Smart Health 27 (2023) 100352

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

BioscoreNet: Traumatic Brain Injury (TBI) detection using a multimodal self-attention fusion neural network and a passive bioscore monitoring framework from smartphone sensor data

Florina Asani, Bhoomi Patel, Srinarayan Srikanthan, Emmanuel Agu *

Worcester Polytechnic Institute, 100 Institute Road, Worcester 01609, MA, USA

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Traumatic Brain Injury
Smartphone sensing
Passive monitoring
Neural networks
Bioscore

ABSTRACT

Traumatic Brain Injury (TBI) affects millions of individuals globally and can cause motor, cognitive and emotional deficits. Emerging research in the field of mobile health has demonstrated that smartphone sensor data can be used to monitor various chronic diseases, which can facilitate early ailment detection and continuous monitoring of recovering patients. The goal of the paper's contributions is two-fold: Proposing (1) A Deep Learning framework that uses smartphone sensor data mobility data (accelerometer, magnetometer, gyroscope, pedometer, pressure, altitude, accessibility) to detect whether a subject has TBI after a head injury. (2) The concept of a TBI Bioscore, a number that quantifies the certainty (0–1) that a subject has TBI. The TBI deep learning detection model uses a self-attention mechanism for multimodal feature fusion and a stacked LSTM for prediction. This model achieves a Balanced Accuracy (BA) of 90.2% and a True Positive Rate (TPR) of 83.3% in correctly identifying TBI instances. The Bioscore is generated using Monte-Carlo Dropout uncertainty estimation. When a users' TBI Bioscores are visualized throughout, it steadily decreases for users who have normal recoveries. The proposed TBI detection and Bioscore framework facilitates population-level passive, continuous and remote TBI screening and monitoring. A glanceable display of color-coded Bioscores can enable a small team of medical staff to monitor a large pool of patients and intervene preemptively.

1. Introduction

Problem: The current healthcare system is schedule-driven with many patients receiving care infrequently. This situation becomes especially problematic for long-term ailments or those that require frequent check-ups during an extended recovery period. Increasing the frequency of health assessments is challenging as many healthcare providers are already understaffed and overworked (Dewart, Corcoran, Thirsk, & Petrovic, 2020). Population-level tools for effective public health surveillance and continuous monitoring are needed to solve problems that are currently costing the healthcare system billions of dollars, including early ailment detection and recovery monitoring. *Early detection of ailments:* if ailing users can be discovered early, patient suffering, hospitalization, deaths and overall costs can be reduced. *Recovery monitoring:* is very useful for patients that have been diagnosed with chronic diseases such as Traumatic Brain Injuries (TBI), cancer, patients with mobility issues or the elderly, which involve an extended recovery phase. Patients' recovery trajectories can be monitored to detect health deterioration early and prevent

* Corresponding author.
E-mail address: emmanuel@wpi.edu (E. Agu).

https://doi.org/10.1016/j.smhl.2022.100352
Received 30 September 2022; Accepted 5 October 2022
Available online 2 November 2022

complications or even death. Without adequate population-level recovery monitoring tools, regressing patients are sometimes discovered and re-hospitalized too late, costing Medicare $26 billion annually (Wilson, 2019).

Background: TBI causes distress to millions of individuals worldwide. Severe TBI cases lead to significant motor, cognitive and emotional deficits. Even mild injuries, which make up to 80% of cases, can lead to post-concussion symptoms which can in turn affect the function of patients. Baldwin, Breiding, and Sleet (2016) Approximately, 70% of TBI cases are caused by sudden impact on the head resulting from falls, traffic and sport accidents among others. In the US only, TBI accounts for 30% of all injury-related deaths. Baldwin et al. (2016). The more severe the brain injury, the more likely a person will experience extensive physical and mental symptoms following the injury. For this reason, and for all other cases it is important to have early identification and recovery monitoring, so that post-concussion symptoms can be treated. Assessment of TBI usually includes a neurological exam, which includes an evaluation of thinking, motor function (movement), sensory function, coordination and reflexes. Smartphone sensor data can capture most of the above-mentioned functions. Emerging mobile health research has demonstrated that smartphone sensor data can be used to detect and monitor chronic diseases, decrease healthcare visits, and encourage healthy behavior (Wang et al., 2014). Smartphone data can be gathered and passively, and analyzed to capture patient health behaviors that are predictive of an individual's health status. Such continuous monitoring is can improve outcomes significantly for chronic health diseases such as TBI.

Our novel TBI Bioscore concept: In addition to detecting smartphone users that have an ailment (TBI), we also focus on screening them passively and continuously. Specifically, we recognize that accurate ailment detection in a noisy world is extremely challenging, and instead use smartphone sensing to identify and flag subjects that are highly likely to have an ailment (e.g. TBI). We assume that as is typical in most healthcare scenarios, that the final determination will be made by trained medical personnel after the subject undergoes detailed examinations. Essentially, we assume a human-in-the-loop scenario, where smartphone sensing (phenotyping) suggests subjects with the highest probability of having an ailment. A nurse or medical staff can then reach out (e.g. quick phone call) to such flagged users to decide whether they need to visit the clinic for confirmatory tests (as shown in Fig. 1). In this paper, we propose a novel bioscore concept, which quantifies the likelihood (certainty) that a subject has a given ailment from sensed smartphone data. For instance, a subject's TBI bioscore is the probability that they have TBI and an influenza bioscore is the probability that subject has influenza. The bioscore concept can be adapted to different ailments, as needed. Ailment-specific bioscore models an be trained and deployed. The bioscore can take values in the 0 to 1 range, with higher values indicating a higher probability that the user has the given ailment (TBI). Such higher bioscores would prompt medical personnel to recommend a clinic visit for further prognosis and treatment. On the other hand, bioscores close to 0 correspond to healthy subjects.

