Predicting Peek Readiness-to-Train of Soccer Players Using Long Short-Term Memory Recurrent Neural Networks

Theodor Wiik[1], Håvard D. Johansen[2], Svein-Arne Pettersen[2], Ivan Baptista[2,3], Tomas Kupka[6], Dag Johansen[2], Michael Riegler[4], and Pål Halvorsen[4,5,6]
[1]*Simula Research Laboratory*, Norway [2]*UIT - The Arctic University of Norway* [3]*Tromsø Idrettslag*, Norway
[4]*SimulaMet*, Norway [5]*Oslo Metropolitan University*, Norway [6]*ForzaSys AS*, Norway

***Abstract*—We are witnessing the emergence of a myriad of hardware and software systems that quantifies sport and physical activities. These are frequently touted as game changers and important for future sport developments. The vast amount of generated data is often visualized in graphs and dashboards, for use by coaches and other sports professionals to make decisions on training and match strategies. Modern machine-learning methods has the potential to further fuel this process by deriving useful insights that are not easily observable in the raw data streams.**

This paper tackles the problem of deriving peaks in soccer players' ability to perform from subjective self-reported wellness data collected using the PMSys system. For this, we train a long short-term memory recurrent neural network model using data from two professional Norwegian soccer teams. We show that our model can predict performance peaks in most scenarios with a precision and recall of at least 90%. Equipped with such insight, coaches and trainers can better plan individual and team training sessions, and perhaps avoid over training and injuries.

***Index Terms*—Machine learning, interdisciplinary sport application, medical documentation, performance prediction**

I. Introduction

International sport is undergoing a revolution, fueled by the rapidly increasing availability of athlete quantification data, sensor technology, and advanced analytic software. Algorithmic analysis of this data might provide vital insights for individual training personalization and injury prevention. Key sport governance organizations like Fédération Internationale de Football Association (FIFA) have approved certain wearables and electronic performance and tracking systems in official football matches, providing a foundation for evidence-based decisions and team performance improvements [1, 2]. In Brazil 2014, the German national soccer team used wearable technology to profile the players, and with these statistics it is believed that coach Joachim Löw made the crucial substitute of Mario Götze who scored the winning goal in the world cup final.[1] Although success stories certainly exist, many important areas of sports quantification remain unexplored.

In our case, we are interested in methods for preventing sport injuries. Over two decades with development of player tracking technologies has given us high-fidelity data streams capturing athlete's movements during on field session and matches in minute detail [3]. Some existing tools, like our PMSys system [4], also capture athlete performance metrics off-field using subjective standard self-reporting schemes. Coaches can consume this data from detailed dashboards and plots visualizing trends and statistics in the captured data; both on the individual level, as shown in Figure 1(b), or as aggregate team data, as shown in Figure 1(a). Still, effective screening programs and methods to predict athletes that are at high risk of suffering a sport injury remain largely missing [5].

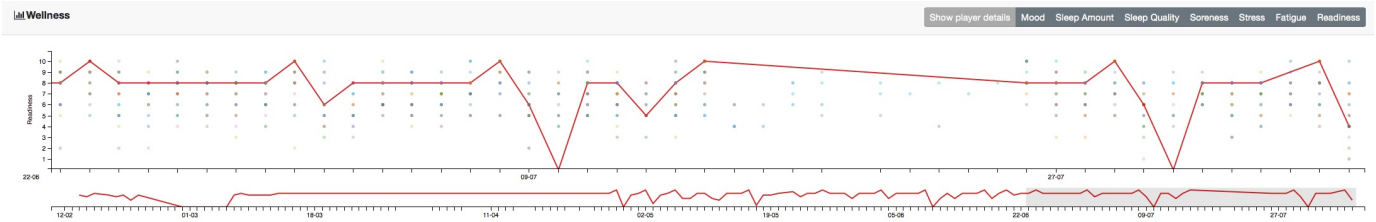
In this paper, we propose that data from systematic longitudinal monitoring of individual athletes' phenotypic and self-reported parameters, like those collected by PMSys, can be used to predict peaks in readiness to train and ultimately prevent injuries. Subjective and self reported data are influenced by individual interpretation and preferences, which can vary over time. As such, there might not exist an exact mapping from reported values to an universal scale common to all players. Self reporting is, however, commonly used and a widely accepted methodology for producing meaningful insights in other fields of research, such as in psychology [6, 7, 8].

