Concussion Predictions of NHL Players: An Analysis of Classification Models

Ahmed Abbas
Faculty of Mathematics, Data Science
University of Waterloo
Waterloo, Canada
a45abbas@uwaterloo.ca

***Abstract*—Among contact sports, ice hockey has the highest rate of concussion incidence [1]. With the rise of modern AI systems, machine learning techniques have become a popular approach in injury predictions [2]. However, there currently lacks an accurate method of determining the risk of concussion in ice hockey, and imbalanced medical datasets present an ongoing challenge in machine learning models, significantly decreasing their performance [3]. This study aims to determine the validity of supervised machine learning classifier models in predicting the next-season concussion risk of a non-goalie National Hockey League (NHL) player, and compare the performances of these models. NHL player statistics and concussion history were compiled from the 2012-2013 season through to 2022-2023 from 2 publically reported databases, in the absence of an official NHL database [4]. Classifiers used consist of decision tree, random forest, k-nearest neighbours, AdaBoost, and XGBoost. Among these classifiers, random forest had the best F1 score of 0.833, with XGBoost as a close second with an F1 score of 0.827.**

***Index Terms*—NHL, machine learning classifiers, concussion, imbalanced dataset, sports medicine, injury prediction**

I. Introduction

Ice hockey is a fast and physical sport, with the National Hockey League (NHL) harboring the most skilled players in the world. Players of this league approach speeds of around 30 kilometers per hour on a regular basis, while exceptionally fast skaters such as Connor McDavid have been clocked eclipsing 40 kilometers per hour. Coupling that terrifying speed with a deep-rooted cultural emphasis on physical contact and aggressive behavior [5], it's easy to grasp why injuries are commonplace in the hockey world.

Among these injuries, concussions account for roughly 10% of them during the NHL regular season [6], [7]. A concussion is clinically defined as "a complex pathophysiologic process affecting the brain, induced by biomechanical forces" such as a direct or indirect blow to the head [8]. In 2018, Navarro et al uncovered that NHL concussions led to shorter career lengths, salary reductions and decreased performance when compared with non-concussed controls [9]. Concussions also bear an economic burden of salary costs lost, and this is estimated to cost the league $42.8 million a year in missed player time, making up 20% of all injury costs [10]. Gaining an increased understanding of the consequences of sports-related concussions has garnered public interest [11].

In 1997, the NHL began collecting data on concussions after jointly establishing the NHL-NHLPA Concussion Program with the NHL Players Association (NHLPA). The goal of the program was to gain a scientific understanding of the injury and to contribute new knowledge to the field [12]. Furthermore, in 2010, the League implemented the NHL Concussion Protocol out of concern for player head injuries. This protocol was updated in 2015 to allow for professionally trained spotters to identify players showing any visible signs of concussion during games and immediately remove them from play [13]. In a world-class league that prioritizes concussion safety and player health, the ability to predict which players are vulnerable to concussion at any given moment can also be a promising strategy. Therefore, machine learning classifiers may be a suitable tool for this purpose.

Due to the rise of modern technology, machine learning has introduced itself as a strong approach in medical prediction [3]. Machine learning, a subset of artificial intelligence, is defined as the application of statistical algorithms that can learn patterns from data and then generalize this to unseen data without explicit human intervention [14]. In other words, machine learning is a technique capable of analyzing big datasets and learning from this data to make predictions about the future [15]. However, medical datasets often suffer from the issue of being imbalanced. An imbalanced dataset is when the proportional class sizes in the dataset differ significantly, causing machine learning classifiers to be partial towards the overrepresented class (majority class) [3]. In our use case, concussions in NHL players are reasonably rare compared to healthy players or players with other injuries, posing a challenge for the classification between the concussion and non-concussion NHL population.

The best known technique for dealing with an imbalanced dataset is the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic data from the minority class along the lines of the minority data [16]. This can improve the accuracy of classifiers for a minority class because the new synthetic examples from the minority class are plausible by being close in feature space to existing examples of the minority class. The original paper on SMOTE by Chawla et al. also suggested to combine random under-sampling of the majority class with SMOTE [17].

Machine learning classifiers demonstrate different behaviors depending on their configuration. Therefore, it is important to tune their parameters when trying to produce accurate predictions [18]. Hyperparameters are not directly learned within a model, they are passed into the model programmatically. Determining the best hyperparameters for the model is usually based on a scoring criteria such as accuracy or F1 score of the model. One method to find the best hyperparameters is grid search, which exhaustively tests every combination of a specified range of parameters, returning the best combination based on the scoring criteria [19].