In our proposed framework, the bioscore can be used for both early ailment detection and recovery monitoring scenarios. In the recovery phase, after a patient has been discharged from a hospital, or after an injury occurred, the bioscore will be initially high. If the subject recovers smoothly and without regressing, their bioscore will decrease over time (e.g. on a weekly or monthly basis, depending on the ailment). Should the bioscore of a subject fail to decrease in an ailment-specific recommended time frame, or even begin to increase, nurses or medical staff monitoring their bioscore will be prompted to investigate further by visiting or calling the subject and suggesting early readmission. On the other hand, in the early ailment detection scenario, a health subject's bioscore will initially be low (or close to zero). If the subject becomes afflicted by an ailment (e.g. influenza), their bioscore will increase as more symptoms manifest, assuming a value of 1 when the bioscore models predict that they almost certainly have the ailment. To operationalize the bioscore, users' smartphone sensor data is continuously gathered throughout the day and pushed to the cloud overnight, where bioscore computation is performed. Bioscores for a hundreds of patients can be displayed on a large glanceable dashboard, facilitating a small team's to monitor a potentially large population and identify users and at-risk communities with the highest bioscore. The development steps are shown in Fig. 2. For instance, a few care providers (e.g. nurses) could monitor the bioscores of an entire college campus of thousands of students, military unit or large nursing home. First thing in the morning when they resume work, they can glance at the bioscore display for the population that was computed overnight. They might choose to call the N users with the highest bioscores every morning, where N is a very small compared to total population of the community. For instance, the 5 users with the highest bioscores out of 1000 patients being monitored can be called. For each case, a 2 min phone call would either rule out confounds (e.g. acting slow because I helped my friend move houses yesterday) or lead the nurse to invite the subject for further tests in the clinic.

Specific example-TBI bioscore: As a concrete example, we illustrate the bioscore concept by using smartphone-sensed data to quantify the [0–1] likelihood that a smartphone users suffered a Traumatic Brain Injury, soon after an injury has been recorded. These users have suffered an injury that may have caused mild, moderate or severe TBI, been treated, released from hospital, and they need to be continuously supervised for possible health deterioration. Our TBI bioscore is computed and visualized in three phases (shown in Fig. 2):

1. *Phase 1, Binary classification of TBI status:* We propose a neural networks model for multi-modal smartphone sensor data to accurately detect subjects' that may have suffered TBI (a concussion) after an injury has occurred. We use a multi-modal self-attention mechanism for weight-based feature fusion of different modalities (accelerometer, gyroscope, magnetometer, pedometer, accessibility, altitude, pressure) which are then fed to a stacked LSTM deep learning model for prediction.
2. *Phase 2, Bioscore generation:* as the certainty associated with binary TBI (Yes/No) classification. A Monte-Carlo dropout mechanism is included in the TBI Detection framework that estimates uncertainty associated with the model's prediction, which is further scaled and mapped to a [0–1] range to convert it to the TBI bioscore.
3. *Phase 3, Bioscore visualization:* Bioscores of a large number of monitored subjected are color-coded and displayed on a dashboard we have created for TBI bioscore visualization over several days, which flags the subjects that need follow-up by medical personnel. This TBI score dashboard can be used for both early ailment detection as well as for monitoring the recovery of a patient after being treated and discharged from hospital.

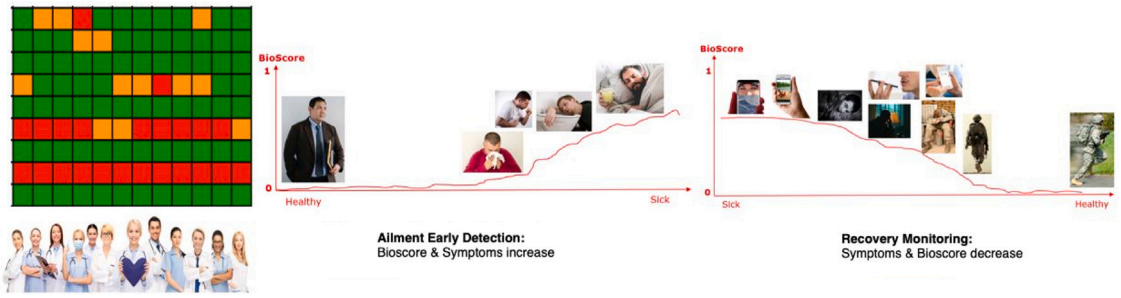

Fig. 1. Our proposed bioscore framework. Left: A small medical team can monitor bioscores of a large cohort on a dashboard. Middle: Bioscore rises from low value to high in early detection scenario. Right: Bioscore decreases in a recovery monitoring scenario.

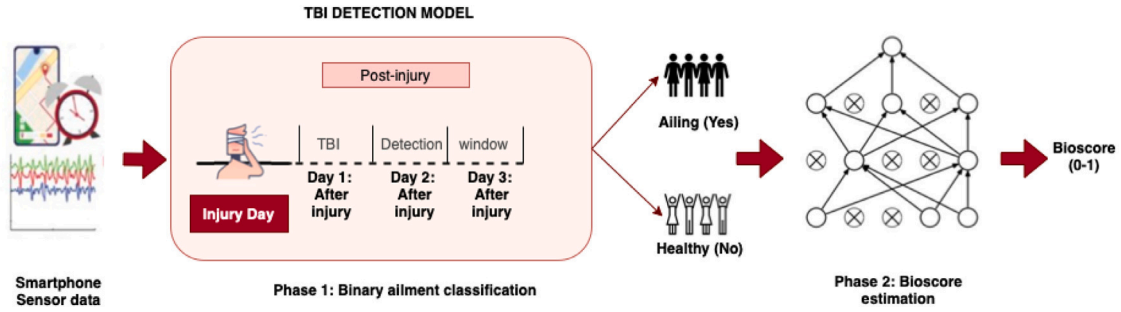

Fig. 2. Bioscore development steps.

Related work: Our TBI bioscore concept is related to two bodies of work. The first body of work includes related research that does TBI detection using smartphone sensor data or any other ailment. And the other body of work, is the work that is closely related to the concept of bioscore or quantified scoring of "illness" levels. *Smartphone sensing and detection of ailments:* An example of a work that belongs to the first group, is Mariakakis et al. (2017). In this paper, Convolutional Neural Networks (CNNs) have been used to detect concussions for athletes. They use the camera as a sensor to monitor pupil movement to detect concussion, which in turn requires active user participation, whereas we use passively generated smartphone sensor data that capture mobility and movement. *The concept of a bioscore* is closely related and draws inspiration from the area of risk quantification of the future occurrence of an adverse event, using human behavior data, which is broadly applicable in many areas such as: fraud detection scores (Bhavani & Amponsah, 2017), risk assessment in criminal justice (Ling & Raine, 2018) and healthcare (Xu et al., 2018). In healthcare, risk quantification tools are used to assess an individual's health status, to identify people who will need help early. To the best of our knowledge, our work is the first to propose a smartphone-sensed score to dynamically quantify the likelihood a user has an ailment (TBI).

