Team #18: Predicting the Outcomes of NBA 3 Point Shots*

Nishant Vishwamitra[1] and Hakeem Jones[2]

Abstract— **The field of basketball analytics has seen a considerable growth in recent times. There has been some preliminary work in the field of basketball analytics using data mining techniques in developing analytics tools and solutions to challenging and important problems. We saw a rise in this field especially since the National Basketball Association (NBA) started making game data publicly available. In our research, we develop a *Support Vector Machine (SVM)* classifier that predicts make or miss of 3 point shots, based on a player's distance from basket, distance from nearest defender and time to shoot. We explain the rationale behind choosing a SVM classifier using powerful datamining techniques such as *Principal Component Analysis (PCA)*. Our work combines data from NBA Play-by-Play statistics and NBA SportVu player movements to formulate a dataset of 573 feature vectors collected over 15 full NBA games. To the best of our knowledge, our paper provides state of the art work in the field of 3 point shot prediction using machine learning techniques by combining multiple NBA data.**

I. INTRODUCTION

Basketball statistics have seen a tremendous growth in recent years and have become a major aspect of game analytics and for formulating game strategies. Player tracking data, such as SportVu have shed a new light on the way game strategies are explored. However, analyzing NBA strategies is a complicated task, due to the numerous influencers in a game [1]. Arbitrarily, the make or miss of a 3 point shot can be influenced by the distance of shooter from the basket, the positions of the defenders, the time that the shooter has got to complete the shot, the positions of the team mates, the quarter, the player's height and so on. Additionally, the make or miss of a 3 point shot can be area specific [6].

Basketball is a game of decisions. It is crucial for basketball researchers, scientists, strategists and game enthusiasts to identify the features of the game that are pivotal in the decision making process of a basketball player. As a trade-off, these features must also be easily *capturable* and *measurable*. With the advent of technology, it is now possible to capture such features that tell the story behind a game decision. The *SportVu* [15], [16] player tracking system, installed in the ceilings of respective arenas provide *fine-grained* player movements. The system consists of 6 cameras that measure each players and the balls location on the x, y, and z axis with 1/25 of a second temporal accuracy. The *Play-by-Play (PBP)* [17] data provides a complete game *narrative* that includes information such as make/miss of a shot, player who attempted a shot and the time at which an attempt is made. By combining the SportVu and the PBP data, an enormous potential for *datamining* basketball trends is unlocked.

There is a shift in the field of basketball analytics, with scientists and statisticians focusing on player tracking data rather than manually observing and recognizing patterns in team strategies. Precise positions of players and the ball in every moment is indeed a very useful tool for recognizing game patterns. As discussed, the game of basketball can be observed as a temporal and spatial problem, based on the player and ball movements and the game time [1], [2], [4]. In this work, we present a *machine learning classifier*, trained to predict the make or miss of a 3 point shot. In our work, we consider both temporal and spatial features as equally important influencers. Although it may appear that all the temporal and spatial features are important to determine the outcome of a three point shot, we consider a subset of these features for our prediction. Using effective machine learning and statistics tools, such as *Principal Component Analysis (PCA)* [11], [12], [13], we determine the most important influencers in the success or failure of a three point shot. We present our findings by representing the projection of our data on the two principal components as a scatter plot. Thus, a visualization of our dataset presents us with a direct understanding of the *data mining* strategy that we wish to formulate for our analytics problem.

The results of visualization using *PCA* gives datamining researchers a lot of information about a particular dataset. *Linear separability* of data is one such important dataset characteristic that can be visualized using *PCA* analysis. State of the art [18], [19], [20], [21] suggests that linear classifiers, such as *Support Vector Machines (SVM)* perform with high sensitivity in cases very the dataset is linearly separable and machine techniques such as *Artificial Neural Networks (ANN)* perform well in cases where the dataset is not linearly separable, or in case the *datamining* scientist looks for patterns using unsupervized learning using ANNs. In our work, we discuss the algorithm that we use for the 3 point shot prediction problem and the *rationale* behind choosing the algorithm using mathematical proof.

