Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier XX.XXXX/ACCESS.XXXX.DOI

Machine learning enabled team performance analysis in the dynamical environment of soccer

S. KUSMAKAR[1], (Member, IEEE), S. SHELYAG[1], Y. ZHU[1], (Member, IEEE), D.B. DWYER[2], P.B. GASTIN[3], AND M. ANGELOVA[1]

[1]School of Information Technology Deakin University, Geelong, Australia, Vic 3125, (e-mail: s.kusmakar@deakin.edu.au)
[2]School of Exercise & Nutrition Sciences, Deakin University, Geelong, Australia, Vic 3125, (e-mail: dan.dwyer@deakin.edu.au)
[3]La Trobe Sport Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia Vic 3086, (e-mail: p.gastin@latrobe.edu.au)

Corresponding author: M. Angelova (e-mail: maia.a@deakin.edu.au).

This work was supported by DSI collaborative research grant RM35517 "Intelligent sensor processing for enhancing defence decision support".

ABSTRACT Team sports can be viewed as dynamical systems unfolding in time and thus require tools and approaches congruent to the analysis of dynamical systems. The analysis of the pattern-forming dynamics of player interactions can uncover the clues to underlying tactical behaviour. This study aims to propose quantitative measures of a team's performance derived only using player interactions. Concretely, we segment the data into events ending with a goal attempt, that is, "*Shot*". Using the acquired sequences of events, we develop a coarse-grain activity model representing a player-to-player interaction network. We derive measures based on information theory and total interaction activity, to demonstrate an association with an attempt to score. In addition, we developed a novel machine learning approach to predict the likelihood of a team making an attempt to score during a segment of the match. Our developed prediction models showed an overall accuracy of 75.2% in predicting the correct segmental outcome from 13 matches in our dataset. The overall predicted winner of a match correlated with the true match outcome in 66.6% of the matches that ended in a result. Furthermore, the algorithm was evaluated on the largest available open collection of soccer logs. The algorithm showed an accuracy of 0.84 in the classification of the 42,860 segments from 1,941 matches and correctly predicted the match outcome in 81.9% of matches that ended in a result. The proposed measures of performance offer an insight into the underlying performance characteristics.

INDEX TERMS Dynamical systems, network science, distribution entropy, football, Kolmogorov complexity, machine learning, performance analysis, Shannon entropy, support vector machines, soccer.

I. INTRODUCTION

IMPROVING comprehension of strategic performance and success in team competition is an important goal in sports science [1]. Data-driven methods can effectively overcome the subjective limitations (manual analysis) of the match and offer better results for football clubs. Quantitative analysis can provide players and coaches with such insight, by allowing them to improve their match and assessment of the event beyond what personal observation can accomplish [2]. Traditionally, methods of performance analysis push the study of one-dimensional and discrete performance indicators towards probabilistic and correlational approaches [3]. However, this results in somewhat limited functional information as it lacks the understanding of the player-to-player interactions that support the actions of players and overall team behaviour.

It is reasonable to expect an analysis of such one-versus-one dynamics in team sports to be insufficient as multiplayer interactions are important in determining success and failure [4]. Therefore, in order to quantify and explain performance, it has been advocated that performance analysis in team sports must also focus on the interactions between players that sustain the overall team behaviour [5], [6]. From the dynamical systems view, the understanding of how the co-ordination emerges from the interaction among the system components, that is, the player-to-player interaction, is the

key to performance analysis [7], [8]. In team sports, performance analysis approaches that consider the interactions of the players in many multiplayer team competitions like football are not well explored [9].

Inspired by empirical studies of networked systems, researchers have recently developed a variety of techniques and models to help us understand player interaction network in sports [10]–[13]. Interaction or passing networks can be constructed from the observation of ball transfer between players. A key challenge is to leverage the interaction networks to gain a functional understanding of the underlying team strategies. For example, by examining the structure of interaction networks, recurrent pass sequences can be identified and linked to a team's playing style [14], [15]. When the emphasis is put at the player level, Duch et al. [16] used the interaction networks to quantify and rank player's contribution relative to the overall team activity.

Due to dissimilarity and diversity in real-world sports data, there is no systematic program for predicting network structure. In addition, there are no particular subsets of diagnostics that are universally accepted [17]. Since team networks are intrinsically subjective and dynamic objects, it is often hard to determine a suitable way of network characterisation that governs team formation [18]. In team sports like football, quantifying player-to-player interaction is the key for understanding the dynamic patterns that generate a scoring opportunity [19]. This motivated us to develop an approach that quantitatively characterises players' interaction in team sports. In this study, a data-driven approach to the study of complex player interactions from event stream data generated during football matches (henceforth referred to as soccer) is employed. The proposed framework can be used to quantify player interactions and connect that with the outcome using a machine learning approach.

Data-driven approaches for soccer analytics are given importance with the availability of the event stream data (*e.g.*, Opta, Wyscout, STATS, SecondSpectrum, SciSports, and StatsBomb). Cintia et al. [15] in their work, extracted pass-based performance measures to learn the correlation to match outcome using a machine learning approach. More recently, Pappalardo et al. [20] in their work employed a machine learning approach to rank players. Their approach is based on computing statistical features from the event stream data for each player, which are then utilised to learn feature weights in a supervised learning framework *i.e.*, relative to the match outcome. The authors then use the learned weights to compute the rating of a player. In another recent study by Decroos et al. [21], the authors have performed a segmental analysis of different match states to extract several associative features of player performance, which are then used to determine the scoring or conceding probability using an ensemble classifier. In contrast to the above-mentioned studies that consider individual player's actions or cumulative team statistics, the proposed study describes a segment of a match using a set of activity and entropy-based quantifiable markers that capture both inter- and intra-player interactions.

To quantify interaction among players in team sports conceived as dynamical systems unfolding in time, it is important to use appropriate measures [22], [23]. The proposed study considers the behaviour of multiple players and the emergent nature of performance to develop pattern-forming dynamics, that is, the dynamic physical relationships that a player may establish with the teammates and opponents to make a goal. We developed a coarse-grain activity model of player-to-player interaction from the possession chain data, that can be used to quantify the dynamic patterns underlying the interaction among players. We used the concepts of information theory retrieval to quantify the complexity of a pattern representing player interactions during sub-segments of the match. Another key challenge from the analytics perspective is the format of the soccer log data, as different vendors use different data formats [21]. Therefore, an analyst has to develop complex pre-processors specific to a dataset. To tackle the challenges posed by the variety of event stream formats and to benefit the data-science community, we propose an approach that uses only a limited amount of information. The proposed approach only uses the possession information, such as player, team, action type, and result from the event stream data. The segmental analysis was thus performed using only the possession information to quantify the team performance and stability in team-dynamics during a specific module, that is, a match segment. Furthermore, based on the derived performance measures we developed a machine learning-enabled decision support system for automated prediction of a team's likelihood of a successful attempt at goal.

II. APPROACH

A. DATASET

In this study we have analysed the dataset from a season of Major League Soccer division of the United States and Canada. The dataset consists of the possession chain data from 13 matches. The interaction information (possession chain) comprises of time and duration of all ball passes and tackles between players. The dataset also includes the nature of the interaction which can be categorised as being between teammates or between opposing players (Table 1). The positional information includes the x-y position of all individuals throughout the entire match ($\sim$90 minutes).

B. COARSE-GRAIN PLAYER INTERACTION MODEL

Given: A set of possession chain information for each match, representing a set of events (pass, shot etc) between players and the game outcome.

A match is split into a number of segments, where each segment represents a phase of the match that begins with either the start of the match or after an attempt (Shot) at the goal and ends with a "*SHOT*" (see Fig. 1). Further, throughout the text the teams in an adversarial relationship during a match were denoted by team-1 and team-2 for each match in the dataset. Using the possession information corresponding to every segment in the match, we propose a coarse-grain model to find quantifiable measures of performance

TABLE 1: An example of ball possession chain data. The table shows a part of a ball possession chain dataset, which represents events in the 1^{st} half of a match.

Team*	Type	SubType	Period	StartTime [s]	From*	To*	Start X (GPS)	Start Y (GPS)
team-1	Set Piece	Kick Off	1	47.4	Player 1		nan	nan
team-1	Pass		1	48.64	Player 1	Player 2	0.5	0.5
team-1	Pass		1	50.81	Player 2	Player 3	0.67	0.49
team-2	Challenge	Aerial Fault Won	1	52.68	Player 4		0.32	0.22
team-1	Challenge	Aerial Fault Lost	1	52.84	Player 3		0.32	0.23
team-1	Ball Lost	Forced	1	54.8	Player 3		0.32	0.23
team-2	Set Piece	Free Kick	1	55.96	Player 5		nan	nan
team-2	Pass		1	57.52	Player 5	Player 6	0.27	0.3
team-2	Pass		1	59.96	Player 6	Player 7	0.3	0.66
team-2	Pass		1	64.44	Player 7	Player 6	0.43	0.9

* The data were deidentified by request of the data owner.

