Comparison of Machine Learning Algorithms for EMG based Muscle Function Analysis

Praveen Kumar
Department of Electronics and Communication
Ramaiah Institute of Technology
Banglore.
E-mail: prempraveen12@gmail.com

Suma K V
Department of Electronics and Communications
Ramaiah Institute of Technology
Banglore.
E-mail: sumakv@msrit.edu

Manjunath C Lakkannavar
Department of Electronics and communication
Ramaiah Institute of Technology
Banglore.
E-mail: manjunathl@msrit.edu

Varun C R
Department of Aerospace and Engineering
Indian Institute of Science
Banglore.
E-mail: raghavendravarun88@gmail.com

2022 4th International Conference on Circuits, Control, Communication and Computing (I4C) | 979-8-3503-9747-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/I4C57141.2022.10057856

***Abstract* — Electromyography (EMG) is the process of measuring the electrical activity that is produced during the muscle activation and is also called as myoelectric signal. This electrical activity can be recorded to analyze the performance of any muscle. Machine learning algorithms can be deployed to predict the presence of abnormality in the muscles. In this study, we have worked on predicting the possibilities of cervicalgia, more commonly known as neck pain, by using distinct time domain features extracted from surface EMG signals recorded from SCM muscle, and performing the comparative analysis of different machine learning binary classifier algorithms like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Naïve Bayes Classifier and Logistic Regression. The feature selection was made using techniques like PCA and Pearson's Correlation Matrix. Maximum classification accuracy of 91.6% was achieved. If predicted well in advance, they can provide valid information to doctors, physiotherapists, yoga therapists or medical practitioners, who can then adapt their diagnosis and treatment for each patient.**

Keywords—EMG, Cervicalgia, SCM, SVM, KNN, Random Forest, PCA, Decision Tree, Naïve Bayes, Logistic Regression, Pearson's Correlation Matrix

I. INTRODUCTION

Electromyography or EMG is a biomechanical process and it is used to measure the activity of the muscle and it is done by placing electrodes in line with the muscle to be analyzed. Then measurement of the electrical signals sent from the nervous system to actuate that particular muscle is done. There are two main types of EMG, namely, surface EMG (sEMG) and intramuscular EMG. The latter is an invasive method which employs needle electrode that is injected into the muscle and it follows the muscle movements. With this method, significantly better accuracy, lower sensor noise, less muscle crosstalk and less distortion are achieved but, the obvious disadvantage is that it is painful and it can cause significant discomfort. Hence intramuscular EMG is rarely used and if it is used it will be for muscle that are deep within the body. So, the preferred method is sEMG where EMG electrodes are placed on the skin by employing a conductive gel and the electrical signals sent from the nervous system to the muscles are picked up by the electrodes. This mode of signal acquisition is easy to perform however, has the disadvantage of high sensor noise, muscle crosstalk and movement artifacts. Appropriate signal processing must be conducted to achieve better quality of signals.

II. RELATED WORK

SCM (Sternocleidomastoid) is one of the most overactive muscles that contributes to poor posture. It causes the head to bend slightly forward called as forward head posture, this problem has discussed in the papers [1],[2]. This muscle runs right from the clavicle and sternum area all way up to the Mastoid Process right behind ear. [3] works on the four neck EMG muscles like right trapezius, left trapezius, right SCM and left SCM. deployed SVM and RFDT machine learning classification models for the prediction and SVM achieved 88% and RFDT yields 81%. Comparison of prediction and data analysis using data driven techniques such as Linear Regression and Artificial Neural Network is conducted. Based on wrist motion and hand motions for EMG signal-based interface and quantified by correlation factor of LR and ANN are 0.62 and 0.57 respectively and estimated normalized RMS error measured 0.22 and 0.23 respectively [4]. Multichannel EMG signal acquisition device is proposed and muscles position have been selected from the palpation method. Again, a comparative analysis for different types classifications like SVM and random forest are presented for different kind of muscle positions. SVM achieves 87% and random forest yields 85% in the paper [5].