In our work, we make the following contributions.

- Compiled a **unique dataset** consisting of 573 feature vectors, by combining player movements and game narrative.
- Provided a **visualization** of our dataset using *PCA*
- Formulate a **machine learning architecture/strategy** based on mathematical *rationale*; and
- Evaluate **results** of our classifier and provide discussions.

The rest of the paper is organized as follows. We introduce

*Both authors are first co-authors of this work

[1]nvishwa@clemson.edu, School of Computing, Clemson University
[2]hakeemj@clemson.edu, School of Computing, Clemson University

the prediction of 3 point shots in the *Problem Statement* section. In the *Data Preprocessing* section, we discuss the methodology behind our data collection and data cleaning strategy. In the *Data Visualization and Analysis* section, we visualize data using data mining tools and analyze the results to choose the best classifier for our dataset. The *Data Classification* describes the details of our classifier. The *Results and Discussion* section provides the classification results of our classifier and baseline comparisons for performance metrics. The *Related Work* section discusses some recent research being carried out in the field of basketball analysis using player tracking data and datamining techniques. We finally conclude this paper with the *Conclusion* section.

II. Problem Statement

A. Predicting Make/Miss of a 3 Point Shot as a Datamining/Machine Learning Problem

The Make/miss of a 3 point shot in an NBA game is not a determinate problem. There are various influencers that contribute to the decision of a player.

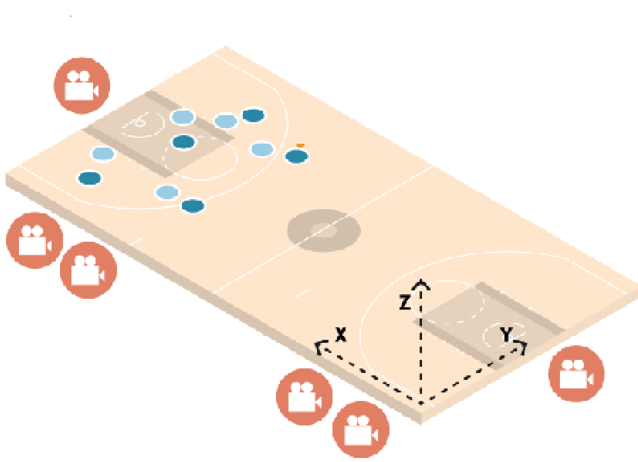

Fig. 1. SportVu Coordinate Plane

The proximity of a player to the basket is an obvious factor that plays a role in the decision making process [4], [7], [1]. As can be seen from Figure 1, a 3 point shot can also be influenced by the distance of a player from the nearest defender, as closeness to a defender might make the player take a different course of actions. The time remaining to make a shot is also a crucial factor, as a player might make an attempt if time is running out by giving less importance to the other factors. Thus in our model, we have used the following features to predict the make/miss of a 3 point shot.

1) Time Remaining to attempt a shot.
2) Distance from the basket.
3) Distance from the nearest defender.

As seen from Figure 1, the player movement tracking data provides us with the x, y and z coordinates of the players and the ball for every *moment* of the game. Using this information, we can calculate the features mentioned above at every attempt of a player for a 3 point shot. This is an non deterministic problem, because these features influence one another, i.e, they are correlated. In the sections that follow, we analyze the pattern of make/miss of a 3 point shot using datamining techniques such as PCA.

B. A Solution Model

We develop a solution model based on machine learning techniques. A make or miss indicate two classes of data, which gives rise to a ***classification*** problem. Our solution model, thus is a *two class, three dimensional classification problem*, founded on reasoning based on datamining approaches which will be discussed in the sections that follow.

