Fall 2020

The application of machine learning models in the concussion diagnosis process

Sujit Subhash

THE APPLICATION OF MACHINE LEARNING MODELS IN THE CONCUSSION DIAGNOSIS PROCESS

by

SUJIT SUBHASH

A THESIS

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

APPLIED MATHEMATICS

2020

Approved by:

Gayla R. Olbricht, Advisor
Robert Paige
Donald Wunsch II

PUBLICATION THESIS OPTION

This thesis consists of the following article, formatted in the style used by the Missouri University of Science and Technology:

Paper I, found on pages 13–34, is intended for submission to Computational Intelligence in Healthcare and E-health (IEEE CICARE) Symposium.

ABSTRACT

Concussions represent a growing health concern and are challenging to diagnose and manage. Roughly four million concussions are diagnosed every year in the United States. Although research into the application of advanced metrics such as neuroimages and blood biomarkers has shown promise, they are yet to be implemented at a clinical level due to cost and reliability concerns. Therefore, concussion diagnosis is still reliant on clinical evaluations of symptoms, balance, and neurocognitive status and function. The lack of a universal threshold on these assessments makes the diagnosis process entirely reliant on a physician's interpretation of these assessment scores. This study aims to show that the implementation of machine learning models can be beneficial to the concussion diagnosis process. While studies on machine learning applications for traumatic brain injuries are gaining traction, previous studies have primarily relied on neuroimaging metrics. The few that used clinical assessment tests have employed only univariate models. This study explores the use of multiple assessment scores in the models and evaluates the importance of each assessment score from the clinical tests. A comprehensive predictive modeling approach was conducted with a number of candidate models and subsampling techniques being evaluated. The findings in this research demonstrate the potential benefits of machine learning models to identify concussed and non-concussed subjects at a 24-48-hour post-injury time point. The results also suggest that not all clinical assessment test scores are of equal importance.

ACKNOWLEDGMENTS

I am greatly indebted to my advisor, Dr. Gayla Olbricht, for giving me the opportunity to pursue my master's degree under her guidance. Her passion for the field of statistics and interests in data science was instrumental in shaping my own, and for this, I am forever grateful. This thesis and my master's degree would not have been possible without the patience and trust she showed in supporting my research. I am thankful to Dr. Robert Paige and Dr. Donald Wunsch II, for bringing their vast experience and knowledge to my advisory committee and guiding me through the challenges I faced in my research.

I am grateful to my family, especially my parents Dr. Subhash Jacob and Anie Subhash, for the love and encouragement. I am thankful to Dr. Tejaswi Materla for motivating and supporting me throughout my master's degree and guiding me through my challenging times.

TABLE OF CONTENTS

Page

SECTION

LIST OF ILLUSTRATIONS

PAPER I Page

LIST OF TABLES

SECTION Page

PAPER I

1. INTRODUCTION

While millions of concussions are diagnosed every year, of particular concern are many concussions that go underreported, or worse undiagnosed. Concussed individuals are susceptible to more severe consequences if their injury is neglected [1]. The series of biological developments that follow a concussion creates a vulnerability for a second injury and can lead to severe neurodegeneration [2].

Although research into concussions dates back to the late 19th century [3], there is not a precise definition of concussion. Clinical evaluations such as SAC, SCAT5, BESS, BSI-18, and ImPACT have been developed to measure the acute symptoms typical of a concussion. However, there is no universal threshold on these clinical tests to characterize or identify a definite concussion. Advanced research into neuroimaging metrics and blood biomarkers have yielded promising results but are not yet ready for deployment at a clinical level. Advanced neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI), have accessibility and cost constraints. Blood biomarkers are yet to show enough evidence to warrant broad-scale implementation as a diagnostic tool. Therefore, clinical assessment tests are currently used in the diagnosis protocol of concussions. Although not required to administer the tests themselves, only physicians can diagnose a concussion, after conducting a holistic review of the patient's assessment metrics and neurological status. This process can consume a physician's time and strain a clinic with limited resources. It also makes it impossible to objectively diagnose concussions at remote locations or clinics without a physician trained to recognize concussions. These limitations motivated the work described in this thesis. Research into machine learning applications in concussions has primarily focused on using DTI and fMRI metrics [4], with few studies that utilize concussion evaluation measurements [5]. Also, most of these studies have relied on univariate models. This study seeks to use multiple clinical concussion assessments and identify the importance of individual assessment scores. The goal is to inspire the application of machine learning models on readily available clinical data in the diagnosis protocol to aid physicians in their judgment by flagging patients suspected of having concussions based on their clinical assessment metrics, even before their neurological examination.

An essential part of predictive modeling is to evaluate a broad range of candidate models. While there is not a clear hierarchy in the models' predictive power, some models perform better than others due to inbuilt characteristics such as bagging and boosting. Another critical consideration that influences model performance is the data-splitting choice during the model-training process. Some of the data-splitting options include validation-split, cross-validation, repeated cross-validation, and bootstrap methods. Section 2 of the thesis provides a brief overview of the candidate models' performance and evaluation metrics. Additionally, the effects of different subsampling techniques that are typically used to address model-performance issues caused by class imbalance are explored for the final models selected from the candidates. Concussions clinics are expected to have a higher frequency of concussed patients coming into the clinic than non-concussed patients. This imbalance is reflected in the relative proportion of classes in the dataset, which poses a challenge to effectively classify both classes. One method to address this challenge is to use subsampling techniques. Some subsampling techniques, such as SMOTE and up-sampling, are explored in Section 2 to investigate these methods' effect on model performance.

The thesis contains the paper intended for publication at the Computational Intelligence in Healthcare and E-health (IEEE CICARE) Symposium. The paper contains a detailed description of the data, data preparation process, and predictive modeling approach for final models used to explore the application of machine learning models with the motivation of promoting their use in the concussion diagnosis process.

This thesis explores machine learning techniques in the concussion diagnosis process, using data from established clinical concussion assessment tests. As the clinical tests require little specialized equipment, it is possible to implement the models explored in this thesis on a large scale. Furthermore, this thesis aims to promote research into applying computational intelligence at a clinical level. The results in the following sections show promising signs for the addition of advanced modeling techniques to the concussion diagnosis protocol.

2. PREDICTIVE MODELING PROCESS

This section presents an overview of the predictive modeling process used in this research study. A detailed description of the data, data preparation, and feature selection process are given in the Paper section of the thesis. A critical step in the predictive modeling process is to identify the type of problem. The problem at hand is a two-class classification problem and requires the consideration of appropriate candidate models. A total of 14 classification models with varying degrees of complexity and interpretability were included in this list of candidates [6]. The class and type of models considered for selection are listed below.

