## **Theoretical Research**

### Literature Review

Formula One is a sport of fine margins where driver and car performance is differentiated to a tenth of a second. A competitive edge for any of these teams over their rivals can pave the way to success throughout the season and improve the finishing position of their cars per race. The introduction of data analytics and more specifically, advanced machine learning algorithms has enabled Formula One teams to process real time data and improve their strategic decision making on a race [1]. The prominence of data analytics in the sport poses the question to whether a race outcome can be accurately predicted and this exploration can be generally divided into three different spaces: psychological, mathematical and machine learning techniques.

When looking at the key concepts and theories around the prediction of sport results, Artificial Neural Networks (ANNs) seem to be the most common unsupervised approach to sport result predictions [2]. An ANN uses interconnected components to translate several inputs into the desired learning output and its power comes from the non-linearity of the hidden neurons, where the weight adjustment here contributes to the final model decision [3] [4]. A study by Bunker and Thabtah [5] [6] examined the application of unlabelled data for sports winner prediction. The ANNs attained a greater accuracy score of 71% once the most suitable metrics for the data had been chosen. [7]

Passfield [8] delves into the emergence of sport analytics and the mix of specialties involved such as mathematics, human physiology and big data. The study provides positive sentiment towards the ability sport science has in reforming knowledge based on player behaviour and sports in general.

The study of Neural Networks in the sport of greyhound racing conducted by Johansson and Sönströd [9] found that the methods performed better at the harder formats and strived for greater value rather than reverting to the simple choice of the favourite in a greyhound race, even when the odds were not supplied to the model for this study. The study demonstrated the complexity at which the Neural Networks perform and their desire to strive for the best possible outcomes when provided with the necessary data to do so.

Another study on greyhound racing predictions is that of Chen, Rinde, Sutahjo, Sommer [10] where two machine learning approaches were deployed, decision-tree and artificial neural networks. This was complemented with three human track experts to compare the algorithm's accuracy against the experts. With the assistance of the experts, the performance variables were reduced from 50 to 10. The results showed that both the decision tree and neural network methods outperformed the human experts for expected profit.

Davoodi, Khateymoori [11] also applied artificial neural networks to predict horse racing results. Data from 100 races at the Aqueduct racetrack in New York from 2010 were selected, where one neural network was used for each horse in a race. The eight features consisted of the horses weight, class of race, horse trainer, jockey, race distance, number of horses in the race, track conditions and weather [12]. These eight features made up the input layer, two hidden layers, and the output layer defined as the horse's finishing time. With over 400 epochs in place, an accuracy of 77% was achieved.

When looking at studies based on athletics and how this data is interpreted based on feature selection, Ofoghi, Zeleznikow, MacMahon, Rehula, and Dwyer [13] published a study regarding the prediction of triathlon results. It emphasised the importance of the data processing phase for this research and the relevant factors that can influence the winner’s performance. When importing the data, the researchers learnt that converting the time variable from HH:MM:SS format to raw seconds helped to emphasise the dominant performance of the leader and other participants against the field.

In the paper of Przednowek, Iskra and Przednowek [14], a similar study is conducted on the athletics discipline and specifically, a 400-metres hurdle race. The data of 21 athletes from the Polish National Team is examined through nonlinear methods [15] and the artificial neural network returns a prediction error of 0.72 seconds, with common predictors for the results also specified for these results such as age, speed endurance and other impacting factors.

Edelmann-Nusser, J.; Hohmann, A.; Henneberg, B. [16] explored the performance of female swimmers in the finals of the 2000 Olympic Games in Sydney by also using nonlinear neural networks (multi-layer perceptrons) through training data from 19 competitions prior to the event, which they found was insufficient in accurately training such networks. [17] The neural models were validated using the ‘leave-one-out’ method, resulting in effective risk-aversion and an error prediction of 0.05s. [18] [19]

Abut, Akay, Daneshvar, Hei [20] undertook the challenge of predicting the racing time accuracy of cross-country skiers. The study concluded that the 3 artificial neural networks used; Multilayer Feed-Forward Artificial Neural Network (MFANN) [21], General Regression Neural Network (GRNN) and Radial Basis Function Neural Network (RBFNN) [22], are fit for making predictions for skiing events due to acceptable error rates from these models.

Another study conducted by Sankaranarayanan, Sattar, Lakshmanan [23] analysed the prediction of a cricket game outcome through regression models while the game was in progress. The application of Ridge regression and attribute bagging algorithms helped to provide a prediction accuracy of 70%, higher than any reported studies in cricket mining. Factors that could influence the game outcome such as the toss, batsmen historical data and other parameters through the use of linear regression and nearest neighbour clustering algorithms [24].

