EFFECTS OF CONCUSSION AND VISUOMOTOR METRICS ON NHL PERFORMANCE: AN EXPLAINABLE AI APPROACH

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Michael Moschitto

June 2023

COMMITTEE MEMBERSHIP

TITLE: Effects of Concussion and Visuomotor Metrics on NHL Performance: an Explainable AI Approach

AUTHOR: Michael Moschitto

DATE SUBMITTED: June 2023

COMMITTEE CHAIR: Sumona Mukhopadhyay, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Joydeep Mukherjee, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Paul Anderson, Ph.D.
Professor of Computer Science

ABSTRACT

Effects of Concussion and Visuomotor Metrics on NHL Performance: an Explainable AI Approach

Michael Moschitto

Cognitive motor integration (CMI), the simultaneous coordination between cerebral function and motor output, is known to deteriorate following a mild traumatic brain injury (mTBI). This thesis explores the relationship between mTBI, CMI, and the performance of elite athletes in the National Hockey League (NHL). The approach focuses on examining the predictive value of various supervised Machine Learning (ML) models with an emphasis on Explainable Artificial Intelligence (XAI) models. Since the ML solution is intended to complement human scouting decisions, we evaluate the experiments based on both interpretability and accuracy on a limited class imbalanced dataset. The contributions of this research are two-fold based on the following research problems: Firstly, the problem of scouting decisions for amateur hockey players to play in the field is addressed by exploring a set of test scores from a neuroscience experiment involving visuomotor performance metrics. Formulated as a supervised binary classification task, results demonstrate that the trained XAI trained models effectively capture the relationship that determines whether amateur hockey players with a history of concussions are likely to play in the NHL. Specifically, we find the best-performing model to be Weighted-Decision Tree trained using all features proposed in this study. Secondly, the effect of previous concussions on scouting decisions is examined by visuomotor metrics and indicators of NHL performance using XAI models. This problem is also formulated as a supervised binary classification task and results show that the trained XAI models are able to predict

concussion history using the visuomotor metrics. While results for this question are inconclusive, we give evidence from current neuroscience literature to support why these models do not reach satisfactory performance. Unlike previous research that mainly relies on physical metrics, our work is novel as it utilizes data derived from a neuroscience test, capturing persistent neurocognitive deficits in elite hockey athletes following concussions.

ACKNOWLEDGMENTS

Thank you to each of the CS faculty, including my committee of Dr. Sumona, Dr. Muhkerjee, and Dr. Anderson, who have patiently contributed to this research and influenced my education at Cal Poly, my roommates for their unwavering encouragement, and my family Mike, Daisy, Eva, and Peter for their constant support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Hockey is a globally popular sport with professional leagues in the US, Canada, Mexico, Russia, Finland, and more, and is played by over 30 million people worldwide. Of the 3 largest professional ice hockey leagues, the National Hockey League (NHL) (US/Canada), the Kontinental Hockey League (Russia), and Swedish Hockey League (Sweden), the NHL is the largest, most talented, and most lucrative. Each year it generates between 2-5 billion dollars in revenue [41] while attracting top hockey talent from around the world. There exist large competitive and business pressures on teams as they seek to be the best each season. NHL teams gain new players through the NHL amateur draft which is composed of 7 rounds of 31 choices per round and totals 217 players [61]. Thus, drafting new players becomes an integral part of team success.

The effect of concussion on the coordination between brain and motor output has been the subject of much research [53, 11, 85, 65, 4]. These studies serve as inspiration for this thesis since the hypothesis designed is that if concussions affect movement output and precision of movement, they may also affect athletic performance. Concussion is defined as the phenomena that occur when biomechanical forces exerted on the head, neck, face, or elsewhere in the body induce a traumatic brain injury [81]. The coordination between the brain and targeted physical movement is called cognitive-motor integration (CMI). Concussions, also referred to as mild traumatic brain injury (mTBI) are shown in multiple instances to negatively affect CMI. The simultaneous

application of cognitive and motor tasks is a prevalent challenge in sports and has been shown to be a good indicator of post-concussion residual deficits [100].

There exists conflicting research into the effects of concussion in the NHL. It has been shown that players who sustain concussions score fewer points the next season, are significantly less likely to play in a full season the following year, and have career earnings reductions [85]. This is corroborated by multiple reviews outlining the long-term cognitive and physical effects of sports-related concussions. Specifically, athletes may experience motor performance deficits over two decades after their last concussion [80] and mild cognitive impairment after retirement [76]. Baker et al. [11] present a study which demonstrates that CMI and decision-making deficits are present in concussed individuals even after they are deemed fit to return to play. These players were cleared to play but they acted less cautiously and demonstrated poor control due to the residual detriment of concussion. Thus, the likelihood they played in the NHL may be affected. Fino et al. found that concussed individuals exhibit larger Dual Task Costs (DTCs) in turning speed and stride time during a planned walking curve marking greater impairment of motor output in those with a prior concussion history than those without [35].

There is also related work exploring the efficacy of machine learning (ML) models in predicting player performance, but present a research gap as they either do not prioritize explainable ML or explainable Artificial Intelligence (XAI) [25, 32, 108] or do not approach prediction from a neuroscience perspective [93, 74, 70]. Explainable AI (XAI) emphasizes model accuracy, fairness, and transparency, and is critical when building trust and confidence in production settings. The alternative to XAI are commonly known as "black-box" models as it is very hard to understand the predictions they make. In this work, ML for sports analytics, explainability is paramount. Our models in practice will be used in tandem with human input to predict whether or

not a player will play in the NHL. Therefore, it is important that model decisions are interpretable as they are used not for outright decisions, but as additional inputs to the traditional NHL scouting process which uses statistics such as goals, assists, shots, player demographics, team environment, defensive acumen, and skating ability. Thus, I present explainable and unexplainable algorithms and discuss trade-offs between their accuracy and explainability.

Mild traumatic brain injury (mTBI) and sports-related concussion (SRC) are prevalent issues as each year the CDC estimates that mTBI accounts for nearly 1.5 million emergency room visits each year with an additional 1.6-3.8 million concussions occurring during sport and recreation activities annually [33]. Thus it is prudent to develop methods of identifying such injuries. Common tests include SCAT5 and Standardized Assessment of Concussion (SAC). SCAT5 is a framework to assess basic memory function, motor, eye, and verbal responses, and "red flag" symptoms. SAC is a brief cognitive test that specifically evaluates orientation, concentration, and memory [5]. However, many SRCs occur without loss of consciousness, memory, or other obvious markers which can make sideline diagnosis difficult, especially since these tests are designed only as preliminary screenings. More intensive tests exist such as ImPACT and computerized tomography (CT) scans, but can be expensive and require prior baseline data. Thus, it is sensible to develop methods of identifying prior concussions.

The task of scouting and drafting players is a challenging task as each year only 49% of players drafted will play at least 1 game in their career [62]. Thus estimating future performance has become a popular area for research with broad applications such as predicting team success based on player actions, player performance in amateur drafts, and player injury prediction [107, 70, 24]. These works, in addition to ours, benefit NHL scouts, team analytics departments, and front-office personnel. Many prior ML approaches use physical metrics or quantitative rankings and statistics to

predict player performance [70, 93, 116, 105]. However, this work takes an alternate approach by using neurological data to predict future performance and concussion instead of physical or quantitative in-game data.

1.2 Research Hypothesis and Questions

The research hypothesis for this thesis is: **Can explainable AI approaches detect and analyze indicators of concussion and its effects on NHL performance in elite athletes?**

Research Questions: As I use metrics from an assessment of visuomotor performance, I break our research into two distinct questions:

- **RQ1:** Are visuomotor performance metrics along with concussion history indicators of an athlete getting a chance to play in an NHL game?
- **RQ2:** Are visuomotor performance metrics reliable indicators of past concussion history?

Problem formulation for RQ1: RQ1 is formulated as a binary classification task where the negative class is labeled with 0 indicating that a player did not play in the NHL and the positive class is indicated by label 1 indicating that the player did play in the NHL.

Research Challenge for RQ1: This is a challenging problem due to the limitations of the dataset that contain only 117 examples or information about the players out of which 60% are the positive class and 40% are the negative class. I choose to balance the dataset using a class weighting approach as the prior probability of any player

playing in the NHL in a given draft year is 49% [61].

Problem formulation for RQ2: This research also investigates the efficacy of our systems to predict whether or not their neurological test scores indicate a prior history of concussions through RQ2. RQ2 is formulated as a binary classification task where the negative class, labeled with 0, indicates no prior concussions, where the positive class is labeled with 1 and indicates *at least one* prior concussion.

Research Challenge for RQ1: RQ2 faces similar constraints due to dataset size and is class imbalanced although less dramatically. The concussion history dataset contains 117 samples of which 53% are of the positive class (prior concussion) while 47% are of the negative class. Thus, the dataset is balanced using a class weighting approach although also investigate classification using without balancing.

1.3 Research Contributions

The novel key contributions are enumerated below:

1. While some prior works use neuroscience scores to assess athletic performance and predict concussion history [4, 84, 124, 102, 15], this work is novel as no studies have used data from this neurological test, developed by Hurtubise *et al.* in [53] named the BrDi™ Test which is a computer-based visuomotor skill assessment task. This test was developed to evaluate the performance of elite athletes with and without concussion on a tracing task designed to test their visuomotor skills.

2. ML models such as Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Linear Tree, Linear Boosting Tree, Light Gradient Boost, Ex-

treme Gradient Boost, and PyTorch Fully Connected Neural Network models were trained and analyzed. Hyperparameter tuning approaches, scaling, and feature selection techniques are employed to predict whether or not a player will play in the NHL and whether or not a player has sustained previous concussions.

3. This thesis presents a human interpretable approach for drafting and performance evaluation for sports analytics by exploring XAI models that allow potential users (coaches, staff, scouts) to trust and understand the results. Ensemble techniques using AdaBoost and Bagging Algorithms were also developed and analyzed. As these models are used in conjunction with human evaluation, I prioritize accuracy and explainability. Weighted Decision Tree tuned using Random Search gives the highest generalization score for classifying players who played in the NHL (RQ1) with significance of **0.603**. Although Weighted Decision Tree tuned using Bayesian Optimization gives higher absolute accuracy of **0.627**, it does not achieve statistical significance as measured using Analysis of Variance (ANOVA). Weighted Decision Tree is also an XAI model and is easy to visualize allowing users to understand, trust, and reason about decisions.

4. For RQ2, weighted SVM gives the highest generalization accuracy for classifying players with previous concussion of **0.612** although does not reach significance. SVM is also an explainable model and allows users to analyze weights to see relevant features. This is discussed in section 4.3.4.

Chapter 2

BACKGROUND: THE BrDi™ TEST, COGNITIVE MOTOR INTEGRATION, AND XAI MODELS

This chapter outlines the relevant background to this thesis including the study from which our dataset is derived, research on Cognitive-Motor Integration (CMI), and summaries of each of the models used. As I prioritize explainable XAI models and compare these against the performance of unexplainable "black box" models, Section 2.3 describes all explainable models used while Section 2.4 describes all "black box" models.

2.1 The BrDi™ Test

This section gives background to the neuroscience experiment, named the BrDi™ Test, from which the data set used in this study was collected. The experiment tests the cognitive-motor integration (CMI) of hockey players using a tracing task and includes multiple variations that mimic motor challenges in hockey scenarios.

The data used in this work is from a study that finds elite hockey athletes exhibit deficits in movement planning, reaction time, and movement accuracy [53]. Hurtubise *et al.* outline an experiment to test CMI in elite hockey players using a tracing task presented in different conditions. The use of a tracing task is a commonly used concussion assessment and has shown in multiple other works to be a good indicator of underlying concussion-related deficiencies in various environments [11, 65, 110]. In the BrDi™ test, participants completed two computer-based CMI tasks on a dual-touch touch-screen laptop. The tasks required participants to slide their index finger

along the screen to move a cursor from a central target to one of four peripheral targets.

To replicate the dynamic physical scenarios in sports, the study employs standard and non-standard conditions during the trials. The standard condition involved participants looking at and reacting to the target presented on the same vertical screen. This is viewed in (A) of Figure 2.1. This resembles an action such as reaching for a coffee cup or shooting at a goal while looking at it. In contrast, the non-standard condition presents two additional challenges, cue reversal, and plane change. During cue reversal, the cursor moves in the opposite direction of the finger movement (180° reversal). During plane change, targets are presented on a vertical screen, and participants respond on a horizontal screen. Part (B) in Figure 2.1 depicts both cue reversal and plane change. This movement is similar to using a brake pedal or a computer mouse. A hockey-specific equivalent of the nonstandard task includes passing the puck to a teammate on one side while navigating an obstacle on the other [53].

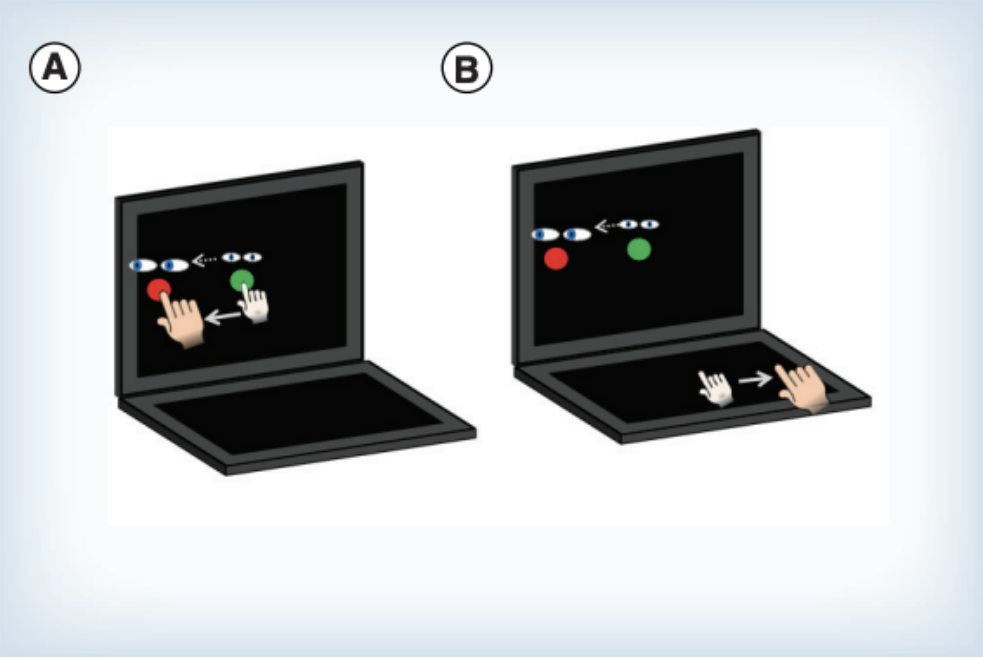

Figure 2.1: Experimental setup used in the BrDi™ test study by Hurtubise *et al.* (2016). Reprinted with permission from [53].

The study was conducted on elite hockey athletes and finds potential lingering deficits of concussion specifically during nonstandard conditions. These findings present similar detriments as seen due to the dual task model and may leave athletes vulnerable to future injury and long-term effects.

2.2 Cognitive Motor Integration

This section details multiple studies outlining how concussions negatively affect coordination between cerebral function and physical movement, i.e., the effect of concussion on Cognitive Motor Integration (CMI).

Baker *et al.* present subjects with a decision to pass through or between two pillars of varying distances. This decision tests participants' medial-lateral center of mass control, essentially their precision of balance, coordination, and decision-making. They show participants with prior history of concussion choose to pass between smaller gaps than those without. This indicates they act less cautiously, exhibit larger variance in their decisions, and demonstrate poor center of mass control. It also has implications for on-field performance and injury as it was conducted with individuals already cleared to play by current concussion Return to Play (RTP) standards [11].

This result is corroborated by a study that uses the dual-task model to expose underlying CMI deficiencies in hockey players. The dual-task model is a method in neuropsychology that requires participants to complete two tasks simultaneously producing challenges called Dual Task Costs (DTC).

Fino et al. found that concussed individuals exhibit larger Dual Task Costs (DTCs) in turning speed and stride time during a planned walking curve marking greater impairment of motor output in those with a prior concussion history than those

without [35]. Specifically, previously concussed participants exhibit slower walking paces and less body inclination indicating a cautious adaptation to increase stability during the task. This reduced ability to perform physical movements may lead to a greater risk of injury in fast-paced athletic environments and in turn, affect the likelihood of playing in the NHL. Furthermore, they find that although they reported no post-concussion symptoms and performed similarly to non-injured controls on a cognitive assessment conducted up to 58 days after the concussion, ice hockey players who had suffered a concussion displayed deficiencies in their ability to perform dual-tasking ice hockey-specific activities [98, 35].

2.3 Explainable AI Models

This section provides a brief explanation of explainable models and why they are useful to this research.

XAI models are methods that allow humans to understand, reason about, and trust outputs [82]. Explainability helps portray accuracy, transparency, and fairness during the decision process. They also allow those who are affected by model decisions to understand and challenge these decisions. As AI becomes more advanced, even those who created the models are unable to reason about results. The alternative to explainable models are commonly referred to as "black box" models and are difficult or impossible to interpret.

To give a hockey-specific example, consider a scout using our classifier to predict whether or not a player will play in the NHL. For a player predicted to play in the NHL, it would be impossible to understand the contributing features using a black-box model. In contrast, an explainable model would show the scout which features are important to the classification (features important to playing in the NHL). It

would therefore be possible to use the model for player development by identifying important features to the classification and creating relevant training goals. Thus the advantage of XAI in this application is clear and may also be applied to injury prevention, player acquisition, and even lineup creation.

2.3.1 Decision Tree

Decision Trees (DT) are supervised XAI models commonly used for both classification and regression tasks and are known for their interpretability, intuitive architecture, and strength when used with both categorical and numerical data. Decision Trees are composed of branches and nodes where the bottom nodes of the tree, containing classification information, are called leaf nodes. At each internal node, an attribute test is performed, based on a splitting criterion, to identify the best way to partition features of the dataset and develop a hierarchy of features from which classifications are made.

Common attribute tests include Gini Index and Information Gain. The Gini Index measures the probability of misclassification for a randomly chosen feature. Therefore, the lower the Gini Index, the lower probability of misclassification. The formula for Gini Index can be given as

$$\text{Gini Index} = 1 - \sum_{i=1}^{j} (p_i)^2 \tag{2.1}$$

where j is the number of classes and p_i represents the probability of an instance belonging to class i, Played in the NHL or Did Not Play in our RQ1 example.

Information gain is the change in Entropy between parent and child nodes where Entropy is defined as the average level of "uncertainty" inherent to the variable's possible outcomes. The formulas for Information Gain and Entropy are given below.

$$\text{Information Gain} = \text{Entropy}_{\text{Parent}} - \text{Entropy}_{\text{Child}} \tag{2.2}$$

$$\text{Entropy} = -\sum_{i=1}^{j} (p_i)^2 \log_2(P_i) \tag{2.3}$$

One of the advantages of DTs is their ability to provide an understanding of the importance of features by measuring each feature's value to the overall classification ability of the model. This allows users of the model to interpret decisions and easily understand factors important to classifications. Specific to our use case, DTs allow model users to identify exactly which BrDi™ scores were used during classification. This provides high interpretability and explainability when reasoning about classifications as decisions are represented in a clear and intuitive manner. A visual depiction of a DT can be seen in Figure 3.4.

2.3.2 Linear Tree

Linear Trees (LT) are variations of Decision Trees that contain linear classifiers, such as Logistic Regression models, at leaf nodes. LTs maintain a similar structure to Decision Trees in which nodes and branches are formed using a measure of feature importance. However, instead of providing class labels at the leaves, LTs have models to create classifications from the points falling to each leaf node. This gives LT the advantage of being able to identify more complex relationships that may not be captured with simple splits. LTs combine the advantages of both cohort-based (Decision

Tree) and regression-based approaches. Additionally, model trees take advantage of ensemble learning by utilizing multiple regressors specific to a single partition of the feature space. This is advantageous as feature groupings are learned automatically and do not require a similarity metric.

2.3.3 Random Forest

Random Forest (RF) is another variation of a Decision Tree in which a group, also known as an ensemble, of Decision Trees is formed to create a classifier that harnesses the power of multiple DTs to make classifications. RF forms an ensemble of DTs using bootstrap sampling which ensures that each Decision Tree is trained on a different subset of the data and also introduces diversity during each split by considering a random subset of features so ensure differences between trees.

Consensus classifications are made by a majority voting scheme among individual DTs. This has the effect of creating a robust classifier and increasing generalization accuracy. Additionally, while individual trees may overfit, the aggregate nature of RF helps reduce the probability of overfitting.