- Linear Models: Logistic regression and support vector machine with a linear kernel (SVM Linear).
- Discriminant Analysis Models: Quadratic Discriminant Analysis (QDA), mixture discriminant analysis (MDA), heteroscedastic discriminant analysis (HDA), flexible discriminant analysis (FDA).
- Nonlinear Models: support vector machine with a radial kernel (SVM Radial), model-averaged neural network (Neural Net).
- Classification Tree Model: Recursive Partitioning (Rpart)
- Bagged/Boosted Tree-Based Models: Stochastic gradient boosting (GBM), C5.0, AdaBoost, random forest (RF), and extreme gradient boosting with DART (XgbDART).

2.1. DATA RESAMPLING TECHNIQUES

Splitting the data into only one training and one test set can lead to high variation in the model's performance [6]. Resampling methods give a more accurate representation of the true fit of a model. Additionally, resampling methods are useful for tuning the model parameters. The following three resampling approaches were used at different stages of this work.

- k-fold Cross-Validation: A k-fold cross-validation approach can be used to reduce the variance that is introduced from selecting the test dataset. In this approach, each fold of data is used as the training set k-1 times and used k times as the test dataset. The model performance is given by the average performance across k-folds of the test data.
- Repeated Cross-Validation: This approach provides an even more robust approximation of the classification model's average performance as the folds in k-fold cross-validation are resampled with replacement.
- Bootstrap: This technique involves random sampling of the training data with replacement. The unselected samples are used to estimate the error rate for each iteration. A modified version of the simple bootstrap called the 632 method addresses the bias created by non-distinct observations in the bootstrap sample by combining the simple bootstrap estimate and apparent error rate [6].

The candidate models were trained using a ten-fold repeated (five times) cross-validation as this approach has good bias and variance properties with reasonable computation time. The performance of the candidate models is shown in Table 2.2.

2.2. PERFORMANCE METRICS

The performance metrics used in this thesis are derived from a confusion matrix [6]. The confusion matrix and calculations for the measurements discussed are shown below in Table 2.1.

Table 2.1 Confusion Matrix

Truth			
Predicted		Positive Class	Negative Class
	Positive Class	True Positive (TP)	False Positive (FP)
	Negative Class	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP+T}{TP+FN+FP+T} \qquad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \qquad (2)$$

$$Specificity = \frac{TN}{FP+TN} \qquad (3)$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{4}$$

$$Kappa = \frac{Accuracy - Expec\quad\ Accuracy}{1 - Expected\ Accuracy} \tag{5}$$

The model selection process often requires the consideration of multiple performance metrics. Given that the dataset used in this study is severely imbalanced in favor of the positive class, it can be seen from the equation (1) that even a model that classifies every single sample as the positive "case" group will have high accuracy. Any model worthy of consideration will need to show an accuracy over the no-information-rate (NIR) baseline. The Kappa metric considers the class distribution in the training set and gives a measurement that takes into account an accuracy obtained by chance. This characteristic makes the Kappa coefficient an informative metric for model-performance measurement on imbalanced datasets. The sensitivity measures the true positive rate, while the specificity measures the models' true negative rate. As there is usually a tradeoff between sensitivity and specificity for imbalanced datasets, balanced accuracy, which is the average of the sensitivity and specificity measurements, indicates the model's overall performance. The objective of the modeling approach in this study is to primarily identify the concussed patients with minimum false negative while identifying enough controls so as to be useful in the diagnosis protocol. However, while sensitivity is more important than specificity when identifying concussed patients, the sensitivity cannot solely serve the purpose of identifying the best performing model as even a null classifier will have perfect sensitivity. The performance metrics need to be considered together in order to gauge a model's performance.

2.3. CANDIDATE MODEL PERFORMANCE

Table 2.2 shows the performance for a ZeroR classifier and the 14 candidate models based on the metrics described in the above section. The ZeroR classifier simply predicts the majority class, and in the context of this paper, classifies every patient as having a concussion. The accuracy of this classifier can be used as a baseline metric to compare the performance of the other models tested.

The results show that linear models (Logistic and SVM Linear) are unable to identify the control group effectively, with very low specificities. While the discriminant analysis models (except for MDA) appear to have the best specificities among the candidate model types, the improved specificity is at the expense of sensitivity. Although the discriminant analysis models are dismissed from further consideration for this study, their ability to identify controls makes them a candidate for inclusion in stacked models.

Nonlinear models (SVM Radial and Neural Net) were also removed from further consideration due to their unsatisfactory performance in terms of identifying controls. It can be observed that the tree-based models are the most appropriate classifiers for the dataset, with the C5.0 algorithm having the best Kappa and accuracy metrics, and the RF, Rpart, and AdaBoost models all having accuracy and Kappa greater than 0.94 and 0.50 respectively. After reviewing the candidate models' performance, the C5.0, Rpart, RF, and XgbDART algorithms were selected as the final models to be used in the paper.

2.4. COMPARING SUBSAMPLING METHODS

Imbalance in a dataset's class frequencies can pose a challenge to train a classification model to recognize both classes effectively. One of the remedies for handling imbalanced data is to use subsampling techniques such as up-sampling and SMOTE [6].

Table 2.2 Candidate Model Performance

Model	Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
ZeroR	0.9233	1	0	0.5	0
Linear Models					
Logistic	0.9282	0.9947	0.1277	0.5612	0.1944
SVM Linear	0.9233	1	0	0.5000	0
Discriminant Analysis Models					
QDA	0.8189	0.8163	0.8511	0.8337	0.3429
MDA	0.9233	0.9965	0.0426	0.5195	0.0672
HDA	0.8467	0.8587	0.7021	0.7804	0.3411
FDA	0.9103	0.9364	0.5957	0.7661	0.4564
Nonlinear Models					
SVM Radial	0.9233	0.99647	0.0426	0.5195	0.0672
Neural Net	0.9331	0.9823	0.3404	0.6614	0.4059
Tree-based Models					
Rpart	0.9429	0.9788	0.5106	0.7447	0.5483
Adaboost	0.9413	0.9806	0.4681	0.7243	0.5196
GBM	0.9233	0.9647	0.4255	0.6951	0.4188
C5.0	0.9543	0.9859	0.5745	0.7802	0.6346
RF	0.9429	0.9841	0.4468	0.7155	0.5166
XgbDART	0.9396	0.9823	0.4255	0.7039	0.4889

The performances for the final models using these techniques are compared with a view of evaluating whether a subsampling method can improve the final model's performance. The bootstrap 632 method was used in the model-validation process when comparing the subsampling methods to the original ratio of class frequencies. The performance comparison results for each of the final models are presented in Tables 2.3, 2.4, 2.5, and 2.6.

Table 2.3 Performance Comparison of Sampling Techniques for C5.0

Method	Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
Original	0.9511	0.9841	0.5532	0.7686	0.6085
Up-sampling	1	1	1	1	1
SMOTE	0.9070	0.9223	0.7234	0.8228	0.4957

Table 2.4 Performance Comparison of Sampling Techniques for Rpart

Sampling Method	Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
Original	0.9429	0.9788	0.5106	0.7447	0.5483
Up-sampling	1	1	1	1	1
SMOTE	0.8532	0.8498	0.8936	0.8717	0.4176

Tables 2.3 and 2.4 show that the up-sampling approach gave perfect classification in the test set using the C5.0 or Rpart algorithms. Despite these results indicating that the up-sampling technique offers superior performance, perfect classification, while encouraging, is suspicious.

