Scalable Infrastructure for Efficient Real-Time Sports Analytics

Håvard D. Johansen*
UiT The Arctic University of Norway

Dag Johansen
UiT The Arctic University of Norway

Tomas Kupka
Forzasys AS, Norway

Michael A. Riegler
SimulaMet, Norway

Pål Halvorsen†*
SimulaMet, Norway

ABSTRACT

Recent technological advances are adapted in sports to improve performance, avoid injuries, and make advantageous decisions. In this paper, we describe our ongoing efforts to develop and deploy PMSys, our smartphone-based athlete monitoring and reporting system. We describe our first attempts to gain insight into some of the data we have collected. Experiences so far are promising, both on the technical side and for athlete performance development. Our initial application of artificial-intelligence methods for prediction is encouraging and indicative.

CCS CONCEPTS

• **Applied computing** → **Consumer health**; **Health informatics**; • **Computing methodologies** → *Supervised learning*; • **Information systems** → Summarization.

KEYWORDS

Sports performance logging, algorithmic analysis, privacy-preserving data collection, artificial intelligence, machine learning

ACM Reference Format:
Håvard D. Johansen, Dag Johansen, Tomas Kupka, Michael A. Riegler, and Pål Halvorsen. 2020. Scalable Infrastructure for Efficient Real-Time Sports Analytics. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425300

1 INTRODUCTION

Mobile and wearable Internet-of-Things (IoT) technologies are being adapted in sports clubs to closely monitor individual athletes during both training and competitions. Dominant clubs have hired inter-disciplinary scientists and experts to collect and analyze data: from training load, wellness, and mood to nutrition to provide evidence-based suggestions to the coaches. Many commercial sports apps already exist, helping athletes maintain meticulous training logs: recording both subjective and objective parameters. With the introduction of mobile and wearable IoT devices, there is still much that can be learned about each individual athlete. Also, there is still a lot of manual and tedious operations associated with development and management of elite athletes, which has been the target research focus for us this last decade attempting to efficiently automate and improve many of these tasks.

We have been in this inter-disciplinary domain the last decade combining technological innovations with sports science—actively participating in the development from pen and paper logging by coaches to advanced automatic Artificial Intelligence (AI) enabled systems. Some of our initial prototypes were developed in collaboration with and used by sports science departments, elite clubs, and national teams, and several of these systems are now adopted by industry and fully deployed in elite sports clubs. This includes non-invasive 24/7 athlete monitoring and logging, and novel video solutions [11]. The pressing research problem is no longer how to obtain enough data from the athletes, but what is the relevant data and how to collect this in a non-invasive, timely, secure and privacy-preserving manner. Managing all this massive heterogeneous data can be more a burden than provide benefits, unless proper analysis can be performed. Hence, our current research focus is to investigate and develop appropriate real-time analysis systems for immediate intervention purposes.

To our surprise, working with elite clubs and national team partners for a decade, our end-users appreciated the most a system that we initially considered trivial from a technological perspective. This was not, for instance, our panorama video analysis system [11] combined with positional sensors [16, 19] nor our real-time analysis of physical performance [1]. Neither was it our video feedback or e-learning multimedia systems [14, 15], but our smartphone-based athlete monitoring system fine tuned over many redesigns and re-implementations. Experiences and on-going research with this system is the focus of this paper. We address this often underestimated component of subjective reporting and the corresponding data analysis process.

In this paper we briefly present the design and implementation of our athlete-monitoring system PMSys: a state-of-the-art tool for collecting and analysing daily subjective reports from individual athletes in sport teams. The system is in daily active use by sport colleges; and junior, elite, and national teams in Norway. Data from a cohort of approximately 400 female soccer players in Norway, Denmark, and Portugal is also used in a research project on female elite performance development. This research project includes researchers from medicine, psychology, nutritional science, sports science, mathematics, and computer science. We point out key findings and our initial experiences from using modern AI methods to predict individual training load from collected data.

*Also affiliated with Forzasys AS, Norway
†Also affiliated with Oslo Metropolitan University, Norway

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425300

2 BACKGROUND

Key sport governance organizations like Fédération Internationale de Football Association (FIFA) have already approved certain wearables and electronic performance and tracking systems in official soccer matches, providing a foundation for evidence-based decisions and team performance improvements [7, 18]. We are in particular interested in the use of technology and data analytics for *injury prevention*. In the 2008–20009 season, the Spanish 1st division had 24 players with 360 player-days absence [8], and in the 2017–2018 season, English Premier League clubs paid £217m in wages to injured players [3]. For Manchester United the average cost per injury was £0.87m, partially driven by high individual salaries. While meeting with English Premier League clubs, we received informal estimates that a central player who is injured for a period quickly accumulates internal loss in the order of £5m. Some speculate that having the second-lowest number of injuries was instrumental for Manchester City to win the 2018 Premier League Championship [21]. One may certainly argue that a team with some of its best players unavailable due to injuries are more likely to lose matches.

