Concussion classification via deep learning using whole-brain white matter fiber strains

Yunliang Cai[1‡], Shaoju Wu[1‡], Wei Zhao[1], Zhigang Li[2], Zheyang Wu[3], Songbai Ji[1,4]*

1 Department of Biomedical Engineering, Worcester Polytechnic Institute, Worcester, MA, United States of America, **2** Department of Biomedical Data Science, Geisel School of medicine, Dartmouth College, Hanover, NH, United States of America, **3** Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA, United States of America, **4** Department of Mechanical Engineering, Worcester Polytechnic Institute, Worcester, MA, United States of America

‡ These authors are co-first authors on this work.
* sji@wpi.edu

Abstract

Developing an accurate and reliable injury predictor is central to the biomechanical studies of traumatic brain injury. State-of-the-art efforts continue to rely on empirical, scalar metrics based on kinematics or model-estimated tissue responses explicitly pre-defined in a specific brain region of interest. They could suffer from loss of information. A single training dataset has also been used to evaluate performance but without cross-validation. In this study, we developed a deep learning approach for concussion classification using implicit features of the entire voxel-wise white matter fiber strains. Using reconstructed American National Football League (NFL) injury cases, leave-one-out cross-validation was employed to objectively compare injury prediction performances against two baseline machine learning classifiers (support vector machine (SVM) and random forest (RF)) and four scalar metrics via univariate logistic regression (Brain Injury Criterion (BrIC), cumulative strain damage measure of the whole brain (CSDM-WB) and the corpus callosum (CSDM-CC), and peak fiber strain in the CC). Feature-based machine learning classifiers including deep learning, SVM, and RF consistently outperformed all scalar injury metrics across all performance categories (e.g., leave-one-out accuracy of 0.828–0.862 vs. 0.690–0.776, and .632+ error of 0.148–0.176 vs. 0.207–0.292). Further, deep learning achieved the best cross-validation accuracy, sensitivity, AUC, and .632+ error. These findings demonstrate the superior performances of deep learning in concussion prediction and suggest its promise for future applications in biomechanical investigations of traumatic brain injury.

Introduction

Traumatic brain injury (TBI) resulting from blunt head impact is a leading cause of morbidity and mortality in the United States [1]. The recent heightened public awareness of TBI, especially of sports-related concussion [2,3], has prompted the Institute of Medicine and National Research Council of the National Academies to recommend immediate attention to address

the biomechanical determinants of injury risk and to identify effective concussion diagnostic metrics and biomarkers, among others [4].

Impact kinematics such as linear and rotational accelerations are convenient ways to characterize impact severity. Naturally, these simple kinematic variables and their more sophisticated variants have been used to assess the risk of brain injury. As head rotation is thought to be the primary mechanism for mild TBI (mTBI) including sports-related concussion, most kinematics metrics include rotational acceleration or velocity, either solely (e.g., rotational injury criterion (RIC), power rotational head injury criterion (PRHIC) [5], brain injury criterion (BrIC) [6], and rotational velocity change index (RVCI) [7]) or in combination with liner acceleration [8].

Kinematic variables, alone, do not provide regional brain mechanical responses thought to cause injury [9]. Validated computational models of the human head are, in general, believed to serve as an important bridge between external impact and tissue mechanical responses. Model-estimated, response-based injury metrics are desirable, as they can be directly related to tissue injury tolerances. Commonly used tissue response metrics include peak maximum principal strain and cumulative strain damage measure (CSDM; [10]) for the whole brain. More recently, white matter (WM) fiber strain [11–14] is also being explored as a potential improvement. There is growing interest in utilizing model-simulated responses to benchmark the performance of other kinematic injury metrics [6,15–17].

Regardless of these injury prediction approaches (kinematic or response-based), they share some important common characteristics. First, they have utilized a single injury dataset for "training" and performance evaluation. Often, this was performed by fitting a univariate logistic regression model to report the area (AUC) under the receiver operating curve (ROC) [8,12,13,18]. However, without cross-validation using a separate "testing dataset", there could be uncertainty how the metrics perform when they are, presumably, deployed to predict injury on fresh, unmet impact cases [12,19]. This is an important issue seemingly under-appreciated, given that AUC especially from a single training dataset provides an average or aggregated performance of a procedure but does not directly govern how a clinical decision, in this case, injury vs. non-injury diagnosis, is made.

Second, an explicit, pre-defined kinematic or response metric is necessary for injury prediction. While candidate injury metrics are typically from known or hypothesized injury mechanisms (e.g., strain), they are derived empirically. For response-based injury metrics, they are also pre-defined in a specific brain region of interest (ROI) such as the corpus callosum (CC) and brainstem. However, they do not consider other anatomical regions or functionally important neural pathways. The commonly used peak maximum principal strain and CSDM describe the peak response in a single element or the volume fraction of regions above a given strain threshold, respectively. However, they do not (and cannot) inform the location or distribution of brain strains that are likely critical for concussion, given the widespread neuroimaging alterations [20] and a diverse spectrum of clinical signs and symptoms [21] observed in the clinic.

Consequently, even when using the same reconstructed American National Football League (NFL) head impacts, studies have found inconsistent "optimal" injury predictors (e.g., maximum shear stress in the brainstem [22], strain in the gray matter and $CSDM_{0.1}$ (using a strain threshold of 0.1) in the WM [18], peak axonal strain within the brainstem [13], or tract-wise injury susceptibilities in the superior longitudinal fasciculus [14]). Most of these efforts are essentially "trial-and-error" in nature as they attempt to pinpoint a specific variable in a particular ROI for injury prediction. However, no consensus has reached on the most injury discriminative metric or ROI. Without accounting for the location and distribution of brain responses that are likely critical to concussion, critical information is lost.

Injury prediction is a binary classification. Besides the baseline univariate logistic regression, there have been numerous algorithmic advances in classification, including feature-based machine learning and, more recently, deep learning [23,24]. Instead of relying on a single, explicit scalar metric that could suffer from loss of information, feature-based machine/deep learning techniques employs multiple features to perform classification. However, despite their successes [23,24], application of feature-based machine/deep learning in TBI biomechanics for injury diagnosis is extremely limited or even non-existent at present. A recent study utilized SVM to predict concussion [25]. However, it was limited to kinematic variables (vs. brain responses) and two injury cases, which did not allow for cross-validation.

Conventional machine learning classifiers such as support vector machine (SVM) and random forest (RF) have been widely used in medical imaging [26,27] and computer vision [28] applications. Deep learning is the most recent advancement in feature-based classification, and it has achieved remarkable success in a wide array of science domains (see [23] for a recent review). This technique has already been successfully applied in numerous neuroimaging analyses, including registration [29], segmentation [30], and WM fiber clustering based on learned shape features [31]. For neurological disease classification, applications include the use of deep Convolutional Neural Network (CNN) for Alzheimer's detection (as reviewed in [32]), and fully connected Restricted Boltzmann Machine to detect mTBI categories based on diffusion tensor imaging (DTI) parameters [33]. However, this technique has not been employed for TBI prediction using brain tissue mechanical responses such as WM fiber strain (i.e., stretch along WM fiber directions). Unlike conventional neuroimages where tissue boundaries readily serve as image features for segmentation and registration, fiber strain responses as a result of mTBI are diffuse [20,34]. This makes it difficult to directly employ CNN-based techniques that are often built on local spatial filters (e.g., 3D CNN designed for tumor segmentation and measurement [35]).

Deep learning techniques are advancing rapidly. Instead of applying the most recent neural network architectures that are still under active development [31,32,36,37], here we chose a more conventional approach to first introduce this important research tool into the TBI biomechanics research field. Implicit features of the entire voxel-wise WM fiber strains were generated from reconstructed head impacts for concussion classification. Performances of the deep learning classifier were compared against baseline machine learning and univariate logistic regression methods in a leave-one-out cross-validation framework. This was important to ensure an objective comparison and to maximize rigor, which has often been overlooked in other biomechanical studies that only reported AUC from a single training dataset [8,12,13,18]. These injury prediction strategies are important extensions to previous efforts, which may provide important fresh insight into how best to objectively predict concussion in the future.

Materials and methods

The Worcester Head Injury Model (WHIM) and WM fiber strain

We used the Worcester Head Injury Model (WHIM; Fig 1 [11,34]) to simulate the reconstructed NFL head impacts [38,39]. Descriptions of the WHIM development, material property and boundary condition assignment, and quantitative assessment of the mesh geometrical accuracy and model validation performances have been published previously. Briefly, the WHIM was created based on high resolution T1-weighted MRI of an individual athlete. DTI of the same individual provided averaged fiber orientations at each WM voxel location [11].

The 58 reconstructed head impacts include 25 concussions and 33 non-injury cases. Identical to previous studies [14,18,39,40], head impact linear and rotational accelerations were

preprocessed before applying to the WHIM head center of gravity (CG) for brain response simulation. The skull and facial components were simplified as rigid-bodies as they did not influence brain responses.

Peak WM fiber strain, regardless of the time of occurrence during impact, was computed at each DTI WM voxel (N = 64272; [34]). For voxels not corresponding to WM, their values were padded with zeroes. This led to a full 3D image volume encoded with peak WM fiber strains (with surface rendering of the segmented WM shown in Fig 1C). They served as classification features for deep neural network training and concussion prediction. The choice of fiber strain instead of more commonly used maximum principal strain was because of its potentially improved injury prediction performance [12,13,34]. As no neuroimages were available for the 58 impact cases, injury detection using a previous deep learning technique based on DTI parameters [33] was not applicable in this study.

