**Literature Review**

An important aspect of this project thus far has been the background research; namely, the examination of other studies that can help to provide insight into how we should carry out this project whilst exposing us to other types of models and their applications with real-world data. This is, therefore, an integral part of the initial stages of the project, which has provided us with numerous justifications of methodologies and architectural decisions, ideas on which directions to explore within the experiment sets and model predictions sets, and proof of concept insofar as the adaptation of recurrent neural networks to the learning from and assessing of human activity recognition.

Deep learning is shown to be of high applicability to the more elaborate behaviours in human activity recognition (HAR), as it substitutes the manually designed feature extraction process with its own internal feature extraction, which is a great improvement over manual feature extraction as human activity lacks the robust physiological basis that benefits other fields such as speech recognition [16]. However, a prominent problem highlighted in previous research was the problem of training long sequences: training a sufficiently large RNN may result in it memorising the entire input-output sequence implicitly, leading to poor generalisation. A proposed solution to this was to introduce ‘breaks’ into the RNN where internal states are reset to 0, while the resets occur after every training batch with a fixed probability of occurrence; this is a technique to bear in mind if our RNN has trouble generalising when the sequence length is particularly long [16]. The importance of utilizing LSTM units to replace some layers of the RNN to solve problems of input-output weight conflict has also been heavily explored in previous research, along with the pros and cons of segmenting what is usually sequence data, with models being trained on segmented data due to the ease of training with corresponding labels but then being able to adapt to work with sequential data [17].

Different ways of utilizing the outputs of RNNs have also been explored: one study takes full advantage of deep RNNs in modelling long-term contextual information of temporal sequences by proposing a hierarchical RNN system for skeletal-based action recognition, where the skeleton is divided into 5 parts. Each part is sent through separate RNNs, the results of which are then further combined in stages until we have only one output. This enables us to capture temporal representations of both low-level body parts and higher-level parts [18]. Bidirectional RNNs are also used here so each part of each sequence utilizes both past and future context [18]. Both of these architectural features need to be kept in mind as a possible implementation in our model to model both low- and high-level features as well as both past and future context.

The combination of using LSTM-based RNNs on the outputs of convolutional neural networks (CNNs) for HAR in videos has also shown to be of great potential, with the ability to take 3D images (i.e. 2D images over time as video) of human movement, extract features from these, and use these to train an RNN which is used to recognize actions based on the temporal evolution of features [19]. These sorts of architectures are significant in that they learn compositional representations in both space (through CNNs) and time (through RNNs), which allow for the images of data to be learned by the models to learn temporal dynamics and perceptual representations and, for sequences of 2D activity data, lead to stronger results for activity classification than similar shallower models where temporal relationships can be more easily lost [20]. This type of architecture (where we would apply a CNN to an ‘image’ of suit-data) may prove to be a viable option if raw measurements or computed statistical values do not serve as viable sources of inputs to the RNNs.

It’s also been shown that this sort of architecture is also extensible to activity recognition in sensor-based data (as opposed to only from images) using a combination of convolutional and LSTM recurrent layers, which specifically has shown promise in the use of accelerometer, gyroscopic, and magnetic sensor data to HAR tasks, while having further justified that LSTM-based models take advantage of learning the temporal feature activation dynamics which CNNs are not fully capable of modelling on their own [21]. Skeleton-based HAR using RNN models with LSTM units have been further explored within [22], where there was an emphasis on the use of joint-based measurements as data and on the use of dropout for regularization of the LSTM neurons to achieve better model performance; this partially influenced our project to focus on joint-based measurements from the suit data (e.g. ‘jointAngle’, ‘jointVelocity’, etc.) and also to include dropout regularization within the ‘rnn.py’ script. Further justification of the importance of analysing joint information of 3D skeletons for HAR within each frame (e.g. looking at each individual joint location within each frame) via spatio-temporal-based LSTM units within an RNN is highlighted in [23], which serves as a possible approach for our system if better contextual information is needed to be drawn from joint angle data.