A Data Mining Approach to Predict Match Outcome in One-Day International Cricket

Waqar Ahmed[1]*, Khurrum Nazir Junejo[1] , Tariq Mahmood[2] and Ghulam Mujtaba[3]

[1]Department of Computer Science, Karachi Institute of Economics and Technology, Karachi, Pakistan
[2]Department of Computer Science, Institute of Business Administration, Karachi, Pakistan
[3]Department of Computer Science, Sukkur Institute of Business Administration, Sukkur, Pakistan

Summary
Accurate prediction of the likelihood of a team to win a game of sports even before it starts is an insight worth millions to coaches, managers, sports analysts, and media persons. In this study, we propose an approach to predict match outcome for the game of Cricket; the second most watched sports in the world. We successfully predict the winner of One-Day International (ODI) cricket match 80% of the time even before the start of the game. To achieve this we consider all ODI matches played between 1971 and 2014 to form a data-set that comprises 11 attributes and 3534 match records. We put comprehensive effort in collection and pre-processing of the raw data to identifying the most decisive attributes for the prediction. We then make use of six well-known machine-learning approaches while experimenting with different intervals, sampling, and attribute selection techniques. Our approach achieves a gain of 25.00% in prediction accuracy with respect to baseline winning ratio of the team. We observe that data of recent matches has a strong influence on the prediction of match outcome. Using our prediction tool, Cricket managers can choose the most appropriate squad for forthcoming ODI match, whereas coach and captain can shape their strategies before the match starts. Furthermore, cricket analysts and media can also use the model for pre-match analysis.

Keywords:
One-Day International, cricket, outcome prediction, classification model, performance evaluation.

1. Introduction

Accurate prediction of the likelihood of a team to win a game of sports even before it starts is an insight worth millions to coaches, managers, sports analysts, and media persons. In this study, different approaches for a novel prediction problem i.e., predicting the outcome of One-Day International (ODI) cricket match is presented. The outcome of an ODI cricket match depends on number of factors related to scoring as well as the athletic strength of playing team members. The process of using past data to predict cricket match outcome has been explored earlier on. Our research question is: "Can the outcome of an ODI be more accurately predicted than current approaches?" Therefore, this study attempts to establish a consistent statistical approach that offers greater accuracy than previous attempts. Secondly, influence of recent matches and the potential of a range of variables that could define the outcome of an ODI cricket match was explored. Thirdly, the effect of different size of training and testing data sets on prediction accuracy using diverse classification models was investigated.

One of the earlier published works on cricket proposed a method of calculation to determine the optimal scoring rate which can be done at any stage of the innings, along with an estimate of the chance of winning in the ODI match[1]. Another study proposed modification of the Duckworth–Lewis resource table to quantify the magnitude of a victory in ODI matches[2] is particularly useful in breaking ties in tournament standings and in quantifying team strength. It is also investigated that Duckworth–Lewis method can be readily modified to predict ODI match outcome while game in progress[3]. The match outcome however cannot be predicted until the match starts, moreover prediction results change radically as match progresses. It has also been concluded that teams generally gain no winning advantage because of winning the toss[4]. Though a technique of learning model for predicting game progression and outcome in ODI has been proposed[5, 6] in cricket data mining literature, prediction accuracies presented are not convincing. Bayesian classifier has also been used to examine how different attributes affect the outcome of an ODI cricket match[7] but the accuracy they achieved is poor. Therefore, in this study, we gathered all factual considerations presented in literature at single point to come up with some more useful attributes and model learning techniques to further improve the prediction accuracy in ODI cricket.

2. Material and Methods

Records of historic ODI cricket matches of last four decades was extracted from an open source database *ESPN-cricinfo*[8]. In-depth statistics of all 22 ODI cricket-playing nations were considered in this study. Collected dataset contains 3534 match records played between January 1971 (when first ODI was played) and October

Manuscript received February 5, 2026
Manuscript revised February 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.2.23

2014 (dataset collected for analysis). Each available match record consists of several attributes, more importantly, ODI #, Date, Team1 and Team2, Country, Ground, Match Type (Day/Night), Batting (First/Second), Scores, Wickets, Runs per over and match result. Model learning and evaluation was carried out using open source predictive analytics platform Rapid Miner Studio[9].

