

To optimize hyperparamters for finetuning, we systematically varied the batch size (4 to 256), learning rate (1e−5 to 1e−3; < 10% of that of the pretraining recommended [5]), and the number of maximum epochs (ranging from 300 to 1000) to yield the smallest validation loss with one random fold in a 10-fold cross-validation scheme. The resulting hyperparameters are given in the Appendix for each finetuning sample size and impact data type. In general, a smaller

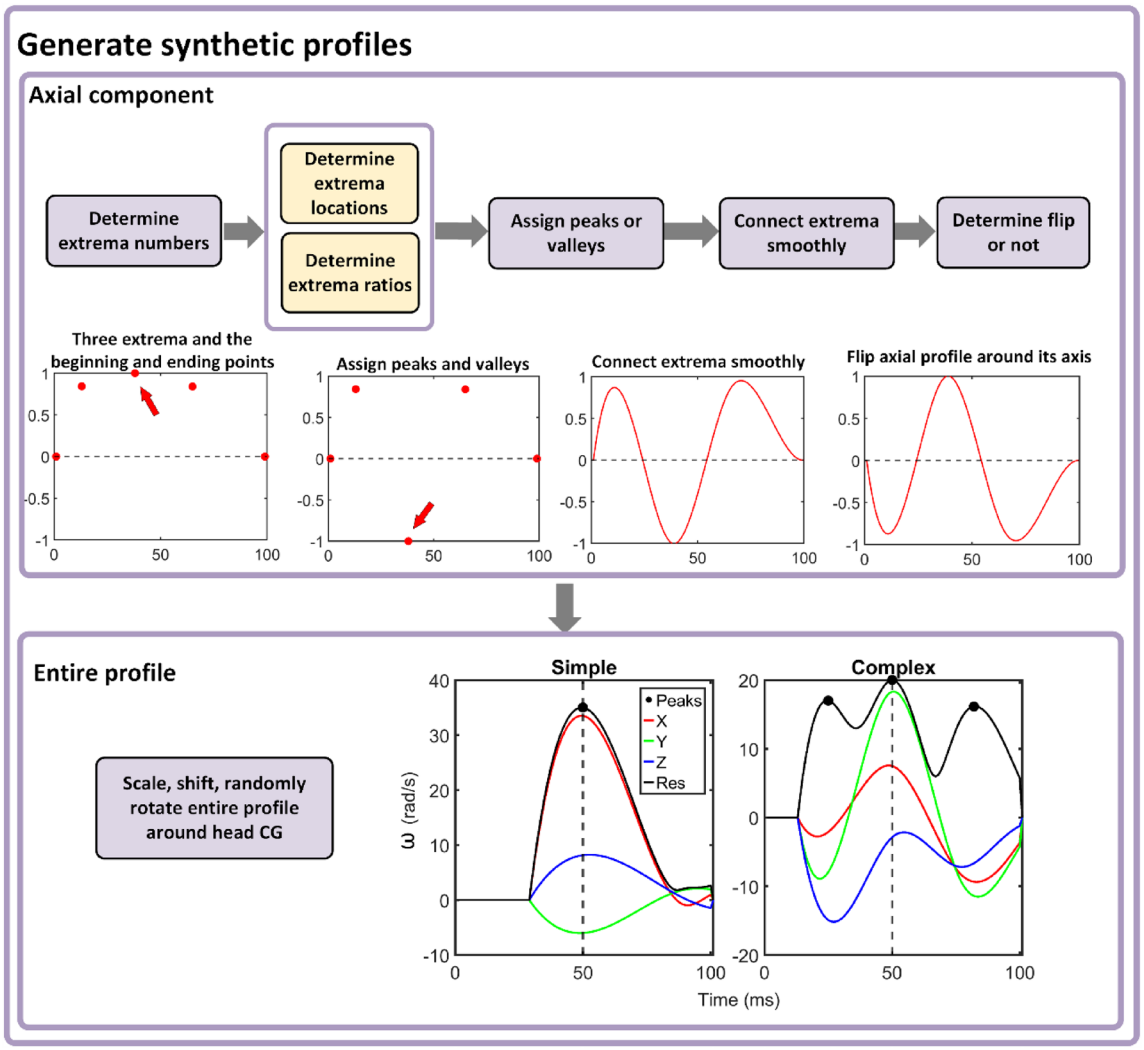

Fig. 4 Procedure for generating synthetic head impact rotational velocity profiles. A zero rotational velocity is assumed at the two ends of the 100 ms impact window. The combined profile is further shifted to leave a non-zero velocity magnitude at the end of the impact window.

training dataset required a smaller learning rate, a smaller batch size, but a larger number of epochs, to converge.

CNN Performance Evaluation

The CNN performance was evaluated based on an independent dataset of head impacts measured from American high school football (HF, $N = 314$, impact duration of 50 ms). Their previous direct model simulations were used to generate resampled voxelwise peak MPS. A relatively short impact duration may not allow the brain to reach peak strain values during the simulation time window [15, 23]. Therefore, we followed a recent study to exclude impacts when the peak rotational velocity occurred near the end of the impact recording window [22]. Specifically, impacts were excluded if the peak resultant rotational velocity occurred within the last 5 ms relative to the temporal window right-handed boundary (as empirically used earlier [15]). This led to 191 impacts for CNN performance evaluation.

Data Analysis

To evaluate TNN estimation accuracy, we randomly generated 100 synthetic impacts and 100 augmented impacts, respectively, for estimation and direct model simulation. Estimation accuracies for both TNN and CNN were evaluated by comparing estimated voxelwise peak MPS with those from direct model simulations in terms of R^2, RMSE and SR. An estimation was said to be successful with sufficient accuracy when the linear regression slope (k) and Pearson correlation coefficient (r), between the estimation and the simulated counterpart, did not deviate from the "perfect

scores" of 1.0 (when identical) by more than a given threshold. In this study, we chose two thresholds of either 0.1 or 0.05 for a more relaxed or a more stringent criterion for assessment, respectively.