Based on collected subjective self-reported data from two professional soccer teams in Norway, we show the effectiveness of using a Long Short-Term Memory (LSTM) recurrent neural network, a common machine learning technique [9], to predict reported training load. With such data, coaches can adjust training load to avoid over training, which is key to keep individual athletes fit.

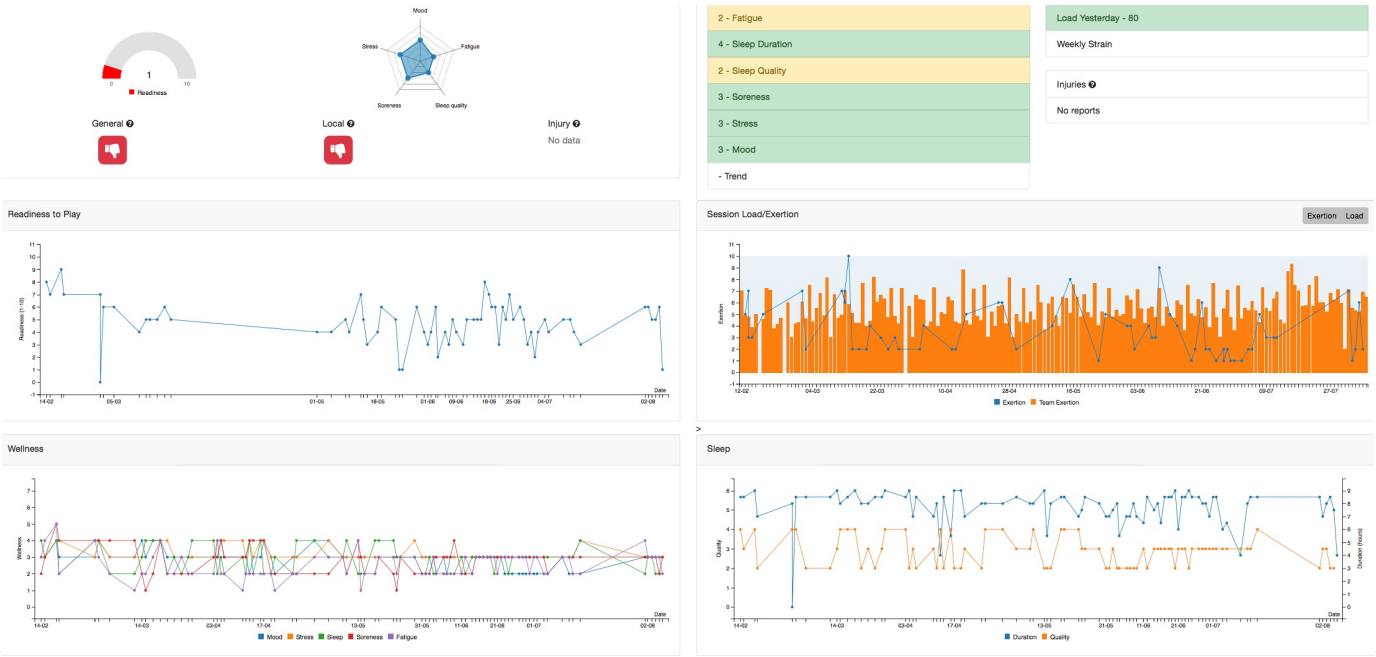
Our findings are promising. From a small amount of data, the system can predict the future very well. On prediction of a player's *readiness to train* on a scale from 1 to 10, we were able to predict positive peaks (values above 8) and negative peaks (values below 3) with a precision and recall above 90%, for both datasets used in the paper. Based on these results, the main contributions of this paper are:

- A novel system to collect, index and visualize data collected from sport professionals.
- Analysis of self reports collected from soccer players using the system to perform prediction of future states, which can be used by coaches.

[1]https://www.verdict.co.uk/world-cup-wearables/



(a) Example of a team plot highlighting one of the players



(b) Example of individual plots

Fig. 1. Dashboard from the PMSys system illustrating athlete and team status

- A method to conduct readiness to train peak prediction using LSTM.

