Insert significance of this literature review to project

I.e. defining human movement patterns into recognisable and comparable information

# Solving Data mining Problems through pattern recognition

## Types of pattern recognition

The fundamental process of performing pattern recognition involves processing the input data and assigning a label to the data based upon its characteristics, which is determined by the pattern recognition system.

The processed label is just one of the many labels that help identify the input data, as it defines one of the many unique characteristics of the processed data. The label is utilized for comparisons of a separate data set to determine whether the new data set matches the processed data set.

The two common categories that the pattern recognition system defines a label are either classification or estimation, where the output label is dependent upon the vector input to the system.

*Insert photo of pattern recognition system i.e. an abstract view similar to textbook*

### Classification

This category is based upon a set of observation parameters that define a finite set of labels, with no sequential order, to the observed input data. The parameters are set prior to the analysis of a sample data set and typically have predefined initial classes otherwise known as a group of labels that have similar or identical characteristics.

### Estimation

In contrast to the classification pattern recognition, this category is more quantitative as it has an infinite set of labels to assign. The estimation category is related to organising numerical input data into either exact values or if it falls within a range of values based upon the defined mapping.

### Uncertainty

Classification errors typically occur when the input data being processed have similarities with two or more separate classes, the system typically assigns the label to the highest probable class. The classification category is dependent upon whether the label matches or not, whereas estimation is simplistic in its approach of assigning the labels to its appropriate or closest matching classes by comparing sequentially throughout all infinite classes until the best match or its closest approximate is found.

## Models

The internals of a pattern recognition system can consist of one or a combination of models for classifying the input data for analytical comparisons, the standardized models are listed below.

### Fixed Models

This model is simply comprised of closed-form equation/s that classify the outputs based upon the mathematical calculation. Prior knowledge is required in developing this particular model in order to formulate the correct mathematical equation.

This model is simplistic however does not cater to varying parameters as the calculation is based upon a known constant for that specific context, the next model addresses this issue.

### Parametric Models

Data-driven models are similar to fixed models in the way they are structured, the key difference is that the equation takes several parameters to produce its output, thus a parametric model. It is far more flexible than the former model as it creates mappings dependent upon these parameters. These mappings once defined, are internally compared against each other in order to select the most viable mapping closest to the desired output.

Insert typical model

### Nonparametric / Data-driven Models

The model centres around handles large volumes of unknown data typically read from data acquisition systems. The model requires an analysed data set to compare the new incoming data with the predefined data set.

### Pre-processing

This particular model is used for complex situations where classifying the input data to the output labels is not possible when the quantity of data is of a larger magnitude than what a nonparametric model can process in an ideal time frame. Though the nonparametric model would be able to complete the task, it would take a sufficient amount of time and data to complete the pattern recognition.

Pre-processing takes into account several key factors to look out for, as a result it decreases the time frame for each calculation to be completed.

Simply put this model takes a large magnitude of data inputs and screens it accordingly to the internal parameters, the outputs of this model is filtered data that is routed typically into a non-parametric model for labelling however it can also be applied to parametric models.

## Learning Algorithm

The main algorithms of concern are the non-parametric models present in the Machine Learning Toolkit (MLT), as the context of this project will require the manipulation of continuous data being input into the system via the IMU.

### Nonparametric

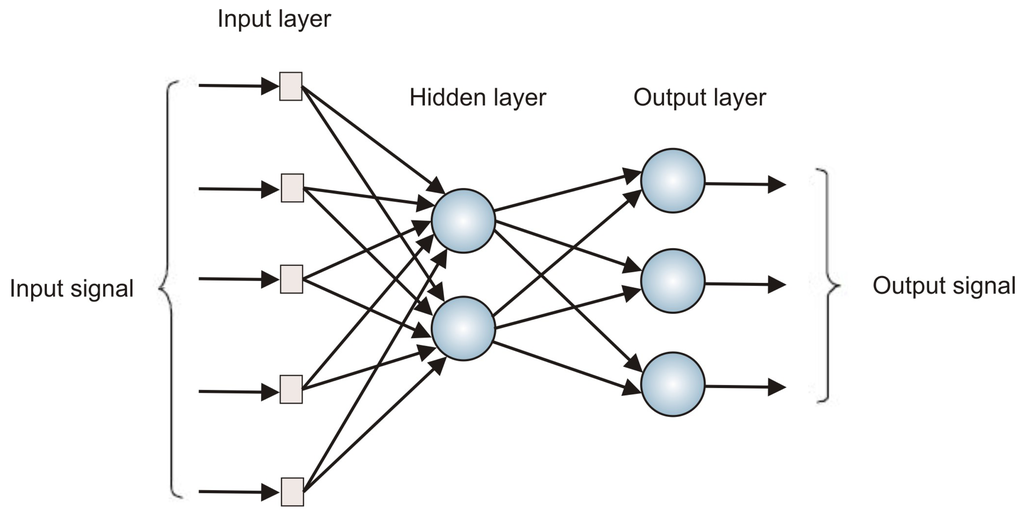
The following algorithms that fall under this category, will group data points based upon the training instructions and the algorithms methodology.