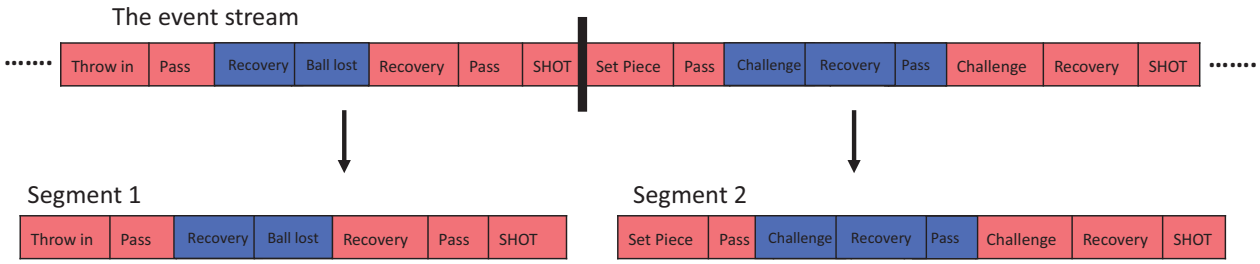


FIGURE 1: Segmentation of the possession chain data. A match is split into different segments of varying length (or the number of events in a segment) ending with a "*SHOT*". The red and blue shaded cells represent possession by different teams. Each segment was individually evaluated for measures of performance.

that demonstrate an associationship with the outcome of that segment, that is, which team (team-1 or team-2) makes an attempt to score by taking a "*SHOT*" at the opposition's goal.

1) Coarse-Grain Models Derived from Possession Chain Data

Each of the match segments was studied separately. The segments represent a sequence of ball possession change events leading to an attempt to score. Each team in a soccer match has 11 players with 3 allowed replacements. Based on the sequence of events in each segment we define two types of coarse-grain models. The first model weighs all the events (e.g. pass, shot at goal, ball lost) equally, whereas the second model weighs events based on their type. More specifically, a higher weight to an event denotes a higher relevance. As we are interested in measures that quantify a successful attempt to score, we assign higher weights to shots and recoveries and lower weights to events like ball lost and faults.

For each segment we first generate a pairwise player matrix, $M_{i,j}$, where $i, j = \{1,\ldots,28\}$, each element of which was initialised to zero. The matrix M contains players of both teams ($i, j = \{1,\ldots,14\}$ and $i, j = \{15,\ldots,28\}$ for team-1 and team-2, respectively) and any element $M_{i,j}$ represents the interaction of the i^{th} player with j^{th} player in the segment. The value of the $M_{i,j}$ element denotes the number of times the players interacted or the number of times the players interacted weighed by the type of event. For example, if player 1 of team-1 passes the ball to player 5 of team-1 the element $M_{1,5}$ of the matrix is incremented by 1 (*i.e.*, $M_{1,5} = M_{1,5} + 1$). Similarly, if player 3 of team 2 recovers the ball in a tackle from player 5 of team 1, then the element $M_{14+3,5}$ of the matrix is incremented by 1 (*i.e.*, $M_{14+3,5} = M_{14+3,5} + 1$). Therefore, the diagonal $14x14$ blocks of the matrix M denote interactions of the players within a team whereas, the off-diagonal blocks represent the inter-team player interactions.

Thus, the matrix M (such that, $M_{i,j} \geq 0\ \forall i, j$) was termed as the interaction matrix. The matrix M represents the connections on the network of players (agents), related to activity-based decision-making to the directed transfer of information (ball) from one agent to another. This coarse-grain interaction model (M), represents the network of connections, accumulated over a sequence of events during a segment of the match. We analysed the interaction between players based on the following approach:

a: Unit increment

Each element $M_{i,j}$ of the interaction matrix is incremented by 1 for an interaction between the i^{th} and j^{th} player of the same team (ball passed) or players of the different team (ball recovery, tackle, ball lost etc.) as follows:

$$M_{i,j} \mapsto M_{i,j} + 1 \tag{1}$$

b: Weighted increment

Each element $M_{i,j}$ of the interaction matrix is weighted by the type of the event E_t. More specifically, we assign a weight to

each event for evaluation of its contribution. Given that we are mostly interested in goal attempts, we introduce a higher weight for shots in comparison to passes, ball losses and other events that lead to loss of ball possession.

$$M_{i,j} \mapsto M_{i,j} + W_{E_t}, \tag{2}$$

where W_{E_t} is the weight corresponding to the event E_t. As we are interested in the likelihood of a successful attempt at goal, we assign a high weight $W_{shot} = 2$ to shots, a low weight $W_{pass} = 0.5$ to passes, and an average weight $W_{shot,pass} = 1$ to all other event types as suggested by Decroos et al. [24].

C. QUANTIFICATION OF TEAM PERFORMANCE FROM COARSE-GRAIN MODELS

To quantify the performance of a team in each segment of the match four measures were proposed:

1) Total Activity Index (TAI)

To quantify the interaction in a segment the matrix M is further divided into four blocks by summing all the elements ($M_{i,j}$) in top left (team-1 $\forall\ i,j\ in\ \{1,\ldots,14\}$), and the bottom right (team-2 $\forall\ i,j\ in\ \{15,\ldots,28\}$), which represent the overall activity of each team that is obtained by summing the activity of all players in a team. The off-diagonal elements represent the interaction between players of both teams. Any row in the top left (team-1 $\forall\ i\ in\ \{1,\ldots,14\}$) or bottom right (team-2 $\forall\ i\ in\ \{15,\ldots,28\}$) block of matrix M represents the interaction of the i^{th} player with the rest of his team. Similarly, any row in the off-diagonal blocks of the matrix M represents player i ($\forall\ i \in \{1,\ldots,14\}$) of team 1 losing the ball to player j ($\forall\ j \in \{15,\ldots,28\}$) of team 2 and vice versa. We introduce the average 2×2 team activity matrix T as follows:

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \tag{3}$$

where each element of matrix T represents the average activity of each block in M, as follows:

$$T_{11} = \sum_{i,j=1}^{N} M_{i,j} \tag{4}$$

$$T_{22} = \sum_{i,j=N+1}^{2N} M_{i,j} \tag{5}$$

$$T_{12} = \sum_{i=1}^{N} \sum_{j=N+1}^{2N} M_{i,j}, \tag{6}$$

$$T_{21} = \sum_{i=N+1}^{2N} \sum_{j=1}^{N} M_{i,j}, \tag{7}$$

where $N = 14$, and $\sum_{i,j=1}^{N} M_{i,j}$ represent a player's activity (team-1 $\forall\ i \in \{1,\ldots,14\}$, and team-2 $\forall\ i \in \{15,\ldots,28\}$, respectively). The overall activity (A_c) of each team in a segment is then calculated as:

$$A_{c1} = \varepsilon \times (T_{11} + T_{21} - T_{12}) \tag{8}$$

$$A_{c2} = \varepsilon \times (T_{22} + T_{12} - T_{21}) \tag{9}$$

where $\varepsilon = \sum_{i,j=1..2} T_{i,j}$ is a normalisation constant.

The total activity index (TAI) of the match is then computed as follows:

$$TAI = A_{c1} - A_{c2} \tag{10}$$

2) Information Entropy as a Measure of Performance

It has been advocated that performance analysis in a team sports should consider the dynamical nature of the match and must consider player-to-player interaction [5], [8]. The stability and consistency of interaction between different players of a team have been considered as a measure of performance in soccer matches [25]. Entropy quantifies the uncertainty coming from the random aspect of the dynamics. Entropy as a measure can be utilised to quantify the consistency of patterns representing player-to-player interaction in the match.

a: Shannon Entropy

Previously, Shannon entropy has been used as a measure of uncertainty in team sports to quantify the variability associated with the movements of players in a match [26]. In this work, we have used Shannon entropy to quantify the patterns representing player-to-player interaction during a segment of the match. Information or Shannon entropy is a measure of the uncertainty or unpredictability in the estimate of the information content of a random variable [27]. The Shannon entropy (H) is defined as follows:

$$H = -\sum_{i=1}^{N} p_i \ln(p_i), \tag{11}$$

where, p_i is the probability of the i^{th} element in the sequence.

b: Kolmogorov Complexity

As an alternative to the probabilistic notion of information content, the Kolmogorov complexity is based on the concept of recursive function [28]. Kolmogorov complexity allows the characterisation of chaotic motion in dynamical systems and the analysis of spatiotemporal patterns [28]. The Kolmogorov complexity $c(N)$ of a sequence with N samples is the length of the shortest binary program that can generate that sequence as output [28], [29]. An appropriate measure of Kolmogorov complexity can be defined by $h(N)$ as follows:

$$h(N) = \frac{c(N)}{b(N)} \tag{12}$$

where $b(N) = N \log_2 N$. In this work, Kolmogorov complexity of the signal was calculated following Kaspar et al. [28].

c: Distribution Entropy

Distribution entropy ($DistEn$) computes the complexity of a time-varying sequence using the distribution of the inter-vector distances [30]. Unlike approximate and sample entropy, $DistEn$ offers high robustness for short length sequences and reduced dependence on pre-determined parameters [30]. $DistEn$ has been previously used in many biomedical applications to quantify the complexity of short length

signals [30], [31]. In the context of soccer, *DistEn* can be used to characterise the complexity of the dynamical network patterns representing the player-to-player interaction during a segment. The *DistEn* of a vector can be defined as:

$$DistEn(m,\tau,\beta) = \frac{1}{\log_2(\beta)} \sum_{i=1}^{\beta} p_i \log_2(p_i) \tag{13}$$

where $\beta = 64$ is the number of bins in the probability distribution, obtained from the data with the lag $\tau = 1$ and embedding dimension $m = 2$. These parameter values are selected based on common recommendations from literature [30].

d: Entropy Derived Performance Indexes

To quantify the complex behaviour in which players interact during a soccer match, three entropy measures were calculated. Based on the type of the entropy three indexes were defined: (1) Shannon entropy index (*SEI*), (2) Kolmogorov complexity index (*KCI*), and (3) Distribution entropy index (*DEI*). Let $s(N)$ denote the entropy for a sequence of length N. We calculate $s(N)$ for each row in each of the four blocks of the interaction matrix M. We introduce a 2×2 matrix S to represent the team entropy/complexity matrix as follows:

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \tag{14}$$

Here, the elements of matrix S represent the averaged entropy/complexity of player-to-player interaction in each block of matrix M, as follows:

$$S_{11} = \frac{1}{N} \sum_{i=1}^{N} h(M_{i,j=1\ldots N}) \tag{15}$$

$$S_{22} = \frac{1}{N} \sum_{i=N+1}^{2N} h(M_{i,j=(N+1)\ldots 2N}) \tag{16}$$

$$S_{12} = \frac{1}{N} \sum_{i=1}^{N} h(M_{i,j=(N+1)\ldots 2N}) \tag{17}$$

$$S_{21} = \frac{1}{N} \sum_{i=N+1}^{2N} h(M_{i,j=1\ldots N}), \tag{18}$$

where $N = 14$.

