Mining Standardized Neurological Signs and Symptoms Data for Concussion Identification

Janani Venugopalan, *Georgia Institute of Technology*
Michelle LaPlaca, *Emory University*
Dongmei Wang, *Emory University*

Journal Title: 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)

Volume: Volume 2017

Publisher: IEEE | 2017-01-01, Pages 285-288

Type of Work: Article | Post-print: After Peer Review

Publisher DOI: 10.1109/BHI.2017.7897261

Permanent URL: https://pid.emory.edu/ark:/25593/vp6j2

Final published version: http://dx.doi.org/10.1109/BHI.2017.7897261

HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2020 July 22.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2017 February ; 2017: . doi:10.1109/bhi.2017.7897261.

Mining standardized neurological signs and symptoms data for concussion identification

Janani Venugopalan[1] **[Student Member, IEEE]**, **Michelle C. LaPlaca, Ph.D.**[2], **May. D. Wang, Ph.D.**[3] **[Senior Member, IEEE]**

[1]Wallace H. Coulter School of Biomedical Engineering, at Georgia Institute of Technology and Emory University, Atlanta, GA, USA

[2]Wallace H. Coulter School of Biomedical Engineering, Petit Institute for Bioengineering and Bioscience at Georgia Institute of Technology and Emory University, Atlanta, GA, USA

[3]Biomedical Engineering Department, Electrical and Computer Engineering, Winship Cancer Institute, Parker H. Petit Institute of Bioengineering and Biosciences, Institute of People and Technology, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Abstract

The Centers for Disease Control estimate that 1.6 to 3.8 million concussions occur in sports and recreational activities annually. Studies have shown that concussions increase the risk of future injuries and mild cognitive disorders. Despite extensive research on sports related concussion risk factors, the factors which are most predictive of concussion outcome and recovery time course remain unknown. In order to overcome the issue of physician bias and to identify the factors which can best predict concussion diagnosis, we propose a multi-variate logistic regression based analysis. We demonstrate our results on a dataset with 126 subjects (ages 12–31). Our results indicate that among 322 features, our model selected 27–29 features which included a history of playing sports, history of a previous concussion, drowsiness, nausea, trouble focusing as measured by a common symptom list, and oculomotor function. The features picked using our model were found to be highly predictive of concussions and gave a prediction performance accuracy greater than 90%, Matthews correlation coefficient greater than 0.8 and the area under the curve greater than 0.95.

I. INTRODUCTION

According to the Centers for Disease Control 1.6 to 3.8 million concussions occur in sports related activities annually [1]. Concussions and sports related head injuries are one of the most common causes of traumatic brain injuries [2], with football injuries accounting for nearly 20% of the concussions [3]. Studies have shown that patients with concussions often present symptoms such as light-headedness, vertigo, cognitive and memory dysfunction, tinnitus, issues with vision, concentration difficulties, nausea, vomiting, headaches, or balance disorders[4]. Symptoms such as sleep irregularities, fatigue, personality changes, an inability to perform usual daily activities, depression, or lethargy [4] may also manifest as

Corresponding author: MD Wang maywang@gatech.edu.

delayed symptoms. Studies have also shown that patients with a concussion are at a relatively higher risk of future concussions and mild cognitive impairments [1].

Concussion is an extremely heterogeneous condition because of the varied injury mechanisms (i.e. location and force of impact) and inherent population heterogeneity (i.e. premorbid conditions, general health), leading to a myriad of symptoms that are—in general—difficult to detect. The Institute of Medicine (IOM) reports that there is limited reporting of mTBI/concussion, contributing to a gap in clinical evidence for accurate diagnosis methods and in predicting the course of recovery and impact on health outcomes [5]. Relatively little is known about the factors that lead to prolonged post-concussive symptoms. There is evidence that there is a relationship between brain injury history (including repeat concussions and interval between concussions) and recovery course and degree of persistent symptoms [6, 7]. Because of the vast number of variables that can contribute to the outcome, it is necessary to use robust analysis techniques that can uncover relationships among inherent variables such as age and medical history and neurological function ranging from vestibular to neurocognitive. Besides knowing which factors contribute most to the outcome, such techniques have the capability of identifying factors that differentiate concussion symptoms from normal physiological variability.