III. Data Preprocessing

Before we could get to running our machine learning techniques to determine the success of a three point shot we needed to find when the shot took place and then extract the aforementioned features: *time to shoot, distance from basket, and distance from nearest defender*. Though the SportVu data provides a wealth of spatial and temporal information, it is almost useless without context. Each SportVu file is essentially a large list of lists (70-100mb), where the game is segmented by roughly 20 second intervals called events that are listed sequentially. Each event contains a player list for each team and 25 moments per second. Each moment contain the quarter, unix time, game/shot clock time in seconds and a list of 11 lists of the ball and each players postion in 3 dimensions. To extract the information relevant to our study we must find the precise moment a shot is taken and then do calculations on the spatial data.

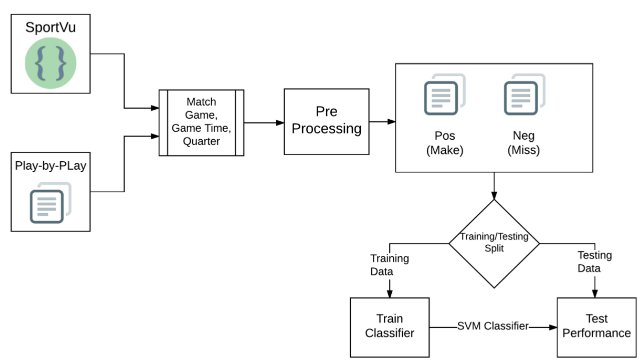

Fig. 2. Schematic of Our Methodology

Fortunately the play by play transcripts from each NBA game are widely accessible. This information from the sideline announcers describes what is happening in the game such as, made or missed shots, steals, blocks etc and what time each play occurred. To get all the 3 point shot instances we parsed the play by play transcripts and recorded the time, quarter, result and shooter for each 3 point shot instance. Using this data as input we were able to search thru the SportVu data for a moment as the player is shooting. Finding an appropriate moment took a bit of fuzzy logic, because the sideline announcers are always temporally off from the

SportVu. This is because a sideline announcer has watch a play, interpret it and then announce it. This delay can often be as much as 10 seconds in some instances. Thus we implemented a logic similiar to below:

1) Find right event number using play by play time
2) In this event loop thru each moment
3) Stop when the moment's time is 10 seconds from play by play, ball's height is inbetween 8-9ft, and ball is less than 3ft away from the shooter

Once we found the right moment we calculated our desired features. We calculated the time to shoot using the minimum of the game clock and shot clock. Distance to nearest defender was found by looping thru all players on the floor with a different teamID than the shooter. Distance from hoop was found by first computing the direction of the ball by examining its trajectory by comparing its current postion with its position in a later moment. With that knowledge we then calculated the shooter's distance from the basket he was aiming at.

This process was ran for every 3 point shot for all the games in our dataset. On average an NBA team shoots at 25 3 pointers per game so each game gave us about 50 shots for our dataset. The process is described in the schematic diagram provided in Figure 2.

IV. Data Visualization and Analysis Using PCA

In machine learning problems, the choice of algorithm is pivotal upon the characteristics of the dataset. *Principal Component Analysis (PCA)* is a powerful datamining technique that allows visualization of a dataset, so that non-trivial decisions can be made regarding the classifier design. Therefore, we use PCA to visualize our data and provide a rationale behind the choice and design of our classifier.

PCA technique mathematically *"chases"* the direction in which there is maximum ***variance*** in the dataset. What this means in that, the component (*feature*) that has the maximum variance is the *principal component* of the dataset. In machine learning, PCA can be effectively applied to identify the feature that is the largest influencer of the data.

In *Principal Component Analysis* of our dataset, we first normalized our dataset by making it 0 mean. Then, we computed the Covariance matrix of our dataset.

$$X = Cov(Y) \tag{1}$$

Our dataset is a 573×3 dimensional matrix. Hence the covariance matrix derived from Equation 1 would be matrix of dimension 3×3. Next, we compute the *Eigen vectors* and the corresponding *Eigen values*, using the Covariance matrix in Equation 1.

$$Evalues, Evectors = Eig(X) \tag{2}$$

The Evalues is a one dimensional matrix. The Evectors is a 3×3 dimensional matrix. The following tables (Tables I and II) depict the Evalues and corresponding Evectors computed for the dataset X.