We start by describing the method used in this paper to capture training and testing data, followed by a description of the datasets. Next, experiments and our main results are presented, before we finally conclude.

II. Method

For this paper, we are using data collected with PMSys [4]: our tool for longitudinal studies on subjective daily parameters in athlete cohorts. The PMSys system consists of a modern smartphone application coupled with our own Open mHealth [10] compliant Data Storage Unit (DSU), running on the Amazon AWS cloud service. The smartphone application is designed to accommodating most in-use smartphone systems, and is currently available for iOS and Android systems. This limits potential selection bias when favouring certain brands.

Using PMSys, athletes can submit new reports every day with little effort, both after training and after matches. For example, the interactive flow for reporting wellness, show in Figure 2, involves seven quick clicks. To ease the process of choosing the correct values for the players, each step on the reporting scales has a defined description. For example, for mood the value 4 is a players normal value. To move up or down the scale, one would have to meet certain criteria. In order to qualify for a value of 2 for mood, a player would have to be more annoyed and easily irritated than usual. The interaction and layout is predictable, and scrolling is avoided on most modern sized screens. With some experience, user will be able to complete this interactive flow in seconds.

Data captured by PMSys is owned by the user who reports it, and is by default not accessible by anyone else. Explicit data sharing policies must be set in our Policy Server Unit (PSU) component, which is a separate process from the DSU. Policies can include coach/trainer access, aggregation functions, time limited access, and are managed centrally by the data owner. Player reports are stored on one or more DSU servers, depending on the level of isolation or replication required. The DSU servers does not contain player names and requires profile information from the PSU to obtain an identity, allowing policies for sharing pseudonymous player data in real time to aggregate functions and deep learning. An overview of the PMSys architecture can be seen in Figure 3.