Despite recommendations to assess concussion in multiple neurological domains, including neurocognitive, balance, and oculomotor, it is largely unknown which factors accurately classify concussion. In order to overcome these challenges and to discover factors which are most indicative of concussion we propose a data driven analysis of clinically used surveys and tests using a multi-variate logistic regression. We demonstrate our results on a dataset consisting of 126 subjects who completed standardized symptom questionnaires and clinical tests spanning neurological domains. Potentially patterns between outcome measures and variables will provide the basis to optimize prediction models.

We structure the remainder of this article as follows. First, a short description of our data source is followed by a detailed description of the preprocessing and data mining approaches in section II. Evaluation, results, and discussion are presented in section III. Finally, the conclusion and future directions of are summarized in section IV.

II. Methods

This is a retrospective study for the classification of subjects into patients with concussions and those without concussions. We used filter-based and sequential feature selection followed by multi-variate logistic regression for the classification analysis. The steps (Fig. 1) include data pre-processing, feature selection, classification, and evaluation. The methods are detailed below:

A. Data Source & Data Pre-Processing

This is a retrospective data analysis using data from 126 subjects ages 17.12 ± 4.03 (mean ± SD) (range = 12;31). The data included 58 female participants (13 concussed) and 68 male participants (13 concussed). We collected our data by recruiting subjects ages 12–31 who had a confirmed concussion and had sought clinical care for persistent concussion symptoms

(all within 6 months of diagnosed concussion). Control subjects were recruited from the community and were age- and fitness-matched, with no concussion in the previous 6 months.

The same clinical tests were applied to both cohorts and included a screening questionnaire with demographics, activity history, and relevant medical history, symptom questionnaire (Post-Concussion Symptom Scale, PCSS), dizziness assessment (Simulator Sickness Questionnaire, SSQ), oculomotor function (Vestibular/Ocular Motor Screening, VOMS)[14], and balance (Sway index and Balance Error Scoring System, BESS, firm and foam). Symptoms were also assessed following administration of VOMS. The data set included a total of 383 variables of which 164 were inclusion/exclusion criteria and free text variables and hence were not included for this study. Of the remaining 219 variables (Table I), 193 were continuous, 12 categorical and 14 binary. We converted the categorical variables into a set of binary features using one hot encoding, and used the binary variables and continuous variables as they were. This gave us a total of 332 features with 0.53% missing data for classification analysis. We imputed the missing data using an alternating least square based PCA [8]. Following data imputation, we performed feature selection followed by classification.

B. Feature Selection

As mentioned above, we have 126 subjects (100 controls, 26 concussions) and 332 features. In order to obtain a good predictive performance, we used feature selection techniques to find the features most indicative of concussions. For feature selection, two approaches can be adopted a) filter-based methods and b) wrapper based methods. Filter-based methods include techniques such as mRMR [9], Wilcoxon ranksum test [10], and differential expression [10], which use the data and labels to find the most influential features. Wrapper based techniques include methods such as sequential forward selection, sequential backward selection, and sequential backward-forward selection [11]. These techniques use the classifiers to select the feature combinations with the most predictive performance.

In our analysis, we test two filter-based methods differential expression (DE) and Wilcoxon ranksum test (WR). At this stage we retained all the features which were differentially expressed when DE was used. Similarly, we used all the features from WR which had a p value $p \leq 0.05$. Following this, the features selected using the filter-based techniques were then further selected using sequential forward selection (SFS) techniques with multivariate logistic regression as the classification method. We selected a total of 50 features using SFS techniques.