Systematic collection of data combined with close personal follow-ups by a well-functioning sports science department that manages the physical performance development of the players, have the potential to prevent injuries, reduce cost, improve performance, and win matches. Such a methodological approach to performance optimization has become vital for modern elite sports.

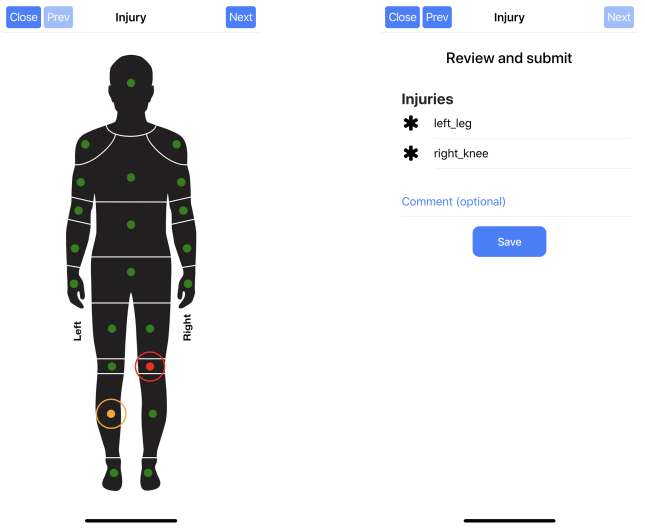
Self-reporting is a widely accepted and used methodology for producing meaningful insights in other research fields, like psychology [6, 20, 22]. Collection of training and competition-exposure data as well as wellness information regarding sleep patterns, mood, nutrition, and muscle soreness on a daily basis needs implementation of modern technology to monitor teams and groups of athletes [13]. Subjective and self-reported data are also influenced by individual interpretation and preferences, which can vary over time. There might not exist an exact mapping from reported values to a universal scale common to all players. As such, subjective sports data is a good candidate for AI analytics.

3 ATHLETE MONITORING

To support subjective parameter logging in sports cohorts, we designed and implemented PMSys: a smartphone-based tool that enables systematic longitudinal monitoring of athletes' phenotypic and self-reported parameters. Athletes may report many useful parameters, like session Rating of Perceived Exertion (sRPE), wellness, injury and illness, session participation, game evaluations—and in these COVID-19 times, an infection symptom check. The system is designed around the governing principle that subjective reports must be captured with little effort and in real-time, while they are fresh and relevant. For instance, for wellness reports, this is the narrow time window after the athlete gets out of bed in the morning and until the first morning training session starts.

3.1 Front-End Applications

Athletes interact with PMSys using our *Reporter App*: a smartphone application, available for both Android and iOS systems. There are significant development and maintenance overhead associated with



(a) Body silhouette **(b) Injury summary**

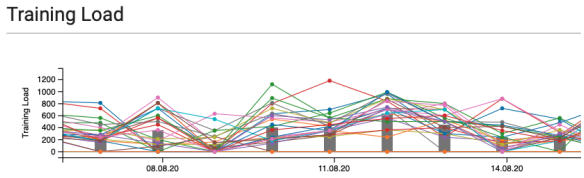
Figure 1: PMSys injury reporting.

smartphone applications compared to traditional web-based services. However, our real-time requirement implies that we needed a tool that is highly available and functional, even without an Internet connection. Because native smartphone applications can store code and state locally on the device, they are better suited to meet these demands than those running in a browser. We further improve availability by designating the smartphone as a the primary node for the user's data, storing reports in a local SQLite database and serializing updates. The data in our cloud backend service are replicas, synchronized with the application on updates and whenever the smartphone comes online. This enables a consistent view among replicas while allowing an athlete to recover personal data in case of damage or loss of his device.