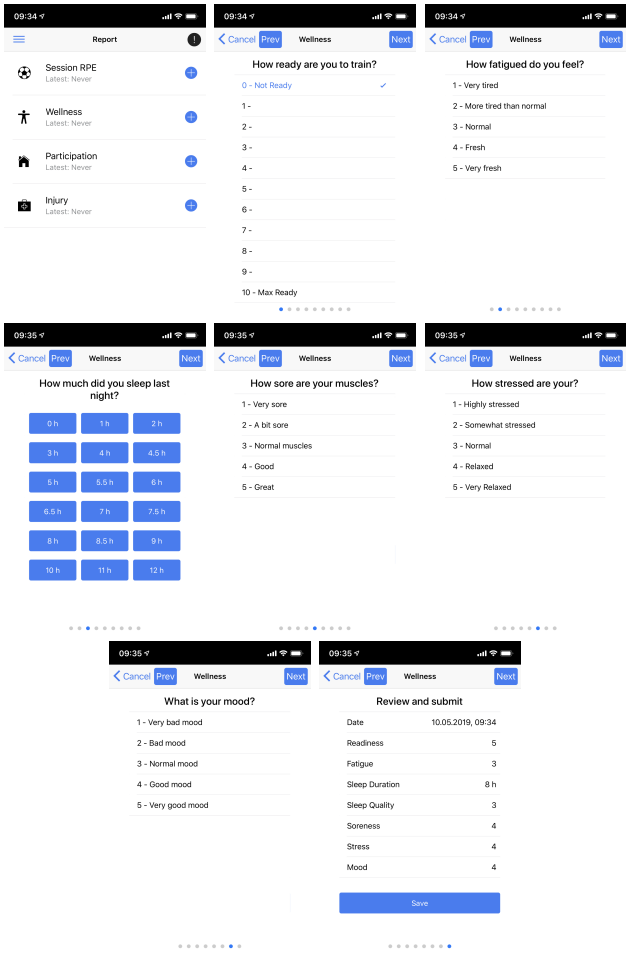


Fig. 2. Entering wellness data into the PMSys smartphone application

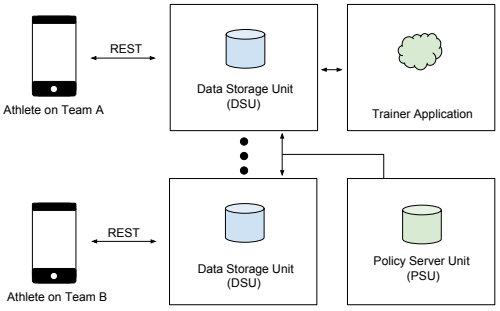


Fig. 3. Architectural overview of the PMSys data storage system

Data collected from player activities routinely measured during the competitive season, it is not subject to approval by an institutional or regional ethics committee [11]. Nevertheless, this project was approved by the Institutional Review Board at UIT–The Arctic University of Norway and by the Norwegian Centre for Research Data. Participation was furthermore based on informed consent from the individual soccer players. Team and player confidentiality was ensured by anonymizing the data.

III. Dataset

Using the data acquisition method outlined in Section II, we created two separate datasets: Team 1 and Team 2. Each dataset consists of self-reported wellness parameters from a Norwegian high-division soccer team. The frequency of the PMSys reporting was on a daily basis, for all participating players. The records contain the players' reported mood, stress levels, sleep duration, sleep quality, fatigue, soreness, and readiness to train (i.e., how fit a player feels). An overview of relevant wellness variables and descriptions captured by PMSys can be found in Table I. All features are measured on a scale from one to five, except sleep duration and readiness. These are reported as time value and on a scale from one to ten, respectively. For all scales, low values are considered least desirable and the higher the better. For example, a mood value of 1 means the player was in a bad mood compared to a mood of 5 that would be a very good mood. Examples on how the scale are presented to the user can be found in Figure 2.

Neither of the datasets were without missing values, meaning not all days from start to finish contained reports from all players. For the initial experiments the missing values were not replaced or removed to get a realistic use case (in reality, the data will always contain some missing values from the players due to for example vacation time, etc.). The Team 1 dataset stretches back from January 2017 to late August 2017 and contains data for 19 different players. The Team 2 dataset contains values from February 2018 to mid June 2018 for 22 players. In total, we have just above 6,000 days of reports included in both datasets. Seen in the context of machine learning this is not much, but we are still able to see some promising results.

IV. Results

The purpose of our experiments was to discover if modern machine-learning methods can be applied to predict future health and fitness states of players. Such predictions can be used by coaches to select the set of players that are to attend a game or a training, and adjust the duration and intensity of their activities. Using the self reported data collected by PMSys, as described above, we attempt to predict the players readiness to train the next day based on the self reported variables mood, stress, sleep quality, fatigue, and soreness.

To train our model, we used Long Short-Term Memory (LSTM): a variation of recurrent neural networks that can memorize certain parts of the data. LSTMs are state-of-the-art for time-series analysis and can also handle missing values quite well, which is a good fit for our task [12]. The LSTM model was designed to take as input a player's reported readiness values, and then output a predicted value for that particular player. The model operates on a day-by-day basis, using one day's values to predict the next. To explore the possibilities inherent in the data, we kept the model small and simple, making it easier to reproduce and interpret the results. Additional data collected in the future might lead to a more complex model.

TABLE I
VARIABLES IN THE DATASET AND THE REPORTED VALUES MEANING

Variable	1	2	3	4	5
Mood	Very bad mood	Bad mood	Normal	Good mood	Very good mood
Stress	Highly stressed	Somewhat Stressed	Normal	Relaxed	Very relaxed
Sleep quality	Insomnia	Restless sleep	Normal	Good	Very restful
Fatigue	Very tired	More tired than normal	Normal	Fresh	Very fresh
Soreness	Very sore	A bit sore	Normal	Feeling good	Feeling great

As hyper parameters for training the LSTM model, we used a sequence number of 36, 30 epochs, batch size of 4, number of layers 4 (input layer, 2 hidden layers, output layer), and as optimizer rmsprop [13]. We also tested more traditional machine-learning approaches with the dataset, such as Random Forrest and linear regression, but could not achieve a statistically significant better results. Therefore we only report the LSTM results in this paper. Even though the dataset size seems small for a deep-learning based approach, we have sufficient data points in the dataset to train the LSTM. The LSTM architecture used in this paper is not very complex having only two hidden layers.