TABLE I
EVECTORS OF THE COVARIANCE MATRIX X

0.99938591	-0.03176084	0.01480062
0.03208233	-0.99924274	0.02201528
-0.0140918	0.0224766	0.99964807

The corresponding Evalues are shown in the table below (Table II).

TABLE II
EVALUES OF THE COVARIANCE MATRIX X

First PC	Second PC	Third PC
4727.19263664	539.23281782	28.11689059

From tables I and II, it can be observed that the first and second Eigen values correspond to the first and second *principal components* of the dataset.

This observation leads to us consider the first and second rows of Table I as the Eigen vectors that correspond to the maximum variance in the dataset. Let this be the matrix $E_{2\times3}$. Next, we must take the ***projection*** of our dataset over the Eigen vectors that correspond to the first and second *principal components*, which as a result provides a two-dimensional data which is convenient to visualize.

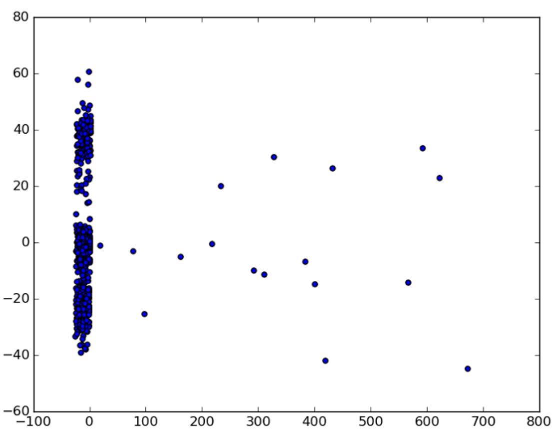

Fig. 3. PCA Analysis Result

$$Projection,\ \ P = E_{2\times3} \times X^{T}_{3\times573} \tag{3}$$

The Equation 3 gives us the projection of our dataset along the *principal components*, a *two-dimensional* dataset that can be plotted and visualized for human inspection of patterns. The Figure 3 depicts the scatter of the two dimensional projection from Equation 3. A discussion of the PCA analysis results is to be made here. From Figure 3, we can see that there are two clusters of data. From the nature of our dataset, we can conclude that the two clusters belong to the shot made and the shot missed.

Upon further inspection, we can observe that the data is for the most part linearly separable, except for outliers that are scattered across outside either of the two clusters. Hence, we can conclude that we can use machine learning classifiers that are effective in linearly separable datasets, such as *Support Vector Machines*. Thus, using PCA, we provide a mathematical proof as to the choice of our classifier. For our use-case, we use the Support Vector Machine classifier with linear transformation. We provide more details about the design and performance of our classifier in the sections that follow.

V. DATA CLASSIFICATION

We use *Support Vector Machine (SVM)* classifier for prediction of 3 point shots. SVM algorithm iteratively finds the hyperplane that linearly separates the data [21]. This hyperplane is found by computing the *support vectors*, which are the vectors in the dataset set that correspond to the maximum distance towards the separating hyperplane. We have experimented with several linear transformations, such as *radial* and *linear* transforms. However, we found no substantial difference between the classification performance, with these different non-linear transformations. Our SVM classification methodology is described as follows [22].

Let us denote our training dataset as,

$$(x_i, y_i), i = 1, ..., n, x_i \in R^d, y_i \in 0, 1 \qquad (4)$$

In Equation 4, x_i denotes the training features and y_i denotes the labels, which in this case is 0 or 1 (negative or positive respectively). R^d denotes the sample space.