CNN estimation accuracies for all pretraining/finetuning configurations were reported and compared with those from the baseline training (i.e., using finetuning samples alone for training without transfer learning from pretraining). A pretraining model alone, without using real-world measured data for training, could place a bottleneck in performance [21, 32]. Therefore, we also reported accuracies of pretrained CNN models based on either synthetic or augmented impacts as a comparison. All CNN trainings were conducted in Python (Intel Xeon E5-2698 with 256 GB and A100 GPU with 80 GB). Generating one batch of pretraining samples of 5000 took one full day (from voxelized relative displacement to strain at every time frame, and finally to voxelized peak strain), most of which was on disk input/output. Each finetuning training took ~ 30 min to complete. All data analyses were executed in MATLAB (Version R2022b).

Results

TNN Accuracy Performances

Figure 5 reports the *k*-*r* plots for TNN-estimated responses. For both impact types, the TNN remained highly accurate in terms of R^2 (0.948–0.967) and RMSE (0.012–0.015), virtually comparable to previous report [35]. However, in terms of the regression slope, the TNN appeared to have slightly over-estimated the peak responses, with the average *k* values 0.05–0.06 above the "perfect" score of 1.0. To further investigate, an example "failed" case for each impact data type was selected for further scrutiny at several time points (Figures A1 and A2 in the Appendix).

CNN Performances

Figure 6 reports CNN estimation performances based on SR (at two success thresholds of 0.1 and 0.05), R^2 and RMSE for different combinations of pretraining (of synthetic impacts) and finetuning samples. Figure 7 reports the same when using augmented impacts as pretraining samples. Figure 8 synthesizes the results into "theoretical trendlines" for the performance metrics relative to pretraining and finetuning dataset sizes (performance trendlines for SR and R^2 similar). Figure 9 compares the accuracy performances using each pretrained CNN model without finetuning. Finally, one-tailed nonparametric Mann Whitney U tests were used to test whether using transfer learning improved model performance for each combination of pretraining and finetuning dataset size over the baseline training. Table 1 summarizes the number of significant results for different combinations of pretraining and finetuning datasets. More detailed results are in Table A2 in the Appendix. Figure A3 illustrates examples of significant or insignificant tests with box plots showing the scatter of independent trials.

Discussion

When deep learning brain models were first developed, it was unclear how many training samples were necessary. A domain-relevant guideline does not exist. Hence, most of them have used thousands of training samples, given that more samples are anticipated to improve accuracy. However, as few as 80 samples were also found sufficient for a much lighter weight fully connected neural network for a porcine brain model (vs. human brain in others), which was designed to conduct sensitivity analysis for controlled cortical impact (3 important input parameters with 4 outputs [26]). A more complex network architecture with larger input and output sizes such as predicting peak strains of the entire brain will likely require a larger training dataset [6, 11, 34, 40]. Nevertheless, using the fewest training samples to achieve a sufficient estimation accuracy can limit the associated computational cost, which is important in practice.

In this study, we employed a large amount of impacts and their direct model simulations already available from previous endeavors to study how accuracy varied with respect to the amount of training samples and the training strategy.

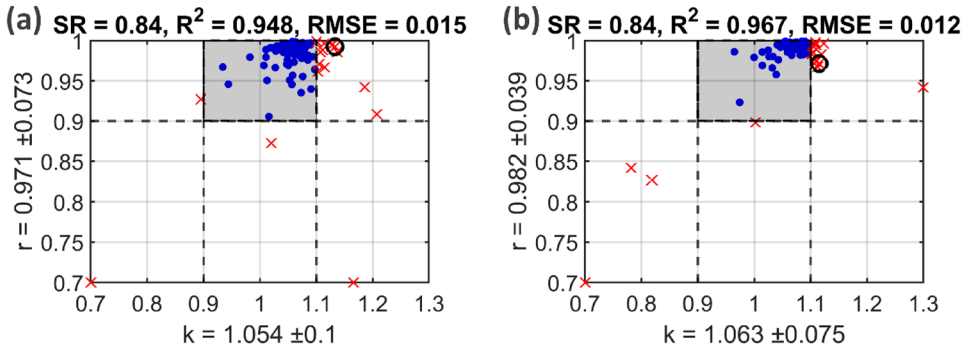

Fig. 5 *k*–*r* plots to report TNN accuracy performances based on 100 randomly generated (**a**) synthetic or (**b**) augmented impacts. Their corresponding SR, R^2 and RMSE are reported. To improve visualization, *k* and *r* values are capped. One example "failed" case for each data type (black circles) is selected for closer scrutiny in the **Appendix**.

