PSL Eye: Predicting the Winning Team in Pakistan Super League (PSL) Matches

Muhammad Humza Siddiqui[1] Muhammad Riaz[2] Muhammad Nasir[3]

Uzma Afzal[4] Sohaib Tariq[5] Tariq Mahmood[6]

Abstract

Pakistan Super League (PSL) is a well-known T20 cricket league with millions of viewers. With this large viewer base, predicting the outcome of PSL matches opens a new research avenue for academic researchers. In this paper, we collect PSL data from relevant sources and generate a validated data set for machine learning experiments. We implement the "PSL Eye" solution which employs Neural Networks (NNs) to predict the match winning team. We preprocess the dataset to eliminate the extra variables then we tune the hyper parameters of NN. After acquiring the optimal values of hyper parameters, we run our NN based PSL Eye to obtain the final results. The overall accuracy of PSL-Eye with testing data set is 82% which is very promising and shows the importance of NN in predicting PSL match outcome.

Keywords: Pakistan Super League, T20, PSL, Prediction, Neural Networks, Tensorflow, Keras, Machine Learning

1 Introduction

Cricket is a bat and ball game played between two teams. At the international level, cricket is played in three different formats, i.e., one-day, T20 (Twenty-20) and test matches. T20 is the most recent and shortened form of cricket restricted to 20 overs. It was introduced by the England and Wales Cricket Board (ECB) in 2003. Several T20 leagues started after 2007 ICC World T20 tournament. Bangladesh Premier League, Big Bash League, Indian Premier League, Pakistan Super League, and Caribbean Premier League are well known and successful T20 leagues[2][15][16].

Pakistan's population are sport loving people and cricket is the most popular sport in the country. In 2009, the Srilankan cricket team was targeted by militants. This tragic incident closed the doors for international cricket in Pakistan. Pakistan Super League (PSL) is a major progress to bring cricket back to Pakistan. It is a successful effort in revival of international cricket and provides an opportunity to train young and talented players. PSL runs between February

[1]Federal Urdu University of Arts Science & Technology, Karachi, Pakistan l siddiqi.humza97@hotmail.com
[2]Federal Urdu University of Arts Science & Technology, Karachi, Pakistan l riazmalik822@gmail.com
[3]Federal Urdu University of Arts Science & Technology, Karachi, Pakistan l nasir03082409229@gmail.com
[4]Federal Urdu University of Arts, Science & Technology, Karachi, Pakistan l uzma.afzal@fuuast.edu.pk
[5]Federal Urdu University of Arts Science & Technology, Karachi, Pakistan l sohaibtariq004@gmail.com
[6]Institute of Business Administration, Karachi, Pakistan l tmahmood@iba.edu.pk

and March every year. Karachi Kings (KK), Islamabad United (ISLU), Lahore Qalandars (LQ), Peshawar Zalmi (PZ), Quetta Gladiators (QG) and Multan Sultan (MS) are the current teams playing in PSL.

It is important to mention that cricket is the second most popular sport of the world with 2.5 billion viewers. Typically, viewers are interested in predictions to see which team will eventually win the match. This high interest of viewers in predicting the outcome of the cricket makes it a potential research avenue for the data science researchers. Several research solutions have been proposed to predict the cricket related variables. In [1], [5], [13] authors present predictive models to predict the players' selection in Indian Premier League (IPL). In [20], a data visualization and prediction tool for IPL data is presented. This HBase tool helps management to select a right team during auction. In [21], machine learning models are trained to forecast the outcome of IPL matches. To the best of our knowledge there is no research for PSL related predictions.

We already discussed that Pakistani are passionate for cricket and PSL fans also want to encourage their team to win the contest with confidence. Keeping all these things in mind, the primary objective of our research is to facilitate the PSL fans, Pakistan Cricket Board (PCB), academicians, researchers and students with the predictions of PSL matches winners. PSL has millions of supporters so it would be an interesting problem to make use of statistics and machine learning to predict the outcome of PSL matches.