ML models are built for classifying the sitting postures like front, back, left and right of a person sitting on a platted chair and the data is collected for analysis. Models were built for this study using five different algorithms like Random Forest, Gaussian Naïve Bayes, Logistic Regression, Support Vector Machine and Deep Neural Network and DNN achieved 98% [6]. Proposed IoT based yoga posture recognition system by employing the deep convolution neural network a sensor based wireless sensor network and invited 18 subjects to perform 26 yoga postures. Evaluated different performance metrics like accuracy, precision, recall and F1-score, for tenfold cross validation revealed average F1-score is 0.9989[7].

Model selection in machine learning is the process of choosing the best suited model for a particular problem that you are trying to solve. Selecting a model depends on various factors such as dataset, favourable results and nature of the model itself. If there is luxury for labelled data we can opt for Supervised algorithms, else an Unsupervised algorithm. For studying any muscle related disorder, the simplest model

can be a binary classifier, which can predict whether the disorder is present or not. With more data, one can venture int a multiclass classifier problem. For our study, a set of people with or without neck pain were selected and labelled accordingly to estimate the presence of neck pain. And hence, we considered some of the popular binary classifiers such Naive Bayes Classifier, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Logistic Regression, and others to determine which of these algorithms will predict the data the most accurately.

Logistic regression is forecasting the results of a set of variables. As a result, the outcome must be a discrete or categorical value. It might be either Yes or No, 0 or 1, false or true etc., but rather than providing the precise result like 0 or 1, it provides the probability value that occur between 0 and 1.

The decision tree is the most effective and popular categorization and prediction approach. Each node in the network of a decision tree, which resembles a flowchart, denotes a check on a characteristic, every branch displays the results of the test, and every leaf node (or terminal node) has a class label.

Naive Because of assumptions are made by the Bayes algorithm are so improbable to be supported by empirical data, the algorithm is regarded as naive. To determine the sum of the individual probabilities of the elements, conditional probability is used. This indicates that, target class variable, the algorithm presumes that perhaps the presence or absence of a particular attribute of a group is independent of the presence or absence of any other characteristic (perfect independent of attributes).

The most often used algorithm for binary classifiers is Support Vector Machine (SVM). The extreme vectors that can assist create the best hyperplane are selected by SVM. The kernel selection becomes crucial here, and we can opt for either linear or non-linear.

KNN is the most useful algorithm in cases where resampling of the dataset is necessary. K is the crucial parameter in the KNN model, some points are to be noted for choosing the best value for K. Error curves can be used for different values of K for the training data and test data. If K is low, then model can lead to overfitting of high variance. For this reason, train error is minimal whereas test error is significant. The test error is reduced if we raise the value of K. Additionally, while picking the value for K, domain knowledge is very helpful. When thinking about classification model issues, it is important to understand that the value of K must be odd.

Compared to Decision Tree Algorithm, Random Forest is an optimized algorithm, which utilizes an ensemble learning. This model consists of many decision trees. A “forest” generated from this model is trained through bootstrap aggregating. The decision tree model's constraints are eliminated as the number of trees increase then increasing the precision of the objective result. Thus, it decreases the issue of dataset overfitting and improves precision. In medical analysis, it is used to predict the risk of a particular disease and is known to give the best results for data like ECG.

III. PROPOSED METHOD

Our study had 30 participants which included 15 Healthy subjects (13 males and 2 female) and 15 subjects who are suffering with neck pain. Target muscle to study the severity of the muscle was chosen to be SCM (Sternocleidomastoid). as suggested by the medical practitioner and Yoga expert from Ramaiah Medical College and Ramaiah Indic Speciality Ayurvedic Hospital, Bangalore, India. From each subject we collected 60 seconds duration of sEMG, both SCM left and SCM right muscles (two signals per subject), at a sampling rate of 700Hz using our custom Multi Channel EMG Acquisition Device [8]. Multi-channel EMG Acquisition device and acquisition of neck EMG signals from the subjects as shown in Fig 2 and Fig 3 respectively. Later, we extracted 16-time domain features on MATLAB. The significance of some of the important parameters which gave us the best results. EMG raw signals can be analysed by using MATLAB tool as shown in Fig 4.

