K-Nearest Neighbor Algorithm: Proposed Solution for Human Gait Data Classification

Shadi Eltanani, Tjeerd Olde Scheper and Helen Dawes
School of Engineering, Computing and Mathematics
Faculty of Technology, Design and Environment
Oxford Brookes University
Oxford, United Kingdom
Email: {seltanani, tvolde-scheper, hdawes}@brookes.ac.uk

***Abstract*—Gait is a well-known motive means for humans. It is both energetically demanding and reflects several of human physical, mental and energetic disorders. Detecting these abnormalities can help medical professionals for better modelling and detection of biosystem chronic diseases, which enable timely treatment of patients and help control of the diseases' spread. In this paper, K-Nearest Neighbour (KNN) machine learning classification algorithm highlights the comparison between the gait patterns of normal healthy individuals and the patients suffering from irregular gait patterns caused by physical disorder conditions, including strapped muscles. Moreover, the Cross Validation test has addressed to examine how accurately the model fits the real-world clinical data. The experimental results show that the KNN algorithm can effectively be a robust learning classifier in classifying normal and abnormal human gait features. The classification performance of our proposed model is 67.7%, and its effectiveness has evaluated at a minimum square error rate.**

***Index Terms*—Human Gait, K-Nearest Neighbour (KNN), Cross Validation, Machine Learning, Square Error Rate, Classification.**

I. Introduction

GAIT is a popular identity in human beings and it basically describes the complex metabolic energy process a person exerts. Several health-related consequences are attributed to the imbalance of metabolic responses the human gait reveals. This chaos contributes to major disorders in the biosystem, including physical immobility (Muscle Atrophy), mental disturbance (Alzheimers, Parkinsons), chronic energetic imbalances (Diabetes, Obesity, Malnutrition), and Neuromuscular deficiency (Cerebral Palsy, Myelomeningocele, Muscular Dystrophy, Rheumatoid Arthritis, Stroke, Poliomyelitis) [1]. Consequently, the analysis of gait signal receives a great attention in the medical field by physicians. This is not only because it helps in diagnosis of biosystem associated diseases, but also assists in the effectiveness of therapeutic solutions and interventions.

Inspired by the recent advances of Machine Learning (ML), several research works attempted to utilize ML algorithms in a number of gait analysis applications in healthcare to potentially diagnose gait disorders or to predict the risk of physical instability or to monitor changes in movement based on gait patterns.

K-Nearest Neighbour (KNN), for instance, is a powerful supervised machine learning technique that depends on learning from data to solve classification problems with high classification performance. The rationale of this learning analysis tool is to model whatsoever data based on a binary criterion, which thereby aims to assign a categorical label to every input sample. Within this context, the binary classification basis of KNN has significant practical implications in early detection and diagnosis of gait pattern abnormalities. This vigorously prompts clinical interventions that can help prevent the imbalance of biosystem chaos and subsequently assist in motor recovery, especially in people suffering from loss of independence due to chronic diseases. This kind of classification could significantly progress to various future KNN application, in particular, as gait diagnosis. The selection of KNN technology, ranked as a trustworthy gait classifier, is primarily driven by its ability to build robust predictive models. This depends on the optimal k nearest Neighbours as a Euclidean distance in a Euclidean space by considering the majority of votes from Neighbours labels. Many prior research work has considered using KNN technique as a gait classifier for various scenarios. The authors in [2], for instance, utilised the KNN classifier to discriminate the individuals gender in the monitoring and surveillance operations. Moreover, the work of [3] reported different machine learning schemes, including the use of KNN approach to detect the neurological abnormalities in brain using various human gait patterns. Furthermore, it is reported in [4] that humans distance walking at the same speed have been recognised in the context of KNN classification process. Nevertheless, the KNN based-rule researched in [5] is introduced to detect healthy gait feature subjects caused by Parkinson disease. Additionally, the KNN principle, based on the distance between the right and left gait skeletal joints, is employed by authors in [6], to build a secure authentication system so that human cognitive behaviors can be recognised. Likewise, the research in [7] presented the KNN as a data classifier to differentiate biometric related-gait features based on collected data from a wearable sensor device, which fundamentally considered the best matched gait signature metrics as a classification measure to identify individuals. The KNN classification rule is implemented in [8] to classify human activities based on their sitting, standing,

978-1-6654-2744-9/21/$31.00

running and walking setups. Similarly, the authors in [9] developed the same approach as in [8] but the classification process was run on Micro-Doppler data of gait features. In [10], the authors addressed the use of KNN approach to classify pedestrian motion data according to the placement of the inertial measurement units (IMUs) device on different parts of human joints, including fixing hand, swinging hand, pocket, and backpack. Adding to that, the research in [11] tested KNN classifier to detect a host of a turning activities of a lower limb prosthesis using a wearable sensor data.