Forecasting future from the past is highly subjective and thus requires extraordinarily expert decision making [10]. Machine Learning (ML) [11] is one of the well-known fields with successful implementations to predict the different variables related to healthcare, software engineering, sports and education [3][6][9]. So, application of ML techniques to PSL data seems justified from this perspective also. From the modern-day ML literature [1], we discovered that the most robust and scalable ML algorithm to learn the complicated patterns related to PSL games, and make predictions about their outcomes is the Neural Network (NN). This selection of neural network is also validated when we analysed the data of Pakistan Super League. Moreover, an important module of the Pakistan Super League is team-analytics and our research can help choice makers throughout the PSL matches to evaluate the strength of a team towards another. The main contributions of our work are as follows:

- We propose the first neural network-based model (which we label as PSL-Eye) to forecast the PSL match winner team.
- We also generate a validated PSL dataset (verified from multiple sources) for the researchers who are interested in working on PSL data.

The rest of the paper is organized as follows. Section II presents the relevant background, Section III presents the data collection and preprocessing. Section IV discusses the PSL Eye and its results. We conclude this paper in Section V.

2 Background

In this section we present the relevant background pertinent to this study. First, we discuss the PSL then we explain neural network, tensorflow and keras.

A Pakistan Super League (PSL)

PSL is a T20 cricket league. It has become one of the top viewing cricketing leagues in the world. It was founded by PCB on 9 September 2015 at Lahore. Its first version was played in UAE. PSL was initiated with 5 teams, i.e., Karachi Kings (KK), Islamabad United (ISLU), Lahore Qalandars (LQ), Peshawar Zalmi (PZ) and Quetta Gladiators (QG). In the third edition which was held in 2018 one more team was introduced in the league named as Multan Sultan (MS). These franchises (teams) are handled and owned by the investors. Initially in 2015, the commercial rights of the league were sold for US$93 million and that for 10 years but according to the sources the market value of the PSL was up to US$300 million in 2017.

The league played in early 2nd and 3rd months of the year. The format of league is double round robin and the playoffs. The PSL is managed by the Pakistan Cricket Board (PCB) head office. Due to security reasons the initial season were completely held in UAE. But from the second season, some matches started to be played in Pakistan also. The main reason to play entire league in Pakistan is to bring the cricket back to home, to fill the empty stadiums and to promote the local talent.

PSL matches have not been the target of any tangible research activity to date. Considering the importance of this local brand and its impact on millions of viewers, our paper presents the first work in this direction.

B Neural Networks (NN)

Today humans are being replaced by computers in the working environment because they can do work more efficiently at much lower cost to businesses [14]. Moreover, computers can adapt and learn according to certain trends, NN helps to make this possible just like human brain. As shown in Fig. 1, neural network itself consists of many small units called neurons. These neurons are grouped into several layers. Unit of one layer interact with the units of the next layer through weighted connections which really adjust connections is a real valued number. A neuron takes the value of a connected neuron and multiplies it with their connections weight. The sum of all connected neuron set in the bias value is then put into an activation function.

NNs have the potential to study and model non-linear and complicated relationships, in real-life, many of the relationships between inputs and outputs are non-linear as well as complex. After getting to know from the initial inputs and their relationships, it can infer unseen relationships on unseen data as well, hence making the model generalize and predict on unseen data. Unlike many other prediction techniques, NN does not impose any restrictions on the input variables (e.g., how they should be distributed).