Our contributions: The following are the key contributions of the paper:

1. A TBI detection framework to accurately detect whether a user has TBI (a concussion) is proposed, which uses a self-attention mechanism for feature fusion of different modalities (accelerometer, gyroscope, magnetometer, pedometer, accessibility, altitude, pressure), by assigning relative weights based on their attention score to create uniform feature representation.
2. In rigorous evaluation, our TBI detection framework was able to accurately identify TBI instances with a True Positive Rate (TPR) of 83.3%, and a True Negative Rate (TNR) of 97% when identifying non-TBI instances, outperforming all baselines.
3. The concept of a bioscore is introduced with a TBI Bioscore as a concrete example. The TBI Bioscore is the likelihood, in the [0–1] range, that a smartphone user has TBI. The Bioscore can be used for population-level ailment screening and early ailment detection as well as subject monitoring during recovery.
4. Monte-Carlo dropout is utilized to estimate the model's uncertainty, which is then scaled and converted to a TBI bioscore.
5. The distribution of subjects' TBI bioscore in the post-injury days is successfully calculated and visualized in a glanceable dashboard.

The remainder of the paper is organized as follows: In Section 2, related work is analyzed and compared to the proposed framework. Section 3 provides background of the techniques utilized. In Section 4, we describe our methodology including the dataset used, the feature extraction process and multi-modal attention based neural network. Our evaluation of the aforementioned frameworks, and results are presented in Section 5. Finally, Section 6 presents our conclusion and future work.

2. Related work

2.1. Smartphone-sensing for ailment detection

Recent research has shown that mobile sensor data can be used for a broad range of passive health assessments such as Parkinson's (Lipsmeier et al., 2018), anxiety (Saeb, Lattie, Kording, & Mohr, 2017), schizophrenia (Torous & Keshavan, 2018), and influenza (Murthy, Asani, Srikanthan, & Agu, 2020), as well as human activity recognition (Fu, Damer, Kirchbuchner, & Kuijper, 2020). These smartphone-based health assessment methods can analyze users' daily routines, habits, and movement patterns as well as to assess their mental and physical health at the individual level without the user being actively involved. Passive smartphone ailment detection without user involvement is more related to this work than work that requires active user involvement. A subset of the TBI work is summarized in Creber et al. work (Creber et al., 2016). A thorough survey of smartphone apps intended to assess TBI is provided, but they found that the apps mainly are used to document, manage, and gather information about various aspects of concussions. This body of work does not automate TBI screening with little to no subject involvement compared to the work proposed in this paper. Mariakakis et al. (2017) propose an automated TBI screening using pupil movement and phone camera for detection, which includes active participation of a subject in screening. In the collected videos, light stimulus with different intensities was setup to be followed by the human eye. In this work, they employ Convolutional Neural Network to measure pupil diameter for gaze tracking from videos. The output of the architecture is the pupil's diameter which can be used to later detect TBI by analyzing the latency between the time of the light stimulus and when the pupil begins to constrict and constructing PLR curves.

Our work, is similar to prior work on smartphone-sensing of ailments as it uses deep learning to analyze sensor data for the same end-goal of identifying an ailment, TBI. The difference from previous TBI detection methods, is the fact that instead of using cameras that seek active participation, in our work, we use passively generated smartphone sensor data to capture mobility patterns as well as user–smartphone interactions to automatically detect and flag possible concussion after injury. The proposed work does not only focus on accurate prediction, but also provides means to monitor the progression or regression of TBI in longer terms.

2.2. Bioscore generation

Health scores The concept of bioscore is related to the area of risk quantification of the future occurrence of an adverse events using human behavior data which is broadly applicable to many areas including fraud detection scores (Bhavani & Amponsah, 2017), risk assessment in criminal justice (Ling & Raine, 2018) and healthcare. In healthcare, risk quantification tools are used to assess an individual's health status, to identify people who will need help early. Prognostic scores quantify an individual's risk of being afflicted with an ailment or condition such as Hodgkin's disease or cancer. Xu et al. introduces a related bioscore that scores the prognostic significance of underlying tumor in breast cancer patients (Xu et al., 2018). More recent examples include emerging research in infectious diseases. The Proximity Index is a risk index that identifies patients at high risk of COVID-19 infection to alert them (Liu, Li, Xu, & Natarajan, 2018). The work proposed in this paper, shares similarities with the previously mentioned work, in the second and final layer of the framework, by proposing a score that from a higher level perspective, quantifies the risk of having TBI. The difference is that the quantified risk is generated through dynamic smartphone-sensed data and is applied for ailment assessment. To the best of our knowledge, the work proposed in this paper, is the first to propose such score. Till date, most prognostic scores usually assess long term risk typically years and not days like the proposed bioscore. Moreover, prognostic scores are derived from lifestyle, biological and genetic risk factors and not from smartphone-sensed behavior data.

Uncertainty methods for generating our bioscore Our bioscore estimation methods are based on neural networks uncertainty quantification methods, specifically the Monte Carlo Dropout proposed by Gal and Ghahramani (Gal & Ghahramani, 2016). They demonstrated that dropout in neural networks can be interpreted as a Bayesian approximation of a well-known probability model: Gaussian process. This method was selected over other proposed Bayesian inference approaches (Blundell, Cornebise, Kavukcuoglu, & Wierstra, 2015; Graves, 2011; Kingma, Salimans, & Welling, 2015; Louizos & Welling, 2017) or deep ensemble methods (Lakshminarayanan, Pritzel, & Blundell, 2016) due to its simplicity and low computational cost. Monte Carlo Dropout works by randomly masking/turning off neurons in the input during both stages of training and testing. This method is used to generate random predictions and interpret them as samples from a probabilistic distribution. In Bioscore estimation, this uncertainty is used to express how certain the classification model is in its binary prediction of whether a subject is recovering from concussion.