Since data is not always linearly separable, we use non-linear transforms to transform the training feature vectors into another dimension where they can be better linearly separable. Let H denote this new transformed dataset. Let ϕ denote this non-linear transform that transforms R^d into H. Therefore,

$$\phi : R^d \rightarrow H \qquad (5)$$

Next, we search for a hyperplane in H, such that a transformed vector $\phi(x_i)$ lies towards one side of the hyperplane H if $y_i = 0$; and $\phi(x_i)$ lies towards the opposite side of the hyperplane H if $y_i = 1$. The equation of this hyperplane can be represented by the following equation. In the Equation 6, "." represents the dot product.

$$\omega.\phi(x) + b = 0 \qquad (6)$$

In the Equation 6, ω is the normal to the hyperplane and $|b|/||\omega||$ is the distance of the hyperplane from the origin. Since there are a finite number of samples, each sample must be at least a β distance away from the hyperplane, for $\beta > 0$.

$$y_i(\phi(x_i).\omega + b) - 1 \geq 0, for\ \ i = 1, ..., n \qquad (7)$$

Though ideally, it may be impossible to separate *all* points with a hyperplane even in H. In this case, we look for the hyperplane that separates training vectors as much as possible, while maximizing a relaxation margin. In order to do this, a relaxation factor ϵ_i is introduced in Equation 7.

$$y_i(\phi(x_i).\omega + b) - 1 + \epsilon_i \geq 0, for\ \ i = 1, ..., n \qquad (8)$$

In Equation 8, ϵ_i is called the slack variable, which is used to measure the degree of misclassification of x_i. Let's say d_1 is the distance from the support vectors to the hyperplane for make of a 3 point shot and d_0 is the distance from the support vectors to the hyperplane for the miss of a 3 point shot. The margin between the two classes is $d_1 + d_0 = d/||\omega||$. In order to maximize the margin, $||\omega||^2$ has to be minimized. A relaxation factor has to be added to account for the slack variables, in the form of $C\sum_i \epsilon_i$, where C is a constant. Thus, this gives rise to the following optimization problem.

Minimize

$$||\omega||^2 + C\sum_i \epsilon_i \qquad (9)$$

Subject to

$$y_i(\phi(x_i).\omega + b) - 1 + \epsilon_i \geq 0, for\ \ i = 1, ..., n \qquad (10)$$

By solving Equations 9 and 10, the values of ω and b can be found out, thus finding the appropriate separating hyperplane.

We split our dataset into 470 training samples and 103 test samples. We use a linear SVM for classification using the above method. In the section that follows, we elaborate over the results of our classification and derive some observations.

VI. RESULTS AND DISCUSSION

We discuss our results and performance of our classifier using the *confusion matrix* in Table III.

TABLE III
CLASSIFIER PERFORMANCE: CONFUSION MATRIX

Confusion Matrix	Predicted NO	Predicted YES
Actual NO	TN=2	FP=47
Actual YES	FN=6	TP=48

From the Table III, we can observe the following classifier metrics.

1) Accuracy = $\frac{TP+TN}{N} = 0.48$
2) Error Rate = $\frac{FP+FN}{N} = 0.51$
3) Recall/Sensitivity = $\frac{TP}{Actual\ YES} = 0.88$
4) Specificity = $\frac{TN}{Actual\ NO} = 0.04$; and
5) Precision = $\frac{TP}{Predicted\ YES} = 0.50$

The low accuracy and precision values maybe attributes to the small size of our dataset. We intend to circumvent this issue by collecting more data, so that a large and rich dataset is produced to predict the make/miss of a 3 point shot with high accuracy and precision. We however, note the large value for recall (0.88). This shows that our classifier will be able to predict new vectors accurately.

Another factor to consider here is the matching of NBA PBP data and SportVu data. On correspondence with Stats [16], we observed that the game clock times on both data were off by a few seconds. That is, they the time do not exactly correspond from one type of data to another. Thus, we get data points that are not exactly made at the time of the shot. This inherent inaccuracy maybe the cause of the low precision and accuracy scores. We intend to mitigate this issue by computing an error rate in the clock times and adjusting the times to match the movements more accurately. This way, we will get a more accurate version of the player movements with time, which would lead to a more robust dataset.

In addition, we intend to come up with an enhanced strategy to correlate the SportVu player tracking data with the Play-by-Play data. This can be accomplished by careful experimentation and analysis of the underlying nature of the two datasets.