The overall complexity for each team in a segment is given by:

$$s_1 = S_{11} + S_{21} - S_{12} \tag{19}$$

$$s_2 = S_{22} + S_{12} - S_{21}, \tag{20}$$

The three entropy derived indexes (*SEI*, *KCI*, and *DEI*) of a segment in the match are then computed as follows:

$$DerivedIndex = s_1 - s_2, \tag{21}$$

where the *DerivedIndex* is *SEI*, *KCI*, *DEI* for s denoting Shannon entropy, Kolmogorov complexity, and distribution entropy, respectively.

D. MACHINE LEARNING APPROACH

The possession chain data from each segment in a match was quantified using the proposed measures, which were then used as features for predicting the team that makes the "*SHOT*" during the segment. In the model training phase, the predictive model was trained using a supervised framework, where each segment ending in a "*SHOT*" was given a label "1" if team-1 makes the shot and a label "2" if the opposition makes the shot. During the testing and validation phase, the learned model was then used to predict the team making the "*SHOT*" in a segmental manner. The outcome of the game (*i.e.* team winning the match) was determined based on the classification of the segments (team-1/team-2) where the "*SHOT*" ends in a goal. For each game, we report the segmental performance and the predicted match outcome (*i.e.* winner of the match). We now describe the classifier and the learning procedure.

1) Support Vector Machine

Support vector machines (SVM) are state-of-art binary state classifiers, which are suited for pattern recognition and classification problems with good robustness to overfitting. Given an *i.i.d.* learning set $\{(x_1,y_1),(x_2,y_2),\ldots,(x_i,y_i)\}$, where $x \in \Re^N$, $y \in \{-1,1\}$, the kernel function maps the input feature space to a high-dimensional space where the data is linearly separable, offering the ability to learn non-linear functions and decision boundaries. The decision function separating the two classes is learned as a hyperplane. The optimisation problem can be formulated as:

$$\min_{\omega,b,\xi} \quad \frac{1}{2}\|\omega\|^2 + \frac{C}{n}\sum_{i=1}^{n} l(\xi) \tag{22}$$

subject to $y_i(\omega \cdot \phi(x) + b) \geq 1 - \xi_i, \forall \quad i \in 1,\ldots,n$, where C is a positive regularisation constant and ξ is the slack term.

By using the Lagrange multiplier techniques, the optimisation problem in SVM is reduced to a dual optimisation problem:

$$\max_{\alpha_k} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j K\langle x_i \cdot x_j \rangle \tag{23}$$

subject to $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $\alpha_i \in [0,C]\ \forall\ i = 1,\ldots,n$.

The learned decision function can then be represented as:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K\langle x_i, x\rangle + b\right), \tag{24}$$

where $K\langle x_i, x_j\rangle$ represents the kernel function. In this study we have used the Gaussian radial basis kernel function.

2) Learning Classification Models

The classification models were trained using a leave-one-out cross-validation approach [32]. Let N represents the total number of matches. In leave-one-out cross-validation approach, the dataset corresponding to a match is left out while the dataset from the remaining matches $(N-1)$ is used for training the SVM classifier. A feature selection using

TABLE 2: Mean, standard deviation, and area under the receiver operator characteristic curve statistics of the indexes *TAI*, *SEI*, *KCI*, and *DEI* derived using two different types of interaction matrix *M* over all matches in the dataset. The statistics show that the proposed indexes have significantly different values corresponding to the team taking the shot at the goal. A positive value of the derived indexes denotes that team-1 takes the shot, while a negative value indicates team-2, relative to whom the indexes are computed.

Measure	Interaction Matrix	Team 1 (mean ± std)	Team 2 (mean ± std)	p-value	AUC
TAI	Unit	0.09 ± 0.2	-0.21 ± 0.27	1.1×10^{-15}	0.79
	Weight	0.13 ± 0.26	-0.21 ± 0.29	5.63×10^{-15}	0.81
SEI	Unit	0.11 ± 0.2	-0.2 ± 0.28	1.51×10^{-15}	0.80
	Weight	0.1 ± 0.23	-0.2 ± 0.27	3.55×10^{-15}	0.78
KCI	Unit	0.04 ± 0.07	-0.02 ± 0.08	6.1×10^{-10}	0.73
	Weight	0.04 ± 0.07	-0.02 ± 0.07	8.1×10^{-10}	0.72
DEI	Unit	0.11 ± 0.2	-0.16 ± 0.26	3.45×10^{-15}	0.79
	Weight	0.11 ± 0.2	-0.18 ± 0.24	8.8×10^{-17}	0.81

TAI: Total activity index; *SEI*: Shannon entropy index; *KCI*: Kolmogorov complexity index; *DEI*: Distribution entropy index.
AUC: area under the receiver operator characteristics curve [34].
p-value: the p-value was calculated using two-tailed Mann Whitney U test and the statistical significance was considered for $p < 0.05$.

Lasso technique was applied on the training set for finding the least correlated and most discriminating features [33], thus ensuring the test data (left-out match) was not a part of feature selection and model learning procedure.

III. RESULTS AND DISCUSSION

In this study, we have analysed each match by segmenting into sequences that end with a "*SHOT*". The possession chain data in each segment was first mapped onto a matrix *M* representing match-integrated ball possession activity of players (Fig. 2). To calculate the estimate of complexity and non-linear dynamics in a match of soccer using the proposed coarse-grain model of teams' activity, we introduced four quantitative measures of team performance (*TAI*, *SEI*, *KCI*, and *DEI*). In addition, a machine learning approach was presented, where we developed machine learning models to predict the outcome of a segment based on the proposed quantitative measures of performance.

We first explain the quantitative measures of performance derived from the proposed coarse-grain model of player interactions network, (A) total activity index (*TAI*), (B) Shannon entropy index (*SEI*), (C) Kolmogorov complexity index (*KCI*), and (D) distribution entropy index (*DEI*), followed by (E) the performance of the proposed machine learning approach and future work.

A. TOTAL ACTIVITY INDEX (*TAI*)

The total activity index (*TAI*) is a measure of a team's activity relative to the other during a segment. Based on the definition of *TAI*, a positive value of *TAI* indicates that team-1 is likely to take the "*SHOT*" at the end of the segment, while a negative value indicates team-2 (Table 2, Fig. 3 (a)). The underlying hypothesis was that the more frequently or longer the players of a team interact during a segment, the more likely it is that this team scores in the particular segment of the match. This was further corroborated by the minimum and the maximum values of *TAI* as seen, for example, in match G_3 (Atlanta United FC (team-1) *vs*. San Jose Earthquakes (team-2), season 2018 (final result: 4-3)) that correspond to the segments when first team-2 was trying to score and then team-1 was trying to equalise by maintaining a higher possession of the ball (the segment ending at 12th and 25th minutes of the match G_3, Fig. 3 (a)). When plotted with respect to the ground truth (*i.e* the outcome of the segment *w.r.t* to the team taking the shot) the distribution of *TAI* is close to normal for both the teams (Fig. 4a, Fig. 4b). The descriptive statistics relating to the performance of *TAI* are shown in Table 2. Results showed significantly different ($p < 0.05$) means for both teams (Table 2). Although a certain overlap could be seen among the *TAI* value ranges derived using the unit and weighted increment matrices (Table 2, Fig. 4a, and Fig. 4b), the area under the receiver operator characteristics curve (AUC) values of 0.79 and 0.81 for *TAI* derived from the unit and weighted increment matrices show a good class separability.

The better performance of the weighted increment matrix shows the introduced bias towards the segment outcome ("*SHOT*") due to the higher weights given for the events that are likely to result in goal attempts in comparison to normal passes. Furthermore, the use of weights provides an alternative evaluation function that offers the opportunity to consider the types of events appearing in a pattern, and the pattern's support to determine its relevance. Finally, the *TAI* derived from the coarse-grain activity model shows good potential as a quantitative measure of performance in a team sport like soccer.