C. Classification using Multivariate Logistic Regression

For classification of the subjects into 2 groups (those with concussion and those without concussion), we used logistic regression. Logistic regression (LR) is the most commonly used model in healthcare. It is used to predict the probability of the outcome of the study by identifying features which are most predictive of the outcome (concussion presence in his case). A model is trained using a feature set = [x1, x2, x3…..xn](where n = # of features

selected). Logistic regression model calculates the probability with which a specific outcome occurs as opposed to other outcomes using the equation given by

$$h_\theta = \frac{e^{\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots \theta_n x_n}}{1 + e^{\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots \theta_n x_n}} \quad (1)$$

The outcome group (y) is assumed to be true (1) when the probability exceeds 0.5. The values of = (θ0, θ1, θ2 … θn) is trained from the training data set by minimizing a cost function given in by the following equation (where m = # of samples)

$$Cost\,Function = \frac{1}{m}\sum_{i=1}^{m}\left[-y^i log\left(h_\theta\left(\vec{x}^i\right)\right)\right| \quad \left|-\left(1-y^i\right)log\left(1-h_\theta\left(\vec{x}^i\right)\right)\right] \quad (2)$$

It is a binary classifier which classifies one group against another. The assumption made by logistic regression is that the outcomes must be discrete.

We evaluated our models using 10 fold cross-validation. The evaluations metrics reported were Matthew's correlation coefficient (MCC), area under the curve (AUC) and accuracy.

III. Results & Discussion

To ensure that the features selected have predictive power we first generated a PCA plot. After that we performed classification with two options for filter-based feature selection (1. Wilcoxon ranksum test, 2. Diffe rential expression). The results of the classification analysis are given below:

A. Test for Separation in the Data

Before we performed classification and obtained features indicative of classification, we performed principal component analysis (PCA) to ensure that the data has separation and the features used for classification have predictive power. The results from PCA (Fig. 2) indicate a good separation between the two clusters with 125 principal components. Since the features gave a good separation, we proceeded to perform the classification and to discover features most indicative of concussions.

B. Classification Results: Wilcoxon Ranksum Test

As mentioned above we first performed filter-based feature selection followed by SFS and classification using logistic regression. The results of logistic regression after Wilcoxon ranksum as the filter-based techniques (Fig. 3) indicate that the performance of the classification drops after 27 features. The MCC, AUC and accuracy values were all found to be above 0.8, 0.9 and 0.95 respectively.

The top features obtained using this method (Table II) include symptoms such as fatigue, dizziness, nausea, mechanism of injury, and oculomotor scores.

C. Classification Results: Differential Expression

The results of logistic regression after DE as the filter-based techniques (Fig. 4) indicate that the performance of the classification drops after 29 features. The MCC, AUC and accuracy values were all found to be above 0.8, 0.9 and 0.95 respectively.

The top features obtained using this method (Table II) include mechanism of injury, total PCSS, fatigue and nausea following dizziness questionnaire (SSQ), oculomotor function (VOMS scores), Balance (Sway index) and nausea due to SSQ.

The common features between both the methods were 23 and included features such as previous history of sports in the past 6 months, recent concussions, drowsiness, total PCSS, % normal, post SSQ focusing, SSQ related nausea, Oculomotor scores related to SSQ, and VOMS scores.

The model and our data indicate that the features such as fatigue after SSQ testing was higher in patients with concussion. Similarly, the VOMS tests also had high distinguishing features for concussion. We found that these features picked up by our model correlated well with those found in the literature [12–14]. In addition, we also show that other factors such as the mechanism of injury, Sway indices and total PCSS are also relevant. Such insights could help that identify concussion patients. However, from our analysis we found that the accuracy and MCC drop drastically after 27–29 features. In the future we will investigate this effect by conducting our analysis on a larger patient population.

In addition, our current method uses logistic regression with standard feature selection techniques such as ranksum test and differential expression followed by sequential feature selection. Feature selection was done to prevent overfitting and to obtain a meaningful list of features. This is important, since the number of features is larger than the number of samples, making the system prone to overfitting. In the future we will compare our models with more complicated models such as support vector machines (SVM), Random Forests, and decision trees for obtain further improvement in results on a larger population.