For training and validation, we use two different methods. First, training on all other players on the team, then predicting the readiness of the chosen player. Second, use most of a player's data to train and then predict on the rest. The drawback of the first method is that the results might be skewed since individuals might have differing behaviour in their patterns. However, the model does get a lot more data to train on. The second method trains the model with the same player it is trying to predict, thus perhaps getting more representative training data, though a lot less.

We also wanted to see what combination of features gave the closest prediction. By trying all possible combinations of features as input, we found that the best result was achieved by using all inputs and differences between the subsets were negligible. Noteworthy though, is that the only feature more or less possible to control is the amount of sleep a player gets. Therefore, this input is of particular interest in trying to predict the players' readiness.

A. Training on one

Training and predicting on same player seemed unfeasible considering the network would only have between 100 to 200 time steps to train on. Intuitively the ideal training data would be a much larger set of data from the same person, but this was not the case with any of the players. Surprisingly, the model still performed fairly well despite the scarceness of training data. The values were off by a visible amount, but the prediction was able to roughly follow the shape of the actual data. In Figure 4(a) and Figure 4(c), we can see the performance of training on single players. As can be seen, the method is not very accurate, but peaks are in general observable.

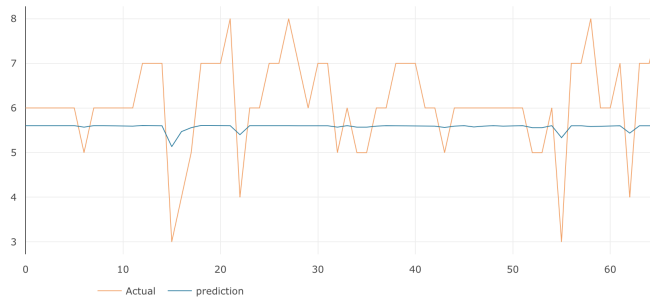
B. Training on all

In training on all players but one, predicting the behavior of that player‘s readiness achieved the best results. Already early in the experiments the model was effectively producing a plot fairly similar in shape as the actual graph. However, the predictions values were off. By tweaking the hyper-parameters of the model we were able to get a closer fit, however still off by a visible amount. Despite this, how closely the prediction follows the general shape of the actual curve were considered as an acceptable preliminary result. Figures 4(b) and 4(d) depict the performance of the model trained on all the players in the team. It can be observed that the prediction is quite accurate and all peaks are clearly visible in the predictions.

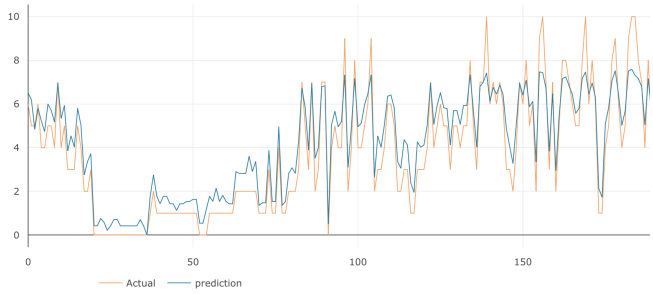
C. Peak detection

Using the LSTM model and the different methods of training described previously, we ended up with four different models: two for each team. These models lead to reasonable results in predicting future values but, most probably due to the lack of training data, precise predictions are not possible. Therefore we extended the experiments to negative or positive peak detection for the readiness to train of a player. Being able to predict peaks in a players perceived readiness to train can be of large interest for a number of reasons. For example, coaches want this information as early as possible so that they can individually fit an exercise program to specific athletes. Being able to predict negative peaks can also be of benefit for a number of reasons. Knowing the day before a match which players will feel down can ease the process of choosing which players should be preferred for a match.

