**Project Overview**

**Motivation for Project and Purpose of Work**

Modern machine learning algorithms and methodologies has seen great success in regards to being applied to biomedical data in order to provide insights that assist specialists and help diagnose potential conditions that may afflict subjects. One example of this is DeepMind’s recent progress with AI-assisted eye scans to detect over 50 different types of diseases potentially in a subject’s eye as accurately as world-leading expert doctors [1]. Deep learning techniques have also been utilized by HeartFlow to help build 3D models of subjects’ hearts and assess the impacts of blockages on blood flow to the heart [2]. Based on these breakthroughs, among others, in recent years, it is very probable that AI-assistance will become the new norm in various clinics and hospitals around the world within the next decade [3]. Taking note of the prominent applicability of artificial intelligence to the analysis of human movement in particular, Imperial College London have undertaken a research initiative in collaboration with Great Ormond Street Hospital to investigate the applicability of various AI techniques to the analysis of subjects with Duchenne muscular dystrophy [4], a form of muscular dystrophy that predominantly affects young males under the age of 12 and severely impacts their movement ability to varying degrees. The hope is that great strides in treatments can be achieved by utilizing artificial intelligence to inform and make decisions on a subject-by-subject basis and potentially draw insights about the condition.

With regards to this project specifically that is undertaken as part of this research initiative, we wish to investigate the applicability of recurrent neural networks (RNNs) to the features of human movement data provided by body suit measurements captured from subjects with Duchenne muscular dystrophy (DMD). This will hopefully provide not only evidence on the applicability of these models to this sort of data, but also should provide important insights into the data itself and of the subjects providing it. These insights from the project will hopefully have a direct positive impact in the ability to assess subjects via the North Star Ambulatory Assessment (NSAA) and possibly provide insights into other features of the condition. Generally, in this project we are trying to gain insights about the body suit data that is captured as ‘.mat’ files (MATLAB data files) through the different measurements that are captured by the 17 sensors of the suit (joint angles, position, accelerometer values, etc.) of NSAAs or 6-minute walks assessments of the subjects by means of sequence modelling using RNNs. This use of sequence modelling is necessary to model the dependencies through time of measurements, and we are more likely to have a robust model if measurement values are treated as NON-independent with respect to time.

The overall goal of the project is therefore to provide a deliverable that includes a complete system that works with varying forms of suit data, learns from it, and provides insights about it, while hopefully being able to be adapted to new subjects, new NSAA assessments of existing subjects, and help inform specialists of the severity of their condition. A significant hope, therefore, is to provide a software solution that positively and directly impacts the lives of those with DMD through hopefully making their assessments easier and more accurate.

**Aims and Objectives**

With the overall aim of the project outlined and with the motivation for undertaking the work provided above, we now turn our attention to covering some of the main aims of the project. These include:

* Building a reasonably good model (with surrounding supporting scripts that are outlined later on) that, when presented with new, unseen ‘.mat’ files of body suit data, can give a reasonably good approximation of individual NSAA activity scores and an overall NSAA score (i.e. the accumulation of all individual activity scores). A prominent limitation currently, however, is the overall lack of data files: we have no more than 50 complete ‘.mat’ files in total for each of the 6-minute walk and NSAA assessments. This is primarily due to the fact that the data collection is currently an ongoing process and a large repository of previously-collected suit data from other subjects with DMD does not appear to exist that’s publicly available. Hence, an implicit requirement of the project is to be able to make the most out of the data we have available, such as using it to train a model to predict different things, use different measurements contained within the ‘.mat’ files, look at applying statistical analysis on the raw data, and so on.
* Being able to use trained model(s) to gain insights into the most influential activities and measurements from the ‘.mat’ files on overall NSAA score and to identify activities that correlate highly with overall assessment. In doing so, it could possibly enable the reduction of 17 activities needed for accurate overall NSAA assessment to far fewer if only a few are needed to correctly assess the subject. The conclusions that we could possibly draw from the project, therefore, hopefully have the potential to aid specialists in the practical undertaking of the assessments through minimizing the amount of testing the subjects have to do.
* Investigating the impact of training models on different types of source data directly. For example, we’d like to see whether or not it’s possible to train models on natural movement behaviour data sets to the same standard as if we were using NSAA data sets when training towards overall and individual NSAA scores. If this were to be the case, then there exists a real possibility of not requiring the NSAA assessments to be completed by subjects at all, and instead simply requiring the subject to undertake natural movement instead, which may be significantly easier and/or more practical for subjects.
* Building models that are trained only on one ‘version’ of assessments of subjects and attempt to generalise to subsequent versions. Here, by ‘version’ we mean an assessment of a subject that takes place at a certain time, with subsequent versions being the same assessment but taken 6-months later on. For example, a subject’s initial NSAA assessment would be stored as the subject name ‘D4’, and when that subject returns 6-months later to undertake their subsequent NSAA assessment the resultant data file would be stored as ‘D4V2’. The hope is therefore to train models on non-‘V2’ files and generalise to newly presented ‘V2’ files. This should provide an advisory tool for any specialists wishing to assess how a subject’s conditioned has progressed during the time between assessments.
* Looking into how possible it is to build models that generalise well to new subjects and the system settings needed to achieve this; by this, we mean models that are able to assess subjects that they have never come across before during training (which differs to the previous bullet point, which looks at new data from existing subjects). Therefore, if this were to be the case then we would be able to extend the applicability of this system to not only new assessments of the existing subjects but brand new subjects to the overall research initiative. There are numerous techniques that we would have to look into if the models have a problem generalizing, and so a large amount of the model predictions sets will focus on this aim.
* Package all the scripts and models necessary for a specialist or any other researcher wishing to use any of the built tools in a way that is easy to use and gives intuitive output. This requires us to construct the system in a way where it is possible to be uses by others outside of the development environment in order to be practically applicable to achieve the aims outlined above.

With our overall project aims outlined above, it’s also useful to cover some of the objectives that we intend to achieve in order to complete these aims, many of which will be investigated within their own experiment set or model predictions set. These include:

* Comparing models built from different measurements (e.g. joint angles, acceleration values, computed statistical values, etc.) on their performance of evaluating unseen sequences of data to an accurate D/HC classification and overall/single-act regression of NSAA scores.
* Evaluating the ideal values for different sequence setups for the data going into the model with respect to the performance for various output types. This includes finding the ideal sequence length, sequence overlap, and discard proportion of frames within the sequences.
* Investigating the ideal number of features needed for the raw measurements and computed statistical values to train a model. This will involve a trade-off, with more features providing more of the inherent variance within the data and fewer features making it easier for the model to learn from.
* Looking into how well models performed when evaluated on files from a different source directory than their own. For example, we’d like to investigate the potential to models built from NSAA files and assess subject files from the natural movement behaviour data set. If this is possible, then with the finished models we could use these to assess a subject based solely on their natural movement, which might be much more practical than requiring the subject to undertake the NSAA assessment.
* Investigating how well models perform when they are familiar with the subject as opposed to when the model has never seen the subject before in training (even if it was trained on different data from the same subject than was used for assessing the subject).
* Assessing the applicability of generalisation techniques that includes downsampling the data, adding Gaussian noise to the data set, concatenation of features for multiple measurement types, and the leaving-out of anomalies within the subjects.