Deep learning: Background

Deep learning has dramatically improved the state-of-the-art in numerous research domains (see a recent review in Nature Methods [23]). However, its application in TBI biomechanics is nonexistent at present. This technique allows models composed of multiple processing layers to learn representations of data with multiple levels of abstraction [23]. A deep learning neural network uses a collection of logical units and their activation statuses to simulate brain function. It employs an efficient supervised update method [41] or an unsupervised network training strategy [42]. This makes it feasible to train a "deep" (e.g., more than 3 layers) neural network, which is ideal for learning large scale and high dimensional data.

For a deep learning neural network, the l-th layer transforms an input vector from its lower layer, a_{l-1}, into an output vector, a_l, through the following forward transformation:

$$x_l = W_l a_{l-1} + b_l \tag{1}$$

$$a_l = \sigma_l(x_l) \tag{2}$$

where matrix W_l is a linear transform describing the unit-to-unit connection between two adjacent, l-th and (l-1)-th, layers, and b_l is a bias offset vector. Their dimensions are configured to produce the desired dimensionality of the input and output, with the raw input data represented by x_0 (Fig 2). The nonlinear normalization or activation function, σ_l, can be defined as
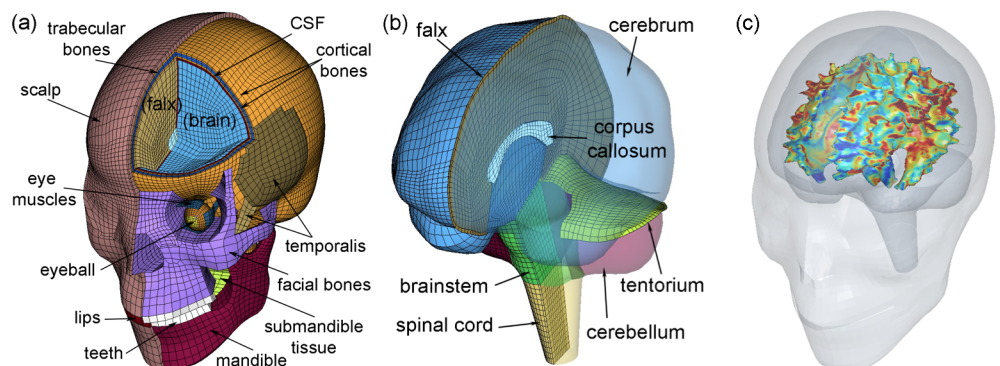
Fig 1. The Worcester Head Injury Model (WHIM). Shown are the head exterior (a) and intracranial components (b), along with peak fiber strain-encoded rendering of the segmented WM outer surface (c). The x-, y-, and z-axes of the model coordinate system correspond to the posterior–anterior, right–left, and inferior–superior direction, respectively. The strain image volume, which was used to generate the rendering within the co-registered head model for illustrative purposes, directly served as input signals for deep learning network training and concussion classification (see Fig 2).

https://doi.org/10.1371/journal.pone.0197992.g001
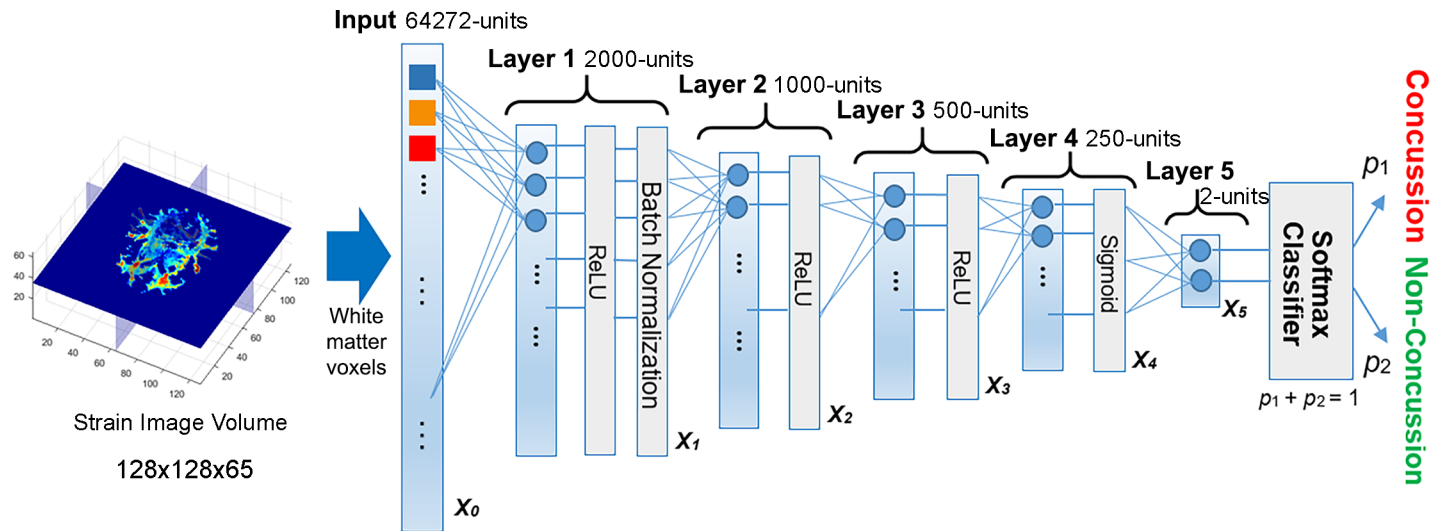
Fig 2. Structure of the deep learning network. The network contained five fully connected layers to progressively compress the fiber-strain-encoded image features, and ultimately, into a two-unit feature vector for concussion classification.

either a Sigmoid or a TanH function [43], or Rectified Linear Units (ReLU) [44] in order to suppress the output values for discriminant enhancement [45] and for achieving non-linear approximation [46]. Upon network training convergence, the optimized parameters, $W = \{W_l\}$ and $b = \{b_l\}$, are used to produce predictions of the cross-validation dataset. More details on the mathematics behind and procedures of deep network training are provided in the **Appendix**.

Deep learning: Network design and implementation

A systematic approach to designing an "optimal" deep learning network is still an active research topic [47]. As a clear rule is currently lacking, trial-and-error is often used to determine the appropriate number of layers and the numbers of connecting units in each layer. Here, we empirically developed a network structure composed of five fully connected layers (i.e., each unit in a layer was connected to all units in its adjacent layers; Fig 2), similarly to that used before [48]. The number of network layers was chosen to balance the trade-off between network structure nonlinearity and regularity.

The numbers of connecting units in each layer or the network dimension also followed a popular pyramid structure [48] to sequentially halve the number of connecting units in subsequent layers (i.e., a structure of 2000-1000-500-250 units for layers 1 to 4; Fig 2). Each layer performed feature condensation transform (Eqs 1 and 2) independently. The final feature vector, x_5, served as the input for injury classification. Table 1 summarizes the dimensions of the

Table 1. Summary of the dimensions of the weights and offset parameters, along with the normalization functions used to define the deep learning network. See **Appendix** for details regarding the normalization functions.

Parameter	Definition
W_l: 2D matrix	W_1: 2000×64272; W_2: 1000×2000; W_3: 500×1000; W_4: 250×500; W_5: 20×250
b_l: 1D vector	b_1: 2000 dim; b_2: 1000 dim; b_3: 500 dim; b_4: 250 dim; b_5: 2 dim
σ_l: normalization function	σ_1: ReLU + batch normalization; σ_2: ReLU; σ_3: ReLU; σ_4: Sigmoid; σ_5: no normalization (i.e., using an identity matrix)

weights, W_l, and offset vectors, b_l, as well as the normalization functions, σ_l, used to define the deep network. In total, the network contained over 1.31×10^8 independent parameters.

For the first three layers (i.e., layers 1 to 3 in Fig 2), ReLU were used that provided a sparser activation than TanH and Signmoid functions to allow faster and more effective training [44]. A batch normalization technique was also used to avoid internal co-variate shift as a result of non-normal distributions of the input and output values. This enhanced the network robustness [49]. In contrast, the last layer prior to classification (layer 4 in Fig 2) adopted a Sigmoid function to normalize output values to [0, 1], which was necessary to facilitate the Softmax classification [50].

Upon training convergence, the initial high dimensional feature vector was condensed into a more compact representation. A Softmax function, S (Eq A1 in **Appendix**), transformed the input feature vector, y, into a final output vector, (p_1, p_2). The corresponding vector values represented the probability of concussion (p_1) and non-injury (p_2), respectively, where $p_1 + p_2 = 1$, by necessity. Concussion was said to occur when $p_1 \geq 0.5$.

The network was trained via an ADAM optimization [51] in Caffe [52]. A number of hyper-parameters needed to be optimized to achieve a satisfactory performance. With trial and error, we selected a gradient descent step size or learning rate of 2×10^{-8} for all network layers, and the gradient descent momentum (i.e., the weight to multiply the gradient from the previous step in order to augment the gradient update in the current step) was set to 0.5. The default parameter values, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$, were used to prevent the weights from growing too fast. The training dataset was divided into a batch size of 5 for training (randomly resampled cases were added when the remaining batch was fewer than 5). The maximum number of epochs was 5000 (an epoch is a complete pass of the full training dataset through the neural network).