IV. Conclusion

The factors which are most predictive of concussions remain unknown. Current concussion research often focusses on single diagnostic domains rather than the combination of multiple modalities that can better predict concussion [15]. In addition, since there exists a vast number of factors that can be affected as a result of concussion, it is necessary to use robust analysis techniques that can uncover relationships among inherent features such as age, medical history and neurological function ranging from vestibular to neurocognitive. This study was designed to overcome these issues as well as the potential physician bias in order to find specific factors that best identify concussions. We demonstrate our results using multi-variate logistic regression based analysis on a dataset of 126 subjects recruited from concussion clinics(age 12 – 31 years). In addition to PCSS, SSQ, and VOMS provoking symptoms, our results also identified other features such as previous history of sports, recent concussions, and individual symptoms (imbalance, drowsiness, sleepiness) which were among the common list of top 23 features. Our model gave a prediction performance

accuracy greater than 90%, Matthews correlation coefficient greater than 0.8 and the area under the curve greater than 0.95.

However, despite the good performance, our results suffer from challenges such as the use of a small dataset which could have had overfitting and the lack of external data set for validation. In the future we will extend this study by testing our models on a larger and an external dataset. In addition, we plan to include other types of tests for analysis.

References

[1]. Daneshvar DH, Nowinski CJ, McKee A, and Cantu RC, "The Epidemiology of Sport-Related Concussion," Clinics in sports medicine, vol. 30, pp. 1–17, 2011. [PubMed: 21074078]

[2]. (2016). Traumatic Brain Injury (TBI). Available: http://www.asha.org/public/speech/disorders/TBI/

[3]. Cantu RC, "RETURN TO PLAY GUIDELINES AFTER A HEAD INJURY," Clinics in Sports Medicine, vol. 17, pp. 45–60, 1/1/ 1998. [PubMed: 9475970]

[4]. Leclerc S, Lassonde M, Scott Delaney J, Lacroix VJ, and Johnston KM, "Recommendations for Grading of Concussion in Athletes," Sports Medicine, vol. 31, pp. 629–636, 2001. [PubMed: 11475324]

[5]. Graham R, Rivara FP, Ford MA, and Spicer CM, Sports-related concussions in youth: improving the science, changing the culture: National Academies Press, 2014.

[6]. Covassin T, Moran R, and Wilhelm K, "Concussion symptoms and neurocognitive performance of high school and college athletes who incur multiple concussions," The American journal of sports medicine, vol. 41, pp. 2885–2889, 2013. [PubMed: 23959963]

[7]. Eisenberg MA, Andrea J, Meehan W, and Mannix R, "Time interval between concussions and symptom duration," Pediatrics, vol. 132, pp. 8–17, 2013. [PubMed: 23753087]

[8]. Kiers HA and ten Berge JM, "Alternating least squares algorithms for simultaneous components analysis with equal component weight matrices in two or more populations," Psychometrika, vol. 54, pp. 467–473, 1989.

[9]. Ding C and Peng H, "Minimum redundancy feature selection from microarray gene expression data," J Bioinform Comput Biol, vol. 3, pp. 185–205, 4 2005. [PubMed: 15852500]

[10]. Saeys Y, Inza I, and Larrañaga P, "A review of feature selection techniques in bioinformatics," bioinformatics, vol. 23, pp. 2507–2517, 2007. [PubMed: 17720704]

[11]. Pudil P, Ferri F, Novovicova J, and Kittler J, "Floating search methods for feature selection with nonmonotonic criterion functions," in In Proceedings of the Twelveth International Conference on Pattern Recognition, IAPR, 1994.

[12]. Chen Y-C, "Postural and cognitive precursors of post-bout motion sickness and concussion-related symptoms in boxers," 2014.

[13]. Collins MW, Kontos AP, Reynolds E, Murawski CD, and Fu FH, "A comprehensive, targeted approach to the clinical care of athletes following sport-related concussion," Knee Surgery, Sports Traumatology, Arthroscopy, vol. 22, pp. 235–246, 2014.