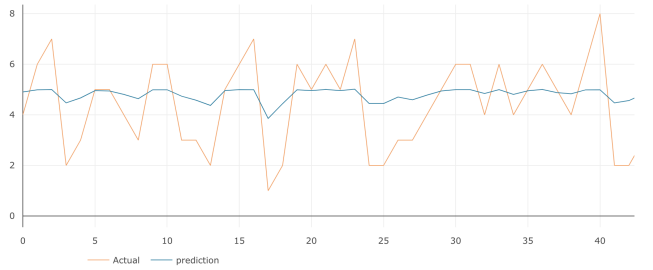
1) Positive Peaks: A positive peak is defined as a readiness to train above 8. As one can see from Figure 5, the prediction is reasonably good at recognizing peaks of eight, nine, and tens. With this specific player, in each instance, a positive deviation of 20% from the average of the entire prediction, meant an actual value of 8, 9 or 10. A good start, however, distinguishing these high values from each other seems to be slightly more difficult for the model. In general, the definite peaks of the prediction align with the definite peaks of the actual values. Unfortunately, the peaks of the prediction are of only slightly higher value than other lower peaks that correspond to lower actual values. The difficulty of differentiating high values from each other greatly diminishes if a player has a higher average value for readiness to train. Players with curves that on average stay high have a larger variation in their positive peaks, thus easier to accurately differentiate 8, 9 and 10s. Being able to differentiate predicted 8s from predicted 9s, is of course desirable, and a requisite in mapping the predicted values to a discrete 0-10 scale. However, in these early experiments, this seemed too difficult for the simple model used.



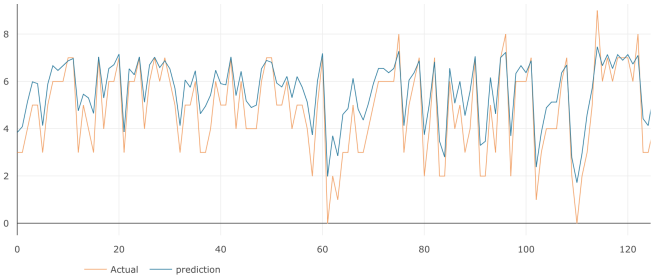
(a) Team 1, training using one specific player.



(b) Team 1, training using all players in the team.



(c) Team 2, training using one specific player.



(d) Team 2, training using all players in the team.

Fig. 4. Example predictions compared to real data for the two different training methods of the LSTM. Subfigure a and b show the player wise and all players model for Team 1 and subfigures c and d the same for Team 2. Each plot shows both the actual data and predictions for a random player from the specified team. The x-axis is a time index per report, and the y-axis is the 0–10 scale of readiness to play.

2) Negative Peaks: A negative peak for readiness to train is a value below 3. As with the positive peaks the prediction effectively predicted negative peaks from the base model.

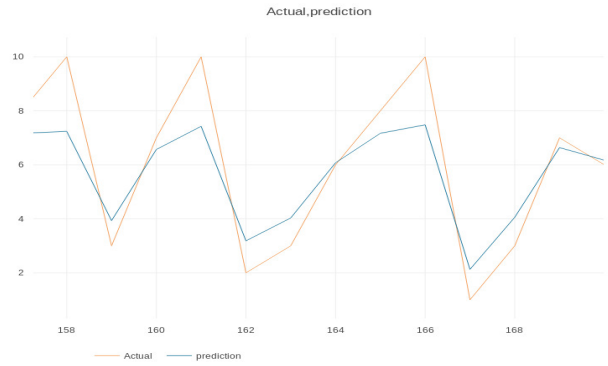


Fig. 5. Example graph for actual and detected positive and negative peaks for readiness to train in the dataset. The x axis is a time index per report, and the y-axis is the 0-10 scale of readiness to train.

From Figure 5, one can observe that the prediction closely follows the drops of the actual graph. Moreover, the prediction was able to differentiate well between low extremes. With this specific player for example, the average corresponding values to the 10s and 9s were 7.434 and 7.278, respectively; a difference of only 0.156, but also the range of the predicted values for the two actual values overlaps. Whereas with the average corresponding values for the negative peaks of 2 and 1 were at 2.952 and 2.001, respectively. Moreover, there was no overlap between the values, meaning the highest values predicted for a 1, was still lower than the lowest predicted value for 2. This makes it a lot easier to create ranges of predicted values that map to actual values of the discrete 0-10 scale.