Table 2.5 Performance Comparison of Sampling Techniques for RF

Sampling Method	Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
Original	0.9494	0.9894	0.4681	0.7287	0.5616
Up-sampling	0.9462	0.9735	0.6170	0.7953	0.6083
SMOTE	0.8923	0.8958	0.8511	0.8734	0.4955

Table 2.6 Performance Comparison of Sampling Techniques for XgbDART

Sampling Method	Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
Original	0.9347	0.9735	0.4681	0.7208	0.4893
Up-sampling	0.9282	0.9488	0.6809	0.8148	0.5540
SMOTE	0.8825	0.8869	0.8298	0.8584	0.4635

The results for the subsampling methods in Tables 2.5 and 2.6 indicate superior Kappa for the up-sampling technique in the RF and XgbDART models. However, the accuracy and sensitivity are higher when using the original data structure. The SMOTE method offered the best balanced-accuracy among the sampling methods for the RF and XgbDART models. However, the lower sensitivity with this subsampling approach eliminated this method from further use in this study. The low sensitivity when using the SMOTE technique can be attributed to the smaller training sample size of the positive class group that is created during this method. Its performance is worth investigating with a larger sample size of the control group. The original data's performance and the up-sampled data are very similar for both the RF and XgbDART models. The up-sampled data had higher specificity but lower sensitivity.

The SMOTE method significantly impacted all the chosen models' sensitivity, and the up-sampling technique negatively affected the sensitivity for the RF and XgbDART models. Although the up-sampling method gave perfect classification on the test set for the C5.0 and Rpart models, further investigation is warranted before recommending this subsampling technique. For the reasons stated above, it was decided that a subsampling technique will not be used to tackle the class imbalance issue inherent in the dataset. Future research can further explore these subsampling techniques along with other remedial techniques such as cost-sensitive training, and unequal class-weights [6].

PAPER

I. PREDICTIVE MODELING OF SPORTS-RELATED CONCUSSIONS USING CLINICAL ASSESSMENT METRICS

ABSTRACT

Concussions represent a growing health concern and are difficult to diagnose and manage, with roughly four million concussions diagnosed every year in the United States. While research in machine learning applications for concussions have focused on the use of advanced metrics such as neuroimages, and blood biomarkers, these metrics are yet to be implemented at a clinical level due to cost, and reliability concerns. Therefore, concussion diagnosis is still reliant on clinical evaluations of symptoms, balance, and neurocognitive status and function. The lack of a universal threshold on these assessments make the diagnosis process entirely reliant on a physician's interpretation of these assessment scores. The aim of this study is to explore and promote the use of machine learning techniques to aid the concussion diagnosis process. The benefits of the models proposed include being able to flag concussed patients even before being seen by a doctor and expanding the scope of concussion diagnosis to remote locations, and areas with limited access to doctors.

1. INTRODUCTION

Concussion, a term that is often used to describe mild traumatic brain injury (mTBI) has several consensus-based definitions [1], and the lack of a universal definition has led to bias in clinical applications. Although the terms concussion and mTBI are frequently used interchangeably, the loosely defined former has some notable distinctions from the latter [2]. The Glasgow Coma Scale (GCS) has traditionally been used to gauge the severity of traumatic brain injuries, with a GCS score between 13-15 indicating an mTBI. However, the GCS score is not suitable for distinguishing severity variation in the mTBI range. It is possible to have a fractured skull or intracranial hemorrhage and still obtain a GCS score between 13-15 [3]. Despite the confusion surrounding the definition of concussion, it is generally agreed that a concussion is a biomechanically induced alteration in brain physiology inducing neurocognitive dysfunction, not necessarily involving a loss of consciousness [1-4]. There is not any disagreement about whether concussions represent a severe problem that is difficult to diagnose and has potentially disabling sequelae [5]. In the United States itself, roughly four million sports and recreation-related concussions occur every year [6].

In 2014, the National Collegiate Athletic Association, together with the Department of Defense established the Concussion Assessment, Research, and Education (CARE) Consortium to address the challenges associated with concussion diagnosis and management, particularly among student-athletes and military cadets [7][8][1]. The ongoing CARE study is the biggest clinical concussion-study in history [1], and has conducted research on traditional and new clinical concussion assessment tools, magnetic resonance imaging (MRI) metrics, and investigated the pathophysiology of concussions through neurological tests, neuroimages, and blood biomarkers [9].

Although advanced concussion assessment approaches have shown promise, they are yet to transition to an application on a clinical scale. Access to advanced MRI techniques for individual patients is limited, and blood biomarkers are yet to show the required level of sensitivity for implementation as a diagnostic tool [5]. Therefore, concussion diagnosis relies on clinical examinations [10] that evaluate symptoms, neurocognitive status and function, and balance. The Brief Symptom Inventory-18 (BSI-18), Standardized Assessment of Concussion (SAC), Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT), and Balance Error Scoring System (BESS) are some of the most commonly used clinical concussion assessments [10]. Less commonly used tests for reaction time, oculomotor and vestibular function include the King-Devick, Vestibular Ocular Motor Screen (VOMS), and Clinical Reaction time [10]. The Sports Concussion Assessment Tool (SCAT5) [11] is a commonly used tool for evaluating concussions in sports that combines assessments for cognitive-measure, balance, and acute symptoms to provide a broad scope of measurements. However, it does not function as a single metric for diagnosis [12].

Clinical examinations can be easily administered by trained proctors or healthcare professionals and do not require a doctor to administer. However, only a physician can diagnose the concussion. While research into machine learning applications in concussions is gaining traction, they have primarily focused on using diffusion tensor imaging (DTI) and functional MRI metrics [13], with few studies relying on concussion evaluation measurements [14]. This paper aims to explore and promote the use of machine learning techniques to aid the diagnosis process, which relies on clinical evaluation metrics. By correctly identifying a majority of concussions before being seen by a doctor, a clinic can quickly flag patients that require immediate attention. Additionally, these models can be implemented in remote areas, including rural or isolated military locations, with limited access to trained physicians to recommend additional examination or care to individuals identified by the model. Moreover, such a tool will enable clinics with limited resources to effectively manage patient care.

2. METHODS

2.1. DATA

The data used in this study are available through the research conducted by the CARE consortium [7][8] and were downloaded on August 22, 2019, through the Federal Interagency Traumatic Brain Injury Research (FITBIR) website [15]. The CARE investigation conducted clinical assessments [10] for all consenting participants at baseline. Individuals diagnosed with a concussion were labeled as a 'case', and clinical assessments were repeated at five time-periods post-injury [7], including <6 hours post-injury, 24-48 hours post-injury, asymptomatic, unrestricted return to play (RTP), and finally six months post-RTP. Matched non-concussed subjects labeled as ‘control’ were also given the clinical-assessment tests following the case subjects' evaluations. The assessments included established Level A tests and emerging Level B tests [10]. Except for ImPACT, which was given at 25 out of 29 sites, the other level A tests were administered at all CARE test locations. The models used in this study only used Level A scores.

