Recurrent Neural Network-based Acute Concussion Classifier using Raw Resting State EEG Data

Karun Thanjavur[1,*], **Arif Babul**[1,+], **Brandon Foran**[2+], **Maya Bielecki**[2+], **Adam Gilchrist**[2+], **Dionissios T. Hristopulos**[3,+], **Leyla R. Brucar**[4,+], **and Naznin Virji-Babul**[4,5,+]

[1]Department of Physics and Astronomy, University of Victoria, Victoria, British Columbia, V8P 5C2, Canada
[2]Department of Computer Science, Middlesex College, Western University, London, Ontario, N6A 5B7 Canada
[3]School of Electrical and Computer Engineering, Technical University of Crete, Chania, Greece
[4]Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, British Columbia, V6T 1Z3, Canada
[5]Department of Physical Therapy, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, V6T 1Z3, Canada
[*]karun@uvic.ca
[+]these authors contributed equally to this work

ABSTRACT

Concussion is a global health concern. Despite its high prevalence, a sound understanding of the mechanisms underlying this type of diffuse brain injury remains elusive. It is, however, well established that concussions cause significant functional deficits; that children and youths are disproportionately affected and have longer recovery time than adults; and recovering individuals are more prone to suffer additional concussions, with each successive injury increasing the risk of long term neurological and mental health complications. Currently, concussion management faces two significant challenges: there are no objective, clinically accepted, brain-based approaches for determining (i) whether an athlete has suffered a concussion, and (ii) when the athlete has recovered. Diagnosis is based on clinical testing and self-reporting of symptoms and their severity. Self-reporting is highly subjective and symptoms only indirectly reflect the underlying brain injury. Here, we introduce a deep learning Long Short Term Memory (LSTM)-based recurrent neural network that is able to distinguish between healthy and acute post-concussed adolescent athletes using only a short (i.e. 90 seconds long) sample of resting state EEG data as input. The athletes were neither required to perform a specific task nor subjected to a stimulus during data collection, and the acquired EEG data was neither filtered, cleaned of artefacts, nor subjected to explicit feature extraction. The LSTM network was trained and tested on data from 27 male, adolescent athletes with sports related concussion, bench marked against 35 healthy, adolescent athletes. During rigorous testing, the classifier consistently identified concussions with an accuracy of >90% and its ensemble-median Area Under the Curve (AUC) corresponds to 0.971. This is the first instance of a high-performing classifier that relies only on easy-to-acquire resting state EEG data. It represents a key step towards the development of an easy-to-use, brain-based, automatic classification of concussion at an individual level.

1 Introduction

Concussion or mild traumatic brain injury (mTBI) is an urgent public health concern. Canadian [1] and US data [2, 3] indicate an annual rate of reported mTBIs of 1,100 per 100,000 people, 75% of which involve children, youths and young adults [4]. The actual rate is likely much higher as individuals suffering from concussion often do not seek medical advice [5, 6]. Despite its high prevalence, a sound understanding of the mechanisms underlying mTBI remains elusive. What is clear is that concussions induce diffuse, heterogeneous, spatially distributed changes in brain structure and function, resulting in cognitive, motor, emotional and behavioral challenges that can persist for many months [7–9]. Children and youths are especially vulnerable: They are disproportionately affected by concussions [10] and take longer than adults to recover [7, 11]. The effects of concussion on their developing brain can lead to compromised brain and mental health, impairing learning, working and socializing [12]. Such disruptions at this critical developmental period in the lives of children and youths can be devastating.

The management of concussion suffers from a key challenge: There is no accepted, objective, brain-based tool to make an initial diagnosis of concussion. The current gold standard involves a clinical assessment by an experienced physician. This approach is difficult to standardize as it is dependent on the individual physician and relies on subjective reporting of the symptoms. Symptoms, however, do not appear to be directly correlated to the patho-physiological mechanisms responsible for the cognitive, emotional and behavioral deficits [13, 14] and several studies [14–23] have shown that structural and functional abnormalities in the brain persist even after symptom resolution, especially in pediatric patients. Manning et al. [20] and Hristopulos et al. [24] hypothesize that the increased vulnerability of children and youths to brain injury, as well as their longer recovery time, are consequences of the effects of concussion being overlaid on brains whose structural and functional organization are undergoing dynamic changes due to development. Moreover, recovering individuals are also at a higher risk of sustaining additional concussions, which can lead to neuro-psychiatric and neuro-degenerative disorders [25, 26]. Consequently, clinicians often find themselves in a quandary when trying to decide when pediatric patients can safely return to play/learn.

Numerous groups have sought to leverage neuroimaging data to not only detect mTBIs reliably and objectively, but also track brain recovery. At the group-level, structural changes in the integrity of white matter in adolescents has been observed using diffusion tensor imaging (DTI) [20, 27–32]. Similarly, studies of youths during "resting state" reveal significant mTBI-induced alterations in the functional organization of the brain with respect to their healthy counterparts, with key features being (i) a shift in spectral profile toward higher frequencies in the frontal brain regions, [33] (ii) an increase in functional connectivity (hyperconnectivity) [17, 34, 35], and (iii) disrupted information flow patterns [24], again particularly in the frontal regions. All of these studies rely on summary measures to characterize functional changes in the brain. However, because the brain injury due to mTBI is diffuse and the resulting changes in the brain structure and function are subtle, the summary measures cannot be used to statistically differentiate between healthy and injured at the *individual* level, leading the Radiological Society of North America to advise that "in pediatric patients, imaging should not be routinely obtained to diagnose mild TBI" because "there is insufficient evidence [to support] the routine clinical use of advanced neuroimaging techniques for diagnosis and/or prognostication at the individual patient level" [36].

Rapid advances in machine learning (ML) and artificial intelligence (AI) methods have opened up a new way to explore neuroimaging data [37–39]. ML evolved from the study of computer vision and computational learning theory, and is a method of data analysis that automates pattern recognition and model building. ML systems learn to identify relevant, often latent, features directly from the data, rendering explicit programming unnecessary. Consequently, they are especially suited for analysis of complex, high-dimensional datasets that themselves are products of often poorly understood dynamics. Focusing specifically on mTBI, several groups have attempted to develop ML algorithms to detect concussions. Examples of such attempts include the use of support vector machine (SVM) to distinguish between healthy and concussed individuals using explicit features derived from resting state magnetoencephalography (MEG) data [40] or resting state electroencephalography (EEG) recordings [41]; the use of random forest [42] and SVM [43] to classify subjects on the basis of structural connectivity features derived from diffusion magnetic resonance imaging (dMRI) data, with Vergara et al. [43] augmenting the dMRI measures with functional magnetic resonance imaging (fMRI)-based resting state functional network connectivity measures; the use of SVM to classify young adults using features from multiple task-related EEG recordings [44]; the use of an ensemble of shallow multilayer perceptron network with one hidden layer to distinguish between healthy and concussed adolescents using descriptors derived from resting state EEG data [45]; and most recently, the use of a deep learning convolution neural network to classify adolescents using EEG recordings of single-trial event-related potentials [46].

The goal of our research is to develop a reliable, accurate, easy-to-use deep learning neural network-based classifier (hereafter, *Conc*ussion *C*lassification *Net*work or *ConCNet*) that can distinguish between healthy and acute post-concussed individuals using only a *short* (i.e. 90 seconds long) sample of *raw resting state EEG* data (hereafter, raw90-rsEEG) as input. Here, "raw" refers to data that has neither been filtered nor cleaned to remove known artifacts; we discuss this further below. We explain our decision to use 90-second long samples in §4. Our decision to select EEG as the modality of choice was informed by a number of considerations, including the fact that (i) it is an inexpensive, non-invasive, direct, fine time resolution probe of the neural activity in individuals; (ii) at the group-level, it can reliably and objectively detect functional changes in the brain due to mTBI [24, 28, 47, 48]; and (iii) it does not require elaborate infrastructure, the introduction of contrast agents and radioisotopes, or exposure to radiation. Additionally, resting state data is easy to collect; the participant is neither required to perform a specific task nor subjected to a stimulus.

One challenge of working with EEG time series, particularly resting state EEG signals, is that they do not consist only of electrical signals from the brain, but also a variety of contaminants and distortions from environmental (e.g. strong power line interference at 60 Hz) and physiological sources (e.g. electrical signals generated by eye and head movements, eye

blinks and cardiac activity). Cleaning the data involves identifying and removing these without altering the signals from the brain. Strategies in use range from fully automated algorithms to mixed schemes where some artifacts are removed via automated algorithms and others are identified and removed manually. The lack of a standardized, consensus cleaning pipeline and more importantly, the lack of well-defined measures for quantifying a dataset's degree of "cleanliness" pose a serious challenge [49–51]. Additionally, commonly used artifact removal schemes are designed to treat known distortions, leaving open the possibility that the data may still be contaminated by latent artifacts. Even in the case of known artifacts, the identification-decontamination strategies are based on specific assumptions about the morphology of the distortions and the manner in which they interact with the signals of interest. These do not always fully capture the diverse ways in which the artifacts are manifest in real recordings [50, 52], and improper cleaning can introduce new distortions and unknown biases [50, 53]. Meisler et al. [54] assert that commonly used data cleaning methods do not necessarily improve brain state classification. In deliberately choosing to work with *raw* data, we circumvent these concerns and instead rely on the deep learning network to learn the relevant aspects of the data while ignoring the rest. In the process, we also avoid having to grapple with the complex problem of feature selection [55–57], which itself can be a source of bias [58–60], as well as confounding effects due to the EEG signals' strongly non-stationary character on the feature values [61, 62].

In designing our classifier, we have adopted a recurrent neural network architecture comprising two layers of bi-directional Long Short Term Memory (LSTM) [63, 64] units. An LSTM-based network is ideally suited for processing an ordered sequence of data points, such as the EEG time series data, where there is a high likelihood of temporal correlations and/or causal relationships between measurements at different time points and where the time lag between these could potentially be large [65]. For a more detailed exposition of the reasons informing our decision to adopt a bi-directional LSTM network, including recent findings from the studies of brain-computer interface systems, we refer the reader to §4. As for training and testing the network, we used a total of 216 EEG samples obtained from 27 male adolescent athletes with sports related concussion within 1 month of injury, who met the concussion diagnostic criteria consistent with the Berlin consensus statement [66], and 280 datasets from 35 age and sex matched healthy athletes. For the concussed participants, the date and time of the direct blow was documented by the team coach as per the consensus statement, and the team physician or an experienced physician with expertise in concussions made the diagnosis of concussion based on the Berlin consensus statement. The number of symptoms and symptom severity of each concussed subject were assessed using either the Sports Concussion Assessment Tool 3 (SCAT3) or the Child Sports Concussion Assessment Tool 3 (Child SCAT3) (see §4).

