AN ANN MODEL FOR PREDICTION OF GOAL SCORES BY INDIVIDUAL TEAM

E. Jones
Department of Computer Science
Modibbo Adamawa University
(MAUTECH) Yola, Nigeria.

B. Y. Baha
Department of Information Technology
Modibbo Adamawa University
(MAUTECH) Yola, Nigeria.

O. Sarjiyus
Department of Computer Science
Adamawa State University
Mubi Nigeria

Abstract- **Football is the most popular sport in the world, played in most countries in the world. The sport has a very large betting industry. The current estimations, which include both the illegal markets and the legal markets, suggest the sports match-betting industry is multi-billion dollar industry. This means the ability to accurately predict outcomes of games can be very lucrative. Although various research works have employed statistical and mathematical models to predict match results, none of the models are able to predict the exact number of goals to be scored by each team in a football match. In this research, an Artificial Neural Network (ANN) model was developed following the Machine Learning Life Cycle. The dataset was obtained from Football-Data.co.uk, a reputable football data website. Backward Elimination technique was used to select the dataset features for the proposed neural network model. The features selected include: h_a= Home or Away, xG= expected goals, xGA = expected goals away team, npxG = expected goals without penalties, npxGA = expected goals without penalties away team, deep = Number of plays in opponent final third, scored = goals scored, missed = goals conceded, result = win, draw or lose date = Date of the match, wins = binary for wins, draws = binary for draws, loses = binary for loses, teamId = team name, matchtime = time of match, tot_goal = total goals team has scored so far, tot_con = total number goals conceded by the team so far, Referee.x = refree name, HtrgPerc = shot on target/total shots – Home, AtrgPerc = shot on target/total shots – Away, matchDay = day of match. The dataset was then split into training and testing set at 75% and 25% respectively. The neural network developed is a multi-layer perceptron classifier implemented by the MLPClassifier class in sklearn. The model was compiled with different parameters to find the model with the highest accuracy relative to the mean squared error. The graph for the accuracy score and mean squared error was plotted and it showed the mean squared error was relatively the same for all the models. The model with the highest accuracy score was selected. The selected model has three (3) hidden layers that consist of 10,10, and 10 neurons with sigmoid optimizer and tanh activation function.. The model ran 1000 epochs and got an accuracy score of 97.92% with MSE of 2.8644, implying that real life games with unknown results can indeed be predicted with a high level of accuracy using machine learning.**

Keywords- **Artificial Neural Network, Dataset, Features, Football, Model.**

I. INTRODUCTION

Football is a uniting sport that has gained global acceptance. This has led to an entire industry built around the game, part of which is predicting match outcomes. The current estimations, which include both the illegal markets and the legal markets, suggest the sports match-betting industry is worth anywhere between $700bn and $1tn [1].

Machine Learning has been applied in various manners and in various fields for predictive purposes. It has been applied in weather prediction, stock market predictions, natural disaster predictions, and even health predictions. Another field that has seen a growing application of machine learning is sports. [2] state that machine learning has been applied in sports predictions ranging from basketball, rugby, cricket, swimming, to horse racing. However, it has mostly been used to predict win/lose outcomes, in other words, predicting who will win in a sports competition, but not exactly specifying the winning conditions (i.e. the specific score). [2] went on to state that sports result prediction is nowadays very popular among fans around the world, mostly due to the expansion of sports betting.

Sports pundits make predictions on the outcome of matches and whole leagues using information and experience gathered from the past. Predicting the results of sports matches is interesting to many, from fans to punters. It is also interesting as a research problem, in part due to its difficulty, because the result of a sports match is dependent on many factors, such as the morale of a team (or a player), skills, current score, etc. So even for sports experts, it is very hard to predict the exact results of sports matches [3].

Mathematical and statistical models work well in the prediction of match outcomes [4]. However, with advancements in computing and computing power, computers can be made to make these predictions using Artificial Neural Networks. This moves the burden of going through large amounts of raw data from the human expert to the machine (computer). [5] predicted the competitive

performance of an elite female swimmer (200-m backstroke) at the Olympic Games 2000 in Sydney using artificial Neural Networks. [6] applied decision tree induction for identifying characteristics in one-versus-one player interactions that drive the outcome in hockey contests. [7] used logistic regression and decision trees for explaining match outcomes in Australian Rules football. [8] presented a mixture of modelers approach to forecast the 2014 NCAA men's basketball tournament.