The nonparametric algorithms fall under two subcategories either unsupervised or supervised learning.

A supervised learning algorithm alters the parameters used to pattern the input data based upon the trained data. The unsupervised learning algorithm clusters the data with no specific order or pattern, the output data is essentially unlabelled.

#### Multilayered Perceptron (MLP)/Backpropagation (BP)

A form of neural network whose original purpose was to model neurological practices, the MLP achieves this objective by combining its architecture with the BP algorithm. This allows it to form nonlinear relationships between the incoming input data and the output labels formed by the algorithm.

In terms of just nonparametric modelling, MLP makes use of several tiers typically consisting of three layers as seen in (fig. reference) but for patterns of greater complexity additional tiers, known as hidden layers are required. For an n number of inputs to the MLP’s first tier, the input layer will consist of an n number of nodes that represent the weight of that input. The next tier otherwise known as the hidden layer, contains nodes where individually the values are the weighted sums of the input nodes due to a sigmoidal function ([Kennedy 1997](#_ENREF_1)). This node summation process continues until the output layer where the label is the weighted sums of the inputs.  http://www.mdpi.com/sensors/sensors-13-15613/article\_deploy/html/images/sensors-13-15613f7-1024.png

The MLP output is used by the BP algorithm to obtain the desired results; this is achieved over several iterations of tuning the weights of the MLP via a mean-squared error (MSE) function ([Kennedy 1997](#_ENREF_1)).

Equation no. reference

where *i* and *j* are the indexes of *Ntrain* patterns of the training set and each output variable respectively. The trained output dij of *j* index output from *i* index pattern and *yij* is the calculated output from the model.

MLP/BP Training Flow Chart

#### Radial Basis Functions (RBF)

Similar to MLP in mapping inputs and outputs based upon weighted sums, the function however uses radial Gaussians (RBF equation) instead of the MSE function ([Kennedy 1997](#_ENREF_1)).

*Equation Reference*

where *y* is the output of the algorithm, *N* are the number of radial Gaussian functions, *X* is the vector input, *i* is the i-th radial Gaussian function, *Meani* is the location of the centre, is the spread, *h* being the overlap factor, *w0* is the bias and *wi* is the weighted sum of that specific radial Gaussian function ([Kennedy 1997](#_ENREF_1)).

RBF is similarly trained as BP however, its processing time can be improved when combined with K Means Clustering, and the combination of the two minimizes the variation between the training patterns to the Gaussian centres. This optimization is achieved by tuning the weights of each Gaussian such that the MSE is minimized on each iteration.

RBF Training Flow Chart

#### K Nearest Neighbours (KNN)

KNN is simplistic in its operation in that it stores the input and output pairs of a training set in a database, the size of the stored database is highly dependent on the dimensionality of the training set. The database acts as a reference for new input patterns however despite the pre-processed training pattern the limiting factor of this algorithm is in the overall dimensionality of the training set, essentially the more patterns in the stored training set the longer the comparison process will take to examine through the training set.

KNN utilises the Euclidean measure as a means of comparison between the patterns within the training to the input pattern, this equation allows the input pattern to find its closest match in the training set, thus its nearest neighbour.

*Equation reference*

where the distance calculated is of *i* dimensions between the input pattern to the closest pattern within the training set.

#### K Means Clustering

K means is a widely utilized algorithm primarily due to its simplicity in assigning or organising data into the nearest cluster. These clusters, otherwise known as classes are user dependent and thus influence the amount of data points per cluster.

These cluster centres are calculated based upon the mean distance () of each individual data point assigned to it.

*Equation reference*

where Ci is the ith cluster, xk is an element of set Ci, vi is the current cluster centre of Ci and is the distance between these two points ([Vathy-Fogarassy & Abonyi 2013](#_ENREF_2)).

Once it is centred the cluster is split in half and the data points previously assigned to the prior cluster become members of the new clusters, where the new clusters are evaluated using the same process as the above i.e. the cluster finds its centre by calculating the mean distance () of the surrounding data points. After the original cluster is split the process will iterate until the number of clusters specified by the user is met. The data points are assigned based upon their distance to the nearest cluster.

K-Means Training Flow Chart from ([Kennedy 1997](#_ENREF_1))

* + 1. Initialize ‘k’ number of clusters
    2. Begin classification of training set.
       1. Associate each data point to its nearest cluster
    3. Calculate current mean (equation no.) of the cluster to find the new centre of the cluster incorporating the new data set distances
       1. Repeat until the mean distance is less than the value specified by user
    4. Split cluster in half and distribute the data points, based upon the highest deviation (equation no.)
       1. Store the cluster centre
       2. Repeat ii) 1) until the ‘k’ number of clusters is achieved

K Means Clustering typically creates a range of 2 to clusters for data estimation purposes however for classification the number of clusters increases depending on the context ([Kennedy 1997](#_ENREF_1)).

Elaborate on the types of Algorithms

Training and Testing

# References

Kennedy, R.L. 1997, *Solving Data Mining Problems Through Pattern Recognition*, Prentice Hall PTR.

Vathy-Fogarassy, Á. & Abonyi, J. 2013, *Graph-Based Clustering and Data Visualization Algorithms*, Springer.