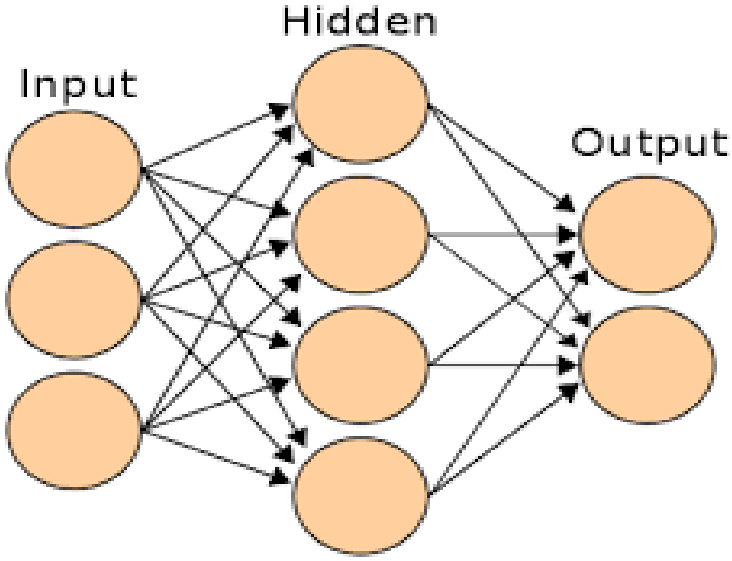


Figure 1: Architecture of Neural Network

C *Tensoflow and Keras*

Tensor flow is an open source library developed by the Google brain team. It's a versatile library but it was originally created for tasks that require heavy numerical computations. For this reason, tensorflow was geared towards the problem of machine learning and deep neural networks. Due to a C, C++ backend tensorflow was able to run faster than pure python code. Tensorflow offers several advantages for an application. It provides both a python and a C++ API. But the python API is more complete and it's generally easier to use. Tensorflow structure is based on the execution of a data flow graph. A data flow graph has two basic units a node represents a mathematical operation and an edge represents a multidimensional array known as a tensor. Tensorflow's flexible architecture allows you to deploy computation on one or more CPUs or GPUs or in a desktop server or even a mobile device. All of this can be done while only using a single API. Tensor flow has built in support for deep learning and neural networks so it's easy to assemble a net assign parameter and run the training process. It also has a collection of simple trainable mathematical functions that are useful for neural networks and any gradient based machine learning algorithm will benefit from tensor flows auto differentiation and sweet a first-rate optimizer. Tensorflow provides a lot of flexibility because it gives you control over the network structure and the functions used for processing [8].

Keras is an interface that allows us to easily access and customize the Machine learning frameworks, including Tensorflow, Microsoft cognitive tool kit CNTK and Theano. These frameworks also known as backends do all the heavy lifting when importing keras in Jupiter notebook.

Using Tensorflow back end an extremely popular choice for programmers. It's a great easy way to start implementing machine learning and specifically deep neural networks. If someone is interested in the basics of neural networks then Keras allows for quick experimentation with deep neural networks and focuses on being user friendly.

3 Data Collection and Preprocessing

In this section, we discuss data collection procedure and the preprocessing of the PSL data set. We also explain the encoding scheme we used for the categorical variables.

We collected the data of four seasons of PSL from different sources. Pakistan Cricket Board (PCB) official website [22], espncricinfo [23] and Cricingif [24]. ESPN, Cricingif provided us all data about match's venue, toss, weather and ball-by-ball record of batting and fielding side. First PSL season started in 2016. Our dataset contains 115 entries. The data is scrapped from the site and maintained in a Comma Separated Values (CSV).

We want to predict the winner teams of the next PSL matches, so we added all the details of the teams, specifically., team, opposition, home team, toss winner, batting performance (team), bowling performance (team), fielding performance (team), batting performance (opposition), bowling performance (opposition), fielding performance (opposition) , weather, pitch, team last matches performance, opposition last matches performance, team result, opposition result, team score, and opposition score to our dataset.

In season 1, there are some missing values in weather and pitch column. So, we used imputation method to fill the missing values of columns according to remaining values of the same columns. According to imputation mostly there is flat value in pitch column and cloudy in weather column in Dubai.

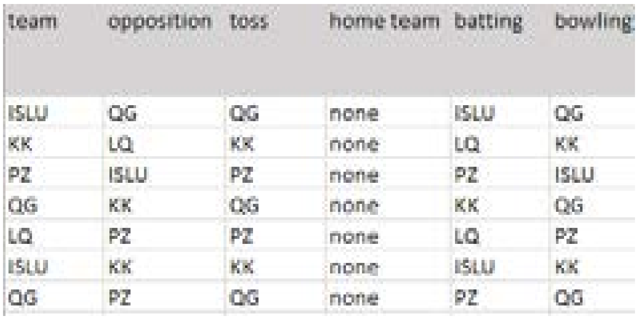
For better understanding and to make the dataset look some way or another jumbled free, abbreviation is used for every team name instead of their complete name. The abbreviations used in the dataset are the official ones. Figure 2 shows these abbreviations.