2 Results

In Table 1, we present the demographic information about the participants in this study. All of the concussed participants met the Berlin criteria and exhibited from 4 to 22 SCAT3 symptoms, at the time of testing. The symptoms most frequently reported included irritability, sensitivity to light, dizziness, fatigue, "don't feel right", and difficulty concentrating/remembering.

The development and assessment of *ConcNet* proceeded in three stages. Here, we briefly summarize these three stages, focusing on the results. Additional details are provided in the Methods section (§4).

Stage 1: Identifying a promising machine learning algorithm for *ConcNet*

A number of different machine learning techniques have been proposed to tackle classification problems (see Kotsiantis et al. [67] and Singh et al. [68] for comprehensive reviews). Prior to settling on a deep learning neural network-based classifier, and specifically an LSTM-based deep learning recurrent neural network, we experimented with other options, including using Support Vector Machine (SVM) to distinguish between healthy and concussed individuals on the basis of features derived from filtered and cleaned, resting state EEG data [41], and using an ensemble of multilayer perceptron networks to distinguish between the two populations on the basis of descriptors derived from filtered but not cleaned resting state data [45]. In all cases, the classifiers were trained and validated using a subset of the available data; tested on the remaining data using a *"blinded", one-shot* protocol, as described in §4; and evaluated on the basis of *Accuracy* and *Consistency*. *Accuracy* (and most of the other commonly used assessment metrics) are defined in Table 2 while *Consistency* is defined below.

Since our aim is to develop a clinically viable classifier, we considered an algorithm as promising only if its *Accuracy* was > 80% during the Stage 1 testing phase. *Accuracy* is the fraction of classifications that are correct. All except the LSTM-based classifier failed to reach this threshold, with the features-based algorithms achieving an accuracy of at the most 65%. The LSTM

classifier, the focus of this paper, achieved $Accuracy = 88.9\%$. The architecture of our first generation *ConcNet*, following Stage 1 training and tuning, is shown in Figure 1.

In *ConcNet*, each input is assigned a probability (P_{mTBI}) by the *softmax* logistic regression (logit) unit at the output layer and by comparing this probability to the user-chosen classification threshold value, given a score of 0 (*healthy*) or 1 (*concussed*). For *ConcNet*, we set the threshold at 0.5 and a participant with a probability score of greater than this threshold (i.e. $P_{\mathrm{mTBI}} > 0.5$) is classified as concussed. *ConcNet*'s performance after two rounds of Stage 1 blind test is shown in Table 3. The confusion matrix summarizing the results is shown in Figure 2. Apart from *Accuracy*, the model also achieved $Recall = 100\%$, $Precision = 80\%$ and $Specificity = 80\%$. *Recall* (also known as *Sensitivity* or *True Positive Rate*) measures the proportion of actual positives (i.e. concussed) that are correctly identified as such or alternatively, it is a measure of the classifier's ability to identify concussed as concussed. *Precision* is the fraction of true positive in the set classified as positive (i.e. the set of both true and false positives); it is a measure of how much trust one can have in the classifier's positive predictions. *Specificity* (also, True Negative Rate) measures the fraction of participants labelled as healthy that are correctly identified as such. The network correctly identified all of the concussed samples but misidentified three raw90-rsEEG samples contributed by three different healthy participants. In this regard, the classifier worked just as one would want a clinical classifier to, correctly identifying all the concussed subjects (i.e. high recall), and although a small number of healthy participants were misclassified (i.e. a slightly lower precision), it can be argued that this is an acceptable trade-off because *ConcNet* is erring on the side of caution.

As illustrated in Table 3, the blind sample comprised more than one (but not identical) raw90-rsEEG samples from some of the subjects. Each sample consists of a segment of synchronous 90 second-long EEG time series. The motivation for including different segments taken from the same subject is to investigate if these are similarly classified (consistency). The classifier's *Consistency* is defined as the ratio of the distinct number of segment pairs contributed by the same subjects that are similarly classified, to the total number of such pairs in the sample. Our blind sample included a total of 12 such pairs of raw90-rsEEG recordings, each pair derived from the same subject. The LSTM network classified 11 segment pairs identically (and correctly) for a *Consistency* rating of 91.7%.

Stage 2: Optimizing the LSTM-based *ConcNet*

Encouraged by the Stage 1 results, we proceeded to systematically optimize and simplify *ConcNet*'s structure without compromising on performance. The final network architecture, shown in Figure 3, is largely the same as *ConcNet* 1.0 with one key difference. It has one less fully connected layer. Specifically, *ConcNet* 2.0 is made up of (1) a sequential input layer that receives the 64-channel raw90-rsEEG data and passes it to (2) the first LSTM layer comprising 100 memory units. The output is handed off to (3) a dropout layer that randomly discards the results from 30 of the 100 LSTM units (i.e. *DOf*=30%). Dropout is a commonly used regularization technique to prevent overfitting. The output from the Dropout layer is passed to (4) a second LSTM layer with 100 bi-directional memory units, then to (5) another dropout layer with $DOf = 30\%$. The series data thus synthesized by the two LSTM+Dropout layers is fed to two fully connected (FC) layers, (6) the first with 8 nodes and (7) the second with 2 nodes, with each of the nodes utilizing the ReLU (see §4) activation function. Given the small number of nodes in these FC layers, the corresponding dropout layers present in *ConcNet* 1.0 (c.f. Figure 1) were no longer necessary. As noted previously, the binary classification and the assignment of a probability happens in the final *softmax* logit layer.

In addition to being somewhat more streamlined than our first generation version, our final network also formally fared a bit better in that the number of false positives dropped from 3 to 2 and correspondingly, the accuracy increased to 92.6%. We emphasize that this formal improvement in the performance is not statistically significant but it did mean that we had accomplished our stated goal: to streamline the network architecture as much as possible while maintaining performance. Going forward, the network architecture and hyperparameters were held fixed for the duration of the experiment.

Stage 3: Assessing *ConcNet*'s Performance

The third and the final stage involved carefully assessing the performance of *ConcNet* 2.0, including quantifying the uncertainties associated with its performance metrics. We were specifically interested in assessing how well *ConcNet* would generalize as the training, validation and testing datasets are varied. To that end, we drew upon the full dataset of available raw90-rsEEG recordings and used stratified balanced sampling strategy (see §4 for details) to train 100 independent *ConcNet* 2.0 networks; that is, networks that are identical copies, in terms of architecture and hyperparameters, of the final network configuration from Stage 2 above.

Table 4 shows the median values (Q_2), as well as the 25% (Q_1) and 75% (Q_3) quartiles, of the metrics commonly used to gauge a classifier's performance. Notably, the *Accuracy* and *Specificity* remain high, with median values at 0.917 ($Q_1 = 0.833, Q_3 = 1.000$) and 0.900 respectively ($Q_1 = 0.850, Q_3 = 1.000$) respectively. All three *Recall* quartiles are equal to 1, while the mean is also reassuringly high at 0.93.

We also examined the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) for *ConcNet*. The latter is a particularly useful measure because while the values of *Accuracy*, *Specificity* and *Recall* depend on the adopted threshold for classification, AUC does not. An ROC curve plots the True Positive Rate (TPR) vs. the False Positive Rate (FPR) for a range of classification thresholds. In the present case, we vary the probability threshold for being classified as concussed from 0-100%. A low classification threshold generally leads to more healthy participants being classified as concussed. Figure 4 shows the ROC derived using the results of the 100 networks. The ensemble median AUC for *ConcNet* is 0.971 ($Q_1 = 0.963, Q_3 = 0.975$). AUC measures the performance across all classification thresholds: An AUC of 0.5 corresponds to a classifier whose accuracy is no better than a coin toss while an AUC of 1.0 corresponds to a perfect test (i.e. 100% recall and 100% specificity). A value of > 0.96 indicates a classifier with excellent discriminatory ability [69].

To investigate *ConcNet* results further, we examined the full set of predicted scores assigned to each subject. We note that a subject is assigned a score only when a raw90-rsEEG recording from that subject is presented to a network as a test case. Figure 5 shows the median of the predicted scores as well as the corresponding 25th and 75th percentiles. For clarity, the results are divided across two panels, with the left panel showing the results for the 27 concussed participants and the right panel summarizing the results for the 35 healthy participants. In each panel, the blue points represent those subjects who were correctly classified. The total number of points may seem to be fewer than expected, given the number of participants in each category; however, this is not the case. Many of the blue points in the *true positive* (TP) and *true negative* (TN) quadrants have very similar scores and are stacked on top of each other. The small error bars associated with these points indicate that the corresponding subjects were assigned similar predicted scores regardless of which network from the ensemble was used or which of their eight raw90-rsEEG recordings was analyzed.

As for the misclassifications, of the 27 concussed subjects, we found that one was nearly always (i.e. consistently) misclassified while the EEG samples of another subject tended to be misclassified more often than not (and therefore, the median fell below our threshold). These two cases appear as red points in the false negative (FN) quadrant of the left panel in Figure 5. As for the healthy subjects, we found that two were misclassified most of the time and one was more frequently misclassified than not. These are shown as red points in the right panel of Figure 5. The scatter in the predicted scores for all of the misclassified cases is generally large.

Analyzing the misclassifications further, we noted that they were *not random occurrences* in the sense that the same five individuals tended to be misclassified regardless of which of the subjects' raw90-rsEEG recordings was used or the network in the ensemble it was processed through. To further illustrate this, in Figure 6, we show the scores as stacked bar charts for all the concussed and healthy samples. For each bar, the blue and red areas represent the fractional number of times that the subject under consideration was respectively correctly or incorrectly classified. Any bar whose red portion occupies >50% of the column corresponds to a sample which was more frequently misclassified. The two concussed and the three healthy "problematic" subjects readily stand out.

Figure 6 also shows a few additional columns with small red segments. These segments represent a small number of networks that misclassified the subjects in question. We found that these failures were also largely due to the 'problematic" samples highlighted above. Since the partitioning of the full dataset into training, validation and testing for each of the 100 networks is done at random, every so often the EEG data from one or more "problematic" cases appeared in the training/validation subsets for some of the networks. This degraded the performance of the affected networks. The moderate value for the *Precision* and *Markedness*, and the spread in their values, are primarily due to the five "problematic" individuals.

The systematic nature of the problem makes us suspect the issue is not with the classifier but with the data. We will return to this matter in §3 and discuss possible reasons for the misclassifications. As a test, we removed the five systematically misclassified individuals from our dataset, and retrained and tested a single network. Not surprisingly, the overall accuracy rose to 98%, an improvement of nearly 5% from our previous Stage 2 results.