Similar to DT, RF calculates feature importance and finds the average decrease in impurity or the average reduction in error across all the trees. This information can be used to gain insight into feature importance and measure the relative importance of features to a classification.

The ensemble nature of RF also lends itself to parallelization to decrease training time on large datasets. While not necessary with our small dataset of 117 rows, parallelization is a common approach to larger-scale problems.

2.3.4 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that uses the concepts of margin and decision boundaries to separate data points into classes. SVM is adept at giving high generalization accuracy for small datasets as the algorithm uses only a subset of the data (named support vectors) to create the decision boundary.

The goal of SVM is to find the hyperplane which maximizes the distance, called the margin, between the hyperplane and the closest data points from each class. This distance is calculated using the support vectors which lie close to the boundary. SVM is able to create decision boundaries of varying shapes via kernel functions. Kernel functions map the input features to higher-dimensional feature spaces where there may be a better separation between targets in the case that features are not linearly separable. As viewed in Table 3.3, the optimal kernel is chosen as linear as the classes may be linearly separable in a higher dimension feature space.

SVM also balances the tradeoff between maximizing the margin and minimizing the classification error using a regularization parameter C. A higher C value results in a smaller margin although potentially fewer misclassifications, while a lower value of C allows for a larger margin at the risk of a greater number of incorrect predictions.

2.3.5 Logistic Regression

Logistic Regression is a supervised machine-learning algorithm for binary classification tasks known for its simplicity and versatility. Logistic Regression models the relationship between features and targets using a linear relationship and can be extended using penalties (L1, L2, Elastic Net) to the loss function.

L1 Regularization (Lasso Regression) adds a penalty term to the LR loss function that encourages the model to have many coefficients that are exactly zero. This helps to reduce the complexity of the model and can prevent overfitting. L2 Regularization (Ridge Regression) adds a penalty term to the loss function that encourages coefficients to remain small instead of exactly zero as in L1 regularization. L2 regularization also helps reduce overfitting and improves generalization performance. Elastic Net is a combination of L1 and L2 regularization techniques that balance the advantages of L1 and L2 regularization and can overcome their limitations.

2.4 Black Box Models

2.4.1 Linear Boosting Tree

Linear Boosting Tree (LBT), is a "black box" supervised model that uses the boosting algorithm to iteratively train an ensemble of weak linear models to form a more robust classifier.

The LBT algorithm is a two-stage process. First, a weak learner is trained on the dataset to make classifications. Second, the errors from the weak learner are input into a decision tree, and the path to the worst leaf is identified. The leaf most influential to the error is used to create a new feature and the algorithm continues from the first step until convergence is achieved. This has the effect of correcting mistakes weak learners make.

LBTs are adaptable in that better-performing models are more heavily weighted in the final classification and weights are also used to focus on difficult-to-classify samples. While prone to overfitting, LBTs also provide hyperparameters to control the complexity of the trees which can help reduce overfitting. This adaptive weighting

also helps address the class imbalance by assigning higher weights to misclassified instances and can help in learning from the minority class. Thus dealing with class imbalance may not be necessary.

While this paper considers LBTs to be "black box" due to their ensemble nature, which is less intuitive than Random Forest, they maintain some level of interpretability to do the use of linear weak learners. While explainability is not as readily available as other XAI models, coefficients of the linear models may still be used to understand the effect of features in the classification process.

2.4.2 Feed Forward Neural Network

Feed forward (FNN), or deep learning (DNN) neural nets, are supervised "black box" algorithms that consist of layers of connected nodes called neurons. Information flows in the forward direction through layers in which each layer calculates a weighted sum of its inputs that is passed to an activation function. This has the effect of introducing non-linearity into the network making DNN effective at capturing complex relationships between features and targets. While there are many frameworks and their corresponding implementation details for creating DNNs, I implement the network used in this work using PyTorch.

In PyTorch, there are 4 key steps to training a DNN: network architecture, loss function, optimizer, and updating weights. Network architecture is first defined using a Python class-based interface. Next, a loss function, specific to the application, must be chosen. In our work, I choose Binary Cross Entropy Loss which is a more numerically stable combination of a sigmoid activation layer and Binary Cross Entropy Loss. Binary Cross Entropy Loss is a common loss function for binary classification as it optimizes the model to estimate accurate probabilities for the positive class, penal-

izes incorrect predictions effectively, and provides efficient gradients for optimization, leading to improved classification performance.

The next choice during development is an optimizer. Optimizers control the algorithm with which weight updates in the Network are performed and implement various techniques to aid convergence accuracy and speed. There are many different types of optimizers such as Adagrad, Adam, and RMSProp.

Finally, weight updates must be performed to backpropagate weights from the output layer through each previous layer using a recursive chain rule. PyTorch does this automatically using its autograd system and thus backpropagation is a simple function call.

2.4.3 Extreme Gradient Boost

Extreme Gradient Boost (XGB), is a popular and powerful machine learning model that draws on the boosting algorithm to sequentially build an ensemble of Decision Trees where each tree is an improvement on the errors of the last tree. XGB utilizes a regularized learning objective that consists of a loss function and penalty term to aid in overfitting and generalization performance. XGB is an improvement to the Gradient Boosting Algorithm as it contains L1 and L2 regularization terms, is scalable due to distributed training and efficient algorithms to find feature splits, and is able to handle missing values. XGB does contain methods for understanding feature importance although I consider it a "black box" algorithm due to its use of an ensemble of Decision Trees.

2.4.4 Light Gradient Boost

Light Gradient Boost (LGBM), is a supervised machine learning model that takes advantage of the gradient boosting algorithm in a similar fashion to LBT and XGB. LGBM is different from XGB as it performs feature splits in a leaf-wise fashion as opposed to XGB's depth-wise split. A depiction of leaf-wise vs depth-wise splits can be seen in Figure 2.2. LGBM is also significantly faster and uses less memory than XGB due to its histogram-based tree construction in contrast to the exact splitting algorithms used in XGBoost.

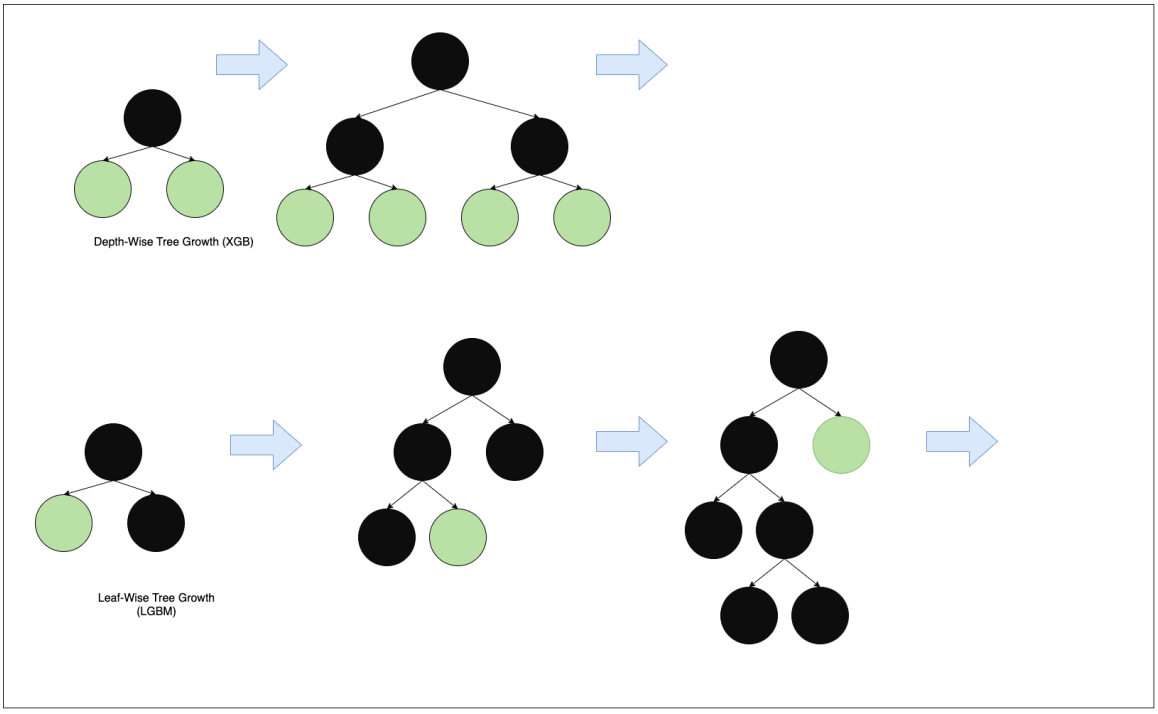

Figure 2.2: Depth-Wise vs Leaf Wise Tree Growth.

2.5 Methods

2.5.1 Data Preparation

This section outlines separate data preparation steps performed before training models.

The original dataset is 150 rows each with 57 features. Features listed data such as randomized id, age, date of birth, height, weight, concussion history (Yes / No), and number of prior concussions. However, I cannot mention players by name as the collected data is anonymized. Additionally, 41 different metrics were collected during the neuroscience test [53]. Many columns contained missing values and while the usual practice would be to fill missing values using the arithmetic mean or other technique, I reason this is not appropriate for this dataset as it is collected from humans. Thus, any metric with missing values is removed. After column removal, any remaining rows with missing values were dropped. Lastly, non-predictive columns such as a player's initials, position, and shooting dexterity were removed leaving a dataset containing 117 rows and 39 features. Splitting data into training and generalization sets gives 93 training rows and 24 generalization rows.

2.5.2 Class Balancing

In this work, the target variable of playing in the NHL contained a class imbalance of 60% who did play in the NHL and 40% who did not play in the NHL. Due to the limited sample size of ≈ 100 rows, I considered popular oversampling algorithms seen in prior works such as Synthetic Minority Oversampling (SMOTE). SMOTE is one of the most popular methods of oversampling and works by imputing new data points for already existing ones. Specifically, SMOTE chooses the k nearest data

points to each point in the minority class, randomly chooses one of the k neighbors, and then creates a synthetic example by creating a point on the line connecting the random neighbor and minority point [22]. This process is repeated until the desired distribution is found. I reason it is not logical to create human data in such as fashion and do not use SMOTE.

Instead, this work chooses a class weighting which is a method of dealing with class imbalance without inputting or removing data points. Class weighting assigns weights to classes in the training dataset such that errors during model training of the minority class are more heavily penalized in the loss function. The formula to compute class weight is given by

$$w_j = \frac{n}{k \cdot n_j} \tag{2.4}$$

where w_j is the weight of class j, n the number of samples, k the number of classes, and n_j is the number of samples in class j. As it relates to our dataset, j represents the binary classification (1: Played in the NHL, 0: Did *not* play in the NHL) and therefore n_j equal to 69 samples for the positive class (Played) and 48 for the negative class (Did not play).

I also investigate undersampling to address the imbalance, however undersampling involves removing data points. Due to the already limited nature of the dataset, undersampling produces worse results than class weighting.

2.5.3 Feature Scaling

This section describes 3 methods of feature scaling used in the pipeline including StandardScaler, MinMaxScaler, and RobustScaler. StandarScaler is a common method of standardization, also known as z-score standardization, that transforms features to have a mean of 0 and a standard deviation of 1. It can help the convergence of algo-

rithms such as Linear/Logistic regression and Support Vector Machine (SVM) that are sensitive to scale although assumes input features are normally distributed. MinMaxScaler scales data using feature minimums and maximums to a range between 0 and 1. Like z-score standardization, it can aid the convergence of certain algorithms but is also sensitive to outliers. Lastly, RobustScaler is a method of scaling that is robust to outliers in the data. It scales each feature by subtracting the median and dividing by the interquartile range (IQR). It scales data points to a mean of 0 and variance of 1 while centering data around the median and scaling it between the 25^{th} and 75^{th} percentiles (IQR). It is useful in situations where distributions are skewed or outliers exist. The formulas for each of the three scaling methods are given below:

StandardScaler:

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{2.5}$$

MinMaxScaler:

$$X_{scaled} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{2.6}$$

RobustScaler:

$$X_{scaled} = \frac{X - median(X)}{IQR} \tag{2.7}$$

where X refers to a feature (measurement variable) μ refers to the arithmetic mean, σ is the standard deviation, and IQR is the interquartile range, i.e., the range between the 25^{th} and 75^{th} percentiles of the data.

2.5.4 Feature selection

Feature selection is the process of determining which features are most important to make a prediction. Small data sets with many features can pose challenges to training

machine learning models so it can be useful to remove unimportant features before training.

As part of tuning, I employ 3 different types of feature selection and feature reduction: feature selection using Linear Support Vector Machine (Linear SVM), ExtraTreesClassifier, and direct selection of 5 features found to be significant during the BrDi™ Test.

Feature selection using Linear SVM and ExtraTreesClassifier are two methods of using the learned weights from a machine learning model to identify important features in the data. Linear SVM is an algorithm that is commonly used for classification and regression tasks. Linear SVM assigns weights to each feature based on its importance to the model making it advantageous for feature selection. These weights can be used to identify relevant features allowing for better model explainability and performance. Similarly, ExtraTreesClassifier is an ML algorithm that constructs multiple decision trees on random subsets of the data and aggregates their predictions to make a final prediction. It can also be used to weight features and perform feature selection. Linear SVM may be preferred when dealing with linearly separable data or when interpretability is important, while ExtraTreesClassifier may be preferred when dealing with high-dimensional data or when non-linear relationships between features and the target variable are suspected.

2.5.5 Scoring

This section discusses the scoring methods used when evaluating classifiers.

Models were tuned using two methods: Sklearn RandomSearchCV and Optuna GridSearchCV (Bayesian Optimization using the Tree-Structured Parzen Estimator (TPE)

algorithm) to search over a parameter space to achieve the best generalization accuracy. I collect three different accuracy metrics: precision, recall, and F1.

Precision, the proportion of positive predictions that are actually correct can be written as $precision = \frac{TP}{TP+FP}$ where TP is the number of true positives and FP is the number of false positives.

Recall, the proportion of correctly identified positive samples can be formulated as $recall = \frac{TP}{TP+FN}$ where TP is the number of true positives and FN is the number of false negatives.

F1 is used to score the models as it balances the tradeoff between precision and recall. F1 score is the harmonic mean of precision and recall and is a better measurement than accuracy, precision, or recall as it represents the balance between correct positive predictions (precision) and correctly identifying positive samples (recall). Accuracy, as measured by the F1 score, is given by the formula:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

I report the weighted average of the F1 score between the positive and negative class. Weighted averaging of scores uses the support (number of samples) for each class to give a more accurate representation of results for imbalanced datasets. Without using the weighted average, higher results might be achieved by models predicting the positive class due to the class imbalance. The formula for calculating the weighted average between the two classes is given below. Subscripts $_0$ and $_1$ denote the two classes of our binary classification: did not play and played in the NHL respectively. TP, FP, FN, and FP are the number of true positives, false positives, false negatives, and false positives.

$$F1_0 = \frac{2 \cdot TP_0}{2 \cdot TP_0 + FP_0 + FN_0}$$

$$F1_1 = \frac{2 \cdot TP_1}{2 \cdot TP_1 + FP_1 + FN_1}$$

The weighted average F1 score is then calculated as:

$$F1_{weighted} = \frac{support_0 \cdot F1_0 + support_1 \cdot F1_1}{support_0 + support_1}$$

where $support_0$ and $support_1$ are the number of samples in the two binary classes, respectively.

I also report precision and recall as some of our models do not achieve ideal training results. Specifically due to our small and imbalanced dataset, many models suffer from overtraining, a bias toward predicting the positive class. Precision is an important metric in this instance as it penalizes models that incorrectly predict the positive class.

2.5.6 Hyperparameter Tuning Frameworks

This section outlines the method by which I construct the prediction pipeline and perform distributed hyperparameter tuning on each model.

There exist various algorithms for finding optimal hyperparameters. In this work, I choose Random Search and Bayesian Optimization using the Tree-Structured Parzen Estimator (TPE) algorithm. Random Search is a method of searching a parameter space in a random fashion. I choose Random Search over Grid Search, searching

all combinations of parameters, as it has been shown that Random Search gives as good or better results than Grid Search by taking advantage of larger parameter configuration space [12]. Random search also uses less computational power than Grid Search. Bayesian Optimization using the TPE algorithm is also implemented as it has been shown to produce better tuning results than Random Search. It is a non-parametric algorithm that models an objective function (computes training accuracy) using probability density functions (PDF). It consists of two phases: exploration and exploitation. In the exploration phase, TPE builds a "good" PDF that models the distribution of hyperparameters that lead to high accuracy scores and a "bad" PDF that leads to low accuracy scores. It then samples points from each PDF to evaluate the objective function. In the exploitation phase, TPE builds a new PDF from the best-performing data points which is used to sample new points for the next iteration of the algorithm.

To find optimal hyperparameters, I create a tuning pipeline and explore the parameter search space using the two aforementioned methods: Random Search and Bayesian Optimization using the TPE algorithm.

The pipeline steps are feature scaling, feature selection, and model prediction. I implement scaling before selection to eliminate biases during selection as some techniques may be sensitive to feature scale. I also test models on an independent hold-out set (generalization set) not used in training to ensure no data is leaked from training to generalization.

To manage limited data while ensuring maximum confidence in generalization results, I repeat tuning for 5 iterations using different random seeds to produce different training/generalization splits. An example of the distributed architecture can be viewed in Figure 2.3.

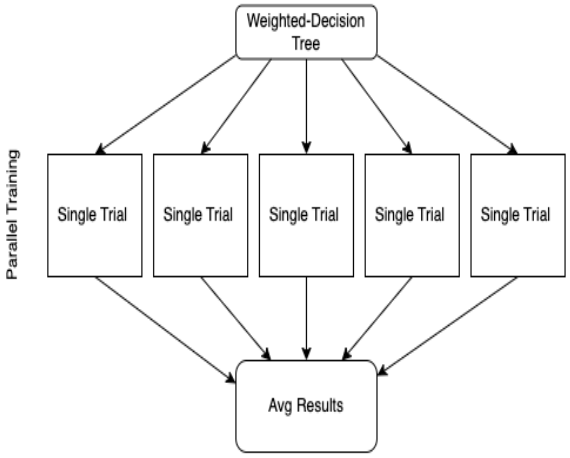

Figure 2.3: Distributed training architecture where each Single Trial contains a pipeline seen in Figure 2.4.

Training is distributed in regard to each seed to ensure minimal training time. Parameter search is performed on 80% of the data using 10-fold cross-validation while 20% is reserved for generalization. Generalization and training splits are stratified to ensure an even distribution of classes across each. To find the right balance of training to generalization size, I compare model results of three different train/generalization ratios (70/30, 80/20, 90/10) and find that 80/20 gives a good balance of generalization and training size. A visual depiction of the pipeline can be viewed in Figure 2.4. In this figure, we describe the pipeline visually including each of the following steps: feature scaling, feature selection, cross-validation, Random Search and Bayesian Optimization, and result collection. The pipeline is contained in each of the "Single Trials" given in 2.3.