This study's primary objective is to use predictive models in classifying the case and control groups at a given time to explore the potential of using machine learning methods to diagnose concussions using only established Level A measurements. The data were filtered to focus on the 24-48 hour time point as it is regarded as a critical period during which concussion patients are symptomatic, and this time point has the most complete data during the acute concussion phase [7]. Also, tests such as the ImPACT test and BSI-18 are not given at the < 6-hour time point. The data contained clinical assessment measures for 2455 participants with 2265 subjects in the concussed case group and 190 subjects in the control group. Although concussion history was not factored into the predictive modeling in this study, it is worth noting that the matched controls included individuals with a history of concussions. The Level A concussion assessment tests are briefly described below.

- Balance Error Scoring System (BESS): BESS [16] is used to assess the effects of mild head injury on static postural stability. The test is conducted on both firm and foam surfaces with the scores for each surface range from 0 to 30, and each increment representing an error.
- Standardized Assessment of Concussion (SAC): The SAC [17] assesses the cognitive status and contains sections on orientation, immediate memory, concentration, and delayed recall. Each section contains a binary scoring system (0=wrong, 1=correct). Composite scores are calculated for each section and then added to give a total score.

- Brief Symptom Inventory-18 (BSI-18): The BSI-18 [18] is a self-reported questionnaire consisting of 18 descriptions of physical and emotional pain symptoms. Individuals are asked to indicate on a scale from 0 (not at all) to 4 (very much) to what extent they are troubled by each symptom. The symptom list consists of three symptom scales: somatization, depression, and anxiety. Each of the scales comprises of six symptoms.
- Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT): The ImPACT test [19] is a neurocognitive test that measures verbal memory, visual memory, visual-motor speed, reaction time, impulse control, and post-concussion symptoms. It can be a useful tool to establish neurocognitive performance post-injury by comparing to baseline (when available) or to scores for similar age groups.

The clinical evaluations listed above provided the features for the predictive models used in this study. A full list of features from the tests is given below in Table 1.

Table 1. Feature List

Assessment	Features Extracted
BESS	BESS.Total.Firm.Error, BESS.Total.Foam.Error
SAC	SAC.Concentration, SAC.Delayed.Recall, SAC.Immediate.Memory, SAC.Orientation*
BSI-18	BSI18.Depression, BSI18.Anxiety, BSI18.Somatization
ImPACT	ImPACT.Total.Symptom, ImPACT.Visual.Motor.Speed, ImPACT.Visual.Memory, ImPACT.React.Time, ImPACT.Verbal.Memory, ImPACT.Impulse.Control*

* Eliminated during the feature selection process.

2.2. DATA PREPARATION

The data preparation process revealed several unusual observations, especially in the control group, that would significantly influence the predictive modeling process. Due to the limited number of control subjects, the extreme observations were replaced with missing values and then imputed. The ImPACT Clinical Interpretation Manual [19] provides guidelines for identifying unusual observations at baseline that can usually indicate deliberate poor performance at baseline. For instance, ImPACT reaction time scores in the range of 0.8 to 1.5 at baseline are usually indicative of sandbagging [19]. Similarly, unusual observations for other ImPACT composite scores include verbal memory less than 70, visual memory less than 60, motor speed less than 25, and impulse control scores higher than 30 [19]. It is recommended that impulse control scores above 20 are reevaluated [20]. Given that the control subjects did not have a concussion, the values suggested by the ImPACT manual were used as a cutoff to replace the unusual observations in the control group. The filters used for the other assessment tests included a BESS foam-error greater than 17 but firm-error less than 7. SAC concentration and delayed recall scores below 3 were also replaced with missing values in the control group. The extreme observations in the case group were not manipulated as it is possible that they suffered from significant cognitive impairment. Only an unusual value of 83 seconds for the ImPACT reaction time score was replaced with a missing value in the case group.

The data contained missing values, with the ImPACT composite scores having the most missing values. The ImPACT total symptom score had 719 missing values, while 699 individuals were missing for the other ImPACT composite scores. It is likely that the missing values for the ImPACT scores can largely be attributed to the ImPACT tests being administered at fewer locations than the other Level A tests. The missing values were imputed using multivariate imputation by chained equations (mice), and the predictive mean matching method was used in the mice algorithm [21]. The imputations were performed separately for the case and control groups to preserve the nature of the groups' distributions. The boxplots representing some of the test scores are shown in Figure 1.