B. SHANNON ENTROPY INDEX (*SE1*)

Shannon entropy gives a measure of uncertainty to quantify the randomness associated with a time-varying signal. The Shannon entropy index (*SEI*) quantifies the underlying variability in player-to-player interaction for a team relative to the other team. The Shannon entropy of a team in a segment would be low (≈ 0) if only few players interact with each

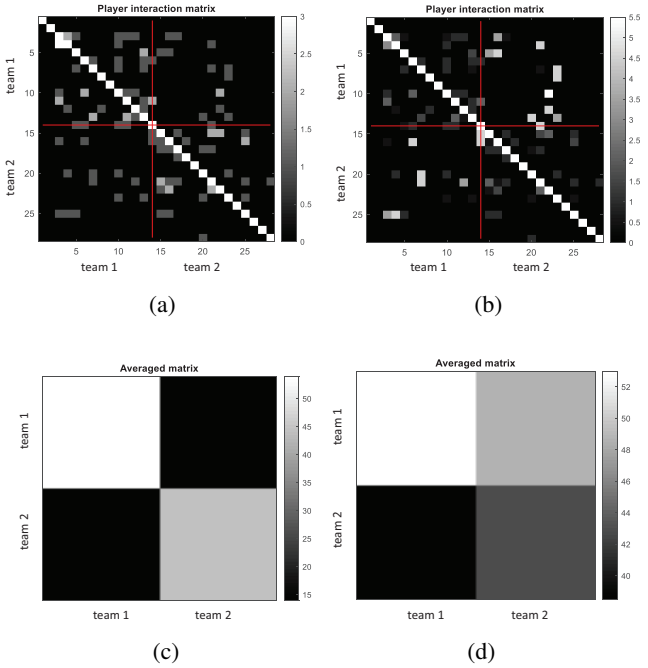


FIGURE 2: (a) Player interaction matrix M, computed for the unit increment for a segment of match G_3. (b) Same as (a) for the weighted increment. (c) Team activity matrix T, computed for the unit increment for a segment of match G_3. (d) Same as (c) for the weighted increment. The self-interaction (main diagonal of the matrix M) has been saturated in the left panel to reveal the interaction between the players. Every element $M_{i,j}$ represents the ball originator and receiver for an event. The main diagonal blocks in (a) and (b) represent the interaction between players of the same team, whereas the top-right and bottom-left corner blocks represent interactions with the players of the opposition team. The lighter the color, the higher the value of activity between the players. For the shown segment, team-1 made an offensive attack against the team-2, which is also evident in the higher activity (lighter color) of team-1 as shown in (c) and (d).

other, thus minimising the randomness and the associated unpredictability, whereas it would be high (≈ 1) if different players are continuously interacting with each other. A higher entropy indicates that there is more uncertainty in pattern-forming dynamics governing the interaction among players. Alternatively, a higher entropy represents that players are not constrained to a specific role and assume a higher tactical role (e.g. players moving both forward, backward, and through the sides of pitch, thus forging more player-to-player interactions). In team sports, a longer possession of the ball is likely to forge more player-to-player interactions especially, during a strategy leading to an offensive on the opposition more players are likely to be involved (e.g. in a match of soccer midfielders, centre forwards, wing forwards can be a part of an attack). Therefore, we hypothesised that the Shannon entropy for the team that is attacking would be higher relative

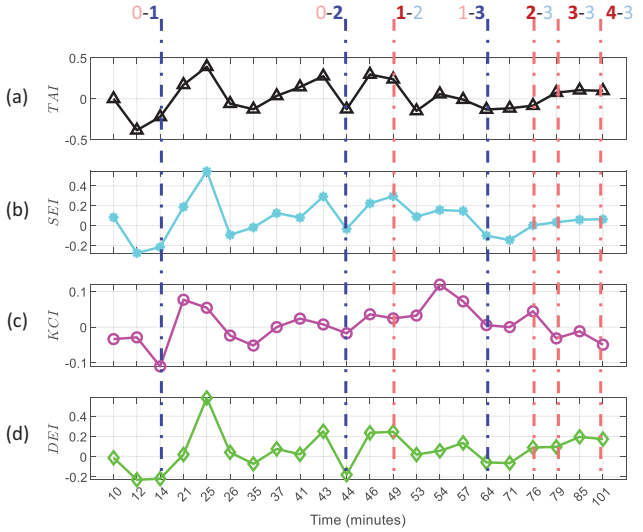


FIGURE 3: Match G_3: Atlanta United FC (team-1) *vs*. San Jose Earthquakes (team-2), season 2018 (final results: 4-3). Temporal evolution of the proposed quantitative markers of performance (a) Total activity index (*TAI*), (b) Shannon entropy index (*SEI*), (c) Kolmogorov complexity index (*KCI*), and (d) Distribution entropy index (*DEI*), derived using the weighted network of connections represented by matrix M. The vertical dashed lines indicate the moments at which a goal was scored in the match (the red (– – –) and blue (– – –) lines represent the goal scored by team-1, and team-2, respectively). Each interval on the timeline represents the time stamp of the segment ending with a "*SHOT*".

to the other. Therefore, as defined in section II-C2d, *SEI* would be > 0 if team-1 is attacking and < 0 if team-2 is attacking (Table 2, Fig. 3 (b)). The minimum and the maximum values of *SEI* denote an offensive behaviour by team-2 and team-1, respectively (segments ending at 12th and 25th minute in Fig. 3 (b)).

Thus, *SEI* can be a good marker indicating when a team makes an offensive against the opposition. The *SEI* index correlated with the segment outcomes, that is whether team-1 or team-2 takes the shot (highest AUC: 0.80, Table 2). The mean *SEI* for team-1 and team-2 were significantly different for both unit and weighted increment matrix (Table 2). A similar observation was made from the distribution when *SEI* was plotted with respect to the true outcome of the segment (Fig. 4c, and Fig. 4d). Based on the descriptive statistics (Table 2), the *SEI* index can be used as a potential marker of a team's performance derived from a coarse-grain network model representing player-to-player interaction.

C. KOLMOGOROV COMPLEXITY INDEX (*KCI*)

The use of Kolmogorov complexity was motivated by the presumption that interaction among players during a segment can be both random or synchronised (if certain players interact more frequently). Let us consider two

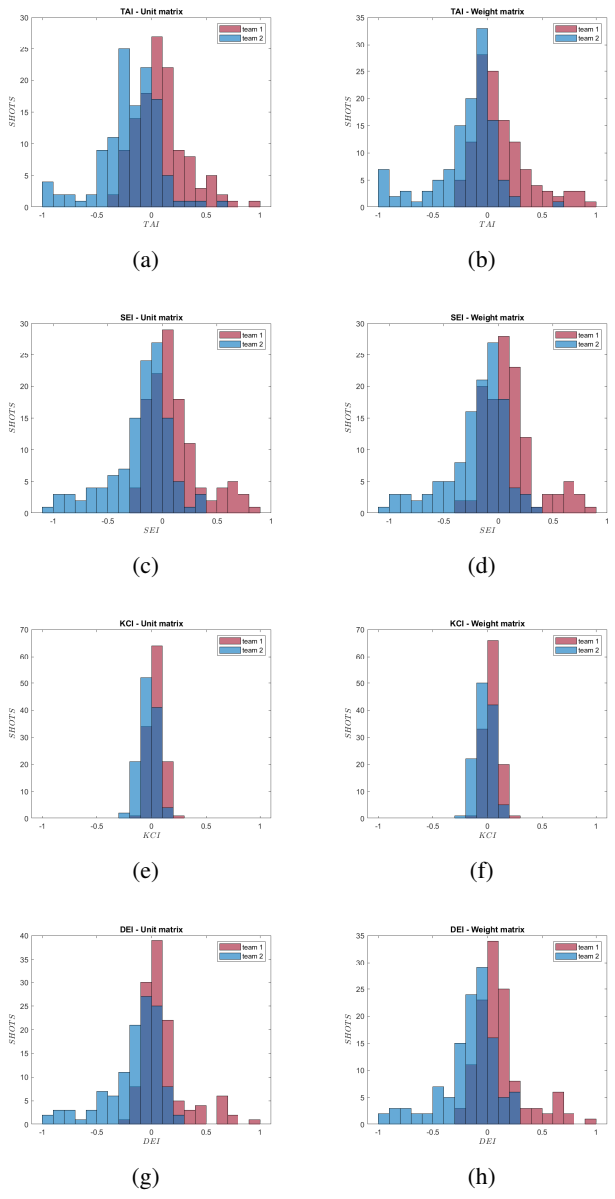


FIGURE 4: A distribution of the proposed measures of performance on "*SHOTS*" with respect to the true segmental outcome, shown for all matches in the dataset; (a-b) *TAI*, (c-d) *SEI*, (e-f) *KCI*, and (g-h) *DEI*, derived using unit and weighted network of connections represented by the interaction matrix M. The distribution of the derived quantitative measures (*TAI*, *SEI*, *KCI*, and *DEI*) was close to normal, with both teams having a significantly different means ($p < 0.05$).