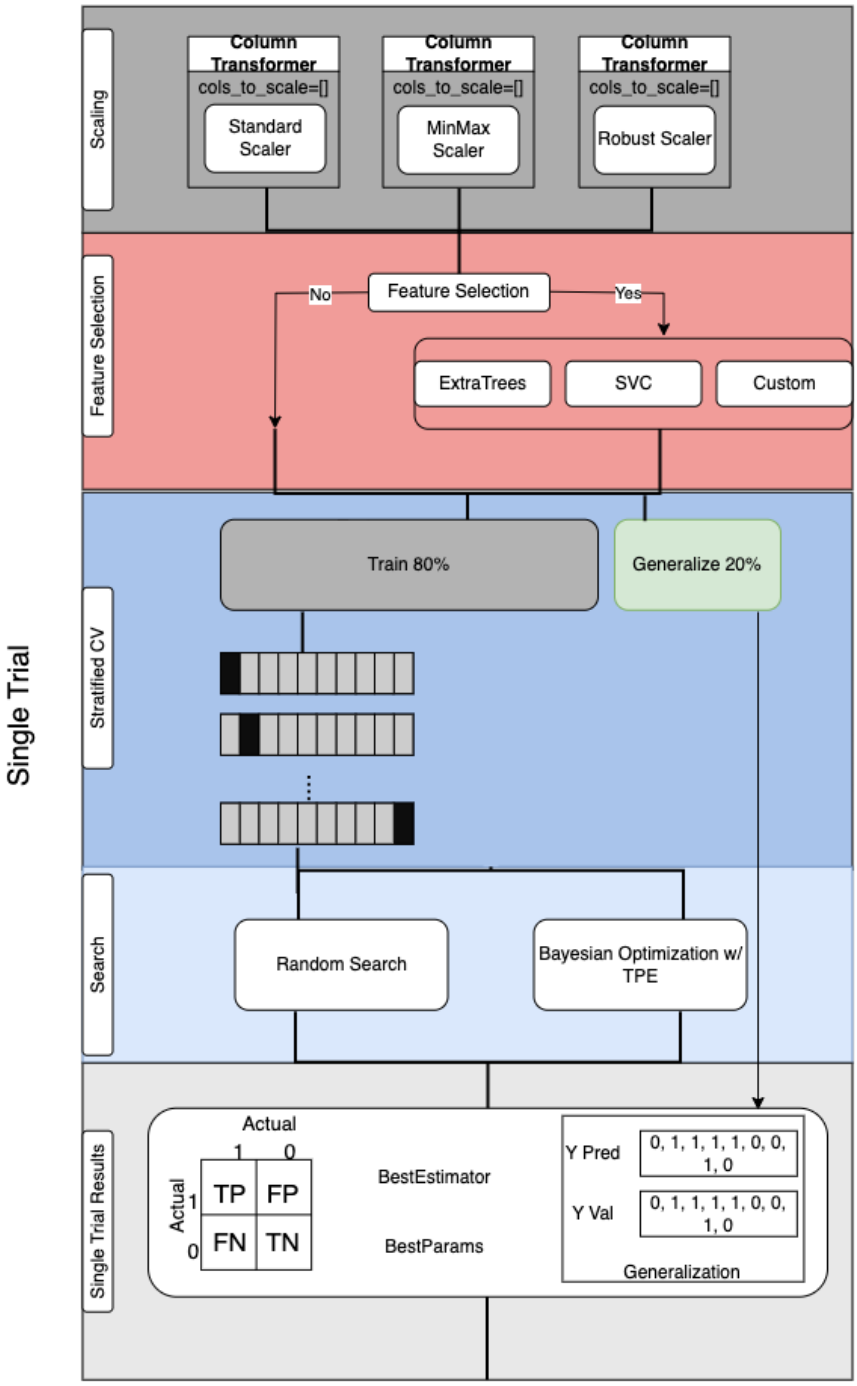

Figure 2.4: Pipeline architecture including feature scaling, selection, parameter search, and results.

Chapter 3

RQ1: ARE VISUOMOTOR PERFORMANCE METRICS AND CONCUSSION HISTORY INDICATORS OF AN ATHLETE PLAYING IN AN NHL GAME?

3.1 Literature Review

This section details ML approaches to hockey performance prediction. Many hockey studies are performed on small detailed datasets [40] and as modern deep learning requires many training examples, standard ML algorithms, such as regression, are popular choices for hockey analytics. Additionally, there has been a lot of progression in the applications of ML using physiological signals [93, 74, 70, 25, 32, 97]. However, the application of ML with visuomotor (neurological) signals in hockey analytics has yet to gain significant momentum.

Many studies have focused on the effect of physiological measures as predictors of performance in the NHL. For example, multiple linear models were trained in [25] on input features such as VO_2max, fatigue index, grip strength, and bench press from the NHL combine. Models were used to predict points scored, time on ice (TOI), time to play in the first NHL game, and career seasons played. Players were separated by positions to control for the positional differences of two target variables TOI and points scored. Results show the potential for individual metrics as a predictor of career success as model accuracy ranged from 2% to 16%.

Lojain *et al.* performs similar research combining demographic, anthropometric, and traditional hockey statistics as inputs to a regression model (Zero-Inflated Negative Binomial (ZINB)) to predict NHL draft order [32]. They compare the effectiveness of features such as defensive awareness, puck protection ability, shooting ability, work

ethic, age, height, and weight to predict games played in the NHL. They find height and weight to not be predictors of future performance in the NHL while age is. They also find that the ability to scan the ice (denoted by vision and seeing the ice) and good "hockey sense" make a player significantly less likely to play in zero NHL games. Both scanning the ice and hockey sense presents a dual-task cost (DTC) which are known to be affected by concussion. This provides further evidence that concussions may affect a player's chances of playing in at least one game.

The authors in [74] created player rankings using expected goals per 60 minutes using ordinary least squares (OLS) and Ridge regression. Hockey analysis in general is limited by relatively low scoring rates. This work uses Fenwick and Corsi ratings, statistics derived from combinations of shots, missed shots, and blocked shots, with other significant predictor variables goals, shots, hits, hits against, and faceoffs. It is novel as it uses Ridge regression and compares the model to results using only goals, shots, Fenwick, and Corsi statistics as predictors of expected goals. It achieves the best performance using Ridge regression which may be explained by Ridge regressor's effectiveness in handling colinearity within predictors.

Perrson *et al.* [93] investigated the effectiveness of different models in predicting the top 10, 25, and 50 percent of players using hockey statistics and NHL video game scores. Data was split by position and features were selected using both a wrapper and filter method and SMOTE to balance classes. Filter methods select attributes with correlation to the target variable but not each other while wrapper methods use supervised learning to test subsets of features. They compare the efficacy of 6 different models: Logistic Regression, Naive Bayes, Bayesian Network, Decision Tree, K-Nearest Neighbor, and Random Forest. Overall, both Bayesian classifiers performed the best. The study is limited by a lack of model parameter tuning and the challenge of finding good defensive metrics.

Work done by [70] uses demographic (age, weight, height), performance (goals scores, plus/minus), career statistics (games played, time on ice), and scouting rank (CSS) to predict the number of games played in a player's first 7 years. Predicting games played in a player's first 7 years is a common approach in hockey analytics [105, 106, 120] as it is the amount of time a team maintains contract control of a player before they are able to reach free agency [2]. Results are measured as a correlation between rankings given by the outputted probability of playing in games and games played in 7 years and the correlation between draft number and games played in 7 years. Rankings given by the outputs of Linear Trees produce a correlation almost twice as strong as ranking players by their draft number (.83 vs .43) while also giving immediate insight into how a player compares to his cohort and what makes him stand out.

This work values the explainability of models as they are used in conjunction with domain expert opinions. Features that make players exceptional can be found by examining feature vectors and feature weights. The feature vector for a single player is given by $(x_1, x_2, \ldots, x_i)$, the average feature vector for all players $(\bar{x_1}, \bar{x_2}, \ldots, \bar{x_i})$, and learned weights from a leaf's regression model $(w_1, x_2, \ldots, w_i)$ for group g. The log odds difference between player i is a randomly selected player in his cohort is given by

$$\sum_{m}^{j=1} \bar{w_j}(x_{ij} - \bar{x_{gi}}) \tag{3.1}$$

This sum provides a measure of the player's rank in comparison to others within the group. By maximizing the value of $\bar{w_j}(x_{ij} - \bar{x_{gi}})$, I can identify the player's strongest features.

Lawrence et al. introduced the Prospect Cohort Success (PCS) in which the groups of players are formed using the Euclidian distance between 4 variables (age, height,

points per game, league) which have been shown to be statistically significant for Canadian Hockey League players. Predictions about future player performance (play in > 100 or > 200 games) are made based on a player's closest peers and used to reference what type of career a player may have. For example, if 22% of a player's peers played 100 games in the NHL, the PCS value for that player is 22% [97]. Lawrence et al. show this method outperforms multivariate regression of the aforementioned variables and is significant in 12/15 global hockey leagues. Additionally, PCS explains more variance than a regression performed in leagues that contribute roughly 80% of NHL players.

Bryson et al. group players by cohort size before performing regression for games played, salary, and points [16]. The grouping of players by cohort size and age is employed to explain how two economic theories (Easterlin's Hypothesis and Welch's theory) affect NHL player labor market outcomes. Grouping players in this way provides explainability to regression models used for prediction.

Schulte et al. mimic the way scouts watch games by grouping and comparing players by style of play. Players are first clustered in an unsupervised manner by style of play using on-ice location data before their impact on scoring the next goal is predicted. Scoring impact predictions are outputted using a Markov decision model [108]. This approach captures useful aspects such as game context (score), game trajectories, and player impact at specific game states although only significant *after* clustering occurs [94].

Gliznitsa *et al.* [40] used Decision Trees to classify hockey players into above or below-average productivity. Productivity is measured by dividing a player's time on ice (TOI) and their points during that time. Decision trees proved effective as they were able to achieve a productivity classification accuracy near 80%. Due to the explainable nature of Decision Trees, researchers are able to draw two important conclusions.

First, they can make a claim that Decision Trees are 80% accurate at classifying players into tiers of productivity based on features such as goals, assists, and years of experience. Second, they can conclude which features were most important by examining the feature splits of the decision tree. They are able to make arguments such as assists per game are more important than goals per game for productivity, and that the highest productivity is linked with only moderate levels of experience. This is important as scouts are able to understand the decisions of these models, and also include model reasoning in their own decisions. Insights from these XAI models also help build trust in their decisions and provide explanations to reason about misclassification.

This thesis draws inspiration from these examples in an attempt to build models that are high in accuracy and also provide explanations for their decisions. As noted, many prior works use physiological features, prior statistics, demographics, or a combination of all three as features for a model. This work highlights a shift from physiological to neurological features such as the scores from the BrDi™ experiment. Traditionally, sports performance has been predicted using physical attributes. Yet, more and more emphasis is placed on the role of the brain in on-field decision-making and output. This work explores predictors that use neurological signals, and detriment to these signals (via concussion), which I hypothesize transitively affect on-field performance and output.

This thesis also identifies two trends in research for predicting hockey performance in that many works use either cohort-based approaches, in which a player is compared to his peers or models in which comparisons of features are made, or regression-based approaches, which emphasize simplicity and create predictions by fitting models to data. Specifically, the authors in [40, 94, 97] all develop classifiers by comparing or grouping players. In contrast, [70, 93, 74, 32] each fit regression-based models to data.

3.2 RQ1 Results

This section presents the results our work. As this is preliminary research using data from the BrDi™ test, I survey a wide array of ML models. Parameter tuning is implemented for 6 explainable models (Decision Tree, Random Forest, Linear Tree, Logistic Regression, Elastic Net, Support Vector Machine (SVC)), and 4 black box models (Extreme Gradient Boost (XGB), Light Gradient Boost (LGBM), Linear Boosting Tree, Deep Learning Neural Network (DNN)). I also include ensemble learners of weak classifiers using bagging and AdaBoost algorithms.

Figure 3.1 is presented to illustrate the flow of data specific to RQ1 throughout our system.

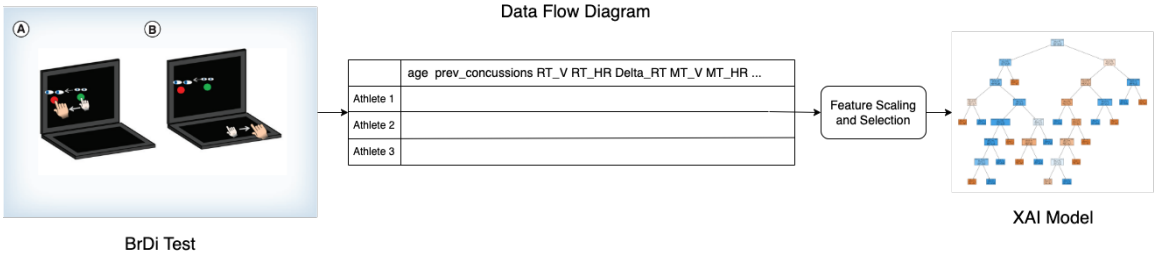

Figure 3.1: Flow of data from BrDi™ test, to tabular DataFrame, to feature scaling and selection, and an ML model.

3.3 RQ1 Random Search

Table 3.1: Average generalization results over 5 trials of tuning using Random Search.

	Train			Generalize		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Model						
Weighted-DecisionTreeClassifier	0.725	0.716	0.713	0.611	0.608	**0.603**
Weighted-LinearTreeClassifier	0.645	0.637	0.639	0.609	0.592	0.593
Weighted-LinearBoostClassifier	0.727	0.718	0.720	0.550	0.533	0.533
Weighted-NeuralNetClassifier	0.621	0.622	0.572	0.549	0.575	0.526
Weighted-SVC	0.669	0.658	0.656	0.530	0.525	0.522
Weighted-LogisticRegression	0.636	0.611	0.613	0.535	0.517	0.514
Weighted-RandomForestClassifier	0.740	0.731	**0.732**	0.519	0.517	0.513
Weighted-ElasticNet	0.644	0.624	0.626	0.534	0.508	0.501
Weighted-LGBMClassifier	0.732	0.714	0.716	0.491	0.475	0.473
Weighted-XGBClassifier	0.350	0.591	0.440	0.340	0.583	0.430
Mean	**0.649**	**0.662**	**0.643**	**0.527**	**0.543**	**0.521**

Results for single classifiers can be seen in Table 3.3. Each model is class balanced using the formula given by EQ. 2.4 and the table is sorted by generalization F1 score. Observe that the Weighted Decision Tree produces the best generalization accuracy (weighted F1 score) of 0.603. It is also of note that Weighted-Decision Tree finds the best generalization accuracy using *all features* proposed in this study as feature selection (None, StandardScaler, MinMaxScaler, RobustScaler) is included as part of the tuning process. Trials are repeated using different random seeds to split data into train, test, and generalization sets in order to give more confidence in model performance. The method of repeating trials is chosen in order to keep as much data as possible for training while still achieving robust generalization trials.

3.3.1 RQ1 Random Search ANOVA

Figure 3.2 shows the results of a one-way Analysis of Variance (ANOVA) for $p < .05$ to compare generalization scores between models tuned using Random Search. ANOVA is a statistical test to compare variances of arithmetic means. In this study, the results of an ANOVA test with $p < .05$ tell us that one of the means of the generalization set results is different with statistical significance. ANOVA is performed using the mean sum of squares and the mean sum of squared errors between generalization results. Weighted Decision Tree performs the best on the generalization set with statistical significance (p=.0013) when tuned with Random Search over 5 trials.

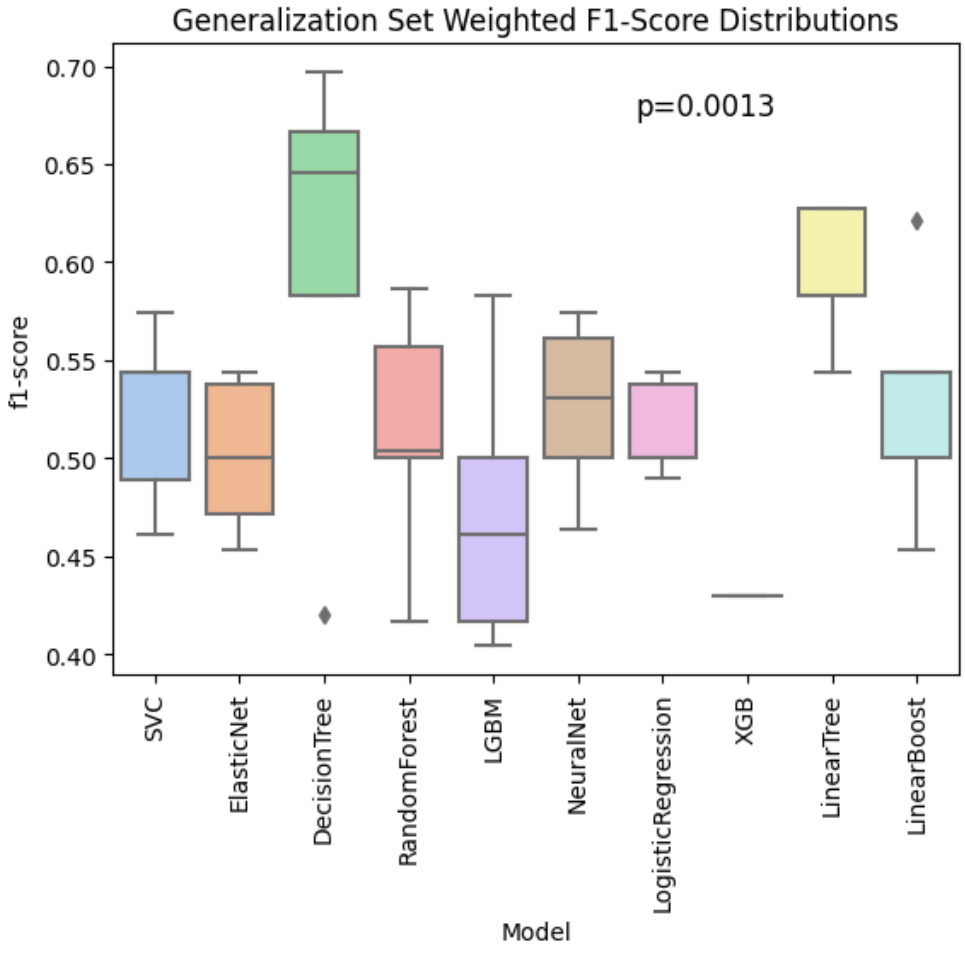

Figure 3.2: ANOVA of generalization results using Random Search.

Weighted Decision Tree tuned with Random Search is selected as the preferred model since ANOVA results using Bayesian Optimization did not achieve statistical significance ($p < .05$). Thus, Weighted Decision Tree tuned using Random Search is the best classifier and is significantly more accurate than the other models.

3.3.2 RQ1 Random Search Black Box vs XAI Models

The difference in accuracy scores between explainable and "black box" models when tuned using Random Search is further examined. Results are seen in Table 3.2. Notice that explainable models produce better results when comparing absolute and mean F1 scores. Although not depicted, similar results hold for tuning using the TPE algorithm as well. As Weighted Decision Tree is the model with the highest generalization accuracy and is also explainable, I conclude it is best suited for this task. However, if it was not, the selection of the best model would create a decision to balance F1 score and explainability and would ultimately depend on how much users value accuracy vs trust, transparency, and fairness.

Table 3.2: Avgerage generalization F1 scores of explainable and "black box" models over 5 trials using Random Search.

Explainable		**Black Box**	
Model	**F1-Score**	**Model**	**F1-Score**
Weighted-DecisionTreeClassifier	**0.603**	Weighted-LinearBoostClassifier	**0.533**
Weighted-LinearTreeClassifier	0.593	Weighted-NeuralNetClassifier	0.526
Weighted-SVC	0.522	Weighted-LGBMClassifier	0.473
Weighted-LogisticRegression	0.514	Weighted-XGBClassifier	0.430
Weighted-RandomForestClassifier	0.513	NaN	NaN
Weighted-ElasticNet	0.501	NaN	NaN
Mean	**0.541**	**Mean**	**0.491**

Table 3.3: Best parameters and score for a single trial for models trained using Random Search.

Model	Best Params	F1-score
DecisionTreeClassifier	{'selector': RandomUnderSampler(random_state=42), 'scaler': 'StandardScaler', 'model__min_samples_split': 13, 'model__min_samples_leaf': 11, 'model__max_features': 'sqrt', 'model__max_depth': 4, 'model__criterion': 'gini', 'model___weight': 'balanced'}	**0.696**
LinearTreeClassifier	{'selector': 'None', 'scaler': 'MinMaxScaler', 'model__n_jobs': -1, 'model__min_samples_split': 80, 'model__min_samples_leaf': 100, 'model__max_depth': 3, 'model__max_bins': 10}	0.626
LinearBoostClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__n_estimators': 20, 'model__max_depth': 1, 'model__ccp_alpha': 0.1}	0.621
RandomForestClassifier	{'selector': 'CustomFeatureSelector', 'scaler': 'None', 'model__n_estimators': 100, 'model__min_samples_split': 6, 'model__min_samples_leaf': 8, 'model__max_features': 'log2', 'model__max_depth': 2, 'model___weight': 'balanced'}	0.586
LGBMClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__subsample': 0.5, 'model__num_leaves': 2, 'model__num_iterations': 300, 'model__min_data_in_leaf': 30, 'model__min_child_samples': 7, 'model__max_depth': 7, 'model__max_bin': 50, 'model__learning_rate': 0.05, 'model__early_stopping_rounds': 3, 'model__colsample_bytree': 0.5, 'model___weight': 'balanced', 'model__boosting_type': 'dart'}	0.583
NeuralNetClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__optimizer__lr': 0.1, 'model__optimizer': 'torch.optim.adam.Adam'}	0.574
SVC	{'selector': 'CustomFeatureSelector', 'scaler': 'None', 'model__kernel': 'linear', 'model___weight': 'balanced', 'model__C': 0.0001}	0.574
ElasticNet	{'selector': 'None', 'scaler': 'StandardScaler', 'model__solver': 'saga', 'model__penalty': 'elasticnet', 'model__l1_ratio': 0.95, 'model___weight': 'balanced'}	0.544
LogisticRegression	{'selector': 'CustomFeatureSelector', 'scaler': 'StandardScaler', 'model__solver': 'liblinear', 'model__penalty': 'l2', 'model___weight': 'balanced', 'model__C': 0.01}	0.544
XGBClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__scale_pos_weight': 10000, 'model__n_estimators': 50, 'model__max_depth': 1, 'model__learning_rate': 0.01, 'model__gamma': 40}	0.429
Mean	NaN	**0.578**

3.3.3 RQ1 Random Search Best Hyperparameters and Scores

Table 3.3 delineates hyperparameters for the model producing the best generalization score in addition to the F1 Score when trained using Random Search. The mean generalization score is higher ($\approx .06$) for a single trial due to reporting average results over 5 trials.