2.1 Dataset

Eleven different datasets were established including a dataset for each of the 10 test-playing nations i.e. Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe. The 11th dataset is a merged dataset of all non-test playing nations i.e. Afghanistan, Bermuda, Canada, East Africa, Hong Kong, Ireland, Kenya, Namibia, Netherlands, Scotland, United Arab Emirates and United States of America titled as others. Merging non-test playing nations, which are not so mature in cricket, is logical since these teams have (as yet) below average performance against test-playing nations. Match records that result in a draw, had no result or abandoned are not included in the dataset.

2.2 Preprocessing

Comprehensive effort was put in collection of raw data and preprocessing for the range of variables that could define the outcome of an ODI cricket match. Several attributes were derived from available database while some of them devised using feature engineering techniques. Following subsections describe some devised attribute with their statistical significance in forecasting problems.

2.2.1 Home Advantage

The role of home advantage has been shown to play a vital role in any analysis of sporting events [10-12]. Winning the coin toss at the outset of a match provides no competitive advantage whereas the advantages of playing one's home field increase the probability of winning in ODI match[13].

2.2.2 Pitch Report

Regrettably, the pitch report of each match is not available in records. Therefore, the behavior of the pitches of all the venues was generalized because the presence of amount of grass and cracks on the pitch may vary on a given day but overall behavior remains same. The pitch behaviors were categorized into slow, bouncy, dry and green pitches. After careful study of pitches, a pitch type to all 157 international cricket grounds were assigned.

2.2.3 Weather Report

Weather can also play a vital role on the outcome of the ODI cricket match. Especially temperature, overcast and humidity. Weather data was collected of all related cities from Weather Underground[14] that spreads overs 6 continents. Unfortunately, weather data is available from 1996 only whereas matches are being played since 1971. Therefore, we devised an attribute *Season* based on World Season Calendar[15] that caters environmental conditions graciously. The structure of this attribute is described in Table 1.

2.2.4 Consecutive Wins Before Current Match

We also considered the sum of consecutive wins of a team before the current match starts (i.e., a defeat resets the counter else plus one for each win in a row). It is an integer type of attribute devised for the first time to predict the match outcome in ODI cricket.

2.3 Target Team

Even though a prediction model can be learned for any nation, in this study we limit our analysis to one test-playing nation. An overall record of ODI matches played by any team is shown in

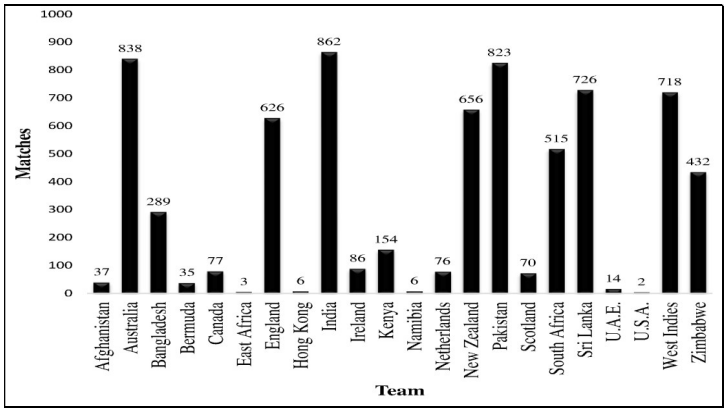


Figure 2

Figure 1. Total number of matches played by each team in One-Day International format

Obviously, a team who has played large number of matches could be handy in learning an effective prediction model. Statistics show that Australia, India and Pakistan have both played 800+ matches but we selected Pakistan as, according to our experience, it has shown outcomes that are more inconsistent and hence poses a bigger challenge for our prediction problem.