Team Name	**Abbreviation**
Peshawar Zalmi	PZ
Islamabad United	IU
Quetta Gladiator	QG
Karachi Kings	KK
Lahore Qalandar	LQ
Multan Sultan	MS

Figure 2: Abbreviation Chart

There are categorical variables in our dataset. Thus, at whatever point there is an absence of numeric value we convert the categorical variables to numeric values by encoding. Self-encoding technique applied to the categorical data values to convert in numeric values. Columns which have categorical values (shown in fig. 3) like team names, pitch, weather, venue are encoded with the numeric values.

The initial dataset had many features. Trying to feed all these features into the model does not make sense. We need only those features which are significant and play a role in our predicting variable, i.e., match winner. Some variables are divided into multiple columns such as teams batting, bowling, and fielding performance. Weights are assigned to categories according to the correlation between those columns and our dependent variables (Categories: Platinum=0.8, Diamond=0.7, Gold=0.6, Silver=0.5, Emerging=0.4, Supplementary=0.5).



team	opposition	toss	home team	batting	bowling
ISLU	QG	QG	none	ISLU	QG
KK	LQ	KK	none	LQ	KK
PZ	ISLU	PZ	none	PZ	ISLU
QG	KK	QG	none	KK	QG
LQ	PZ	PZ	none	LQ	PZ
ISLU	KK	KK	none	ISLU	KK
QG	PZ	QG	none	PZ	QG

Figure 3: Image from Dataset of categorical variables

For calculating each team's performance points, we formulated several equations to get our variables in single format rather than multiple column with multiple values (as shown in Fig. 4). We took all variables and distributed them in all three departments of game (batting, bowling, fielding) to make one equation for each department. The equations are as follows:

$$Bt.P = (c*w/d) \tag{1}$$

$$Bl.P = (c*w/d) \tag{2}$$

$$F.P = (ct.t*w/t.wk.t)+(d.c*w/r.wk)+(r.o*w/t.wk.t) \tag{3}$$

Where,

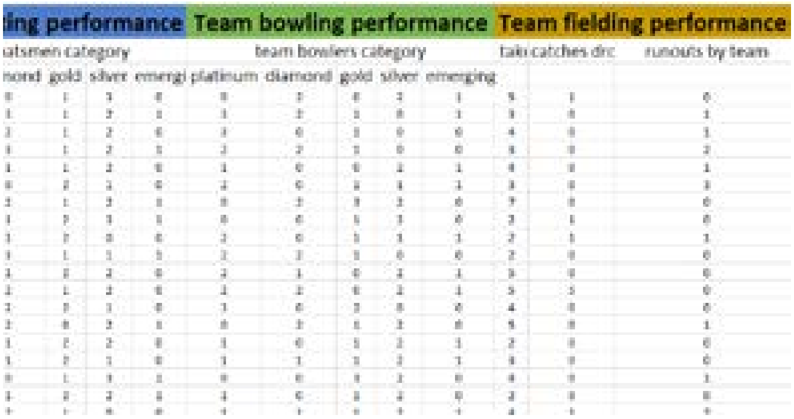
Bt.P = batting performance
Bl.P = bowling performance
F.P = fielding performance
c = category (of players)
w = weight
d = depth (no of regular batsmen or no of regular bowlers in teams)
ct.t = catches taken
t.wk = total wicket(s) (lost)
t.wk.t = total wickets taken
d.c = drop catches
r.wk = remaining wicket(s)
r.o = runout

Depth: depth here is the total number of players of particular department played in a particular match. Sum of total number of players in all categories of a department in a match is the depth of that column.