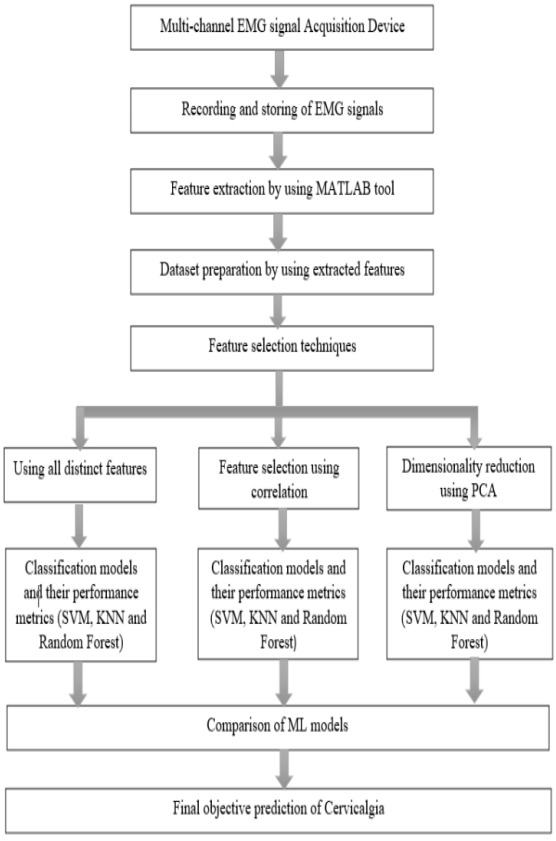

Fig 1: Flowchart of EMG Signal Analysis

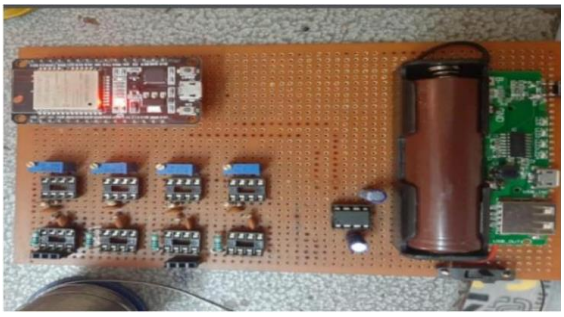

Fig 2: Multi-channel EMG Acquisition device

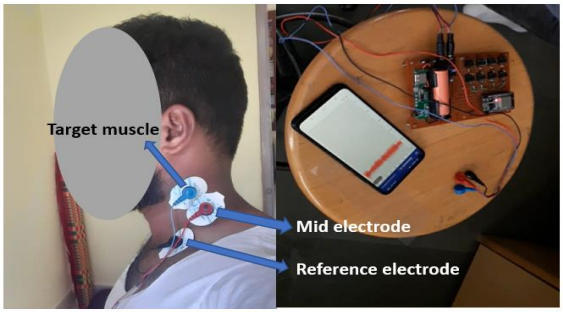

Fig 3: Acquisition of EMG signal from the participants

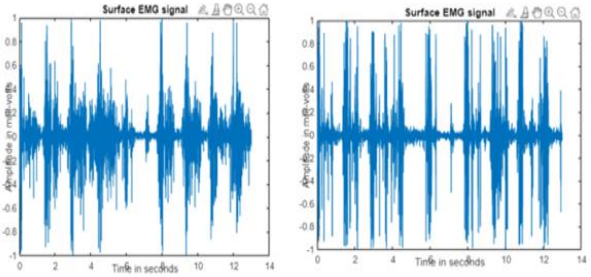

Fig 4: EMG raw signals viewed on MATLAB

Table 1: Time domain novel features and their importance

Time Domain Feature	Importance
Slope Sign Change (SSC)	SSC which measures signal frequency and it is the quantity of variation in the slope of the EMG waveform that occur within an analysis frame.
Mean Absolute Value (MAV)	It is a common characteristic utilised in myoelectric control applications and provides a convenient method for detecting the intensity of muscular contractions.
Simple Square Integral (SSI)	SSI manipulate strength of the surface EMG signal.
Average Amplitude Change (AAC)	It gauges the wave's vertical separation from the norm.
Myopulse Percentage Rate	It is the average of the Myopulse output where the absolute magnitude of the EMG signal is greater than the predetermined threshold level.
Enhanced Wavelength (EW)	The multiplexing technique used by enhanced wavelength multiplexes several carrier signals into a single signal by employing various wavelengths.
Wavelength Amplitude (WA)	It specifies the maximum amplitude of the EMG signal.