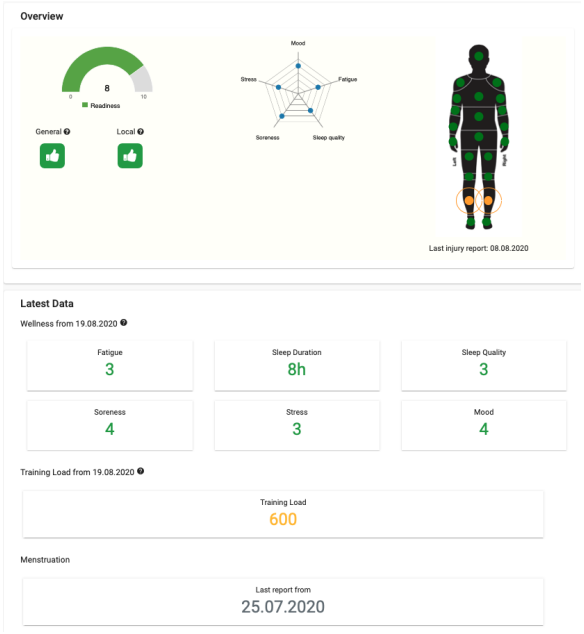
Our requirement for fresh and relevant data also required us to prioritize the user experience towards reducing the time spent to report rather than on whistles and bells. Over time, we have simplified questions, reduced the text, and improved input interfaces. For example, going from a tedious set of 11 complex questions for injuries to create an injury score, to be repeated for each injury, we have devised a much simpler interface based on a body silhouette as depicted in Figure 1. Compared to a couple of minutes for injury reporting earlier, reporting with the new interface is minimized on average to about 12–15 seconds.

For many report types, like sRPE and wellness, that are based on scientifically validated questionnaires, optimizations like the one we did for injuries cannot be adapted. In such cases, our goal has been to reduce button clicks and eliminate scrolling by using simple layouts with large buttons. This reduces required interactions and attentions, and helps athletes automate the process of submitting reports in time.

Team personnel, like coaches and physicians, can interact with athlete data using our *Trainer Portal*: a web-hosted Single-Page Application (SPA) implemented using the Angular framework [9]. The portal displays submitted data, both in a team view and in a



(a) Team overview



(b) Single player

Figure 2: PMSys Trainer Portal samples.

single-player view, from all team members who have granted access. Included team views include visualizations of injuries, illnesses, and session participation (see Figure 2a). It also plots important training-load indicators, like daily and weekly load, acute load, chronic load, acute chronic workload ratio, monotony, and strain. Single player view gives individual details (see Figure 2b). The portal also allows push messages to be sent, directly or on a schedule, as reminders to report. This has proven to be very efficient for increased participation.

3.2 Cloud Services

The backend system consists of several microservices that run on the Amazon AWS public cloud. All reports recorded by our Reporter App is synchronized with our Open mHealth [10] compliant Data Storage Unit (DSU). Because our DSU is critical for report submission, it is kept functionally minimal, stable, and independent from other system components, except for a cloud-hosted PostgreSQL database for persistent state.

In accordance with the Open mHealth standard, submitted reports are JSON data-point objects, each made up of a `header` structure and a body structure. The header contains various field, the most important are: `id`, `schema_id`, and `user_id`. To improve security and privacy, the `id` identifies the report, but is required to be unique only in combination with the user the identity `user_id`. The body can be of any type. The structure of the body is application specific, and set by the `schema_id` field.

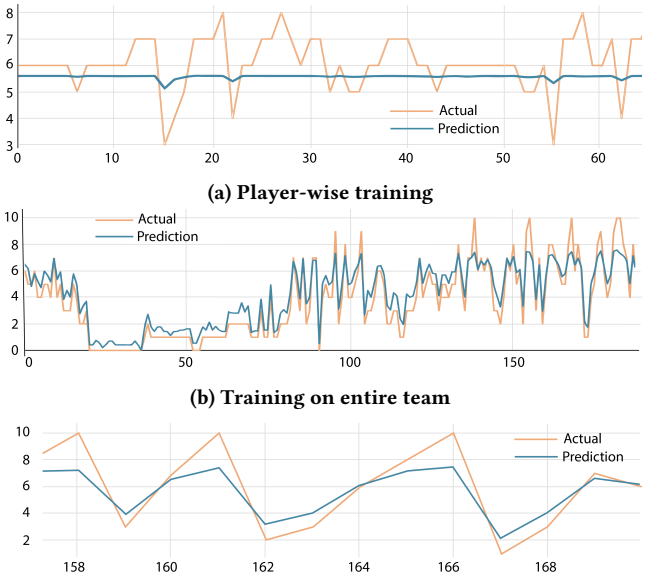
We want to avoid having schema-specific logic within our DSU component, as that would make it difficult to add new report types. Instead, the DSU considers the bodies as opaque objects, inserted into the database without further processing or interpretation. Interpretations of the data-point bodies are only done in the Trainer Portal, Reporter App, and in report-specific analytics functions running in the backend services.