Existing research has explored utilizing machine learning to predict NHL game outcomes. In 2019, Gu et al. detailed an ”expert system” that can predict NHL game outcomes through big data and machine learning [20]. In 2015, Demers expressed that support vector machines were superior to relevance vector machines in the context of predictive ability in the NHL playoffs [21]. More recent research has investigated the application of machine learning techniques to injury prediction in the NHL. In 2020, Luu et al. described advanced ML models such as XGBoost and random forest have a strong capability in predicting next-season NHL player injury [4]. However, this previous research lacks the use of machine learning to predict relatively uncommon yet serious injuries in the NHL, such as concussions. This is likely due to the trouble associated with imbalanced datasets.

To allow for data-driven concussion prevention by leveraging available player data, this study aims to explore and evaluate the performance of various supervised machine learning classifiers in predicting next-season concussion occurrence for non-goalie NHL players. The classifiers will predict next-season concussion occurrence based on features such as concussion history and player performance statistics.

II. Methodology

A. Dataset and Data Preprocessing

Concussion data for non-goalie NHL players from the years 2012 to 2023 was provided by Kaggle, uploaded by Joseph McLauchlan [22]. He sourced this data from player profile pages on The Sports Network and Sports Forecaster. Relevant features from this dataset included player name, position, NHL season the concussion was sustained, and number of games missed due to the concussion. The total number of concussions a player sustained and the total number of games missed was calculated by season, shortening the dataset to one row per player per season. In other words, each season was treated independently from every other so past concussions were not carried forward to future years.

A second dataset of player performance statistics over the same stretch of time was provided by MoneyPuck, a free to use public dataset [23]. These two datasets were then matched with each other using player name and NHL season year, resulting in a dataset of player statistics and concussion history for each season from 2012 to 2023. The target feature for each record was the event that the player had a concussion in the next season. As a result, the 2023 season was dropped from the dataset due to having no 2024 concussion data. All data wrangling operations were completed using the Python (version 3.12.2) pandas (version 2.2.2) library.

At this point it was acknowledged that the resulting dataset was heavily imbalanced as expected, with only 255 out of 8,993 player records being labeled as sustaining a next-season concussion. This would affect machine learning model performance significantly [3]. Nevertheless, giving up on this approach would be admitting defeat, so it was decided to use a combination of SMOTE and random under-sampling. The minority class was over-sampled using SMOTE to have 10% of the number of examples of the majority class. Next, random under-sampling was used to reduce the number of examples of the majority class to have 50% more than the minority class. These operations were performed by the SMOTE and RandomUnderSampler classes respectively from the imbalanced-learn library (version 0.12.4) in Python, using a random_state parameter of 42. This resulted in a final dataset of 2,619 player records, out of which 873 sustained a next-season concussion.

B. Feature Selection

The ”position” feature was remapped to an integer from 1 to 4 for center, left wing, right wing, and defense respectively. Features in the dataset were assessed for multicollinearity using the variance inflation factor (VIF) in an ordinary-least-squares regression context using the Statsmodels library (version 0.14.4) in Python. A high VIF indicates that features are highly correlated with each other, which can impact the reliability of a model. Features with the maximum VIF were removed sequentially until all features had a VIF of less than 10. This reduced the number of features in the dataset from 152 to 21. The final features selected for players and their respective VIFs are detailed in Table I.

C. Classifiers

The supervised machine learning classifiers used in this study include decision tree (DT), random forest (RF), k-nearest neighbors (KNN), and AdaBoost. These models were implemented using the DecisionTreeClassifier, RandomForestClassifier, KNeighborsClassifier, and AdaBoostClassifier classes respectively from the scikit-learn library. In addition, the XGBoost classifier was also used, created using the XGBoost Python library (version 2.1.1).

D. Hyperparameter Tuning

To achieve highest performance from the classifiers, hyperparameter tuning was done with the grid search cross-validation approach using the GridSearchCV function in scikit-learn. Cross-validation is used under each configuration combination for a classifier to assess the scoring of the model under those settings. In cross-validation, the dataset is partitioned into multiple subsets (folds), using one of the folds as a validation set and the rest of the data as a training set. This process is repeated for each different fold. In GridSearchCV, F1 score was used as the scoring criteria, and 5 folds were used for cross-validation. The range of hyperparameters used