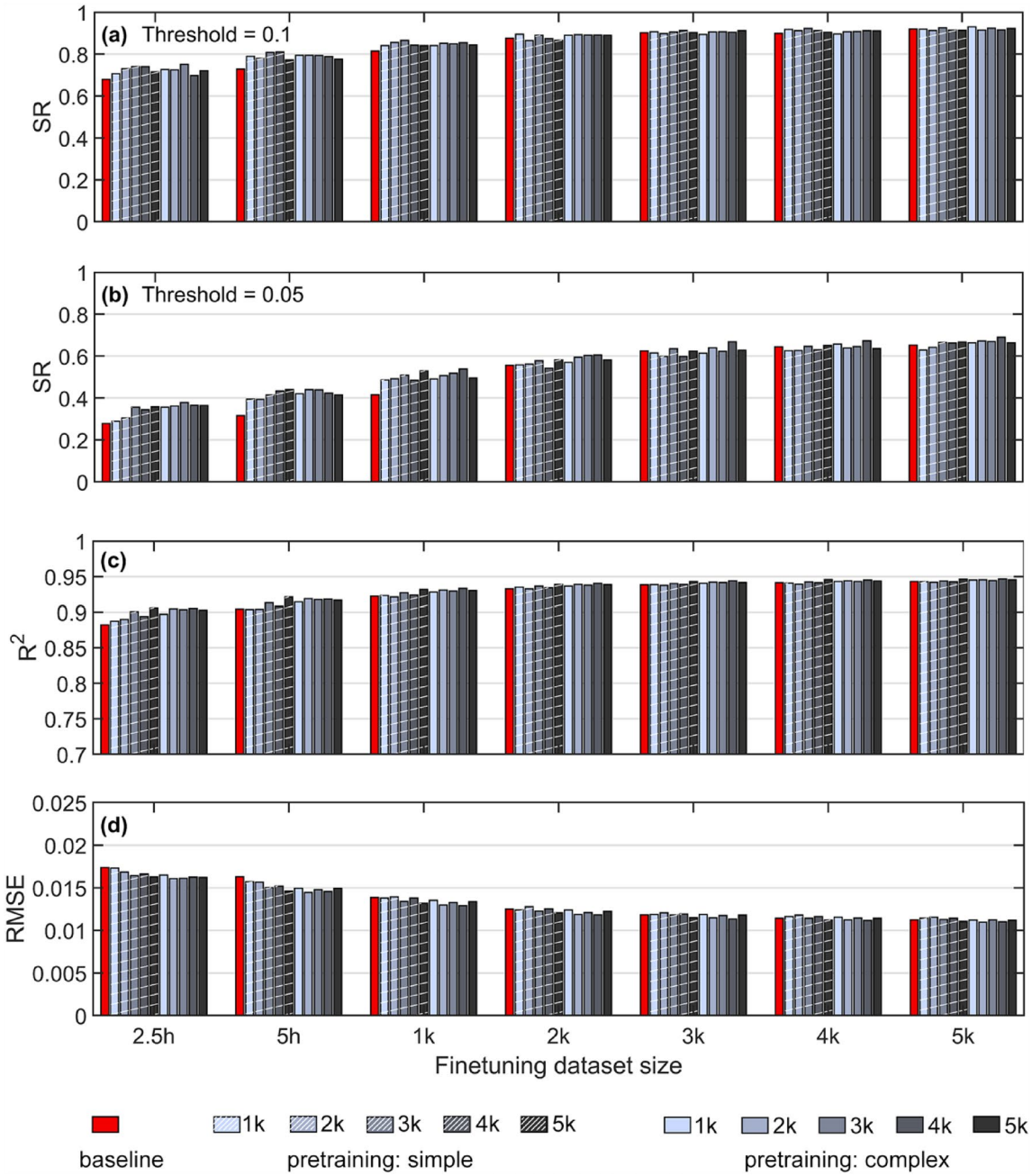

Fig. 6 Summary of CNN estimation accuracy in terms of SR (with success threshold of either 0.1 or 0.05, **a** and **b**, respectively), (**c**) R^2, and (**d**) RMSE using synthetic samples for pretraining. The pretraining samples are further divided into "simple" vs. "complex" impacts according to the number of extrema points in the resultant rotational velocity profiles.

Findings from this study may not directly benefit deep learning models already built, but they could provide insight into an economical approach for developing a *future* deep learning brain model either at the global or the microscale level, when fresh training samples are necessary that would require costly simulations using the updated baseline FE model. These efforts could contribute to a wider adoption of deep learning brain models in the future.

We find that pretraining was effective at improving a CNN model estimation accuracy, especially when the finetuning dataset was relatively small (e.g., < 1000 samples; Figures 6–7). The trends are clearer in the synthesized theoretical trendlines (Figure 8). Statistical tests further indicated that pretraining was most effective in terms of R^2 as significant improvements occurred in most of the pretraining-finetuning combinations (e.g., 47 and 64 out of a total of 70