This paper implements a KNN learning algorithm to automatically classify normal and pathological gait data patterns for a number of individuals. The model is evaluated on a real-world medical data, where the results showed that the highest classification performance associated with significant low error rates that makes it the best classifier depends on the scale size of the data split.

The rest of this paper is organized as follows. In section II, the theory of KNN approach is described. In section III, the methodology of our experiment is reported. In section IV, the simulation results are analysed and discussed. And finally, some conclusions are drawn in section V.

II. K-Nearest Neighbour (KNN)

A. Overview of KNN

The KNN is a non-parametric supervised machine learning classification technique. The principle of KNN relies on computing the Euclidean distance between the test (unknown data points) and the training data samples. Let $\mathbf{x} \in \mathbb{R}^{n\times d}$=$(x_1, ..., x_n)$ be the matrix of features, where n is the number of training samples and d is the number of features. For a given an arbitrary point in the unknown samples set x_o, the Euclidean distance in the feature plane $\mathbb{R}^p$, where $p = 2$ is a real number, can be formulated as:

$$d_i = \|\mathbf{x_r} - \mathbf{x_o}\|_p = \left(\sum_{i=1}^{n} |x_i - x_o|^p\right)^{\frac{1}{p}} \tag{1}$$

To classify a number of features into M classes, then the outcome of classified entities can be presented as Ω=$\{\Omega_1, ..., \Omega_m\}$, where $1 \leq m \leq d$. Choosing the k training samples with the minimum distance to the unknown data point x_o, the KNN algorithm calculates the number of Neighbours assigned to each data class $l \in \mathbb{R}^{1\times d}$=$(l_1, ..., l_d)$ existing in the training set $\mathbf{S}_r$=$\{(x_1, l_1), ..., (x_n, l_d)\}$, where $\mathbf{x}_r \in \mathbb{R}^{n\times 1}$=$(x_1, ..., x_n)$ is the training example associated with $\mathbf{S}_r$. Each member in $\mathbf{S}_r$ corresponds to a class label in Ω. The process is fundamentally based on estimating the conditional probability for each class as an empirical fraction. This can mathematically be given by:

$$\mathbf{P}_r = p\big[m(l) \in l \mid \mathbf{x} = x_o\big] = \frac{1}{k} \sum_{i\in\mathcal{N}(l,\mathbf{S}_r)} \mathbb{I}(\mathbf{x}_r \in l) \tag{2}$$

where $\mathcal{N}(l, \mathbf{S}_r)$ are the indices of the k nearest data samples to l in the training set $\mathbf{S}_r$. $\mathbb{I}(.)$ is an indicator function expressed as:

$$\mathbb{I}(w) = \begin{cases} 1, & \text{if } w \text{ is True} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

III. Methodology

In this section, all gait features were examined using the implemented KNN algorithm. Software routines were developed in Matlab R2018b for analysing the gait data samples and to perform several tests, including tests to examine effects of tuning the k parameter with the aid of cross validation approach on classification performance. The generalisation performance of the KNN model was determined by estimating the average of the mean square error of training and cross validation tests against various iterated values of k parameter. To explore the KNN model and its role for gait features classification, the considered dataset involves the gait patterns of five different persons. In particular, the distribution features of the dataset monitored during continuous walking steps were utilised to develop the KNN model and to test classification performance of the model. The KNN model aims for the automatic recognition of ill-apparent (strapped) and well-apparent (normal) gait types from their respective gait-patterns.

A. MOReS Dataset

The Movement, Occupational and Rehabilitation Sciences (MOReS) centre at the Faculty of Health and Life Sciences at Oxford Brookes University has provided a MOReS dataset that includes two features groups of gait data Accelerometer sensory (well-apparent and ill-apparent) patterns for 5 different persons. The normal well-apparent walk represents no known injuries or abnormalities in gait patterns, in contrast to the stiffness represented in the ill-apparent or strapped walk.