TABLE I
SELECTED FEATURES AND THEIR VIFS

Feature	VIF
position	3.155476
iceTimeRank	4.636949
secondaryAssists	6.455156
reboundGoals	4.955718
playStopped	4.526071
penalties	8.282339
hits	5.589574
takeaways	8.405989
lowDangerGoals	4.324064
mediumDangerGoals	5.054751
highDangerGoals	7.832568
defensiveZoneGiveaways	6.667384
expectedGoalsFromActualReboundsOfShots	5.630439
faceoffsWon	1.858770
penaltyMinutesDrawn	7.181589
shotsBlocked	9.913655
onIceAgainstReboundGoals	7.619589
expectedGoalsForAfterShifts	9.607867
expectedGoalsAgainstAfterShifts	6.031173
gamesMissedFromConcussion	1.671085
numberOfConcussions	1.705655

TABLE II
RANGE OF PARAMETERS USED IN GRIDSEARCHCV

Parameter	Range
dt.max_depth dt.min_samples_split dt.min_samples_leaf dt.criterion	[3, 5, 10, None] [2, 5, 10] [1, 2, 5] ["gini", "entropy"]
rf.n_estimators rf.max_depth rf.min_samples_split rf.min_samples_leaf rf.max_features	[50, 100, 200] [None, 10, 20] [2, 5, 10] [1, 2, 4] ["sqrt", "log2", None]
knn.n_neighbors knn.weights knn.p	[3, 5, 7, 9] ["uniform", "distance"] [1, 2]
AdaBoost.n_estimators AdaBoost.learning_rate	[50, 100, 150] [0.01, 0.1, 1.0]
XGBoost.n_estimators XGBoost.learning_rate XGBoost.max_depth XGBoost.subsample	[50, 100, 150] [0.01, 0.1, 1.0] [3, 5, 7] [0.8, 1.0]

in GridSearchCV is outlined in Table II. Any parameters not listed are using the default in their respective Python libraries.

Table III contains the final hyperparameters after tuning with the grid search cross-validation approach.

TABLE III
FINAL HYPERPARAMETERS USED IN MODELS

Parameter	Value
dt.max_depth dt.min_samples_split dt.min_samples_leaf dt.criterion	None 10 1 "entropy"
rf.n_estimators rf.max_depth rf.min_samples_split rf.min_samples_leaf rf.max_features	100 20 2 1 "log2"
knn.n_neighbors knn.weights knn.p	3 "distance" 1
AdaBoost.n_estimators AdaBoost.learning_rate	150 1.0
XGBoost.n_estimators XGBoost.learning_rate XGBoost.max_depth XGBoost.subsample	150 0.1 7 0.8

E. Machine Learning Model: Validation

All machine learning modeling was performed on a Lenovo Legion Pro 5i laptop with a 13th Gen Intel Core i7 2.10 GHz processor. After determining the best hyperparameters, a repeated stratified k-fold cross-validation approach was used to evaluate the model. Stratified means that each fold of the cross-validation split will have the same class distribution as the dataset, which in our case is 1:2. Three repeats of 10-fold cross-validation was used, therefore fitting and evaluating 30 models of the dataset for each classifier. Each fold had the data split into two sets: 80% as the training set and 20% as the test set. This was implemented using the RepeatedStratifiedKFold class from scikit-learn with a random_state parameter of 1.

Evaluation metrics were calculated for each of the 30 models, and the mean and standard deviation were computed to arrive at a final score for the classifier. Metrics calculated consist of accuracy, precision, recall, and F1 score. The accuracy of the model defines the percentage of players correctly classified by the model. Precision is the percentage of players that the model correctly identifies as sustaining a next-season concussion out of all players the model predicted to sustain a concussion. Recall can be considered the flip of precision, describing how often the model correctly identifies next-season concussions out of all players that actually sustained a next-season concussion. F1 score is the weighted average of precision and recall. All metrics were calculated using scikit-learn, and the mean and standard deviation were computed using the NumPy Python library (version 2.1).

Figures 1 and 2 summarize the entire predictive model process.

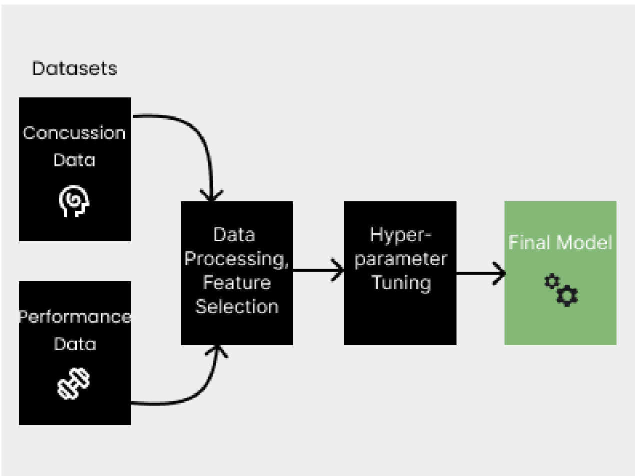

Fig. 1. Diagrammatic representation of finding the best performing model.