Category: Category in our dataset represents the class of player that which class does he belongs. PSL has six categories of players, i.e., platinum, diamond, gold, silver, emerging and supplementary. We mentioned the total number of players from a particular category in our data set. It shows that how many players from all categories are representing a particular team in a particular match in all three departments which are batting, balling and fielding.

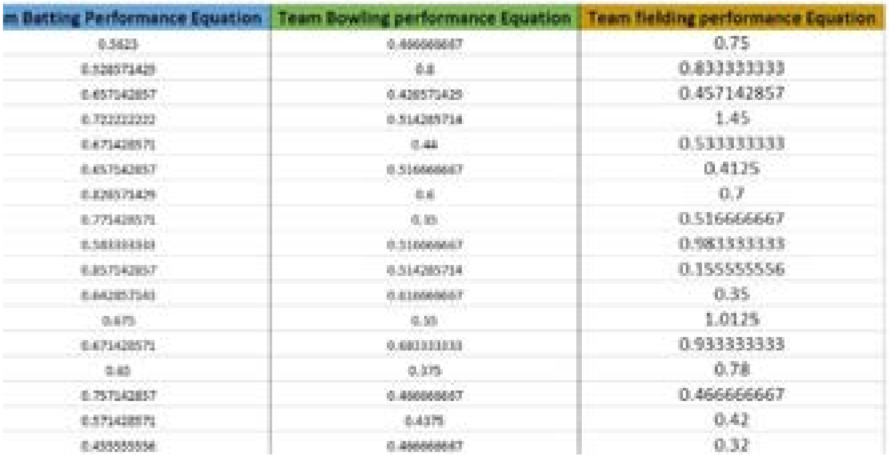
Weight: weight here is a value given to independent variable according to the importance of that independent variable towards the dependent variable. Higher the importance higher the value. It depends on how much that independent variable going to effect on the dependent variable. It determines the weightage of the independent variable in the equation.

Figure 5 presents an image from the preprocessed and validated data set. This dataset is available at [7] and can be shared with other researchers.



ing performance				Team bowling performance					Team fielding performance		
atsmen category					team bowlers category				taki catches drc		runouts by team
nond	gold	silver	emergi	platinum	diamond	gold	silver	emerging			
0	1	1	0	0	2	0	2	1	5	1	0
1	1	2	1	1	2	1	0	1	3	0	1
2	1	2	0	1	0	1	0	0	4	0	1
1	1	2	1	2	2	1	0	0	5	0	2
1	1	2	0	1	0	0	2	1	4	0	1
0	2	1	0	2	0	1	1	1	3	0	3
2	1	2	1	0	2	3	2	0	7	0	0
1	2	1	1	0	0	1	1	0	3	1	0
1	2	0	0	2	0	1	1	1	2	1	1
1	1	1	1	2	2	1	0	0	2	0	0
1	2	2	0	2	1	0	2	1	3	0	0
2	1	2	0	2	2	0	2	1	5	1	0
2	2	1	0	1	0	2	0	0	4	0	0
2	0	2	1	0	2	1	2	0	5	0	1
1	2	2	0	1	0	1	2	1	2	0	0
1	2	1	0	1	1	1	2	1	3	0	0
0	1	1	1	0	0	1	2	0	4	0	1
1	2	2	1	1	0	1	2	0	2	0	0

Figure 4: Image before preprocessing



m Batting Performance Equation	Team Bowling performance Equation	Team fielding performance Equation
0.5623	0.466666667	0.75
0.528571429	0.8	0.833333333
0.657142857	0.428571429	0.457142857
0.722222222	0.514285714	1.45
0.671428571	0.44	0.533333333
0.657142857	0.516666667	0.4125
0.828571429	0.6	0.7
0.771428571	0.35	0.516666667
0.583333333	0.516666667	0.983333333
0.857142857	0.514285714	0.155555556
0.642857143	0.616666667	0.35
0.675	0.35	1.0125
0.671428571	0.683333333	0.933333333
0.65	0.375	0.78
0.757142857	0.466666667	0.466666667
0.571428571	0.4375	0.42
0.455555556	0.466666667	0.32

Figure 5: Final data set – An Image

We also plot the graph of the team and opposition performances columns to compare the strength. Fig. 6 represents the scatter plot of batting, bowling and all-rounder performances of team and opposition. On x-axis of graphs shows team's batting, bowling and fielding strength and opposition on y-axis of the graph. In every graph each point indicates the co-relation between two variables.