3. Background

3.1. Self-attention mechanism

In our work, we use attention mechanism to weight different modalities of features differently. Attention is a mechanism for dynamically highlighting important parts of input data. It works based on the intuition that humans focus on only a relatively small region when processing a large amount of information. The concept of attention was first introduced with the architecture of Transformers from Vaswani et al. (2017) The self-attention mechanism weights different positions of a single sequence relative to each other in order to compute a representation of the same sequence. Self attention can be formalized with the following mathematical equations:

$$h_{t,t'} = \tanh\left(x_t^T W_t + x_{t'}^T W_x + b_t\right) \tag{1}$$

$$e_{t,t'} = \sigma\left(W_a h_{t,t'} + b_a\right) \tag{2}$$

$$a_t = \mathrm{softmax}\left(e_t\right) \tag{3}$$

$$l_t = \sum_{t'} a_{t,t'} x_{t'} \tag{4}$$

3.2. LSTM for time series classification

To exploit the temporal patterns within smartphone-sensor data, we used LSTMs to classify whether a subject has TBI. Long Short-Term memory network, or LSTM, is a type of Recurrent Neural Network (RNN) (Hochreiter & Schmidhuber, 1997). This network uses previous time events to inform or classify the later event. An LSTM module has a cell state and internal mechanisms called gates which provides them with the power to selectively learn or retain information from other units. They can basically help regulate the flow of information. The three gates that an LSTM network has are the forget gate, input gate and output gate that use hyperbolic tangent and sigmoid activation function. It is in the forget gate that the information that needs to be forgotten is controlled when a new information enters the network (Van Houdt, Mosquera, & Nápoles, 2020).

3.3. Monte-Carlo dropout uncertainty estimation

To generate TBI bioscore, as the likelihood(certainty) (0–1) that a subject has TBI, we use Monte Carlo Dropout uncertainty estimation. Monte-Carlo dropout is used to model uncertainty in the predictions. It works by running multiple forward passes on the trained network with different dropouts. To estimate the uncertainty of one sample we assimilate the predictions from multiple networks with different dropout. By computing mean and variance on these assimilated predictions we get an estimate of the model's posterior distribution and uncertainty pertaining to that sample which is given by:

$$P = \frac{1}{T}\sum_{i=0}^{T} f_n^{di}(x) \tag{5}$$

$$c = \frac{1}{T}\sum_{i=0}^{T}\left[f_n^{di}(x) - P\right]^2 \tag{6}$$

where P is Predictive posterior mean and c is Uncertainty.

4. Methodology

4.1. TBI dataset description

The dataset used in this work, was collected by Charles River Analytics as part of DARPA's Warfighter Analytics for Smartphone Healthcare (WASH) program. The participants provide consent for the collection of passive data gathered through sensors on their cell phone and potential wearable devices with overlapping sensors. The smartphone sensor data used in the dataset are: accelerometer, gyroscope, magnetometer, pedometer, accessibility sensors (Assistive Touch, Dark system colors etc.), altitude and pressure. The participants considered were all iOS device users. All participants are prompted to complete a baseline questionnaire regarding their health, mood, physical activity and smartphone usage. Among these questionnaires, there are questions that are specifically asked to identify TBI and any injury that may have caused a concussion. The subjects that have been diagnosed with TBI, were identified using the answers to the following questions:

- Were you in any accidents this week?
- Did you hit your head at all this week?
- Did you become unconscious due to the injury?
- Did you see a doctor for your head injury?
- Did the doctor say you have a concussion?

A total number of 1522 subjects who have reported an impact in their head was recorded. Out of that number only 67 instances had reported a visit to the doctor and 44 total instances were diagnosed with concussion. The sensor data collected were X,Y and Z-axis values along and their associated timestamp from the accelerometer, gyroscope and magnetometer. The other sensor data gathered were pressure and altitude, the battery state, pedometer logs and accessibility features (Assistive Touch, Bold Text, Closed Captioning, Darker System Colors, Voice over, Mono System Audio, Reduced Transparency, Shake to Undo, Speaking the screen, speaking the selection and switch control)

The TBI bioscore is generated in four stages: (1) Feature Creation that includes the following sub-stages: (a) Data Collection and Preprocessing, (b) Data Segmentation in different windows and, (c) statistical feature extraction for that segment (2) Feature fusion through attention scores, (3) Deep Learning modeling to accurately detect subjects who could have TBI, (4) Bioscore generation through predictions obtained in stage 3. The architecture overview is shown in Fig. 3.

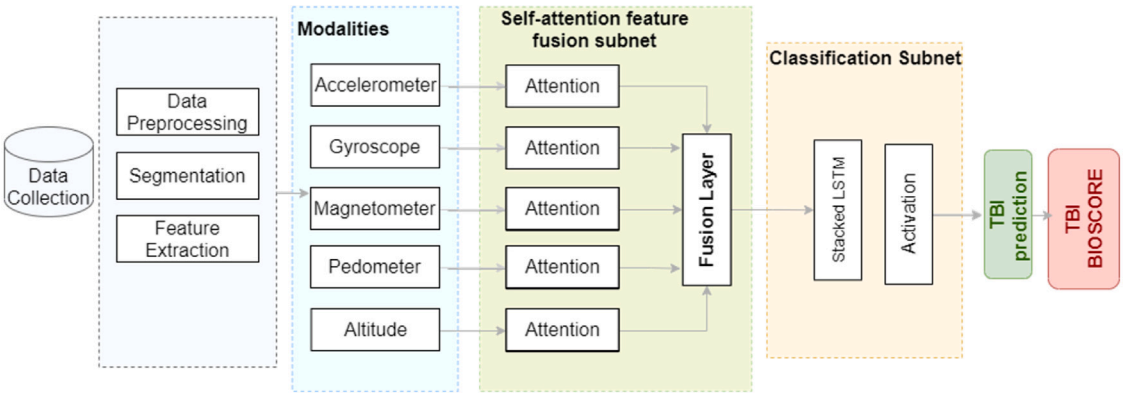

Fig. 3. Overview of our TBI Detection and bioscore generation approach.

Table 1
Statistical features extracted from all sensors.