Early stopping for optimal training

An optimal number of training epochs achieves the best cross-validation accuracy at the minimum computational cost. However, this is not feasible to determine for fresh, unmet cases. Here, we monitored the validation accuracy of the training dataset to empirically determine a stopping criterion. Specifically, three training trials (in a leave-one-out cross-validation framework, see below) were generated to observe the convergence behaviors of the training and validation error functions (internally, 10% of the training dataset were used for validation within the deep learning training iterations; Eq A4 in **Appendix**; Fig 3). The training error function asymptotically decreased with the increase in the number of epochs. The validation error function initially decreased, as expected, but started to increase after sufficient epochs, indicating overfitting has occurred.

These observations suggested the use of an "early stopping criterion" [50] to ensure sufficient training with a minimum number of epochs. Initially, 300 training epochs were empirically used to monitor the validation error convergence behavior [53]. If validation error did not decrease, the network training was considered as failed due to a poor initialization and the training would terminate. With the chosen learning rate 2×10^{-8}, we found that the network always converged within [1000, 5000] epochs, which was set as an admissible range of epochs.

Concussion classification and performance evaluation

An objective evaluation of the concussion classification performances was important. Previously, a repeated random subsampling framework was employed to split the injury cases into independent and non-overlapping training and testing datasets [14]. Given the relatively small sample size (N = 58), here we adopted a leave-one-out cross-validation for performance
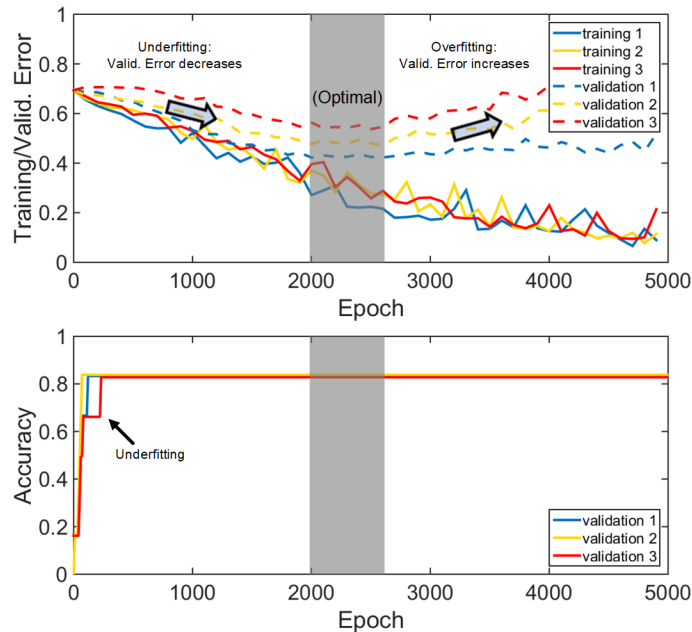
Fig 3. Illustration of the training and validation error functions and the corresponding validation accuracy.
(Top): Error functions from three deep learning training trials; (Bottom): the corresponding validation accuracy
(based on the 10% training dataset used for validation internally), vs. training epochs for three randomly generated
trials. Maximum validation accuracies based on validation datasets were achieved using an early-stopping criterion
after 2000 epochs.

evaluation. This maximized the training dataset so that to allow mimicking a real-world injury
prediction scenario by potentially optimizing the prediction on a fresh, unmet head impact.

Performance comparison against baseline machine learning classifiers

Conventional SVM and RF were selected as baseline classifiers to benchmark the performance
of deep learning. Typically, a machine learning classifier requires an explicit feature selection
to reduce input dimensionality and remove redundant, irrelevant, and noisy features from the
input data in order to improve performance [54]. However, there is no standard approach for
feature selection. For example, while the F-score approach is common for SVM [55], RF offers
feature selection by itself [56]. In contrast, an explicit feature selection is not necessary in deep
learning [57] as this is automatic during the optimization to maximize the input-output corre-
lation. For completeness, here we conducted classification first without feature selection using
the entire dataset as input to provide a reference performance for each classifier. After feature
selection, they were compared in a more typical scenario for the two baseline machine learning
techniques.

To avoid the classical feature selection bias problem [58], independent feature selections
were performed for each leave-one-out cross-validation trial. Specifically, only the training
dataset (N = 57), excluding the testing data point, were used for feature selection, with either
the F-score or RF-based approach. Using the recommended strategy [55], the F-score ap-
proach retained approximately 4% of features (N = 2566; empirically determined to yield the
highest cross-validation accuracy for SVM). For the RF-based method, a simplified variant of
the conventional "gini" importance ranking approach [56] was used. A total of 5000 randomly
initialized runs of RF were first conducted so that all of the voxels had a chance to serve as an
important feature. After each run, the top 1% highest ranked features were retained to vote on

a voxel-wise basis. The top 1% most frequently voted voxels (N = 643) among all of the runs were finally selected. The "top 1% criteria" were similarly determined empirically to yield the highest cross-validation accuracy for RF. After the 58 independent feature selections, a probability map was generated based on the frequency of each WM voxel selected as an important feature for classification.

A linear kernel was used for SVM [55]. For RF, the numbers of decision trees and depths were determined empirically to maximize cross-validation accuracy. They were 45 and 64 without feature selection, and 75 and 8, or 75 and 12, respectively, when using the F-score or RF for feature selection. As RF depended on a random initialization, 100 RF trials were conducted for each training/injury prediction. For deep learning with feature selection, a smaller neural network with 5 fully connected layers of dimensions of 500-250-125-60-2 was designed to accommodate the substantially reduced feature size, which resulted in 4.85×10^5 independent parameters. The learning rate was adjusted to 1×10^{-6}. Other hyper-parameters remained unchanged.

Performance comparison against scalar injury metrics

In TBI biomechanics research, univariate logistic regression is the most commonly used method to report the AUC of a single training dataset [5,6,8,13,22]. They rely on a scalar response metric which is essentially a single, pre-defined feature. The following four injury metrics were used for further performance comparison: Brain Injury Criteria (BrIC [6]; a kinematic metric found to correlate the best with strain-based metrics in diverse automotive impacts [15]), CSDM [10] for the whole brain (CSDM-WB) and the CC (CSDM-CC) based on maximum principal strain, as well as peak WM fiber strain in the corpus callosum (Peak-CC; [13,14]). The critical angular velocities for BrIC depend on the model used. For WHIM, they were 30.4 rad/s, 35.6 rad/s, and 23.5 rad/s along the three major axes, respectively [16]. For CSDM, an "optimal" strain threshold of 0.2 was used, which was to maximize the significance of injury risk-response relationship for the group of 50 deep WM regions using the same reconstructed NFL injury dataset [14].

Upon training convergence or after fitting, all classifiers generated a probability score for each of the impact case in the training and testing datasets. For deep learning, this was p_1 in Fig 2 (Eq A1 in **Appendix**), which allowed constructing an ROC to report AUC (*perfcurve.m* in Matlab). For each classifier, an AUC for each training dataset was calculated based on 57 impact cases for each of the 58 independent injury predictions (as necessitated by the leave-one-out cross-validation framework). An average AUC was then reported. In contrast, a single AUC value for the testing dataset was obtained based on the probability scores of the 58 independent predictions.

Data analysis

Simulating each head impact of 100 ms duration in Abaqus/Explicit (Version 2016; Dassault Systèmes, France) required ~50 min on a 12-CPU Linux cluster (Intel Xeon E5-2680v2, 2.80 GHz, 128 GB memory) with a temporal resolution of 1 ms. An additional 9 min was needed to obtain element-wise cumulative strains (single threaded). The classification framework was implemented on Windows (Xeon E5-2630 v3, 8 cores, 16 GB memory) with GPU acceleration (NVidia Titan X Pascal, 12 GB memory). Training each deep neural network typically required ~15 min and ~7 min for the two deep networks, respectively, but subsequent injury prediction was real-time (<0.01 sec).

For all the concussion classifiers, their performances were compared in terms of leave-one-out cross-validation accuracy, sensitivity, and specificity, as well as AUCs for both the training
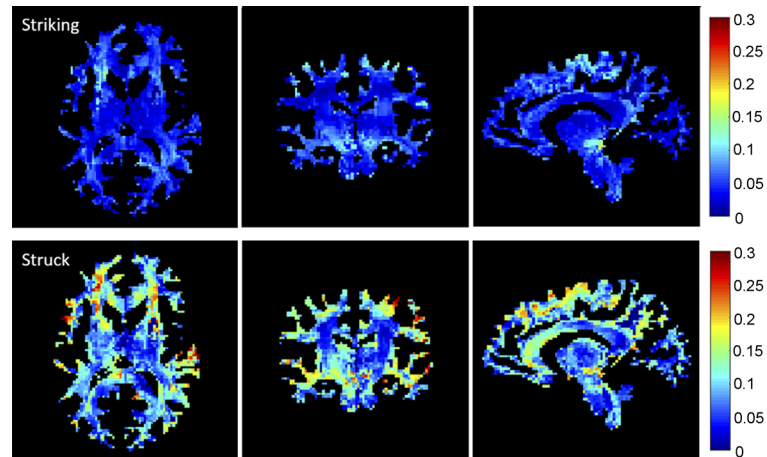
Fig 4. Cumulative WM fiber strains on representative orthogonal planes for a pair of striking (non-injury) and struck (concussed) athletes.

https://doi.org/10.1371/journal.pone.0197992.g004

and testing datasets. In addition, results from bootstrapped samples were further provided to account for the variation in prediction as compared with the traditional leave-one-out procedure [58]. All data analyses were conducted in MATLAB (R2017b; Mathworks, Natick, MA).