3 Discussion

One of major current gaps in concussion management is determining when an athlete has recovered from the initial injury after a sports related concussion. Health care providers often have to make difficult decisions regarding when an athlete is fit to safely return to sport/activity based on self-reporting of symptoms and their severity, and clinical testing. Self-reports, however, are highly subjective and also influenced by socio-cultural factors [70, 71]. Moreover, emerging literature is showing that neurophysiological changes persist well after clinical recovery [15–23], suggesting that athletes may be returning to activity before their brains have fully recovered and may therefore be at increased risk of additional brain injury [66]. In fact, a history of prior concussions is one of the strongest predictors of future concussions [71–73]. Specifically, high school athletes with prior concussions have $>2\times$ higher risk of suffering subsequent concussions [72, 73] and with each successive injury, have a growing risk for sustaining long term neurological and mental health complications. Brain based measures of neurophysiological recovery (based on fMRI, DTI and EEG) have so far only been able to detect differences in the brain at a group level. The results presented here demonstrate that *ConcNet*, our LSTM-based deep learning network for concussion classification, is a promising engine for an objective, brain-based platform for identifying acute post-concussed individuals using only their resting state EEG data. Admittedly, the EEG data used to train and test the network was acquired from only male adolescent athletes (ages: 10–18), and this is a limitation of the present study. Nonetheless, *ConcNet*'s excellent performance serves as a proof of concept demonstrating that it is possible to train a deep learning network to recognize acute post-concussed individuals, with a high degree of accuracy, using only their resting state EEG data.

Encouraged by our current results, we were actively expanding our database to include female athletes as well as younger and older subjects until interrupted by the Covid-19 pandemic. Several studies find that female athletes have an inherent elevated risk for concussion compared to their male counterparts, differ in the types (and severity) of symptoms experienced, and take longer to become symptom free [70, 74–76]. However, in the absence of an objective tool for confirming a concussion, it is difficult to establish whether this is a genuine difference or the result of a bias related to how males and females report concussions and concussion symptoms [70, 74]. *ConcNet*-like platform offers an opportunity to expand the current understanding of how sex, among other factors, affects injury epidemiology and outcomes.

Of the 27 concussed subjects considered in this study, 25 subjects were consistently classified correctly while 2 subjects were tended to be misclassified as healthy. We have already noted that adjusting the classification threshold slightly would have reduced the number of misclassifications to 1; however, at the moment, we do not have a clear basis for doing so. Moreover, we are intrigued by the systematic differences in the probabilistic scores assigned to the 2 individuals in questions, compared to the scores assigned to the other 25 subjects. Three possible explanations stand out:

Firstly, we cannot outrightly exclude the possibility that the failures reflect *ConcNet*'s small but finite probability of misclassifying. However, since the misclassifications are not randomly distributed across the subjects and that most of the misclassifications are associated with the raw90-rsEEG datasets contributed by the five problematic subjects, we find it hard to accept that these are products of random errors.

Secondly, the two subjects may have been misdiagnosed at the outset. Our criteria for identifying subjects with concussion were (i) documentation of the date and time of a direct blow by the team coach, and (ii) a diagnosis by an experienced physician. An error would have had to have happened at both stages and while we cannot exclude the possibility, we think it is also unlikely.

The third possibility is that either the subjects were well on their way to recovering from their injury at the time their brains were scanned or that their brains were only minimally affected by the injury. In reviewing the two misclassified concussed subjects' other available data, we found that they were indistinguishable from the other concussed subjects except for the fact that they were both scanned at 1 month post injury, and comprised 2 of the 3 individuals in our sample at the edge of our time window, which lends itself to the "on the way to recovery" conjecture. All the other subjects were scanned at 1 week post injury. In light of this, perhaps a more pertinent question is not why the two subjects were misclassified but rather, why was the third subject in the 1-month post injury cohort not misclassified. We speculate that the "correctly" classified member of the cohort may have suffered a more pronounced injury or may be experiencing a more prolonged recovery. Neuroimaging studies find that the brain's response to an mTBI, as suggested by a significant spread in the measures derived from neuroimaging probes of the brain's structure and function, varies considerably from subject to subject both at the acute phase as well as in the brain's recovery trajectory [77, 78]. In fact, this same explanation may also explain the three subjects in the healthy group that were misclassified as concussed. Longitudinal neuroimaging studies show that in some individuals, mTBI induced changes in the brain can persist for several months and even past a year [15–23]. These persistent alterations are seen in EEG, fMRI and

DTI data. We suspect that our misclassified "healthy" may be one of these individuals who may have suffered one or more concussions previously and are exhibiting persistent changes. We did not request our subjects' concussion history since there is considerable evidence showing that self-reported histories are prone to errors [70, 71]. Ultimately, the resolution of this puzzle requires growing our relatively shallow dataset with a larger sample that encompasses the full range of diversity present during the acute post concussion phase.

Another fascinating question emerging from our study is: how is *ConcNet* able to leverage raw, highly nonstationary, EEG time series data to distinguish between concussed and healthy adolescent subjects? Given that an LSTM-based network's main advantage is that it can identify temporal correlations and causal relations within a sequential data stream, we suspect that *ConcNet* may be picking up alterations in the temporal dynamics of brain due to mTBI. There is evidence showing that concussions do just this: By applying the sliding window analysis to resting state fMRI data, Muller and Virji-Babul [79] identified three distinct brain states and showed that while healthy adolescents switched dynamically between the three brain states and spend approximately the same amount of time in each brain state, concussed adolescents spent the majority of time in only one brain state. The possibility of using neural networks, like our LSTM-based *ConcNet*, to probe brain dynamics is an exciting new frontier that we are starting to explore.

To summarize, in this paper, we introduced *ConcNet*, a high-performing deep learning Long Short Term Memory (LSTM)-based recurrent neural network that is able to distinguish between healthy and acute post-concussed adolescent athletes using only a short (i.e. 90 seconds long) sample of resting state EEG data as input. The athletes were neither required to perform a specific task nor subjected to a stimulus during data collection, and the acquired EEG data was neither filtered, cleaned of artefacts, nor subjected to explicit feature extraction. The classifier's final network architecture is shown in Figure 3. *ConcNet* was trained and tested using a total of 216 EEG datasets, obtained from 27 male, adolescent athletes with sports related concussion within 1 month of injury, and 280 datasets from 35 age and sex matched healthy athletes. Using a stratified balanced sampling strategy, we trained an ensemble of 100 independent networks in order to carry out a detailed assessment of the classifier. *ConcNet* achieved median values of 0.917, 0.900 and 1 on *Accuracy*, *Specificity* and *Recall* respectively (see Table 4 for more details). The magnitudes of all three are reassuringly high. These metrics do, however, depend on the adopted threshold for classification. Presently, we use a threshold of 0.5 to distinguish between concussed and healthy, with participants who are assigned a probability score greater than the threshold (i.e. $P_{\mathrm{mTBI}} > 0.5$) being classified as concussed. Adopting a different threshold, like $p_{\mathrm{mTBI}} > 0.4$, increases the metric values and reduces the associated scatter. We also evaluated *ConcNet*'s Area Under the Curve, a performance metric that does not depend on the adopted threshold for classification. The resulting ensemble-median value of 0.971 confirms that in the context of the population group that *ConcNet* was trained to classify, the classifier does very well at distinguishing between the concussed and the healthy subjects. This is the first instance of a high-performing, objective, brain-based, platform for automatic classification of concussion at an individual level that relies only on easy to acquire resting state EEG data. It represents a key step towards addressing one of the fundamental challenges in concussion management: the dependence on highly subjective self-reporting of symptoms when determining whether an athlete has suffered a concussion or has recovered from one.

4 Methods

Participants

The EEG recordings used in this study were collected from a total of 62 male adolescent athletes. Of these, 27 individuals [mean age = 13.5 years; standard deviation (SD): ±2.6 years], had suffered a sport-related concussion, were within one month of injury, and met the concussion diagnostic criteria consistent with the Berlin consensus statement [66]. The other 35 were healthy, active, age- and sex-matched cohort of athletes [mean age: 14.7 years; SD: ±2.1 years]. We chose the time window of one month based on the 2017 Berlin Consensus Statement's expected time for clinical recovery after injury in children and adolescents [80]. Here, concussion is defined as a traumatic brain injury caused either by a direct blow to the head, face, neck, or elsewhere on the body that leads to an impulsive force transmitted to the head, resulting in changes in one or more of the following clinical domains: (i) physiological (e.g. neck pain, balance problems, headache, fatigue), (ii) cognitive (e.g. difficulty with attention, feeling in a "fog"), (iii) emotional (e.g. irritability, sadness) and (iv) behavioural (i.e. sleep/wake disturbances). In the case of the concussed participants, the date and time of the direct blow was documented by the team coach as per the consensus statement, and the team physician or a physician with expertise in concussion made the diagnosis of concussion again as per the Berlin consensus statement. Individuals with focal neurologic deficits, pathology and/or those on prescription medications for neurological or psychiatric conditions were excluded from this study. Our study was approved by

the University of British Columbia Clinical Research Ethics Board (Approval number: H17-02973). All participants provided assent and the adolescents' parents gave written informed consent for their children's participation under the approval of the ethics committee of the University of British Columbia and in accordance with the Helsinki declaration.

EEG recordings

Data collection

Five minutes of resting state EEG data were collected while participants had their eyes closed. We used a 64-channel HydroCel Geodesic Sensor Net (EGI, Eugene, OR) connected to a Net Amps 300 amplifier [81]. The signals were referenced to the vertex (Cz) and recorded at a sampling rate of 250 Hz. The scalp electrode impedance values were typically less than 50 $k\Omega$. To eliminate any transients at the start and the end of a data collection session (due, for example, to subjects taking a few seconds to settle down, etc.), 1000 data points were removed from the beginning and the end of each time series; this corresponds to removing data with a total duration of 8 seconds. We did not filter or otherwise clean the data to remove artefacts due to line noise, eye blinks and motion as is typically done (c.f. Porter et al. [82], Rotem-Kohavi et al. [83, 84], Munia et al. [19]).

Data segmentation and augmentation

We segmented the 64-channel EEG data into three 90 seconds long consecutive sets of raw90-rsEEG. We further augmented our data compilation by extracting, from each subject's five minutes long 64-channel data stream, five additional 90 seconds long, synchronous segments with randomly chosen starting points. Thus, in total, we extracted 8 sets of raw90-rsEEG recordings per participant. EEG signals are well known to be strongly non-stationary on timescales greater than $\sim$0.25 s, exhibiting significant variations, and even shifts, in the statistical properties of the signal over time [61, 62]. This behaviour poses a significant challenge for extracting stable features from the signal as well as for designing reliable brain-computer interface systems within the context of real environments [85–88]. In the present study, we leverage this behaviour to generate multiple, effectively independent, realizations of the raw90-rsEEG datasets from each of the participants.