3) Peak detection methods: To determine if the prediction was a positive or negative peak from the LSTM model prediction, we performed two different post processing methods. For the first methods, we simply defined a value as a positive or negative peak if it was below or above a certain margin of the defined maximum. For this method, we tested three different values 1, 2, and 3. For the second method, we calculated the average of all time steps and classified a predicted value as a positive or negative peak if the value were above or below a certain percentage of the average. Both methods are very simple, and for future work, it might be interesting to apply another step of machine learning on the peak detection part itself. Nevertheless, as the results show, peak detection using this methods works very well.

4) Peak Detection Results: For the peak results, we are reporting weighted averages of the standard metrics precision, recall, and F1 score. In addition, we also provide the number of true positives, false positives, true negatives and false negatives. Table II lists the results for the different models and different post processing methods for peak detection. In general, we can observe that models trained on the whole team data lead to better results than the models trained on individual players (0.993 vs. 0.864 for Team 1 and 0.987 vs. 0.92 for Team 2). This is very interesting and can be explained either by the fact that a model trained on the whole team has more data and secondly that dynamics within the teams are

TABLE II

RESULTS FOR NEGATIVE AND POSITIVE PEAK DETECTION. (THE POST PROCESSING METHOD SETTING AS DESCRIBED IN SECTION IV-C3, TRAINED IS TRAINING MODEL, TOTAL PEAKS IS THE TOTAL NUMBER PREDICTIONS TO MAKE, POSITIVES IS THE NUMBER OF POSITIVE PEAKS, NEGATIVES IS THE NUMBER OF NEGATIVE PEAKS, WPREC IS THE WEIGHTED PRECISION, WREC IS THE WEIGHTED RECALL AND WF1 IS THE WEIGHTED F1 SCORE.)

Dataset	Post processing method	Trained	Total peaks	Total Positives	Total Negatives	True Positives	True Negatives	False Positives	True Negatives	WPrec	WRec	WF1
Team 1	3,3	all	1366	506	860	503	854	6	3	0.993	0.993	0.993
Team 1	2,2	all	1366	506	860	503	854	6	3	0.993	0.993	0.993
Team 1	1,1	all	1366	506	860	316	721	139	190	0.756	0.759	0.756
Team 1	3,3	player	484	227	257	177	242	15	50	0.872	0.866	0.864
Team 1	2,2	player	484	227	257	174	239	18	53	0.86	0.853	0.852
Team 1	1,1	player	484	227	257	87	198	59	140	0.591	0.589	0.572
Team 2	3,3	all	685	554	131	549	127	4	5	0.987	0.987	0.987
Team 2	2,2	all	685	554	131	549	127	4	5	0.987	0.987	0.987
Team 2	1,1	all	685	554	131	493	84	47	61	0.849	0.842	0.845
Team 2	3,3	player	144	126	18	118	14	4	8	0.926	0.917	0.92
Team 2	2,2	player	144	126	18	81	4	14	45	0.756	0.590	0.656
Team 2	1,1	player	144	126	18	9	3	15	117	0.331	0.083	0.11
Team 1	ave	all	1366	506	860	503	784	76	3	0.949	0.942	0.943
Team 1	20,10	all	1366	506	860	272	723	137	234	0.722	0.728	0.721
Team 1	40,20	all	1366	506	860	149	623	237	357	0.543	0.565	0.55
Team 1	60,30	all	1366	506	860	50	448	412	456	0.352	0.365	0.358
Team 1	ave	player	484	227	257	227	218	39	0	0.931	0.919	0.919
Team 1	20,10	player	484	227	257	19	156	101	208	0.302	0.362	0.318
Team 1	40,20	player	484	227	257	19	96	161	208	0.217	0.238	0.226
Team 1	60,30	player	484	227	257	19	79	178	208	0.191	0.202	0.196
Team 2	ave	all	685	554	131	549	127	4	5	0.987	0.987	0.987
Team 2	20,10	all	685	554	131	213	127	4	341	0.846	0.496	0.528
Team 2	40,20	all	685	554	131	0	115	16	554	0.033	0.168	0.055
Team 2	60,30	all	685	554	131	0	85	46	554	0.025	0.124	0.042
Team 2	ave	player	144	126	18	102	14	4	24	0.888	0.806	0.832
Team 2	20,10	player	144	126	18	0	3	15	126	0.003	0.021	0.005
Team 2	40,20	player	144	126	18	0	3	15	126	0.003	0.021	0.005
Team 2	60,30	player	144	126	18	0	2	16	126	0.002	0.014	0.003

also influencing the performance of the players and therefore taking the whole teams' data into account lead to a better understanding (this is also visible in Figure 4).