Results

Strain-encoded whole-brain image volume

Fig 4 illustrates and compares peak WM fiber-strain-encoded images on three orthogonal planes for a pair of striking and struck (non-injured and concussed, respectively) athletes involved in the same head collision. Without feature selection, deep learning directly utilized all of the strain-encoded WM image features for training and concussion classification.

Deep learning vs. SVM and RF

Without an explicit feature selection, deep learning outperformed both SVM and RF in accuracy and specificity, with sensitivity slightly lower than that of RF (Table 2). SVM performed the worst in all categories. Feature selection improved the performances of both SVM and RF in all categories, regardless of the specific feature selection approach. However, only RF-based feature selection slightly improved the accuracy over deep learning, at the cost of slightly poorer specificity (Table 3). Fig 5 shows the probability maps indicating the frequency of each WM voxel serving as an important feature for classification using either the F-score or RF-based approach. Features identified by the former was more substantial because 4% of all WM voxels were selected from each trial, vs. only 1% for the latter method. For the RF-based

Table 2. Summary of results. Shown are leave-one-out cross-validation accuracy, sensitivity, and specificity based on the testing dataset for the three feature-based machine learning classifiers. No feature selection was conducted and WM voxels of the entire brain were used for classification. For RF, the 95% confidence intervals (CI) were also reported based on the 100 random trials.

	Deep learning	SVM	RF
Accuracy	**0.845**	0.724	0.810 (0.759–0.862)
Sensitivity	0.760	0.640	**0.800** (0.699–0.880)
Specificity	**0.909**	0.788	0.818 (0.772–0.879)

https://doi.org/10.1371/journal.pone.0197992.t002

Table 3. Performance summary. Shown are accuracy, sensitivity, and specificity of the three feature-based classifiers when using either the F-score or RF-based approach for feature selection prior to classification (95% CI for RF in parentheses).

|	Deep learning		SVM		RF	
	F-score	**RF-feat**	**F-score**	**RF-feat**	**F-score**	**RF-feat**
Accuracy	0.845	**0.862**	0.828	0.828	0.828 (0.793–0.862)	0.842 (0.810–0.862)
Sensitivity	**0.880**	0.840	0.800	0.760	0.768 (0.720–0.831)	0.787 (0.760–0.840)
Specificity	0.818	0.879	0.848	0.879	0.873 (0.848–0.909)	**0.883** (0.849–0.909)

https://doi.org/10.1371/journal.pone.0197992.t003

method, the right superior longitudinal fasciculus (SLF-R) and left external capsule (EC_L) were two dominant regions often selected for classification.

Feature-based machine learning vs. scalar injury metrics using univariate logistic regression

The best performing deep learning, SVM and RF classifiers in terms of accuracy were obtained with RF-based feature selection (Table 4). They all had significantly higher performances in all categories than the scalar injury metrics from univariate logistic regression. Deep learning continued to perform the best among all classifiers in AUC using the testing dataset. With the leave-one-out scheme, ROCs were produced for each classifier based on the testing dataset (Fig 6). Two additional ROCs corresponding to the best and worst AUCs, respectively, were also produced for each classifier from the training datasets, as were typically reported in TBI biomechanical studies ([8,12,13,18]; Fig 7).

The classifiers were further compared using an out-of-bootstrap approach [59] based on 100 individual bootstrapped samples (Table 5), where unselected samples in each trial served as testing dataset for validation. Deep learning had the best accuracy and AUC, along with the smallest .632+ error [58].

Finally, all classifiers were further evaluated by re-running the training/testing using randomized concussion/non-concussion labels. For each classifier based on 50 individual trials, none of the resulting mean accuracy, sensitivity, or specificity was significantly different from 0.5 ($p > 0.05$ via two-tailed t-tests). This indicated all classifiers had reliable prediction power, no selection bias was likely, and that they indeed captured features important for classification [60].
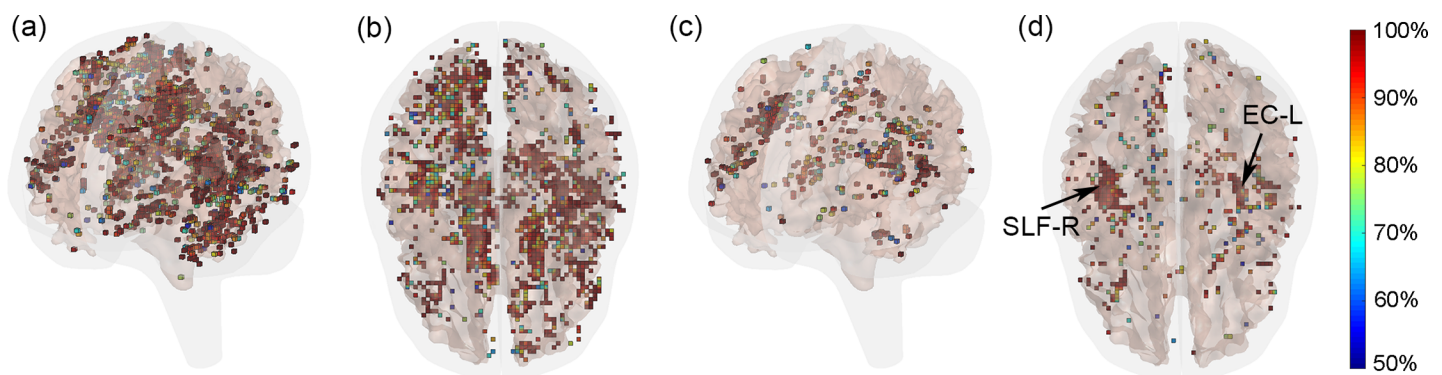
Fig 5. Probability maps for WM voxels selected. (a and b): using the F-score or (c and d) RF-based approach based on 58 independent feature selections. In each trial, the two approaches selected 4% and 1%, respectively, of the WM voxels as features. To improve visualization, only voxels with a probability greater than 50% (i.e., selected by at least 29 times) are shown. For the RF-based approach, SLF-R and EC-L were two dominant regions often selected for classification. See Matlab figure (S1 Fig) in the **supplementary** for interactive visualization.

https://doi.org/10.1371/journal.pone.0197992.g005

Table 4. Performance summary of the best performing feature-based classifiers (all with RF feature selection) as well as of the four scalar metrics from univariate logistic regression. Accuracy, sensitivity, specificity and AUC were reported based on the 58 separate injury predictions in the leave-one-out cross-validation framework. The average AUC measures (and 95% CI) for the training datasets were also reported.

	Deep learning	SVM	RF (95% CI)	BrIC	CSDM-WB	CSDM-CC	Peak-CC
Accuracy	**0.862**	0.828	0.842 (0.810–0.862)	0.776	0.741	0.776	0.690
Sensitivity	**0.840**	0.760	0.787 (0.760–0.840)	0.640	0.640	0.760	0.600
Specificity	0.879	0.879	**0.883 (0.849–0.909)**	0.879	0.818	0.788	0.758
AUC-Testing	**0.892**	0.872	0.856	0.781	0.786	0.771	0.737
AUC-Training average (95% CI)	0.967 (0.933, 0.978)	0.963 (0.951, 0.981)	**1.000 (1.000, 1.000)**	0.805 (0.797, 0.831)	0.838 (0.831, 0.860)	0.815 (0.807, 0.843)	0.770 (0.760, 0.791)

https://doi.org/10.1371/journal.pone.0197992.t004

Discussion

Developing an accurate and reliable injury predictor is one of the cornerstones in TBI biomechanics research for decades. Much of the work has so far focused on developing a single, scalar metric to describe impact severity and to predict injury. Numerous kinematics and model-estimated response variables have been proposed. Nonetheless, an "optimal" injury metric remains elusive and does not yet exist. However, a single scalar metric may not be sufficient for mTBI, including concussion, given the widespread neuroimaging alterations [20] and a diverse spectrum of clinical signs and symptoms [61] observed in the clinic. In this study, instead of similarly attempting to pinpoint an explicit response measure pre-defined in a specific ROI, we employed voxel-wise WM fiber strains from the entire brain as implicit features for injury prediction. The classical injury prediction was formulated into a supervised classification. Deep learning automatically distilled the most discriminative features from the strain-encoded image volumes for concussion classification. This was in sharp contrast to the current common approach in which a pre-defined scalar feature was essentially "hand-picked" to fit a univariate logistic regression for classification.
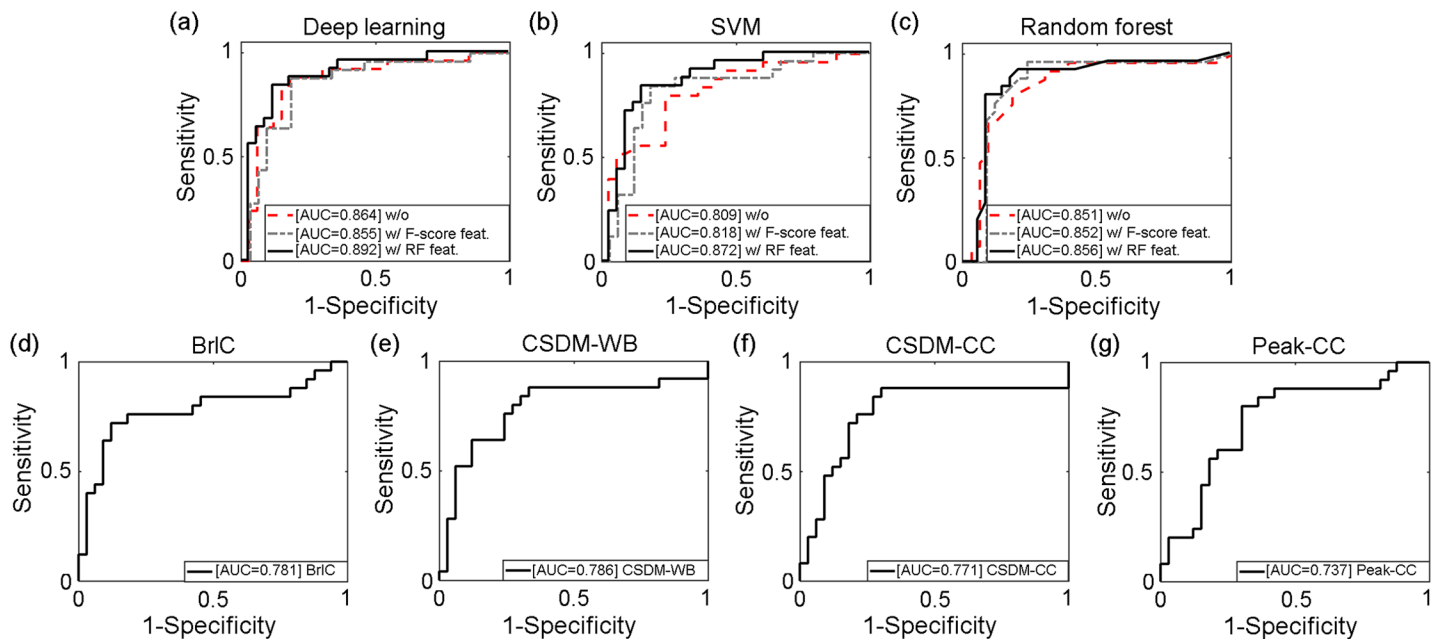
Fig 6. Comparisons of ROCs based on the testing dataset for the total of 7 classifiers.

