https://www.findaphd.com/phds/project/sport-and-exercise-science-fully-funded-swansea-university-and-swim-wales-phd-scholarship-the-application-of-advanced-data-analytics-to-monitor-training-load-and-predict-success-in-international-swimming/?p157949&fbclid=IwAR38IQuAavgAXHKnQLt\_Ks7shhm7eQ8NRdKTAmjU0aZCxkMKiYiDVvuiBCo

Applicants for research degrees must attach a research proposal for your proposed project - maximum of 1500 words including references. If the proposal has already been determined (e.g. you are applying in response to an advertised studentship), please attach a copy.

It is recommended that the proposal includes some or all of the following:

* A title (or area) for the proposed research project
* Aims and objectives of the research
* An outline of the proposed methodology, including information on the research sample and methods of data collection
* A list of questions to be addressed by the research
* Significance of research
* A bibliography and brief summary of research that has already been undertaken in this field

This study will gather data from at least 30 international swimmers of both sexes in order to achieve a high confidence interval with all participants signing an informed consent form.

The experiment will take place in the pool of the University of Swansea. The participants will be instrumented with a wearable inertial measurement unit (IMU) for data acquisition and later processing. Between the wearable module in this case a micro controller unit (MCU) and the base station a 433-MHz half-duplex radiofrequency (RF) communication link, based on the LoRa Peer-to-Peer (P2P) technology will be implemented.

The MCU will manage information of a three-axis accelerometer, gyroscope, magnetometer, heart rate sensor and a pulse oximeter sensor (SPO2). A memory card will also be added to the wearable device for redundancy, along with a haptic stimulation transducer which allows interaction with the swimmers by means of commands received from the base station.

Heart rate will also be measured through a chest band ( Polar T31™ Coded band) which will detect and transmit data through a 5.3 kHz radio signal. Additionally, non-invasive sensors like MAX30102 sensors will also be placed on the skin surface of the swimmers (e.g fingertips) in order to calculate SP02.

The acquisition sampling rate will be kept constant at four samples per second, and all data will be sent to the base station every second. Accelerometer, and gyroscope data will be sampled at the same frequency of 433 Hz using a full scale set at ±4 g and ±500deg· s−1, respectively. Heading data will be acquired between 0 and 7, heart rate in bpm and SPO2 in percentage.

The acquired data will then be used to build a stroke classification dataset in order to test and validate the swimming analytics framework. The system will integrate the data of the biosensors and allow online recognition of the swimming style providing important post-analysis insights of the swimmer’s performance.

Training data can also be gathered through Full HD video cameras which will be fully synchronized. Some of them will be located underwater, in the roof looking down on the pool and a side camera to track starts.

**Data collection**

The training process should last around 90-100 weeks and should be divided into different preparation macrocycles, including final competitions. Macrocycles should consist of 8-14 weeks of training preparation and 1-2 weeks of competitions.

The number of competitive performances on a specific event of each swimmer should be listed down and training loads should be documented in three zones of swim training intensities and two categories of dryland training. The three zones of training intensity should be controlled by lactate testing during the training process.

The categories of training should be similar to all swimmer athletes; The three categories of training should be in this order. 1) Compensation and maintenance aerobic training (2-3 mmol/L blood lactate), 2) Developmental and overload aerobic endurance training (4-6 mmol/L blood lactate), 3) Anaerobic and speed training (6-20 mmol/L blood lactate). The dryland training should consist of dryland strength training and dryland general conditioning training. All three different training loads will be quantified in km/week while dryland trainings in hours.

The competitive performance of the athletes will be transformed according to the FINA points.

The first analysis should determine the influence of the 2-week taper phase prior to all competitions (Model A). The second analysis should determine the influence of the high load training phase three (Model B) and four weeks prior to the competitions and lastly the third analysis should determine the influence of a 4-week phase prior to the competitions (Overall Model).

For Models A and B, Multi-layer Perceptron will consist of 10 input neurons, 2 hidden neurons and 1 output neuron. While the overall model will be computed as the mean of Model A and B. For the multi-layer perceptron to be trained it is necessary to have at least 40 data sets with each consisting of one competitive performance and the accompanying 10 training loads of 2 weeks. This data will be used to pre-train the neural network.

The Leave-One-Out Cross-Validation procedure should be used to make predictions on the data not used to train the model. The procedure should use the number of competitions – 1 to train the network. The predicted competitive performance should be compared with the real competitive performance and the error computed (error – modeled performance – real performance). This procedure should be used for each data set and each model. Then the overall model should be computed as the mean value of modeled performances. Mean error and standard deviation of the modeled performances should also be calculated.

Questions: Use of Cameras and IMU?

What will we measure during training? Distance covered or stroke rate , strokes, turn, dives ?

**Data collection**

Training Load – km [file:///C:/Users/Lenovo/Downloads/Edelmann\_Hohmann\_Henneberg%20(2).pdf](about:blank)

The competitive performance of the athletes will be transformed according the FINA points.