2.4 Attribute Selection

Identification of most decisive attributes to predict match outcome is the fundamental objective of this study. After preprocessing, the established dataset contains following attributes:

1. Country
2. Ground
3. Match Type (Day/Night)
4. Batting (First/Second)
5. Pitch Report
6. Home Ground
7. Consecutive Wins
8. Opponent
9. ODI#
10. Season
11. Date
12. Match Result (Classification Label)

A dataset may contain attributes that provide little power to classify instances even in some cases, these attributes negatively affect the classification accuracy. Therefore, in order to eliminate such attributes the brute-force search method (an attribute selection method that evaluates all possible combinations of the input features, and then finds the best subset) was used to identify the most decisive attributes to learn prediction model.

The dataset (all matches played by Pakistan) was fed to six well-known classifiers[16, 17] including **k-Nearest Neighbor**, Neural Network, Decision Tree (ID3), Random Forest, Logistic Regression and Naïve Bayes. It was found that five attributes give best performance with every classification model hence they are finally selected for further analysis. Those five attributes are listed below:

1. Consecutive Wins
2. Opponent
3. Ground
4. Home Ground
5. Match Type (Day/Night)

Table 1. The structure* of attribute Season based on World Season Calendar included in every match record; however, the value depends upon date and venue** of the ODI match

Interval	Seasons	
	Cricket Grounds in Northern Hemisphere	Cricket Grounds in Southern Hemisphere
21st Dec to 21st Mar	Winter	Summer
22nd Mar to 20 Jun	Spring	Autumn
21st Jun to 19th Sep	Summer	Winter
20th Sep to 20th Dec	Autumn	Spring

*Specific weather conditions, temperatures and length of the day define season in a year. In World Season Calendar, the year is divided into four seasons of 13 weeks each: spring, summer, autumn and winter.

**Seasons in the Southern Hemisphere are opposite to the seasons in the Northern Hemisphere.

2.5 Sampling Technique

It is a technique of dividing members of the population into homogeneous subgroups. We used it to split data into training, validation and testing sets as shown in Table 2. Such proportionate allocation uses a sampling fraction in each of the subpopulation that is proportional to that of the entire population. For instance, if the population S consists of m examples in the male subpopulation and f examples in the female subpopulation (where m + f = S), then the relative size of the two samples (x1 = m/S*(test-set size), x2 = f/S*(test-set size)) should reflect the proportion in test-set. This enables one to sample even the smallest and most inaccessible subgroups in the population that suits our case. The technique offers higher statistical precision compared to simple random sampling because the variability within the subgroups is lesser compared to the variations when dealing with the entire population.

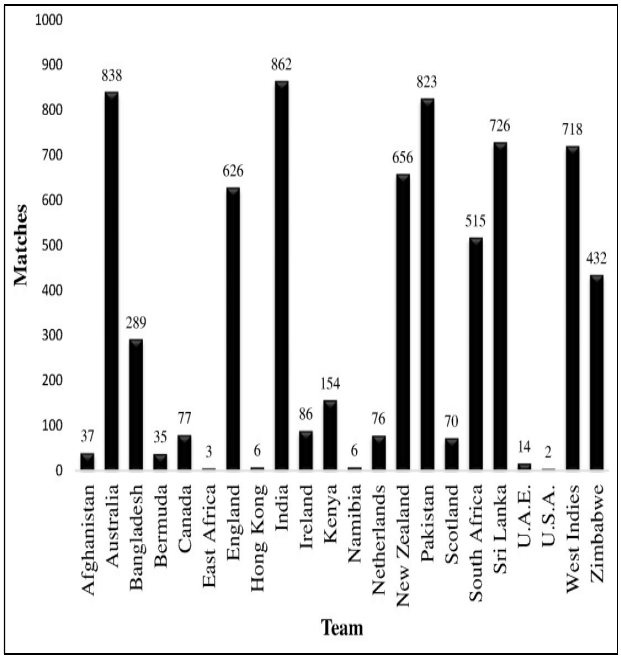


Figure 2. Total number of matches played by each team in One-Day International format

Table 2: Stratified sampling of dataset to establish Train-set, Validation-set and Test-set considering all match records* played between 1971 and 2024.