B. Dataset Splitting

The crucial part of KNN, however, is to randomly split dataset into training and testing sets. This is to ensure that both sets have the even and fair distribution and that they are independent of each other, and to provide a promising results as to the real performance. One portion of data can be used for fitting and evaluating on the other set, which allows an unbiased estimation of generalisation mean error rate.

C. Cross Validation

Cross validation is a standard test, which is commonly used to avoid biasing the data performance results and to test the ability of the classification model depending on several combinations of the testing and training sets. In this process, the performance of the prediction of each gait class is maintained by a systematic exclusion of a small portion of gait data during the training process, whereas the excluded gait data points can be used to test the trained model. This cycle of data validation is repeated until each of data points is included in the testing data set. Since the number of data points available in gait dataset is limited to 6000 gait trials

that correspond to 5 different persons, it is vital to validate gait data using different scales and observing the associated error rate of each data split portion. It is important to run MOReS dataset over different data splitting scenarios to draw a conclusion on the accuracy of KNN classification model.

D. Measuring Mean Square Error Rate

To gauge how well the KNN classification model algorithm works, it is a necessitate to measure the mean square error (MSE) rate of original data samples against the predicted data points to be able to measure how costly the predictive model is. In practise, this means that the model well-performed when the average error rate $\mathcal{E}$ reduces to minimal. The MSE can be given by:

$$\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y_i}\right)^2 \tag{4}$$

where y_i is the original data samples (N) and $\widehat{y_i}$ is the predicted data points.

E. Optimal Value of K in KNN Model

As the name of KNN model suggests, it is an important consideration to define an optimal value of k parameter in KNN model, which defines the number of neighbours to consider when classifying the data points and impacts the complexity and the effectiveness of the model. Moreover, the optimum value of the k depends on various factors, including the size of the dataset and the distribution of data points in their coordinates space. Also, the k variable has to be best selected explicitly using the method of cross validation, in order to produce a minimum rate of classification error and to reflect on the accuracy of the KNN model.

IV. Simulation Results

In this experiment, a 70-fold cross validation test was applied in which 6000 gait data points were divided into approximately 86 subsets, with a majority of 90% of data are allocated for training whereas the remaining 10% are included for the testing scenarios. This setup shows a minimal test and cross validation error rates. Additionally, the optimal value of $k = 4$ is observed, where the best performance of KNN model is maintained. To reflect on the performance metric of KNN model, the mean square error quantifiable approach is assessed, where the percentage of unclassified data points is averaged 100 runs over total number of training and testing data sample sets in order to ensure that fair results were produced. Figure 1 is an example of MOReS data Histogram distribution of a well-apparent (normal) and an ill-apparent (strapped) subjects. This illustrates the relationship between normal and strapped gait patterns, grouped by their corresponding labels. It clearly reveals some qualitative differences between these two groups, such as increased variability, significance decrease in central tendency, and high obvious skewness (skewed to the right) in the normal well-apparent plots. The features experimented from Figure 1 were used to KNN model as well as to test its capability to classify or

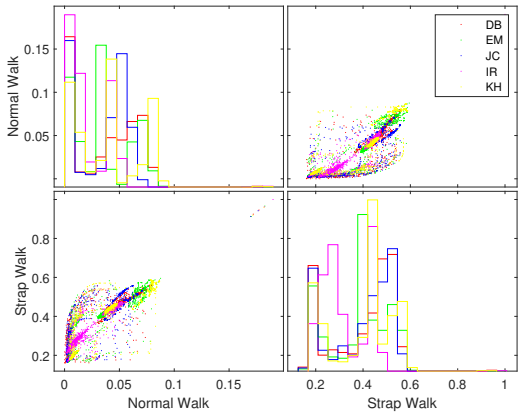

Fig. 1: Pre-trained class labels associated with their medical conditions

discriminate between the gait patterns that correspond to the 5 different data labels, which are identified as EM, IR, KH, DB, and JC respectively. These labels are defined in this way due to data protection rules and regulations. The prediction outcome of the trained KNN model is shown in Figure 2. It shows that the 2D space of predicted data is sparse enough in comparison to that of the original data distribution. Noticeably, all label patterns were clearly classified. Due to limited mobility of strap walk caused by stiffness, the spread of some features of the predicted strap walk data is remarkably clustered close at some interval points and becomes more dense than their counterparts of their predicted normal walk classified labels.