**Data Analysis**

The first analysis should determine the influence of the 2 week taper phase prior to all competitions. The second analysis should determine the influence of the high load training phase three and four weeks prior to the competitions and lastly the third analysis should determine the influence of a 4 week phase prior to the competitions.

These models should consist of 10 input neurons, 2 hidden neurons and 1 output neuron. The 10 input neurons are necessary to account for 2 weeks with each of 5 training loads for both models.

The Leave-One-Out Cross-Validation procedure should be used to make predictions on the data not used to train the model. The procedure should use the number of competitions – 1 to train the network. The predicted competitive performance should be compared with the real competitive performance and the error computed (error – modeled performance – real performance). This procedure should be used for each data set and each model. Then the overall model should be computed as the mean value of modeled performances. Mean error and standard deviation of the modeled performances should also be calculated.

**Wearable sensors**

Physiological information such as heart rate, breathing effort, or oxygen saturation is very important to understand if an athlete is training at their physical limits, allowing the coach to adapt and improve the training session parameters, which in turn improves the athlete’s performance. Therefore, in this system, heart rate and oxygen saturation are also measured by means of the two following sensors.

The most conventional approach used in the majority of the research works is based on inertial sensors placed on the athlete, including a Micro Controller Unit (MCU) inside a sealed waterproof case [[4](about:blank#B4-sensors-21-05162),[15](about:blank#B15-sensors-21-05162),[16](about:blank#B16-sensors-21-05162)].

A less invasive approach is to employ video cameras, usually placed underwater in the swimming pool [[17](about:blank#B17-sensors-21-05162),[18](about:blank#B18-sensors-21-05162)]. Such video systems can be installed in the field with no need for encumbering the athletes with carrying additional hardware and are able to track motion and to analyse the performance from a mechanical perspective

Nevertheless, wearable devices are both less expensive and easier to set up in different scenarios and locations and allow for the acquisition of physiological data such as heart rate, breath rate, body temperature, motion, and position information, among others.

**IMU**

Several swimmers were monitored and the approach detected swimming bouts, laps, and swimming technique at the macro level. A statistic correlation between the sensor values and the output results was proposed.

The main contribution is related to the commercial availability of such system for real time analysis by coaches and teams.

**Swimming Analytics Framework**

Considering voting, three different combination techniques have been used: majority voting, average of probabilities, and product of probabilities [[28](about:blank#B28-sensors-21-05162)].

Training data is used in model construction, while test data is used to evaluate the performance of the model. The predictions provided by the model in the test data are then compared with the correct labels, and thus, the accuracy of the classifier is calculated.

Another approach that tends to compensate the rigidity of splits is cross validation, where experiments are often repeated with different random splits into training and testing datasets. Usually, the dataset is randomly split and evaluated from 5 to 30 times, and the mean and variance of the criteria are reported. Techniques such as cross validation and k-fold validation help guard against randomness, in particular data splits, and allow for sounder results. The k-fold validation involves splitting the data in k parts (folds), using (k−1) parts for training and the remaining part for testing. This is repeated k times, considering all possible testing sets, one at a time. There are multiple advantages, such as considering that every example from the original dataset has the same chance of appearing in the training and testing set, or the ability to perform validations when data are scarce.

### 5.3. Post-Processing Decision Support Tool

The results of the acquisition and learning processes can be used after post-processing to support coaches’ decisions on athletes’ training and performance evaluation, namely by taking advantage of information as the swimming style, turns, average speed, and stroke counting, which are available in real time.

The framework was deployed through a computational application developed in Python 3.8.0 using Visual Studio Code, allowing for visualisation of the data transmitted by the wearable module in real time as well as analysis

presents an example of the real-time received data used by the algorithm. Namely, using accelerometer and gyroscope information, along with heart rate data, it is possible to show turn detection and stroke type, allowing for decision support and correlation between values.

Considering the results obtained, it is revealed that. for stroke classification purposes, a window size of one second seems to better grasp the swimmer’s movement, as it achieves a 92.42% F1

the use of stacking with Naïve Bayes as a meta-classifier has the best performance. It is also important to note that those feature representations use less sensor data than others, which seems to denote that stacking with Naïve Bayes, as a meta-classifier, grasps movement better when less features are used, which is an important insight.

Considering only the 1 s feature representation, the one that presents the best results regarding stroke classification, Random Forest seems to be the best classifier when using ensembles, which might be related to the robustness of the Random Forest algorithm itself.

**Random forest**

All statistics were calculated using R software [[23](about:blank#pone.0254538.ref023)] and implemented with the base and randomForest packages to fit linear regression and random forest models, respectively. The parameters of the random forest were tuned by making use of a cross-validation based technique. Five-fold cross validation was run 100 times in conjunction with a grid search for selecting model parameters including the number of variables to sample at each split in the tree, and the number of variables sampled as candidates at each split in the tree. Given the randomly sampled nature of random forests, repeated evaluations provide a more robust selection for the tuning parameters

**Macro average F1 and Random forrest classifier.**

Training Load – Multi Layer Neural network for training load (km).

Average stroke and Speed counting -