Bailey, Clarke [25] also explored machine learning methods to predict the outcome of a cricket match between two teams and their respective features. The feature selection within this study is relatively small and does not offer a direct comparison between this study where ten individual teams will be examined and the feature selection is expected to be higher. [26]

Various studies are available within the sphere of European football and the machine learning methods applied to this data. Tax and Joustra [27] researched a result prediction system for the Dutch Eredivisie league by testing different combinations of dimensionality reduction techniques and classification algorithms. Joseph, Fenton, Neil [28] published a study based solely on the prediction of game results for the Tottenham Hotspur team during the years of 1995-1997 using a Bayesian network which was constructed by a domain expert and found that the domain knowledge [29] of this expert helped the network to outperform machine learning models that are constructed based on data analysis.

Goddard [30] approached the football prediction challenge from a different angle to most by comparing the modelling of goals scored versus the modelling of win-draw-lose results, rather than just using the previous match results. He concluded that a hybrid model achieved the best possible prediction based on the 25 years worth of data collated from English football matches. The features chosen here include the importance of certain games and the geographical distance between the two opponents.

Hucaljuk [31] used data from the Champions League tournament to set a target accuracy of 60% from their Naive Bayes, K-nearest neighbour and Random Forest techniques. The results outperformed this target accuracy by almost 9% but Hucaljuk outlines the possibility of improving feature selection to include individual player data in the quest to achieve greater results in future works. This would obviously increase the workload significantly and require more resources to achieve this but has greater implications for overall accuracy return from the algorithms.

Broich [32] used data from 153 games within the German Bundesliga to highlight the importance of efficiency when looking at the significance parameters for a match outcome [33]. The number of shots and passes taken by a team, along with the number of touches in the game were deemed important variables also. The study also provided insight that the expected impact of the distance covered by a team did not significantly influence the winning attribute.

Kampakis published two studies which looked at data in relation to the English Premier League. The first study [34] used Twitter posts to predict the outcome of these Premier League matches and proved that the social media platform has useful information for this objective. The author reported that the Twitter-based model performed better than historical data and statistics but increased its accuracy further when mixed with other prediction models.

The second study [35] involved collaborating with two Premier League clubs to predict player injuries in football. This included using machine learning algorithms to predict the recovery time needed for a specific injury, where correlation-based feature selection was used to improve the performance and reliability of this model. The prediction of a player getting an injury was studied through the relationship between the hours of training and matches played by the player with the number of injuries recorded. With the consent of the club, the GPS player data was analysed to see if overtraining or fatigue were contributors to an injury and also to predict if a player is threatened by injury based on the collated data.

Both Rue [36] and Pollard [37] deployed the iterative simulation technique Markov Chain Monte Carlo [38] to predict the outcome of English football matches. A dynamic structure was used in their models to facilitate the updating of parameters once more game results were fed into the existing dataset.

Ulmer, Fernandez, Peterson [39] discussed the use of machine learning algorithms within Premier League data to predict match outcomes based on the 2012-2014 seasons. The collated results confirmed that the outcome of a draw within these results were hard to predict and these predictions were detrimental to the accuracy of the prediction model. The algorithms used included stochastic gradient descent [40], linear support vector machines and random forest [41].

McCabe, Trevathan [42] took a broader approach to their sports prediction results through four sports; Football, Australian Rules Football, Rugby Union and Rugby League. The artificial neural network had 20 input nodes, 10 nodes in the hidden layer and 1 output node. The feature selection consisted of features that were common to all sports included, with any specific to just the one sport not included [43]. The algorithm returned an average performance of 67.5% in this study.

Haghighat, Rastegari, Nourafza [44] conducted a similar study into the data mining techniques for predictions of sport results. They discuss various classification methods such as Artificial Neural Networks, Support Vector Machines, Bayesian Method, Decision Trees, Fuzzy System and Logistic Regression [45] [46] [47]. The study outlined the advantages and disadvantages of each method used and could not provide an overall best method due to the different sports tested in the paper.

The use of American sport data has also been included in this review due to its wide availability and both Purucker [48] and Khan [49] used statistical data from the National Football League in America to deploy Artificial Neural Networks for the prediction of NFL games, to which the outcomes proved both equal and better than the experts respectively through a combination of different classification algorithms.

At college level, Leung and Joseph [50] challenge the use of traditional statistical comparison between the two competing teams of that given year, as players' skillsets are more interchangeable due to the turnover of players at this level. They adopt a prediction method based on the historical results of American college football teams overall. Another college football study by Delen, Cogdell, Kasap [51] found that regression-based predictions were better than classification-based ones by roughly 11% with the data provided.