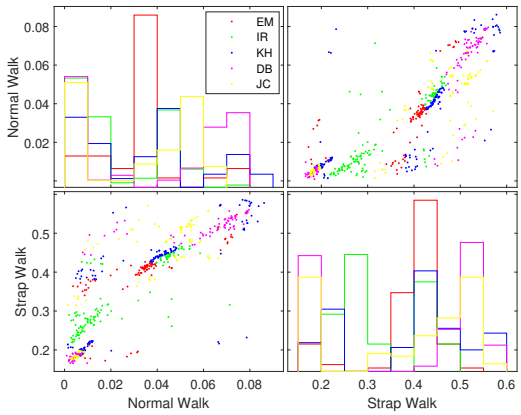

Fig. 2: Predicted class labels associated with their medical conditions

The ability of KNN model to classify the associated data labels based on their corresponding features is depicted in the confusion matrix of Figure 3. In this layout, the correctly classified labels are located on the main diagonal from top left to bottom right that correspond to the number of times the two true and predicted class labels agree. It shows that how far the predicted labels are deviated from their corresponding actual

average or mean values. It is also a measure of the joint variability of each data label with itself, and an indication of test data points that have been well-classified (positive correlation) by the model. The off-diagonal matrix elements represent the incorrectly classified (negative correlation) test data samples by the classifier. It is obvious that there is a strong correlation between IR and KH data labels across the diagonal line as well as a similar correlation in medical conditions patterns between the other three labels. The row-normalized row of

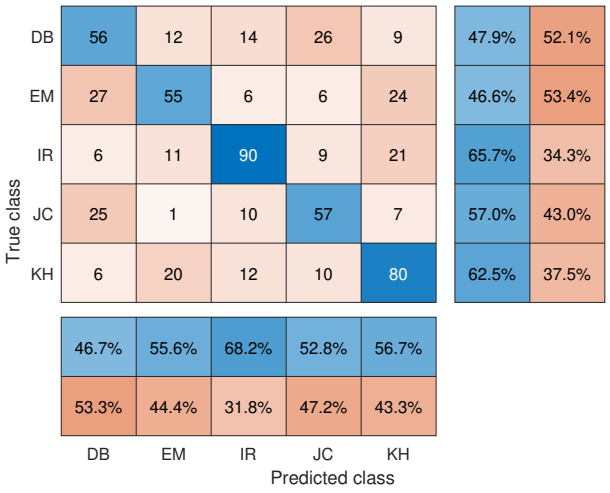

Fig. 3: Confusion Matrix

the main matrix displays the percentages of true positive rates and false positive rates for each true class, in contrast to the column-normalized column which displays the percentages of correctly (positive predictive values) and incorrectly classified (false predictive rates) patterns for each predicted class. The overall accuracy of the correct classified data patterns is approximately 67.7%. In Figure 4, the KNN model is run

Fig. 4: MSE rates with respect to various k values

100 times for several values of k. During this setup, the cross validation, training and testing tests were observed against the classification mean square error is evaluated accordingly. When the value of $k > 4$, the error rate increases significantly and hence the algorithm overfits. The cross validation error reduces to a minimum threshold when $k = 4$, whereas the lowest value of the testing error is recorded when $k = 10$. At this scenario of data splitting ratio, the KNN model showed reasonably good accuracy and consistency with significant minimum classification error rates. Both cross validation and testing error rates have been observed at several splitting ratios setup as seen in Figure 5.