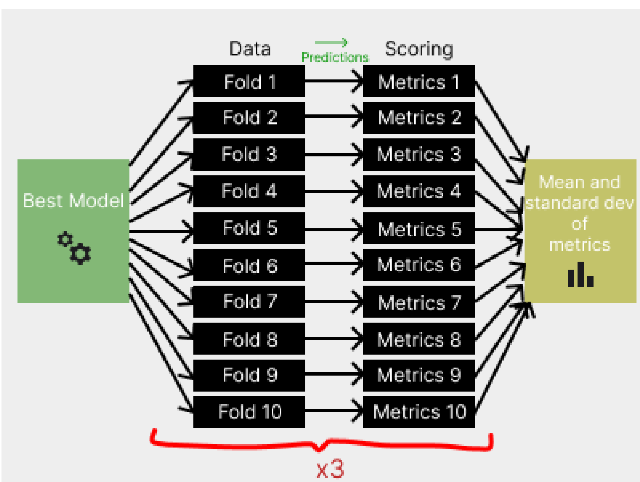

Fig. 2. Diagrammatic representation of evaluating the model.

III. Results and Discussion

A. Results

Table IV shows the accuracy, precision, recall, and F1 score of each classifier for predicting next-season concussion. Random forest achieved the highest accuracy, precision, and F1 score, achieving 0.901, 0.952, and 0.833 respectively, with XGBoost not far behind in all categories. K-nearest neighbors had the highest recall by a wide margin with 0.919.

TABLE IV
Classification Model Performance Results

Model	Accuracy	Precision	Recall	F1 Score
DT	0.814 ± 0.024	0.723 ± 0.044	0.722 ± 0.047	0.721 ± 0.036
RF	0.901 ± 0.017	0.952 ± 0.028	0.741 ± 0.042	0.833 ± 0.030
KNN	0.815 ± 0.024	0.660 ± 0.030	0.919 ± 0.032	0.768 ± 0.026
AdaBoost	0.804 ± 0.022	0.833 ± 0.055	0.519 ± 0.049	0.638 ± 0.044
XGBoost	0.897 ± 0.015	0.933 ± 0.028	0.745 ± 0.045	0.827 ± 0.028

B. Discussion

The goal of this study was to determine the validity of using machine learning models in predicting next-season concussion risk in non-goalie players, and comparing the performance of the models. It was hypothesized that machine learning classifiers can be a suitable tool in concussion prevention by providing the NHL with data-driven insights. Such a tool could provide team physicians and coaching staff with the opportunity to provide preventative care [4].

From two publicly available resources, a dataset was compiled detailing non-goalie NHL player performance statistics and previous season concussion history. Features exhibiting multicollinearity were removed by assessing them using the variance inflation factor. Machine learning classification models were tuned for hyperparameters using cross-validation grid search. These final were applied to the final dataset to predict next-season concussion with a high accuracy. Random forest and XGBoost classifiers provided the best performance in terms of accuracy, precision, and F1 score.

In comparison to past research conducted by Luu et al. on predicting next-season injury, the results align in this study, wherein random forest and XGBoost were among the most accurate classifiers. These models provided strong performance on predicting next season concussion with F1 scores of 0.833 and 0.827 respectively. A surprising result was k-nearest neighbors exhibiting interesting behavior with the best recall of 0.919, despite having a comparatively lower precision score of 0.660. This means that k-nearest neighbors often falsely predicts players to sustain a next-season concussion (false positives), but it very rarely falsely predicts players that do not sustain a next-season concussion (false negatives). The opposite story is true for AdaBoost.

In 2011, the one and only detailed study from the NHL-NHLPA Concussion Program emerged. Benson et al. reported a mean of 80 concussions per season with a median time of six days of playing time lost after examining 559 physician-diagnosed regular season concussions [12]. The application of machine learning and big data has started to transform medicine, and it will be critical for doctors to adjust and improve [24]. Physicians do not necessarily need to understand the technical concepts behind preparing machine learning models, but their practices have the potential to be improved with the ability to interpret the outputs of models.