The Artificial Neural Network approach has been applied in predicting football match results by different researchers for different leagues. For example, [4] used Artificial Neural Networks to predict the outcomes of one week of the Iran Pro League (IPL) 2013-2014 football matches, where the data obtained from the past matches in the seven last leagues were used to make better predictions for the future matches. [9] did a study on "Artificial Intelligence in Sports Prediction". In the study, they used a Multi-layer Perceptron to predict the outcome of sports games, among which was the English Premier League (EPL). [10] wrote a Thesis on 'Football Match Prediction using Deep Learning' which investigated the deep learning method Recurrent Neural Networks (RNNs) for predicting the outcomes of football matches and the test results showed that deep learning may be used for successfully predicting the outcomes of football matches.This research therefore focuses on the prediction of the precise number of goals to be scored in a single football match which, by implication, predicts the winner.

II. REVIEW OF RELATED LITERATURE

Machine learning is a major field in the field of computing that has become more popular as technology has advanced. According to [11] machine learning is a domain of artificial intelligence that strives for enabling machines carry out their jobs skillfully through the application of intelligent programs. It is a natural extension of the intersection of the fields of Statistics and Computer Science. Machine learning uses a variety of algorithms that iteratively learn from data to improve, describe data, and predict outcomes. As training data are absorbed by the algorithms, it is then possible to produce more accurate models based on that data. A machine learning model is he output generated when you train your machine learning algorithm with data. After the model is trained, when you provide the model with an input, it gives you an output. According to [12], a machine learning algorithm that is predictive will create a predictive model. They further stated that when the predictive model is provided with data, it makes a prediction based on the data that trained the model.

[9] did a study on "Artificial Intelligence in Sports Prediction" which was a study on using a Multi-layer Perceptron to predict the outcome of sports games. The researchers obtained data from various sources which covers four major league sports: the Australian Football League (AFL), Super Rugby (Super 12 and Super 14), the Australian National Rugby League (NRL), and English Premier League football (EPL) from way back as 2002. [9] pointed out that the data included noise such that there were details that influenced the contest outside of those captured in the dataset. They used a Multi-layer Perceptron (MLP) to model the features, and for the learning algorithm, they chose back-propagation, which was pointed out to be "slightly more effective than conjugate-gradient method. The MLP they used was three-layered with nineteen input units, one for each feature (or twenty, when Player Availability is included), ten hidden units and a single output unit. The output unit was normalized to be a value between zero and one inclusive. The output values for the two teams competing in each game were calculated and the team which had the highest output value (i.e., the highest confidence that the team would be victorious) was taken as the predicted winner. Their results showed that for the English Premier League (EPL) the best performance was at 58.9%, while the average and worst performace were at 54.6% and 51.8% respectively.

[13] conducted a research to predict football match result Prediction using neural networks and deep learning. In the research, the authors deduced better features from the results of the previous matches that the team played and took into consideration the current form of the team to predict the accurate result of the game. Each dataset had around 380 records with around 60 attributes. The researchers then used a Recurrent Neural Network (RNN) to develop their model. They used basicLSTM cell from the tensorflow library to performed experiments on the dataset. The model was run on a various set of hyper-parameters to find out the best model. The researchers found that the accuracy achieved by the prediction system that worked with the LSTM form of RNNs showed huge improvement with a test accuracy of 80.75%.

[13] did a study that applied neural network algorithm in predicting football match outcome based on player ability index. The researcher applied three algorithms that included convolutional neural network (ANN), random forest (RF) and support vector machine (SVM) to predict the result of a football match, and then compared the accuracy of algorithms. [13] found that the accuracy of these three methods is all between 54% and 58%. The researcher further compared the result to the accuracy of the predictions of the famous football analyst of BBC, Mark Lawrenson, which is only 52%. In addition, the researcher also compared the accuracy of the convolution neural network which is at 58% to the prediction accuracy of the authoritative football gambling organization Pinnacle Sports, which is only 55%. Showing that the convolution neural network is more accurate in the predictions.

The studies by [9, 12, 13] all discuss prediction models that predict the outcome of a football match as either win, lose, or draw. However, the studies do not explicitly specify the conditions for the victory or loss of one team over another. Thus, this research explored the use of artificial neural network to explicitly predict the number of goals to be scored in a football match. Thus, this paper goes a step further in showing how to predict in explicit terms, the number of goals to be scored by each team in a match using an artificial neural network.

III. METHODOLOGY

The preliminary stages of buidling artificial neural network model involves the gathering of the data that will make up the dataset. Data was obtained from Football-data.co.uk, a reputable football data website known to have accurate data on all matches in the big leagues such as the English Premier League (EPL), UEFA Champions League, and other leagues.