https://doi.org/10.1371/journal.pone.0197992.g006
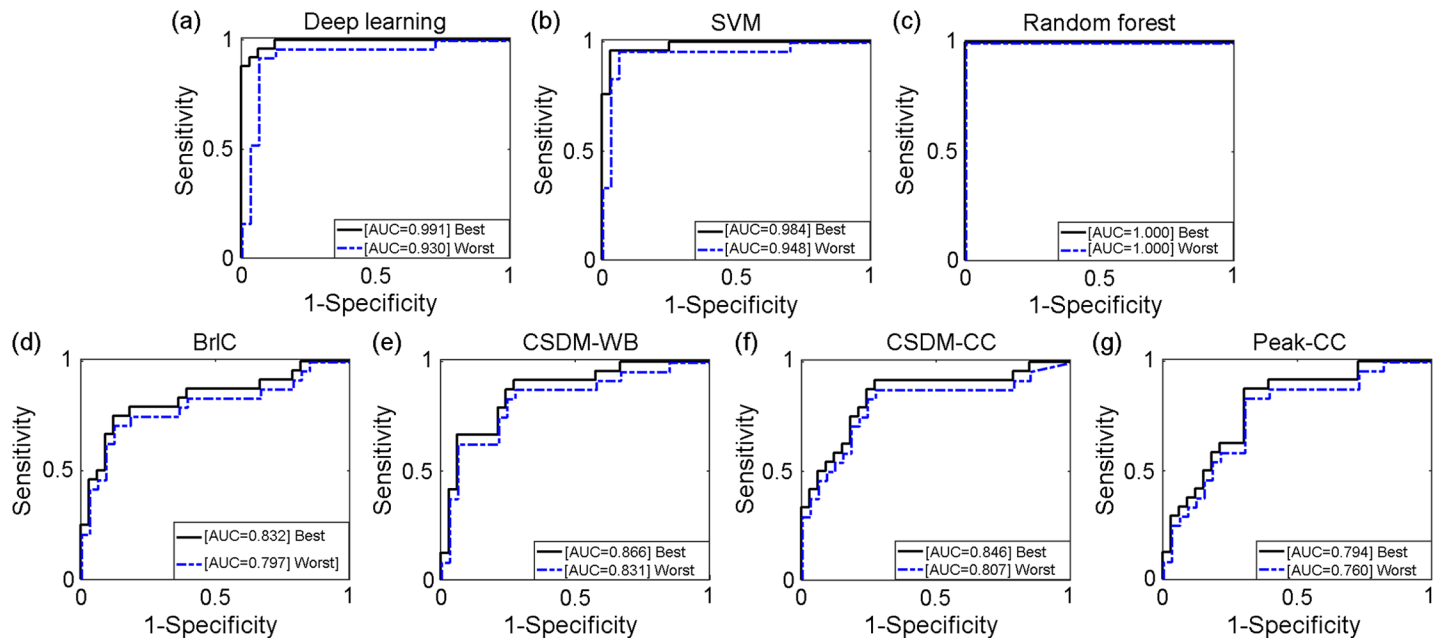
Fig 7. Comparisons of ROCs based on the training datasets. For the deep/machine learning techniques, only results from those with the RF-based feature selection are shown. The two ROCs correspond to the best and worst AUC, respectively.

Our results demonstrated that all of the injury metrics evaluated here were able to capture important features in the data to inform classification, based on the randomized label tests. However, the feature-based machine learning, including deep learning, significantly outperformed all of the scalar injury metrics selected here, in nearly all of the performance categories with either the leave-one-out (Table 4) or bootstrapped cross-validation approach (Table 5). Only BrIC had a better specificity using the bootstrapped cross-validation (Table 5). Among the feature-based classifiers, deep learning also outperformed the other two baseline approaches in cross-validation accuracy regardless of whether features were first selected (Tables 2–5). Both the F-score and RF-based approaches improved the performances of SVM and RF. However, the latter was more effective for the RF classifier with increased accuracy, sensitivity and specificity (Table 3). The RF-based feature selection also improved the accuracy and sensitivity for the deep learning classifier, but at a cost of lowering specificity (Tables 2 and 3). In terms of the improved .632+ error rate, feature-based metrics continued to outperform other scalar metrics, with deep learning being the best (Table 5).

In terms of AUC that is widely used in current TBI biomechanics research [8,12,13,18], the training dataset consistently generated larger scores than their counterparts using the testing dataset (Table 4), with RF even achieving a perfect AUC score of 1.0 (Fig 7). All feature-based

Table 5. Summary of results. Shown are out-of-bootstrap accuracy, sensitivity, specificity, and AUC (mean and 95% CI [59]), along with .632+ error [58] based on 100 bootstrapped trials using the best performing feature-based classifiers (all with RF feature selection) as well as those of the scalar metrics from univariate logistic regression.

	Deep learning	SVM	RF	BrIC	CSDM-WB	CSDM-CC	Peak-CC
Accuracy	**0.800** (0.650–0.915)	0.767 (0.614–0.903)	0.784 (0.628–0.903)	0.781 (0.636–0.950)	0.75 (0.600–0.882)	0.754 (0.583–0.882)	0.678 (0.522–0.833)
Sensitivity	0.766 (0.414–1.000)	0.713 (0.308–1.000)	**0.769** (0.348–1.000)	0.665 (0.333–1.000)	0.671 (0.375–1.000)	0.734 (0.300–1.000)	0.588 (0.250–1.000)
Specificity	0.835 (0.617–1.000)	0.820 (0.554–1.000)	0.806 (0.565–1.000)	**0.878** (0.667–1.000)	0.816 (0.600–1.000)	0.784 (0.500–1.000)	0.763 (0.533–1.000)
AUC-Testing	**0.850** (0.729–0.979)	0.846 (0.714–0.977)	0.847 (0.712–0.986)	0.803 (0.643–0.974)	0.835 (0.664–1.000)	0.818 (0.600–0.967)	0.778 (0.568–0.926)
.632+ error	**0.148**	0.176	0.163	0.207	0.246	0.227	0.292

classifiers significantly outperformed the scalar metrics using either the testing (Fig 6) or train-
ing (Fig 7) dataset (average AUC of 0.873 vs. 0.769 for the testing dataset, vs. AUC of 0.977 and
0.807 for the training dataset, respectively). However, deep learning achieved the highest AUC
based on testing dataset (Table 4 and Fig 6).

The highest AUC from a single training dataset using the latest KTH model (of 0.9655,
when using peak WM fiber strain in the brainstem serving as the predictor [13]) was compara-
ble to that of the three feature-based predictors reported here (range of 0.963–1.000; Table 4).
However, no objective performance comparison can be made here as no cross-validation was
performed in that study.

Feature selection

Deep learning does not typically require an explicit feature selection [57], as it is performed
implicitly during the iterative training process. However, feature selection was important for
both SVM and RF, without which SVM had a rather poor performance. This was likely a typi-
cal "curse of dimensionality" due to the small sample size that led to data overfitting [55], espe-
cially since a simple linear kernel was used for classification. Both feature selection methods
were effective in improving performance. The RF-based approach consistently identified
regions in SLF-R and EC-L as important features (Fig 5). Incidentally, SLF-R was also found to
be one of the most injury discriminative ROIs based on injury susceptibility measures via
logistic regression [14]. The consistency here suggested concordance between the different
classification approaches based on the same dataset. However, caution must be exercised when
attempting to extrapolate this finding to other subject groups, particularly given that neuroi-
mages corresponding to a single subject were used here for the group of subjects that did not
account for individual variability. A subject-specific study would be desirable to address these
limitations in the future, which was not feasible here.