3.3.4 RQ1 Random Search Ensemble Models

Table 3.4 shows the results of grouping classifiers into ensembles using bagging and boosting algorithms. Ensemble methods perform poorly and are also less explainable than single classifiers. Therefore, I maintain that Weighted Decision Tree is still the best classifier for this use case. Table 3.4 is sorted by generalization F1 score.

Table 3.4: Average ensemble model results over 5 trials of tuning using Random Search.

	Train			Generalize		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Model						
Weighted-BaggedLogisticRegression	0.593	0.617	0.587	0.578	0.600	**0.565**
Weighted-BaggedElasticNet	0.657	0.615	0.610	0.609	0.558	0.550
Weighted-AdaBoostLogisticRegression	0.573	0.630	0.577	0.513	0.567	0.517
Weighted-AdaBoostElasticNet	0.634	0.606	0.608	0.540	0.508	0.504
Weighted-BaggedDecisionTree	0.715	0.710	0.704	0.497	0.500	0.486
Weighted-BaggedSVC	0.659	0.641	0.642	0.479	0.467	0.467
Weighted-AdaBoostSVC	0.350	0.591	0.440	0.340	0.583	0.430
Weighted-AdaBoostDecisionTree	0.913	0.895	**0.895**	0.440	0.425	0.423
Mean	**0.637**	**0.663**	**0.633**	**0.500**	**0.526**	**0.493**

3.4 RQ1 Bayesian Optimization

Results from hyperparameter tuning using Bayesian Optimization are presented in Table 3.5. Each model is class balanced using the formula given by EQ. 2.4 and Table 3.5 is sorted by generalization F1 score. Bayesian Optimization yields slightly higher accuracy on both training and generalization sets than Random Search.

Table 3.5: Average generalization results over 5 trials of tuning using Bayesian Optimization.

	Train			Generalize		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-Score**
Model						
Weighted-DecisionTreeClassifier	0.718	0.710	0.710	0.640	0.633	**0.627**
Weighted-NeuralNetClassifier	0.740	0.705	0.706	0.581	0.575	0.574
Weighted-ElasticNet	0.693	0.684	0.686	0.566	0.558	0.559
Weighted-LGBMClassifier	0.744	0.731	0.733	0.548	0.533	0.529
Weighted-RandomForestClassifier	0.763	0.753	**0.754**	0.515	0.508	0.510
Weighted-SVC	0.565	0.544	0.547	0.528	0.500	0.497
Weighted-LogisticRegression	0.642	0.682	0.653	0.483	0.525	0.495
Weighted-XGBClassifier	0.722	0.701	0.634	0.486	0.550	0.469
Mean	**0.698**	**0.689**	**0.678**	**0.544**	**0.548**	**0.532**

Additionally, Weighted Decision Tree with Bayesian hyper-parameter optimization gives higher accuracy with a weighted F1 score of 0.627. However, when comparing the generalization results of tuning methods using ANOVA, Weighted Decision Tree does not achieve significantly better accuracy ($p < .05$) when tuned with Bayesian Optimization. Therefore, one cannot trust the results of tuning using Bayesian Optimization and maintain that Weighted Decision Tree using Random Search is the "best" model. DNN and Elastic Net models using Bayesian Optimization also exhibit improvement ($\approx 5\%$) in scores despite similar parameters used in both searches. It is also of note that Linear Tree and Linear Boosting Implementations are not included in Table 3.5 due to hardware limitations discussed in Section 5.1.

3.4.1 RQ1 Bayesian Optimization ANOVA

This section presents the results of a one-way ANOVA on generalization distributions of models tuned using Bayesian Optimization. Weighted-Decision Tree as evidenced by Figure 3.3 gives the best F1 score. However, as indicated by the p-statistic of

.1389 for an ANOVA $p < .05$, it is inconclusive that Weighted Decision Tree performs better than all other models tuned using Bayesian Optimization.

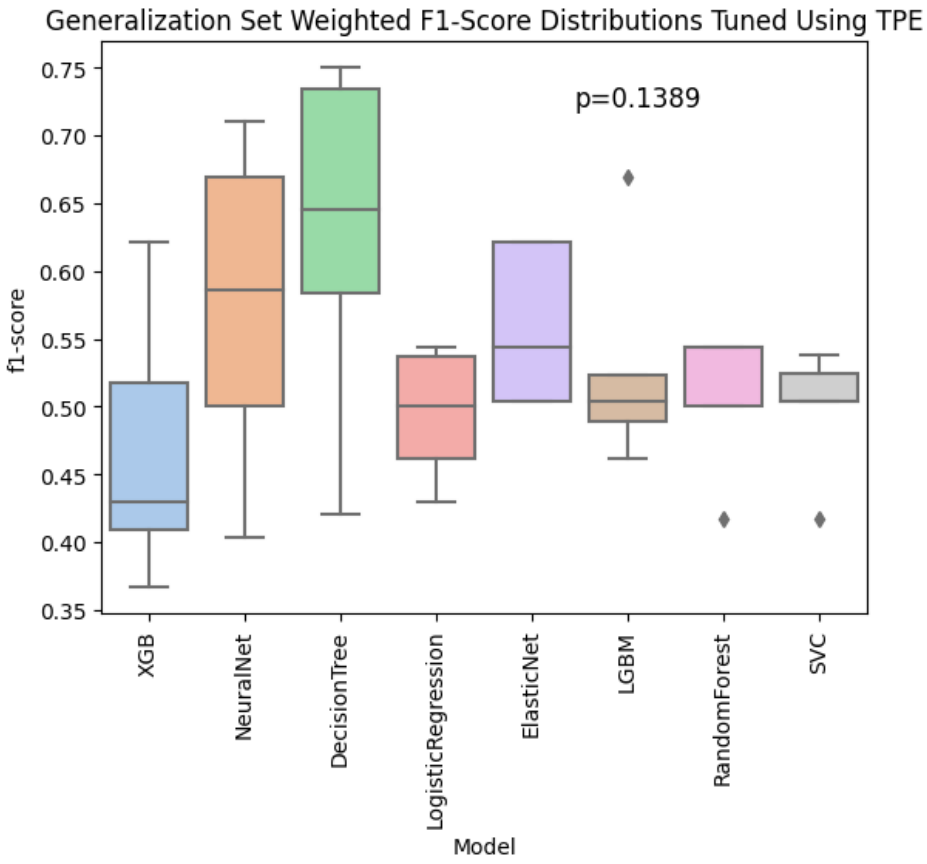

Figure 3.3: ANOVA of generalization results using TPE.

3.4.2 RQ1 Bayesian Optimization Black Box vs XAI Models

Results between "black box" and XAI models, when tuned using Bayesian Optimization using the TPE algorithm, are compared in Table 3.6. In a similar fashion to the comparison made using Random Search, Table 3.6 displays slightly higher generalization scores for XAI models trained using TPE.

Table 3.6: Average generalization F1 scores of explainable and "black box" models over 5 trials using Bayesian Optimization and the TPE algorithm.

Explainable		Black Box	
Model	**F1-Score**	**Model**	**F1-Score**
Weighted-DecisionTreeClassifier	**0.627**	Weighted-NeuralNetClassifier	**0.574**
Weighted-ElasticNet	0.559	Weighted-LGBMClassifier	0.529
Weighted-RandomForestClassifier	0.510	Weighted-XGBClassifier	0.469
Weighted-SVC	0.497	NaN	NaN
Weighted-LogisticRegression	0.495	NaN	NaN
Mean	**0.537**	**Mean**	**0.524**

3.4.3 RQ1 Bayesian Optimization Best Hyperparameters and Scores

Table 3.7 shows the best hyperparameters and generalization scores for models trained using Bayesian Optimization. Best-trial weighted F1-scores improve over Random Search best-trial F1-scores by $\approx 5\%$. Impressive tuning improvements showing the power of Bayesian Optimization are given by increases generalization accuracy in DNN (14%), ElasticNet (8%), and LGBM (9%) models. Overall, all models showed improvement except SVM and RandomForest. A concise comparison of best trials between Random Search and Bayesian Optimization can be seen in 3.8.

Table 3.7: Best parameters and score for a single trial for models trained using Bayesian Optimization.

Model	Best Params	F1-Score
DecisionTreeClassifier	{'scaler': 'StandardScaler', 'selector': RandomUnderSampler(random_state=42), 'model__criterion': 'gini', 'model__max_depth': 5, 'model__min_samples_split': 11, 'model__min_samples_leaf': 14, 'model__max_features': 'sqrt', 'model___weight': 'balanced'}	**0.750**
NeuralNetClassifier	{'scaler': 'StandardScaler', 'model__optimizer__lr': 0.51, 'model__optimizer': 'torch.optim.adam.Adam', 'model__module__hidden_size': 5, 'model__max_epochs': 200}	0.709
LGBMClassifier	{'scaler': 'StandardScaler', 'selector': 'None', 'model__boosting_type': 'dart', 'model__num_leaves': 9, 'model__learning_rate': 0.04414583361009596, 'model__max_depth': 5, 'model__min_data_in_leaf': 33, 'model__min_child_samples': 16, 'model___weight': 'balanced'}	0.668
ElasticNet	{'selector': 'None', 'scaler': 'StandardScaler', 'model__l1_ratio': 0.34415374598165616, 'model__penalty': 'elasticnet', 'model__solver': 'saga'}	0.621
XGBClassifier	{'scaler': 'StandardScaler', 'model__learning_rate': 1.6250654624469127, 'model__max_depth': 3, 'model__n_estimators': 200, 'model__gamma': 40, 'model__scale_pos_weight': 100}	0.621
RandomForestClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__n_estimators': 300, 'model__max_depth': 10, 'model__min_samples_split': 6, 'model__min_samples_leaf': 12, 'model__max_features': 'sqrt', 'model___weight': 'balanced'}	0.544
LogisticRegression	{'selector': 'None', 'scaler': 'StandardScaler', 'model__C': 0.04927338420293337, 'model__penalty': 'l2', 'model__solver': 'liblinear', 'model___weight': 'balanced'}	0.544
SVC	{'selector': 'None', 'scaler': 'StandardScaler', 'model__C': 1.051, 'model__gamma': 81, 'model__kernel': 'linear'}	0.537
Mean	NaN	**0.624**

Table 3.8: Comparison of best single trials for both Random Search and Bayesian Optimization for RQ1.

Model	Bayesian Optimization	Random Search
DecisionTreeClassifier	**0.750**	**0.696**
LinearTreeClassifier	NAN	0.626
LinearBoostClassifier	NAN	0.621
RandomForestClassifier	0.544	0.586
LGBMClassifier	0.668	0.583
NeuralNetClassifier	0.709	0.574
SVC	0.537	0.574
ElasticNet	0.621	0.544
LogisticRegression	0.544	0.544
XGBClassifier	0.621	0.429
Mean	**0.624**	**0.578**

3.5 RQ1 Explainability

As this thesis also values explainability, due to our models being used in conjunction with human decisions, I take into account the explainability of different models. Specifically, results from a Weighted Decision Tree model can be visualized and viewed in Figure 3.4 which is a subset of a larger and more complex visualization.

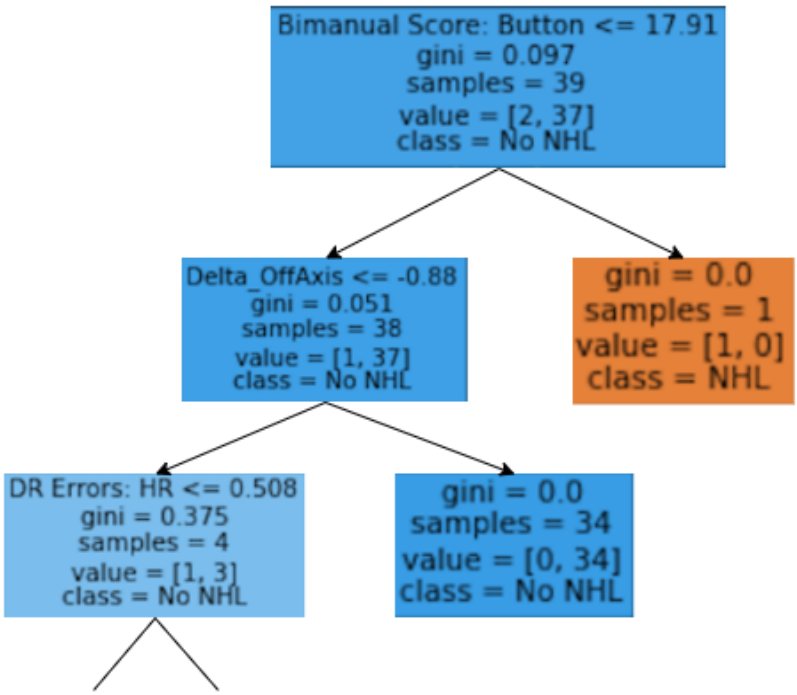

Figure 3.4: A zoomed-in section of a larger representation of a Weighted Decision Tree.

In Figure 3.4 the classification features can be ordered by importance to the classification in a top down fashion. In a situation where two models have similar accuracy, the explainable model is prioritized as it gives a way to reason about the decision. This builds trust in model results through transparency in decisions, aids in future decisions, mitigate potential biases, and aids developers and users of the model in future work. For example, in Figure 3.4 feature *Delta_OffAxis* is more important

than *DR Errors: HR* as the tree splits on the former column first. This would be difficult or impossible to identify with a "black box" model.

3.6 RQ1 Evaluation Result

Evaluation Result: I conclude the XAI model Weighted Decision Tree tuned with Random Search hyper-parameter optimization is able to classify whether or not someone played in the NHL using their concussion history and visuomotor scores with an accuracy of approximately 60%.

3.7 RQ1 Discussion

This research question (RQ1) sought to determine the effects of concussion, indicated through performance on a visuomotor test named BrDi™, on performance in the NHL. Specifically, whether visuomotor metrics are indicators of players playing in *at least* one NHL game. The BrDi™ test is a computer-based eye– hand coordination task was used to examine several kinematic variables [53]. To achieve this, I create multiple distributed hyperparameter tuning pipelines to test various machine and deep learning models including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Linear Tree, Linear Boosting Tree, Light Gradient Boost, Extreme Gradient Boost, and PyTorch Fully Connected Neural Network models in addition to multiple scaling and tuning techniques.

As our models are to be used in practice with human decisions, I emphasize explainable models (XAI) and compare performance between XAI and "black box" approaches as seen in Tables 3.2 and 3.3. While all models, except for XGB and LGBM, tuned with Random Search improve on the prior probability (49%) of a drafted player

playing in the NHL, I choose Weighted-Decision Tree as the best model (F1: .603) as it reaches significance as compared to the distributions of all other models over 5 repeated trials. Weighted-Decision Tree tuned using Bayesian Optimization achieves better performance (F1: 0.627) although is not significantly better than other models tuned using Bayesian Optimization. I also see improvement when comparing Random Search and Bayesian Optimization tuning algorithms alluding to the difficulty of tuning and training.

These models do not produce impressively high accuracies as seen in prior that predict sports performance [40, 74, 70]. There may be multiple explanations for this such as the way scores are reported, trials are repeated, tuning methods, and underlying neurological factors in the brains of elite athletes. In order to fairly report model performance, the weighted F1 score between both the positive (Played in the NHL) and negative (Did Not Play in the NHL) classes is reported. As evidenced by the tendencies of these models to predict the positive class, simply reporting the F1 score without weighting would give much higher results. Trials are repeated to improve confidence in generalization results at the additional cost of decreased accuracy. The increase in performance in models trained with Bayesian Optimization denotes both the difficulty in choosing hyperparameters and implies future work may be able to increase generalization scores through additional tuning efforts. Due to the exploratory nature and number of models chosen in this research, I balance the choice to focus on one model and explore various options.

Many of the models are an improvement to the probability that a drafted player will play in the NHL which points to the utility of using cognitive-motor integration (CMI) performance to predict future athletic success. As noted in [53], although the brains of elite athletes used in this study have ostensibly recovered from mild injury, there may still be deficits in the connections between areas required to successfully

integrate thought and action. That is to say, I suggest that the effects of concussion may be reliable indicators used in predicting athletic success. These effects may be *more* pronounced than our results imply as the effects of concussion are tempered in elite athletes. Research suggests that one of the reasons that elite athletes are able to perform at such a high level is due to protection from concussion [92, 91]. Specifically, Pellman *et al.* find that NFL athletes exhibit no neurocognitive deficits in the week following a concussion while high school athletes do exhibit persistent symptoms. Hurtubise *et al.* propose an alternative hypothesis in which a superior frontoparietal network of elite athletes allows for a greater motor control reserve thus limiting the effect of concussion. In other words, the brains of elite athletes are more efficient than novices. The superiority of the brains of elite athletes is a well studies phenomena and is named neurocognitive efficiency. I posit the neurocognitive efficiency of the elite athletes in this study as a limitation to training accurate classifiers and describe this in detail in Section 5.1.3. Interestingly, when repeated in non-elite (varsity, adolescent, youth) athletic populations [27, 14], BrDi™ test results show an increased detriment of CMI as compared to elite-level athletes. Thus, these models may be more accurate in predicting the effect of concussion on the performance of these sub-elite populations.

Due to the method in which trials are repeated, I are confident in the generalizability of our models. Thus, they are presented as useful tools for scouts, training staff, and other front-office executives interested in player performance analysis. For example, scouts may utilize our models as inputs to a broader argument of whether or not a given player will provide value to a team. Utilizing their visuomotor scores and the outputs of our model, they may be able to give reasons why or why not a player should be drafted given their chances of playing in a game. While the experiment is framed as a classification indicating 0 or 1, it is also possible to see the probability a player will play in a game. All classifications are based on a probability being above or below a threshold, and so a scout would be able to say a player has a 20% chance

of playing by using this probability. Similarly, a member of the training staff may use the outputted probability for injury prediction. For example, if a model outputs a high probability of playing in a game, a scout might decide the player is indeed cleared to play after sustaining a concussion. This is extremely important as current return-to-play standards are not sensitive enough to uncover lingering CMI deficits [53]. Research has found increased activation in the prefrontal and posterior parietal regions during a working memory task performed in retired NHL players in spite of equivalent behavioral performance in comparison to nonconcussed subjects [31]. Additionally, due to the explainable nature of our models, training staff and scouts are able to see which areas (features) athletes lack and which are important to the prediction.

In this research question, I examine the effects of concussion as predictors of athletic performance through the development of multiple machine learning pipelines. Although the models do not achieve impressive accuracy by typical machine learning standards, the results may be explained by neurological factors and the methods of this study. Nevertheless, the results are valuable to both team officials and the greater sports science research community as they may point towards markers of elite athletic performance.

Chapter 4

ARE VISUOMOTOR PERFORMANCE METRICS RELIABLE INDICATORS OF PAST CONCUSSION HISTORY?

This chapter outlines work creating classifiers to predict prior concussion history. Specifically, the following research question is established: **Are visuomotor performance metrics reliable indicators of past concussion history?** This question is formulated as a binary classification task where the negative class, labeled as 0, indicates the athlete has no prior concussion history, while the positive class, labeled as 1, indicates the athlete has sustained *at least* one prior concussion.

4.1 RQ2 Literature Review

This section outlines previous research attempts to classify prior concussion history. There exist two main approaches to developing classifiers including using features from electroencephalogram (EEG) data and visuomotor metrics similar to those used in our work.

4.1.1 Visuomotor Metrics to Classify Concussion

This section outlines prior attempts to classify concussions using visuomotor metrics similar to those used in this work.

Kelty *et al.* in [65] develop a tracing task to assess visuomotor performance and develop a Logistic Regression classifier to predict past concussion history. In total, 205 concussed participants and 209 healthy control participants completed the exper-

iment. The experimental design included a rotary tracing task in which participants were tasked with following a target as closely as possible as it moved around a circle with the goal of staying within the target region for as long as possible.