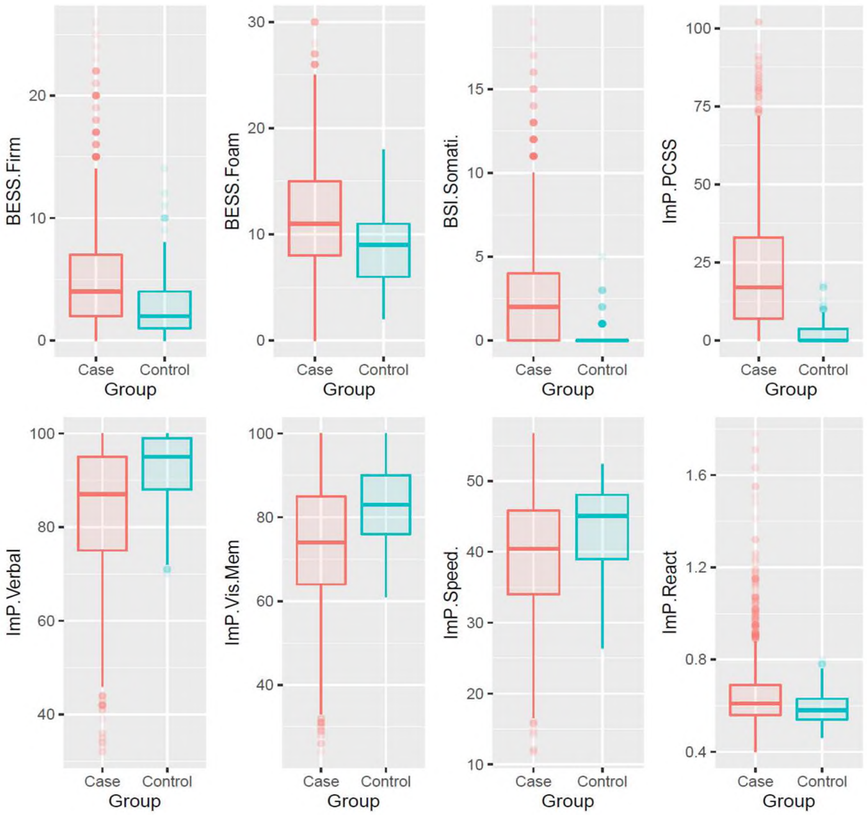

Figure 1. Boxplots of Assessment Scores for Case and Control Groups

2.3. PREDICTIVE MODELING APPROACH

The imbalanced nature of the dataset with 2265 samples in the positive case class and 190 samples in the negative control class makes this a challenging classification problem. Typically complications caused by an imbalance in the classes can be addressed by subsampling techniques such as up-sampling the minority class, down-sampling the majority class, or hybrid methods including synthetic minority oversampling technique (SMOTE) and random oversampling examples (ROSE) [22]. However, there are a few issues with these subsampling techniques. The SMOTE, ROSE, and down-sampling techniques sacrifice some training samples of the positive class. Given the limited data and the fact that identifying concussions is more important than identifying the controls, it was decided that these subsampling methods were unsuitable for this exploratory study. The up-sampling technique is also not suitable as it alters the natural state of the data. The authors wanted to train the data to reflect the imbalance that is likely visible among patients, particularly student-athletes undergoing concussion diagnosis. Due to the reasons stated above, a stratified split was used in this study. The data were split to have 75% of the data in the training set, and 25% of the data in the holdout set in order to have sufficient samples of both classes in the training and evaluation process.

An initial exploration of candidate models revealed that linear and discriminant analysis models were ineffective in separating the case and control classes. Four models with diverse computational and modeling complexities were selected from the candidate models set. The final models selected for this study are briefly described below.

- C5.0: The C5.0 algorithm is an advanced version of the C4.5 algorithm [23] with boosting and unequal penalties for different types of errors [22]. The

C5.0 algorithm can function as a rules-based or tree model. The tuning process for the C5.0 method in the caret package [24] of R can include the selection between rule-based and tree-based models. The C5.0 determines predictor importance by identifying the percentage of training samples that fall into all the terminal nodes after a split [22]. The model also has an option for dropping noninformative predictors through a process called winnowing, but this selection does not always improve the error rate. Winnowing can be added as a tuning parameter to choose the full predictor set, and a pruned predictor set.

- Recursive Partitioning (Rpart): The Rpart is a classification and regression tree method that builds the classification tree by first identifying the feature that best splits the data according to a node purity criteria and building binary trees until no further improvement in performance is observed [25]. A tuning parameter called the cost-complexity (cp) parameter can be incorporated into the model building process to construct a pruned tree to counter the tree-based classifier's tendency to overfit the training data.
- Random Forest (RF): Random Forest [26] is an ensemble model of decision trees that train learners in parallel on different samples of data. Then, the votes of each tree are combined to obtain a predicted class. A random subset of the predictors is selected to grow decorrelated trees. The tuning parameter mtry determines the number of predictors selected. The number of trees needed for good performance is dependent on the number of predictors, with more trees giving more stability to the variable importance estimates [27]. The random

forest classifier is robust to overfitting and can produce competitive results compared to powerful boosting algorithms [26].

- Extreme Gradient Boosting with DART (Dropouts meet Multiple Additive Regression Trees) booster (XgbDART): The extreme gradient boosting model is a boosted ensemble tree-based algorithm built on the gradient boosting algorithm, and has similarities to the random forest model. However, unlike the random forest model where trees are created independently, the gradient boosting algorithm creates trees dependent on prior trees [22]. The XgbDART model uses the DART [28] booster technique, which incorporates dropouts for the ensemble trees in the extreme gradient boosting algorithm. The XgbDART algorithm learns from the existing trees in the ensemble to compensate for shortcomings in the prior trees.

The bootstrap 632 method [29] resampling technique with 500 resamples was used in the model training process as this method effectively reduces the bias and variance in performance. This method is a variation of the bootstrap method that addresses the bias created by non-distinct observations in the bootstrap sample by combining the simple bootstrap estimate and apparent error rate [22]. Due to the extreme imbalance in the data and the small sample size of the control group, the Kappa metric was chosen as the metric for tuning the model. The Kappa metric considers the class distribution in the training set and gives a measurement that takes into account an accuracy obtained by chance. This characteristic makes the Kappa coefficient an informative metric for model performance measurement on imbalanced datasets.

For a given assessment test, the total scores were dropped from the model feature list in favor of their subcomponent scores to assess the importance of each element in a clinical assessment metric. Fifteen features were available for use in the model after dropping the total scores for the SAC, BESS, and BSI-18 tests. The Boruta [30] feature selection algorithm was implemented to identify all the relevant variables to use in the selected models. The Boruta algorithm is based on the random forest model and uses an iterative approach to identify important and nonimportant features by comparing the variable to randomly created shadow attributes [30]. The results from the Boruta algorithm can be seen in Figure 2.

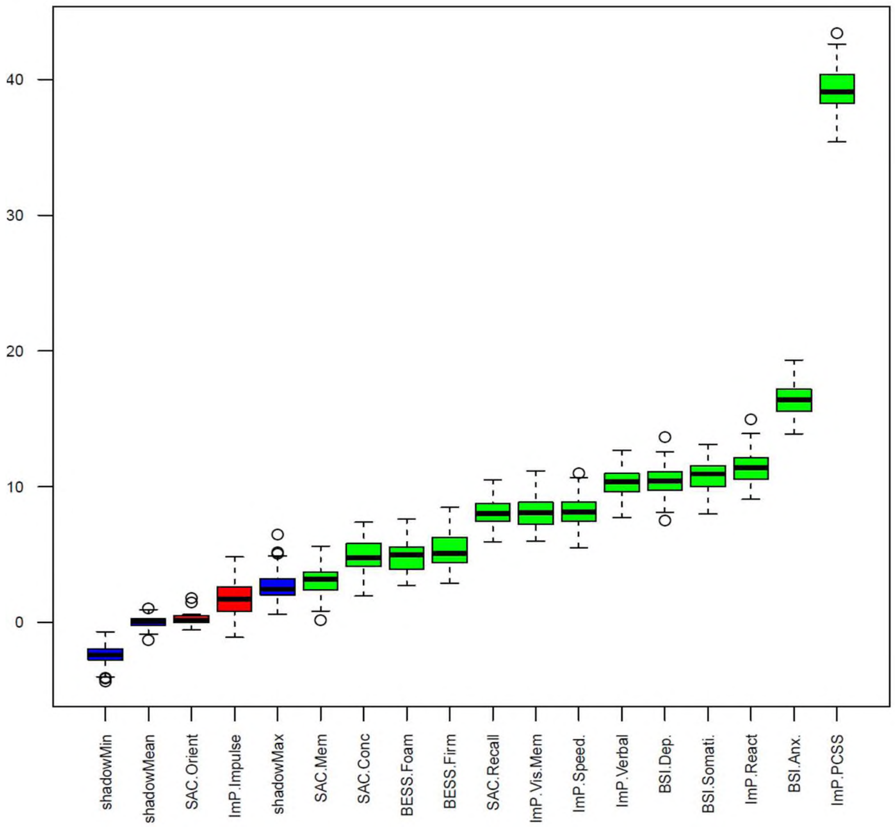

Figure 2. Feature Selection Results from the Boruta Algorithm

The green and red plots represent the Z-scores of the selected and rejected features respectively, while the blue plots are the shadow attributes. The algorithm identified the SAC orientation score and ImPACT impulse control scores as unimportant attributes. Hence, these two features were removed from further consideration, leaving a final feature list of 13 clinical assessment scores for use in the classification models (Table 1).