A. Time domain Analysis of sEMG signals

Time domain analysis is generally used for muscle force detection tool to measure the performance of the muscle weariness. Two existing time domain parameters are used to analyze EMG raw signal. That are Mean Absolute Value (MAV) and Root Mean Square (RMS) and these two parameters can be calculated by using equation (1) and equation (2) respectively [9].

$$MAV = \frac{1}{N}\sum_{k=1}^{N}|x_k| \quad (1)$$

$$RMS = \sqrt{\frac{1}{N}\sum_{k=1}^{N}x_{k,}^{2}} \quad (2)$$

Time domain features had the advantage of simple and effective calculation of the signal characteristics[10]. some impartant TDF and their importance as shown in Table 1.

B. Data Preparation

In this study a novel surface EMG dataset was created using appropriate number of channels. In this work 30 participants are involved, data set contains total 60 samples, which includes both 15 healthy subjects and 15 postural correction subjects each subject having two signals or two samples. In the data set '1' is labeled as healthy subjects and '0' is labeled as postural subjects or unhealthy subjects. 'id' variable which indicates the signals collected from the subject's id as '1 to 60'.

The following methodology has been considered to proceed with the machine learning classification models for prediction of cervicalgia or neck pain. In this study we were used three methods like 1. Using all 16 distinct features, 2. Using correlation and 3. Using principal Component Analysis (PCA).

1) Autocorrelation

The concept of autocorrelation describes the relationship between one or more variables. Those factors can also be input data aspects that were used in the forecasting of the target attribute. Correlation is a statical techniques which computes how one variable changes in relation with the other variable or other feature. Correlation is a bi variate analysis measures which describes the association between the different features. If two variables are closely correlated, then we can predict one variable from the other. Correlation plays a important role in locating an important dependent variables. It also used for the foundation for various modeling techniques. Proper correlation analysis leads to better understanding of given data. Positive correlation can be positively correlated with each other. If the two variables are negatively correlated with each other which means that when the value of one variable increase, then the value of other variables decreases. If there is no correlation between the two variables or two features which means that one variable can increase or decrease, then the value of other variable does not increase or decrease.

2) Principal component Analysis (PCA)

PCA is a method of reducing the dimensionality of datasets. minimize information loss while also improving interpretability. If we run large dataset or dataset which having the greater number of features then we can run it through the PCA, by using PCA we can reduce the variables or features.

Some important properties of PCA are

- Number of principal components is always less than or equal to the number of attributes.
- the priority of principal components decreases as their numbers increase.
- Principal components are orthogonal.

Step involved in the principal component analysis as shown in Fig 5.

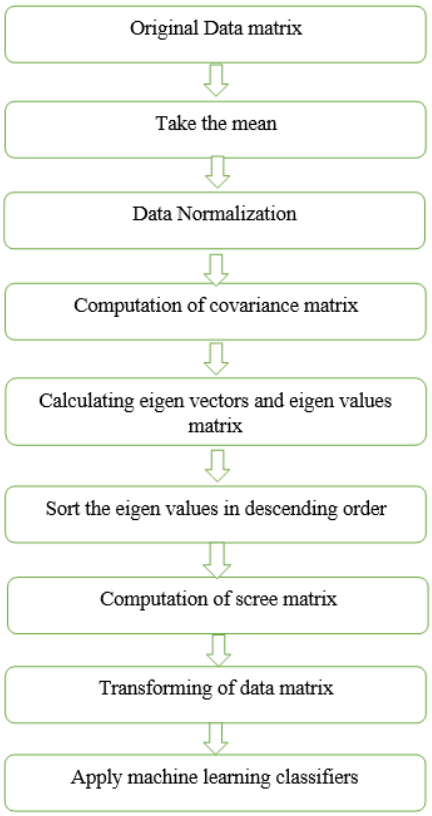

Fig 5: Flow chart for principal component Analysis

PCA performs the following operations in order to evaluate the principal components for a given dataset.