[14]. Mucha A, Collins MW, Elbin R, Furman JM, Troutman-Enseki C, DeWolf RM, et al., "A brief vestibular/ocular motor screening (VOMS) assessment to evaluate concussions preliminary findings," The American journal of sports medicine, p. 0363546514543775, 2014.

[15]. McCrory P, Meeuwisse WH, Kutcher JS, Jordan BD, and Gardner A, "What is the evidence for chronic concussion-related changes in retired athletes: behavioural, pathological and clinical outcomes?," British journal of sports medicine, vol. 47, pp. 327–330, 2013. [PubMed: 23479493]

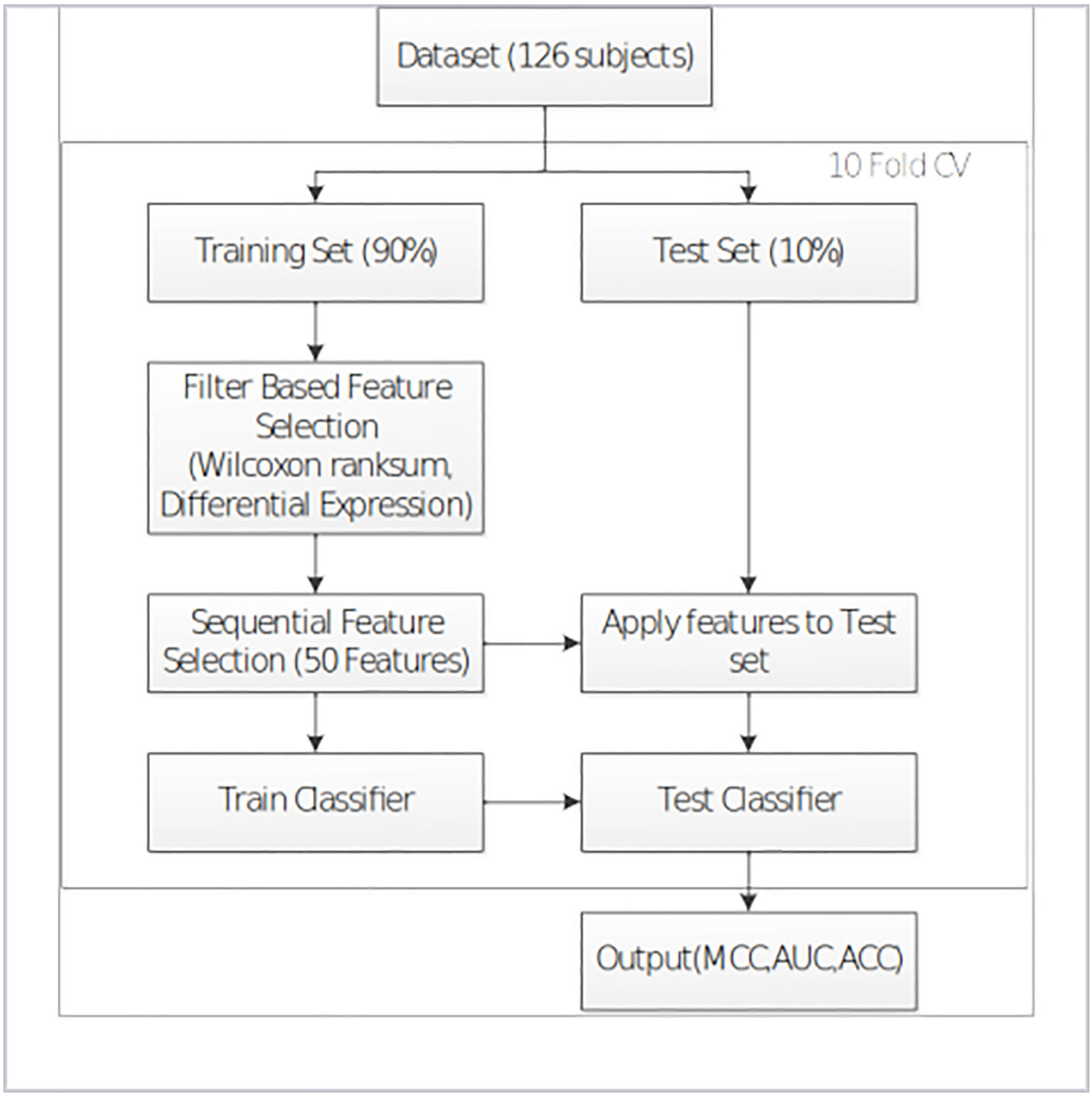

Figure 1.
Steps in data mining

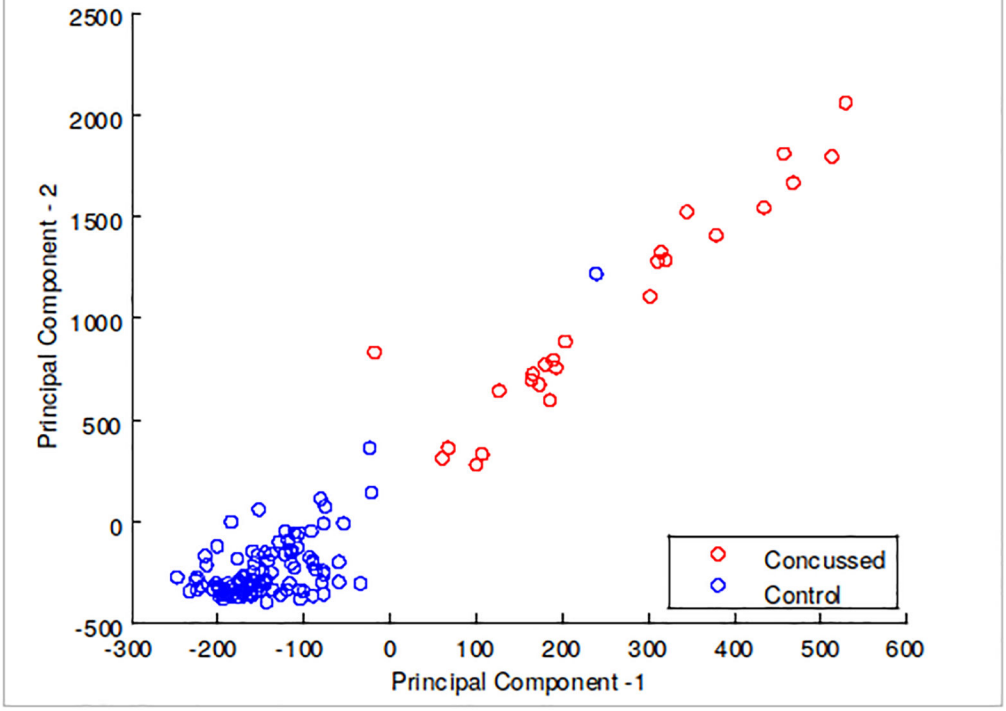

Figure 2.
Results of principal component analysis

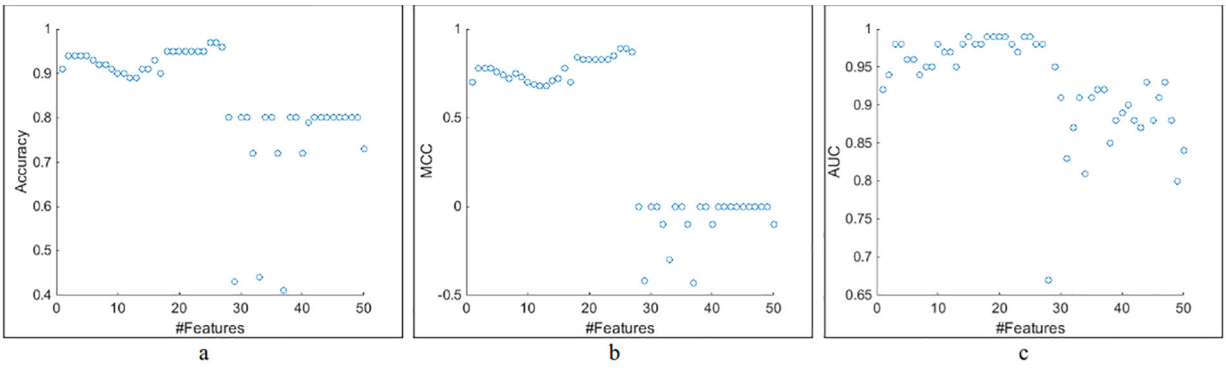

Figure 3.