Feature-based classifiers vs. scalar injury metrics

Feature-based machine/deep learning classifiers utilized multiple features for classification.
It started from the entire voxel-wise WM fiber strains. With data-driven feature-selection
aimed at reducing redundant information in the input and to avoid data overfitting, multiple
features were retained for subsequent classification to maximize performance. In contrast,
scalar injury metrics relied on a single response variable often empirically pre-defined. Kine-
matic injury metrics, including BrIC, are typically constructed by using the peak magnitudes
of linear/rotational acceleration or velocity, and their variants. They characterize impact sever-
ity to the whole brain, but are unable to provide tissue response directly. While a head FE
model estimates tissue responses throughout the brain, only the peak response magnitude of a
single element in a pre-defined ROI (e.g., peak-CC) or a dichotomous volume fraction above a
certain threshold (e.g., CSDM-WB and CSDM-CC) is used for injury prediction. Similar to
kinematic injury metrics, critical information is lost on the location or distribution of peak
brain responses, even though such information is already available. Because of these inherent
limitations with scalar injury metrics, it was not surprising that all of the feature-based classifi-
ers significantly outperformed all of the scalar injury metrics, regardless of the performance
category.

Compared with scalar injury metrics, deep learning was the extreme opposite as it utilized
information from all of the WM voxels of the entire brain as input for classification. The
technique has also been successfully applied to three-dimensional neuroimages for injury
and severity detection [33]. Conceivably, this may enable a multi-modal injury prediction
combining both biomechanical responses (e.g., strain-encoded image volume in Fig 4) and

corresponding neuroimages such as DTI of the same subjects to improve injury prediction performance. This is beyond the capabilities of any kinematic or strain-based injury metrics currently in use.

In addition, a strain threshold was necessary to dichotomize the brain ROI volumes for CSDM measures. An "optimal" strain threshold of 0.2 was previously determined by maximizing the significance of risk-response relationship for the group of 50 deep WM ROIs [14]. While adjusting the strain threshold could provide additional fitting flexibility to further improve the injury prediction performances of the scalar injury metrics, it may also lead to inconsistencies in threshold when each individual ROIs were used for injury prediction. Similarly to the ill-advised effort in reaching the absolute "best" performance with deep learning, this is undesirable, as the strain threshold is related to the physical injury tolerance found from actual *in vivo/in vitro* injury experiments. Importantly, deep learning and the two baseline machine learning classifiers have consistently outperformed all scalar injury metrics using univariate logistic regression. Therefore, this suggests strong motivation for further investigation into the use of the more advanced feature-based concussion classifiers in the future.

Comparison with previous findings

With the same injury dataset, Zhao and co-workers analyzed the injury susceptibilities and vulnerabilities of the entire deep WM ROIs and neural tracts [14]. A univariate logistic regression of each individual ROI/neural tract was conducted to report accuracy, sensitivity, specificity, and training AUC averaged from 100 trials in a repeated random subsampling cross-validation framework. A direct comparison was not feasible here because a leave-one-out cross-validation scheme was adopted in this study instead. Nevertheless, deep learning continued to outperform or at least to be comparable to the performances of each individual ROI/neural tract (e.g., accuracy of 0.862 with deep learning vs. 0.852 using point-wise injury susceptibility in SLF-R). However, unlike the previous study that required registering the FE model to a WM atlas to identify ROIs/neural tracts, no registration or segmentation was necessary with deep/machine learning that used the entire WM voxels as input. In addition, the previous study relied on dichotomized injury susceptibilities, which depended on a strain threshold similarly to the CSDM metrics selected here. This was unnecessary with deep/machine learning.

Another study identified Peak-CC to considerably outperform BrIC in AUC using all of the reconstructed NFL impacts as a single training dataset (0.9488 vs. 0.8629 [13]). Here we reported the opposite (average AUC of 0.770 vs. 0.805 for Peak-CC and BrIC in the training dataset, respectively; Table 4). This suggested disparities between the two head injury models and their analysis approaches. Perhaps most notably, the two models differ in material properties (isotropic, homogeneous vs. anisotropic for the WM). In addition, they have different brain-skull boundary conditions (nodal sharing via a soft layer CSF vs. frictional sliding), mesh resolution (average size of 3.2 mm vs. 5.8 mm), method to calculate fiber strain (projection of a strain tensor vs. assigning averaged fiber directions directly to FE elements), and even segmentation of the CC [11].

Nevertheless, improving a model's injury predictive power is a constant process. Together with more well-documented real-world injury cases, further comparison of injury prediction performances across models is important to understand how best to improve. A high AUC in a training dataset does not necessarily indicate the same high level of AUC or other performance categories using the testing dataset. For example, CSDM-WB had a higher AUC in training (average value of 0.838, vs. 0.805 and 0.815 for BrIC and CSDM-CC; Table 4), but it performed worse in cross-validation accuracy (0.714 with CSDM-WB vs. 0.776 for BrIC and

CSDM-CC). Therefore, it is important that future studies utilize cross-validation, rather than training or fitting, performances for objective evaluation and comparison.

Limitations

The superior performances of the deep learning and baseline machine learning classifiers were encouraging. However, it must be recognized that only one head FE model and a single injury dataset were employed here for performance evaluation and comparison. As even validated head models could produce discordant brain responses [62], it is important to further evaluate whether similar performance gains are possible with estimated brain responses from other head injury models. In addition, errors in the reconstructed head impact kinematics [38] are well-known, and the resulting uncertainties in model results, and implications in injury prediction due to under-sampling of non-injury cases [13,18,22] have been extensively discussed. Further, this dataset does not consider the cumulative effects from repetitive sub-concussive head impacts, the importance of which is becoming realized. Therefore, the deep learning classifier trained here may not be readily applicable to other injury datasets and a fresh training is necessary.

Importantly, a feature-based deep learning classifier has not been applied to TBI biomechanics before, despite its numerous recent successes across a wide array of scientific domains [23]. The deep learning approach and cross-validation framework established here may set the stage for continual development and optimization of a response-based injury predictor in the future. With further cross-validation using more independent injury datasets, the value of deep learning in TBI biomechanical investigations will be better studied.

Nevertheless, limitations with deep learning and its challenges in brain injury biomechanical studies are also noted. First, empirical experience is often necessary to design the network structure, as a clear guideline is lacking. The fact that RF with feature selection outperformed deep learning in sensitivity (when no features were explicitly selected) may indicate that the deep neural network architecture may not be optimal, and there could still be room for improvement. Second, unlike scalar injury metrics relying on explicit features, deep learning behaves much like a "black box" without an obvious physical interpretation of the its internal decision mechanism. Therefore, although an explicit feature selection was not necessary with deep learning, it may still be valuable to provide insight into the most injury discriminative features.

Third, the rather small dataset available in our study (58 cases) also placed a practical challenge in adopting some more advanced network architectures that often require a large dataset. This was especially true as we used strain-encoded neuromimages rather than conventional MRI for classification. For example, most conventional neuroimage-based disease classifications utilize structural MRI from the large Alzheimers Disease Neuroimaging Initiative (ADNI) database for training and testing, with data sample size typically of hundreds or thousands [32]. For neural network-based classifiers, often they employ various deep CNNs augmented by a convolutional autoencoder [63], a transfer learning technique from a pretrained network [64] such as a modified VGG network [65], a 3D CNN architecture [32], or multimodal stacked deep polynomial networks [66]. Deep neural networks could also be applied to small datasets, e.g., by using transfer learning from pretrained, related tasks [67]. However, pretrained tasks based on strain-encoded neuroimages do not currently exist, and it merits further exploration whether knowledge from pretrained, unrelated tasks from other image modalities can be applied to strain neuroimages. Regardless, our work may serve as a starting baseline to motivate further development of advanced deep neural networks appropriate for concussion classification using strain neuroimages, either alone or in combination with other neuroimage modalities [32,33,66].

Finally, the limitation of WHIM using isotropic, homogeneous material properties of the brain was discussed [34]. In addition, a generic head model and the corresponding neuroimages of one individual, rather than subject-specific head models and individualized neuroimages, were used to study a group of athletes. Inter-subject variation in neuroimaging and uncertainty in strain responses on an individual basis could not be evaluated. Nevertheless, a generic model is a critical steppingstone towards developing individualized models and to couple with their own neuroimages for more personalized investigations in the future. This is analogous to the typical 50th percentile head models currently in use that do not yet directly correspond to detailed neuroimages [14].

Conclusion

We introduced a deep learning classifier into biomechanical investigations of traumatic brain injury. The technique utilized voxel-wise white matter fiber strains of the entire brain as input for concussion prediction. Based on reconstructed NFL head impacts, we showed that feature-based classifiers, including deep learning and two baseline machine learning classifiers, outperformed all of the four selected scalar injury metrics in all performance categories in a leave-one-out cross-validation framework. Deep learning also achieved higher performances than the two baseline machine learning techniques in cross-validation accuracy, sensitivity, and AUC. The deep neural network developed here was by no means optimal or was ready for deployment in a more typical, general population. Nevertheless, the superior performances of deep learning and conventional feature-based machine learning in concussion prediction, especially relative to the commonly used scalar injury metrics via univariate logistic regression, suggest its promise for future applications in biomechanical investigations of traumatic brain injury.