The initial dataset consisted of 43 features but was reduced to 22 features after the Backward Elimination technique was applied to select the relevant features. The resulting dataset consisted of 22 features with 576 datapoints as shown in Table 1. The dataset included features that included: tot_goal (total goals team has scored so far), tot_con (total goals team has conceded so far),

HtrgPerc (shot on target/total shots – Home), AtrgPerc (shot on target/total shots – Away), xG (expected goals), xGA (expected goals away team) etc.

Table 1: Sample dataset showing the selected features

	h_a	xG	xGA	npxG	npxGA	deep	scored	missed	result	date	wins	draws	loses	npxGD
1	h	2.23456	0.842407	2.23456	0.842407	11	4	1	w	2019-08-09 20:00:00	1	0	0	1.392153
2	a	0.842407	2.23456	0.842407	2.23456	5	1	4	l	2019-08-09 20:00:00	0	0	1	-1.392153
3	a	3.18377	1.2003	2.42264	1.2003	9	5	0	w	2019-08-10 12:30:00	1	0	0	1.22234
4	h	1.2003	3.18377	1.2003	2.42264	1	0	5	l	2019-08-10 12:30:00	0	0	1	-1.22234
5	h	1.34099	1.59864	1.34099	1.59864	4	1	1	d	2019-08-10 15:00:00	0	1	0	-0.25765
6	a	0.855516	0.670022	0.855516	0.670022	5	3	0	w	2019-08-10 15:00:00	1	0	0	0.185494
7	h	0.909241	1.08752	0.909241	1.08752	0	3	0	w	2019-08-10 15:00:00	1	0	0	-0.178279
8	h	0.87159	1.2246	0.87159	1.2246	5	0	0	d	2019-08-10 15:00:00	0	1	0	-0.35301
9	a	1.2246	0.87159	1.2246	0.87159	5	0	0	d	2019-08-10 15:00:00	0	1	0	0.35301
10	a	1.59864	1.34099	1.59864	1.34099	6	1	1	d	2019-08-10 15:00:00	0	1	0	0.25765

Table 2: Sample dataset showing the selected features

teamId	matchtime	tot_goal	tot_con	Referee.x	HtrgPerc	AtrgPerc	matchDay
Liverpool	20	4	1	M Oliver	0.466666666666667	0.416666666666667	Fri
Norwich	20	1	4	M Oliver	0.466666666666667	0.416666666666667	Fri
Man City	12	5	0	M Dean	0.6	0.642857142857143	Sat
West Ham	12	0	5	M Dean	0.6	0.642857142857143	Sat
Bournemouth	15	1	1	K Friend	0.230769230769231	0.375	Sat
Brighton	15	3	0	C Pawson	0.272727272727273	0.6	Sat
Burnley	15	3	0	G Scott	0.4	0.272727272727273	Sat
Crystal Palace	15	0	0	J Moss	0.333333333333333	0.3	Sat
Everton	15	0	0	J Moss	0.333333333333333	0.3	Sat
Sheffield United	15	1	1	K Friend	0.230769230769231	0.375	Sat

The dataset was normalised usnig the *MinMaxScaler* function in the SciKit Learn library. For training and testing the model, the dataset was split into training and testing datasets at 75% and 25% respectively.

Machine Learning libraries that include pandas, numpy, seaborn, matplotlib, Scikit Learn, and XGBoost were used for the data preprocessing and normalization, data visualization, model training and testing, and performance evaluation.