Despite the strengths of machine learning, the results of a classification model should not be the only deciding factor in determining player concussion risk. The automatic response should not be reducing a player's ice time if they are classified as being vulnerable to concussion. Medical professionals and coaches ought to combine their own knowledge and experience with the outputs to effectively come to a decision on how to handle a player's health.

This study is also not without its limitations. The main limitation was the usage of an imbalanced dataset and attempting to negate this imbalance with synthetic data generated through SMOTE. While SMOTE is an effective method of generating extra examples of the minority class and improving classification performance [25], there are some cases where performance might not improve. In the paper *To SMOTE, or not to SMOTE?* [26], the authors described how state-of-the-art (SOTA) classifiers such as random forest and XGBoost do not achieve improved prediction performance after balancing the dataset. Therefore, the results of the SOTA classifiers used in this study should be taken with a heavy grain of salt. Another limitation is the small amount of data. Only 255 player records were labeled as sustaining a next-season concussion, which results in a small amount of data that SMOTE can generate synthetic samples from. A larger dataset might expose a different perspective of the features on this minority class. Finally, there's a limitation in the lack of publicly available data surrounding the NHL. Most NHL teams hire analysts that record custom metrics [27], and these can surpass the quality of metrics that are publicly accessible. Access to a larger number of features could fine-tune the process of feature selection, and thus improve model performance.

A potential area of interest for future study could be to use Shapley Additive Explanations (SHAP) plots to ascertain the impact of each feature on NHL concussions. A feature with a higher SHAP value would indicate that feature predicts higher concussion probability and vice versa. This study can also be adjusted by propagating concussion history to future seasons, instead of only taking into account the concussion history of the previous season. This would have a potential effect on the prominence of the gamesMissedFromConcussion and numberOfConcussions features in the models. Another area of interest could be to repeat this study for other major sports leagues such as the National Football League (NFL), where concussions are also an injury of high concern. It will be intriguing to see if NFL concussions result in an imbalanced dataset as well.

IV. Conclusion

To summarize, this study compared the performance of several machine learning classification algorithms in predicting next-season concussion in non-goalie NHL players. The best performing classifiers were discovered to be random forest and XGBoost, which provided strong accuracy and F1 scores. This demonstrates the potential of machine learning to improve preventative concussion measures in the NHL. However, it is critical to acknowledge the downsides of imbalanced datasets and the usage of synthetic data. Future research areas may focus on predicting a more broad scope of injuries, injuries in different sports, addressing the challenges posed by imbalanced datasets, and SHAP value analysis to portray the impact of various features on concussions.

References

[1] D. Goodman, M. Gaetz, and D. Meichenbaum, "Concussions in hockey: there is cause for concern:," Medicine & Science in Sports & Exercise, vol. 33, no. 12, pp. 2004–2009, Dec. 2001, doi: 10.1097/00005768-200112000-00005.

[2] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC Med Res Methodol, vol. 19, no. 1, p. 64, Dec. 2019, doi: 10.1186/s12874-019-0681-4.

[3] H. Kaur, H. S. Pannu, and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions," ACM Comput. Surv., vol. 52, no. 4, pp. 1–36, Jul. 2020, doi: 10.1145/3343440.

[4] B. C. Luu et al., "Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An Analysis of 2322 Players From 2007 to 2017," Orthopaedic Journal of Sports Medicine, vol. 8, no. 9, p. 2325967120953404, Sep. 2020, doi: 10.1177/2325967120953404.

[5] M. D. Cusimano et al., "Aggression, Violence and Injury in Minor League Ice Hockey: Avenues for Prevention of Injury," PLoS ONE, vol. 11, no. 6, p. e0156683, Jun. 2016, doi: 10.1371/journal.pone.0156683.

[6] C. D. McKay, R. J. Tufts, B. Shaffer, and W. H. Meeuwisse, "The epidemiology of professional ice hockey injuries: a prospective report of six NHL seasons," Br J Sports Med, vol. 48, no. 1, pp. 57–62, Jan. 2014, doi: 10.1136/bjsports-2013-092860.

[7] R. A. Wennberg and C. H. Tator, "Concussion Incidence and Time Lost from Play in the NHL During the Past Ten Years," Can. j. neurol. sci., vol. 35, no. 5, pp. 647–651, Nov. 2008, doi: 10.1017/S031716710000946X.