Wilcoxon ranksum test results with the feature numbers being increased to a max value of 50 a) Accuracy values b) MCC values c) AUC values. These results indicate that # features to be used after SFS = 27

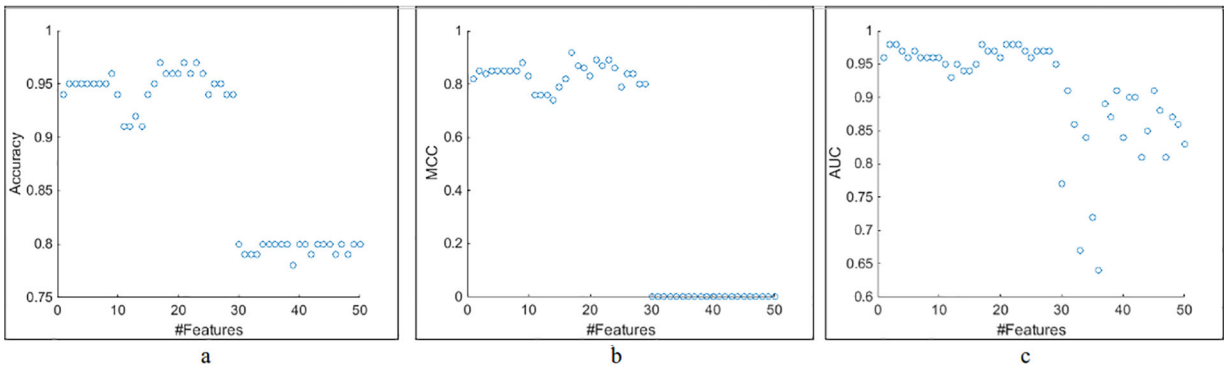

Figure 4.

Differential expression test results with the feature numbers being increased to a max value of 50 a) Accuracy values b) MCC values c) AUC values. These results indicate that # features to be used after SFS = 29

TABLE I.

Features Types in Dataset

TABLE II. Feature Type	**TABLE III. Example**	**TABLE IV. # Variables**	**TABLE V. # Features**
TABLE VI. Binary	TABLE VII. Gender, Cohort (Label), Presence of Impact, Medication Taken, Recent Concussion	TABLE VIII. 14+1 label	TABLE IX. 14 + 1 label
TABLE X. Categorical	TABLE XI. Race (8), Handedness(2), PH-Sport(11), Sport involved(11), Position(17), Context(7), Mechanism(10), FAM migraine individual(5)	TABLE XII. 12	TABLE XIII. 125
TABLE XIV. Quantitative	TABLE XV. Height, weight, dizziness, light sensitive, fatigue, % normal.	TABLE XVI. 193	TABLE XVII. 193
TABLE XVIII. **Total**	TABLE XIX.	TABLE XX. **219**	TABLE XXI. **332**

TABLE XXII.

TABLE XXIII.

Top 10 Features after Classification

SN	Features Wilcoxon Test	Features Differential Expression
1	Post SSQ Fatigue	Unknown Mechanism of Injury
2	Unknown Mechanism of Injury	Total PCSS
3	Total PCSS	Post SSQ Fatigue
4	Sway index Avg	VOMS headache (HA) vertical saccades
5	VOMS HA svertical saccades	VOMS HA VOR vertical
6	VOMS HA convergence	Sway index Avg
7	VOMS HA baseline	Total SSQ Oculomotor
8	VOMS HA VOR vertical	VOMS HA convergence
9	VOMS HA horizontal saccades	VOMS HA baseline
10	VOMS HA VOR horizontal	VOMS HA horizontal saccades