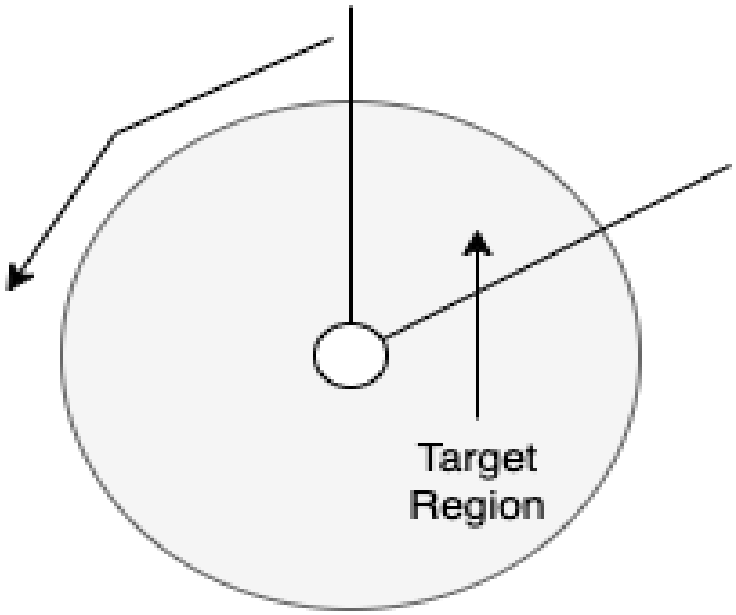

Figure 4.1: A visual depiction of the experimental design in [65].

The calculated measures of error were as follows:

- Instantaneous Error: Distance from the radius of the circle the stylus was at any given time. Error was calculated as both the mean of the error signal and variance of the error signal.
- Percentage Time Outside Target Region: The time the stylus spent outside the target region relative to the total time of the trial.
- Cumulative Micropause Duration (CMD): The sum of the time the stylus did not move. CMD is a common indicator of lingering deficits as seen in [53].
- Complexity Index: A measure of difficulty and randomness in the movement of the target region.

The study found 4 of 5 tracing measures were indicators of recovering from a concussion and that CMD has a positive relationship as an indicator of prior concussion as used with logistic mixed effects modeling [65].

The authors in [34] used a hand dynamometer device, used to measure grip strength, to create a visuomotor tracking task. 30 participants, 16 with prior concussion and 14 control were asked to squeeze the device to match a visual, variable target force for 3 minutes [34]. They parameterize response data into 3 features indicating how participants used errors in position and velocity to correct force output and the cumulative pause duration, similar to CMD. These features were used to train a Gaussian Process (GP) model. A Gaussian process model is a probabilistic machine-learning approach that makes use of infinitely many Gaussian random variables to define a distribution over functions. It gives a flexible framework for classification tasks and is adept at capturing complex patterns in the data. Overall, a model trained using all three features achieved sensitivity of 87% and a specificity of 93%.

Dalecki *et al.* [27] investigate CMI deficits in children and adolescents with concussion history. The experiment makes use of the BrDi™ Test described in section 2.1 where participants slide their finger on a dual-screen laptop to match targets. They find that participant performance on the BrDi™ Test did not match that of healthy controls until 2 years post-injury. Additionally, the features found to be significant using repeated measures mixed ANOVA are inputted to a step-wise discriminant analysis and found to classify previously concussed and healthy individuals with accuracy of 70%.

Al-Mfarej in [7] examines a bimanual motor control assessment in which participants were instructed to raise and lower both arms within a visual boundary at a specific frequency of 1 Hz. In total, 37 time-domain and 48 frequency-domain features were collected. Examples of time-domain features include peak-to-peak distance and max-

imum velocity while examples of frequency-domain features include spectral entropy and power spectrum. A variety of machine learning models were trained including Logistic Regression with either L1 or L2 penalties, SVM with linear kernel, Naïve Bayes, Random Forest, Adaboost and XGboost classifiers. The highest model accuracies were achieved by Logistic Regression (86%) and Adaboost (88%) classifiers.

In [119], Wilkerson *et al.* develop a Logistic Regression model able to classify prior concussion history in elite Olympic-level athletes with 84% accuracy (measured using area under the curve (AUC)). Multiple tests were performed including visual-motor reaction time and whole-body reactive agility assessments. The visual-motor reaction time test was performed by asking participants to press randomly illumined buttons on a grid. The whole-body agility tests involved athletes responding to visual stimuli presented on a virtual reality display by executing horizontal or diagonal movements. Despite being limited by a small sample size (35) and lack of female athletes, the work does indicate deficits in whole body reactive moment as a signal of prior concussion.

To summarize, there exist many prior works investigating the relationship between visuomotor tests as predictors of concussion [119, 7, 34, 27, 65]. The studies show variable levels of classification accuracy, but generally give evidence of visuomotor deficits after concussion. There are also many different methods of testing cognitive-motor integration and visuomotor skill such as using fine motor tasks such as tracing [27, 65], more granular assessment such as grip strength and bimanual motor control [34, 7], and whole body assessment [119].

4.1.2 Electroencephalography for Classifying Concussions

This section outlines prior work in which electroencephalogram (EEG) data has been used to classify concussions. EEG is a common method of assessing cerebral function-

ing due to its excellent temporal sensitivity as a direct recording of electrical brain activity [13]. EEG recordings are often measure multiple brain wave frequencies such as alpha, beta, and theta. For example alpha waves, measured between 8 and 12 Hz, occur doing periods of relaxation and low concentration. In contrast, beta waves, measured between 12 and 30 Hz, are common in conscious waking states and are signals corresponding to alertness and attentiveness. Measuring multiple brain wave frequencies is referred to as multichannel EEG.

The authors in [19] develop a classifier for identifying residual functional deficits after concussion using SVM and a multichannel EEG dataset. 61 total subjects were measured using 19 electrodes and 2 reference electrodes and were asymptomatic as per a variety of post SRC tests. Subjects were measured in 3 conditions: seated with eyes closed, standing on a firm surface with eyes closed, and on a foam surface with eyes closed. After feature selection, 10 features were inputted to SVM with a linear kernel to produce a classification accuracy of 77%. However, it is worth noting the authors report much higher sensitivity (96%) than selectivity (69%) indicating the potential for models to bias the positive class (overfit).

Research done in [55], explores multimodal classification of prior concussion history using neurocognitive, EEG, and clinical symptom data (headache, dizziness). Neurocognitive data was collected using subsets of the ANAM suite [121], reaction time tests for motor output speed, Code Subsitution Learning to test associative learning and memory, and visuomotor reaction time and decision making tests [55]. EEG data was collecting in an eyes closed resting condition using 7 channels. The three feature groups were used to train Linear Discriminant Functions using a Genetic Algorithm and leave-one-out cross validation. Results, measured using sensitivity, specificity, and AUC are 73%, 85%, and 85% respectively.

Jacquin *et al.* [56], in a study that succeeds the previous work done in [55] implement concussion classification in a similar fashion using EEG, neurocognitive performance scores, and standard concussion assessments. The experiment combines the results from a reaction time test, 7 channel EEG, and clinical symptom assessments. The authors trained Genetic Algorithms and Logistic Regression using L1 (Lasso) penalty to obtain generalization results of sensitivity (85%), specificity (76%), and AUC (0.89). While the study did enhance explainability through post-hoc feature analysis using model weights, it did not prioritize explainability such as I do in this research.

Poltavski *et al.* in [95] explore the efficacy of using the ratio of two types of brain waves, theta and gamma, as predictors of concussion in youth athletes. The study was conducted with 81 13-18 year old athletes of which 18 had sustained prior concussion. 9 channel EEG recordings were collected during testing using the Nike SPARQ Sensory Training Station used to test static visual acuity, contrast sensitivity, depth perception, near-far quickness, dynamic visual acuity, perception span, eye-hand coordination, "Go/No-Go", and hand reaction time. A mixed model ANOVA was used to analyze power spectral density ratios between theta and gamma waves and found a significant effect of history of concussion. Thus, showing the ratio of EEG signals as a predictor of concussion. While the study did not train machine learning classifiers, it still stands to corroborate prior works [55, 56, 19] to show the effectiveness in using EEG for concussion prediction. The study also uses principle component analysis (PCA) to show an underlying dimension that effects deficits in multiple visuomotor metrics collected in the study.

While EEG data is not used in our research, it is a common method of visuomotor assessment and concussion prediction and has been shown to produce high classification accuracies. Various machine learning models are used to create classifiers with EEG data such as Logistic Regression, SVM, and Genetic Algorithms. It is also of

note that many of these studies report sensitivity and specificity instead of precision and recall. Sensitivity and specificity are commonly used in medical applications and are subtly different than precision and recall. Sensitivity and specificity place more emphasis on if the *sample* was positive or negative whereas precision and recall emphasize whether the *classifier predicts* the positive or negative class. For example, high specificity means that if the sample is negative, there is high probability the model with classify it as negative. In contrast, high precision means that if the model predicted a sample as positive, there is a high chance it is actually positive. Thus, analyzing classification accuracy can be approached from the true value of the sample (sensitivity/specificity) vs how it can be approached by what the model predicts (precision/recall).

4.2 RQ2 Results

This section presents the results for the second research question: **Are visuomotor scores and player performance indicators of past concussion?** Multiple machine learning models are surveyed including 6 explainable models (Decision Tree, Random Forest, Linear Tree, Logistic Regression, Elastic Net, Support Vector Machine (SVC)), and 4 black box models (Extreme Gradient Boost (XGB), Light Gradient Boost (LGBM), Linear Boosting Tree, Deep Learning Neural Network (DNN)). Additionally, bagging and Adaboost algorithms of various weak classifiers are implemented in an attempt to increase accuracy scores.

In a similar fashion to Figure 3.1, I present Figure 4.2. As I am now interested in classifying prior concussion history, the input feature "previous_concussions" is replaced with "NHL" indicating if a player did or did not play in the NHL.

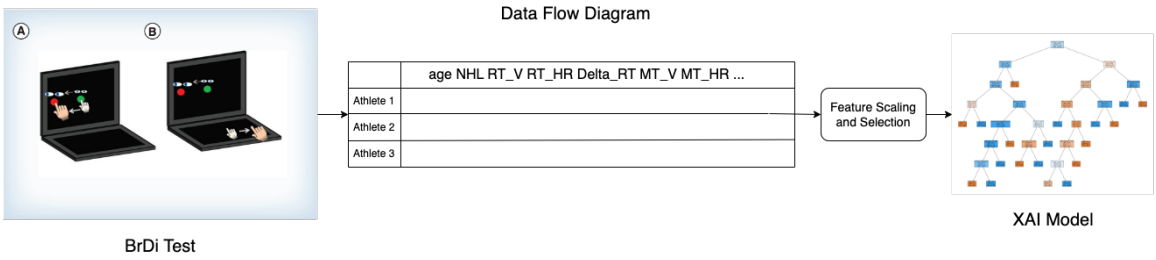

Figure 4.2: Flow of data from BrDi™ test, to tabular DataFrame, to feature scaling and selection, and an ML model.

4.2.1 RQ2 Random Seach

This section presents results for models tuned using Random Search. Results for single classifiers can be seen in Table 4.1.

Table 4.1: Average generalization results over 5 trials of tuning using Random Search.

	Train			Generalize		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Model						
Weighted-ElasticNet	0.621	0.622	0.619	0.514	0.517	**0.512**
Weighted-RandomForestClassifier	0.801	0.800	0.800	0.510	0.508	0.505
Weighted-LGBMClassifier	0.821	0.819	**0.819**	0.493	0.492	0.486
Weighted-LogisticRegression	0.598	0.596	0.596	0.490	0.483	0.480
Weighted-DecisionTreeClassifier	0.704	0.695	0.692	0.454	0.450	0.444
Weighted-LinearBoostClassifier	0.631	0.630	0.630	0.389	0.400	0.391
Weighted-LinearTreeClassifier	0.719	0.718	0.718	0.387	0.400	0.389
Weighted-NeuralNetClassifier	0.619	0.544	0.485	0.367	0.442	0.350
Weighted-XGBClassifier	0.661	0.535	0.425	0.208	0.450	0.284
Mean	**0.686**	**0.662**	**0.643**	**0.424**	**0.460**	**0.427**

The table is sorted by generalization score and each model is class balanced using EQ 2.4. I discern that Elastic Net produces the best generalization accuracy of .512. Using a similar approach to work done in 3.2, model results are reported as an average

over 5 trials to give confidence in model performance while balancing data used for training and generalization.

4.2.2 RQ2 Random Search ANOVA

To identify models that outperformed counterparts with significance, a one-way ANOVA ($p < .05$) is performed to compare generalization scores. ANOVA is a statistical test to compare variances of arithmetic means. In this study, the results of an ANOVA test with p $< .05$ dictate that one of the means of the generalization set results is different with statistical significance. ANOVA is performed using the mean sum of squares and the mean sum of squared errors between generalization results. The results of ANOVA performed with all models is seen in Figure 4.3 and gives $p = .001$ indicating significance.

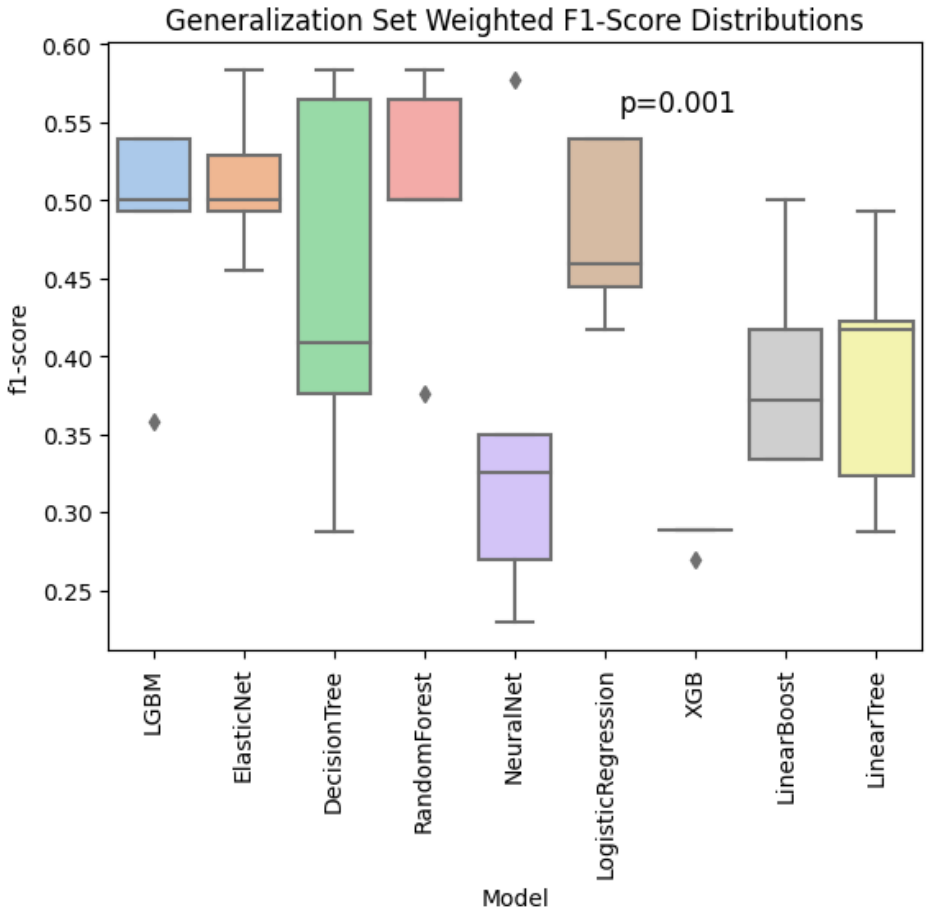

Figure 4.3: One-way ANOVA of all models trained using Random Search.

However, due to two models (Neural Net and XGB) performing much worse than all others, the ANOVA test is re-run excluding these two.

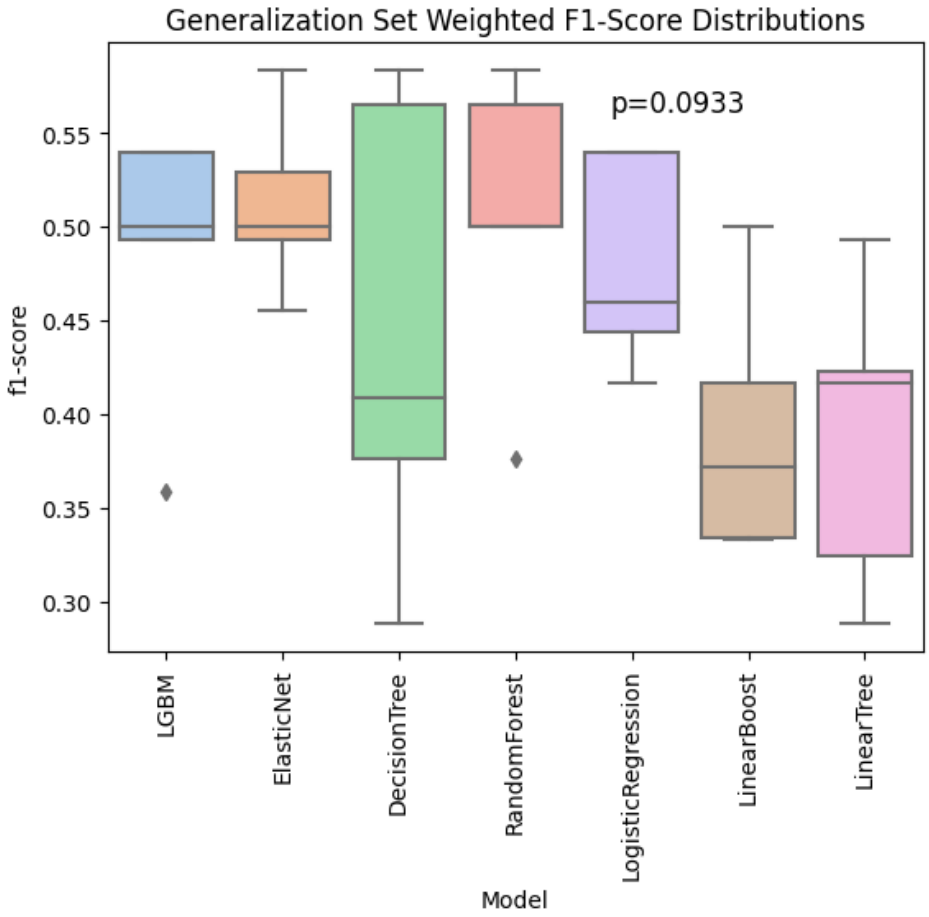

Figure 4.4: One-way ANOVA excluding Neural Net and XGB.

Results are seen in 4.4. In this instance, $p = 0.0933$ indicating no models outperformed others with significance. Thus, it is inconclusive as to whether a single model outperforms all others with significance.

4.2.3 RQ2 Random Search Black Box vs XAI Models

To further the understanding of trained models, I compare "black box" and XAI models. A comparison of the two types can be seen in Table 4.2

Table 4.2: Average generalization F1 scores of explainable and "black box" models over 5 trials using Random Search.

Explainable		Black Box	
Model	**F1-Score**	**Model**	**F1-Score**
Weighted-ElasticNet	**0.512**	Weighted-LGBMClassifier	**0.486**
Weighted-RandomForestClassifier	0.505	Weighted-LinearBoostClassifier	0.391
Weighted-LogisticRegression	0.480	Weighted-NeuralNetClassifier	0.350
Weighted-DecisionTreeClassifier	0.444	Weighted-XGBClassifier	0.284
Weighted-LinearTreeClassifier	0.389	NaN	NaN
Mean	**0.466**	**Mean**	**0.378**

Notice suboptimal performance in NeuralNet, XGB, Linear Tree, and LinearBoost classifiers when trained with Random Search although this is greatly improved upon using Bayesian Optimization.

4.2.4 RQ2 Random Seach Best Hyperparameters and Scores

Table 4.3 illustrates how Random Forest, Elastic Net, and Decision Tree all perform the best in a single trial with a weighted F1 of 0.58.

Table 4.3: Best parameters and score for a single trial for models trained using Random Search.