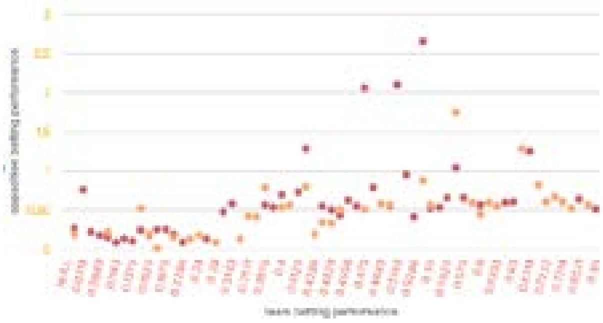


Figure 6A: Graph of batting performances

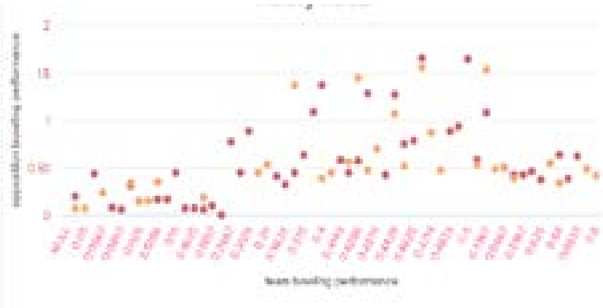


Figure 6B: Graph of bowling performances

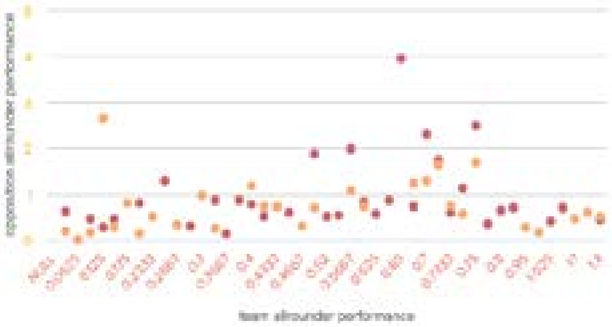


Figure 6C: Graph of all-rounder performances

Control chart is also used to figure out the yearly performance of the teams played in PSL from 1st season started in 2016 till the last season played in 2019 (Figure 7).

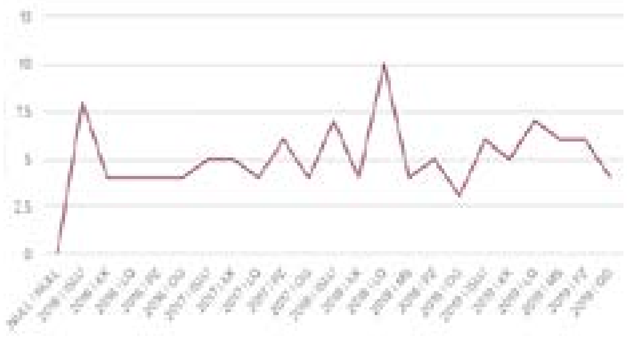


Figure 7: Yearly Team Performances

4 PSL EYE: Modeling And Results

In this section we discuss the modeling of PSL eye. Neural network algorithms apply mostly on numeric data and our dataset is primarily numeric in nature. In our problem domain, our model is used to distinguish match winner or loser, it is a binary classification problem. As already mentioned we have 115 records in our dataset.

Before the actual modeling, we applied weights (as per their significance) to predictors in order to achieve a successful and accurate NN implementation. We also grouped the sub variables (discussed in Section III) into single predictor column. For example, batting performance of the team is calculated by summing six columns (five categories and one batting depth). Weights are also used to simplify our network connections. Objective of the weight is to minimize the error. It is an input to neurons, and it is always 1.