Following privacy-preserving guidelines, each report is owned by the user who generated it, as specified in the `header.user_id` field. The DSU follows a narrow security policy that allows users only to insert and access their own data points. Users are authenticated using trusted third-party OpenID Connect (OIDC) services and are only granted access to data points where `header.user_id` matches the authenticated subject in their access token [17]. Bearer tokens in combination with a third-party authentication provider, like Microsoft or Google, are well suited for our purpose as it allows the DSU to authorize requests without storing directly identifiable information [4], like names or email addresses. Although this approach cannot guarantee anonymity, identification becomes more difficult, allowing safer access to automated backend functions that conduct real-time statistical analysis and AI-based predictions.

Because of the security restrictions in the DSU, the Team Portal must access data using a separate Team Service component. The Team Service grants access based on OIDC access tokens carried with requests, similarly to the DSU. However, team membership and role assignment credentials must be encoded in the token. For this, we provide a Policy Service for managing team roosters, staff lists, and player profiles. This service must be consulted by the OpenID provider to embed the correct credentials during authentication.

4 PEAK PREDICTION USING AI ANALYTIC

To investigate the potential of automatic analysis, we performed an initial experiment [23] to evaluate if current machine-learning methods can be applied to predict future health and fitness states of players. We needed a state-of-the-art machine-learning model for time-series analysis that can memorize certain parts of the data, and Long Short-Term Memory (LSTM) seemed to be a good fit for our task [5]. We used data from a Norwegian elite soccer team collected over several months (January–August) for 19 players. For this initial experiment, we only used the readiness parameter. Although we did not expect our dataset to contain sufficient data to build an accurate machine-learning model, we are still able to observe some promising first insights. For our experiments, we used a small LSTM model with 4 layers: the input layer, 2 hidden layers, and the output layer. The model is designed to take as input a player's reported readiness values and output a predicted value for that particular player on a day-by-day basis, using one day's values to predict the next. To train the LSTM model, we used a sequence number of 36, 30 epochs, batch size of 4, and the rmsprop optimizer [12] as hyper-parameters.



(a) Player-wise training

(b) Training on entire team

(c) Zoom-in on actual detected peaks

Figure 3: LSTM predictions compared to real data for a random player. The x-axis is a time index per report, and the y-axis is the readiness to play.

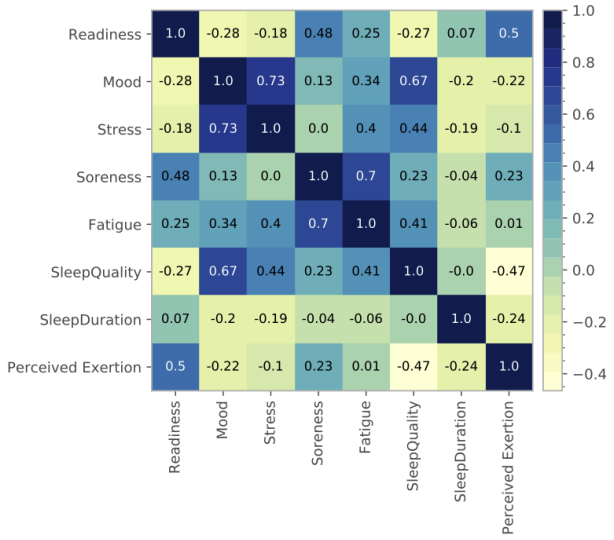


Figure 4: Correlation matrix for the soccer dataset.

For training and validation, we used two different methods. First, we used most of a player's data to train and then predict the rest. Second, we trained the model on all other players on the team, then predicting the readiness of the chosen player. The results are shown in Figure 3a and Figure 3b, respectively. From Figure 3a, we observe even though some peaks can be identified that the amount of data for one single player is not enough to make a useful model having only a few hundred data points. However, by adding more training data from other players on the same team, we observe in Figure 3b that the prediction is quite accurate, and all peaks are clearly visible in the predictions. This is expected as the dynamics within a team influences the performance of single players as they often have the same training schedules.