Statistical Feature	Equation
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
Standard deviation	$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - x)^2}$
Variance	$\sigma^2 = \left[\sum (x-\mu)^2\right] \quad /N$
Kurtosis	$\mathrm{KV} = \frac{1}{\mathrm{N}}$ $\sum_{i=1}^{\mathrm{N}} \left(\frac{\mathrm{x}_1 - \mathrm{x}}{\sigma}\right)^4$
Peak to peak	$PPV = \mathrm{MAX}(M) - \mathrm{MIN}(M)$
Skew	$SV = \frac{1}{N}$ $\sum_{I=1}^{N} \left(\frac{x_i - x}{\sigma}\right)^3$
Interquartile range	$IRQ = \frac{3}{4(n+1)}\mathrm{th} - \mathrm{term}$ $\frac{1}{4(n+1)}\mathrm{th}$ - term
Accessibility features	Boolean value, indicating whether the feature has been turned on and used during that timeframe

4.2. Feature extraction

The purpose of the framework is to accurately detect TBI instances soon after an injury occurred (one day after an injury, two, or even three days). Therefore, the data collected per subject only utilizes the specified time frame. Prior to extracting features, the data was segmented with different sliding window sizes, corresponding to a time frame of one day, two days or three days of data. For every time frame, with its associated window sized and 50% overlap, statistical features were computed and feature were created. Statistical features such as: mean, variance, peak to peak, interquartile range, skewness, kurtosis and standard deviation were computed for all sensor modalities (accelerometer, gyroscope, magnetometer, pedometer, accessibility sensors (Assistive Touch, Dark system colors etc.), altitude and pressure), using the equations in Table 1. For accelerometer, gyroscope and magnetometer the Mean, Variance, Standard deviation, Interquartile range, Kurtosis, Peak and Skewness was computed for all three axis (x,y,z), and in turn 21 features for each of them were created. For pedometer 15 features were computed, which were the following: Mean Active Pace, Floor Ascend Count, Floor Descend Count, Kurtosis, Skew, Peak, Interquartile range of distance and step alongside the Mean, standard deviation and variance of steps as well as the total distance. The same statistical features were computed for pressure and altitude modalities. Each of these modalities had 7 features in total. A total of 92 statistical features were created and 14 boolean features that correspond to the accessibility features (Assistive Touch, Dark system colors etc.) After these features have been computed and extracted for every time frame and window-size, unimodal features are created, cleaned, and normalized, to be then fed as an input to the attention subnet for feature fusion. This subnet is described in more details in the following section.

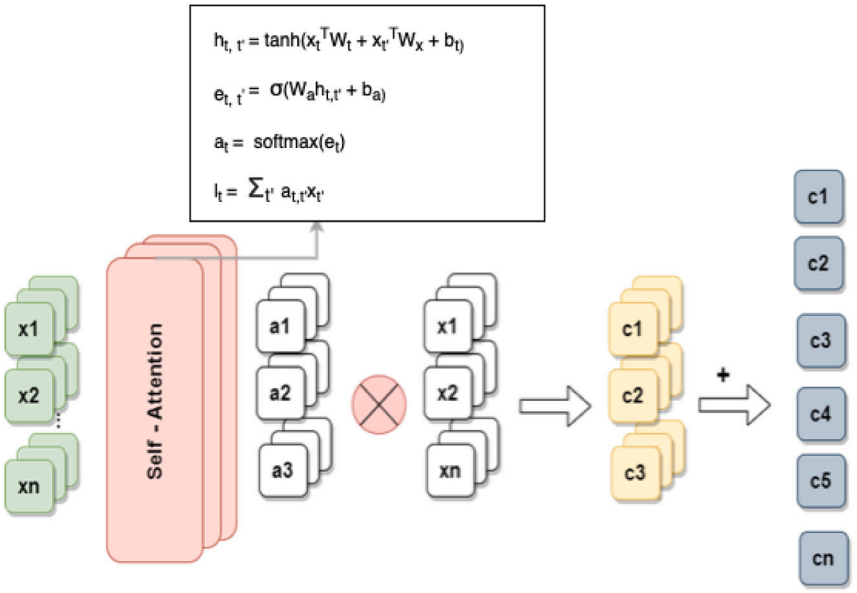

Fig. 4. Self-attention multimodal fusion module.

4.3. Self-attention subnet for feature fusion

When using multimodal data, it is common that not all modalities contribute equally to the classification task at hand. To enhance their contribution and prioritize the important modalities, this subnet employs a self-attention network for feature fusion based on their importance (Liu et al., 2018). For one sequence of features with different timesteps, self-attention computes a score that relates the different features of the sequence. It shows which part of the sequence do other positions pay more attention to. As shown in Fig. 4, the self-attention mechanism takes unimodal sensor feature vectors [st1, st2,... stk] as input, so for every modality (sensor type) such as accelerometer, gyroscope, magnetometer, pedometer, altitude, pressure and accessibility it takes the features across all timesteps, and outputs a vector of attention scores [at1, at2 ..atk] for every modalities' features, using the self-attention layers formalized in Eqs. (1), (2), (3) and (4).

$$a_t = self\_attention(st) \tag{7}$$

$$f_t = \sum_{t'} \frac{1}{k} a_{t,t'} s_{t'} \tag{8}$$

These scores are used as feature weights by multiplying them with the relative modality weight and concatenating to create the final feature vector, ft.

4.4. Stacked LSTM for classification

After self-attention fusion subnet, the output which represents the final set of weighted features, is fed to a stacked LSTM Structure (two layers). Since LSTM operates on sequence data, multiple stacked layers add levels of abstraction of input observations over time. The first LSTM layer, provides a sequence of outputs to the following LSTM layer. Stacking LSTM hidden layers makes the model deeper, which helps to better approximate the data and have more successful results. A deeper architecture is widely used and it is shown that for specific tasks it performs better than shallow architectures. For example, Sutskever, Vinyals, and Le (2014) report that a 4-layers deep architecture was essential in achieving good performance in an encoder–decoder framework for the machine-translation task. Since our goal is binary TBI classification, the activation function of the output layer used in the framework is sigmoid. The output later generated a score between 0 and 1 of whether concussion patterns are detected in the subject.

$$\text{label} = \text{sigmoid}(W * x + b) \tag{9}$$

Fig. 5 depicts of our TBI detection model. The output of the prediction is then passed and used to generate the bioscore as explained in the following section.