[8] M. Aubry et al., "Summary and Agreement Statement of the 1st International Symposium on Concussion in Sport, Vienna 2001:," Clinical Journal of Sport Medicine, vol. 12, no. 1, pp. 6–11, Jan. 2002, doi: 10.1097/00042752-200201000-00005

[9] S. M. Navarro et al., "Short-Term Impact of Concussion in the NHL: An Analysis of Player Longevity, Performance, and Financial Loss," Journal of Neurotrauma, vol. 35, no. 20, pp. 2391–2399, Oct. 2018, doi: 10.1089/neu.2017.5611.

[10] L. Donaldson, B. Li, and M. D. Cusimano, "Economic burden of time lost due to injury in NHL hockey players," Inj Prev, vol. 20, no. 5, pp. 347–349, Oct. 2014, doi: 10.1136/injuryprev-2013-041016.

[11] J. T. Parsons and C. Baugh, "The evolving landscape of policies, rules, and law in sport-related concussion," in Handbook of Clinical Neurology, vol. 158, Elsevier, 2018, pp. 257–267. doi: 10.1016/B978-0-444-63954-7.00025-2.

[12] B. W. Benson, W. H. Meeuwisse, J. Rizos, J. Kang, and C. J. Burke, "A prospective study of concussions among National Hockey League players during regular season games: the NHL-NHLPA Concussion Program," Canadian Medical Association Journal, vol. 183, no. 8, pp. 905–911, May 2011, doi: 10.1503/cmaj.092190.

[13] E. Andrews, T. R. Jildeh, M. J. Abbas, K. Lindsay-Rivera, J. Berguson, and K. R. Okoroha, "Concussions in the National Hockey League: Analysis of Incidence, Return to Play, and Performance," Orthopaedic Journal of Sports Medicine, vol. 10, no. 1, p. 23259671211052069, Jan. 2022, doi: 10.1177/23259671211052069.

[14] S. A. Bini, "Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?," The Journal of Arthroplasty, vol. 33, no. 8, pp. 2358–2361, Aug. 2018, doi: 10.1016/j.arth.2018.02.067.

[15] A. L. Beam and I. S. Kohane, "Big Data and Machine Learning in Health Care," JAMA, vol. 319, no. 13, p. 1317, Apr. 2018, doi: 10.1001/jama.2017.18391.

[16] D. Ramyachitra and P. Manikandan. “IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS : A REVIEW.” International Journal of Computing and Business Research. vol. 5, no. 4, Jul. 2014.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” jair, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[18] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, “Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data,” Ecological Modelling, vol. 406, pp. 109–120, Aug. 2019, doi: 10.1016/j.ecolmodel.2019.06.002.

[19] P. Liashchynskyi and P. Liashchynskyi, “Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS,” Dec. 12, 2019, arXiv: arXiv:1912.06059. doi: 10.48550/arXiv.1912.06059.

[20] W. Gu, K. Foster, J. Shang, and L. Wei, “A game-predicting expert system using big data and machine learning,” Expert Systems with Applications, vol. 130, pp. 293–305, Sep. 2019, doi: 10.1016/j.eswa.2019.04.025.

[21] S. Demers, “Riding a probabilistic support vector machine to the Stanley Cup,” Journal of Quantitative Analysis in Sports, vol. 11, no. 4, Jan. 2015, doi: 10.1515/jqas-2014-0093.

[22] J. McLauchlan, ”NHL Concussion Data 2012-2023,” Kaggle, Sep. 2024, [Dataset], Available: https://www.kaggle.com/datasets/josephmclauchlan/nhl-concussion-data-2012-2023. [Accessed: November 8, 2024]

[23] P. Tanner, ”2012-2023,” MoneyPuck, Dec. 2024, [Dataset], Available: https://moneypuck.com/data.htm. [Accessed: November 8, 2024]

[24] Z. Obermeyer and E. J. Emanuel, “Predicting the Future — Big Data, Machine Learning, and Clinical Medicine,” N Engl J Med, vol. 375, no. 13, pp. 1216–1219, Sep. 2016, doi: 10.1056/NEJMp1606181.

[25] D. Elreedy and A. F. Atiya, “A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance,” Information Sciences, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.

[26] Y. Elor and H. Averbuch-Elor, “To SMOTE, or not to SMOTE?,” May 11, 2022, arXiv: arXiv:2201.08528. doi: 10.48550/arXiv.2201.08528.

[27] S. W. Sheps, “Corsi, Fenwick and Gramsci: How bloggers and advanced analytics are changing the National Hockey League,” International Review for the Sociology of Sport, vol. 55, no. 8, pp. 1192–1211, Dec. 2020, doi: 10.1177/1012690219869192.