Our choice to use 90-second long segments was the result of a compromise between wanting to use a short EEG sample on one hand, and ensuring that the sample is long enough to capture very low frequency brain activity, i.e. oscillations with frequencies down to $\sim 0.01 - 0.02$ Hz (also known as infraslow oscillations or ISO). The latter was motivated by two considerations: (a) EEG studies find that, compared to healthy controls, concussed athletes exhibit altered activity in the delta band ($0.5 - 4$ Hz) [19, 33, 89], with increasing divergences towards to lowest frequencies. (b) Among adolescents, one of the resting state networks strongly impacted by concussion is the default mode network (DMN) [20, 48]. This network, as well as several other resting state networks, support very low frequency, large amplitude oscillations [90]. The power in these frequency modes has been shown to scale approximately inversely with frequency [91]; this $1/f$ background has a profound impact on neuron dynamics and communications [92].

Behavioural Data

When the EEG data were collected, the number of symptoms and symptom severity of each of the concussed subjects were also evaluated using either the Child Sports Concussion Assessment Tool 3 (Child SCAT3), if the injured athlete was younger than 13 years of age (https://bjsm.bmj.com/content/bjsports/47/5/263.full.pdf), or using the Sports Concussion Assessment Tool 3 (SCAT3), if 13 years or older (https://bjsm.bmj.com/content/bjsports/47/5/259.full.pdf). Both SCAT3 and Child SCAT3 are standardized concussion and concussion symptom assessment tools. SCAT3 lists 22 symptoms that can be scored from 0 (none) to 6 (severe) while Child SCAT3 lists 20 symptoms with scores ranging from 0 (none) to 3 (often). The overall symptom severity score of the injured athletes is the sum of the individual symptom ratings. SCAT3 allows for a maximum score of 132 while the highest possible Child SCAT3 score is 60.

Network Architecture

As noted in §2, we had previously experimented with several popular machine learning classification algorithms, including the Support Vector Machine (SVM) algorithm [41] and the simplest type of multilayer perceptron (MLP) neural network [45]. The

latter was a three-layer network consisting of an input layer that receives the data, a single hidden layer that processes the data, and an output layer that classifies based on the results from the hidden layer. The input for both attempts consisted of a set of statistical measures (hereafter, *features*) extracted from each individual's set of 64 resting state EEG time series data. Neither of these two attempts yielded classification accuracy above 65%, a finding that was not a complete surprise. It is well known that measures extracted from different segments of time duration Δt, extracted from even a single non-stationary EEG time series collected at one sitting, can vary significantly [61, 62, 87]. However, prior to extracting the features, we had filtered the data [82–84] to remove signal drift, an approach that is commonly adopted to mitigate weak non-stationary behaviour. This works well enough when one is interested in *population-level* differences between concussed and healthy groups [24, 28, 47, 48]. However, the distribution functions of features for the healthy and concussed individuals overlap, which hampers attempts to use the features to assess *individual* subjects. The problem is further compounded by the fact that EEG time series is strongly non-stationary and simple procedures, like filtering, do not completely remove the non-stationarity of features. This tendency is the source of the difficulties faced by researchers pursuing reliable brain-computer interfaces [86, 88]. In light of this, and in the interest of ensuring that the data processing pipeline is as simple as possible, we felt it best to use the observed (raw) EEG signals and rely on a deep network to learn the signals.

As for the choice of the deep learning network architecture, we took our cue from the brain-computer interface (BCI) literature and adopted a recurrent neural network architecture comprising two layers of bi-directional Long Short Term Memory (LSTM) [63, 64] units. In the context of EEG-based BCIs, Recurrent Neural Networks (RNN) — and in particular, a 2-layer bi-directional LSTM network — have been found to be better at accommodating non-stationary behaviour of EEG time series and consequently, yield higher accuracy in classification tasks [88]. The key feature is the inclusion of *memory units*. As a sequence is processed, the memory of the results associated with prior time points persists in these units and informs the processing of current and future inputs. This facility for capturing long-term dependencies is the reason why LSTMs are better able to learn time-series data due to complex nonlinear dynamics. Our network employs bi-directional memory units [93]. These integrate information from both past and future time steps to compute the activation signals as well as the corrections that are applied to the weights during the back propagation step. Bi-directionality enhances network stability, speeds up training, and results in improved performance in comparision to uni-directional networks [94]. Stability is especially important when training with sparse datasets, as in our case. This coupled with the fact that an LSTM network is better able to handle sequential data made this architecture an obvious first choice.

Apart from the LSTM layers, our network also incorporates fully connected (FC) layers. Each node on a FC layer receives input from all of the nodes on the preceding layer. These inputs are multiplied by the weights and then summed along with a constant term. The result is then mapped to an output using a activation function. We have opted to use the *ReLU* (Rectified Linear Unit) activation function in our network because it has a much greater dynamic range than offered by other commonly used activation functions and therefore enables efficient learning in deep learning environments, and because its non-linear character facilitates the learning of complex dependencies within the data.

Network Training and Validation

Stage 1: For the *blinded, one-shot* protocol used in this stage (see §2), *ConcNet* was trained using only a subset of the available dataset: that is, all raw90-rsEEG segments harvested from 44 (25 healthy and 19 concussed) of the total of 62 subjects; the remaining were reserved for the blind test. We divided this training data into two groups, the training and the validation sets, using a modified 80:20 prescription, with the training set comprising all raw90-rsEEG segments from 15 randomly chosen concussed subjects and an equal number of randomly chosen healthy subjects. The balance of the data were used for validation. We did not allow for any overlap between the two sets; all the raw90-rsEEG segments from any one subject either belonged to the training or the validation set, but not both. Since the data segmentation described in §4 provided eight data segments per subject, the training:validation split resulted in 120 segments each from concussed and control subjects for training, and 32 concussed and 80 control segments for validation.

To train the network, we adopted the *mini-batch* approach: The training proceeded in steps, each using 20 randomly selected, raw90-rsEEG segments, which were processed through *ConcNet*. The difference between the labeled classification and that predicted by *ConcNet*, summed over the 20 samples in the mini-batch was used to define the *loss function*; the smaller the loss, the better the performance. During the subsequent *backpropagation* step, all the weights in the network were updated proportional to the gradient of the loss function with respect to each weight. Validation of the training process was done after every five mini-batches, providing a metric to assess the progress of the training. These steps were continued till all the training

samples had been exhausted. This marked the end of an *epoch*. At this point, the training data was reshuffled (i.e., another set of 15 concussed and 15 healthy subjects were drawn from the available data) and the mini-batch training process was repeated. The cycles were repeated for 25 epochs, by which time the loss function had levelled off, and further training resulted in little or no improvement of performance. At this point, the loss curves from training and validation were used to tune the hyperparameters. For *ConcNet*, these hyperparameters included the number of LSTM layers, the number of LSTM units per layer, the number of fully connected layers and the number of nodes in each, the regularization scheme and dropout rate used, the choice of optimizer, and the learning rate. For *ConcNet*, we adopted *Adam*, an adaptive learning rate optimization algorithm introduced by Kingma & Ba [95], which has been shown to be matched for non-stationary objectives, and works well even with noisy and sparse datasets.

We repeated the process described, including the continued tuning of the hyperparameters via an iterative search, while monitoring both the training and the validation loss curves. Training was stopped when the loss curves started to decrease monotonically and crossed a preset threshold ($< 10^{-4}$ in our case). We did not pursue further explorations of nearly degenerate network configurations.

The full set of training iterations was carried out on a single GPU assisted cluster hosted by *Compute Canada* and took less than 30 minutes. An attempt to train the network on desktop computer equipped with a single GPU took about the same time. Given the file size and computation requirements of the training and validation datasets, the training stage needs at least one GPU, either in a compute cluster or even in a desktop computer. However, since the test data sample size is much smaller, testing the trained network could be done even on a laptop without GPU assistance.

With the hyperparameters satisfactorily tuned, the present Stage 1 *ConcNet* was tested, as noted in Section 2, via *"blinded", one-shot* protocol, where the test data were provided to the software team without labels. The test segments were sourced from subjects that *ConcNet* had never been exposed to previously nor was it taught the data after the fact. The network results were recorded, and then the labels were unmasked for analysis.

Stage 2: During this stage, our aim was to optimise and simplify the network architecture as much as possible while maintaining performance. All of the data was made available and used to construct three groups: a training data set comprising the eight raw90-rsEEG segments from each of 21 randomly selected concussed and 21 randomly healthy participants, for a total of 336 samples; a validation set consisting of the eight raw90-rsEEG from each of the 3 concussed and 3 healthy participants, again randomly selected from the balance of the participants (48 samples in total); and a testing set made up of the rest. This is roughly an 80:10:10 division within the context of a stratified balanced sampling strategy, a commonly used approach for mitigating potential biases that may arise when the number of members in the groups under consideration are dissimilar. In the present instance, we require that the number of concussed and healthy participants in the training and validation sets are matched. The training and validation procedure was similar to that of Stage 1. Once training was completed, the performance was assessed using the test dataset. When faced with degenerate network configurations, we preferred those with fewer layers.

Stage 3: To further assess *ConcNet*'s performance as well as quantify the uncertainties in the performance metrics, we trained 100 independent networks with identical architecture and hyperparameters to final *ConcNet* configuration established at the end of Stage 2. Each network was trained using the same procedure adopted for Stage 2. Each of the 100 networks was trained using a separate, randomly selected realization of training, validation and testing sets. For this reason, the internal weights of each of the networks at the end of training, and hence their individual performance, are expected to vary from network to network, thus yielding the uncertainties associated with each of the performance metrics. Over the course of Stage 3, we maintained a detailed log of which concussed and healthy participants were being used to train, validate and test each network, as well as the resulting classification scores and the predicted labels. Here, the classification scores refer to the probabilities output from the final softmax layer. We also calculated the performance metrics for each of the networks. This type of assessment is critical for determining how well *ConcNet* would generalize to a larger, more heterogeneous datasets.

References

1. Langer, L., Levy, C. & Bayley, M. Increasing incidence of concussion: True epidemic or better recognition? *The J. Head Trauma Rehabil.* **35** (2020).
2. Langlois, J. A., Rutland-Brown, W. & Wald, M. M. The epidemiology and impact of traumatic brain injury: A brief overview. *The J. Head Trauma Rehabil.* **21** (2006).

3. Solomon, G. S. Sports neuropsychology: Assessment and management of traumatic brain injury. *The J. Neuropsychiatry Clin. Neurosci.* **20**, 504–504, DOI: 10.1176/jnp.2008.20.4.504 (2008). https://doi.org/10.1176/jnp.2008.20.4.504.