Opponent	Matches (vs. Pak)	Train-set (80%)	Validation-set (10%)	Test-set (10%)
Australia	92	74	9	9
Bangladesh	32	26	3	3
England	72	58	7	7
India	126	100	13	13
New Zealand	89	71	9	9
Others	26	20	3	3
South Africa	71	57	7	7
Sri Lanka	142	114	14	14
West Indies	126	100	13	13
Zimbabwe	47	37	5	5

*The sets were used in Setting_1 and similar ratios were used for all other five settings. However, No. of matches played against Pakistan depends upon the interval considered for a particular setting.

2.6 Experimental Setup

Despite the fact that data is available from January 1971, due to the continuous modification in cricket rules, earlier matches should be disregarded for better prediction results. Obviously, with the passage of time, teams become matured and develop strengths in all areas (batting, bowling and fielding) with experience. For instance, against any opponent, Pakistan could not capture most wins in early 165 matches i.e., Pakistan lost most of the matches before 1990. In Figure 2 an improved performance can be seen ahead in time by Pakistan against the same opponents.

Since Pakistan has now emerged as a highly matured and competitive team, in order to classify new matches, earlier instances should be excluded when Pakistan's ODI cricket team had totally opposite performance trend (as shown in Figure 2). However, the question is: how many matches should be excluded? As winning rate not only varies with opponent to opponent but also a team do has a psychological pressure/confidence of its prior performance associated with every opponent. Therefore, datasets with following intervals were investigated to determine optimal interval of dataset for final model.

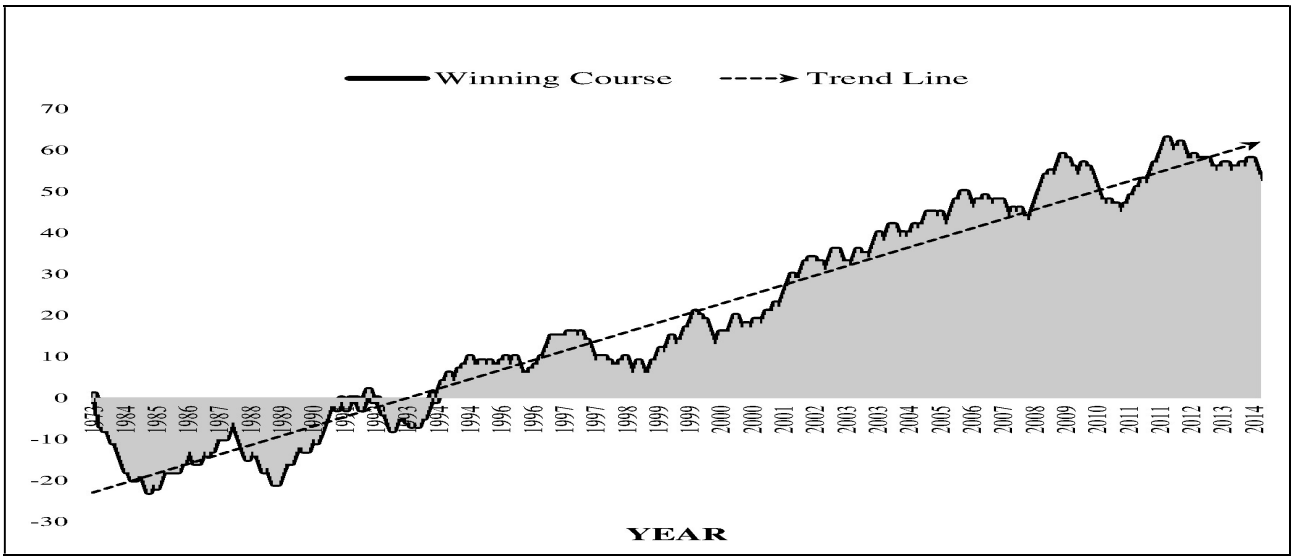


Figure 2: Winning course of Pakistan against all opponents. Rule defined to create the graph is simple, start with zero and +1 for each win else -1 for a loss (Rise in course shows successive wins and fall shows successive losses.). Y-axis in graph is for winning pattern and X-axis for the year in which an ODI match was played