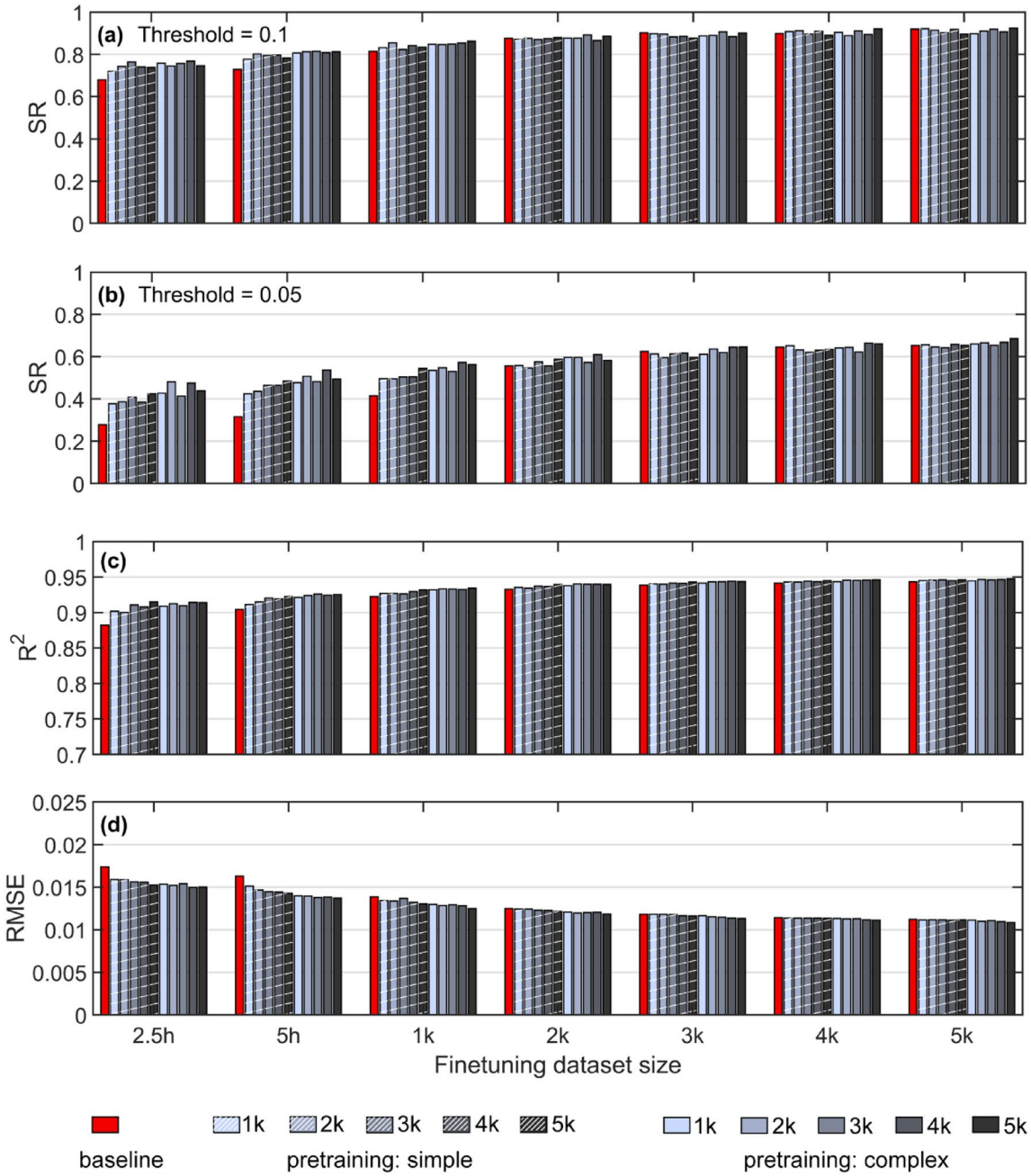

Fig. 7 Summary of CNN estimation accuracy in terms of SR (with success threshold of either 0.1 or 0.05, **a** and **b**, respectively), (**c**) R^2, and (**d**) RMSE using augmented impact samples for pretraining.

combinations were significant using synthetic impacts and augmented impacts, respectively; Table 1). Both synthetic and augmented impacts were similarly effective in pretraining in terms of SR, but augmented impacts improved R^2 and RMSE more than the synthetic impacts (by up to 2% and 3–8%, respectively). This suggested synthetic impacts may have some different features than those in the real-world. From the more detailed Table A2, there was some trend that pretrained models were more beneficial when the finetuning dataset was relatively small, and a large pretrained dataset was usually preferred. However, there was no clear trend between simple vs. complex pretraining impacts.

Given these observations, pretraining is recommended if finetuning samples are relatively few (~1000 or less). However, when sufficient finetuning samples are available (e.g., > 2000), baseline training without pretraining is

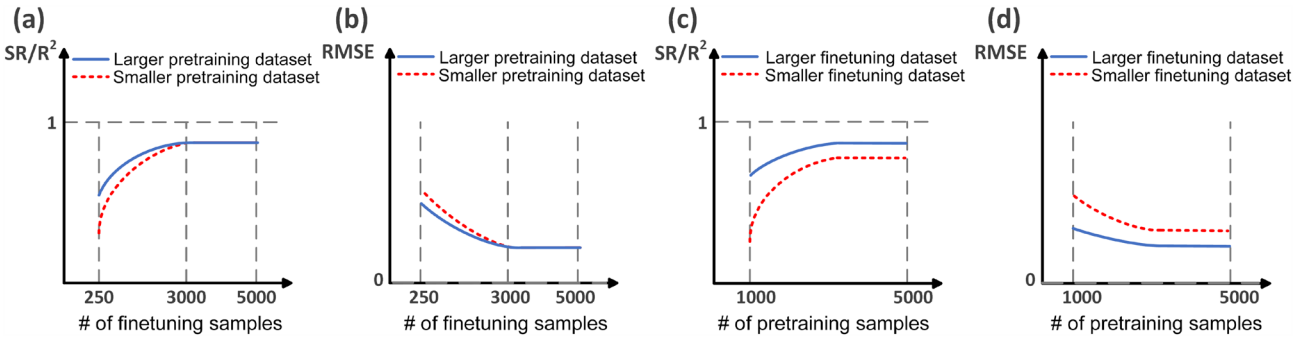

Fig. 8 Synthesized "theoretical trendlines" to illustrate performance variations in terms of (**a, c**) SR/R^2 and (**b**, **d**) RMSE with respect to (**a, b**) finetuning and (**c, d**) pretraining dataset sizes. SR and R^2 have an upper limit of 1.0, while RMSE has a lower limit of 0.0 (when predictions are identical to target values). A larger or a smaller dataset refers to > 3000 or < 1000, respectively. With sufficient finetuning samples (>3000 as in a, b), performances converge to the same value regardless of the pretraining sample size. Pretraining is most effective when finetuning sample size is relatively small (i.e., < 1000).

sufficient. In fact, when the size of finetuning samples is even larger (e.g., > 4000), pretraining could occasionally degrade the CNN performance as indicated by one-tailed Mann Whitney U test. From a biomechanical perspective, focusing on head rotational kinematics and taking advantage of the head/brain geometrical symmetry property relative to the mid-sagittal plane would also reduce the training samples required.