4. Concussion Ontario, Characterizing access to concussion care in Ontario. Published by Concussion Ontario/Ontario Neurotrauma Foundation (2017). "http://concussionsontario.org/access-to-care/concussion-data/survey-of-concussionmtbi-care-in-brain-injury-clinics-and-services-in-ontario/".

5. Daneshvar, D. H., Nowinski, C. J., McKee, A. C. & Cantu, R. C. The epidemiology of sport-related concussion. *Clin. Sports Medicine* **30**, 1–17, DOI: 10.1016/j.csm.2010.08.006 (2011).

6. Cassidy, J. *et al.* Incidence, risk factors and prevention of mild traumatic brain injury: results of the who collaborating centre task force on mild traumatic brain injury. *J. Rehabil. Medicine* 28–60 (2004).

7. Barlow, K. M. *et al.* Epidemiology of postconcussion syndrome in pediatric mild traumatic brain injury. *Pediatrics* **126**, e374–e381 (2010).

8. Dalecki, M., Albines, D., Macpherson, A. & Sergio, L. E. Prolonged cognitive–motor impairments in children and adolescents with a history of concussion. *Concussion* **1**, CNC14 (2016).

9. Mayer, A. R. *et al.* Advanced biomarkers of pediatric mild traumatic brain injury: Progress and perils. *Neurosci. & Biobehav. Rev.* **94**, 149–165 (2018).

10. Moore, R. D., Kay, J. J. & Ellemberg, D. The long-term outcomes of sport-related concussion in pediatric populations. *Int. J. Psychophysiol.* **132**, 14–24, DOI: 10.1016/j.ijpsycho.2018.04.003 (2018).

11. Toledo, E. *et al.* The young brain and concussion: imaging as a biomarker for diagnosis and prognosis. *Neurosci. & Biobehav. Rev.* **36**, 1510–1531, DOI: 10.1016/j.neubiorev.2012.03.007 (2012).

12. Keith, O. Y. & Taylor, H. G. Neurobehavioural outcomes of mild head injury in children and adolescents. *Pediatr. Rehabil.* **8**, 5–16 (2005).

13. Saatman, K. E. *et al.* Classification of traumatic brain injury for targeted therapies. *J. Neurotrauma* **25**, 719–738 (2008).

14. Mayer, A. R. *et al.* Diffusion abnormalities in pediatric mild traumatic brain injury. *J. Neurosci.* **32**, 17961–17969 (2012).

15. Slobounov, S., Sebastianelli, W. & Hallett, M. Residual brain dysfunction observed one year post-mild traumatic brain injury: Combined EEG and balance study. *Clin. Neurophysiol.* **123**, 1755 – 1761, DOI: https://doi.org/10.1016/j.clinph.2011.12.022 (2012).

16. Westfall, D. R. *et al.* Increased brain activation during working memory processing after pediatric mild traumatic brain injury (mtbi). *J. Pediatr. Rehabil. Medicine* **8**, 297–308 (2015).

17. Borich, M., Babul, A.-N., Yuan, P. H., Boyd, L. & Virji-Babul, N. Alterations in resting-state brain networks in concussed adolescent athletes. *J. Neurotrauma* **32**, 265–271 (2015).

18. Munia, T. T. K. *et al.* Preliminary results of residual deficits observed in athletes with concussion history: Combined EEG and cognitive study. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 41–44 (2016).

19. Munia, T. T., Haider, A., Schneider, C., Romanick, M. & Fazel-Rezai, R. A novel EEG based spectral analysis of persistent brain function alteration in athletes with concussion history. *Sci. Reports* **7**, 1–13 (2017).

20. Manning, K. Y. *et al.* Multiparametric MRI changes persist beyond recovery in concussed adolescent hockey players. *Neurology* **89**, 2157–2166 (2017).

21. Chung, A. W., Mannix, R., Feldman, H. A., Grant, P. E. & Im, K. Longitudinal structural connectomic and rich-club analysis in adolescent mTBI reveals persistent, distributed brain alterations acutely through to one year post-injury. *Sci. Reports* **9**, 1–11 (2019).

22. Churchill, N. W., Hutchison, M. G., Graham, S. J. & Schweizer, T. A. Mapping brain recovery after concussion. *Neurology* **93**, e1980–e1992, DOI: 10.1212/WNL.0000000000008523 (2019). https://n.neurology.org/content/93/21/e1980.full.pdf.

23. Manning, K. Y. *et al.* Linked MRI signatures of the brain's acute and persistent response to concussion in female varsity rugby players. *NeuroImage: Clin.* **21**, 101627 (2019).

24. Hristopulos, D. T., Babul, A., Babul, S., Brucar, L. R. & Virji-Babul, N. Disrupted information flow in resting-state in adolescents with sports related concussion. *Front. Hum. Neurosci.* **13**, 419, DOI: 10.3389/fnhum.2019.00419 (2019).

25. Bigler, E. D. Traumatic brain injury, neuroimaging, and neurodegeneration. *Front. Hum. Neurosci.* **7**, 395 (2013).

26. Gupta, R. & Sen, N. Traumatic brain injury: a risk factor for neurodegenerative diseases. *Rev. Neurosci.* **27**, 93–100 (2016).

27. Borich, M., Makan, N., Boyd, L. & Virji-Babul, N. Combining whole-brain voxel-wise analysis with in vivo tractography of diffusion behavior after sports-related concussion in adolescents: a preliminary report. *J. Neurotrauma* **30**, 1243–1249 (2013).

28. Virji-Babul, N. *et al.* Diffusion tensor imaging of sports-related concussion in adolescents. *Pediatr. Neurol.* **48**, 24–29, DOI: 10.1016/j.pediatrneurol.2012.09.005 (2013).

29. Yallampalli, R. *et al.* Acute white matter differences in the fornix following mild traumatic brain injury using diffusion tensor imaging. *J. Neuroimaging* **23**, 224–227, DOI: 10.1111/j.1552-6569.2010.00537.x (2013).

30. Yuan, W., Wade, S. L. & Babcock, L. Structural connectivity abnormality in children with acute mild traumatic brain injury using graph theoretical analysis. *Hum. Brain Mapp.* **36**, 779–792, DOI: 10.1002/hbm.22664 (2015).

31. Murdaugh, D. L. *et al.* Longitudinal changes in resting state connectivity and white matter integrity in adolescents with sports-related concussion. *J. Int. Neuropsychol. Soc.* **24**, 781–792, DOI: 10.1017/S1355617718000413 (2018).

32. Wu, T. *et al.* A preliminary report of cerebral white matter microstructural changes associated with adolescent sports concussion acutely and subacutely using diffusion tensor imaging. *Brain Imaging Behav.* **12**, 962–973, DOI: 10.1007/s11682-017-9752-5 (2018).

33. Balkan, O., Virji-Babul, N., Miyakoshi, M., Makeig, S. & Garudadri, H. Source-domain spectral eeg analysis of sports-related concussion via measure projection analysis. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4053–4056, DOI: 10.1109/EMBC.2015.7319284 (2015).

34. Newsome, M. R. *et al.* Functional connectivity is altered in concussed adolescent athletes despite medical clearance to return to play: a preliminary report. *Front. Neurol.* **7**, 116, DOI: 10.3389/fneur.2016.00116 (2016).

35. Manning, K. Y. *et al.* Multiparametric MRI changes persist beyond recovery in concussed adolescent hockey players. *Neurology* **89**, 2157–2166, DOI: 10.1212/WNL.0000000000004669 (2017).

36. RSNA statement on traumatic brain injury (TBI) imaging (Updated: 9/13/2018). Published by The Radiological Society of North America (RSNA) (2018). "https://www.rsna.org/uploadedFiles/RSNA/Content/Role_based_pages/Media/RSNA-TBI-Position-Statement.pdf".

37. Davatzikos, C. Machine learning in neuroimaging: progress and challenges. *NeuroImage* **197**, 652 (2019).

38. Thomas, A. W., Heekeren, H. R., Müller, K.-R. & Samek, W. Analyzing neuroimaging data through recurrent deep learning models. *Front. Neurosci.* **13**, 1321 (2019).

39. Erus, G., Habes, M. & Davatzikos, C. Machine learning based imaging biomarkers in large scale population studies: A neuroimaging perspective. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, 379–399 (Elsevier, 2020).

40. Vakorin, V. A. *et al.* Detecting mild traumatic brain injury using resting state magnetoencephalographic connectivity. *PLoS Comput. Biol.* **12** (2016).

41. Garg, S., Yeung, A., Garudadri, H. & Virji-Babul, N. A comparison of EEG power spectral and wavelet features in concussed cohorts using support vector machine. presentation, 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society on "Biomedical Engineering: A Bridge to Improve the Quality of Health Care and the Quality of Life", Milan, Italy, August 25-29 (2015). "http://150.162.46.34:8080/embc-2015/papers/20563441.pdf".

42. Mitra, J. *et al.* Statistical machine learning to identify traumatic brain injury (TBI) from structural disconnections of white matter networks. *NeuroImage* **129**, 247–259 (2016).

43. Vergara, V. M., Mayer, A. R., Damaraju, E., Kiehl, K. A. & Calhoun, V. Detection of mild traumatic brain injury by machine learning classification using resting state functional network connectivity and fractional anisotropy. *J. Neurotrauma* **34**, 1045–1053 (2017).

44. Cao, C., Tutwiler, R. L. & Slobounov, S. Automatic classification of athletes with residual functional deficits following concussion by means of EEG signal using support vector machine. *IEEE Transactions on Neural Syst. Rehabil. Eng.* **16**, 327–335, DOI: 10.1109/TNSRE.2008.918422 (2008).

45. Keshavarz-Hedayati, B., Dimopoulos, N., Babul, A., Yeung, A. & Virji-Babul, N. Neural network-based classification of concussed and control groups from eeg. poster presentation, 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society on "Biomedical Engineering: A Bridge to Improve the Quality of Health

Care and the Quality of Life", Milan, Italy, August 25-29 (2015). See also Keshavarz-Hedayati, Babak (2019), Studies on Applications of Neural Networks in Modeling Sparse Datasets and in the Analysis of Dynamics of CA3 in Hippocampus (Doctoral dissertation, University of Victoria, Victoria, British Columbia, Canada), Chapter 2, Section 2.8.2, "http://venus.library.uvic.ca/handle/1828/10742?show=full".

46. Boshra, R., Ruiter, K. I., DeMatteo, C., Reilly, J. P. & Connolly, J. F. neurophysiological correlates of concussion: Deep learning for clinical assessment. *Sci. Reports* **9**, 1–10 (2019).