VII. Related Work

There has been significant research in the field of basketball analytics. One of the first works in this field was done by the authors of [4], by applying the SportVu motion capture data to the analysis and statistics of a basketball game. This work proved to be a seminal work in this field, ushering in a new age of basketball statistics by using player tracking data provided by SportVu.

Authors is [1] represent decisions made in a basketball game by using player tracking data provided by SportVu. Common moves such as dribble, shot etc are evaluated. In addition, they present their modeling framework, applying it to every moment in SportVu data to come up with new metrics and predictions of the game. In another paper [6], the authors elaborate over predicting spatial regions on the court using SportVu data. In this way, they present a model to estimate the real estate of a court based on player movements. They present new metrics in their work.

Using player tracking data provided by SportVu, authors in [8] develop models to explain how shooters get open. They establish that, provided on ball defensive pressure reduces shooting percentage, how an offense can get shooters open. By demonstrating that frequency of defensive role swaps can be used to predict open shots, they use this important finding to measure the effective of a team's defense. They also hypothesize a method for querying archived games for such metrics by using the SportVu player tracing data.

Authors of [7] present a model that learns basketball statistics such as rebounds, using player tracking data instead of just numbers from box score. By analyzing player tracking data using datamining techniques, they show that rebound events can be expressed in three dimensions: Positioning, Hustle and Conversions. They also come up with new metrics to better analyze a basketball game using player tracking data provided by SportVu.

The application of artificial neural network in machine learning problems has seen enormous success in recent times, especially due to the proliferation of inexpensive general processing units (GPUs). The work in [10] describes the use of LSTM (Long short term memory) networks, a type Recurrent Neural Networks, to operate on *big* SportVu data to contrast and perform analysis of basketball plays, especially the classification of offensive plays. LSTM allows the learning of complex on-court interaction of the players. To emphasize the effective of their method, they demonstrate how well the LSTM model performs when different plays belonging to totally different seasons are provided as inputs to the network. An innovative basketball coaching assistance tool has been developed in [9], for analyzing game plays. The authors use machine learning to train a classifier to label ball screen plays according to defensive response strategies such as Over, under etc.

Authors of [2] describe the use of SportVu data in predicting defensive plays using neural networks. In this work, there is a shift in focus from offensive plays to defensive plays. They present a model that provides a new and revolutionary number of defensive metrics that are not just limited to the overall defensive ability such as blocks and steals. They use of combination of statistical modeling techniques with player tracking data, such as data provided by SportVu and visualization techniques to effectively provide better characterization of defensive tactics in a basketball game.

However, as we can observe from the state of the art, we have no significant research in predicting the make or miss of 3 point shots. Though a lot of work has been done by using player tracking data provided by SportVu, very little work has been done by combining the player tracking data with the Play-by-Play narrative to get new insights into the game of basketball, although the correlation of the two types of data sources also gives rise to new challenges. Visualization is an important part of datamining. Yet, very few research has focused on developing novel methods for data visualization using player tracking data. Additionally, most work have used the player tracking data as-is, to perform datamining related analytics. Pre-processing data by combining multiple sources is more meaningful from this aspect, as a rich, high quality data is bound to give better results than a huge amount of fine-grained, un-processed data.

VIII. Conclusion

In this work, we have developed a novel machine learning classifier based on *Support Vector Machine* to predict make or miss of NBA 3 point shots, using player tracking data provided by SportVu and game narrative provided by Play-by-Play. We have provided an in-depth methodology behind our data pre-processing strategy, to develop a rich dataset in three dimensions. Furthermore, we provide a visualization and analysis of our dataset using*Principal Component Analysis*, a powerful datamining tool used in visualization and dimensionality reduction. Using the PCA analysis results, we establish the *linearly separable* nature of our dataset, as use this as mathematical rationale behind using hyperplanar classifiers such as *Support Vector Machines*. We present the performance results of our classifier using a confusion matrix, to calculate classifier performance metrics such as *accuracy, sensitivity, recall* etc. Finally, we discuss some state of the art that have been influenced us to present our work in the field of basketball analytics.