Model	Best Params	F1-score
RandomForestClassifier	{'selector': 'None', 'scaler': 'None', 'model__n_estimators': 50, 'model__min_samples_split': 10, 'model__min_samples_leaf': 4, 'model__max_features': 'sqrt', 'model__max_depth': 2, 'model___weight': 'balanced'}	**0.583**
DecisionTreeClassifier	{'selector': 'CustomFeatureSelector', 'scaler': 'StandardScaler', 'model__min_samples_split': 13, 'model__min_samples_leaf': 11, 'model__max_features': 'sqrt', 'model__max_depth': 4, 'model__criterion': 'gini', 'model___weight': 'balanced'}	**0.583**
ElasticNet	{'selector': 'CustomFeatureSelector', 'scaler': 'MinMaxScaler', 'model__solver': 'saga', 'model__penalty': 'elasticnet', 'model__l1_ratio': 0.99, 'model___weight': 'balanced'}	**0.583**
NeuralNetClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__optimizer__lr': 0.1, 'model__optimizer': 'torch.optim.adam.Adam'}	0.578
LogisticRegression	{'selector': 'CustomFeatureSelector', 'scaler': 'StandardScaler', 'model__solver': 'liblinear', 'model__penalty': 'l2', 'model___weight': 'balanced', 'model__C': 0.3}	0.539
LGBMClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__subsample': 0.5, 'model__num_leaves': 2, 'model__num_iterations': 400, 'model__min_data_in_leaf': 25, 'model__min_child_samples': 5, 'model__max_depth': 7, 'model__max_bin': 25, 'model__learning_rate': 0.1, 'model__early_stopping_rounds': 7, 'model__colsample_bytree': 0.5, 'model___weight': 'balanced', 'model__boosting_type': 'dart'}	0.539
LinearBoostClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__n_estimators': 20, 'model__max_depth': 1, 'model__ccp_alpha': 0.1}	0.500
LinearTreeClassifier	{'selector': 'None', 'scaler': 'MinMaxScaler', 'model__n_jobs': -1, 'model__min_samples_split': 20, 'model__min_samples_leaf': 20, 'model__max_depth': 3, 'model__max_bins': 10}	0.493
XGBClassifier	{'selector': 'None', 'scaler': 'None', 'model__scale_pos_weight': 10000, 'model__n_estimators': 50, 'model__max_depth': 5, 'model__learning_rate': 0.5, 'model__gamma': 40}	0.288
Mean	NaN	**0.521**

4.2.5 RQ2 Random Search Ensemble Models

Table 4.4 depicts model results for ensemble methods trained using Random Search. The table is sorted by generalization accuracy and all models are class balanced using EQ 2.4. Unlike results seen for RQ 1, various ensemble algorithms outperform single classifiers. Specifically, BaggingClassifierRandomForestClassifier produces the highest generalization accuracy of models trained using Random Search.

Table 4.4: Average ensemble model results over 5 trials of tuning using Random Search.

	Train			Generalize		
Model	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
BaggingClassifierRandomForestClassifier	0.729	0.723	0.722	0.567	0.567	**0.564**
BaggingClassifierDecisionTreeClassifier	0.818	0.815	0.814	0.512	0.508	0.504
AdaBoostClassifierDecisionTreeClassifier	0.980	0.978	0.979	0.495	0.483	0.478
AdaBoostClassifierRandomForestClassifier	0.998	0.998	**0.998**	0.476	0.492	0.474
BaggingClassifierLogisticRegression	0.582	0.555	0.515	0.570	0.517	0.473
AdaBoostClassifierLogisticRegression	0.446	0.533	0.454	0.416	0.500	0.425
Mean	**0.759**	**0.767**	**0.747**	**0.506**	**0.511**	**0.486**

4.3 RQ2 Bayesian Optimization

Model training is repeated using Bayesian Optimization and results are seen in Table 4.5. The table is sorted by generalization score and all models are balanced using EQ 2.4. Note that SVM produces the best classification accuracy of 0.612. Similar to results seen in RQ1, when comparing Table 4.5 and Table 4.1 there exists drastic improvement in average generalization accuracy. This is attributed to the power of Bayesian Optimization for hyperparameter tuning.

Table 4.5: Average generalization results over 5 trials of tuning using Bayesian Optimization.

	Train			Generalize		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Model						
Weighted-SVC	0.593	0.589	0.586	0.616	0.617	**0.612**
Weighted-DecisionTreeClassifier	0.703	0.690	0.682	0.537	0.533	0.520
Weighted-LinearBoostClassifier	0.666	0.665	0.665	0.512	0.517	0.512
Weighted-ElasticNet	0.639	0.639	0.637	0.514	0.517	0.511
Weighted-LGBMClassifier	0.858	0.856	**0.856**	0.515	0.508	0.508
Weighted-NeuralNetClassifier	0.739	0.714	0.701	0.496	0.517	0.483
Weighted-RandomForestClassifier	0.853	0.852	0.851	0.478	0.475	0.473
Weighted-XGBClassifier	0.713	0.690	0.644	0.530	0.517	0.454
Weighted-LogisticRegression	0.522	0.572	0.537	0.442	0.483	0.447
Mean	**0.698**	**0.696**	**0.684**	**0.516**	**0.520**	**0.502**

4.3.1 RQ2 Bayesian Optimization ANOVA

Results of a one-way ANOVA ($p < .05$) are presented for generalization distributions of models tuned using Bayesian Optimization. Figure 4.5 illustrates results of the ANOVA test ($. < .05$) are inconclusive as $p = .2966$. However, it is of note the box plot shows a wide distribution of scores which validates our use of multiple trials. For example, the single best trial for SVM trained using Bayesian Optimization is 0.79 despite the average over 5 trials of 0.62.

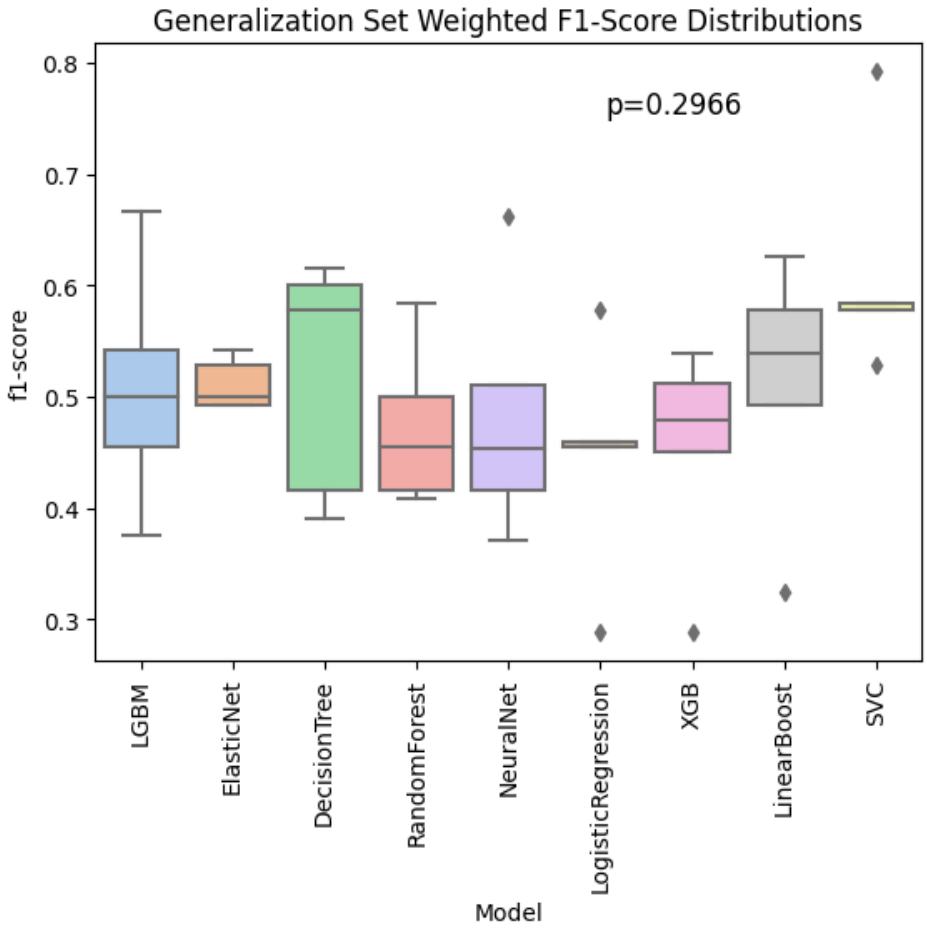

Figure 4.5: One-way ANOVA of all models trained using Bayesian Optimization

4.3.2 RQ2 Bayesian Optimization Black Box vs XAI Models

Table 4.6 compares the results of "black box" vs XAI models when trained using Bayesian Optimization. Explainable models perform better on average, although are largely influenced by SVC.

Table 4.6: Avgerage generalization F1 scores of explainable and "black box" models over 5 trials using Bayesian Optimization.

Explainable		Black Box	
Model	**F1-Score**	**Model**	**F1-Score**
Weighted-SVC	**0.612**	Weighted-LinearBoostClassifier	**0.512**
Weighted-DecisionTreeClassifier	0.520	Weighted-LGBMClassifier	0.508
Weighted-ElasticNet	0.511	Weighted-NeuralNetClassifier	0.483
Weighted-RandomForestClassifier	0.473	Weighted-XGBClassifier	0.454
Weighted-LogisticRegression	0.447	NaN	NaN
Mean	**0.513**	**Mean**	**0.489**

4.3.3 RQ2 Bayesian Optimization Best Hyperparameters and Scores

Much better results for single trials as compared to models trained with Random Search are viewed in Tables 4.7 and 4.8. Although this thesis could have reported the best accuracy for this research question as 0.792, averages are reported to improve confidence in model scores.

Table 4.7: Best parameters and score for a single trial for models trained using Bayesian Optimization.

Model	Best Params	F1-Score
SVC	{'selector': 'CustomFeatureSelector', 'scaler': 'StandardScaler', 'model__C': 1.551, 'model__gamma': 51, 'model__kernel': 'sigmoid'}	**0.792**
LGBMClassifier	{'scaler': 'StandardScaler', 'model__boosting_type': 'dart', 'model__max_bin': 22, 'model__num_leaves': 10, 'model__learning_rate': 0.08362420395484067, 'model__max_depth': 3, 'model__min_data_in_leaf': 22, 'model__min_child_samples': 16, 'model___weight': 'balanced'}	0.667
NeuralNetClassifier	{'scaler': 'StandardScaler', 'model__optimizer__lr': 0.31000000000000005, 'model__optimizer': 'torch.optim.adam.Adam', 'model__module__hidden_size': 9, 'model__max_epochs': 200}	0.662
LinearBoostClassifier	{'selector': 'CustomFeatureSelector', 'scaler': 'MinMaxScaler', 'model__n_estimators': 100, 'model__max_depth': 3, 'model__ccp_alpha': 0.013902048606045738}	0.626
DecisionTreeClassifier	{'scaler': 'StandardScaler', 'selector': RandomUnderSampler(random_state=42), 'model__criterion': 'gini', 'model__max_depth': 7, 'model__min_samples_split': 14, 'model__min_samples_leaf': 14, 'model__max_features': 'sqrt', 'model___weight': 'balanced'}	0.615
RandomForestClassifier	{'selector': 'None', 'scaler': 'StandardScaler', 'model__n_estimators': 450, 'model__max_depth': 8, 'model__min_samples_split': 5, 'model__min_samples_leaf': 10, 'model__max_features': 'log2', 'model___weight': 'balanced'}	0.583
LogisticRegression	{'selector': 'CustomFeatureSelector', 'scaler': 'StandardScaler', 'model__C': 0.03633093196904575, 'model__penalty': 'l2', 'model__solver': 'saga', 'model___weight': 'balanced'}	0.578
ElasticNet	{'selector': 'CustomFeatureSelector', 'scaler': 'StandardScaler', 'model__l1_ratio': 0.04912925395458216, 'model__penalty': 'elasticnet', 'model__solver': 'saga'}	0.542
XGBClassifier	{'scaler': 'StandardScaler', 'model__learning_rate': 1.160267468320288, 'model__max_depth': 3, 'model__n_estimators': 150, 'model__gamma': 35, 'model__scale_pos_weight': 50}	0.539
Mean	NaN	**0.623**

Similar to RQ1, Bayesian Optimization provides an impressive improvement in single best trials in relation to Random Search. A comparison of single best trials is seen in Table 4.8.

Table 4.8: Comparison of best single trials for both Random Search and Bayesian Optimization for RQ2.

Model	Bayesian Optimization	Random Search
Weighted-SVC	**0.792**	NAN
Weighted-LGBMClassifier	0.667	0.539
Weighted-NeuralNetClassifier	0.662	0.578
Weighted-LinearBoostClassifier	0.626	0.500
Weighted-DecisionTreeClassifier	0.615	**0.583**
Weighted-RandomForestClassifier	0.583	**0.583**
Weighted-LogisticRegression	0.578	0.539
Weighted-ElasticNet	0.542	**0.583**
Weighted-XGBClassifier	0.539	0.288
Mean	**0.623**	**0.521**

4.3.4 RQ2 Explainability

This section gives examples of methods to understand the best XAI model: SVM. As seen in Table 4.7, SVM with sigmoid kernel, Standard Scaler, and custom feature selector (5 features) produces the best results. This pipeline is reproduced in order to calculate feature importance. Figure 4.6 charts the permutation feature importance in a bar plot. Permutation importance is a technique for model inspection useful for non-linear estimators. Essentially, it is the decrease in model scores when a single feature value is randomly shuffled. Thus, the relationship between feature and target (Concussion History) is removed, and model performance due to a single feature is estimated [90].

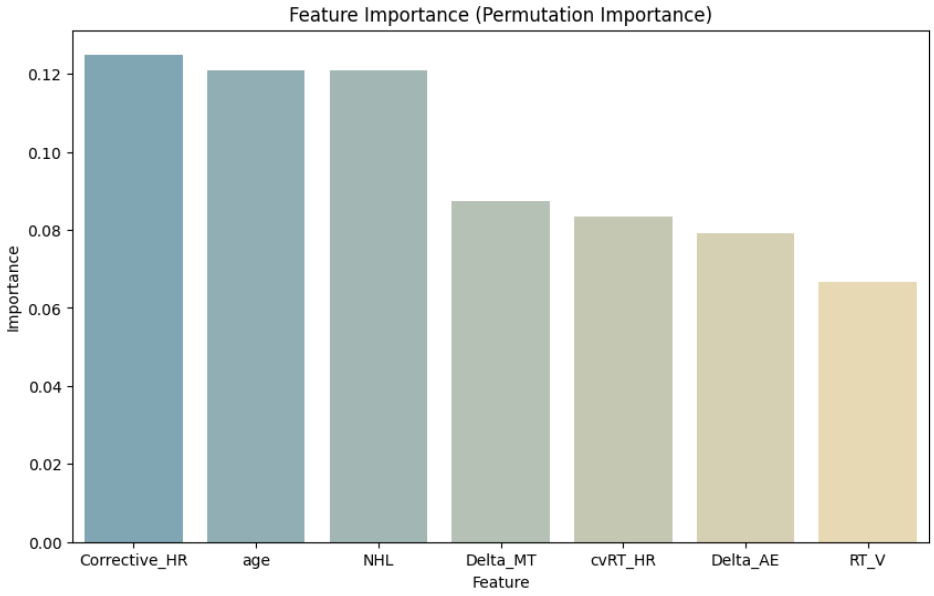

Figure 4.6: Bar plot of SVM feature permutation importance

However, this technique is not particular to SVM and can be used with any classifier "black box" or explainable. To capture the true explainable nature of SVM, I graph the decision boundary, described in Section 2.3.4, over top of the generalization set. It is important to note this is not a perfect representation of the original SVM. In order to graph the decision boundary in two dimensions, the same best-scaling and feature selection algorithms found using Bayesian Optimization are applied, but the input feature dimensionality is reduced to 2 dimensions. This is done using Principle Component Analysis (PCA). Once the features in two dimensions are obtained, the target is predicted using the best hyperparameters displayed in Table 4.7 for SVM and results are graphed. Results are seen in Figure 4.7. It is of note that the model does not perform as well as seen in Table 4.5 due to dimensionality reduction using PCA.

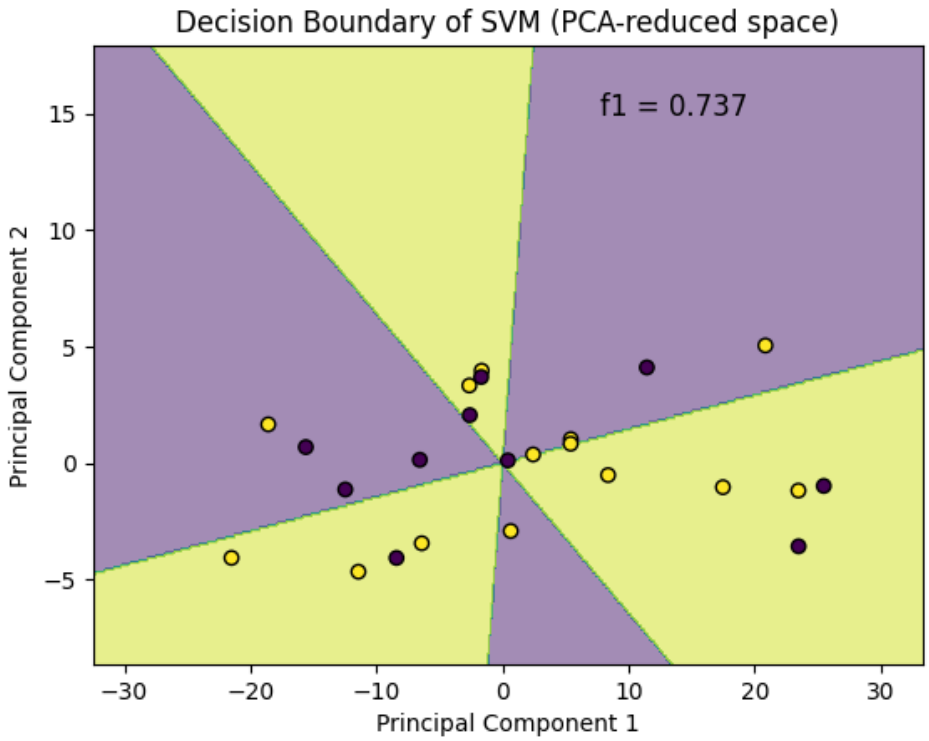

Figure 4.7: Scatter plot of Sigmoid SVM and predicted datapoints.

4.4 RQ2 Evaluation Result

Due to all one-way ANOVA tests failing to reach significance, it is not possible to conclude there is a single best model for **RQ2 - Are visuomotor metrics indicators of past concussion history?**

4.5 RQ2 Discussion

The second research question sought to explore the efficacy of using visuomotor metrics to classify past concussion history. This was accomplished using the machine and deep learning pipelines described in 2.5.6. Due to inconclusive results when comparing model distributions using one-way ANOVA tests, a claim cannot be made about the

best model suited for this task. However, there are still many important takeaways from this work.

The method by which scores are reported is paramount. An SVM classifier is trained to $\approx 80\%$ accuracy, yet is inconclusive due to repeating trials. Yes, a single trial may have been reported, **however, this would be a misleading result.** Additionally, notice two additional ways to visualize model results. Explainability results seen in RQ1 are furthered by charting SVM feature weights and decision boundaries. While direct insight into features is not clear in Figure 4.7 due to PCA, the figure still serves to **build trust in our model.** Should future work reach conclusive results, this work has established a reliable framework for assessing and utilizing the explainability of various models.

There may be mechanisms to understand why the performance of the models is lackluster as compared to prior works and our first research question. As described in the original paper, concussions were self-reported and all athletes were cleared to play by physicians and current return-to-play concussion standards [53]. In contrast in RQ1, which athletes played in the NHL is known for certain, and thus the relationship may be harder to discern for this question. The paper also states that significant detriment in visuomotor performance was only exhibited in the nonstandard condition which involved queue reversal and plane change as outlined in 2.1. However, data is not filtered by condition and all BrDi™ test metrics of both the standard and nonstandard conditions are utilized.

The models do not achieve the accuracy seen in prior works that use visuomotor metrics to predict concussion. There may be logical explanations for this. I characterize in 3.7 how the effects of concussion are milder in elite athletes as compared to non-elite athletic populations. In literature, this phenomenon is named neurocognitive efficiency. I hypothesize difficulties in training models for this research question may

be due to the neuroprotective effects of the brains of elite athletes. Dalecki *et al.* in [27], perform the BrDi™ test on adolescents and use this data to train a classify to predict prior concussion history, yet only achieve accuracy of 70%. This is an interesting result when compared to mine as I postulate the detriments of concussion are *greater* in adolescents as compared to elite athletes. Thus, there is a basis to rationalize why the models in this research would not perform similarly.