47. Borich, M., Makan, N., Boyd, L. & Virji-Babul, N. Combining whole-brain voxel-wise analysis with in vivo tractography of diffusion behavior after sports-related concussion in adolescents: a preliminary report. *J. Neurotrauma* **30**, 1243–1249, DOI: 10.1089/neu.2013.3269 (2013).

48. Borich, M., Babul, A.-N., Yuan, P. H., Boyd, L. & Virji-Babul, N. Alterations in resting-state brain networks in concussed adolescent athletes. *J. Neurotrauma* **32**, 265–271, DOI: 10.1089/neu.2013.3269 (2015).

49. Daly, I. *et al.* What does clean eeg look like? In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3963–3966 (IEEE, 2012).

50. Islam, M. K., Rastegarnia, A. & Yang, Z. Methods for artifact detection and removal from scalp EEG: A review. *Neurophysiol. Clinique/Clinical Neurophysiol.* **46**, 287–305 (2016).

51. Jiang, X., Bian, G.-B. & Tian, Z. Removal of artifacts from EEG signals: a review. *Sensors* **19**, 987 (2019).

52. Bono, V., Jamal, W., Das, S. & Maharatna, K. Artifact reduction in multichannel pervasive EEG using hybrid WPT-ICA and WPT-EMD signal decomposition techniques. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5864–5868 (IEEE, 2014).

53. Widmann, A. & Schröger, E. Filter effects and filter artifacts in the analysis of electrophysiological data. *Front. Psychol.* **3**, 233 (2012).

54. Meisler, S. L., Kahana, M. J. & Ezzyat, Y. Does data cleaning improve brain state classification? *J. Neurosci. Methods* **328**, 108421 (2019).

55. Munson, M. A. & Caruana, R. On feature selection, bias-variance, and bagging. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 144–159 (Springer, 2009).

56. Krawczuk, J. & Łukaszuk, T. The feature selection bias problem in relation to high-dimensional gene data. *Artif. Intell. Medicine* **66**, 63–71 (2016).

57. Neumann, U. *et al.* Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Min.* **9**, 36 (2016).

58. Sahiner, B., Chan, H.-P., Petrick, N., Wagner, R. F. & Hadjiiski, L. Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size. *Med. Phys.* **27**, 1509–1522, DOI: 10.1118/1.599017 (2000). https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.599017.

59. Smialowski, P., Frishman, D. & Kramer, S. Pitfalls of supervised feature selection. *Bioinformatics* **26**, 440–443, DOI: 10.1093/bioinformatics/btp621 (2009). https://academic.oup.com/bioinformatics/article-pdf/26/3/440/16896706/btp621.pdf.

60. Krawczuk, J. & Łukaszuk, T. The feature selection bias problem in relation to high-dimensional gene data. *Artif. Intell. Medicine* **66**, 63–71 (2016).

61. Kaplan, A. Y., Fingelkurts, A. A., Fingelkurts, A. A., Borisov, S. V. & Darkhovsky, B. S. Nonstationary nature of the brain activity as revealed by eeg/meg: Methodological, practical and conceptual challenges. *Signal Process.* **85**, 2190–2212, DOI: https://doi.org/10.1016/j.sigpro.2005.07.010 (2005).

62. Klonowski, W. Everything you wanted to ask about EEG but were afraid to get the right answer. *Nonlinear Biomed. Phys.* **3**, 2 (2009).

63. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780, DOI: 10.1162/neco.1997.9.8.1735 (1997). https://doi.org/10.1162/neco.1997.9.8.1735.

64. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *arXiv e-prints* arXiv:1808.03314 (2018). 1808.03314.

65. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).

66. McCrory, P. *et al.* Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in Berlin, October 2016. *Br. J. Sports Medicine* **51**, 838–847, DOI: 10.1136/bjsports-2017-097699 (2017).

67. Kotsiantis, S. B., Zaharakis, I. & Pintelas, P. Supervised machine learning: A review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **160**, 3–24 (2007).

68. Singh, A., Thakur, N. & Sharma, A. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315 (Ieee, 2016).

69. Fan, J., Upadhye, S. & Worster, A. Understanding receiver operating characteristic (ROC) curves. *Can. J. Emerg. Medicine* **8**, 19–20 (2006).

70. Mollayeva, T., El-Khechen-Richandi, G. & Colantonio, A. Sex & gender considerations in concussion research. *Concussion* **3**, CNC51 (2018).

71. Schmidt, J. D. *et al.* Age at first concussion influences the number of subsequent concussions. *Pediatr. Neurol.* **81**, 19–24 (2018).

72. Schulz, M. R. *et al.* Incidence and risk factors for concussion in high school athletes, north carolina, 1996–1999. *Am. J. Epidemiol.* **160**, 937–944 (2004).

73. Van Pelt, K. L. *et al.* A cohort study to identify and evaluate concussion risk factors across multiple injury settings: findings from the care consortium. *Inj. Epidemiol.* **6**, 1–11 (2019).

74. Dick, R. W. Is there a gender difference in concussion incidence and outcomes? *Br. J. Sports Medicine* **43**, i46–i50, DOI: 10.1136/bjsm.2009.058172 (2009). https://bjsm.bmj.com/content/43/Suppl_1/i46.full.pdf.

75. Berz, K. *et al.* Sex-specific differences in the severity of symptoms and recovery rate following sports-related concussion in young athletes. *The Physician Sportsmed.* **41**, 58–63, DOI: 10.3810/psm.2013.05.2015 (2013). https://doi.org/10.3810/psm.2013.05.2015.

76. Zuckerman, S. L. *et al.* Effect of sex on symptoms and return to baseline in sport-related concussion. *J. Neurosurgery: Pediatr.* **13**, 72–81 (2014).

77. Churchill, N. W. *et al.* The first week after concussion: Blood flow, brain function and white matter microstructure. *NeuroImage: Clin.* **14**, 480 – 489, DOI: https://doi.org/10.1016/j.nicl.2017.02.015 (2017).

78. Churchill, N. W., Hutchison, M. G., Graham, S. J. & Schweizer, T. A. Scale-free functional brain dynamics during recovery from sport-related concussion. *Hum. Brain Mapp.* **41**, 2567–2582, DOI: 10.1002/hbm.24962 (2020). https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.24962.

79. Muller, A. M. & Virji-Babul, N. Stuck in a state of inattention? functional hyperconnectivity as an indicator of disturbed intrinsic brain dynamics in adolescents with concussion: a pilot study. *ASN Neuro* **10**, DOI: 10.1177/1759091417753802 (2018).

80. McCrory, P. *et al.* Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in berlin, october 2016. *Br. J. Sports Medicine* **51**, 838–847 (2017).

81. Virji-Babul, N. *et al.* Changes in functional brain networks following sports-related concussion in adolescents. *J. Neurotrauma* **31**, 1914–1919, DOI: 10.1089/neu.2014.3450 (2014).

82. Porter, S. *et al.* Changes in brain-behavior relationships following a 3-month pilot cognitive intervention program for adults with traumatic brain injury. *Heliyon* **3**, e00373, DOI: 10.1016/j.heliyon.2017.e00373 (2017).

83. Rotem-Kohavi, N. *et al.* Network analysis of perception-action coupling in infants. *Front. Hum. Neurosci.* **8**, 209, DOI: 10.3389/fnhum.2014.00209 (2014).

84. Rotem-Kohavi, N., Oberlander, T. F. & Virji-Babul, N. Infants and adults have similar regional functional brain organization for the perception of emotions. *Neurosci. Lett.* **650**, 118–125, DOI: 10.1016/j.neulet.2017.04.031 (2017).

85. Pratiher, S., Chattoraj, S. & Mukherjee, R. Stationplot: A new non-stationarity quantification tool for detection of epileptic seizures. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 499–503 (IEEE, 2018).

86. Shen, Y.-W. & Lin, Y.-P. Challenge for affective brain-computer interfaces: Non-stationary spatio-spectral EEG oscillations of emotional responses. *Front. Hum. Neurosci.* **13**, 366, DOI: 10.3389/fnhum.2019.00366 (2019).

87. Raza, H., Rathee, D., Zhou, S.-M., Cecotti, H. & Prasad, G. Covariate shift estimation based adaptive ensemble learning for handling non-stationarity in motor imagery related EEG-based brain-computer interface. *Neurocomputing* **343**, 154–166 (2019).

88. Gu, X. *et al.* EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *arXiv preprint arXiv:2001.11337* (2020).

89. Thompson, J., Sebastianelli, W. & Slobounov, S. EEG and postural correlates of mild traumatic brain injury in athletes. *Neurosci. Lett.* **377**, 158–163 (2005).

90. Watson, B. O. Cognitive and physiologic impacts of the infraslow oscillation. *Front. Syst. Neurosci.* **12**, 44, DOI: 10.3389/fnsys.2018.00044 (2018).

91. Demanuele, C., James, C. J. & Sonuga-Barke, E. J. Distinguishing low frequency oscillations within the $1/f$ spectral behaviour of electromagnetic brain signals. *Behav. Brain Funct.* **3**, 1–14 (2007).

92. Sobie, C., Babul, A. & de Sousa, R. Neuron dynamics in the presence of $1/f$ noise. *Phys. Rev. E* **83**, 051912, DOI: 10.1103/PhysRevE.83.051912 (2011).

93. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Process.* **45**, 2673–2681 (1997).

94. Wakuya, H. & Zurada, J. M. Bi-directional computing architecture for time series prediction. *Neural Networks* **14**, 1307–1321 (2001).

95. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv e-prints* arXiv:1412.6980 (2014). 1412.6980.

Acknowledgements

Author contributions statement

A.B. and D.T.H. conceived the work; N.V.B. and L.B. collected and prepared the data; A.B., K.T., B.F., M.B. and A.G contributed to the design of the network; K.T., B.F., M.B. and A.G. trained and tested the network; A.B., N.V.B., K.T., and D.T.H. interpreted the results; A.B. and K.T. drafted the paper; A.B., K.T., D.T.H. and N.V.B. substantially revised the paper. All authors reviewed the manuscript.