Supporting information

S1 Fig. A Matlab figure showing the probability maps of selected white matter voxels (above 50% probability) using either the F-score or RF-based approach.
(FIG)

Acknowledgments

Appendix

Deep learning backpropagation for supervised training

An objective error function from the previous network layer (see **Fig 2**) can be used to maximize the input-output correlation either in an unsupervised [68] or a supervised [69] manner to minimize training error. Here, we used a supervised method for concussion classification, as supported by Caffe [52]. A Softmax classifier [50] based on condensed feature vector was adopted. Mathematically, this classifier is defined as:

$$S_x(j) = \frac{e^{x(j)}}{\sum_k e^{x(k)}} \qquad (A1)$$

where $x(j)$ and $x(k)$ are the j-th and k-th element of the feature vector, x, respectively, obtained from the trained network (output from the final layer). The classifier was trained by minimizing the Cross-Entropy error function relative to the known data label, $t(k)$, of either 0 or 1 (representing concussion or non-injury, respectively, in our study) for a training dataset, x, and its corresponding classifier output, S_x [70]:

$$E(x) = -\sum_k [t(k)\log S_x(k) + (1 - t(k))\log(1 - S_x(k))] \tag{A2}$$

The total error, $E = \sum_x E(x)$, for the training dataset served as the objective function for training via a backpropagation algorithm, as described below.

For a deep learning network with parameters, $W = \{W_l\}$ and $b = \{b_l\}$, the error function in Eq A2 can be represented as $E(W,b)$, which quantifies the classification error between the predicted and ground-truth labels. Deep network training is to optimize W and b in order to minimize the error, E. An efficient approach is through a backpropagation algorithm [69]. First, the network performs a *forward propagation* (Eqs 1 and 2) to produce classification and the error function value. For a network of L layers, the gradient of the error function with respect to x_l at the l-th layer ($l \leq L$), $\delta_l = \nabla_{x_l} E$, can be iteratively computed via the following *backpropagation*:

$$\delta_L = \nabla_{a_L} E \odot \sigma'_L(x_L) \tag{A3}$$

$$\delta_l = W_{l+1}^T \delta_{l+1} \odot \sigma'_l(x_l) \tag{A4}$$

where $\odot$ is the element-wise product. These gradients are used to minimize E via a gradient descent algorithm. Eqs A3 and A4 are derived by the chain rule in calculus, and the mathematical details can be found in standard neural network textbook (e.g., [50] Chap. 4.7). After computing δ_l, the gradients with respect to W and b are finally obtained:

$$\nabla_{b_l} E = \delta_l \tag{A5}$$

$$\nabla_{W_l} E = \delta_l \otimes a_{l-1}^T \tag{A6}$$

where $\otimes$ represents the tensor product. The following pseudo algorithm describes the training process for a network of L layers.

1. Input training set X;

2. For each training sample x in X:

   a. Compute the forward transformations (Eqs 1 and 2) from layer 2 to L

   b. Compute the error, $\delta_L(x)$, (Eq A3) for layer L

   c. Compute backpropagation, $\delta_L(x)$, (Eq A4) from layer L to 2

3. Gradient descent (for a given step size, $\lambda > 0$):

   a. Update W: $W_l \leftarrow W_l - \frac{\lambda}{|X|}\sum_x \delta_l(x) \otimes a_{l-1}^T(x)$ from Layer L to 2

   b. Update b: $b_l \leftarrow b_l - \frac{\lambda}{|X|}\sum_x \delta_l(x)$ from Layer L to 2

The training continues until the network is converged to generate optimized network parameters, W and b, which are then fixed to perform classification on the cross-validation dataset.

Author Contributions

References

1. CDC. Report to Congress on Traumatic Brain Injury in the United States: Epidemiology and Rehabilitation. 2015. 10.1161/HYPERTENSIONAHA.111.186106

2. Chrisman S, Schiff M, Chung S, Herring S, Rivara F. Implementation of concussion legislation and extent of concussion education for athletes, parents, and coaches in washington state. Am J Sports Med. 2014; 42: 1190–96. https://doi.org/10.1177/0363546513519073 PMID: 24510067

3. Zonfrillo M, Kim K, KB A. Emergency department visits and head computed tomography utilization for concussion patients from 2006 to 2011. Acad Emerg Med. 2015; 22.

4. NRC I. Sports-related concussions in youth: improving the science, changing the culture [Internet]. Graham R, Rivara FP, Ford MA, Spicer CM, editors. Washington, DC; 2014. Available: http://www.ncbi.nlm.nih.gov/books/n/nap18377/pdf/

5. Kimpara H, Iwamoto M. Mild traumatic brain injury predictors based on angular accelerations during impacts. Ann Biomed Eng. 2012; 40: 114–26. https://doi.org/10.1007/s10439-011-0414-2 PMID: 21994065

6. Takhounts EGG, Craig MJJ, Moorhouse K, McFadden J, Hasija V. Development of Brain Injury Criteria (BrIC). Stapp Car Crash J. 2013; 57: 243–66. Available: http://www.ncbi.nlm.nih.gov/pubmed/24435734 PMID: 24435734

7. Yanaoka T, Dokko Y, Takahashi Y. Investigation on an Injury Criterion Related to Traumatic Brain Injury Primarily Induced by Head Rotation. SAE Tech Pap. 2015; 2015-01-1439. https://doi.org/10.4271/2015-01-1439Copyright

8. Rowson S, Duma SM. Brain Injury Prediction: Assessing the Combined Probability of Concussion Using Linear and Rotational Head Acceleration. Ann Biomed Eng. 2013; 41: 873–882. https://doi.org/10.1007/s10439-012-0731-0 PMID: 23299827

9. King AI, Yang KH, Zhang L, Hardy W, Viano DC. Is head injury caused by linear or angular acceleration? IRCOBI Conference. Lisbon, Portugal; 2003. pp. 1–12.

10. Bandak FA, Eppinger RH. A three- dimensional finite element analysis of the human brain under combined rotational and translational acceleration. Stapp Car Crash J. 1995; 38: 145–163.

11. Zhao W, Ford JC, Flashman LA, McAllister TW, Ji S. White Matter Injury Susceptibility via Fiber Strain Evaluation Using Whole-Brain Tractography. J Neurotrauma. 2016; 33: 1834–1847. https://doi.org/10.1089/neu.2015.4239 PMID: 26782139

12. Sullivan S, Eucker SA, Gabrieli D, Bradfield C, Coats B, Maltese MR, et al. White matter tract-oriented deformation predicts traumatic axonal brain injury and reveals rotational direction-specific vulnerabilities. Biomech Model Mechanobiol. 2014; 14: 877–896. https://doi.org/10.1007/s10237-014-0643-z PMID: 25547650

13. Giordano C, Kleiven S. Evaluation of Axonal Strain as a Predictor for Mild Traumatic Brain Injuries Using Finite Element Modeling. Stapp Car Crash J. 2014;November: 29–61. PMID: 26192949

14. Zhao W, Cai Y, Li Z, Ji S. Injury prediction and vulnerability assessment using strain and susceptibility measures of the deep white matter. Biomech Model Mechanobiol. 2017; 16: 1709–1727. https://doi.org/10.1007/s10237-017-0915-5 PMID: 28500358

15. Gabler LF, Crandall JR, Panzer MB. Assessment of Kinematic Brain Injury Metrics for Predicting Strain Responses in Diverse Automotive Impact Conditions. Ann Biomed Eng. 2016; 44: 3705–3718. https://doi.org/10.1007/s10439-016-1697-0 PMID: 27436295

16. Zhao W, Kuo C, Wu L, Camarillo DBDB, Ji S. Performance evaluation of a pre-computed brain response atlas in dummy head impacts. Ann Biomed Eng. 2017; 45: 2437–2450. https://doi.org/10.1007/s10439-017-1888-3 PMID: 28710533

17. Kuo C, Wu L, Zhao W, Fanton M, Ji S, Camarillo DB. Propagation of errors from skull kinematic measurements to finite element tissue responses. Biomech Model Mechanobiol. 2017; 17: 235–247. https://doi.org/10.1007/s10237-017-0957-8 PMID: 28856485

18. Kleiven S. Predictors for Traumatic Brain Injuries Evaluated through Accident Reconstructions. Stapp Car Crash J. 2007; 51: 81–114. 2007-22-0003 [pii] PMID: 18278592

19. Anderson AE, Ellis BJ, Weiss J a. Verification, validation and sensitivity studies in computational biomechanics. Comput Methods Biomech Biomed Engin. 2007; 10: 171–84. https://doi.org/10.1080/10255840601160484 PMID: 17558646

20. Bigler ED, Maxwell WL. Neuropathology of mild traumatic brain injury: Relationship to neuroimaging findings. Brain Imaging Behav. 2012; 6: 108–136. https://doi.org/10.1007/s11682-011-9145-0 PMID: 22434552

21. Duhaime A-C, Beckwith JG, Maerlender AC, McAllister TW, Crisco JJ, Duma SM, et al. Spectrum of acute clinical characteristics of diagnosed concussions in college athletes wearing instrumented helmets. J Neurosurg. 2012; 117: 1092–1099. https://doi.org/10.3171/2012.8.JNS112298 PMID: 23030057

22. Zhang L, Yang KHH, King AII. A Proposed Injury Threshold for Mild Traumatic Brain Injury. J Biomech Eng. 2004; 126: 226–236. https://doi.org/10.1115/1.1691446 PMID: 15179853

23. LeCun Y, Bengio Y, Hinton G. Deep learning. Nat Methods. 2015; 521: 436–444. https://doi.org/10.1038/nmeth.3707

24. Greenspan H, Ginneken B van, Summers RM. Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Trans Med Imaging. 2016; 35: 1153–1159. https://doi.org/10.1109/TMI.2016.2553401