vectors $s_x = \{0,0,1,0,0,1,0,0,1,0,0,1,0,0\}$, and $s_y = \{0,1,0,0,0,1,0,1,0,0,1,0,0,0\}$ of length 14 each (only a maximum of 11 players of each team were active during any segment of the match; however, a pattern length of 14 was considered as a soccer match can have a maximum of 3 substitutes.), that represents interaction pattern of players s_x and s_y. The value of the i^{th} element in vectors s_x and s_y represents the number of times the i^{th} player ($\forall i \in \{1,...,14\}$) interacted with player s_x or s_y, including any self-interaction. Both the sequences s_x and s_y have the same Shannon entropy of 2 and $DistEn$ of 0.143 ($m = 2$, $\beta = 64$), whereas both have a different Kolmogorov complexity (0.81 and 1.63, respectively). Sequence s_x has a pattern composed of units $\{0,0,1\}$ in recursion, whereas sequence s_y has no obvious pattern, thus s_y has a higher complexity. In the context of soccer, if the players are interacting in a synchronised manner, that is, few particular players are part of a strategy (offensive or defense), such patterns would be represented by simpler sequences with lower complexity or unpredictability. From a coaches point of view, it is important to assess the dynamics of pattern formation occurring in each segment of the match to decode the underlying strategy [22]. The proposed Kolmogorov complexity index derived from the player-to-player interaction network (matrix M) gives a quantitative measure of local numerical relations in which the dynamics of a teams pattern formation varies relative to the other team. For example, if certain players are only restricted to particular parts of the playing pitch as in the formation 4:4:3, the player-to-player interactions in such a segment would be represented by a less complex patterns like s_x. On the other hand, if a team allocates more players in sub-segments of a match to prevent opposition's attacking move (*i.e.*, a defensive strategy) or to create an offensive move at opposition's goal, the player-to-player interactions would be represented by more complex patterns without any recursive sub-patterns as shown by s_y. Therefore, Kolmogorov complexity derived (*KCI*) index captures the complexity of patterns that is different from Shannon entropy derived index or *SEI*. *KCI* showed a good correlation when plotted with the segmental outcome of a match (AUC (0.73), Table 2). A *KCI* value > 0 favoured team-1, while a $KCI < 0$ indicates a shot taken by team-2 (Fig. 3 (c)). For match G_3, the segment ending at the 54th minute represents a case when team-1 is making an offensive against the opposition to level the scores at $2-2$, which is shown by the maximum value of *KCI* at the 54th minute (Fig. 3 (c)). The *KCI* index followed a distribution close to normal, when plotted on the true segment outcome, that is, with respect to the team taking the "*SHOT*" (Fig. 4e, and 4f), with a significantly different mean values for both the teams (Table 2). *KCI* is a measure to quantify the regularity of complex patterns in which players interact during team sports. It gives a numerical relation in which the dynamics of a team's pattern formation varies over the segments of a match. *KCI* can allow coaches to discover, identify and quantify segments during a match, when a team interacts in more complex or rather synchronised patterns.

D. DISTRIBUTION ENTROPY INDEX (*DEI*)

The distribution entropy (*DistEn*) measures the complexity of patterns governing player-to-player interactions by taking into account the hidden information in the state-space *via* estimating the probability density of inter-vector distances. A chaotic sequence has the maximum *DistEn*, thus patterns of player-to-player interaction with high variability would be characterised by a high *DistEn* (≈ 1) and *vice versa*. The distribution entropy derived index (*DEI*) quantifies the chaotic patterns underlying player-to-player network of interaction for a team relative to the other. Similar to Shannon entropy-derived index, a $DEI > 0$ would indicate a higher variability (associated with an attacking move) in patterns governing player-to-player interaction for team-1, when computed relative to team-2 (Fig. 3 (d)). However, a particular advantage of *DEI* over *SEI* is that *SEI* can be affected by the variance of the sequences representing player interaction patterns, while *DEI* is derived using a probability density function with fixed bin number (β). Thus, *DEI* is more robust as it considers inter-vector distances. Let us consider two vectors $s_x = \{0,0,1,0,0,0,1,1,1,1,3,1,1,0\}$, and $s_y = \{0,0,1,0,0,1,0,2,0,0,5,0,1,0\}$ that represent the interaction patterns of player x and player y during a segment of the match, and having the same number of total interactions ($\sum s_x = \sum s_y = 10$). The 5 in s_y represents the number of times the 11th player of the team interacted with player y during the segment. The Shannon entropy for s_x, and s_y is 2.84, and 1.96, respectively, whereas *DistEn* is 0.27, and 0.36, respectively. Pattern s_y represents that player y interacts more frequently (5 times) with the 11^{th} player, which results in high inter-vector distances thus leading to a higher *DistEn* value. The interaction pattern represented by s_y indicates the events, when player y is continuously interacting with a particular player (11th player in the team). This pattern might signify an underlying strategy where the players' (defender/midfielder) are interacting with a particular player (forward) as a part of a strategy to generate a scoring opportunity. *DistEn* provides an ability to encode such patterns, thus distribution entropy-derived index (*DEI*) can be a good marker to characterise the complexity of player-to-player interaction patterns, such that the underlying strategy can be quantified as a measure of a team's performance during a segment of the match. When plotted with respect to the outcome of the segment, the *DEI* values were normally distributed (Fig. 4g, and 4h). The mean *DEI* values for the teams in the adversarial relationship were significantly different (Table 2), and a good class separability was achieved with an AUC of 0.79, and 0.81 for *DEI* derived using unit and weighted interaction network of players.

E. PERFORMANCE OF THE MACHINE LEARNING APPROACH

We developed an automated machine learning model to predict the outcome of a match segment using the proposed measures of performance quantification (*TAI*, *SEI*, *KCI*, and *DEI*). Our machine learning approach showed a mean sensitivity of 78.3% (95% confidence interval (CI): 70.3% - 85.3%), a specificity of 73.8% (95% CI: 69% - 80.2%) and an overall accuracy of 75.2% in predicting the segmental outcomes of the matches. Although our dataset comprised of only 13 matches, it should be noted that we performed a segmental analysis on segments of different duration (segments ending with "*SHOT*"), resulting in a sizeable number of samples (241 temporal segments) for training and validating the machine learning classifier. In addition, our approach is based on a robust cross-validation approach that ensures no bias of the learned model to the ground truth.

The predicted segmental outcomes for all the matches in the database are shown in Table 3. One match G_{13} ended in draw. Among the rest, the predicted outcome correlated with the ground truth (*i.e.*, the winner of the match) in 8 (66%) of 12 matches. Furthermore, the application of the automated segmental analysis is not limited to the overall match outcome. It also helps to analyse the underlying local prediction statistics. The outcome of our developed prediction model on a complete match is shown in Fig. 5. The prediction models give the segmental likelihood of an attempt to goal for both the teams (Fig. 5 (a)). In the particular match shown in Fig. 5, team-1 (shown in red) won the match 4-3. The segments where a goal was scored are marked with $*$ (segments 3, 11, 13, 17, 19, 20, and 22 shown in Fig. 5). It can be seen in Fig. 5 that segments where team-1 has scored a goal (segments 13, 19, 20, and 22) have a higher likelihood. Similarly, the likelihood for team-2 is higher in the segments where they scored a goal. Segments 9, 14, 18, and 19 show the where both the teams are engaged in gaining the possession of the ball as they want to equalise. The segments where the predicted outcome (Fig. 5 (b)) did not match the true outcome (Fig. 5 (c)) are the ones, where the possession of the ball is continuously changing between the teams. To quantify the minority of segments, where the model does not provide a sufficient agreement with the ground truth data, in future we would incorporate more sophisticated measures by introducing player labels (forwards, mid-fielders, defence) to understand player-to-player interaction using concepts of mutual information retrieval [36].

F. VALIDATION AND COMPARISONS ON PUBLIC DATASET

To elaborate on the efficacy of the proposed approach a thorough performance evaluation was carried out on the largest available public dataset of soccer logs [35]. This dataset comprised event logs (possession chain data) from 1,941 matches of 7 major competitions (Table 4). The proposed machine learning approach showed an overall sensitivity of 83.5%, a specificity of 83.4%, F_1 score of 83.7%, and an AUC of 0.84 in classifying a total of 42,860 segments ending in "*SHOT*" (*i.e.*, whether team-1 or team-2 makes the "*SHOT*" at the goal) (Table 4). The match outcome (segments leading to a goal) correctly correlated in 1,202 (81.9%) of 1,467 matches that ended in a result. The performance measures with the 95% CI (calculated over all matches in a competition) for each competition are also reported (Table 4). The European

TABLE 3: The predictive performance of our developed machine learning models. Shown are the segments predicted in favour of a team with the overall prediction accuracy, the predicted winner, and the true match results.

Team	Segments won - team 1 (Predicted)	Segments won - team 2 (Predicted)	Classification Accuracy	Predicted Winner*	Match Winner
G_1	5	7	0.75	team 2	team 2
G_2	10	11	0.62	team 2	team 1
G_3	13	9	0.77	team 1	team 1
G_4	7	7	0.71	team 1	team 1
G_5	12	7	0.78	team 1	team 1
G_6	8	12	0.70	team 2	team 1
G_7	11	14	0.72	team 2	team 2
G_8	9	5	0.71	team 1	team 1
G_9	17	10	0.74	team 1	team 1
G_{10}	6	15	0.71	team 2	team 1
G_{11}	8	5	0.69	team 1	team 2
G_{12}	7	7	0.71	team 1	team 1
G_{13}	8	11	0.84	team 2	DRAW

* For all matches ending in a result, the predicted outcome of the segments ending in goal was used to decide the predicted winner of the match. If both the teams have the same number of predicted winning segments, the team with the higher possession time was considered the winner.