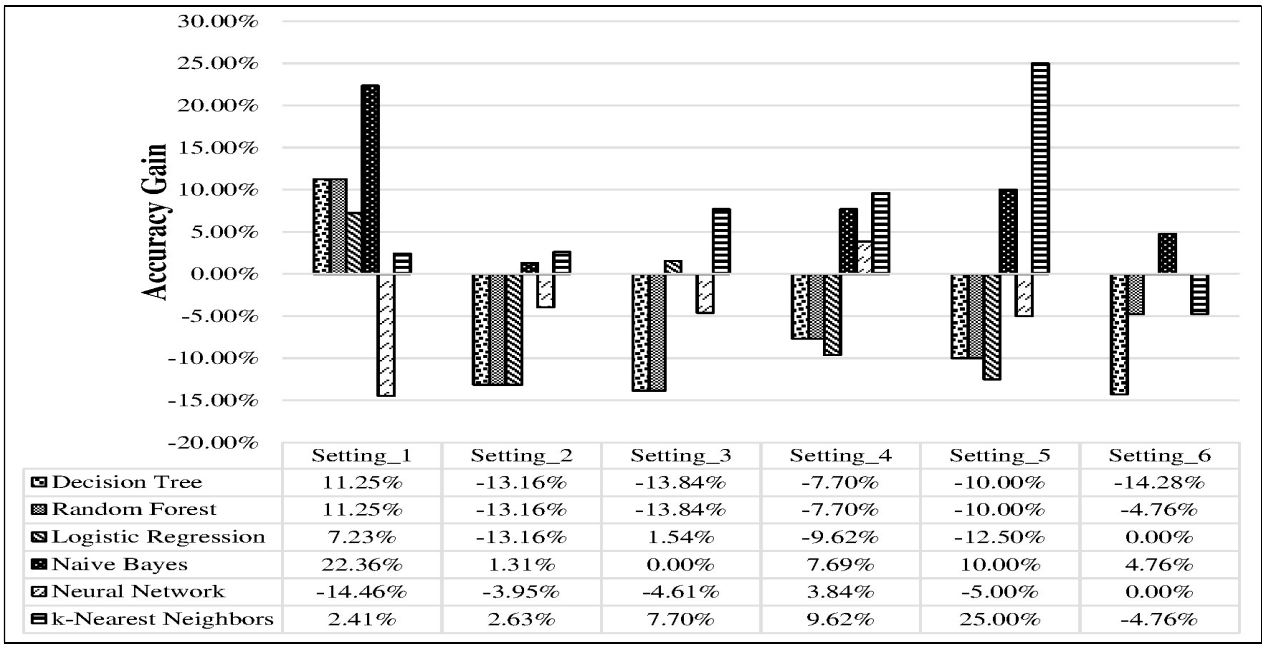
1. Dataset form 1971 to 2025
2. Dataset form 1985 to 2025
3. Dataset form 1990 to 2025
4. Dataset form 1995 to 2025
5. Dataset form 2000 to 2025
6. Dataset form 2005 to 2025

We call each interval consideration a setting. In each setting the dataset was divided into training, validation and test-set for every classification model. The 80% dataset was used to construct training-set in order to train the classification model. Whereas 10% dataset was used to construct validation set. Validation set, which is independent from the training-set, is used for model's parameter tuning/selection and to avoid over fitting. However, after parameter tuning, validation set was then merged with training-set to form a new training-set of 90% examples for performance evaluation. The rest 10% dataset, which is absolutely unseen to the prediction model (like future matches), was used in online learning setting to evaluate the performance of trained model. Online learning is a technique of model learning in which we include tested instance of test-set in to train-set to update our prediction model for future/remaining test instances. While evaluating classification model, performance was also compared with baseline to highlight the gain factor.