Talukder [52] explored the possibility of injury within the National Basketball Association (NBA) players and used machine learning techniques to create a model for injury prevention during a game. The model achieved high accuracy figures for short-term injury prediction and ranked injuries based on the economical expense caused and the impact a player would have on the team’s performance if he was not strategically rested instead.

Kuehn [53] devised a framework for evaluating NBA players skillset and how these compliment the overall performance of the team. A probabilistic model for possessions, which considered factors such as the opponents lineups and in-play events revealed that players are mostly considered on their individual statistics in the sport and does not place enough emphasis on their complementary contributions to the overall team performance.

Wiseman [54] used various regression algorithms to predict the winning golf score on a PGA event based on the scores after round 1. Through linear regression, neural network, Bayesian linear regression and decision tree regression, the prediction of the winning score was accurate 67% of the time [55]. Feature selection included the round 1 leading score, round 1 average score, course par, major event and course yardage. The results were validated on 2016 scores based from the models training of data from 2004 to 2015.

Although the study of Formula one is limited, there is motorsports literature available through the examination of American NASCAR racing to support this. Pfitzner & Rishel [56] delved into the possibility of reliable predictors existing in the outcome of a race through correlation analysis and found that certain factors such as qualifying speed, pole position and driver experience. One of the main conclusions from this study was the effect of team membership i.e., number of team members being two or three, however, this will be irrelevant in this study due to the number of drivers per team being fixed to two. A similar study performed by Allender [57] found a correlation between the starting position of each driver and their experience to their finishing position in a race through the regression model used. This study questions the element of flair over experience which will be an interesting point of exploration for this research.

Depken, Mackay [58] offer an insight into the structure of a team motorsport like Formula One when they analysed the NASCAR Sprint Cup Series from 2005 to 2008, using multiple regression models. The conclusion from this was multi-car teams perform better in overall team finishing positions and cooperation.

The research papers so far have focused on data generated from the players within their respective sports. In Formula One, however, the car itself has more than 250 sensors attached when competing in both qualifying and championship races. In a recent video created by the Mercedes-AMG Petronas team on their YouTube channel [59], they discuss their typical data generation output over a race weekend. Up to 1TB (terabyte) of data is recorded per car and they estimate 30MB of live data is generated per lap. The sensors also vary in function to comply with regulatory requirements for a tyre pressure monitoring system, along with providing real-time data on temperature, speed, tyre degradation and pressure.

An article published by Alex Woodie [60] also features the head of IT with the Mercedes team who highlights the volume of data generated through these sensors over a given season. With over 10TB of data being collected during this time, the data is stored using cloud storage array solutions and poses an issue with how many eyes can sift through this data. Instead, the team has adopted an approach to look at the anomalies within this data to gain greater insight and drive their decision making throughout the season.

To continue the focus on the collective contribution, Bell, Smith, Sabel and Jones [61] delve into the efforts of both driver and team and reach the conclusion that the team effect greatly outweighs the drivers effect (86% contribution) and quantifies the increase in driver changes, where as team effects have a ‘legacy effect’ and are steady based on accumulated experience.

Eichenberger, Stadelmann [62] published a study on the best Formula One driver over a 57 year period, from 1950 to 2006. The authors took the view that driver performance is dependent on the capability of their car and their innate talent. Through the application of linear regression techniques, variables such as collisions, engine failures and disqualification were removed from the study. Two control variables were also introduced to provide balance and robustness, that being the team partner driver and home advantage for a driver when racing in their own country.

A comparative analysis between both Formula One and NASCAR has been conducted by Silva, Silva [63] to outline both the similarities and differences between both disciplines. Qualifying and race data from the 2009 season is explored to achieve this. One clear difference between the sports is the fact that NASCAR has double the number of races and drivers in a season. The similarities include the point scoring system for driver and constructor championships, as well as the qualifying laps scheduled before race day to determine each driver's starting position on the grid. Four key variables were defined to gain insight into race predictions; Qualifying, Practice, Points and Results. The research confirmed that the Qualifying variable, which held the results of each driver’s qualification time, had the greatest significance in predicting Formula One results. To compare this with NASCAR, a combination of the Practice and Result data formed the strongest prediction for NASCAR. One key difference to note here is the rules around rain occurring during a NASCAR race, where the race event is cancelled.

Graves, Reese, Fitzgerald [64] used a Bayesian hierarchical framework to analyse the NASCAR racing results. They assess the driver’s abilities and predict their future standings based on the rate at which their skills have improved throughout a given season. The study also provides insight into the existence of track specialists - some drivers perform much better on certain tracks than others.