TNN Performance

The TNN remained highly accurate when estimating either the synthetic or augmented impacts, with R^2 of 0.948–0.967 and RMSE of 0.012–0.015 (approximately 6–7.5% of the injury threshold of 0.2 previously established based on the reconstructed NFL impacts [42]). The R^2 was slightly lower than the value of > 0.99 achieved at the time when the MPS was at peak [35], in part, because it was the accumulated peak strains used for evaluation in this study.

Interestingly, the SR was only 84% when using the success threshold of 0.1 (Figure 5), which was somewhat lower than earlier studies where > 96% could be achieved [11, 22]. While the majority of testing impacts had a Pearson correlation coefficient, r, above 0.95, most of them had a k value greater than 1.0, with an average of 1.054–1.063. This suggests a general over-estimation of 5–6% for this dataset evaluated here.

Notably, R^2 and RMSE are commonly used for accuracy assessment in TBI biomechanical studies [9, 31, 34, 38]. The SR is a relatively recent accuracy metric [11]. It is possible to have a perfect R^2 of 1.0, but the accuracy in terms of RMSE could still be poor (e.g., consider two samples where the values of one are exactly twice of the other). Therefore, R^2, alone, may not be sufficient to quantify accuracy between two samples of the same variable. While a non-zero RMSE could reflect the discrepancy between the two samples, it does not indicate whether an overall over- or under-estimation occurs. In this case, the regression slope, k, could provide valuable additional insight (e.g., detecting a 5–6% over-estimation for the dataset evaluated here). Therefore, using multiple accuracy metrics is recommended for a comprehensive understanding of the agreement among different methods.

The error in peak MPS seemed to be the accumulation of errors when converting from estimation of relative brain-skull displacement into strain over time (see Figures A1 and A2 in the Appendix). By design, the TNN was trained to estimate voxelized relative brain-skull displacement over time so that to significantly reduce the amount of data to handle (vs. directly estimating the complete strain tensor) [35]. Further evaluation of TNN accuracy in terms of accumulated peak strain is necessary to determine if this strain over-estimation is specific to the testing dataset employed here or more systematic. In the latter case, a simple scaling could mitigate the systematic error.

Impact Samples

Without finetuning, pretraining models also had some capability in estimating brain strains. However, their accuracies were considerably lower than those with additional finetuning, and they were also largely invariant with respect to the amount of pretraining samples used (Figure 9). This was in contrast to CNNs with additional finetuning, where accuracy generally improved with the increase in finetuning sample size. For example, the SR for pretraining models (with the success threshold of 0.1) was approximately 0.5 and 0.6, respectively, when using synthetic and augmented datasets, respectively. This was considerably lower than 0.8–0.9 when at least 500 finetuning samples were used in training.

Pretrained CNNs with augmented impacts consistently outperformed those with synthetic data. This again suggests