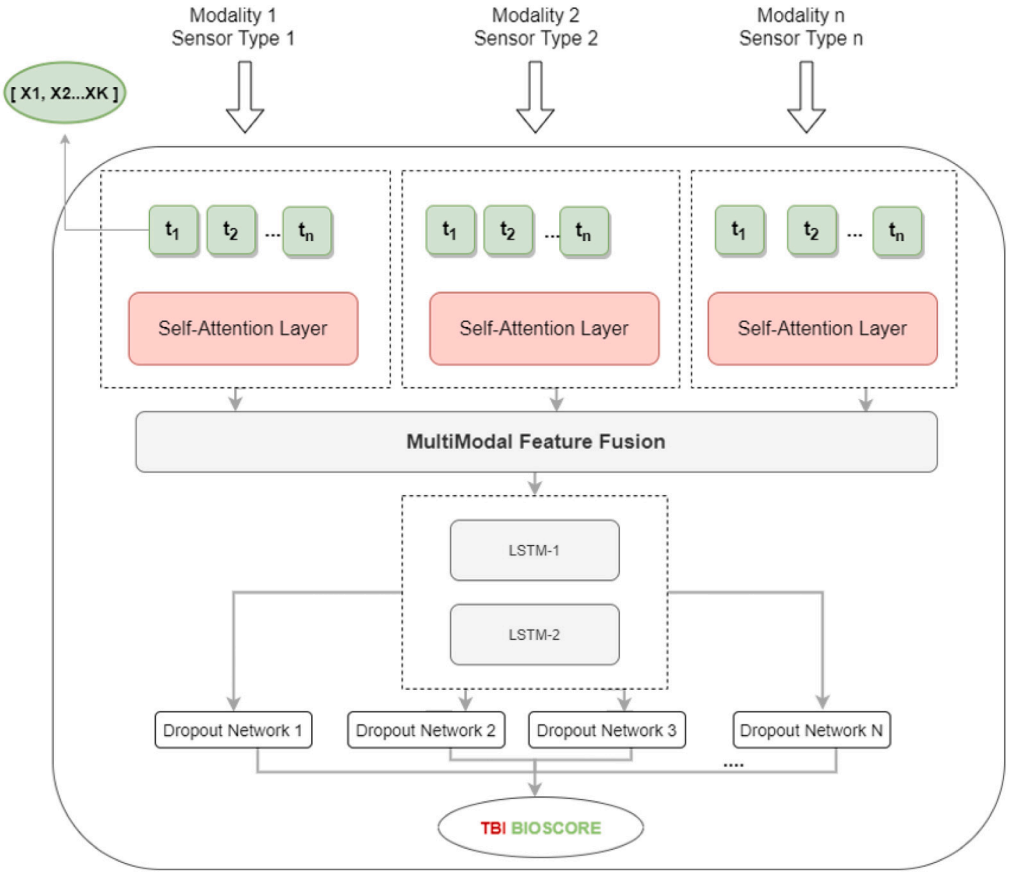

Fig. 5. TBI Bioscore generation architecture.

4.5. TBI bioscore generation

Once the network is trained, Monte Carlo dropout is applied to obtain N networks mimicking the trained network as illustrated in Fig. 2. Thereafter, the predictions from these N networks are ensembled. By computing mean and variance on these assimilated predictions we get an estimate of the model's posterior distribution and uncertainty pertaining to a data point as shown in

$$\text{Uncertainty} = \frac{1}{N}\sum_{i=0}^{N}[f_n^{di}(x) - \underbrace{\left(\frac{1}{N}\sum_{i=0}^{N} f_n^{di}(x)\right)}_{\text{Posterior mean}}]^2 \tag{10}$$

Using the calculated uncertainty, we calculate the certainty of the prediction for every subject:

$$\text{certainty} = 1 - \text{Uncertainty} \tag{11}$$

For predictions belonging to the class of TBI subjects, the certainty (bioscore) in the range (0.5, 1] is scaled as follows:

- A TBI subject predicted with 100% certainty, will have a bioscore closer to 1.
- A TBI subject predicted with lower certainty will have a bioscore closer to 0.5 For predictions belonging to the class of non-TBI subjects, scale certainty is in range [0, 0.5], the bioscore is scaled as follows:
  - A non-TBI subject, predicted with 100% certainty, will have a bioscore closer to 0.
  - A non-TBI subject, predicted with low certainty, will have a bioscore closer to 0.5

This TBI bioscore is then computed daily and used to monitor a subject's health status over time.

5. Evaluation and results

The evaluation of the proposed passive TBI bioscore generation framework is done using Fbeta measures, balanced accuracy, true positive rate, true negative rate, false positive rate, and false negative rate. The target label is TBI (Yes/No), a binary target

Table 2
Comparing different attention mechanisms and Neural Network architectures.

Attention mechanisms	Classifier	BA	TPR	TNR	FBeta, Beta = 2
No attention	MLP	0.78	0.667	0.893	0.8
	LSTM	0.802	0.712	0.892	0.691
	RNN	0.70	0.50	0.914	0.50
	Stacked LSTM	0.819	0.667	0.971	0.727
Attention	Stacked LSTM	0.823	0.731	0.915	0.81
Self-attention	**Stacked LSTM**	**0.902**	**0.833**	**0.971**	**0.833**

*TPR–True Positive Rate # TNR–True Negative Rate # BA–Balanced Accuracy.

whose value is 0 when the subject does not have TBI, and 1 for subjects with TBI. Because the dataset was imbalanced, the focus when evaluating the model was to improve upon the true positive rate and accurately predict TBI instances. The approach was finally evaluated using the following metrics:

- F-beta Measure is a generalized F-measure that adds 'beta' as a configuration parameter. A smaller beta value (0.5) gives more weight to precision and less to recall, whereas a larger beta value, such as 2.0 gives less weight to precision and more weight to recall. It is useful in our case because recall is an important metric (correctly classifying TBI instances, therefore we used a beta score of 2.0.

$$F - \text{Beta} = \frac{\left(1 + \text{beta}^2\right) * \text{Precision} * \text{Recall}}{\text{beta}^2 * \text{Precision} + \text{Recall}} \tag{12}$$

- F-measure is calculated as the harmonic mean of the model's precision and recall, giving each the same weighting. It is basically a calculation of the ratio of correctly predicted positive samples divided by the total number of positive examples that could be predicted.

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

- True Positive Rate (TPR): is calculated as:

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

- True Negative Rate (TNR): is calculated as:

$$TPR = \frac{TP}{TP + FN} \tag{15}$$

- Balanced Accuracy (BA): is calculated as the average of the proportion corrects of each class individually. It is a good metric for imbalanced datasets:

$$BA = \frac{TPR + TNR}{2} \tag{16}$$

A time frame of 24 h of data, one day after the injury, with a window size of 2 h and 50% overlap was selected as the optimal feature creating frame. The models are evaluated using RNN and LSTM sequence to sequence architectures with combinations of self-attention and simple attention layer. LSTM outperformed RNN and was selected as a baseline model to improve the results upon Since there are multiple modalities of data, experiments using early-fusion vs late fusion techniques were conducted. It was noted that some modalities affect the result more than others, and therefore an attention mechanism was employed to weight these modalities and create uniform representations. Self-attention mechanism was proven to significantly improve the True positive Rate and was selected as a module in the final architecture. Further hyperparameter tuning was done to achieve to an optimal result. The LSTM Network to create a deeper network and to handle class imbalance every misclassification was penalized in the loss function, by assigning class weights, as a ratio of negative classes to positive classes. The learning rate used was 0.0001, the optimizer used was Adam while the loss function that best suits our model is binary cross-entropy. To make the model even more robust, Monte Carlo Dropout ensemble methods were used on both training and testing and the average score of 50 model probabilities was used as a final score (see Table 2).