Observing the promising analysis above, we then tried to predict peaks more accurately, meaning when the athlete feels at the top and bottom, indicated by a readiness score of 8–10 and 1–3, respectively. Figure 3c shows a zoomed-in overview. Even if the actual readiness value often is off, the trend of the peak is detected with high precision with percent-values in the high 90s for both precision and recall if we add some post-processing.

5 VARIABLE CORRELATION

Next, we explore how reported training and wellness parameters interact by computing a correlation matrix for our dataset. The correlations, shown in Figure 4, are generated from the average scores across all players in the dataset.

Lighter colors represent a weaker or negative correlation between a variable pair, while a darker color indicates the opposite. When considering readiness as the target variable, a positive correlation can be seen when compared to the variables soreness and fatigue. In a way, these observations are reasonable, considering how energy levels and muscle soreness potentially affect physical effort. On the other hand, looking at stress, there is a positive correlation between mood, fatigue, and sleep quality. A higher score on the stress level implies that a player is very relaxed, and the mood, fatigue, and sleep quality seem to correlate positively with this. Hence, a good mood, low fatigue, and better sleep quality reduce the overall stress level among players. Moreover, although these variables positively correlate with the target variables, there are other indicators that are negatively correlated as well. For instance, the relationship between readiness and mood is negative, which indicates that the readiness of players is affected negatively by their mood levels. Similarly, the sleep quality of players affects their readiness negatively. In general, for both the stress and readiness scores, the negative coefficients are closer to zero, which may indicate weaker correlations. Nevertheless, they may still be contributing input features in the classification task itself.

6 CONCLUSIONS

Success in modern elite soccer is increasingly impacted by advancements in technology [2]. Based on a decade of experience in the intersection of computer science, sports science, and medicine, we have developed PMSys: a smartphone-based tool and backend system for athlete monitoring. The system has seen wide use in various soccer teams, from academy youth teams, to national (Norway U21 and senior) and international elite teams. We have observed that coaches have adjusted training routines based on data collected by PMSys, potentially avoiding injuries, and making teams more aware of the impact of details, like the effect of going to bed earlier.

To not fall into disuse, tools like PMSys must be scalable, secure, and easy to use. Most importantly, both athletes and team personnel must see clear benefits of using such systems. Towards this end, we have shown that an initial analysis of PMSys data using AI for prediction is encouraging and indicative.

REFERENCES

[1] K. Andreassen, D. Johansen, H. Johansen, I. Baptista, S. A. Pettersen, M. Riegler, and P. Halvorsen. 2019. Real-time Analysis of Physical Performance Parameters in Elite Soccer. In *Proceedings of the International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–6.

[2] Simon Austin. 2020. Monchi: Big Data is the future of football. Training Ground Guru: https://trainingground.guru/articles/monchi-big-data-is-the-future-of-football.

[3] BBC. [n.d.]. *Premier League clubs paid £217m in wages to injured players in 2017-18.* Retrieved 2020-09-17 from https://www.bbc.com/sport/football/45045561

[4] Arnar Birgisson, Joe Gibbs Politz, Ulfar Erlingsson, Ankur Taly, Michael Vrable, and Mark Lentczner. 2014. Macaroons: Cookies with Contextual Caveats for Decentralized Authorization in the Cloud. In *Network and Distributed System Security Symposium*.

[5] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 6085.

[6] Cora L Craig, Alison L Marshall, Michael Sjorstrom, Adrian E Bauman, Michael L Booth, Barbara E Ainsworth, Michael Pratt, ULF Ekelund, Agneta Yngve, James F Sallis, et al. 2003. International physical activity questionnaire: 12-country reliability and validity. *Medicine and science in sports and exercise* 35, 8 (2003), 1381–1395.

[7] Michele Di Mascio, Jack Ade, Craig Musham, Olivier Girard, and Paul S Bradley. 2018. 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).

[8] Ismael Fernández Cuevas, Pedro Carmona, Manuel Quintana, Javier Salces, Javier Arnaiz-Lastras, and Antonio Barrón. 2010. Economic costs estimation of soccer injuries in first and second spanish division professional teams. In *Proceedings of the Annual Congress of the European College of Sport Sciences (ECSS)*.