3. RESULTS

The predictive modeling results on the holdout set containing 566 concussed (case) and 47 non-concussed (control) samples are presented in this section. The test set's imbalanced nature with a prevalence of 92.33% was assumed to represent the real potential rate of concussion among athletes after a head impact. The final tuning parameters obtained from the bootstrap training method for each classification model tested are given in Table 2.

Table 2. Final Tuning Parameters for the Classification Models

C5.0	Rpart	RF	XgbDART

trials = 100
model = tree
winnow = FALSE
	cp = 0.01398601	mtry = 11
		ntree = 5000
		(ntree was set manually)
			nrounds = 709
			max_depth = 9
			eta = 0.6213674
			rate_drop = 0.3279209
			skip_drop = 0.8211969
			min_child_weight = 1

The confusion matrices from the classification models on the holdout set are presented in Table 3. It can be seen that the random forest model had the best performance in terms of identifying the concussed case group, and the C5.O algorithm correctly identified the most controls from the models tested. All of the models presented in this paper were successfully able to identify most of the case group members, and the C5.0 and Rpart models correctly classified more than half of the controls.

Table 3. Confusion Matrices for the Classification Models

C5.0				Rpart			
Truth				Truth			
Predicted		Case	Control	Predicted		Case	Control
	Case	557	20		Case	554	23
	Control	9	27		Control	12	24
RF				XgbDART			
Truth				Truth			
Predicted		Case	Control	Predicted		Case	Control
	Case	559	25		Case	558	27
	Control	7	22		Control	8	20

The ZeroR classifier simply predicts the majority class, and in the context of this paper, classifies every patient as having a concussion. The accuracy of this classifier can be used as a baseline metric to compare the performance of the other models tested. The other metrics used to assess model performance include the Kappa, specificity, sensitivity, balanced accuracy, and F2 scores.

The sensitivity measures the true positive rate of classification on the test set, while the specificity measures the models' true negative rate. The F2 score is a weighted averaged of the precision and sensitivity of a model’s performance such that false negatives are more important than false positives. As the primary goal of the modeling approach is to correctly identify as concussions while minimizing false negatives, the F2 score can be considered an important metric for evaluating model performance.

Table 4. Performance Metrics for the Classification Models

	Acc.	Sens.	Spec.	Balanced Acc.	F2	Kappa	Training Time
ZeroR	0.9233	1	0	0.5	0.9904	0	0
C5.0	0.9527*	0.9841	0.5745	0.7793	0.9803	0.6257	14.62 mins
Rpart	0.9429*	0.9788	0.5106	0.7447	0.9750	0.5483	13.30 secs
RF	0.9478*	0.9876	0.4681	0.7279	0.9814	0.5528	50.06 mins
XgbDAR T	0.9429*	0.9859	0.4255	0.7057	0.9793	0.5050	5.23 hours

* Accuracy > 0.9233 at 0.05 significance level. Acc: Accuracy, Sens: Sensitivity, Spec: Specificity, Balanced Acc: Balanced Accuracy

The performance metrics of the evaluated classification models are presented in Table 4. The C5.0, Rpart, random forest, and XgbDART models all had a significantly better accuracy than the ZeroR classifier. The C5.0 model had the highest balanced accuracy (77.93%), with a 98.41% sensitivity and specificity of 57.47%. This model also had the highest Kappa metric with the XgbDART showing the lowest Kappa. As expected, the ZeroR classifier with perfect sensitivity has the highest F2 score. However, this model does not add value by merely classifying every subject as having a concussion, and this can be seen in the zero specificity of the ZeroR classifier.

The random forest had the highest F2 score among the models evaluated with the C5.0 model following closely. The relative performance of the Rpart model is particularly impressive, considering that it was computationally the least expensive model in terms of time. The tuning process took the least amount of time for the Rpart model and the most amount of time for the XgbDART model.

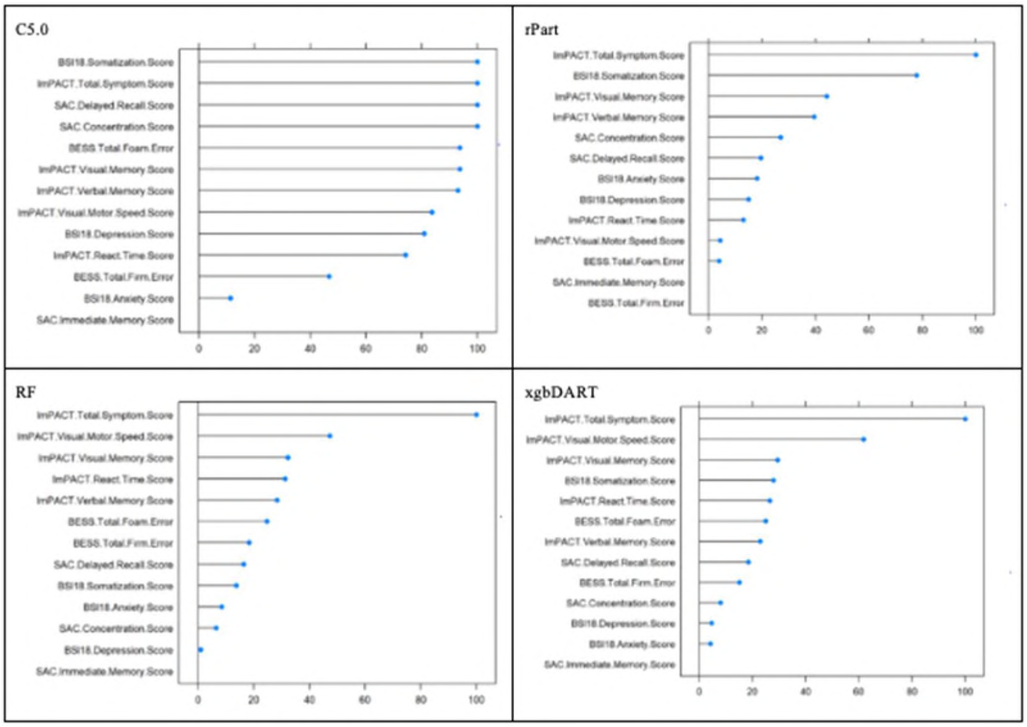

Figure 3. Variable Importance Plots for the Classification Models

The variable importance of the assessment scores for each model is presented in Figure 3. The x-axis on the figure represents the scaled importance of the features, with 100 representing the most important feature. The BSI-18 somatization, ImPACT symptom, SAC concentration, and delayed recall scores were essential features in the C5.0 algorithm model. In contrast, the BESS firm, BSI-18 anxiety, and SAC immediate memory scores were the least important features for that model. In the Rpart model, the ImPACT Symptom, BSI-18 somatization, and ImPACT visual and verbal scores were ranked (in order) as the top features. The BESS foam, SAC immediate memory, and BESS firm scores were the least important variables in the Rpart model. The random forest model's top five features were all composite scores of the ImPACT test, with the symptom and motor speed composite scores being the most important features. The SAC concentration score, BSI-18 depression, and SAC immediate memory scores were the least important in the random forest model. The ImPACT symptom score was again the most important predictor and the SAC immediate memory score the least important predictor in the XgbDART model.