Feature Importance: Fig. 6, visualizes the ranked feature importances for all weighted fused modalities. The most important features are the mobility features such as pedometer, accelerometer and magnetomer. High pedometer kurtosis value indicates that there are outliers in the dataset, in our case the TBI instances which compared to non-TBI instances have significantly lower distance and steps traveled.

Statistical Differences feature values of TBI vs. non-TBI subjects: When analyzing features that show how active and how much does a subject move around, by calculating the sum distance and the average steps throughout the day after an injury, we can see significantly different patterns for TBI vs Non-TBI users as shown in Fig. 7. By examining the general distribution of the mean in the sensors, we can observe mobility differences in the mobility patterns of TBI vs. Non-TBI Subjects. As Figs. 8 and 9 show, TBI users show less movement the day after the injury.

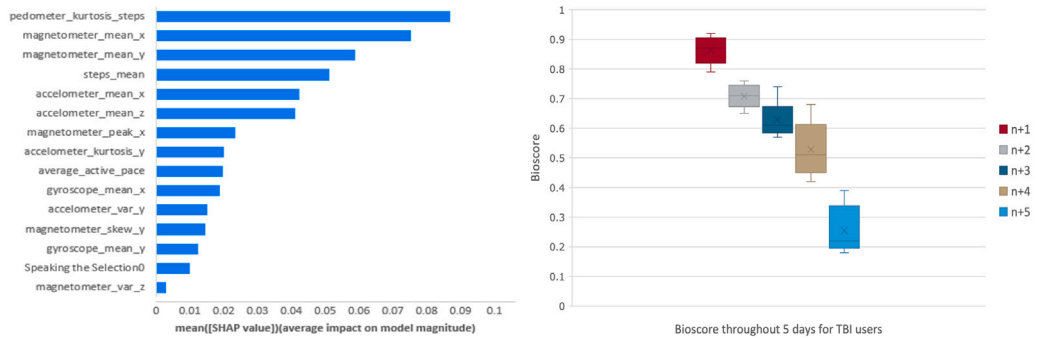

Fig. 6. Feature importance for all modalities combined [left] and boxplots showing bioscores on days n + 1 to n + 5 for all users who had an injury on day n[right].

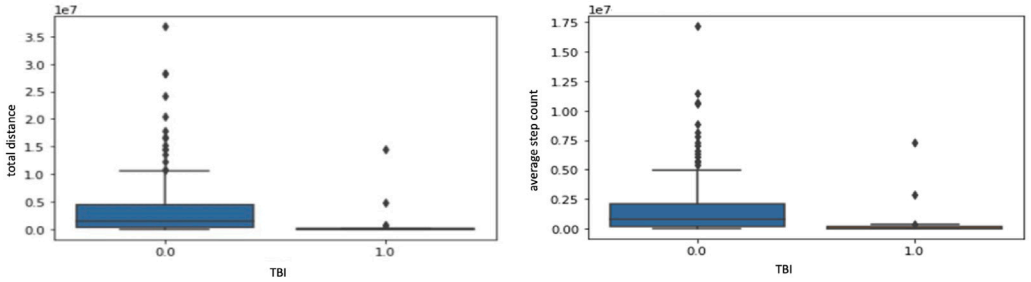

Fig. 7. Statistical differences in total distance feature values of TBI vs. non-TBI subjects.

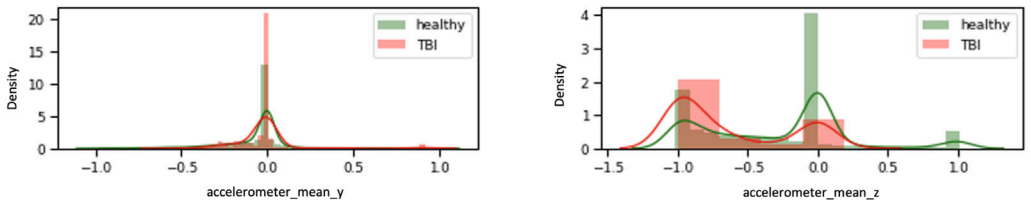

Fig. 8. Statistical differences in accelerometer mean features of TBI vs. non-TBI subjects.

Bioscore: The results found in this work demonstrate that the bioscore concept is feasible and can that a glanceable display of color-coded bioscores can assist overworked healthcare workers in the identification and the timely treatment of ill individuals in a large cohort, as well as in recovery monitoring. As shown in Fig. 6, when analyzing the bioscores of users who have had an injury, every day for 5 days post-injury, an interesting and valuable emerging pattern is observed. On the first day after injury, there is low variance as almost all the subjects are accurately predicted to have been injured, and in the following 4 days, we note that the bioscore starting to decrease, for some subjects faster than others. This shows how even a mild concussion interferes with user's ability to function as normal, which is reflected in his/her mobility patterns and phone usage behavior.

Bioscore Dashboard: The glanceable dashboard visually summarizes the well-being of a large population of users in a workplace, organization, sporting event, military unit, and any large communities where overall monitoring of their health is essential. (as shown in Fig. 6). This dashboard also facilitates a granular analysis of every users bioscore providing valuable historical data on the subjects' ailment trajectory in the prior 5 days and one current or future bioscore, depending on the ailment at hand. It provides a good interface of the TBI detection model for bioscore generation, and it can be generalized to other ailments as well (see Fig. 10).