3. Results

The dataset used in each setting comprise of five attributes (mentioned in previous section) and number of records based on respective considered intervals. Parameters of each classifier was tuned using validation-set then kept to compute prediction accuracy using test-set. A comprehensive information about number of records used to train and evaluate the effectiveness of the prediction model with a particular setting are shown in Table 2.

In Setting_1, whole dataset from 1971 to 2025 was considered that comprises of 823 records. Figure 3 shows that Naïve Bayes outperforms all other classifiers with performance gain of 22.36% w.r.t baseline. In Setting_2, dataset from 1985 to 2025 comprising 749 records, k-Nearest Neighbor (k-NN) outperforms all other classifiers with performance gain of 2.63%. In Setting_3, dataset from 1990 to 2025 comprising 641 records, k-NN outperforms all other classifiers with performance gain of 7.70%. In Setting_4, dataset from 1995 to 2025 comprising 529 records, k-NN outperforms all other classifiers with performance gain of 9.62%. In Setting_5, dataset from 2000 to 2025 comprising 380 records, k-NN outperforms all other classifiers with performance gain of 25%. In Setting_6, dataset from 2005 to 2014 comprising 220 records, Naïve Bayes outperforms all other classifiers with performance gain of 4.76%.



	Setting_1	Setting_2	Setting_3	Setting_4	Setting_5	Setting_6
Decision Tree	11.25%	-13.16%	-13.84%	-7.70%	-10.00%	-14.28%
Random Forest	11.25%	-13.16%	-13.84%	-7.70%	-10.00%	-4.76%
Logistic Regression	7.23%	-13.16%	1.54%	-9.62%	-12.50%	0.00%
Naive Bayes	22.36%	1.31%	0.00%	7.69%	10.00%	4.76%
Neural Network	-14.46%	-3.95%	-4.61%	3.84%	-5.00%	0.00%
k-Nearest Neighbors	2.41%	2.63%	7.70%	9.62%	25.00%	-4.76%

Figure 3: Gain in prediction accuracy with respect to baseline offered by six well-known classifiers using six different settings.

Results shown in Table 3 reveals that k-NN can offers highest accuracy in 4 out of 6 settings than five other classifiers that exhibits its worth as predictor of match outcome in ODI cricket. Though Naïve Bayes does offer sufficient accuracy in Setting_1, it is still less than what k-NN offers in Setting_5.

4. Discussion

This study brings a distinct contribution to the cricket mining literature relating to the new time-series prediction problem i.e., predicting the outcome of One Day International cricket match. To answer our research question "Can the outcome of an ODI be more accurately predicted than current approaches?", we adopted several unique approaches for dataset formation and classification model learning and established a worthy statistical method that allows one to predict the winner of matches with 80% accuracy which is far greater than the work previously presented in literature. Hence, our question is answered in affirmative. Whole study revolves around formation of precise dataset and then finding the most relevant attributes out of it. Comprehensive effort has been putted in collection and preprocessing of the raw data. Consequently an attribute was discovered that has never been used in literature which is "Consecutive wins before current match". Vital attributes were identified through exhaustive analysis that gave rise to such distinguish results.

This study reveals that, being a simple algorithm, k-Nearest Neighbor has outperformed five other renowned classification algorithms that has not been proved in literature yet as far as prediction of ODI match outcome is concerned. It has also been verified that recent data has strong influence on the prediction of match outcome. This influence gets even more affective if we consider team's performance in immediate-past matches that was well exploited using online model learning technique.