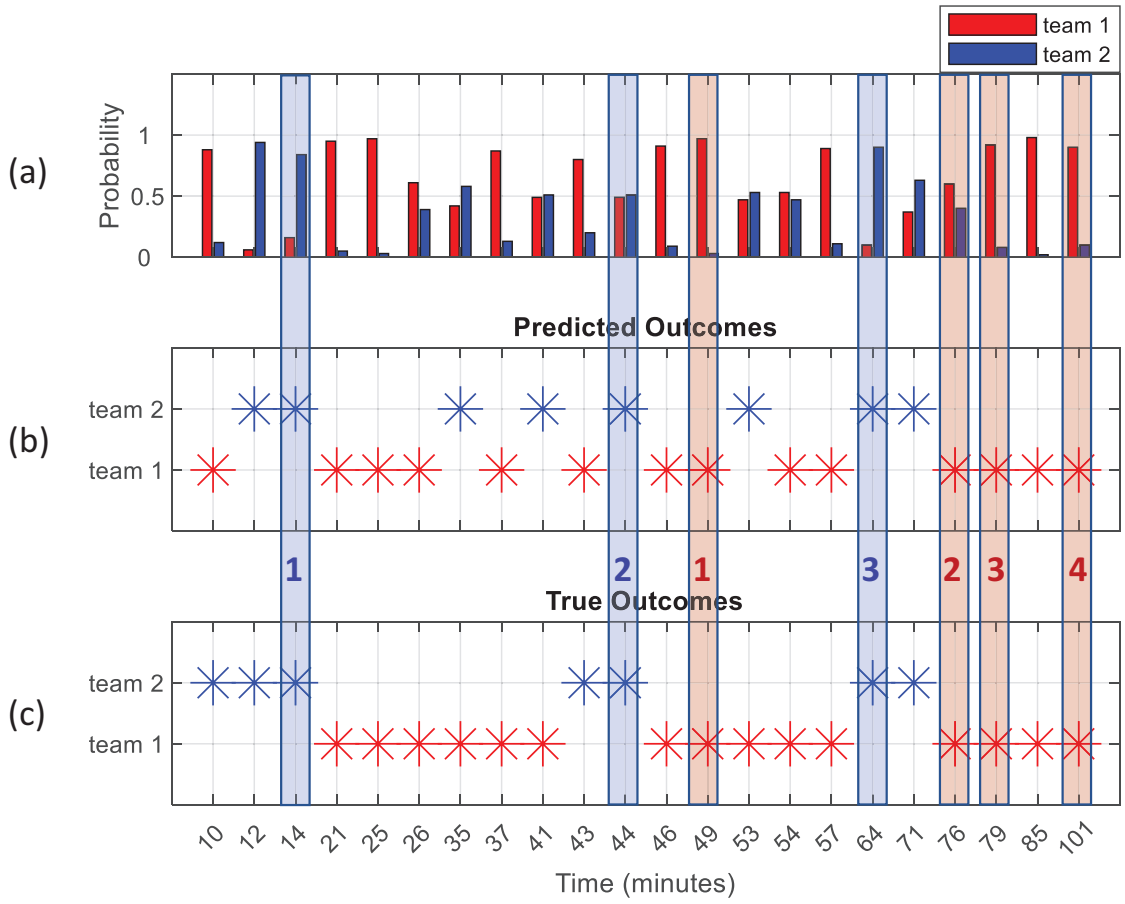


FIGURE 5: Match G_3: Atlanta United FC ($team-1$) *vs*. San Jose Earthquakes ($team-2$), season 2018 (final results: 4-3). The segmental analysis for match G_3, showing the predicted outcomes for every segment compared to the true outcomes; (a) segmental likelihoods, (b) predicted outcomes, and (c) the true match outcome, which is the team taking the "*SHOT*" at the oppositions goal at the end of the segment. The vertical bars on the match timeline show the segments where a team scored a goal. Red bars indicate team-1 and blue bars indicate goals scored by team-2. The intervals on the timeline indicate the time stamp corresponding to the segments ending with a "*SHOT*".

TABLE 4: Validation of the proposed approach on the largest open collection of soccer logs from 7 major competitions [35]. Shown are the total number of matches, matches ending in result, the number of analysed segments, mean performance measures with 95% [CI: confidence interval] over all matches of each competition, and the accuracy depicting the percentage of matches where the predicted winner correlated with the true match outcome.

Competition	Total Matches	Matches with Results	Segments Ending in "*SHOT*"	Statistical measures of segmental performance				Correctly Predicted Match Outcome†
				Sensitivity† (95% [CI])	Specifivity† (95% [CI])	F_1 score† (95% [CI])	AUC† (95% [CI])	
European Cup 2016	51	40	1195	77.4% (71.9%-79.6%)	85.8% (82.9%-89.4%)	82.5% (77.7%-85.3%)	0.82 (0.79-0.84)	72.5% (29/40)
French First Division	380	282	8277	85.5% (83.9%-87.0%)	81.0% (78.9%-82.8%)	83.4% (82.0%-84.6%)	0.83 (0.82-0.84)	80.4% (227/282)
German First Division	306	221	6869	84.7% (82.6%-86.2%)	83.2% (81.0%-84.9%)	82.8% (81.3%-84.0%)	0.84 (0.83-0.85)	81.9% (181/221)
Italian First Division	380	297	8758	87.1% (85.5%-88.4%)	82.3% (80.6%-84.0%)	84.7% (83.5%-85.9%)	0.85 (0.84-0.86)	85.5% (254/297)
Spanish First Division	380	293	7933	84.7% (83.0%-86.2%)	82.9% (80.9%-84.4%)	82.9% (81.7%-84.1%)	0.84 (0.83-0.85)	78.1% (232/293)
World Cup 2018	64	55	1413	85.9% (82.5%-88.2%)	81.6% (77.6%-84.5%)	85.1% (82.5%-86.9%)	0.83 (0.81-0.85)	70.1% (39/55)
English First Division	380	279	8415	79.4% (77.4%-81.3%)	87.3% (85.4%-88.5%)	84.7% (83.4%-85.8%)	0.83 (0.82-0.84)	86.1% (240/279)
Overall	1941	1467	42860	83.5%	83.4%	83.7%	0.84	81.9% (1202/1467)

† Sensitivity: the percentage of segments correctly classified for team-1 in each match; Specificity: the percentage of segments correctly classified as belonging to team-2 in each match; F_1 score: the harmonic mean of precision and recall; 95% [CI]: AUC: the area under the ROC curve for the segmental analysis of each match [34]; 95% confidence interval [CI] computed for measures of segmental performance over all matches in each competition; Correctly Predicted Match Outcome: the percentage of matches where the predicted winner of the match correlated with the true match outcome.

cup 2016, and the Italian first division had the lowest and the highest AUC for segmental analysis among the 7 competitions. Overall the AUC of the segmental performance was close to the overall AUC for each of the 7 competitions, which shows the consistent performance of the proposed approach across the competitions. The Italian first division had the highest number of segments ending in "*SHOT*" (8,758) from the 380 matches. The proposed machine learning approach resulted in an AUC of 0.85 (95% [CI]: 0.84 - 0.86) in correctly classifying the segments. Furthermore, the predicted match winner correlated with the ground truth in 254 (85.5%) of the 297 matches of the Italian first division that ended in a result. The overall performance of the proposed approach on 1,941 matches and 42,860 segments shows the efficacy of the proposed quantitative markers (*TAI*, *SEI*, *KCI*, and *DEI*) of a team's performance.

Furthermore, to elaborate the efficacy of the proposed quantifiable markers of team performance, we compared the results of the proposed approach with studies that employ a machine learning approach for evaluating performance [15], [20], [21]. A direct comparison of the proposed segmental analysis approach can be done with the study by Decross et al. [21]. They employed a segmental analysis to learn the importance of players actions based on the outcome of a match state (*e.g.* success in taking a "*SHOT*" at opponents goal). On the contrary, Pappalardo et al. [20] defined a feature vector for each team and modelled the outcome of the match (Win/Loss) using a linear support vector machines classifier. As both the studies [20], [21] use different datasets, therefore, to ensure a direct comparison of the proposed approach the algorithms by Decross et al. [21], and Pappalardo et al. [20] are run on the soccer logs from 1,941 matches of 7 competitions [35]. The model estimation and the learning task was performed using a leave-one-out cross-validation approach as explained in section II-D. Additionally, the results on the public dataset were compared with the study by Cintia et al. [15], who analysed the match outcome using pass-based performance indicator (H-indicator) and evaluated the performance on the German, Spanish, Italian, and English division leagues.

For a comparison of the segmental performance, the algorithm by Decross et al. [21] was used to model the segments ending in a "*SHOT*" with an XGBoost classifier [37]. The algorithm by Decross et al. [21] showed an overall AUC of 0.83 (Table 5). In comparison, the proposed approach showed a similar performance with an overall AUC of 0.84 on 42,860 analysed segments (Table 5). Further, for a comparison of the correctly predicted match outcome, the algorithm by Pappalardo et al. [20] was employed. The algorithm by Pappalardo et al. [20] could correctly classify the match outcome in 1176 (77.5%) of 1467 matches that ended in a result among the 7 competitions (Table 5). In comparison, the proposed approach could correctly classify the match-winner in 1,202 of 1,467 matches. Furthermore in comparison to the study by Cintia et al. [15] who reported a mean accuracy 0.55 in correctly predicting the match outcome, the proposed approach showed a higher mean accuracy of 0.82 (German: 0.81, Spanish: 0.78, Italian: 0.85, and English division: 0.86) in correctly predicting the match outcome. The improved performance of the proposed approach shows the robustness and

TABLE 5: Comparisons of the proposed approach with some recent machine learning-based studies.