[9] Google Inc. [n.d.]. *One framework. Mobile & desktop.* Retrieved 2020-09-19 from https://angular.io/

[10] David Haddad, Ida Sim, Simona Carini, Emerson Farrugia, et al. [n.d.]. *Want To Use Mobile Health Data AND Have It To Make Sense?* Retrieved 2020-09-17 from http://openmhealth.org

[11] Pål Halvorsen, Simen Sægrov, Asgeir Mortensen, David K.C. Kristensen, Alexander Eichhorn, Magnus Stenhaug, Stian Dahl, Håkon Kvale Stensland, Vamsidhar Reddy Gaddam, Carsten Griwodz, and Dag Johansen. 2013. BAGADUS: An Integrated System for Arena Sports Analytics – A Soccer Case Study. In *Proceeding of the International Conference on Multimedia Systems (MMSys)* (Oslo, Norway). 48–59.

[12] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. Rmsprop: Divide the gradient by a running average of its recent magnitude. Coursera lecture 6e, Neural networks for machine learning.

[13] Terje Bektesevic Holmlund, Peter W. Foltz, Alex S. Cohen, Håvard D. Johansen, Randi Sigurdsen, Pål Fugelli, Dagfinn Bergsager, Jian Cheng, Jared Bernstein, Elizabeth Rosenfeld, and Brita Elvevåg. 2018. Moving psychological assessment out of the controlled laboratory setting: Practical challenges. *Psychological Assessment* 31, 3 (2018), 292–303. https://doi.org/10.1037/pas0000647

[14] D. Johansen, P. Halvorsen, H. Johansen, Håkon Riiser, C. Gurrin, B. Olstad, C. Griwodz, Åge Kvalnes, J. Hurley, and T. Kupka. 2011. Search-based composition, streaming and playback of video archive content. *Multimedia Tools and Applications* 61 (2011), 419–445.

[15] Dag Johansen, Håvard Johansen, Tjalve Aarflot, Joseph Hurley, Åge Kvalnes, Cathal Gurrin, Sorin Zav, Bjørn Olstad, Erik Aaberg, Tore Endestad, Haakon Riiser, Carsten Griwidz, and Pål Halvorsen. 2009. DAVVI: A Prototype for the next Generation Multimedia Entertainment Platform. In *Proceedings of the ACM International Conference on Multimedia (MM)* (Beijing, China). 989–990. https://doi.org/10.1145/1631272.1631482

[16] Håvard D. Johansen, Svein Arne Pettersen, Pål Halvorsen, and Dag Johansen. 2013. Combining Video and Player Telemetry for Evidence-based Decisions in Soccer. In *Proceedings of the International Congress on Sports Science Research and Technology Support (icSPORTS)*. 197–205. https://doi.org/10.5220/0004676101970205

[17] M. Jones, J. Bradley, and N. Sakimura. 2015. *JSON Web Token (JWT).* RFC 7519. Internet Engineering Task Force (IETF). 1–30 pages. https://tools.ietf.org/html/rfc7519

[18] Kwok Ng and Tatiana Ryba. 2018. The Quantified Athlete: Associations of Wearables for High School Athletes. *Advances in Human-Computer Interaction* 2018 (2018).

[19] Svein A. Pettersen, Håvard D. Johansen, Ivan A. M. Baptista, Pål Halvorsen, and Dag Johansen. 2018. Quantified Soccer Using Positional Data: A Case Study. *Frontiers in Physiology* 9 (2018), 866. https://doi.org/10.3389/fphys.2018.00866

[20] Stéphanie A Prince, Kristi B Adamo, Meghan E Hamel, Jill Hardt, Sarah Connor Gorber, and Mark Tremblay. 2008. 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* 5, 1 (2008), 56.

[21] Reuters Staff. 2018. *Soccer-Rising cost of Premier League injuries raises fixture concerns-study.* Reuters. Retrieved 2020-09-17 from https://www.reuters.com/article/soccer-england-injury/soccer-rising-cost-of-premier-league-injuries-raises-fixture-concerns-study-idUSL8N1Q3493

[22] Arthur A Stone, Christine A Bachrach, Jared B Jobe, Howard S Kurtzman, and Virginia S Cain. 1999. *The science of self-report: Implications for research and practice.*

[23] Theodor Wiik, Håvard D. Johansen, Svein-Arne Pettersen, Ivan Baptista, Tomas Kupka, Dag Johansen, Michael Riegler, and Pål Halvorsen. 2019. Predicting Peek Readiness-to-Train of Soccer Players Using Long Short-Term Memory Recurrent Neural Networks. In *Proceedings of the International Conference on Content-Based Multimedia Indexing (CBMI)* (Dublin, Ireland).