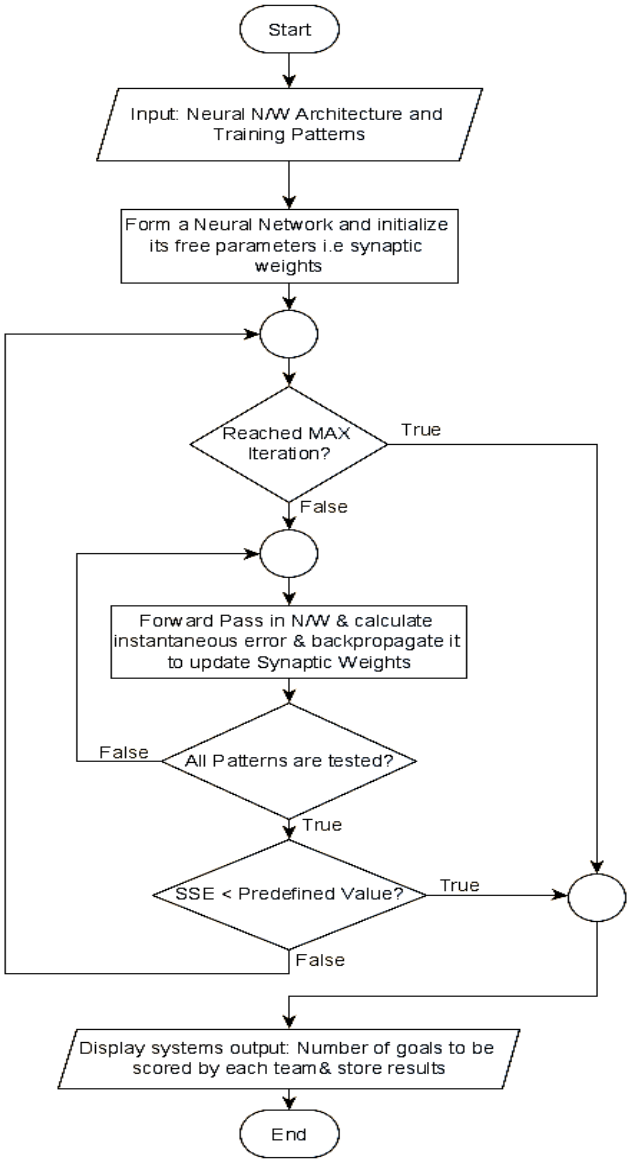


Fig 1: Flowchart of the proposed system.

IV. RESULTS

To test the model, it was compiled with varying parameters as shown in Table 2. The hidden layer was altered between 1,2,3, and 4 layers. The activation function for the hidden layer was alternated between relu and tanh. The optimizer used for the different runs are the Adam and sigmoid optimizers. The number of iterations (epochs) ranged between 100 and 1000, with the learning rate at either constant or adaptive. The model was fed in batches of 10, 20, 30, and 50. The model with the highest accuracy score of 97.92% consisted of 3 hidden layers of 10, 10, and 10 neurons each, compiled with the tanh activation function at 1000 epochs with a constant learning rate.

Figure 2 plots out the accuracy of each model against its Mean Squared Error (MSE). The graph shows a relative consistency in the MSE for all models tested, ranging from 2.7418 to 2.868. The accuracy of all the models ranged from 59.03% to 97.92%. This means the error margin for all the models tested remained relatively the same even as the accuracy varied more widely. Taking these factors into account, the model with the highest level of accuracy of 97.92% is selected.

Table 3: ANN Trials with Different Parameters

S/N	Hidden Layer	Number of Neurons	Hidden Layer Activation	Optimizer	Accuracy %	Epochs	Learning Rate	Batch	f'Mean Square Error
1	1	5	relu	Adam	84.03	100	constant	10	2.7979
2	2	5 5	relu	Adam	93.75	100	adaptive	10	2.8186
3	2	20 10	relu	Adam	93.75	1000	adaptive	20	2.868

4	2	40 20	tanh	Adam	90.28	1000	constant	30	2.8671
5	**3**	**10 10 10**	**tanh**	**sigmoid**	**97.92**	**1000**	**constant**	**30**	**2.8644**
6	3	40 40 40	tanh	sigmoid	96.53	1000	adaptive	50	2.8653
7	3	50 30 30	relu	sigmoid	59.03	100	adaptive	50	2.7418
8	3	50 30 30	relu	sigmoid	96.53	1000	adaptive	50	2.7418
9	4	40 40 30 20	tanh	sigmoid	85.42	100	constant	50	2.8232

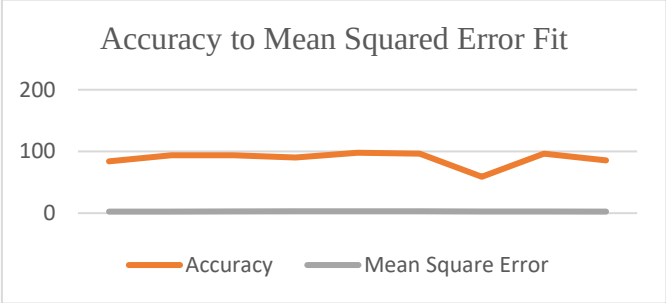
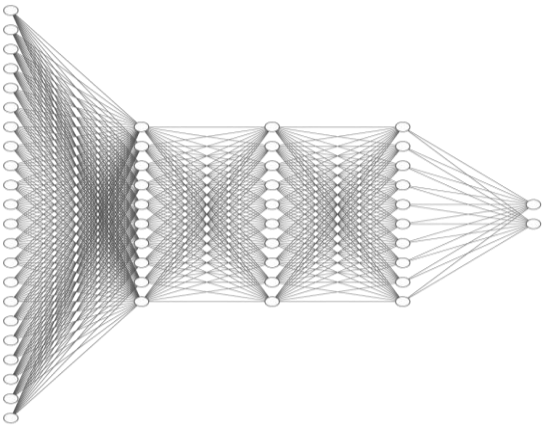