Competition	Segmental Performance (AUC)†		Correctly Predicted Match Outcome†	
	Proposed Approach	Decross *et al.* [21]	Proposed Approach	Pappalardo *et al.* [20]
European Cup 2016	0.82 (0.79-0.84)	0.81 (0.78-0.85)	72.5% (29/40)	67.5% (27/40)
French First Division	0.83 (0.82-0.8)	0.83 (0.82-0.85)	80.4% (227/282)	81.5% (230/282)
German First Division	0.84 (0.83-0.85)	0.83 (0.81-0.85)	81.9% (181/221)	80.5% (178/221)
Italian First Division	0.85 (0.84-0.86)	0.85 (0.83-0.86)	85.5% (254/297)	79.4% (236/297)
Spanish First Division	0.84 (0.83-0.85)	0.83 (0.82-0.85)	78.1% (232/293)	81.5% (239/293)
World Cup 2018	0.83 (0.81-0.85)	0.82 (0.77-0.89)	70.1% (39/55)	70.1% (39/55)
English First Division	0.83 (0.82-0.84)	0.84 (0.83-0.85)	86.1% (240/279)	81.3% (227/279)
Overall	0.84	0.83	81.9% (1202/1467)	77.5% (1176/1467)

† AUC: the area under the ROC curve for the segmental analysis of each match [34]; Correctly Predicted Match Outcome: the percentage of matches where the predicted winner of the match correlated with the true match outcome.

efficiency of the proposed quantitative markers (*TAI*, *SEI*, *KCI*, and *DEI*) in capturing a team's underlying performance characteristics (Table 5).

The performance of the proposed approach can be attributed to the use of kernelised classifier and the non-linearity of the proposed indices like *SEI*, *KCI*, and *DEI* that can quantify the underlying non-linear dynamics of player interaction. Based on the performance validation on external dataset and comparison with recent studies, it can be concluded that the proposed approach offers a data-driven framework for evaluating a team's performance in a segmental manner, offering the potential for predictive analytics in sport sciences using data science research.

G. INTERPRETABILITY FOR SPORTS ANALYSTS

Our analysis shows that the interaction between players is essential for generating a scoring opportunity. To outline the applicability of the proposed features, we use histogram plots representing each player's feature values that are derived from the player interaction matrix M for a segment of the match (Fig. 6).

The histograms illustrate the level of interaction of each player, when their team has possession of the ball. The players that are more frequently involved in the ball possession have feature values above the team's mean value, which is represented by the dashed line $(--)$ as shown in Fig. 6. Across this match segment, team-1 performs better than team-2, because the segment ends with team-1 having a successful attempt at scoring *i.e.*, a "*SHOT*" at goal. Six players of team-1 (P_2, P_5, P_8, P_9, P_{10}, and P_{11}) maintain a level of interaction (as indicated by the rectangular box in Fig. 6a) above the team's mean, which is higher than team-2, where only three players are above the team's mean (as indicated by the rectangular box in Fig. 6b).

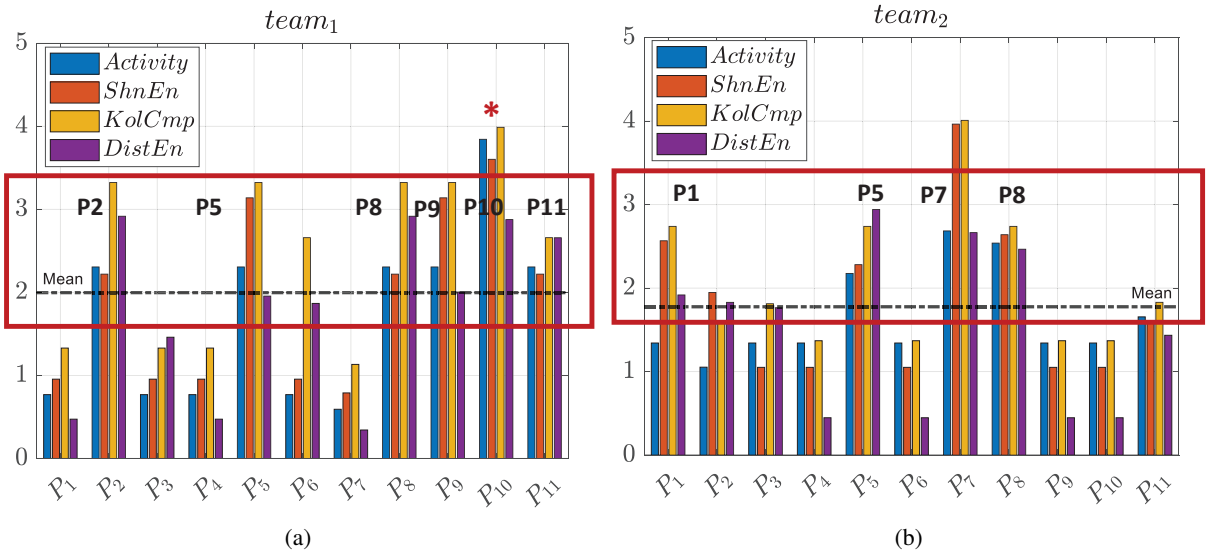
The feature *Activity* shows the players that are more frequently involved in a team's ball possession activity. The remaining features (*ShnEn*, *KolCmp*, and *DistEn*) were also above average for more players in team-1 than for team-2. Sports analysts can interpret this as an association between the complexity of passing between players and the likelihood of having a shot at goal. In other words, when a team has possession of the ball, there may be a benefit in making a relatively large number of passes between a large proportion of the team, as they move the ball towards their opponent's goal post.

The usual analysis of an opponent's tactics is a resource-intensive procedure, as most tactical analyses are performed by manually reviewing the match videos or scouting matches in-person to identify the players that are constantly part of the ball possession activity and are involved in generating scoring opportunities [24]. The features used in the present analysis may enable the automatic identification of such players using a data-driven approach. For example, player P_{10} in team-1 is one such player who had the highest ball possession activity during the shown segment of the match (indicated by an $*$ in Fig. 6a). Identifying the players that are more frequently involved in match states that end with an attempt at scoring *i.e.*, a "*SHOT*" at goal, may assist sports analysts and team staff to develop strategies suited to an opponent's playing style.

The proposed study presents different characteristics of a team's performance during a segment of a match that ends with a "*SHOT*" on the goal. Although, there are different ways to define match segments (*e.g.* a segment ending with the ball going out, a foul etc.) the purpose of the study was to identify the characteristics leading to an attempt at scoring a goal. Therefore, in this study, we analysed segments ending with a "*SHOT*" on the goal, which is also a limitation of the study. Furthermore, the influence of match location, quality of opposition, match type *etc*. were not controlled for while developing the predictive models. Thus, further research is required to investigate the effects of these variables to further enhance the understanding of teams and players performances.

IV. CONCLUSION

Our study proposes information theory-derived quantifiable measures of performance that can uncover the dynamic patterns underlying team sports like soccer. The study provides first evidence of a machine learning-enabled approach for



(a) (b)

FIGURE 6: The feature histograms showing players activity, Shannon entropy (*ShnEn*), Kolmogorov complexity (*KolCmp*), and Distribution entropy (*DistEn*) derived from the player interaction matrix M. The derived parameters are normalised to ensure feature commensurablity. Shown here for a segment of match G_3 (a) team-1, and (b) team-2. The dotted horizontal lines in subplots (a), and (b) represent the mean across all the players of a team. The rectangular box in histogram plot for team-1 (subplot (a)) indicates the players (P_2, P_5, P_8, P_9, P_{10}, and P_{11}) who maintain a ball possession activity that is higher than the team's mean. Player P_{10}, who makes the shot at goal had the highest interaction with the rest of the players during team-1 ball possession activity. In contrast, the ball possession activity of team-2 is mainly among players P_1, P_5, P_7, and P_8.

automated predictive analysis of performance in a segmental manner, offering the potential for uncovering local numerical markers of team performance. Our developed predictive models show a mean accuracy of 75.2% in predicting the segmental outcome of the likelihood of team making a successful attempt to score a goal on our dataset comprising 13 matches. In addition, the segmental outcomes could predict the correct overall winner in 66.6% of the matches that resulted in a winner. Furthermore, the validation on an external dataset comprising 42,860 segments from 1,941 matches showed the robustness of the approach. Finally, the study demonstrates that the analysis we present can help uncover the pattern dynamics of a team's network derived using possession chain data, by quantitatively analysing measures of performance that have a specific distribution and that can be used to predict the performance of a team.

ACKNOWLEDGMENT

The authors thank Dr Alexander Kalloniatis and Dr Tim Wilkin for the useful and constructive comments during the development of the material for this paper. The authors acknowledge the funding from the Defence Science Institute, Australia.

AUTHOR CONTRIBUTIONS STATEMENT

S.K. performed the analysis and wrote the manuscript. S.S., Y.Z., and M.A. contributed to the analysis of the results. D.D. provided the dataset. D.D. and P.G. helped in formulating the study's significance from a sports perspective. All authors contributed to and reviewed the manuscript.

REFERENCES

[1] T. McGarry, "Applied and theoretical perspectives of performance analysis in sport: Scientific issues and challenges," International Journal of Performance Analysis in Sport, vol. 9, no. 1, pp. 128–140, 2009.

[2] I. Franks, "Use of feedback by coaches and players," Science and football III, pp. 267–278, 1997.

[3] C. Wright, C. Carling, and D. Collins, "The wider context of performance analysis and it application in the football coaching process," International Journal of Performance Analysis in Sport, vol. 14, no. 3, pp. 709–733, 2014.