In Batting and Bowling Performance, where category is player category column value, weight which we are assigned above, and depth is the number of batsmen or bowlers in playing eleven. In Batting and Bowling Performance, where category is player category column value, weight which we are assigned above, and depth is the number of batsmen or bowlers in playing eleven. After this calculation we have a modified dataset which is used to train NN for PSL match winner predictions.

We modeled the given match prediction problem to predict the wining possibility for the Pakistan super league teams, i.e., KK, LQ, QG, IU, PZ and MS. We modeled the neural network using Sequential Model which is simplest type of neural. The model is coded in python by using TensorFlow and Keras.

In our model, the name of parameters is set as: epochs as epochs, learning rate, batch size as batch size and output as results some values we put directly. We used sigmoid activation function because this is binary classification problem sigmoid scales the output on the scale of 0 to 1.

We have performed all these experiments on a windows 10 machine with Intel core i7-7th gen CPU, 8GB RAM. For optimal results, NN hyper parameters such as input neurons, hidden layers, output neurons, time steps, batch size, and optimizer are needed to be set. These parameters are important for generating results and good accuracy. We tuned these parameters for our dataset. For this, we simply used trial and error methodology which is a right way to fix these parameters.

Our NN based PSL Eye model generates optimal results with the following values of hyper parameters. 2 hidden layers are used to construct PSL NN. Batch size is set to 10, the batch size is the number of units deal before the model is learned. For activation function in hidden layers we used "relu" and, "sigmoid" for result layer because relu provides probability and sigmoid provide results in the binary form either 0 or 1. We also used Adam Optimizer. Number of epochs is set to 20. Table I presents the accuracies of the model with different setting of hyper parameters.

Table 1: Model Accuracy

No #		Hyper parameters and results	
	Epochs	batch size	Results
1	16	10	60.98%
2	17	10	75.61%
3	18	10	79.08%
4	19	10	82.93%
5	20	10	80.34%

After acquiring the optimal values for these hyper parameters, we re-run the NN with PSL data to obtain the final result. Our model generates the results of winning team with 79%, 80% and 82% accuracies on validation, training and testing data set respectively. Table II shows the details of the results using an image from the results file. In comparison with a research Table III presents a comparison of existing research work PSL-Eye.

Table 2: Result of PSL EYE

Validation Results		Training Results		Testing Results	
Actual	Predicted	Actual	Predicted	Actual	Predicted
Win	Loss	Loss	Loss	Loss	Win
Win	Win	Loss	Loss	Win	Win
Win	Win	Loss	Win	Win	Loss
Loss	Loss	Win	Win	Win	Win
Loss	Loss	Win	Win	Win	Win
79%	80%	82%			
			Overall Accuracies		

Table 3: PSL-EYE: A Comparison

Research	Problem	Solution	Accuracy (%)
PSL-Eye	PSL match winner	Neural Net	82
[21]	IPL match winner	Neural Net	72.6
[1]	Cricketer Performance	Neural Net	49-77

We also developed a web interface [25] of PSL Eye and uploaded these results to share with research community. It is a user-friendly interface with complete introduction and background of our work.

5 Conclusion and Future Work

A validated dataset for Pakistan Super League (PSL) is generated to run the Machine learning (ML) experiments. Experiments are run on the data set and, a Neural Network (NN) based solution (PSL Eye) is proposed to predict the winning team of the match. PSL eye model generates results with 82 % accuracy on testing data set. In future we will run experiments on PSL data with other ML algorithms to improve the accuracy of PSL Eye.

References

[1] Iyer SR, Sharda R. Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications. 2009 Apr 1;36(3):5510-22.

[2] Petersen C, Pyne DB, Portus MJ, Dawson B. Analysis of Twenty/20 cricket performance during the 2008 Indian Premier League. International Journal of Performance Analysis in Sport. 2008 Nov 10;8(3):63-9.