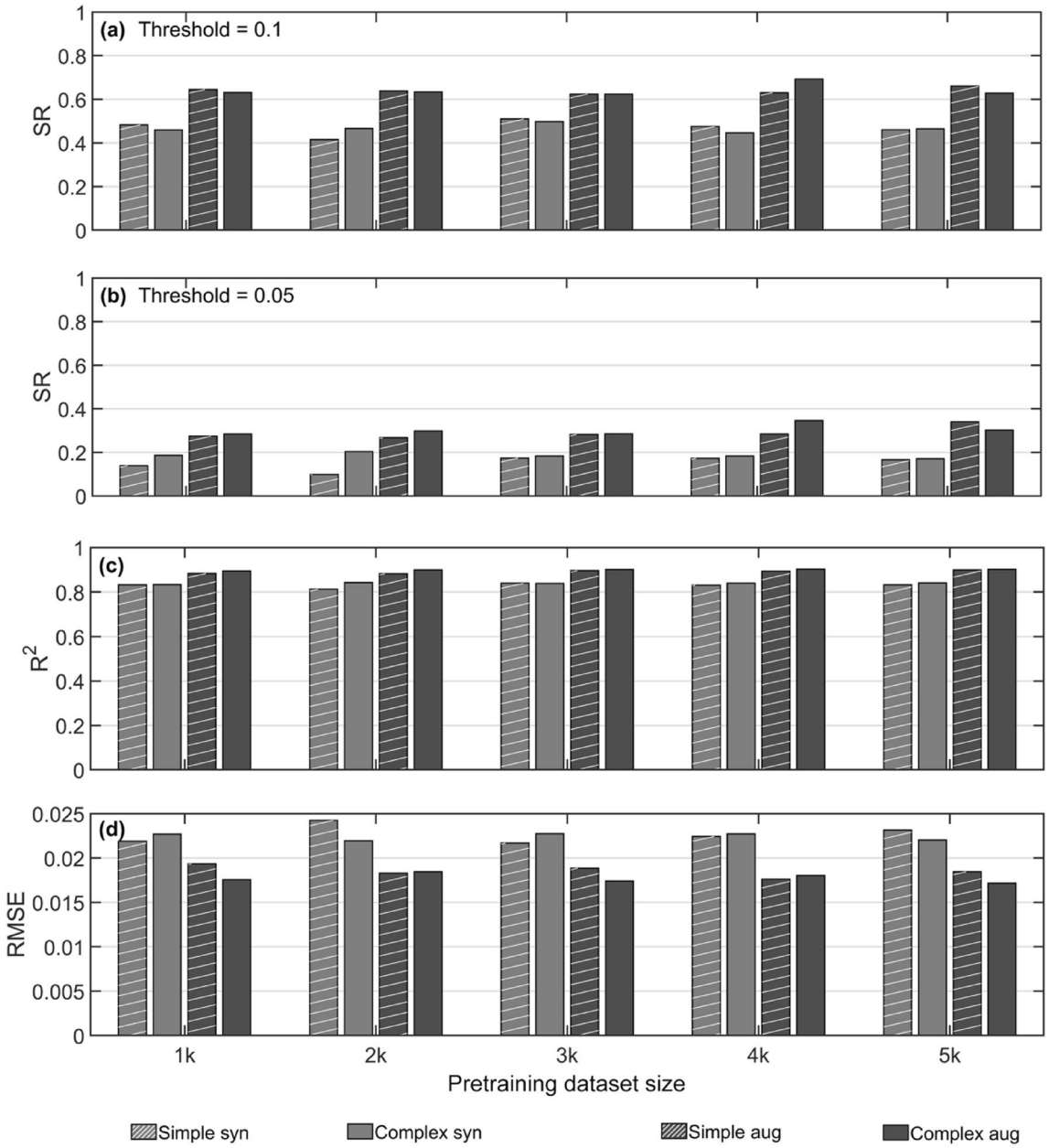

Fig. 9 Pretrained CNN performances in terms of SR, R^2 and RMSE using four pretraining impact data types with a range of sample sizes, without additional finetuning.

Table 1 Summary of the number of Mann Whitney U tests reporting significant improvement ($p < 0.05$) in CNN prediction accuracy when using a pretrained model compared to baseline training without transfer learning

	SR (threshold = 0.1)	SR (threshold = 0.05)	R^2	RMSE
Syn (simple/complex)	21 (9/12)	31 (17/14)	47 (25/22)	23 (15/8)
Aug (simple/complex)	23 (10/13)	33 (17/16)	64 (32/32)	32 (13/19)

For each combination of pretraining and finetuning datasets (of size of 5 and 7, respectively), the total number of tests is 35 (= 5 × 7). For example, for 9 out of 35 combinations using simple synthetic impacts, pretraining significantly improves prediction accuracy in terms of SR (with threshold of 0.1), but significant improvement occurs in terms of R^2 for 25 out of them. "Syn": synthetic impacts; "Aug": augmented impacts

some differences between the two types of impacts. The synthetic impacts assumed a zero rotational velocity at both ends of the 100 ms impact window to facilitate data generation. After impact profile shifting, a non-zero velocity usually occurred at the end of the impact window, similar to real-world impacts. Nevertheless, more work is necessary to devise pretraining samples that best represent real-world impacts, perhaps, by considering features in both time and frequency domains [1, 7].

Impact Kinematic Profiles

We have intentionally and empirically categorized pretraining impacts into "simple" vs. "complex" based on the number of extrema found in the resultant profile. However, it was not obvious whether there were any performance differences between the two, either with or without additional finetuning. This observation may, at least in part, be due to the relatively "simple" independent testing samples that had a considerably shorter impact duration than other augmented and measured impacts (50 ms, vs. ~ 100 ms). The utility of the "simple" vs. "complex" profiles require further investigation in the future, e.g., when dealing with longer impact durations such as those in automotive impacts.

However, it is important to ensure loading profiles to capture the complete or at least the majority of the transient impact process. For all real-world head impacts, there is a basic expectation that the head will eventually come to a stop. Therefore, both acceleration and deceleration are expected. When focusing on rotational acceleration peak magnitude (as in earlier studies [19]) without considering deceleration, it is questionable that the resulting simulated brain strains are trustworthy [15, 23]. This could compromise strain-based injury detection and risk assessment, which is likely more of a concern when considering the spatial distribution of brain strains.

Generality of Findings

Impacts in this study were mostly from American football, either measured on the field or reconstructed in laboratory. The synthetic impacts were also generated based on these measured impacts. There is some evidence suggesting kinematic differences from impacts in different sports or events (e.g., the same kinematic-based injury criteria correspond to different strain levels [39]). However, the findings from this study may be more universal across sports and level (e.g., collegiate vs. high school) from several perspectives.

First, the TNN and a separate multi-task CNN (i.e., separating the continuous impact duration into multiple segments to facilitate training, which shares the same neural network architecture as those in the current study) have shown to be accurate for a range of impact types exhaustively chosen from the literature [35]. Therefore, it is likely that the CNN model will remain applicable to other head impact types, such as ice-hockey, lacrosse, and soccer. One explanation is that random data augmentation may provide a range of impact kinematic profiles that could represent those in different sports. Their kinematic features will likely overlap in the "feature space" (albeit automotive impacts may have larger differences due to their typically much longer impact durations, which would pose additional challenges [33]). Both rotational velocity peak magnitudes and their temporal locations are considered, as their influence on brain strain is universal [4, 29].

Second, we expect that transfer learning may be universally effective for other sports, another baseline FE model other than the WHIM, and even microscale axonal injury models [28, 41]. This is because they all use time series data such as impact kinematics or axonal stretch history as input. A monotonic input-output relationship is anticipated, where larger input values would lead to elevated response magnitudes. Therefore, a pretrained model that has already learned the basic mapping would likely reduce the number of finetuning samples required for training a deep learning model.

Limitations

This study was limited to using CNN for training and assessment. While the specific findings/recommendations may not be directly applicable to other neural network architectures such as U-Net [6] and fully connected neural networks [26, 40], we anticipate that at least the approach herein remains applicable. A generative adversarial network (GAN) may also be a more advanced alternative to generate synthetic impact profiles, as recently demonstrated to augment material microstructural patterns from a relatively small training dataset [20]. Regardless, the utility of TNN for rapid generation of pretraining samples from these augmented impacts for neural network training likely remains.