Table 3: Prediction accuracy offered by top most classifiers using six different settings*. Where baseline is a percentage of loss of Pakistan in records included in Test-set i.e., if we predict a loss for each match in Test-set, we will get respective prediction accuracy

	Training Set Records	Test Set Records	Baseline (%)	Accuracy (%)					
				DT	RF	LR	NB	ANN	kNN
Setting_1	740	83	55.42	66.67	66.67	62.65	**77.78**	40.96	57.83
Setting_2	673	76	56.58	43.42	43.42	43.42	57.89	52.63	59.21

	Training Set Records	Test Set Records	Baseline (%)	Accuracy (%)					
				DT	RF	LR	NB	ANN	kNN
Setting_3	576	65	56.92	43.08	43.08	58.46	56.92	52.31	64.62
Setting_4	477	52	53.85	46.15	46.15	44.23	61.54	57.69	63.46
Setting_5	340	40	55.00	45.00	45.00	42.50	65.00	50.00	**80.00**
Setting_6	199	21	57.14	42.86	52.38	57.14	61.9	57.14	52.38

4.1 Practical Applications

Pakistan Cricket Board (PCB) can use our prediction model to assess the merits of certain strategies of play. The term strategy refers to the systematic plan of action taken by a team e.g. the coin toss (the captain winning the toss has an important decision to make; whether to bat or field first), Field placement, choosing bowlers, batting order, batting shot selection and sharing the strike. More importantly, a predicted match outcome right before the match starts can considerably change team's moral and confidence. For instance, if our tool predicts a Win for coming match, team could land confident in ground with a proper game plan and if it predicts a Loss, they could adjust their strategies accordingly by being more alert and careful while playing to turn the match in must win game. Furthermore, it could help cricket analysts and media essentially to discover winning pattern of Pakistan cricket team against all other opponents and pre-match analysis.

4.2 Future Work

In this study, a statistical analysis was carried out to predict match outcome, suggests several new avenues for future research work. As discussed in previous sections, attributes like Temperature, Humidity, Event and Wind speed could not be used in analysis due to their large number of missing values. The prediction accuracy can further be improved by taking account of such attributes. Similar study of predicting match outcome can also be carried out for other teams as well. Moreover, a cricket board must be concerned to improve team's performance in other game formats e.g., test and Twenty20 too. Therefore, analogous analysis can also be carried out for both formats.

5. Conclusions

In conclusion, our findings suggest five most decisive attributes and effective model learning techniques that offers match outcome prediction accuracy up to 80% (with gain of 25% w.r.t baseline) in ODI cricket. It is also discovered that the k-Nearest Neighbor is a most effective classifier to predict match outcome in ODI cricket as long as performance of last 15 years of a team is considered.

Acknowledgments

The authors would like to thank all colleagues who contributed to this study especially in data collection.

References

[1] S. R. Clarke, "Dynamic programming in one-day cricket-optimal scoring rates," *Journal of the Operational Research Society,* pp. 331-337, 1988.

[2] B. M. De Silva, G. R. Pond, and T. B. Swartz, "Applications: Estimation of the Magnitude of Victory in One-day Cricket RMIT University, Mayo Clinic Rochester and Simon Fraser University," *Australian & New Zealand Journal of Statistics,* vol. 43, pp. 259-268, 2001.

[3] M. Bailey and S. R. Clarke, "Predicting the match outcome in one day international cricket matches, while the game is in progress," *Journal of sports science & medicine,* vol. 5, p. 480, 2006.

[4] P. Allsopp and S. R. Clarke, "Rating teams and analysing outcomes in one-day and test cricket," *Journal of the Royal Statistical Society: Series A (Statistics in Society),* vol. 167, pp. 657-667, 2004.

[5] V. V. Sankaranarayanan, J. Sattar, and L. V. Lakshmanan, "Auto-play: A Data Mining Approach to ODI Cricket Simulation and Prediction," 2014.

[6] N. Tariq Mahmood, Talha, Mirza and Ahmed, "Predicting Cricket Performance through Data Mining," in *In Proceedings of the 2013 IEEE International Conference On Computer and Emerging Technologies*, Khayrpur, Pakistan, 2013.

[7] A. Kaluarachchi and A. S. Varde, "CricAI: A classification based tool to predict the outcome in ODI cricket," in *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on*, 2010, pp. 250-255.