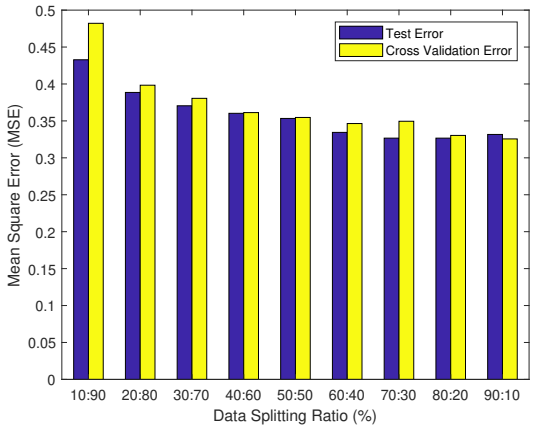

Fig. 5: MSE rates against different data splitting ratios

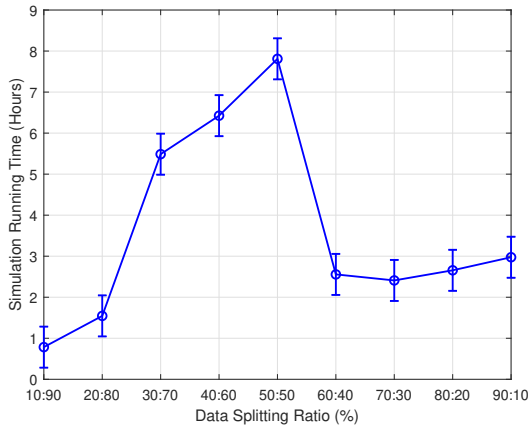

Fig. 6: Simulation time versus various splitting ratios

The best splitting ratio is when the testing and training sets of MOReS data divided quarterly, where the testing and cross validation error rates become minimum as well. It is worth noting that simulation time of the KNN model is observed. Figure 6 illustrates that time of data training increases gradually as the split ratio does. At the halfway scale of MOReS data splitting, the time of the data trained reaches the maximum peak, then it starts to decrease down. This error bar plot of Figure 6 generates a vertical error bar plot at each data point. The range of values in this error plot evaluates the lengths of each error bar above and below the data points, so that the total error bar lengths are twice in length the error values.

V. Conclusion

In this paper, an automated human gait classifier based on a robust machine learning tool, K-Nearest Neighbour is proposed. It clearly proves that useful data gait features can be extracted using Histogram distribution that can effectively and automatically classify well-apparent (normal) and ill-apparent (strapped) gait patterns. The accuracy of this model is demonstrated by means of estimating the average error rate between predicted and actual data patterns. The overall performance of the KNN model was nearly around 67.7%. This is due to significant similarity in the numerical values of the sensory collected measurements during walking period that were susceptible to wrong classification in the model. KNN model has a great potential for future medical applications in terms of identifying and predicting several health conditions.

References

[1] DeLuca PA. Gait analysis in the treatment of the ambulatory child with cerebral palsy. Clin Orthop Relat Res. 1991 Mar;(264):65-75. PMID: 1997253.

[2] P. B. Shelke and P. R. Deshmukh, "Gait Based Gender Identification Approach," 2015 Fifth International Conference on Advanced Computing and Communication Technologies, 2015, pp. 121-124, doi: 10.1109/ACCT.2015.66.

[3] P. Patil, K. S. Kumar, N. Gaud and V. B. Semwal, "Clinical Human Gait Classification: Extreme Learning Machine Approach," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1-6, doi: 10.1109/ICASERT.2019.8934463.

[4] M. W. Rahman and M. L. Gavrilova, "Kinect gait skeletal joint feature-based person identification," 2017 IEEE 16th International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC), 2017, pp. 423-430, doi: 10.1109/ICCI-CC.2017.8109783.

[5] J. P. Félix et al., "A Parkinson's Disease Classification Method: An Approach Using Gait Dynamics and Detrended Fluctuation Analysis," 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), 2019, pp. 1-4, doi: 10.1109/CCECE.2019.8861759.

[6] M. W. Rahman and M. L. Gavrilova, "Kinect gait skeletal joint feature-based person identification," 2017 IEEE 16th International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC), 2017, pp. 423-430, doi: 10.1109/ICCI-CC.2017.8109783.

[7] P. Tahafchi and J. W. Judy, "Freezing-of-Gait Detection Using Wearable Sensor Technology and Possibilistic K-Nearest-Neighbor Algorithm," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 4246-4249, doi: 10.1109/EMBC.2019.8856480.

[8] H. Huang, X. Li and Y. Sun, "A triboelectric motion sensor in wearable body sensor network for human activity recognition," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016, pp. 4889-4892, doi: 10.1109/EMBC.2016.7591823.

[9] Y. Lin and J. Le Kernec, "Performance Analysis of Classification Algorithms for Activity Recognition Using Micro-Doppler Feature," 2017 13th International Conference on Computational Intelligence and Security (CIS), 2017, pp. 480-483, doi: 10.1109/CIS.2017.00111.

[10] P. Kasebzadeh, K. Radnosrati, G. Hendeby and F. Gustafsson, "Joint Pedestrian Motion State and Device Pose Classification," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 8, pp. 5862-5874, Aug. 2020, doi: 10.1109/TIM.2019.2958005.

[11] Z. Liu, W. Lin, Y. Geng and P. Yang, "Intent pattern recognition of lower-limb motion based on mechanical sensors," in IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 651-660, 2017, doi: 10.1109/JAS.2017.7510619.