[3] Bilal M, Asif S, Yousuf S, Afzal U. 2018 Pakistan General Election: Understanding the Predictive Power of Social Media. In2018 12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS) 2018 Nov 24 (pp. 1-6).

[4] Kaluarachchi A, Aparna SV. CricAI: A classification based tool to predict the outcome in ODI cricket. In2010 Fifth International Conference on Information and Automation for Sustainability 2010 Dec 17 (pp. 250-255).

[5] Prakash CD, Patvardhan C, Singh S. A new category based deep performance index using machine learning for ranking IPL cricketers. Int. Jl. of Electronics, Electrical and Computational System IJEECS ISSN. 2016 Feb.

[6] Salman M, Qaisar S, Qamar AM. Classification and legality analysis of bowling action in the game of cricket. Data Mining and Knowledge Discovery. 2017 Nov 1;31(6):1706-34.

[7] PSL DATA REPOSITORY. [online] Available at: https://github.com/nasir03082409229/psl-eye.git [Accessed 30 Oct. 2019].

[8] Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc."; 2017 Mar 13.

[9] Nafees U, Parveen S, Zahid K, Afzal U. A Tensorflow Based Neural Network to Predict Drone Strikes in Pakistan. In2018 12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS) 2018 Nov 24 (pp. 1-6). IEEE.

[10] Sankaranarayanan VV, Sattar J, Lakshmanan LV. Auto-Play: A data mining approach to ODI Cricket simulation and prediction. InProceedings of the 2014 SIAM International Conference on Data Mining 2014 Apr 28 (pp. 1064-1072). Society for Industrial and Applied Mathematics.

[11] Kumar G. Machine learning for soccer analytics. KU Leuven. 2013. Kumar, Gunjan. "Machine learning for soccer analytics." KU Leuven (2013).

[12] Portus MR, Farrow D. Enhancing cricket batting skill: implications for biomechanics and skill acquisition research and practice. Sports Biomechanics. 2011 Nov 1;10(4):294-305.

[13] Saikia H, Bhattacharjee D, Lemmer HH. Predicting the performance of bowlers in IPL: an application of artificial neural network. International Journal of Performance Analysis in Sport. 2012 Apr 1;12(1):75-89.

[14] Battiti R, Villani A, Le Nhat T. Neural network models for intelligent networks: deriving the location from signal patterns. Proceedings of AINS. 2002 May 8.

[15] Kalgotra P, Sharda R, Chakraborty G. Predictive modeling in sports leagues: an application in Indian Premier League. Available at SSRN 2465300. 2013 Apr 28.

[16] Kansal P, Kumar P, Arya H, Methaila A. Player valuation in indian premier league auction using data mining technique. In2014 international conference on contemporary computing and informatics (IC3I) 2014 Nov 27 (pp. 197-203). IEEE.

[17] Passi K, Pandey N. Increased Prediction Accuracy in the Game of Cricket using Machine Learning. arXiv preprint arXiv:1804.04226. 2018 Apr 9.

[18] Jhanwar MG, Pudi V. Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. InMLSA@ PKDD/ECML 2016 Sep 19.

[19] Prakash CD, Patvardhan C, Lakshmi CV. Data Analytics based Deep Mayo Predictor for IPL-9. International Journal of Computer Applications. 2016 Oct;152(6):6-10.

[20] Singh S, Kaur P. IPL visualization and prediction using HBase. Procedia computer science. 2017 Jan 1;122:910-5.

[21] Lamsal R, Choudhary A. Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning. arXiv preprint arXiv:1809.09813. 2018 Sep.

[22] Pakistan Cricket Board official website [online] Available at: https://www.pcb.com.pk/ [Accessed 30 Oct. 2019].

[23] ESPN Cricinfo [online] Available at: http://www.espncricinfo.com/ [Accessed 30 Oct. 2019].

[24] CRICINGIF [online] Available at: https://www.cricingif.com/series/1278/pakistan-super-league-psl. [Accessed 30 Oct. 2019].

[25] PSL EYE: [online] Available at: .https://psl.hairnet.com.pk/ [Accessed 30 Oct. 2019]