Fig 2: Accuracy to Mean Squared Error Fit.

V. CONCLUSION

The selected neural network is a Multi-layer Perceptron classifier implemented by the MLPClassifier class in sklearn. It is an estimator available as a part of the neural_network module of sklearn for performing classification tasks using a multi-layer perceptron. This model optimizes the log-loss function using LBFGS or stochastic gradient descent. The model consists of three hidden layers made up of 10, 10, 10 neurons. The activation function for the model is the tanh activation function. For the optimizer, the sigmoid function was used. In this model, weights were adjusted automatically depending on the number of neurons fed into the model and the activation function used after each iteration.



Input Layer ϵ R^{22} Input Layer ϵ R^{10} Input Layer ϵ R^{10} Input Layer ϵ R^{10} Input Layer ϵ R^{2}

Fig 3: 22-10-10-10-2 ANN Model Structure.

The chosen model performed very well with an accuracy score of 97.92% and a mean squared error of 2.8642. The model was timed at 9.1 seconds at 1000 epochs. This shows that artificial neural networks can be applied to predict the exact number of goals to be scored in a football match with a high degree of accuracy.

REFERENCES:

1. Keogh, F., & Rose, G. *Football betting - the global gambling industry worth billions 2013*. Retrieved November 3, 2021, from BBC: https://www.bbc.com/sport/football/24354124
2. Stekler, H. O., Sendor, D., & Verlander, R. Issues in sports forecasting. *International Journal of Forecasting, 26*(3), 606-621 2010.
3. Arabzad, S. M., Araghi, M. T., Sadi-Nezhad, S., & Ghofrani, N. Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. *International Journal of Applied Research on Industrial Engineering*, 1(3), 159–179 2014.

4. Schumaker, R. P., Solieman, O. K., & Chen, H. *Sports Data Mining.* New York 2010: Springer.
5. Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. Modeling and Prediction of Competitive Performance in Swimming Upon Neural Networks. *European Journal of Sport Science, 2*(2), 77-89 2002.
6. Morgan, S., Williams, M. D., & Barnes, C. Applying decision tree induction for identification of important attributes in one-versus-one player interactions: A hockey exemplar. *Journal of Sports Sciences, 31*(10), 1031-1037 2013.
7. Robertson, S., Back, N., & Bartlett, J. D. Explaining match outcome in elite Australian Rules football using team performance indicators. *Journal of Sports Sciences*, 34(1), 1-8 2015.
8. Yuan, L. H., Liu, A., Yeh, A., Kaufman, A., Reece, A., Bull, P., & Bornn, L. A mixture-of-modelers approach to forecasting NCAA tournament outcomes. *Journal of Quantitative Analysis in Sports, 11*(1), 13-27 2015.
9. Mccabes, A., & Trevathen. Artificial intelligence in sport prediction, fifth international conference on information technology: New generation (itng 2008), pp. 1194 – 1197, doi: 10.1109/ITNG.2008.2003.
10. Petterson, D., & Nyquist, R. *Football Match Prediction using Deep Learning* [Master's thesis, Chalmers University Of Technology, Gothenburg] 2017. Chalmers Publication Library. https://publications.lib.chalmers.se/records/fulltext/250411/250411.pdf
11. Mohammed, M., Khan, M. B., Bashier, E., & Bashier, M. *Machine Learning: Algorithms and Applications.* Boca Raton: CRC Press 2017.
12. Hurwitz, J., & Kirsch, D. *Machine Learning For Dummies.* New Jersey 2018: John Wiley & Sons, Inc.
13. Tiwari, E., Sardar, P., & Jain, S. Football Match Result Prediction Using Neural Networks and Deep Learning. *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO).* Chen, H. Z. (2019). Neural Network Algorithm in Predicting Football Match Outcome Based on Player Ability Index. *Advances in Physical Education*, 9, 215-222 2020. Noida