In addition, we have also focused on spatially detailed and voxelwise peak strains of the entire brain. However, a "scalar" value such as peak strain of the whole brain or cumulative strain damage measure (CSDM) [31] remains commonly used at present [8, 14]. While it is trivial to compute a scalar value from a spatially detailed peak strain distribution, developing a deep learning brain model directly for a single scalar output or a vector output for a few selected anatomical regions [10] would dramatically reduce the number of outputs (~20 thousand here vs. 1 or a few). Without the need to consider brain strain spatial distribution, a simpler deep learning architecture may suffice that would also require considerably fewer training samples to limit the demand of computational resources.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10439-023-03354-3.

Acknowledgements Funding from the NSF award under Grant No. 2114697 is acknowledged.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Arrue, P., N. Toosizadeh, H. Babaee, and K. Laksari. Low-rank representation of head impact kinematics: a data-driven emulator. *Front. Bioeng. Biotechnol.* 8:1–11, 2020.
2. Bayly, P. V., A. Alshareef, A. K. Knutsen, K. Upadhyay, R. J. Okamoto, A. Carass, J. A. Butman, D. L. Pham, J. L. Prince, K. T. Ramesh, and C. L. Johnson. MR imaging of human brain mechanics in vivo: new measurements to facilitate the development of computational models of brain injury. *Ann. Biomed. Eng.* 2021. https://doi.org/10.1007/s10439-021-02820-0.
3. Bian, K., and H. Mao. Mechanisms and variances of rotation-induced brain injury: a parametric investigation between head kinematics and brain strain. *Biomech. Model. Mechanobiol.* 2020. https://doi.org/10.1007/s10237-020-01341-4.
4. Carlsen, R. W., A. L. Fawzi, Y. Wan, H. Kesari, and C. Franck. A quantitative relationship between rotational head kinematics and brain tissue strain from a 2-D parametric finite element analysis. *Brain Multiphysics*. 2:100024, 2021.
5. Dao, T. T. From deep learning to transfer learning for the prediction of skeletal muscle forces. *Med. Biol. Eng. Comput.* 57:1049–1058, 2019.
6. Deck, C., N. Bourdet, A. Trog, F. Meyer, V. Noblet, and R. Willinger. Deep learning method to assess brain injury risk. *Int. J. Crashworthiness*. 2022. https://doi.org/10.1080/13588265.2022.2130600.
7. Escarcega, J. D., A. K. Knutsen, R. J. Okamoto, D. L. Pham, and P. V. Bayly. Natural oscillatory modes of 3D deformation of the human brain in vivo. *J. Biomech.* 119:110259, 2021.
8. Fahlstedt, M., F. Abayazid, M. B. Panzer, A. Trotta, W. Zhao, M. Ghajari, M. D. Gilchrist, S. Ji, S. Kleiven, X. Li, A. N. Annaidh, and P. Halldin. Ranking and rating bicycle helmet safety performance in oblique impacts using eight different brain injury models. *Ann. Biomed. Eng.* 49:1097–1109, 2021.
9. Gabler, L. F., J. R. Crandall, and M. B. Panzer. Assessment of kinematic brain injury metrics for predicting strain responses in diverse automotive impact conditions. *Ann. Biomed. Eng.* 44:3705–3718, 2016.
10. Ghazi, K., M. Begonia, S. Rowson, and S. Ji. American Football Helmet Effectiveness Against a Strain-Based Concussion Mechanism. *Ann. Biomed. Eng.* 50:1498–1509, 2022.
11. Ghazi, K., S. Wu, W. Zhao, and S. Ji. Instantaneous whole-brain strain estimation in dynamic head impact. *J. Neurotrauma*. 38:1023–1035, 2021.
12. Guo, M.-H., Z.-N. Liu, T.-J. Mu, and S.-M. Hu. Beyond self-attention: External attention using two linear layers for visual tasks. 2021.
13. Hernandez, F., L. C. Wu, M. C. Yip, K. Laksari, A. R. Hoffman, J. R. Lopez, G. A. Grant, S. Kleiven, and D. B. Camarillo. Six degree-of-freedom measurements of human mild traumatic brain injury. *Ann. Biomed. Eng.* 43:1918–1934, 2015.
14. Ji, S., M. Ghajari, H. Mao, H. Kraft, Reuben, M. Hajiaghamemar, M. B. Panzer, R. Willinger, M. D. Gilchrist, S. Kleiven, and J. D. Stitzel. Use of brain biomechanical models for monitoring impact exposure in contact sports. *Ann. Biomed. Eng.* 50:1389–1408, 2022.
15. Ji, S., S. Wu, and W. Zhao. Dynamic characteristics of impact-induced brain strain in the corpus callosum. *Brain Multiphys.* 3:100046, 2022.
16. Ji, S., and W. Zhao. A pre-computed brain response atlas for instantaneous strain estimation in contact sports. *Ann. Biomed. Eng.* 43:1877–1895, 2015.
17. Ji, S., and W. Zhao. Displacement voxelization to resolve mesh-image mismatch: application in deriving dense white matter fiber strains. *Comput. Methods Programs Biomed.* 213:106528, 2022.
18. Ji, S., W. Zhao, Z. Li, and T. W. McAllister. Head impact accelerations for brain strain-related responses in contact sports: a model-based investigation. *Biomech. Model. Mechanobiol.* 13:1121–1136, 2014.
19. King, A. I., K. H. Yang, L. Zhang, W. Hardy, and D. C. Viano. Is head injury caused by linear or angular acceleration? 2003.
20. Kobeissi, H., S. Mohammadzadeh, and E. Lejeune. Enhancing mechanical metamodels with a generative model-based augmented training dataset. *J. Biomech. Eng.* 144:121002, 2022.
21. Kortylewski, A., A. Schneider, T. Gerig, B. Egger, A. Morel-Forster, and T. Vetter. Training Deep Face Recognition Systems with Synthetic Data. 1–8, 2018.
22. Lin, N., S. Wu, and S. Ji. A morphologically individualized deep learning brain injury model. *J. Neurotrauma (in press)*. 2023. https://doi.org/10.1089/neu.2022.0413.
23. Liu, Y., A. G. Domel, N. J. Cecchi, E. Rice, A. A. Callan, S. J. Raymond, Z. Zhou, X. Zhan, Y. Li, M. M. Zeineh, G. A. Grant, and D. B. Camarillo. Time window of head impact kinematics measurement for calculation of brain strain and strain rate in American Football. *Ann. Biomed. Eng.* 2021. https://doi.org/10.1007/s10439-021-02821-z.
24. Madhukar, A., and M. Ostoja-Starzewski. Finite element methods in human head impact simulations: a review. *Ann. Biomed. Eng.* 47:1832–1854, 2019.
25. Meaney, D. F., B. Morrison, and C. R. Bass. The mechanics of traumatic brain injury: a review of what we know and what we need to know for reducing its societal burden. *J. Biomech. Eng.* 136:021008, 2014.
26. Menichetti, A., L. Bartsoen, B. Depreitere, J. Vander Sloten, and N. Famaey. A machine learning approach to investigate the uncertainty of tissue-level injury metrics for cerebral contusion. *Front. Bioeng. Biotechnol.* 9:0201008, 2021.
27. Miller, L. E., J. E. Urban, M. A. Espeland, M. P. Walkup, J. M. Holcomb, E. M. Davenport, A. K. Powers, C. T. Whitlow, J. A. Maldjian, and J. D. Stitzel. Cumulative strain-based metrics for predicting subconcussive head impact exposure–related imaging changes in a cohort of American youth football players. *J. Neurosurg.* 29(4):387–396, 2022.
28. Montanino, A., X. Li, Z. Zhou, M. Zeineh, D. B. Camarillo, and S. Kleiven. Subject-specific multiscale analysis of concussion: from macroscopic loads to molecular-level damage. *Brain Multiphys.* 2:100027, 2021.
29. Post, A., E. S. Walsh, T. B. Hoshizaki, and M. D. Gilchrist. Analysis of loading curve characteristics on the production of brain deformation metrics. *Proc. Inst. Mech. Eng. Part P J. Sport. Eng. Technol.* 0:1–8, 2012.
30. Sanchez, E. J., L. F. Gabler, A. B. Good, J. R. Funk, J. R. Crandall, and M. B. Panzer. A reanalysis of football impact reconstructions for head kinematics and finite element modeling. *Clin. Biomech.* 64:82–89, 2018.
31. Takhounts, E. G., S. A. Ridella, R. E. Tannous, J. Q. Campbell, D. Malone, K. Danelson, J. Stitzel, S. Rowson, and S. Duma. Investigation of traumatic brain injuries using the next generation of simulated injury monitor (SIMon) finite element head model. *Stapp Car Crash J.* 52:1–31, 2008.