Hammeke *et al.* in [48] perform a study of NCAA (college) football players after concussion diagnosis. They find decreased cortical activation in brain regions associated with attention during the 'acute' (up to 4 days after injury) phase, but increased activation after the resolution of visible symptoms ('subacute' phase). They suppose that satisfactory performance on standardized clinical tests during the subacute phase was the result of improved neurocognitive performance. However, they reasoned the performance improvement was due to compensatory cognitive mechanisms as evidenced by the increase in cortical activation. These findings suggest that performance deficits may appear with an increase in task difficulty that exceeds the capability of the compensatory activity [53]. Put another way, there is a positive relationship between task difficulty and evidence of concussion in visuomotor metrics. This may explain other works that exceed model accuracies found in this work as they perform visuomotor assessments that may be considered more difficult [34, 7, 119]. For example, Wilkerson *et al.* produce a classifier able to discern concussions in Olympic-level athletes with 84% accuracy, yet use reaction time and whole body agility assessments. This task may be more difficult than the task outlined in the BrDi™ Test and thus explain superior results.

Should our models have achieved more accurate results, they would serve as practical tools for injury diagnosis and prevention. As previously stated, current return-to-play standards are often not sensitive enough to capture underlying visuomotor deficits.

Our models, paired with the BrDi™ assessment, may serve as a more sensitive indicator of injury in the future.

While our models in their current state are not readily available tools for injury diagnosis, our work still gives valuable conclusions. Not only do I provide a ready-to-use framework for future training efforts, but also give examples of how the method by which accuracies are reported can affect the perception of results. Our work also corroborates the neuroprotective qualities of elite athletes and the need for challenging visuomotor assessment while ultimately advancing the fields of neuroscience and medical assessment.

Chapter 5

THREATS TO VALIDITY

5.1 Threats to Validity

This section describes threats to the validity of this study including dataset size, training hardware configurations, and noise. These are explained in detail as follows:

5.1.1 Dataset Size and Overfitting

Dataset size was a significant limitation of this study as many ML and DL techniques require large amounts of data to capture complex relationships between features (visuomotor scores) and targets (Played in the NHL). Some ML models, such as Weighted Logistic Regression (LR) and Weighted SVM are known to learn feature weights with small amounts of data. I trained LR models with multiple penalties including L1 (Lasso Regression), L2 (Ridge Regression), and Elastic Net penalties to create accurate classifiers that avoid the overfitting (overtraining) problem.

L1 Regularization adds a penalty term to the LR loss function that encourages the model to have many coefficients that are exactly zero. This helps to reduce the complexity of the model and can prevent overfitting. L2 regularization adds a penalty term to the loss function that encourages coefficients to remain small instead of exactly zero as in L1 regularization. L2 regularization also helps reduce overfitting and improves generalization performance.

Elastic Net is a combination of L1 and L2 regularization techniques that balance the advantages of L1 and L2 regularization and can overcome their limitations. It can

select relevant features like L1 regularization and also handle correlated features like L2 regularization.

SVM is also a powerful classifier when used with small datasets. I train SVM models with 3 kernels, that affect the shape of the decision boundary, including linear, radial basis function, and sigmoid kernels. However, I see in Table 3.3 and Table 3.5 that Weighted Decision Tree still performed best on the generalization set compared to all the other models.

As seen in Table 3.3 and Table 3.5, many models achieve higher training scores than generalization (independent hold-out set) scores. From an ML perspective, our results are inconclusive due to overtraining. Yet predicting which players selected in the amateur draft will play in the NHL is a difficult task. Only 49% of players will go on to play in an NHL game despite scouts taking a much larger number of factors than used in this research into account. I exhibit success in this regard as almost all models achieve accuracy scores $> 49\%$. In the ideal scenario, I would observe identical or very similar +- $\approx 5\%$ scores from which I could conclude training results are replicable with higher confidence on unseen datasets.

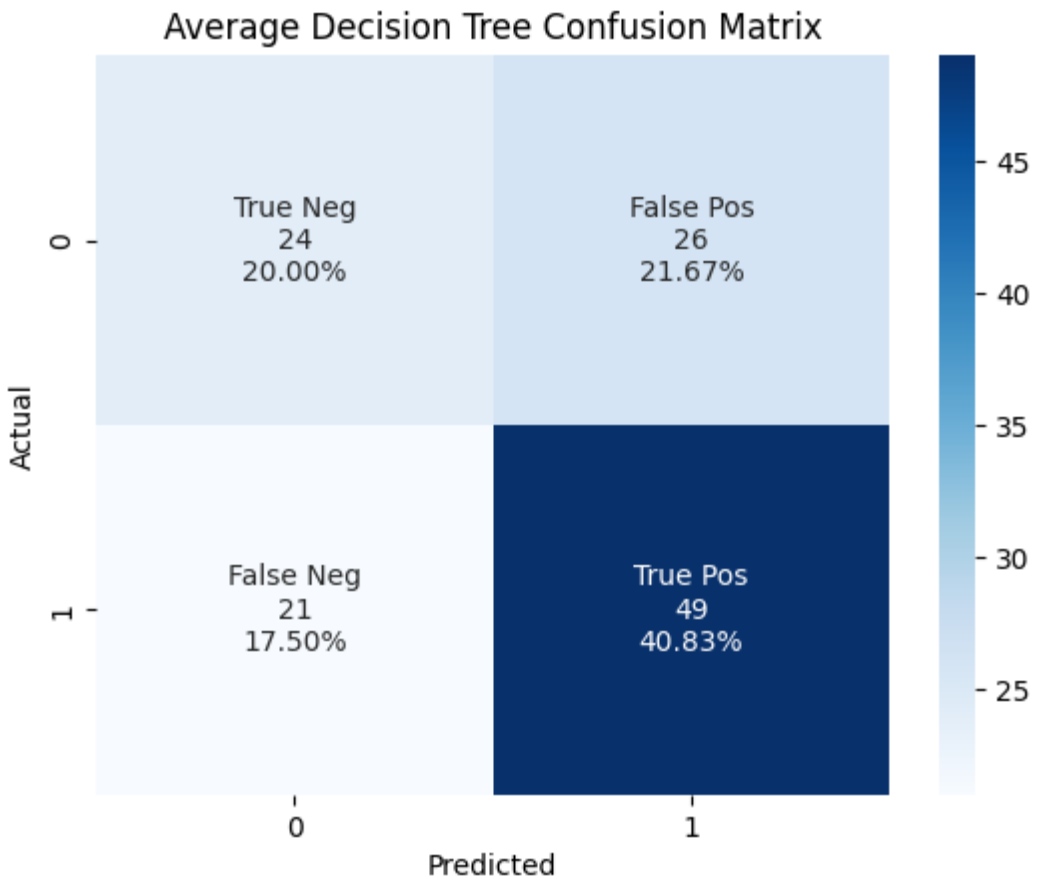

Figure 5.1: Confusion matrix for the best XAI model over 5 trials - Weighted Decision Tree.

Imperfect accuracy scores are reflected in confusion matrices of generalization results such as Figure 5.1 where I see predictions on the bottom left to top right axis indicating a misclassification. In Figure 5.1 the true classes are on the y-axis, and predicted classes are on the x-axis. Label 1 represents Played in the NHL and label 0 represents Did Not Play in the NHL. Predictions only on the top left to bottom right diagonal would demonstrate 100% generalization accuracy. Although some single splits achieve very high generalization accuracy, such as those seen in Figure 5.2, I report generalization accuracy as an average among 5 splits to give greater confidence in results despite lower scores. This difference in scores can be viewed by comparing Figure 5.1 to Figure 5.2. I see that the confusion matrix for the single best trial of

the Weighted Decision Tree looks much better than the average generalization results of Weighted Decision Tree even though they are the same classifier.

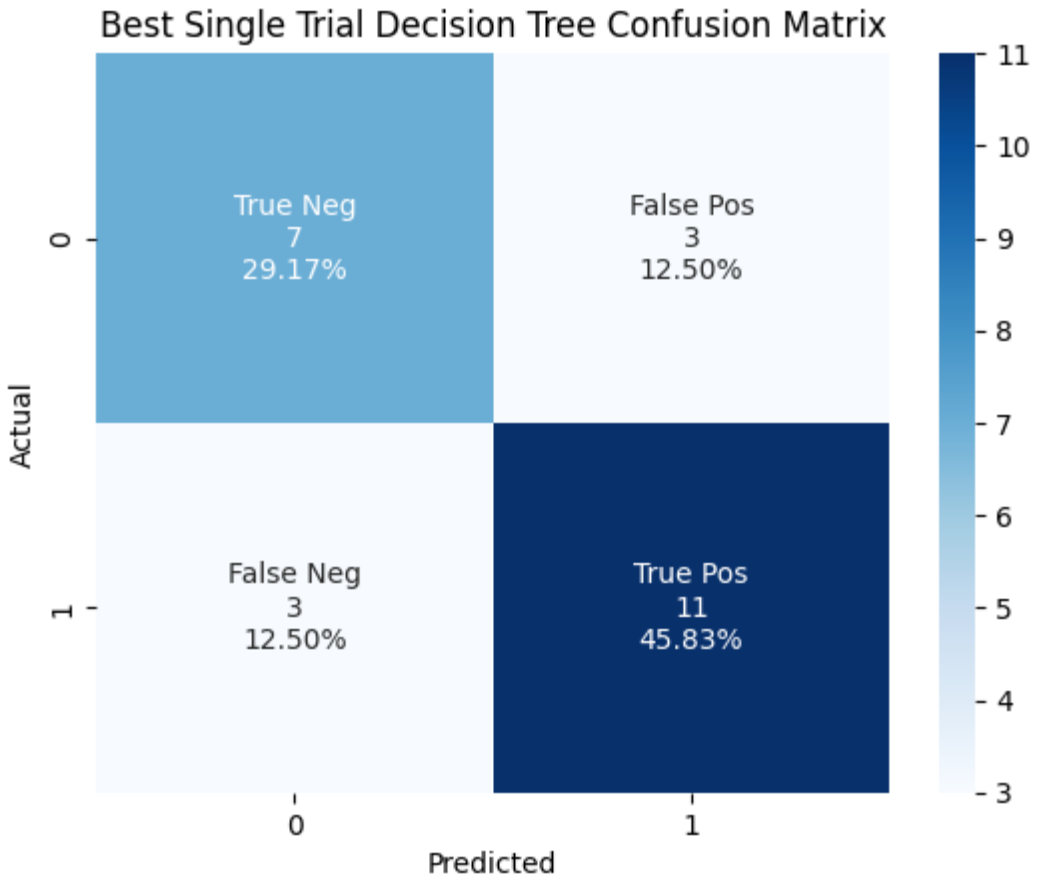

Figure 5.2: Confusion matrix for Weighted Decision Tree which produced the best generalization score for a *single trial*.

5.1.2 Dataset Noise

Noise in the dataset may be a significant factor in regard to training difficulty due to the immense number of reasons a player may not play in the NHL. For example, a player may get hurt due to a non-concussive injury, may not develop skills as projected during the draft, simply suffer from the highly skilled and competitive nature of the NHL, or even leave over contract disputes.

For example, Jesse Niinimäki was drafted 15th overall (1st Round) by the Edmonton Oilers. Scouts felt he could be an undervalued talent moving to the NHL from the Finnish men's hockey league. Despite being drafted very high in the amateur draft, his success in the Finnish league did not translate to the NHL. He played in 24 games in the American Hockey League (minor league to the NHL) before leaving the NHL to play overseas. While Jesse Niinimäki was not included in our dataset, he is an example of the difficulty in valuing amateur players and a testament to the challenge of predicting future success.

While some of these examples occur more frequently than others, the number of variables that factor into playing is clear. To control for some of these factors data such as historical statistics, injuries, and skill rankings may be included in future research.

5.1.3 Neurocognitive Efficiency

While the focus of this work is machine learning and not neuroscience, I present developments from modern neuroscience research as an explanation for the difficulty in creating effective classifiers. As described in [49], neurocognitive efficiency is the ratio of work output (motor behavior) to neuromotor effort; essentially the amount of work one's brain must do to achieve a desired physical movement. Evidence of neurocognitive efficiency is seen across multiple sports including golf, volleyball, baseball, and marksmanship [38, 114, 83, 28]. In each of these studies, EEG data was recorded to measure brain activity between experts and novices and shows that experts achieve better performance (motor output) despite lower cortical activation. The studies point toward a decrease in cortical activity related to better visuomotor performance. Thus, I hypothesize that elite athletes, suffering from concussion-induced cognitive deficits, may possess efficient enough mechanisms to introduce variance in their vi-

suomotor performance. This in turn adds noise into our dataset and may enhance the difficulty of creating accurate classifiers.

5.1.4 Limited Hardware Configuration:

The study was limited by hardware configurations. The training was performed using a state-of-the-art 64 Core Threadripper 3990X CPU and NVIDIA RTX A6000 GPU accessed via a distributed remote cluster. However, the distributed server cluster used for training has session timeout limits of 30 minutes at which point the session may be restarted on a different physical machine. For most pipeline configurations, this was not an issue as training was completed within the 30-minute window due to high CPU power. For some models, including Linear Tree and Linear Boosting Tree when trained using Bayesian Optimization, training time exceeded 30 minutes and would not complete if moved to a different physical server. This is the reason that Linear Tree and Linear Boosting Tree results are not included in Table 3.5. Additionally, the Jupyter Lab instance used to host notebooks was not CUDA enabled and thus the GPU was not available for Pytorch DNN training.

5.1.5 AI, Ethics, and Sports: A Commitment to Human Rights

While not a direct limitation of this study, we address a basic commitment to human rights that all artificial intelligence and machine learning models must respect. Specifically, we posit this commitment not as a limitation to the research itself, but to our models used in a production setting.

We acknowledge that the inclusion of AI in sports, although increasingly widespread and powerful, is not without limitations. Specifically as given in [103], AI in sports "could lead to misuses such as cheating, illegal betting, and particularly, to manip-

ulating competitions and dismissing the human value." We put forward that our models should only be used in practice should they uphold the following values of Human Dignity, Autonomy and Protection of the Self, Equality, and Democracy. Furthermore, we summarize a framework for an appropriate application of AI given in [103].

- **Principle of Monmaleficence:** AI should respect its duty to not cause harm to athletes, real or potential, by commission or omission.
- **Principle of Beneficence:** Those who use AI in sports should do so only for the benefit of athletes and others.
- **Principle of Fairness:** The application of AI should be equal to all members and should propagate throughout sports in a way that minimizes discrimination and bias.
- **Principle of Transparency and Accountability:** There exists an onus on those who implement AI in sports to be transparent, share decisions and data, and be accountable for both human decisions informed by AI and AI decisions themselves.

We give the aforementioned principles to advise and caution future users of our models to implement them only with the utmost respect for human rights.

Chapter 6

CONCLUSION AND FUTURE WORK

This section gives the conclusions to my work in addition to suggestions for how successive future efforts to this research may benefit.

6.1 Conclussion

This paper investigates various ML models and their efficacy in using data from a neuroscience test to classify which athletes actually played in an NHL game based on their BrDi™ (visuomotor) performance metrics. As this is exploratory research, this thesis explores a variety of models, preprocessing, and tuning techniques. The thesis finds the best explainable model is Decision Tree and the best black box model is Linear Boosting Tree with generalization scores of .60 and .53 respectively in regards to RQ1. RQ2 gives inconclusive results. However, model tuning results are shown to be inconclusive as many models suffer from overtraining, the ability for a model to generalize to unseen data accurately.

The benefits of explainable models are also discussed when used in conjunction with human decisions. Despite "black box" models achieving generally lower accuracy scores than XAI models, it is illustrated how XAI model visualization allows users to better comprehend results. Explainable models help users to trust decisions, provide accountability behind decisions, and give insights into predictions. Thus, in some scenarios, users may prefer to implement explainable models even if there are black box alternatives that provide better accuracy results.

The final contribution of this thesis lies in the development of automated pipelines that support distributed training of two types: Random Search (SkLearn) and Bayesian Optimization using the Tree-Structured Parzen Estimator algorithm. This is done in order to streamline future work efforts. Simplicity and compatibility are ensured with support for both traditional machine learning models (SVM, Decision Tree, Logistic Regression.) and also Pytorch DNN models.

6.2 Future Work

This section describes potential methods to build on our current research.

Many of the ML models in this research suffer from overtraining, in which they do not perform as well on generalization splits as training. To address the issue of overtraining, I used scaling techniques, hyperparameter tuning and multiple train/test splits (70/30, 80/20, 90/10), cross-validation techniques, and sampling methods.

There are multiple approaches that may be useful to address the overtraining problem. For example, overtraining may also be improved by adding other features that contribute to a player's chances of playing in the NHL. Features such as scouting rank, scoring rate, points, and country are public metrics and have been shown to produce effective classifiers [105, 70, 45].

It may also be useful to repeat this study with multiple modalities and incorporate data such as EEG. Shown in prior works to be effective classifiers of concussion, direct brain wave time series data may be useful to incorporate and has been shown to create successful classifiers also used in this study such as SVM and LR.

With the inclusion of more data, it may be possible to report results from only single trials. In our work, 5 trials are reported to give more confidence in generalization

scores at the cost of accuracy results. This is demonstrated by comparing Table 3.3 and Table 3.3. Reporting single trials shows better accuracy across many different models and may be valid in future work.

Better results may also be achieved by building off the most successful models in this work and continuing training efforts. A breadth of models are prioritized in this thesis due to its exploratory nature while future efforts may concentrate on only models above a threshold accuracy. I develop robust yet easy-to-use pipelines so that future work may focus more heavily on model tuning using our existing architecture for both machine and deep learning.

Many past studies use fewer features than this one and thus future work may also benefit from a greater emphasis on feature selection. While feature selection using multiple techniques is performed, the classifiers may suffer from the Curse of Dimensionality. As stated in [79], the Curse of Dimensionality, describes the fact that as the number of input features increases, more data is needed for models to generalize sufficiently well. Despite the fact the number of features used in our work (39) is not typically considered highly dimensional (usually around 100), the limited number of training samples may contribute to the issue. Thus successor research may benefit from techniques such as forward feature selection, chi-square analysis, Fisher's Score, Principal Component Analysis, and so forth.

Additional methods of simplifying input features may include grouping players in some way and training ML models on groups of similar players in a cohort-based approach similar to [70, 97, 40, 94]. For example, players may be clustered based on playing position, date of injury, number of concussions, or reasons for leaving the NHL other than injury.

Future work may also benefit from building on our success of intelligent hyperparameter tuning Bayesian Optimization. Through comparison between Tables 3.3 and 3.5, it is clear that Bayesian Optimization outperforms Random Search for our dataset. Genetic Algorithms, and similar techniques, have been shown to produce more accurate classifiers in prior experiments and may provide improvement over our methods.

BIBLIOGRAPHY

[1] Cal Poly Github.

[2] Collective bargaining agreement.

[3] *Effects of Sport Related Concussion on Academic Performance in High School Athletes*. PhD thesis.

[4] Neuroassessment in sports: An integrative approach for performance and potential evaluation in athletes. *Frontiers in Psycholog*, 13.

[5] The standardized assessment of concussion (sac).

[6] E. Aidman. Cognitive fitness framework: Towards assessing, training and augmenting individual-difference factors underpinning high-performance cognition. *Frontiers in Human Neuroscience*, 13, 2020.

[7] D. Al-Mfarej. Quantifying upper-limb bimanual coordination performance using machine learning techniques for concussion screening. Master's thesis, University of Waterloo, 2021.

[8] N. J. Aldrich, L. Alfieri, P. Brooks, and H. R. Tenenbaum. Does discovery-based instruction enhance learning? a meta-analysis. 2007.

[9] L. Alfieri, P. J. Brooks, N. J. Aldrich, and H. R. Tenenbaum. Does discovery-based instruction enhance learning? *Journal of educational psychology*, 103(1):1, 2011.

[10] R. B. Ammons. Acquisition of motor skill: Iii. effects of initially distributed practice on rotary pursuit performance. *Journal of Experimental Psychology*, 40(6):777–787, 1950.

[11] C. S. Baker and M. E. Cinelli. Visuomotor deficits during locomotion in previously concussed athletes 30 or more days following return to play. *Physiological Reports*, 2(12), 2014.

[12] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[13] J. W. Britton, L. C. Frey, J. L. Hopp, P. Korb, M. Z. Koubeissi, W. E. Lievens, E. M. Pestana-Knight, and E. St Louis. Electroencephalography (eeg): An introductory text and atlas of normal and abnormal findings in adults, children, and infants. 2016.

[14] J. A. Brown, M. Dalecki, C. Hughes, A. K. Macpherson, and L. E. Sergio. Cognitive-motor integration deficits in young adult athletes following concussion. *BMC sports science, medicine and rehabilitation*, 7(1):1–12, 2015.

[15] J. M. Bruce, R. J. Echemendia, W. Meeuwisse, M. G. Hutchison, M. Aubry, and P. Comper. Development of a risk prediction model among professional hockey players with visible signs of concussion. *British journal of sports medicine*, 52(17):1143–1148, 2018.