4. DISCUSSION

4.1. CONCLUSIONS

This study evaluated the performance of machine learning models, specifically C5.0, Rpart, random forest, and XgbDART, on identifying concussed individuals in an imbalanced dataset by using scores from the BSI-18, BESS, SAC, and ImPACT assessment tools as features. Few studies have explored using the Level A tools in a multifactor model to separate concussed individuals from non-concussed subjects at a given time point. This study showed that it is possible to use the established concussion assessment metrics in machine learning models to identify a majority of concussed individuals correctly. The C5.0 model had the best balanced-accuracy, making this a good choice for trials at clinics looking to optimize their resources. The random forest model had the highest F2 score, making this the preferred model for a cautious classification approach seeking to minimize the misclassification of the concussed group. The computational efficiency of the simple Rpart model makes this model a good choice for quick classification. Although the highest specificity of the models evaluated was only 57.45%, the results show much promise for further exploration of using machine learning techniques in clinical settings. The results also support the use of such modeling approaches in remote locations without a doctor. Rather than merely assuming that anyone coming into a clinic has a concussion, using these models can provide some type of objective classification supported by data.

When viewed across all models, the variable importance plots indicate that the ImPACT symptom score is the most critical variable. The SAC immediate memory score is the least important predictor of the 13 features used to construct each model. The SAC orientation score and ImPACT impulse control scores are nonessential predictors according to the Boruta feature selection method. It can also be observed from the variable importance plots that the BESS foam score is more important than the BESS firm score. Of the BSI-18 scores, it can be seen that the somatization score is more important than the depression or anxiety scores. Overall, it appears that the ImPACT test generally has important scores.

Recent research on using machine learning models in concussion diagnosis has concentrated on using advanced metrics, such as neuroimaging and biomarkers. However, the application of these methods on a large clinical level is still limited. Clinical evaluations are already conducted on a large scale, and therefore, the models explored in this study can be easily implemented and tested across any clinic with these data. The results from each of the models show encouraging signs to promote the use of these classification models in the diagnosis protocol.

4.2. LIMITATIONS AND FUTURE WORK

A limitation of this study is that only student-athletes and their level A measurements were used in the modeling process. Some factors such as gender, concussion history, cause of injury, and age-group that can potentially add to the models' predictive power were not considered in this study. A further limitation is that a large number of missing values for the ImPACT scores were imputed. Using a more complete dataset with a larger number of control subjects and incorporating the factors listed above that were excluded in this study can improve the machine learning models' predictive ability. Also, the use of change scores was not considered in this study, but it can be useful in the future for evaluating athletes when baseline scores are available. Additionally, the expected cost of misclassifications can be explored to identify optimum thresholds for applications in clinics with limited resources. This study's results motivate future research to explore applying deep learning models to identify concussions using clinical assessment metrics.

REFERENCES

1) NCAA Sport Science Institute. (2017). DIAGNOSIS AND MANAGEMENT OF SPORT-RELATED CONCUSSION BEST PRACTICES.

2) Chancellor, S. E., Franz, E. S., Minaeva, O. V., & Goldstein, L. E. (2019). Pathophysiology of Concussion. Seminars in Pediatric Neurology, 30(Grade 3), 14–25. https://doi.org/10.1016/j.spen.2019.03.004

3) Giza, C. C., & Kutcher, J. S. (2014). An introduction to sports concussions. CONTINUUM: Lifelong Learning in Neurology, 20(6 Sports Neurology), 1545.

4) Kazl, C., & Torres, A. (2019, July). Definition, Classification, and Epidemiology of Concussion. In Seminars in pediatric neurology (Vol. 30, pp. 9-13). WB Saunders.

5) Snyder, A. R., & Giza, C. C. (2019). The Future of Concussion. Seminars in Pediatric Neurology, 30, 128–137. https://doi.org/10.1016/j.spen.2019.03.018

6) Katz, B. P., Kudela, M., Harezlak, J., McCrea, M., McAllister, T., Broglio, S. P., & CARE Consortium Investigators. (2018). Baseline performance of NCAA athletes on a concussion assessment battery: a report from the CARE Consortium. Sports Medicine, 48(8), 1971-1985.

7) Broglio, S. P., McCrea, M., McAllister, T., Harezlak, J., Katz, B., Hack, D., ... & CARE Consortium Investigators. (2017). A national study on the effects of concussion in collegiate athletes and US military service academy members: the NCAA–DoD concussion assessment, research and education (CARE) consortium structure and methods. Sports medicine, 47(7), 1437-1451.

8) http://www.ncaa.org/sport-science-institute/topics/ncaa-dod-care-consortium

9) http://www.ncaa.org/sport-science-institute/topics/care-consortium-findings

10) Broglio, S. P., Katz, B. P., Zhao, S., McCrea, M., McAllister, T., & CARE Consortium Investigators. (2018). Test-retest reliability and interpretation of common concussion assessment tools: findings from the NCAA-DoD CARE Consortium. Sports Medicine, 48(5), 1255-1268.

11) Echemendia, R. J., Meeuwisse, W., McCrory, P., Davis, G. A., Putukian, M., Leddy, J., ... & Schneider, K. (2017). The sport concussion assessment tool 5th edition (SCAT5): background and rationale. British Journal of Sports Medicine, 51(11), 848-850.

12) Garcia, G. G. P., Broglio, S. P., Lavieri, M. S., McCrea, M., McAllister, T., & CARE Consortium Investigators. (2018). Quantifying the value of multidimensional assessment models for acute concussion: an analysis of data from the NCAA-DoD Care Consortium. Sports Medicine, 48(7), 1739-1749.

13) Cantu, R. C. (2019). History of Concussion Including Contributions of 1940s Boston City Hospital Researchers. Seminars in Pediatric Neurology, 30, 2–8. https://doi.org/10.1016/j.spen.2019.03.002

14) Navarro, S. M., Sokunbi, O. F., Haeberle, H. S., Schickendantz, M. S., Mont, M. A., Figler, R. A., & Ramkumar, P. N. (2017). Short-term outcomes following concussion in the NFL: a study of player longevity, performance, and financial loss. Orthopaedic journal of sports medicine, 5(11), 2325967117740847.

15) https://fitbir.nih.gov/

16) researchers and clinicians at the University of North Carolina's Sports Medicine Research Laboratory. (n.d.). Balance Error Scoring System (BESS).