Standardization: first we start with the standardization then we have a covariance matrix computation and use that to generator our eigen vectors which is the feature vector, eigen vectors are like translation for moving the data. the range of the qualities must be standardized as the next step in the standardization process. lie within similar boundary. This process involves removal of the mean from the variable and scaling the data with respect to the standard deviation.

Covariance matrix computation in a multidimensional dataset, the correlation between any two or more attributes is expressed using a covariance matrix. Understanding how well the input data set's characteristics differ from the mean in relation to one another is the main objective of this stage. or to determine whether they are connected in any way. Due of their strong correlation, characteristics are redundant in their information. The covariance matrix must be computed in order to determine these relationships.

Feature Vector: this is just a matrix that contains the eigen vectors of the elements that we choose to keep as the columns. Here, we decide whether we must keep or disregard the significant principal components that we have generated in the previous steps. Higher the principal components higher is the variance and vice versa.

IV. Results and Discussions

Performing various classification models like SVM, KNN and Random Forest with three different methods.

Method-1: Training the model using all 16 features from the dataset,

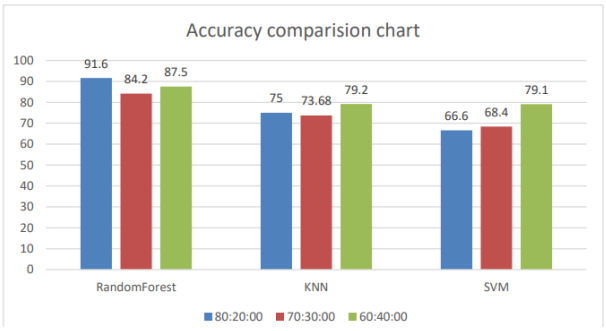

Fig 6: Accuracy comparion of SVM, KNN and Random Forest using all 16 features

From the above Fig 6, we can conclude that the random forest achieves the maximum accuracy (91.6%) at the split ratio of 80:20.

Method-2: Training the model using reduced features (correlation factors).

Correlation matrix visualization before feature selection as shown in below Fig 7. It shows that there some features that has a very high correlation with our target variables values and also some of the features have negate e correlation with the target value and some have positive.

Features are reduced by setting the arbitrarily threshold as 0.4, so we get the features like SSC, EW and WA and some features are removed by checking the multicollinearity. Then try to reduce those features having high VIF until each feature has VIF is less than 5 and with some tries 16 features are reduced to 3 features are MPRS, AAC and SSI.

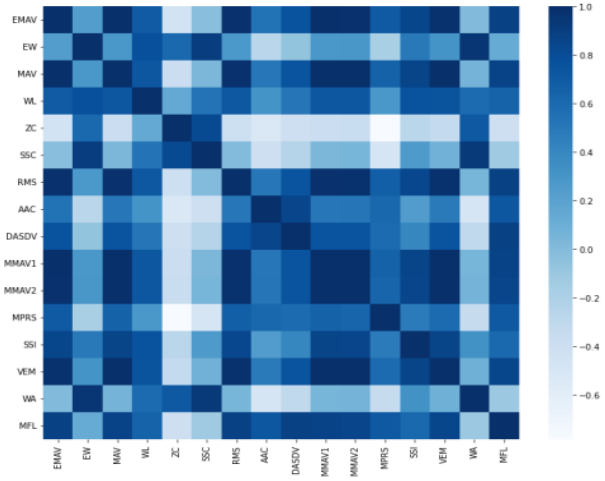

Fig 7: Correlation matrix visualization

By training the models using those reduced features we achieve accuracies as follows.

Table 2: Accuracy comparision of SVM, KNN and Random Forest using reduced features.

Models	Accuracy (%)	Precision	Recall
Random Forest	91.6	0.93	0.91
KNN	66.5	0.66	0.66
SVM	66.6	0.66	0.62

By referring above Table 2, we can conclude that random forest yields the highest accuracy that is 91.6%.

Method-3: Training the model by using Principal Component Analysis (PCA).