Additional information

Competing interests
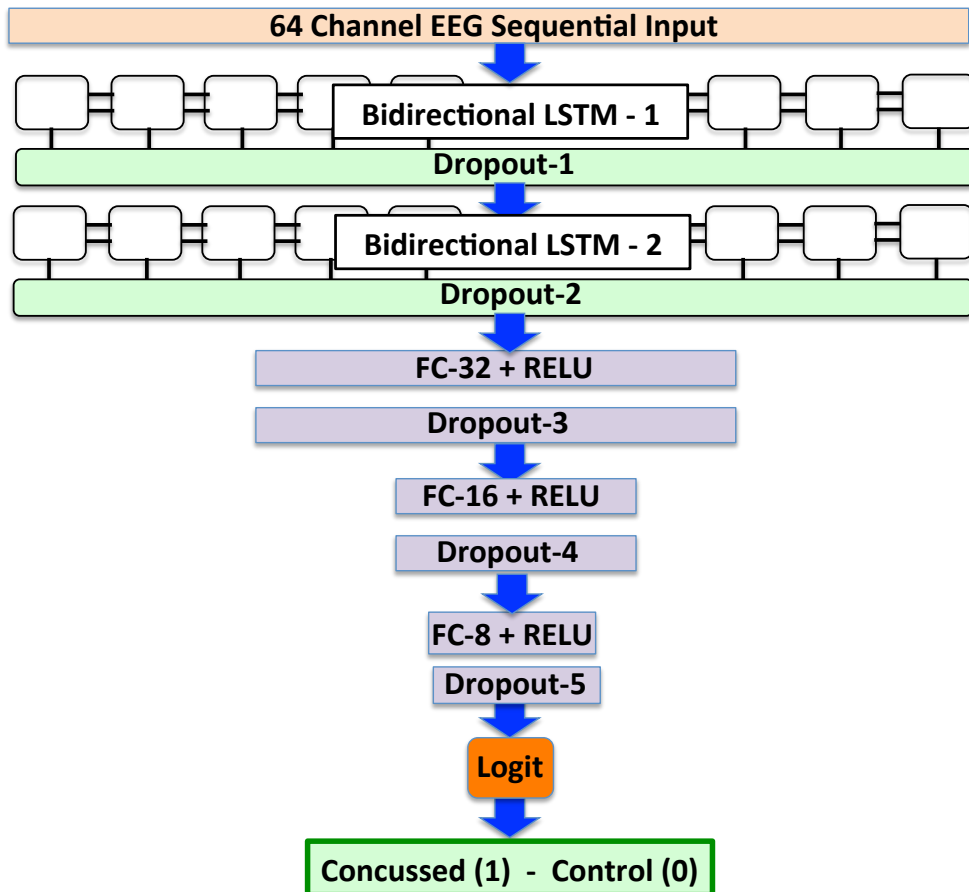
The authors declare no competing interests.

Figure 1. The architecture of our first generation *ConcNet*, a recurrent neural network architecture comprising two LSTM layers. The 64-channel EEG input is fed through a sequential input layer to the first layer with bi-directional Long Short Term Memory (LSTM) units. The output is then passed to a dropout layer for regularization to reduce overfitting, and then to the second LSTM layer that is similar to the first, another dropout layer, followed by three sets of fully connected and dropout layers. The fully connected layers have 32, 16 and 8 units and use a Rectified Linear Unit activation functions. The output is passed to a *softmax* logistic regression (logit) classifier, where it is assigned a probability and classified as *concussed* or *healthy*.
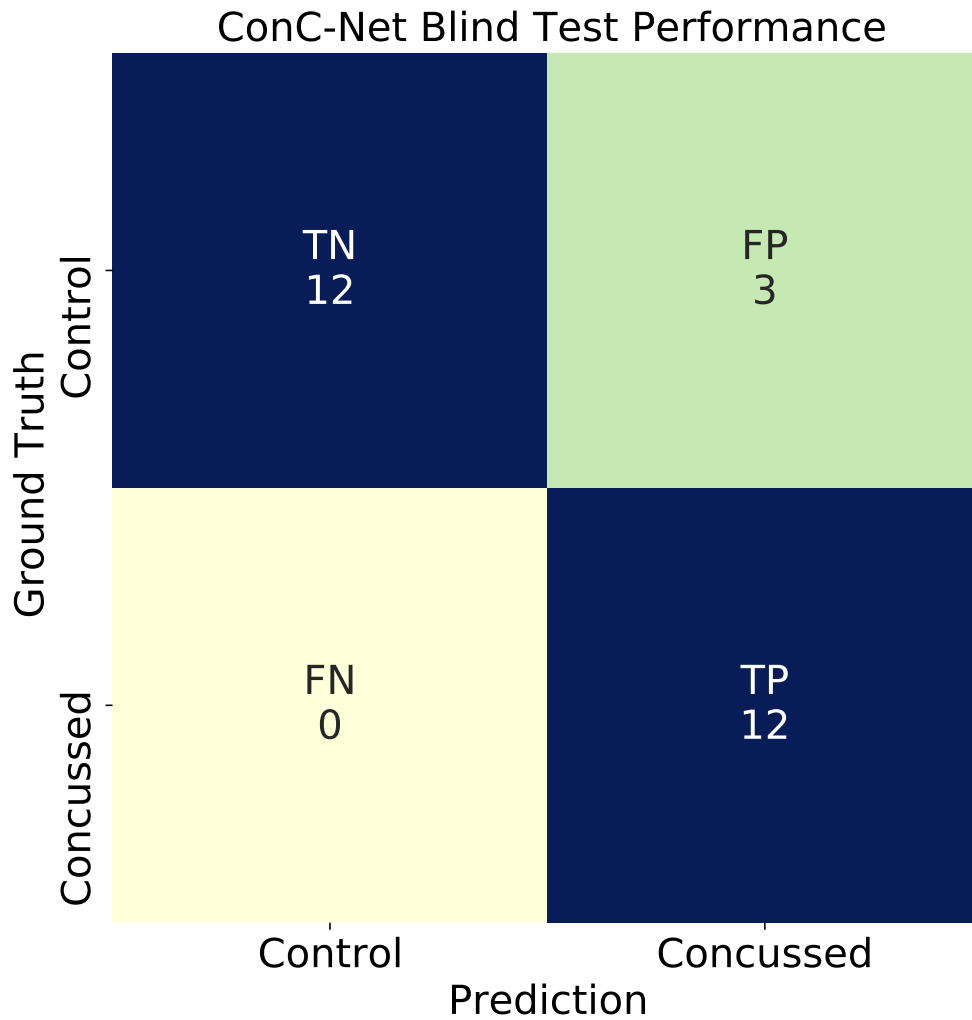
ConC-Net Blind Test Performance