[8] (2014, 15 October). *ESPN Cricinfo*. Available: http://www.espncricinfo.com/

[9] (2014, 15 October). *Rapid Miner Studio*. Available: https://rapidminer.com/

[10] W. C. Winter, W. R. Hammond, N. H. Green, Z. Zhang, and D. L. Bliwise, "Measuring circadian advantage in Major League Baseball: a 10-year retrospective study," *Int J Sports Physiol Perform,* vol. 4, pp. 394-401, 2009.

[11] P. Fowler, R. Duffield, and J. Vaile, "Effects of Domestic Air Travel on Technical and Tactical Performance and Recovery in Soccer," *International journal of sports physiology and performance,* vol. 9, pp. 378-386, 2014.

[12] B. Cunniffe, K. Morgan, J. Baker, M. Cardinale, and B. Davies, "'Home Vs Away'Competition: Effect on Psychophysiological Variables in Elite Rugby Union," *International journal of sports physiology and performance,* 2015.

[13] B. M. De Silva and T. B. Swartz, *Winning the coin toss and the home team advantage in one-day international cricket matches*: Department of Statistics and Operations Research, Royal Melbourne Institute of Technology, 1998.

[14] (2014, 15 October). *Weather Underground*. Available: http://www.wunderground.com/

[15] I. Asimov, "The tragedy of the moon," *The tragedy of the moon., by Asimov, I.. New York, NY (USA): Doubleday, 16+ 220 p.,* vol. 1, 1973.

[16] T. Weise, S. Achler, M. Göb, C. Voigtmann, and M. Zapf, "Evolving Classifiers–Evolutionary Algorithms in Data Mining," *Kasseler Informatikschriften (KIS) vol,* pp. 1-20, 2007.

[17] M. Panda and A. Abraham, "Hybrid evolutionary algorithms for classification data mining," *Neural Computing and Applications,* vol. 26, pp. 507-523, 2015.

Mr. Waqar Ahmed received his B.E. (Electronics Engineering) degree in 2012, M.S. (Electronics Engineering) in 2015 from PAF-Karachi Institute of Economics & Technology, Pakistan. He is currently pursuing doctorate in Electronics Engineering at PAF-KIET. He has worked on three funded research projects related to FPGA Architecture, Application and CAD tools. Pattern Classification and Machine learning are also his area of research where he exploited PCA, Fisher Vectors and Neural Network etc., for different research problems.

Dr. Khurram Nazir Junejo is an Assistant Professor at Karachi Institute of Economics and Technology. Previously, he worked at Singapore University of Technology and Design as a Post Doc Fellow on data driven approaches for securing cyber physical systems. His research interests also include text classification, sentiment analysis, and web navigation behavior prediction. Currently he is supervising three PhD, and four MS students in these areas. He is also the winner of prestigious Discovery Challenge Award (2006) held by European Conference on Machine Learning.

Dr. Tariq Mahmood is an Assistant Professor at the Institute of Business Administration (IBA), Karachi, Pakistan. His research interests include Big Data Analytics, Data Science, and Deep Learning. He heads the Big Data Analytics Research Lab (BDA-LAB) at IBA. He has several His research particularly focuses on comparison of diverse BDA tools and DL algorithms, along with creation of personalized BDA solutions for specific corporate domains. Dr. Mahmood also maintains active collaboration with industry over development of Hadoop solutions and data science projects. He also focuses on promoting his research interests in the corporate sector through conducting workshops and developing collaborations.

Ghulam Mujtaba received his Master degree in Computer Science from National University, FAST, Karachi, Pakistan. He is also a gold medalist in his masters. He is currently enrolled as a Ph. D. student in the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur Malaysia. He is also working as Assistant Professor in Sukkur Institute of Business Administration (Sukkur IBA), Sukkur, Pakistan since 2006 to date. He has vast experience in teaching and research. Before Sukkur IBA, he was working in well-known software house in Karachi, Pakistan for 4 years. He has also published several articles in academic journals indexed in well reputed databases such as ISI-indexed and Scopus-indexed.