Principal Component Analysis is a dimentionality reduction data with the minimum loss of information in the dataset. The cumulative variance of first three principal components as shown in Fig 8.

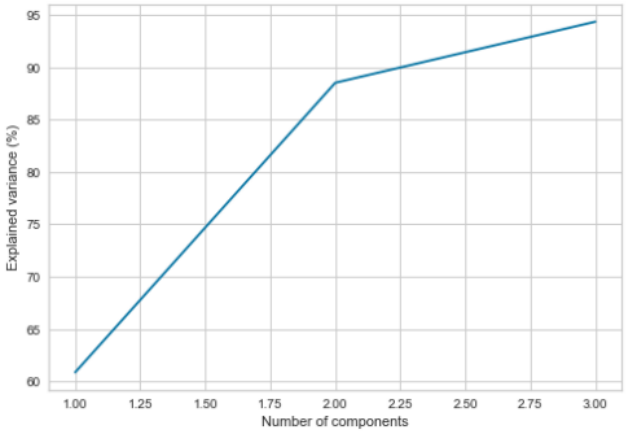

Fig 8: Cumulative variance of first 3 principal components.

By applying these transformed principal components to train the classification models like SVM, KNN and Random Forest. Table 3 gives the Accuracy comparision of SVM, KNN and Random Forest using PCA

Table 3: Accuracy comparision of SVM, KNN and Random Forest using PCA

Models	Accuracy (%)	Precision	Recall
Random Forest	83.3	0.875	0.875
KNN	75	0.741	0.833
SVM	91.6	1.000	0.87

Finally, we can conclude that comparative analysis of classification models like SVM, KNN and Random Forest by using above discussed all three methods in Table 4.

Table 4: Comparative Analysis of SVM, KNN and Random Forest by using all three methods

Models	Using all 16 features	Using correlation	Using PCA
Random Forest	91.6	91.6	83.3
KNN	79.2	66.2	75
SVM	79.1	66.6	91.6

Table 5: Hyperparameters chosen to train the Algorithm.

SVM – Support Vector Machine	C = 1 Kernel = Radial Basis Function Degree = 3 Gamma = Scale Tolerance = 0.001 Cache Size = 200 Class Weight = None Decision Function = OVR Random State = None
K-Nearest Neighbors	Number of Neighbours = 5 Weight = 'Uniform' Algorithm = 'Auto' Leaf Size = '30' Power Parameters = 2 Metric = 'Minkowski'
Random Forest	Number of trees in the forest = 100 Criterion function = Gini Minimum Samples Split = 2 Minimum Samples Leaf = 1 Maximum Features = Square Root Max Leaf Node = None Warm Start = False

Table 6: Comparison between proposed method and previous method

Paper Studies	Approach	Accuracy
[3]	SVM	88%
	RFDT	81%
[4]	LR	62%
	ANN	57%
[5]	SVM	87%
	RF	85%
Proposed	SVM	91.6%
Proposed	RF	91.6%

Table 6 shows performance comparison between proposed method and previous with other two techniques based on accuracy,

Scikit Learn Documentation has the following models accessible.

In situations with large features and little training data, SVM performs better than KNN. Compared to kNN, SVM is simpler to understand and requires less computation. Table 5 gives the hyperparameters chosen to train the Algorithm. The classification accuracy and speed of SVM are both noticeably higher than those of kNN. With regard to Random Forest, it excels at recognising neck pain because it has a specific affinity for time-domain continuous features and can handle a variety of numerical, categorical, and continuous variables, much like our dataset does. As a result, Random Forest may perform better than other curve-based algorithms if there is significant nonlinearity between the independent variables. Consequently, we believe that the RF algorithm learned more effectively than the other algorithms due to the strong non-linearity of the time domain characteristics.

So, what might have caused the other classifiers to yield lesser accuracy than the above three?

In decision trees, a tiny change in the data can have a significant impact on the decision tree's structure, resulting in a result that differs from what people would typically see. When forecasting the outcome of a continuous variable is the primary objective, decision trees perform less well. Because we are using time continuous features decision tree miserably fails in predicting the real nature of the data. Whereas, Naïve Bayes classifier makes the unusual but unrealistic assumption that all predictors (or variables) are independently. For correlation analysis, this fails because our assumption of features being highly positively correlated tend to provide the best results becomes contradictory. In Logistic Regression, time Domain sEMG analysis is a highly non-linear problem and Logistic Regression has a very linear decision surface, so naturally it will fail in the analysis. Also, it makes no assumptions about distributions.