[4] D. Araujo, K. Davids, and R. Hristovski, "The ecological dynamics of decision making in sport," Psychology of sport and exercise, vol. 7, no. 6, pp. 653–676, 2006.

[5] P. S. Glazier, "Game, set and match? substantive issues and future directions in performance analysis," Sports medicine, vol. 40, no. 8, pp. 625–634, 2010.

[6] L. Vilar, D. Araújo, K. Davids, and C. Button, "The role of ecological dynamics in analysing performance in team sports," Sports Medicine, vol. 42, no. 1, pp. 1–10, 2012.

[7] K. Davids, D. Araújo, and R. Shuttleworth, "Applications of dynamical systems theory to football," Science and football V, vol. 537, p. 550, 2005.

[8] B. Travassos, K. Davids, D. Araújo, and T. P. Esteves, "Performance analysis in team sports: Advances from an ecological dynamics approach," International Journal of Performance Analysis in Sport, vol. 13, no. 1, pp. 83–95, 2013.

[9] H. Folgado, K. A. Lemmink, W. Frencken, and J. Sampaio, "Length, width and centroid distance as measures of teams tactical performance in youth football," European Journal of Sport Science, vol. 14, no. sup1, pp. S487–S492, 2014.

[10] P. Passos, K. Davids, D. Araújo, N. Paz, J. Minguéns, and J. Mendes, "Networks as a novel tool for studying team ball sports as complex social systems," Journal of Science and Medicine in Sport, vol. 14, no. 2, pp. 170–176, 2011.

[11] C. Braham and M. Small, "Complex networks untangle competitive ad-

vantage in australian football," Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 28, no. 5, p. 053105, 2018.
[12] J. M. Buldú, J. Busquets, J. H. Martínez, J. L. Herrera-Diestra, I. Echegoyen, J. Galeano, and J. Luque, "Using network science to analyse football passing networks: dynamics, space, time and the multilayer nature of the game." Frontiers in psychology, vol. 9, p. 1900, 2018.
[13] J. Ramos, R. J. Lopes, and D. Araújo, "WhatâĂŹs next in complex networks? capturing the concept of attacking play in invasive team sports," Sports medicine, vol. 48, no. 1, pp. 17–28, 2018.
[14] L. Gyarmati and X. Anguera, "Automatic extraction of the passing strategies of soccer teams," arXiv preprint arXiv:1508.02171, 2015.
[15] P. Cintia, F. Giannotti, L. Pappalardo, D. Pedreschi, and M. Malvaldi, "The harsh rule of the goals: Data-driven performance indicators for football teams," in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2015, pp. 1–10.
[16] J. Duch, J. S. Waitzman, and L. A. N. Amaral, "Quantifying the performance of individual players in a team activity," PloS one, vol. 5, no. 6, p. e10937, 2010.
[17] R. Hristovski, K. Davids, D. Araujo, and P. Passos, "Constraints-induced emergence of functional novelty in complex neurobiological systems: a basis for creativity in sport," Nonlinear Dynamics-Psychology and Life Sciences, vol. 15, no. 2, p. 175, 2011.
[18] J. Buldú, J. Busquets, I. Echegoyen et al., "Defining a historic football team: Using network science to analyze guardiolaâĂŹs fc barcelona," Scientific reports, vol. 9, no. 1, pp. 1–14, 2019.
[19] P. Silva, R. Duarte, P. Esteves, B. Travassos, and L. Vilar, "Application of entropy measures to analysis of performance in team sports," International Journal of Performance Analysis in Sport, vol. 16, no. 2, pp. 753–768, 2016.
[20] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, "Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 5, pp. 1–27, 2019.
[21] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1851–1861.
[22] L. Vilar, D. Araújo, K. Davids, and Y. Bar-Yam, "Science of winning soccer: Emergent pattern-forming dynamics in association football," Journal of systems science and complexity, vol. 26, no. 1, pp. 73–84, 2013.
[23] R. Duarte, D. Araújo, H. Folgado, P. Esteves, P. Marques, and K. Davids, "Capturing complex, non-linear team behaviours during competitive football performance," Journal of Systems Science and Complexity, vol. 26, no. 1, pp. 62–72, 2013.
[24] T. Decroos, J. Van Haaren, and J. Davis, "Automatic discovery of tactics in spatio-temporal soccer match data," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 223–232.
[25] K. Davids, C. Button, D. Araújo, I. Renshaw, and R. Hristovski, "Movement models from sports provide representative task constraints for studying adaptive behavior in human movement systems," Adaptive behavior, vol. 14, no. 1, pp. 73–95, 2006.
[26] P. Silva, R. Duarte, J. Sampaio, P. Aguiar, K. Davids, D. Araújo, and J. Garganta, "Field dimension and skill level constrain team tactical behaviours in small-sided and conditioned games in football," Journal of sports sciences, vol. 32, no. 20, pp. 1888–1896, 2014.
[27] C. E. Shannon, "A mathematical theory of communication," Bell system technical journal, vol. 27, no. 3, pp. 379–423, 1948.
[28] F. Kaspar and H. Schuster, "Easily calculable measure for the complexity of spatiotemporal patterns," Physical Review A, vol. 36, no. 2, p. 842, 1987.
[29] A. Lempel and J. Ziv, "On the complexity of finite sequences," IEEE Transactions on information theory, vol. 22, no. 1, pp. 75–81, 1976.
[30] P. Li, C. Liu, K. Li, D. Zheng, C. Liu, and Y. Hou, "Assessing the complexity of short-term heartbeat interval series by distribution entropy," Medical & biological engineering & computing, vol. 53, no. 1, pp. 77–87, 2015.
[31] P. Li, C. Karmakar, J. Yearwood, S. Venkatesh, M. Palaniswami, and C. Liu, "Detection of epileptic seizure based on entropy analysis of short-term eeg," PloS one, vol. 13, no. 3, p. e0193691, 2018.
[32] G. C. Cawley, "Leave-one-out cross-validation based model selection criteria for weighted ls-svms," in The 2006 ieee international joint conference on neural network proceedings. IEEE, 2006, pp. 1661–1668.
[33] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.
[34] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases." Radiology, vol. 148, no. 3, pp. 839–843, 1983.
[35] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," Scientific data, vol. 6, no. 1, pp. 1–15, 2019.
[36] D. R. Brillinger, "Some data analyses using mutual information," Brazilian Journal of Probability and Statistics, pp. 163–182, 2004.
[37] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

SHITANSHU KUSMAKAR received the Masters degree in clinical engineering from IIT Madras, India, in 2012, and the Ph.D. degree in Electrical and Electronic Engineering from The University of Melbourne, Australia, in 2019. He was a Melbourne India Postgraduate Scholar at University of Melbourne. His doctoral research focused on developing an artificial intelligence (AI)-based system for automated detection of epileptic seizures using a wearable sensing device. He is currently a research fellow in complex data analytics at School of Information Technology, Deakin University, Australia. His research interests include artificial intelligence, machine learning, cognitive computing, health-AI, and time-series analysis.

SERGIY SHEYLAG received his Ph.D. degree from the University of Göttingen, Germany, in 2004. He worked as a research fellow with the Department of Applied Mathematics of the University of Sheffield (UK), with the Department of Mathematics and Physics of Queen's University, Belfast (UK), was a Future Fellow with the Department of Mathematical Sciences of Monash University (Australia) and worked as a senior lecturer at the University of Northumbria, Newcastle (UK). Currently, Dr. Shelyag is a senior lecturer in the School of Information Technology, Deakin University, Victoria, Australia. His current research interests are in mathematical and computational modelling of complex physical processes, data analytics, machine learning and dynamical systems.

YE ZHU received his Ph.D. degree in Artificial Intelligence with a Mollie Holman Medal for the best doctoral thesis of the year from Monash University (Australia) in 2017. He is a lecturer with the School of Information Technology, Deakin University, Australia since 2019. He joined Deakin University in 2017 as a research fellow of complex system data analytics. His research works focus on clustering analysis, anomaly detection, and their applications for pattern recognition and information retrieval.

DAN DWYER received his Ph.D degree from Griffith University in Queensland. He has worked as a research active-academic at the University of Tasmania and the University of Newcastle. He has also worked as a Sport Scientist at the Victorian Institute of Sport in Melbourne. He is currently a Senior Lecturer in Applied Sport Science at Deakin University and a member of the Centre for Sport Research. His research focusses on the measurement, analytics and prediction in sport.

PAUL GASTIN is Professor and Head of Sport and Exercise Science at La Trobe University. His teaching and research focuses on innovation in sport science and coaching to enhance the performance of people and organisations across the sport participation spectrum. Paul has worked in Olympic/Paralympic and professional sport in Australia and overseas holding senior positions at the Victorian Institute of Sport, the UK Sports Institute and UK Sport. He is an ESSA Fellow and an accredited level 2 Sport Scientist and High Performance Manager.

MAIA ANGELOVA received her Ph.D. from the University of Sofia. Dr Angelova is currently a Professor of Data Analytics and Machine Learning in the School of Information Technology, Deakin University, Victoria, Australia. Prior to that, she was a Lecturer in Physics in Somerville College, University of Oxford and Professor of Mathematical Physics at Northumbria University, United Kingdom until 2016. She is the director of Data to Intelligence Research Centre and leads Data Analytics Research Lab at Deakin University. Her research interests include mathematical modelling, data analytics, time series, machine learning, dynamical systems with applications to health.