Furthermore, it can be observed that the team that collected data for a longer period (Team 1) achieves better performance than the one with less data (0.993 vs. 0.987). Nevertheless, this difference is minimal and might not be statistical relevant.

In terms of the two different post-processing methods and their margins, we can observe that the method taking the average into account performs worse than the method looking at a fixed margin. For Team 1, the difference is around 4 percent and for Team 2 they perform equally good. For both methods, increasing the margins or percentage leads to worse results and for the player based method fixed margins work better. The reason for that is most probably the fact that for using the average in a precise way more training data is required, otherwise the average is not very accurate (depicted in the difference between the results for Team 1 and Team 2).

V. CONCLUSION

In this paper, we have made a contribution to the silent revolution in international sports having an increasing availability of athlete quantification data by developing an advanced analytic components over various data types. As a first step to validate the potential of our innovative idea, we chose to focus on the subjective wellness reports of soccer players.

Because we study time ordered discrete data-points, we opted to use Long Short-Term Memory (LSTM) recurrent neural network technique for this paper: a state-of-the-art machine-learning method for time series analysis. Using data from two professional soccer teams, we show that LSTM can be applied to predict future peeks in a soccer player's readiness-to-train. Overall, the performance of our models are promising. Based on a small amount of training data, the system can predict the future very well. For both datasets, we predict positive and negative peaks with a precision and recall above 90%.

ACKNOWLEDGEMENTS

This work was supported in part by the Norwegian Research Council project numbers 263248/O70 and 250138.

REFERENCES

[1] M. Di Mascio, J. Ade, C. Musham, O. Girard, and P. S. Bradley, "Soccer-specific reactive repeated-sprint ability in elite youth soccer players: Maturation trends and association with various physical performance tests," *The Journal of Strength & Conditioning Research*, 2018.

[2] K. Ng and T. Ryba, "The quantified athlete: Associations of wearables for high school athletes," *Advances in Human-Computer Interaction*, vol. 2018, 2018.

[3] A. Rossi, E. Perri, A. Trecroci, M. Savino, G. Alberti, and M. F. Iaia, "Characterization of in-season elite football trainings by gps features: The identity card of a short-term football training cycle," in *Proc. of IEEE ICDMW*, Dec 2016, pp. 160–166.

[4] Forzasys AS and the Corpore Sano center. (2018) Pmsys webpage. [Online]. Available: http://forzasys.com/pmsys.html

[5] R. Bahr, "Why screening tests to predict injury do not work—and probably never will…: a critical review," *British Journal of Sports Medicine*, vol. 50, no. 13, pp. 776–780, 2016. [Online]. Available: https://bjsm.bmj.com/content/50/13/776

[6] C. L. Craig, A. L. Marshall, M. Sjorstrom, A. E. Bauman, M. L. Booth, B. E. Ainsworth, M. Pratt, U. Ekelund, A. Yngve, J. F. Sallis *et al.*, "International physical activity questionnaire: 12-country reliability and validity," *Medicine and science in sports and exercise*, vol. 35, no. 8, pp. 1381–1395, 2003.

[7] A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, and V. S. Cain, *The science of self-report: Implications for research and practice*. Psychology Press, 1999.

[8] S. A. Prince, K. B. Adamo, M. E. Hamel, J. Hardt, S. C. Gorber, and M. Tremblay, "A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 5, no. 1, p. 56, 2008.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] O. mHealth. (2018) Open mhealth. [Online]. Available: http://openmhealth.org

[11] E. M. Winter and R. J. Maughan, "Requirements for ethics approvals," *Journal of Sports Sciences*, vol. 27, no. 10, pp. 985–985, 2009, pMID: 19847681.

[12] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[13] G. Hinton, N. Srivastava, and K. Swersky, "Rmsprop: Divide the gradient by a running average of its recent magnitude," *Neural networks for machine learning, Coursera lecture 6e*, 2012.