17) McCrea, M., Kelly, J. P., Randolph, C., Kluge, J., Bartolic, E., Finn, G., & Baxter, B. (1998). Standardized assessment of concussion (SAC): on-site mental status evaluation of the athlete. The Journal of head trauma rehabilitation, 13(2), 27-35.

18) Wideman, T. H., Sullivan, M. J. L., Inada, S., McIntyre, D., Kumagai, M., Yahagi, N., Turner, J. R., Upton, J., Burns, R. J., Rothman, A. J., Michie, S., Johnston, M., Nakashima, M., Vedhara, K., Dawe, K., Wong, C., Gellman, M. D., Brimmer, D., Zielinski-Gutierrez, E., … Woltz, P. (2013). Brief Symptom Inventory. Encyclopedia of Behavioral Medicine, 269–270. https://doi.org/10.1007/978-1-4419-1005-9_3

19) Lovell, M. (2007). ImPACT 2007 (6.0) clinical interpretation manual. Pittsburgh, PA: ImPACT Applications Inc.

20) ImPACT. (2007). ImPACT Interpretation Manual.

21) Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40–49. https://doi.org/10.1002/mpr.329

22) Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

23) Quinlan, J. R. (2006). Bagging, Boosting , and C4.5. 725–730. https://www.aaai.org/Papers/AAAI/1996/AAAI96-108.pdf

24) https://topepo.github.io/caret/subsampling-for-class-imbalances.html#

25) Kattan, M., Chun, F. K.-H., Graefen, M., Haese, A., & Karakiewicz, P. I. (2012). Recursive Partitioning. Encyclopedia of Medical Decision Making, 1–60. https://doi.org/10.4135/9781412971980.n280

26) Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

27) Pavlov, Y. L. (2019). Random forests. Random Forests, 1–122. https://doi.org/10.1201/9780367816377-11

28) Rashmi, K. V., & Gilad-Bachrach, R. (2015). DART: Dropouts meet multiple additive regression trees. Journal of Machine Learning Research, 38, 489–497.

29) Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American statistical association, 78(382), 316-331.

30) Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. J Stat Softw, 36(11), 1-13.

SECTION

3. CONCLUSIONS

This thesis was motivated by the potential of machine learning models to support the diagnosis protocol for concussions. This thesis aims to explore and promote the use of these advanced modeling techniques at a clinical level, using widely available concussion evaluation metrics. To accomplish this, a detailed predictive modeling approach was followed to classify concussed and non-concussed patients. A wide variety of classification models, including linear, nonlinear, and tree-based models, were evaluated as candidates to identify the most appropriate model-type for this classification problem. It was observed that tree-based classification models, including boosted and bagged models, are more suitable to classify the dataset containing the clinical concussion test features of 2265 concussed and 190 non-concussed student-athletes. Concussion clinics typically have a higher relative proportion of concussed patients coming into the clinic. Therefore, subsampling techniques were explored to remedy the class-imbalance issue that is innate to clinical concussion data. Although no subsampling technique was chosen in the final implementation of the chosen models, the results in Section 2 show each subsample approach’s pros and cons.

The findings in the Paper section of the thesis demonstrate the potential benefits of using tree-based classifiers to identify concussed and non-concussed subjects at a 24-48-hour post-injury time point. The research also suggests that not all clinical assessment test scores are of equal importance. It was found that the ImPACT symptom score is an essential assessment metric to identify concussed patients while the SAC orientation and ImPACT impulse control scores are nonessential features. It was also observed that the SAC immediate memory score is of low importance.

The results observed in this thesis show promising signs for the clinical implementation of machine learning techniques, particularly tree-based classification models, in the concussion diagnosis process. While the application of advanced machine learning models in TBI research has been gaining momentum, especially on advanced neuroimaging data, this thesis's research showed that it is beneficial to use these models on routine clinical evaluations. The study advocates for further exploration of machine learning techniques using clinical assessment metrics. Clinical assessment scores are easier to collect than imaging or biomarker data, thereby offering a platform for large scale clinical exploration and implementation of these models. Implementing machine learning models to identify concussed patients will also expand the scope of clinical evaluations to remote locations and clinics without physicians trained to identify concussions.

Based on this study's results, further research can explore the implementation of these models on larger datasets with few missing values and use of additional factors such as concussion history, gender, cause of injury, and age that can improve the predictive power of the machine learning models. The encouraging results can motivate exploring techniques to handle imbalanced datasets such as cost-sensitive training, unequal class-weights, and alternate-thresholds for models. The subsampling techniques can also be explored further to remedy model training issues caused by the imbalanced data. Section 2 provides a foundation for this investigation. The use of change scores rather than raw scores can also be investigated when baseline metrics are available. Furthermore, using data collected from concussion clinics, the expected cost of misclassification can be used to tune appropriate model parameters to optimize model performance tailored to match a clinic's expectations.

BIBLIOGRAPHY

[1] Broglio, S. P., McCrea, M., McAllister, T., Harezlak, J., Katz, B., Hack, D., ... & CARE Consortium Investigators. (2017). A national study on the effects of concussion in collegiate athletes and US military service academy members: the NCAA–DoD concussion assessment, research and education (CARE) consortium structure and methods. Sports medicine, 47(7), 1437-1451.

[2] Giza, C. C., & Kutcher, J. S. (2014). An introduction to sports concussions. CONTINUUM: Lifelong Learning in Neurology, 20(6 Sports Neurology), 1545.

[3] Cantu, R. C. (2019). History of Concussion Including Contributions of 1940s Boston City Hospital Researchers. Seminars in Pediatric Neurology, 30, 2–8. https://doi.org/10.1016/j.spen.2019.03.002

[4] Sakai, K., & Yamada, K. (2019). Machine learning studies on major brain diseases: 5-year trends of 2014–2018. Japanese Journal of Radiology, 37(1), 34–72. https://doi.org/10.1007/s11604-018-0794-4

[5] Garcia, G. G. P., Broglio, S. P., Lavieri, M. S., McCrea, M., McAllister, T., & CARE Consortium Investigators. (2018). Quantifying the value of multidimensional assessment models for acute concussion: an analysis of data from the NCAA-DoD Care Consortium. Sports Medicine, 48(7), 1739-1749.

[6] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

VITA

Sujit Subhash received his B.E. in Mechanical Engineering from M.S. Ramaiah Institute of Technology in June, 2010, and joined Missouri University of Science and Technology as a graduate student in the Engineering Management & Systems Engineering Department in August, 2012. In April, 2014, he was awarded the 2013-2014 Outstanding M.S. Graduate Student Research Award of the Engineering Management & Systems Engineering Department at Missouri University of Science and Technology. In December, 2014, he received his M.S. in Engineering Management from Missouri University of Science and Technology, Rolla, Missouri. He enrolled in the master's program at the Department of Mathematics and Statistics in January, 2019, and graduated with a M.S. in Applied Mathematics in December, 2020.