Figure 2. Confusion Matrix summarizing the results from the blind test carried out in Stage 1.
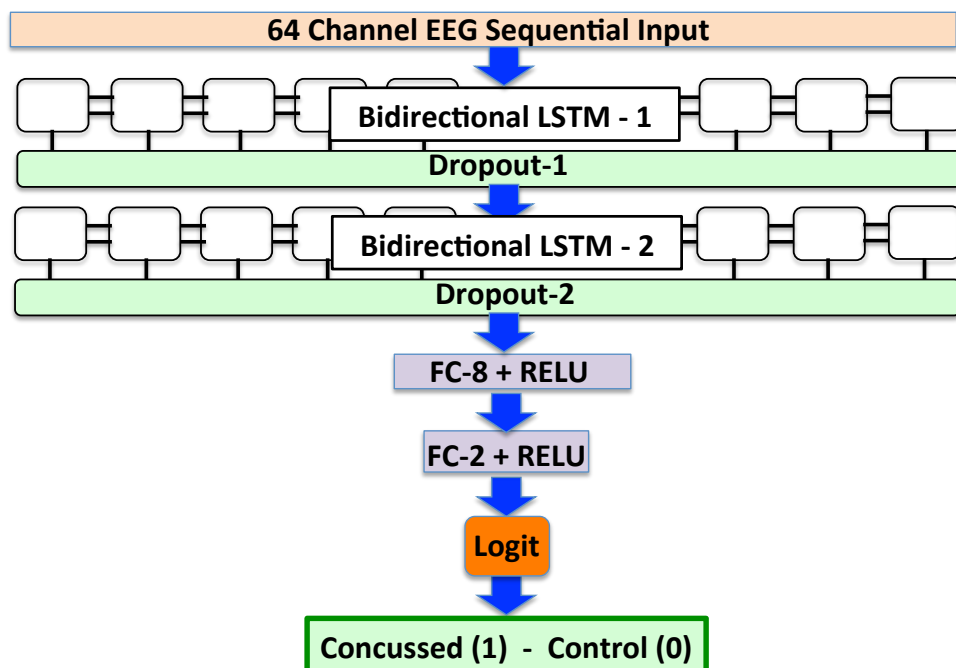
Figure 3. The architecture of our final streamlined version of *ConcNet* (*ConcNet* 2.0). The architecture is similar to that shown in Figure 1 except that the version shown here only 2 fully connected layers, with 8 and 2 units, and no associated dropout layers.
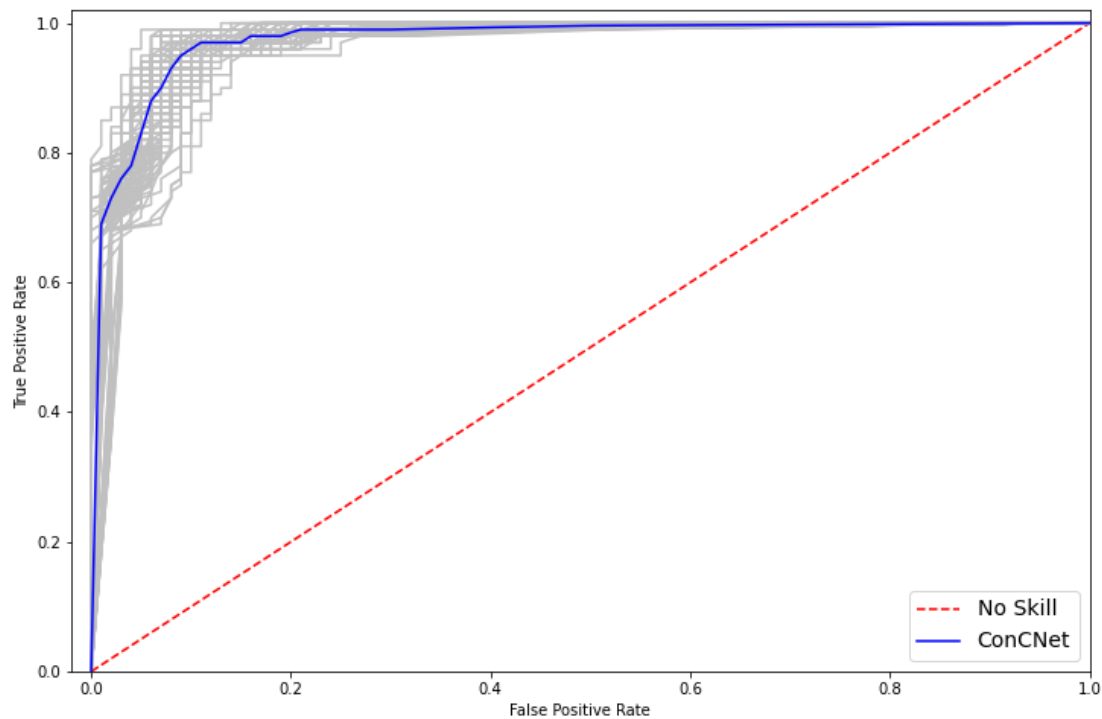
Figure 4. The Receiver Operating Characteristic (ROC) curves for the 100 LSTM networks. The grey curves show the results for each of the 100 networks and the blue curve shows the median. The ensemble median Area Under the Curve (AUC) is 0.971 (the 25% and the 75% percentile are 0.963 and 0.975, respectively). A classifier with no better accuracy than chance would be expected to have an AUC of 0.5, while an AUC of 1 corresponds to a classifier with perfect accuracy. A value of > 0.96 indicates a classifier with excellent discriminatory ability [69].
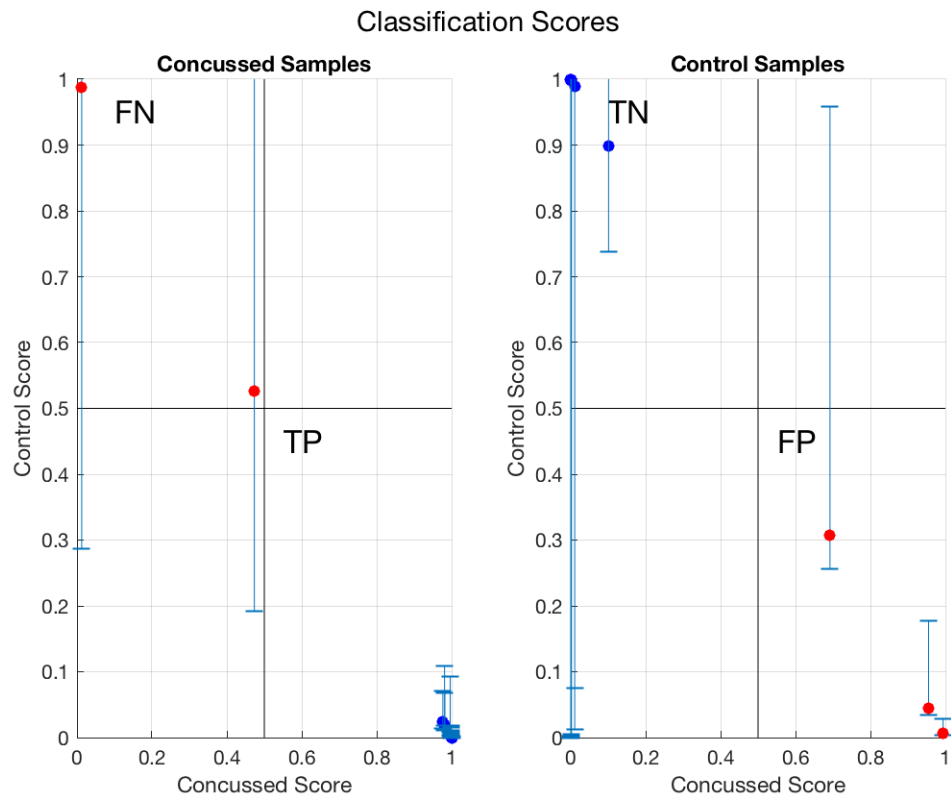
Figure 5. The median value of the probabilities, P_{mTBI}, assigned by the *softmax* logistic regression (logit) unit to each concussed and healthy participant whenever the sample was used for testing a network in the ensemble, plotted along the $x-$axis, and the median value of $P_{\mathrm{Hlth}} \equiv (1 - P_{\mathrm{mTBI}})$ plotted on the $y-$axis. The $y-$error bars show the 25th and 75th percentiles of the predicted probability scores. The left panel shows the results for the 27 concussed subjects, while the right shows the results for the 35 healthy individuals. Blue points represent samples which were properly classified based on our criterion of $P_{\mathrm{mTBI}} > 0.5$. The red points correspond to the individuals whose median score resulted in a misclassification.
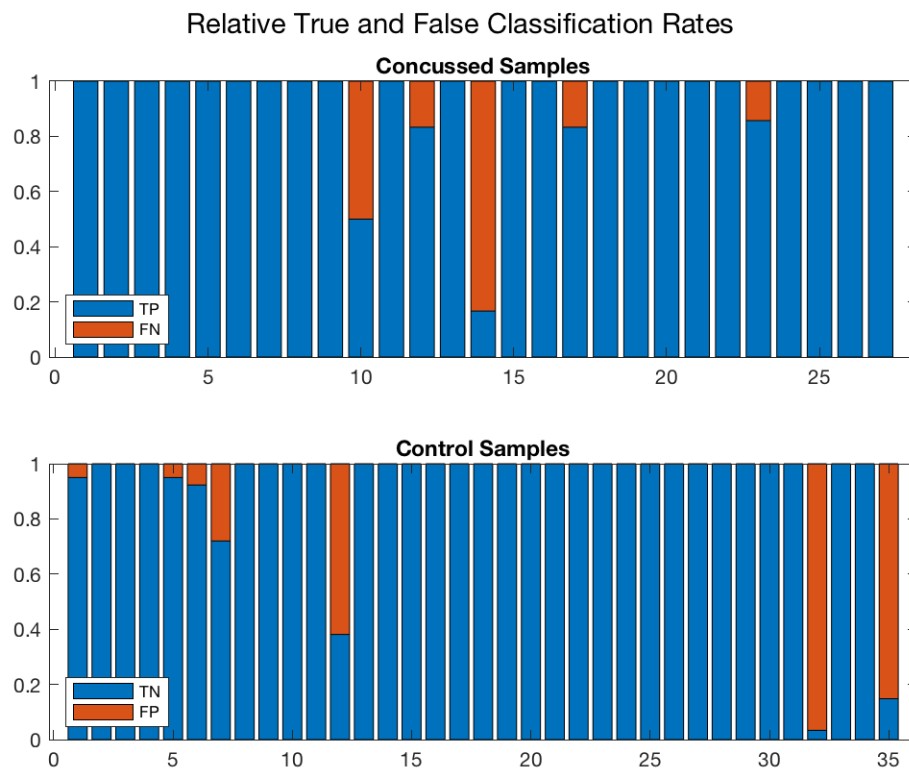
Figure 6. A different representation of the ensemble results for the 27 concussed and 35 healthy samples. The blue and red regions in each column illustrate the relative fractions of times that participant was properly or wrongly classified by the ensemble of networks. Three healthy and two concussed "problematic" subjects tended to be systematically misclassified by the networks in the ensemble. A handful of other participants were sometimes misclassified by one or two of the networks. On further examination, the training of the latter involved an EEG sample from one of the "problematic" individuals.

Demographic information	Healthy	Concussed
Age in years (SD)	14.7 (2.1)	13.5 (2.6)
Sex	100% male	100% male
Time since concussion		100% within one month
SCAT: Number of symptoms (SD)		9.2 (6.2)
SCAT: Symptom severity (SD)		21.7 (18.7)
Child SCAT: Number of symptoms (SD)		13.0 (5.4)
Child SCAT: Symptom severity (SD)		23.2 (13.6)

Table 1. Demographic information for the participants involved in this study. SD stands for standard deviation.

Performance Metric $=$ Definition	Alternate name	
$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$		
$Recall = \frac{TP}{TP+FN}$	also True Positive Rate (TPR)	
$Precision = \frac{TP}{TP+FP}$	also Positive Predictive Value (PPV)	
$Specificity = \frac{TN}{TN+FP}$	also True Negative Rate (TNR)	
$Miss\ Rate = \frac{FN}{FN+TP}$	also False Negative Rate (FNR)	$FNR \equiv 1 - TPR$
False Discovery Rate (FDR) $= \frac{FP}{FP+TP}$		$FDR \equiv 1 - PPV$
False Positive Rate (FPR) $= \frac{FP}{FP+TN}$		$FPR \equiv 1 - TNR$
Negative Predictive Value (NPV) $= \frac{TN}{FN+TN}$		
$Informedness = TPR + TNR - 1$		
$Markedness = PPV + NPV - 1$		

Table 2. Common measures used in this paper to gauge the performance of classification tasks. In the formula above, TP (True Positives) is the number of correctly classified concussed individuals; TN (True Negatives) is the number of correctly classified healthy participants; while FP and FN are the numbers of falsely classified healthy (i.e. False Positives) and concussed (i.e. False Negatives) participants, respectively.

Blind Test Subject	Actual Status	Segment 1	Segment 2	Segment 3	Result			
					TN	FP	FN	TP
Test 1	Concussed	✓	✓	✓				3
Test 2	Healthy	✓	✓	✓	3			
Test 3	Healthy	✓			1			
Test 4	Healthy	✓	✓		2			
Test 5	Healthy	✓			1			
Test 6	Healthy			✓	1			
Test 7	Concussed		✓					1
Test 8	Healthy	✓	✓		2			
Test 9	Concussed	✓		✓				2
Test 10	Healthy	✓			1			
Test 11	Concussed		✓	✓				2
Test 12	Healthy	✓	✗		1	1		
Test 13	Concussed	✓		✓				2
Test 14	Healthy	✗				1		
Test 15	Concussed			✓				1
Test 16	Healthy	✗				1		
Test 17	Concussed			✓				1
				Total:	12	3	0	12

Table 3. Details of the Stage 1 blind test results for *ConcNet*. The table shows the number of distinct segments from any one participant were that were present in the blind test sample, and how each of these were classified. All except one pair were consistently — and correctly — classified. Final column gives the results in terms of the number of True Negative (TN), False Positive (FP), False Negative (FN) and True Positive (TP). These are further summarized in the Confusion Matrix in Figure 2.

Metric	Q_2	Q_1	Q_3	Metric	Q_2	Q_1	Q_3
Accuracy	0.917	0.833	1.000	*Area Under Curve (AUC)*	0.971	0.963	0.975
Recall (TPR)	1.000	1.000	1.000	*Informedness*	0.900	.800	1.000
Precision (PPV)	0.667	0.500	1.000	*Markedness*	0.667	0.500	1.000
Specificity (TNR)	0.900	0.850	1.000	*Miss Rate* (FNR)	0.000	0.000	0.000
False Discovery Rate (FDR)	0.333	0.000	0.500	*False Positive Rate* (FPR)	0.100	0.000	0.150

Table 4. The ensemble median (Q_2), the 25% percentile (Q_1), and the 75% percentile (Q_3), of the various network metrics distributions. The statistics are derived from the 100 networks ensemble test, quantifying the performance of *ConcNet* 2.0. The results for the five participants who were systematically misclassified have *not* been excluded.