32. Tremblay, J., A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.* 2018-June:1082–1090, 2018.
33. Wu, S., W. Zhao, S. Barbat, J. Ruan, and S. Ji. Instantaneous brain strain estimation for automotive head impacts via deep learning. *Stapp Car Crash J.* 65:139–162, 2021.
34. Wu, S., W. Zhao, K. Ghazi, and S. Ji. Convolutional neural network for efficient estimation of regional brain strains. *Sci. Rep.* 9:17326, 2019.
35. Wu, S., W. Zhao, and S. Ji. Real-time dynamic simulation for highly accurate spatiotemporal brain deformation from impact. *Comput. Methods Appl. Mech. Eng.* 394:114913, 2022.
36. Wu, S., W. Zhao, Z. Wu, J. C. Ford, L. A. Flashman, T. W. McAllister, J. Hu, and S. Ji. Subject-specific Head Injury Models via Scaling Based on Head Morphology: Initial Finding. , 2019.
37. Yang, K. H., J. Hu, N. A. White, A. I. King, C. C. Chou, and P. Prasad. Development of numerical models for injury biomechanics research: a review of 50 years of publications in the Stapp Car Crash Conference. *Stapp Car Crash J.* 50:429–490, 2006.
38. Zhan, X., Y. Li, Y. Liu, N. J. Cecchi, O. Gevaert, M. M. Zeineh, G. A. Grant, and D. B. Camarillo. Piecewise multivariate linearity between kinematic features and cumulative strain damage measure (CSDM) across different types of head impacts. *Ann. Biomed. Eng.* 2022. https://doi.org/10.1007/s10439-022-03020-0.
39. Zhan, X., Y. Li, Y. Liu, A. G. Domel, H. V. Alizadeh, S. J. Raymond, J. Ruan, S. Barbat, S. Tiernan, O. Gevaert, and M. M. Zeineh. The relationship between brain injury criteria and brain strain across different types of head impacts can be different. *J. R. Soc. Interface*. 18(179):20210260, 2021.
40. Zhan, X., Y. Liu, S. J. Raymond, H. V. Alizadeh, A. G. Domel, O. Gevaert, M. M. Zeineh, G. A. Grant, and D. B. Camarillo. Rapid estimation of entire brain strain using deep learning models. *IEEE Trans. Biomed. Eng.* 9294:1–11, 2021.
41. Zhang, C., and S. Ji. Sex differences in axonal dynamic responses under realistic tension using finite element models. *J. Neurotrauma* (in press), 2023.
42. Zhao, W., Y. Cai, Z. Li, and S. Ji. Injury prediction and vulnerability assessment using strain and susceptibility measures of the deep white matter. *Biomech. Model. Mechanobiol.* 16:1709–1727, 2017.
43. Zhao, W., and S. Ji. Brain strain uncertainty due to shape variation in and simplification of head angular velocity profiles. *Biomech. Model. Mechanobiol.* 16:449–461, 2017.