6. Conclusion and future work

TBI causes millions of individuals distress and it can lead to significant motor, cognitive and emotional deficit. In order to facilitate continuous, passive, population-level ailment monitoring, we proposed a bioscore framework that enables a small team of medical personnel to detect ailment occurrence early and monitor the recovery trajectory of large numbers of patients. When there are subjects diagnosed with chronic diseases such as Traumatic Brain Injuries, cancer, surgery, or patients with mobility issues and elderly that involve an extended recovery phase, health deterioration is possible. Without adequate population-level

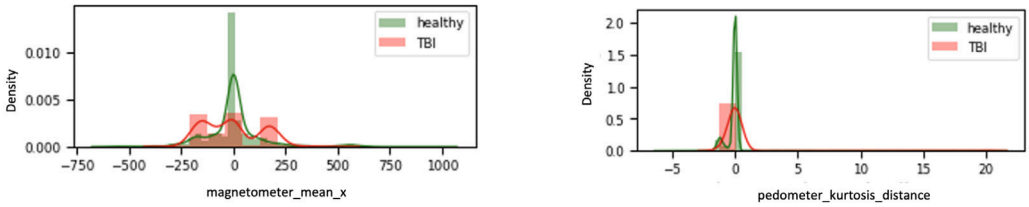

Fig. 9. Statistical differences in magnetometer mean and kurtosis features of TBI vs. non-TBI subjects.

Fig. 10. Dashboard illustrating the bioscore of users.

recovery monitoring tools, regressing patients are sometimes discovered and re-hospitalized too late. Our novel bioscore estimates the probability (certainty) (0–1) that a subject has a given ailment based on their smartphone-sensed behavior, social interactions, mobility and communication patterns. We illustrate this novel concept by passively phenotyping smartphone users, generating their TBI bioscore as the probability that a subject has concussion or TBI, after an injury. 48 h of post-injury data was analyzed using Stacked LSTMs with Self-attention for feature fusion, to perform binary classification of whether the subject has concussion. Then their Bioscore is calculated by using Monte Carlo Dropout to calculate the model's uncertainty that is further converted to a certainty probability measure. The proposed model achieves a balanced accuracy of 90.2% with a True Positive Rate in correctly identifying TBI subjects of 83.3%. We successfully calculated and visualized the distribution of users' TBI bioscores on a glanceable dashboard in the days following an injury. Our bioscore framework facilitates recovery trajectory monitoring for patients in the recovery stage, as well as early detection of infected users, enabling reducing contagion. Since the framework supports multiple modalities, in the future it can be extended to use not only smartphone sensor data of mobility traces, but also audio or even images simultaneously for different ailment predictions. This will enable the framework to be flexibly used to monitor these various ailments and curb casualties.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Emmanuel Agu reports financial support was provided by Worcester Polytechnic Institute.

Data availability

The data that has been used is confidential.

Acknowledgments

This material is based on research funded by DARPA, USA under agreement number FA8750-18-2-0077. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

Baldwin, G., Breiding, M., & Sleet, D. (2016). Using the public health model to address unintentional injuries and TBI: A perspective from the centers for disease control and prevention (CDC). *NeuroRehabilitation*, *39*(3), 345–349.

Bhavani, G., & Amponsah, C. T. (2017). M-Score and Z-Score for detection of accounting fraud. *Accountancy Business and the Public Interest*, *1*(1), 68–86.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In *Proc. ICML* (pp. 1613–1622). PMLR.

Creber, R. M. M., Maurer, M. S., Reading, M., Hiraldo, G., Hickey, K. T., & Iribarren, S. (2016). Review and analysis of existing mobile phone apps to support heart failure symptom monitoring and self-care management using the mobile application rating scale (MARS). *JMIR MHealth and UHealth*, *4*(2), Article e5882.

Dewart, G., Corcoran, L., Thirsk, L., & Petrovic, K. (2020). Nursing education in a pandemic: Academic challenges in response to COVID-19. *Nurse Education Today*, *92*, Article 104471.

Fu, B., Damer, N., Kirchbuchner, F., & Kuijper, A. (2020). Sensing technology for human activity recognition: A comprehensive survey. *IEEE Access*, *8*, 83791–83820.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc ICML* (pp. 1050–1059). PMLR.

Graves, A. (2011). Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*, *24*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Kingma, D. P., Salimans, T., & Welling, M. (2015). Variational dropout and the local reparameterization trick. *Advances NIPS*, *28*, 2575–2583.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv preprint arXiv:1612.01474.

Ling, S., & Raine, A. (2018). The neuroscience of psychopathy and forensic implications. *Psychology, Crime & Law*, *24*(3), 296–312.

Lipsmeier, F., Taylor, K. I., Kilchenmann, T., Wolf, D., Scotland, A., Schjodt-Eriksen, J., et al. (2018). Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 parkinson's disease clinical trial. *Movement Disorders*, *33*(8), 1287–1297.

Liu, K., Li, Y., Xu, N., & Natarajan, P. (2018). Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730.

Louizos, C., & Welling, M. (2017). Multiplicative normalizing flows for variational bayesian neural networks. In *Proc ICML* (pp. 2218–2227). PMLR.

Mariakakis, A., Baudin, J., Whitmire, E., Mehta, V., Banks, M. A., Law, A., et al. (2017). PupilScreen: using smartphones to assess traumatic brain injury. *Proceedings of the ACM IMWUT*, *1*(3), 1–27.

Murthy, S. N., Asani, F., Srikanthan, S., & Agu, E. (2020). DeepSEAS: Smartphone-based early ailment sensing using coupled LSTM AutoEncoders. In *2020 IEEE int'l conf. big data* (pp. 4911–4918). IEEE.

Saeb, S., Lattie, E. G., Kording, K. P., & Mohr, D. C. (2017). Mobile phone detection of semantic location and its relationship to depression and anxiety. *JMIR MHealth and UHealth*, *5*(8), Article e7297.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).

Torous, J., & Keshavan, M. (2018). A new window into psychosis: The rise digital phenotyping, smartphone assessment, and mobile monitoring. *Schizophrenia Research*, *197*, 67–68.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, *53*(8), 5929–5955.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wang, J., Wang, Y., Wei, C., Yao, N., Yuan, A., Shan, Y., et al. (2014). Smartphone interventions for long-term health management of chronic diseases: an integrative review. *Telemedicine and E-Health*, *20*(6), 570–583.

Wilson, L. (2019). MA patients' readmission rates higher than traditional Medicare, study finds. *HealthcareDive*, Retrieved May, 20, 2021.

Xu, L., Zhang, Z., Liu, Q., Zhou, B., Liu, Y., Xiang, Q., et al. (2018). Validation of CPS+ EG, neo-bioscore, and modified neo-bioscore staging systems after preoperative systemic therapy of breast cancer: Protocol of a retrospective multicenter cohort study in China. *Thoracic Cancer*, *9*(11), 1565–1572.