When looking at factors that can influence the outcome of a Formula One race, the study performed by Biemann, Liu, Zeng and Huang [65] used reinforcement learning to identify recurrent patterns and LSTM (long short-term memory) architecture to gain achievements in the model outcome. This study offered an alternative approach to addressing the weather calculations for this study which was not approached in the other literature.

Choo, C.L.W. [66] published a research paper which compared the insights gained from the work of Tulabandhula, Rudin [67] which focused on pit stop strategy and tyre changes based on a given track within the season. The suggestion of adding more features surrounding pit stops and pit crew performance for further insight was a novel suggestion here.

Aversa, Cabantous, Haefliger [68] collated data from the infamous final race of the 2010 Formula One season, where Ferrari’s decision support system provided incorrect conclusions and in turn, influenced their race strategy to deny Fernando Alonso a world title as expected. The Ferrari team chose to align their pit stop strategy with their rivals at the time, Red Bull, and pitted the car shortly after their competitor. This non-data driven strategy forced Alonso to re-join the circuit in the midst of a condensed field and hinder his ability to progress past other cars and ultimately, lose the expected championship title.

Looking at the relationship between Formula One and technology and referencing the research question of the potential benefits machine learning algorithms can have on the decision making in the sport, Jenkins [69] investigates the relationship between a team’s competitive advantage and their ability to adapt to changes in technological knowledge or discontinuities. The study finds that competitive advantage is achieved through a team's capability to pivot through successive technological discontinuities. The two capabilities are defined as dynamic and sustaining capabilities.

Aversa, Furnari, & Haefliger [70] link to this study when analysing the relationship between the business model and racing performance. The business models in Formula One are broken down into four; Internal Knowledge Transfer, External Knowledge Transfer, Formula One Supply and Talent. The four models evaluate the collaboration between Formula One teams and manufacturers, the sales of technology externally to other industries, the transferability between Formula One teams regarding supply and the scouting systems set for investing in both existing and future talent.

The authors focus specifically on the Red Bull and Williams racing teams and place the spotlight on their key factors such as financial, technological development and knowledge and human resources. The conclusion offers positive sentiment to the Red Bull strategy to invest in younger drivers and allocate costs to technological advancements instead.

Rosso, Rosso [71] examines the relationship between various race factors such as the weather and tyre types from the data distributed after the 2016 Monaco Grand Prix. They applied quantile regression [72] to study these factors in greater detail and the role the weather conditions had to play on the outcome of this specific race.

Sundar [73] provides a great insight into the 2005 Monaco Grand Prix.Kimi Raikkonen, a Finnish driver. was the fastest driver on Saturday’s qualifying and started the race from first place. He was competing with the title contender Fernando Alonso at the time in second place. It was on lap 23 when Minardi’s driver Christijan Albers crashed into a wall and the Safety Car was deployed. It is an unwritten rule in F1 that when the Safety Car is on track, the team strategy should always be to pit their cars. This is due to the fact that cars have to slow down once a safety car has been issued and follow this car around the track at a reduced pace, thus, reducing the impact of time in the pit. Alonso pitted immediately as it was expected, but surprisingly Raikkonen stayed on the track. This decision was made by McLaren chief strategist Neil Martin, after assessing all the data available. This decision to stay out on track was made within a minute of receiving the notification that a safety car had been deployed. Through the use of data analytics in this scenario, the team chose to take a risky decision and go against the grain of pitting the car. In the end, the decision bore fruit as Raikkonen gained a 35 second lead on Alonso and could afford to pit later in the race for the advantage of new tyres and still keep a 13 second lead when doing so. It proved an excellent decision post-race and is a prime example of data-driven decision making within sports.

Finally, a recent study conducted by Amazon Web Services [74] provided an insight into their partnership with Formula One and how the use of Machine Learning techniques and high performance computing are transforming the sport. It highlights the millions of data points being generated per second from the 300 sensors attached to a Formula One car and how the sport is becoming data-driven. The study also provides background analysis to factors which have been identified for this research such as tyre and driver performance. However, the machine learning algorithms used have not been disclosed due to their partnership with the sport and do not give guidance to the setup, training and evaluation of these models.

The literature mentioned above emphasises the importance of feature generation and having domain knowledge in order to successfully apply this. The use of Neural Networks were applied in most of these studies and has increased in popularity for a prediction model similar to the one being proposed here. Regarding the literature available for Formula One, there are few studies published but the NASCAR motorsport studies provide some insight into the algorithms that could be used. The NASCAR studies do not provide the necessary parallels in terms of whether a driver should make a pit stop since a single tire compound is used and this would lead to a new approach being taken for the application of Formula One racing.