V. Conclusion and Future Scope

In the current times, with most professions demanding constant usage of computer over stretched period of time, significant set of people are facing neck pain or cervicalgia. This prompt for an early detection and treatment with physiotherapy, yoga therapy, etc. The medical community as well as patients may benefit greatly from this use of effective technology assistance. In this study, we have acquired dataset of sEMG from both healthy controls and subjects requiring postural correction. Various features have been extracted. Three stages of experimentation are done namely, classification using all sixteen features, classification using reduced feature set of three features obtained based on correlation coefficient and finally, classification based on features as per Principal Component Analysis. Performance

of machine learning algorithms such as SVM, KNN and Random Forest classification models for prediction of cervicalgia is compared. Highest classification accuracy of 91.6% is achieved. In continuation of this work further, we can train the mentioned models and predict the type of cervicalgia. Limitations of this work is lack of sufficient EMG data. time domain signals are known to be stochastic, in supervised machine learning models, we identified randomly accurate that accuracy scores might have satisfied. To overcome this, we can adapt unsupervised learning models to learn the randomness of the data.

REFERENCES

[1] Do, Youn Lee, Chan Woo Nam, Youn Bum Sung, Kyoung Kim, and Hae Yong Lee. "Changes in rounded shoulder posture and forward head posture according to exercise methods." Journal of physical therapy science 29, no. 10 (2017): 1824-1827.

[2] Nobari, Meisam, Seyed Asadullah Arslan, Mohammad Reza Hadian, and Behnaz Ganji. "Effect of corrective exercises on cervicogenic headache in office workers with forward head posture." Journal of Modern Rehabilitation 11, no. 4 (2017): 201-208.

[3] Suma K V, Lakshmi Shrinivasan, Varun C R, Nagendra B N. Analysis of Muscle Function for Postural Correction Using Surface EMG Signals, Webology (ISSN: 1735-188X) Volume 19, Number 2, February 2022. Pp. 2796 – 2808.

[4] Pan, Lizhi, Dustin L. Crouch, and He Huang. "Comparing EMG-based human-machine interfaces for estimating continuous, coordinated movements." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, no. 10 (2019): pp. 2145-2154

[5] Pancholi, Sidharth, and Amit M. Joshi. "Portable EMG data acquisition module for upper limb prosthesis application." *IEEE Sensors Journal* 18, no. 8 (2018): pp. 3436- 3443.

[6] Adane Gelaw, Tariku, and Misgina Tsighe Hagos. "Posture Prediction for Healthy Sitting using a Smart Chair." arXiv e-prints (2022): arXiv-2201.

[7] Gochoo, Munkhjargal, Tan-Hsu Tan, Shih-Chia Huang, Tsedevdorj Batjargal, Jun- Wei Hsieh, Fady S. Alnajjar, and Yung-Fu Chen. "Novel IoT-based privacy-preserving yoga posture recognition system using low-resolution infrared sensors and deep learning." *IEEE Internet of Things Journal* 6, no. 4 (2019): pp. 7192-7200.

[8] .Varun, C. R., Sumant A. Gunagi, K. A. Venugopal, Shrinath Darur, and K. V. Suma. "Multi Channel Acquisition and Scope of Analysis of Surface EMG signals." In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pp. 1-5. IEEE, 2020.

[9] Zawawi, TNS Tengku, A. R. Abdullah, W. T. Jin, R. Sudirman, and N. M. Saad. "Electromyography signal analysis using time and frequency domain for health screening system task." International Journal of Human and Technology Interaction (IJHaTI) 2, no. 1 (2018): 35-44.

[10] Support Vector Machine-Based EMG Signal Classification Techniques: A Review Diana C. Toledo-Pérez 1 , Juvenal Rodríguez-Reséndiz 2,* , Roberto A. Gómez-Loenzo 2 and J. C. Jauregui-Correa 2.