[16] A. Bryson, R. Gomez, and T. Zhang. All-star or benchwarmer? relative age, cohort size and career success in the nhl. In *Breaking the Ice*, pages 57–91. Springer, 2017.

[17] T. A. Buckley, K. N. Bryk, K. L. Van Pelt, S. P. Broglio, S. A. East, S. L. Zuckerman, and A. W. Kuhn. Concussion and national hockey league player performance: an advanced hockey metrics analysis. *Journal of athletic training*, 54(5):527–533, 2019.

[18] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[19] C. Cao, R. L. Tutwiler, and S. Slobounov. Automatic classification of athletes with residual functional deficits following concussion by means of eeg signal using support vector machine. *IEEE transactions on neural systems and rehabilitation engineering*, 16(4):327–335, 2008.

[20] S. Chaiken, P. Kylionen, and W. Tirre. Organization and components of psychomotor ability. 1999.

[21] W.-L. Chang and Y. Liu. Exploring the roles of employee attitudes, self-efficacy, and literacy on the intention to use artificial intelligence in the workplace. *Journal of Organizational Computing and Electronic Commerce*, 31(1):29–47, 2021.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[23] B. H. Chow, A. M. Stevenson, J. F. Burke, and E. E. Adelman. the effect of concussion on batting performance of major league baseball players. *Open Access Journal of Sports Medicine*, Volume 10:41–48, 2019.

[24] J. Claesson, E. Häglund, and P. Staberg. Injury prediction in elite ice hockey using machine learning. 2018.

[25] J. N. Cohen, K. M. Thompson, V. K. Jamnik, N. Gledhill, and J. F. Burr. Relationship of fitness combine results and national hockey league performance: A 25-year analysis. *International Journal of Sports Physiology and Performance*, 17(6):908–916, 2022.

[26] J. Cunningham, S. P. Broglio, M. O'Grady, and F. Wilson. History of sport-related concussion and long-term clinical cognitive health outcomes in retired athletes: A systematic review. *Journal of Athletic Training*, 55(2):132–158, 2020.

[27] M. Dalecki, D. Albines, A. Macpherson, and L. E. Sergio. Prolonged cognitive–motor impairments in children and adolescents with a history of concussion. *Concussion*, 1(3):CNC14, 2016.

[28] C. Del Percio, C. Babiloni, M. Bertollo, N. Marzano, M. Iacoboni, F. Infarinato, R. Lizio, M. Stocchi, C. Robazza, G. Cibelli, et al. Visuo-attentional and sensorimotor alpha rhythms are related to visuo-motor performance in athletes. *Human brain mapping*, 30(11):3527–3540, 2009.

[29] V. Dutot, V. Bhatiasevi, and N. Bellallahom. Applying the technology acceptance model in a three-countries study of smartwatch adoption. *The Journal of High Technology Management Research*, 30(1):1–14, 2019.

[30] M. P. Education. The science of strength: How data analytics is transforming college basketball.

[31] C. Esopenko, P. Kumar, C. Alain, T. Chow, R. McIntosh, S. Strother, and B. Levine. Neuroimaging during a working memory task in remote traumatic brain injury: evidence from nhl alumni. In *JOURNAL OF COGNITIVE NEUROSCIENCE*, pages 74–74. MIT PRESS 55 HAYWARD STREET, CAMBRIDGE, MA 02142 USA, 2013.

[32] L. Farah. Evaluating the efficacy of talent identification and development in the national hockey league entry draft. 2022.

[33] M. Faul, M. M. Wald, L. Xu, and V. G. Coronado. Traumatic brain injury in the united states: emergency department visits, hospitalizations, and deaths, 2002-2006. 2010.

[34] M. S. Fine, P. S. Lum, E. B. Brokaw, M. S. Caywood, A. J. Metzger, A. V. Libin, J. Terner, J. W. Tsao, J. N. Norris, D. Milzman, et al. Dynamic motor tracking is sensitive to subacute mtbi. *Experimental brain research*, 234:3173–3184, 2016.

[35] P. C. Fino, M. A. Nussbaum, and P. G. Brolinson. Locomotor deficits in recently concussed athletes and matched controls during single and dual-task turning gait: preliminary results. *Journal of neuroengineering and rehabilitation*, 13(1):1–15, 2016.

[36] P. Freire. *Pedagogy of the oppressed*. Bloomsbury publishing USA, 2018.

[37] I. Gal and L. Ginsburg. The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, 2(2), 1994.

[38] G. Gallicchio, A. Cooke, and C. Ring. Practice makes efficient: Cortical alpha oscillations are associated with improved golf putting performance. *Sport, exercise, and performance psychology*, 6(1):89, 2017.

[39] G. García Botero, F. Questier, S. Cincinnato, T. He, and C. Zhu. Acceptance and usage of mobile assisted language learning by higher education students. *Journal of Computing in Higher Education*, 30:426–451, 2018.

[40] M. Gliznitsa and N. Silkina. Using decision trees to determine the important characteristics of ice hockey players. In *International Russian Automation Conference*, pages 359–369. Springer, 2021.

[41] C. Gough. Nhl league revenue 2005-2021, Aug 2022.

[42] R. Grashow, M. G. Weisskopf, K. K. Miller, D. M. Nathan, R. Zafonte, F. E. Speizer, T. K. Courtney, A. Baggish, H. A. Taylor, A. Pascual-Leone, et al. Association of concussion symptoms with testosterone levels and erectile dysfunction in former professional us-style football players. *JAMA neurology*, 76(12):1428–1438, 2019.

[43] G. Gray, D. Evans, P. Cornwell, F. Costanzo, and B. Self. Toward a nationwide dynamics concept inventory assessment test. In *2003 Annual Conference*, pages 8–1202, 2003.

[44] G. Green, K. M. Pollack Porter, A. Kvit, S. Conte, J. D'Angelo, A. Valadka, and F. C. Curriero. Examining batting performance after a sports-related concussion among major league baseball position players. *The American Journal of Sports Medicine*, 49(3):790–797, 2021.

[45] A. C. Greene. The success of nba draft picks: Can college careers predict nba winners?, Nov 2015.

[46] M. N. Haider, J. J. Leddy, S. Pavlesen, M. Kluczynski, J. G. Baker, J. C. Miecznikowski, and B. S. Willer. A systematic review of criteria used to define recovery from sport-related concussion in youth athletes. *British Journal of Sports Medicine*, 52(18):1179–1190, 2017.

[47] R. J. Haier, B. V. Siegel, K. H. Nuechterlein, E. Hazlett, J. C. Wu, J. Paek, H. L. Browning, and M. S. Buchsbaum. Cortical glucose metabolic rate correlates of abstract reasoning and attention studied with positron emission tomography. *Intelligence*, 12(2):199–217, 1988.

[48] T. A. Hammeke, M. McCrea, S. M. Coats, M. D. Verber, S. Durgerian, K. Flora, G. S. Olsen, P. D. Leo, T. A. Gennarelli, and S. M. Rao. Acute

and subacute changes in neural activation during the recovery from sport-related concussion. *Journal of the International Neuropsychological Society*, 19(8):863–872, 2013.

[49] B. D. Hatfield, K. J. Jaquess, L. Lo, and H. Oh. The cognitive and affective neuroscience of superior athletic performance. *Handbook of Sport Psychology*, page 487–512, 2020.

[50] G. L. Herman. *The development of a digital logic concept inventory.* University of Illinois at Urbana-Champaign, 2011.

[51] V. N. Hoi. Understanding higher education learners' acceptance and use of mobile devices for language learning: A rasch-based path modeling approach. *Computers & Education*, 146:103761, 2020.

[52] T.-M. Hung, T. W. Spalding, D. L. Santa Maria, and B. D. Hatfield. Assessment of reactive motor performance with event-related brain potentials: attention processes in elite table tennis players. *Journal of Sport and Exercise Psychology*, 26(2):317–337, 2004.

[53] J. Hurtubise, D. Gorbet, Y. Hamandi, A. Macpherson, and L. Sergio. The effect of concussion history on cognitive-motor integration in elite hockey players. *Concussion*, 1(3):CNC17, 2016.

[54] M. G. Hutchison, P. Comper, W. H. Meeuwisse, and R. J. Echemendia. A systematic video analysis of national hockey league (nhl) concussions, part i: Who, when, where and what? *British Journal of Sports Medicine*, 49(8):547–551, 2013.

[55] A. Jacquin, S. Kanakia, D. Oberly, and L. S. Prichep. A multimodal biomarker for concussion identification, prognosis and management. *Computers in biology and medicine*, 102:95–103, 2018.

[56] A. E. Jacquin, J. J. Bazarian, D. J. Casa, R. J. Elbin, G. Hotz, D. M. Schnyer, S. Yeargin, L. S. Prichep, and T. Covassin. Concussion assessment potentially aided by use of an objective multimodal concussion index. *Journal of concussion*, 5:20597002211004333, 2021.

[57] I. Jeffreys and J. Moody. *Strength and conditioning for sports performance*. Routledge, 2021.

[58] R. Jenkins. Technology ethics review: Gradimages facial recognition at graduation ceremony, May 2020.

[59] JLikens.

[60] A. Jobin, M. Ienca, and E. Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1:389–399, 2019.

[61] W. Jones. How the nhl entry draft works: A complete guide, May 2021.

[62] W. Jones. What percentage of nhl draft picks make it to the nhl?, Feb 2021.

[63] A. Jović, K. Brkić, and N. Bogunović. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205, 2015.

[64] J. Kamins, E. Bigler, T. Covassin, L. Henry, S. Kemp, J. J. Leddy, A. Mayer, M. McCrea, M. Prins, K. J. Schneider, et al. What is the physiological time to recovery after concussion? a systematic review. *British journal of sports medicine*, 51(12):935–940, 2017.

[65] D. G. Kelty-Stephen, M. Qureshi Ahmad, and L. Stirling. Use of a tracing task to assess visuomotor performance for evidence of concussion and recuperation. *Psychological Assessment*, 27(4):1379–1387, 2015.

[66] A. W. Kuhn, S. L. Zuckerman, D. Totten, and G. S. Solomon. Performance and style of play after returning from concussion in the national hockey league. *The American journal of sports medicine*, 44(8):2152–2157, 2016.

[67] P. A. Lachenbruch. Analysis of data with excess zeros. *Statistical methods in medical research*, 11(4):297–302, 2002.

[68] H. M. Leet-Pellegrini. Conversational dominance as a function of gender and expertise. In *Language*, pages 97–104. Elsevier, 1980.

[69] T. Lehmus Persson, H. Kozlica, N. Carlsson, and P. Lambrix. Prediction of tiers in the ranking of ice hockey players. In *International workshop on machine learning and data mining for sports analytics*, pages 89–100. Springer, 2020.

[70] Y. Liu, O. Schulte, and C. Li. Model trees for identifying exceptional players in the nhl and nba drafts. *Machine Learning and Data Mining for Sports Analytics*, page 93–105, Apr 2019.

[71] D. Long, T. Blunt, and B. Magerko. Co-designing ai literacy exhibits for informal learning spaces. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, 2021.

[72] D. Long and B. Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16, 2020.

[73] Lucas. What is the nhl hockey draft? how does it work? - bs hockey, Jun 2022.

[74] B. Macdonald. An expected goals model for evaluating nhl teams and players. In *Proceedings of the 2012 MIT Sloan Sports Analytics Conference*, 2012.

[75] T. Machado, D. Gopstein, A. Nealen, O. Nov, and J. Togelius. Ai-assisted game debugging with cicero. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2018.

[76] G. Manley, A. J. Gardner, K. J. Schneider, K. M. Guskiewicz, J. Bailes, R. C. Cantu, R. J. Castellani, M. Turner, B. D. Jordan, C. Randolph, and et al. A systematic review of potential long-term effects of sport-related concussion. *British Journal of Sports Medicine*, 51(12):969–977, 2017.

[77] J. Manyika, S. Lund, M. Chui, J. Bughin, J. Woetzel, P. Batra, R. Ko, and S. Sanghvi. Jobs lost, jobs gained: Workforce transitions in a time of automation. *McKinsey Global Institute*, 150, 2017.

[78] L. Margulieux, P. Denny, K. Cunningham, M. Deutsch, and B. R. Shapiro. When wrong is right: The instructional power of multiple conceptions. In *Proceedings of the 17th ACM Conference on International Computing Education Research*, pages 184–197, 2021.

[79] S. Marsland. *Machine learning: an algorithmic perspective*. CRC press, 2015.

[80] D. N. Martini and S. P. Broglio. Long-term effects of sport concussion on cognitive and motor performance: a review. *International journal of psychophysiology*, 132:25–30, 2018.

[81] P. McCrory, W. Meeuwisse, J. Dvorak, M. Aubry, J. Bailes, S. Broglio, R. C. Cantu, D. Cassidy, R. J. Echemendia, R. J. Castellani, et al. Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in berlin, october 2016. *British journal of sports medicine*, 51(11):838–847, 2017.

[82] A. Mojsilovic. Explainable ai (xai), Aug 2019.

[83] H. Nakamoto and S. Mori. Experts in fast-ball sports reduce anticipation timing cost by developing inhibitory control. *Brain and cognition*, 80(1):23–32, 2012.

[84] H. Nakata, M. Yoshie, A. Miura, and K. Kudo. Characteristics of the athletes' brain: evidence from neurophysiology and neuroimaging. *Brain research reviews*, 62(2):197–211, 2010.

[85] S. M. Navarro, R. W. Pettit, H. S. Haeberle, S. J. Frangiamore, N. M. Rahman, L. Farrow, M. Schickendantz, and P. N. Ramkumar. Short-term impact of concussion in the nhl: An analysis of player longevity, performance, and financial loss. *Journal of neurotrauma*, 35 20:2391–2399, 2018.

[86] A. L. Neustadtl, W. K. Bukowski, A. Neustadtl, and D. Milzman. Performance after concussion in national hockey league players. *Journal of athletic training*, 56(4):404–407, 2021.

[87] D. T. K. Ng, J. K. L. Leung, K. W. S. Chu, and M. S. Qiao. Ai literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, 58(1):504–509, 2021.

[88] N. H. Nguyen, D. T. Nguyen, B. Ma, and J. Hu. The application of machine learning and deep learning in sport: Predicting nba players' performance and popularity. *Journal of Information and Telecommunication*, 6(2):217–235, 2021.

[89] B. Olds, R. Streveler, R. Miller, and M. Nelson. Using a delphi study to identify the most difficult concepts for students to master in thermal and transport science (session 1531). In *Proceedings of the 2003 American Society for Engineering Education Annual Conference & Exposition*, 2003.

[90] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[91] E. J. Pellman, M. R. Lovell, D. C. Viano, and I. R. Casson. Concussion in professional football: recovery of nfl and high school athletes assessed by computerized neuropsychological testing—part 12. *Neurosurgery*, 58(2):263–274, 2006.

[92] E. J. Pellman, D. C. Viano, I. R. Casson, C. Arfken, and H. Feuer. Concussion in professional football: players returning to the same game—part 7. *Neurosurgery*, 56(1):79–92, 2005.

[93] T. Persson, H. Kozlica, N. Carlsson, and P. Lambrix. Prediction of tiers in the ranking of ice hockey players. In *International workshop on machine learning and data mining for sports analytics*, pages 89–100. Springer, 2020.

[94] S. Pettigrew. Assessing the offensive productivity of nhl players using in-game win probabilities. In *9th annual MIT sloan sports analytics conference*, volume 2, page 8, 2015.

[95] D. Poltavski, K. Bernhardt, C. Mark, and D. Biberdorf. Frontal theta-gamma ratio is a sensitive index of concussion history in athletes on tasks of visuo-motor control. *Scientific Reports*, 9(1), 2019.

[96] V. Potluri, T. Grindeland, J. E. Froehlich, and J. Mankoff. Ai-assisted ui design for blind and low-vision creators. In *the ASSETS'19 Workshop: AI Fairness for People with Disabilities*, 2019.

[97] M. Puck. Prospect cohort success – evaluation of results, Oct 2015.

[98] N. Reed, P. Fait, K. Zabjek, and M. Keightley. Concussion and concurrent cognitive and sport-specific task performance in youth ice hockey players: a single-case pilot study. *Journal of Neurology & Neurophysiology*, 4(5), 2013.

[99] Y. Register and A. J. Ko. Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics. In *Proceedings of the 2020 ACM Conference on International Computing Education Research*, pages 67–78, 2020.

[100] J. K. Register-Mihalik, A. C. Littleton, and K. M. Guskiewicz. Are divided attention tasks useful in the assessment and management of sport-related concussion? *Neuropsychology review*, 23:300–313, 2013.

[101] Reinl. The ai future of front-end development, Mar 2023.

[102] S. Rowson and S. M. Duma. Brain injury prediction: assessing the combined probability of concussion using linear and rotational head acceleration. *Annals of biomedical engineering*, 41:873–882, 2013.

[103] A. C. Sampedro. The case of ai in sport: Some ethical concerns at play. *Diagoras: International Academic Journal on Olympic Studies*, 5:18–29, 2021.

[104] A. Sarkar, A. D. Gordon, C. Negreanu, C. Poelitz, S. S. Ragavan, and B. Zorn. What is it like to program with artificial intelligence? *arXiv preprint arXiv:2208.06213*, 2022.

[105] M. Schuckers. Draft by numbers: using data and analytics to improve national hockey league (nhl) player selection. In *MIT Sloan Sports Analytics Conference*, pages 1–8, 2016.

[106] M. E. Schuckers and S. Argeris. You can beat the "market": Estimating the return on investment for nhl team scouting. *Journal of Sports Analytics*, 1(2):111–119, 2015.

[107] O. Schulte. Valuing actions and ranking hockey players with machine learning. In *Linköping Hockey Analytics Conference*, pages 2–9, 2022.

[108] O. Schulte, Z. Zhao, M. Javan, and P. Desaulniers. Apples-to-apples: Clustering and ranking nhl players using location information and scoring impact. In *Proceedings of the MIT Sloan Sports Analytics Conference*, 2017.

[109] E. L. Singman. Automating the assessment of visual dysfunction after traumatic brain injury. *Med Instrum*, 1(1):3, 2013.

[110] N. Smeha, R. Kalkat, L. E. Sergio, and L. M. Hynes. Sex-related differences in visuomotor skill recovery following concussion in working-aged adults. *BMC Sports Science, Medicine and Rehabilitation*, 14(1), 2022.

[111] K. Sokol and P. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020.

[112] P. Tingling, K. Masri, and M. Martell. Does order matter? an empirical analysis of nhl draft decisions. *Sport, Business and Management: An International Journal*, 2011.

[113] A. To. Sports analytics:, Dec 2020.

[114] B. Tomasino, M. Maieron, E. Guatto, F. Fabbro, and R. I. Rumiati. How are the motor system activity and functional connectivity between the cognitive and sensorimotor systems modulated by athletic expertise? *Brain research*, 1540:21–41, 2013.

[115] E. Tshukudu and Q. Cutts. Understanding conceptual transfer for students learning new programming languages. In *Proceedings of the 2020 ACM conference on international computing education research*, pages 227–237, 2020.

[116] D. Voyer and E. F. Wright. Predictors of performance in the national hockey league. *Journal of Sport Behavior*, 21(4):456, 1998.

[117] C. Wang and Y. Zhang. Exploring factors that affect learners' engagement with ai courses: The role of social influence. In *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2020.

[118] S. Wheeler. Wheeler: A guide to scouting and evaluating nhl draft prospects, Jun 2022.

[119] G. B. Wilkerson, D. C. Nabhan, and R. T. Crane. Concussion history and neuromechanical responsiveness asymmetry. *Journal of athletic training*, 55(6):594–600, 2020.

[120] D. Wilson. *Mining NHL draft data and a new value pick chart*. PhD thesis, Carleton University, 2016.

[121] J. Woodhouse, D. J. Heyanka, J. Scott, A. Vincent, T. Roebuck-Spencer, K. Domboski-Davidson, K. O'Mahar, and R. Adams. Efficacy of the anam

general neuropsychological screening battery (anam gns) for detecting neurocognitive impairment in a mixed clinical sample. *The clinical neuropsychologist*, 27(3):376–385, 2013.

[122] B. Xie and M. Li. Exploring factors affecting ai literacy among college students in china: A structural equation modeling approach. *International Journal of Emerging Technologies in Learning*, 14(8):163–178, 2019.

[123] K. Yarrow, P. Brown, and J. W. Krakauer. Inside the brain of an elite athlete: the neural processes that support high achievement in sports. *Nature Reviews Neuroscience*, 10(8):585–596, 2009.

[124] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, and D. Zhang. Cognitive workload recognition using eeg signals and machine learning: a review. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.