25. Hernandez F, Wu LC, Yip MC, Laksari K, Hoffman AR, Lopez JR, et al. Six Degree-of-Freedom Measurements of Human Mild Traumatic Brain Injury. Ann Biomed Eng. 2015; 43: 1918–1934. https://doi.org/10.1007/s10439-014-1212-4 PMID: 25533767

26. Fan Y, Shen D, Davatzikos C. Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM. Med Image Comput Comput Interv–MICCAI 2005. 2005; 3749: 1–8. https://doi.org/10.1007/11566465_1

27. Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage. 2013; 65: 167–175. https://doi.org/10.1016/j.neuroimage.2012.09.065 PMID: 23041336

28. Bishop CM. Pattern Recognition and Machine Learning. Pattern Recognition. 2006. https://doi.org/10.1117/1.2819119

29. Wu G, Kim M, Wang Q, Gao Y, Liao S, Shen D. Unsupervised deep feature learning for deformable registration of MR brain images. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013. pp. 649–656. https://doi.org/10.1007/978-3-642-40763-5_80

30. De Brébisson A, Montana G. Deep neural networks for anatomical brain segmentation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2015. pp. 20–28. 10.1109/CVPRW.2015.7301312

31. Gupta V, Thomopoulos SI, Rashid FM, Thompson PM. FiberNET: An ensemble deep learning framework for clustering white matter fibers. bioRxiv. 2017; https://doi.org/10.1007/978-3-319-66182-7_63

32. Khvostikov A, Aderghal K, Benois-Pineau J, Krylov A, Catheline G. 3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies. 2018; Available: http://arxiv.org/abs/1801.05968

33. Cai Y, Ji S. Combining Deep Learning Networks with Permutation Tests to Predict Traumatic Brain Injury Outcome. In: Crimi A, Menze B, Maier O, Reyes M, Winzeck S, Handels H, editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Athens, Greece: Springer International Publishing AG 2016; 2016. pp. 259–270. https://doi.org/10.1007/978-3-319-55524-9_24

34. Ji S, Zhao W, Ford JC, Beckwith JG, Bolander RP, Greenwald RM, et al. Group-wise evaluation and comparison of white matter fiber strain and maximum principal strain in sports-related concussion. J Neurotrauma. 2015; 32: 441–454. https://doi.org/10.1089/neu.2013.3268 PMID: 24735430

35. Kamnitsas K, Ferrante E, Parisot S, Ledig C, Nori A V, Criminisi A, et al. DeepMedic for Brain Tumor Segmentation. Brainlesion Glioma, Mult Sclerosis, Stroke Trauma Brain Inj. 2016; https://doi.org/10.1007/978-3-319-55524-9_14

36. Ioannidou A, Chatzilari E, Nikolopoulos S, Kompatsiaris I. Deep Learning Advances in Computer Vision with 3D Data. ACM Comput Surv. 2017; 50: 1–38. https://doi.org/10.1145/3042064

37. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017. pp. 60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026

38. Newman JA, Beusenberg MC, Shewchenko N, Withnall C, Fournier E. Verification of biomechanical methods employed in a comprehensive study of mild traumatic brain injury and the effectiveness of American football helmets. J Biomech. 2005; 38: 1469–81. https://doi.org/10.1016/j.jbiomech.2004.06.025 PMID: 15922758

39. Newman J, Shewchenko N, Welbourn E, Welbourne E, Welbourn E. A proposed new biomechanical head injury assessment function-the maximum power index. Stapp Car Crash J. 2000; 44: 215–247. PMID: 17458729

40. Zhao W, Ji S. Brain strain uncertainty due to shape variation in and simplification of head angular velocity profiles. Biomech Model Mechanobiol. Springer Berlin Heidelberg; 2017; 16: 449–461. https://doi.org/10.1007/s10237-016-0829-7 PMID: 27644441

41. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323: 533–536. https://doi.org/10.1038/323533a0

42. Salakhutdinov R, Hinton GE. Deep Boltzmann Machines. Proc 12th Int Conf Artif Intell Statics. 2009; 448–455. 10.1109/CVPR.2009.5206577

43. Lecun Y, Bottou L, Orr GB, Muller K-R. Efficient backprop. Neural networks: tricks of the trade. 1998. pp. 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

44. Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. Proc 27th Int Conf Mach Learn. 2010; 807–814. 10.1.1.165.6419

45. Ruck D, Rogers S, Kabrisky M, Oxley M. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. IEEE Trans. 1990;

46. Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control signals Syst. 1989;

47. Schmidhuber J. Deep Learning in neural networks: An overview. Neural Networks. 2015. pp. 85–117. https://doi.org/10.1016/j.neunet.2014.09.003 PMID: 25462637

48. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science (80-). American Association for the Advancement of Science; 2006; 313: 504–507.

49. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:150203167. 2015; 1–11. https://doi.org/10.1007/s13398-014-0173-7.2

50. Bishop CM. Neural networks for pattern recognition. J Am Stat Assoc. 1995; 92: 482. https://doi.org/10.2307/2965437

51. Kingma DP, Ba JL. Adam: a Method for Stochastic Optimization. Int Conf Learn Represent 2015. 2015; 1–15. http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503

52. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. ACM International Conference on Multimedia. 2014. pp. 675–678. 10.1145/2647868.2654889

53. Bengio Y. Practical recommendations for gradient-based training of deep architectures. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2012; 7700 LECTU: 437–478. https://doi.org/10.1007/978-3-642-35289-8-26

54. Guyon I, Elisseeff A. Feature Extraction, Foundations and Applications: An introduction to feature extraction. Stud Fuzziness Soft Comput. 2006; 207: 1–25. https://doi.org/10.1007/978-3-540-35488-8_1

55. Chen Y, Lin C. Combining SVMs with Various Feature Selection Strategies. Strategies. 2006; 324: 1–10. https://doi.org/10.1007/978-3-540-35488-8_13

56. Breiman L. Random forests. Mach Learn. 2001; 45: 5–32. https://doi.org/10.1023/A:1010933404324

57. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013; 35: 1798–1828. https://doi.org/10.1109/TPAMI.2013.50 PMID: 23787338

58. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Second Edi. Hastie T, Tibshirani R, Friedman J, editors. Springer; 2008.

59. Beleites C, Baumgartner R, Bowman C, Somorjai R, Steiner G, Salzer R, et al. Variance reduction in estimating classification error using sparse datasets. Chemom Intell Lab Syst. Elsevier; 2005; 79: 91–100. https://doi.org/10.1016/J.CHEMOLAB.2005.04.008

60. Nisbet R, Iv JE, Miner G. Handbook of Statistical Analysis and Data Mining Applications: Robert Nisbet, John Elder IV, Gary Miner: 9780123747655: Amazon.com: Books. 2009.

61. Duhaime A-C, Beckwith JG, Maerlender AC, McAllister TW, Crisco JJ, Duma SM, et al. Spectrum of acute clinical characteristics of diagnosed concussions in college athletes wearing instrumented helmets: clinical article. J Neurosurg. 2012; 117: 1092–9. https://doi.org/10.3171/2012.8.JNS112298 PMID: 23030057

62. Ji S, Ghadyani H, Bolander RP, Beckwith JG, Ford JC, McAllister TW, et al. Parametric Comparisons of Intracranial Mechanical Responses from Three Validated Finite Element Models of the Human Head. Ann Biomed Eng. 2014; 42: 11–24. https://doi.org/10.1007/s10439-013-0907-2 PMID: 24077860

63. Vu TD, Yang HJ, Nguyen VQ, Oh AR, Kim MS. Multimodal learning using convolution neural network and Sparse Autoencoder. 2017 IEEE Int Conf Big Data Smart Comput BigComp 2017. 2017; 309–312. 10.1109/BIGCOMP.2017.7881683

64. Wang S, Shen Y, Chen W, Xiao T, Hu J. Automatic Recognition of Mild Cognitive Impairment from MRI Images Using Expedited Convolutional Neural Networks. International Conference on Artificial Neural Networks. Springer; 2017. pp. 373–380.

65. Billones CD, Demetria, Olivia Jan Louville D Hostallero DED, Naval. PC. Demnet: A convolutional neural net- work for the detection of alzheimer's disease and mild cognitive impair- ment. In Region 10 Conference (TENCON), 2016 IE. Region 10 Conference (TENCON), 2016 IEEE. IEEE; 2016. pp. 3724–3727.

66. Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease. IEEE J Biomed Heal Informatics. 2018; 22: 173–183. https://doi.org/10.1109/JBHI.2017.2655720 PMID: 28113353

67. Wagner R, Thom M, Schweiger R, Palm G, Rothermel A. Learning Convolutional Neural Networks From Few Samples. Available: https://pdfs.semanticscholar.org/07be/dfb53304ccab4929a7226b6fd4900f50e2a9.pdf

68. Hinton GE, Osindero S, Teh Y-W. A Fast Learning Algorithm for Deep Belief Nets. Neural Comput. 2006; 18: 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527 PMID: 16764513

69. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient Based Learning Applied to Document Recognition. Proc IEEE. 1998; 86: 2278–2324. https://doi.org/10.1109/5.726791

70. Tibshirani R. A Comparison of Some Error Estimates for Neural Network Models. Neural Comput. 1996; 8: 152–163. https://doi.org/10.1162/neco.1996.8.1.152