IX. Future Work

As future work, we intend to extend this methodology to predict the performance of individual NBA players in terms of their 3 point shot performance. We want to provide novel metric that measure the NBA shot success of a player using player tracking data and machine learning techniques.

Acknowledgment

We would like to thank Dr. Wang for the informative Cpsc 8560, Spring 2017 course on Datamining. Clemson University is acknowledged for generous allotment of compute time on Palmetto cluster.

References

[1] Cervone, Dan, et al. ”POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data.” 8th Annual MIT Sloan Sports Analytics Conference, February. Vol. 28. 2014.

[2] Franks, Alexander, et al. ”Counterpoints: Advanced defensive metrics for nba basketball.” 9th Annual MIT Sloan Sports Analytics Conference, Boston, MA. 2015.

[3] McQueen, Armand, Jenna Wiens, and John Guttag. ”Automatically recognizing on-ball screens.” MIT Sloan Sports Analytics Conference (SSAC). 2014.

[4] YonggangNiu, Haojie Huang, and Huanbin Zhao. ”Application of the Sport VU Motion Capture System in the Technical Statistics and Analysis in Basketball Games.” Asian Sports Science: 45.

[5] Goldsberry, Kirk. ”Courtvision: New visual and spatial analytics for the nba.” 2012 MIT Sloan Sports Analytics Conference. 2012.

[6] Cervone, Daniel, et al. ”A multiresolution stochastic process model for predicting basketball possession outcomes.” Journal of the American Statistical Association 111.514 (2016): 585-599.

[7] Maheswaran, Rajiv, et al. ”Deconstructing the rebound with optical tracking data.” MIT Sloan Sports Analytics Conference. 2012.

[8] Lucey, Patrick, et al. ”How to get an open shot: analyzing team movement in basketball using tracking data.” MIT SSAC (2014).

[9] McIntyre, Avery, et al. ”Recognizing and Analyzing Ball Screen Defense in the NBA.” Ann Arbor 1001: 48104.

[10] Wang, Kuan-Chieh, and Richard Zemel. ”classifying NBA offensive plays using neural networks.” Proc. MIT SLOAN Sports Analytics Conf. 2016.

[11] Jolliffe, Ian. Principal component analysis. John Wiley & Sons, Ltd, 2002.

[12] Wold, Svante, Kim Esbensen, and Paul Geladi. ”Principal component analysis.” Chemometrics and intelligent laboratory systems 2.1-3 (1987): 37-52.

[13] Abdi, Herv, and Lynne J. Williams. ”Principal component analysis.” Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

[14] Remmert, Hubert. ”Analysis of group-tactical offensive behavior in elite basketball on the basis of a process orientated model.” European Journal of Sport Science 3.3 (2003): 1-12.

[15] Tamir, Michael, and Gal Oz. ”Real-Time Objects Tracking and Motion Capture in Sports Events.” U.S. Patent Application No. 11/909,080.

[16] STATS, L. ”SportVU.” (2015).

[17] Kubatko, Justin, et al. ”A starting point for analyzing basketball statistics.” Journal of Quantitative Analysis in Sports 3.3 (2007): 1-22.

[18] Osuna, Edgar, Robert Freund, and Federico Girosi. ”An improved training algorithm for support vector machines.” Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop. IEEE, 1997.

[19] Yegnanarayana, B. Artificial neural networks. PHI Learning Pvt. Ltd., 2009.

[20] Hopfield, John J. ”Artificial neural networks.” IEEE Circuits and Devices Magazine 4.5 (1988): 3-10.

[21] Medin, Douglas L., and Paula J. Schwanenflugel. ”Linear separability in classification learning.” Journal of Experimental Psychology: Human Learning and Memory 7.5 (1981): 355.

[22] Spasova, Velislava, Ivo Iliev, and Galidiya Petrova. ”Privacy Preserving Fall Detection Based on Simple Human Silhouette Extraction and a Linear Support Vector Machine.” International Journal Bioautomation 20.2 (2016).