Feature Extraction using MATS toolbox

Vinuthira Gandhi Chandrasekaran, Ancilia Dmello and Sweta Agrawal

***Abstract*— The EEG and ECG data recorded from a distracted and undistracted driver is analyzed and various measures or features are extracted, e.g. Statistical, Linear, Non-Linear, Frequency, and Modelling. The aim is to build a model for classifying the signal as normal or distracted and to rank the features on how best they classify the data.**

I. Introduction

The main mission of National Highway Traffic Safety Administration (NHTSA) is to ‘prevent accidents, save lives, prevent injuries, and reduce economic costs due to road traffic crashes.’ Distracted driving is a significant and difficult safety problem to consider.

Distracted driving is any activity that could divert a person's attention away from the primary task of driving [Source: ‘http://www.distraction.gov/’ Official Website for US Distraction Driving]. These distractions may be dangerous for driver, passenger, or any bystander. There might be number of distractions while driving like Texting, Using a cell phone or smartphone, Eating and drinking, Talking to passengers, Reading, Using a GPS navigation system, Watching a video, Adjusting a radio/ CD player/ MP3 player etc. But, the most alarming distraction is text messaging and using a cell phone or smartphone as it requires visual, manual, and cognitive attention from the driver.

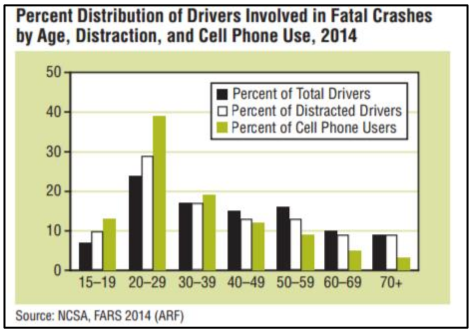
Source: National Center for Statistics and Analysis (NCSA), FARS 2014 Annual Report File (ARF)

	Crashes	**Drivers**	**Fatalities**
Total	**29,989**	**44,583**	**32,675**
Distraction-Affected (D-A)	2,955 (10% of total crashes)	3,000 (7% of total drivers)	3,179 (10% of total fatalities)
Cell phone in Use	385 (13% of D-A crashes)	398 (13% of distracted drivers)	404 (13% of fatalities in D-A crashes)

Table 1. Fatal Crashes, Drivers in Fatal crashes, and Fatalities, 2014

To support above statement, NHTSA released distracted driving statistics in April 2016. Figure 1 provides information on crashes, drivers, and fatalities involved in fatal distraction-affected crashes in 2014. There were a total of 29,989 fatal crashes in the United States involving 44,583 drivers causing fatal crashes where 32,675 people were killed. As shown, there were total of 385 fatal crashes 398 distracted drivers and total 404 fatalities reported to have involved the use of cell phones as distractions (13% of all fatal distraction-affected crashes, drivers and fatalities). This makes cell phone as the major reason for distracted driving on American roadways.

Figure 2 provides the distribution of drivers age for total number drivers involved in fatal crashes, distracted drivers involved in fatal crashes, and distracted drivers on cell phones during fatal crashes.



Source: NCSA, FARS 2014 (ARF)

Fig. 1. Distribution of drivers involved in fatal crashes

Along with National Highway Traffic Safety Administration of the USA, ‘National Safety Council Report’ states that the use of a cell phone impairs a driver’s ability as much as driving drunk. The sensitivity of the issue “Using a Cell Phone while driving” and its consequences are the prime reason of us going for this project.

There are many applications like finance, research and bio-medical analysis which require computing linear, nonlinear and other measures from big datasets. There are many commercial software and freeware which have implemented these measures but none of them allows all the measures to be extracted at once. MATS toolbox provides a very versatile and interactive GUI to select, run and view many measures either collectively, randomly or selectively, at once.

II. Motivation

There are four major motivations which focus on the methods of traffic safety using the result of Feature Extraction using MATS toolbox (FEMT).

A. Legislation and policy-making

Informed legislation and policy-making in traffic rules is generally based on empirically rigorous research data. Results of FEMT will support traffic safety research based on real-world on-road traffic data and are, accordingly, relevant to legislators and other law makers for framing rules of the road. In many countries, legislators and policy-makers can use these results to support the writing of safety rules and regulations. Legislators and other policy-makers in the traffic safety domain need a solid foundation in research to make effective laws and policies. Detail analysis of FEMT can provide ample amount of information about the prevalence of inattention and distraction, pointing to these as major factors contributing to crashes and near-crashes on road. To summarize, FEMT can be used by policy makers and experts to make effective traffic regulations for their respective country or state or county.

B. Driver Education and behavior based safety

Data collected and processed in FEMT can be used to improve the efficiency of driver education. FEMT can provide more insight about the effects of distracted driving and these details can be included in driver education and training program by elaborating clear context to student drivers. Research on FEMT can address the identification of factors that contribute to the occurrence of crashes and other safety critical events and that can be included in driver education and training program to educate new drivers on this important aspect of road safety.

C. Advanced Driver Assistance Systems (ADAS)

FEMT can be used as a key factor for the development of ADAS in modern vehicles. To develop new ADAS features and/or improve current systems, we need to improve our understanding of driver behavior in real traffic. And this real time driver distraction data can be provided by us for ADAS features to act. Generally ADAS features are classified as Preventive ADAS Features which are intended to avoid crashes and Corrective ADAS Features which are designed to mitigate the consequences of the driver distraction/negligence on road. In both the above features, data analysis provided by FEMT will be key factor for designers to design safety features in vehicle under ADAS.

D. Safety ratings for drivers and insurance

FEMT collected for individual driver can be used to decide safety rating of each driver. This data will help decide the driver's behavior, distraction avoidance index, drive stress quotient etc. of individual driver. This data will allow us to make profile for each driver and can be used as driver's feedback which will drivers to self-evaluate themselves. These drivers profile can help insurance provider to decide driver's driving history or diligence on the road.

III. Previous Works

EEG and ECG signals have been widely used to investigate brain and heart activities in healthcare. These data are valuable for the medical field as they are used to differentiate abnormal conditions and prognosticate ailments in a patient. Data analysis of these signals have been around for a while and researchers have experimented with different methods of feature extraction and selection from these signals. The following sections briefly describes different methods of feature extraction and selection adopted by others researchers.

A. Using Wavelet transforms and Neural Network

Pari Jahankhani et al. [5] have analyzed EEG data from patients with epilepsy during seizure and classified them. They used the Discrete Wavelet transform (DWT) to represent the time frequency distribution of the signals by generating 5 sub- bands each having 4 dispersion measures. This method captures transient features, which occur during epileptic periods and localizes them in time and frequency accurately. [5] The labels and classes are provided as it is supervised learning. To reduce the number of features, Rough Sets and PCA (Principal Component Analysis) was used. A Neural Network classifier, Learning Vector Quantization (LVQ 2.1) was employed to classify unknown EEG signals as normal or epileptic. The steps adopted in [5] is mentioned in Figure 2. A maximum accuracy of 100 percent was achieved.

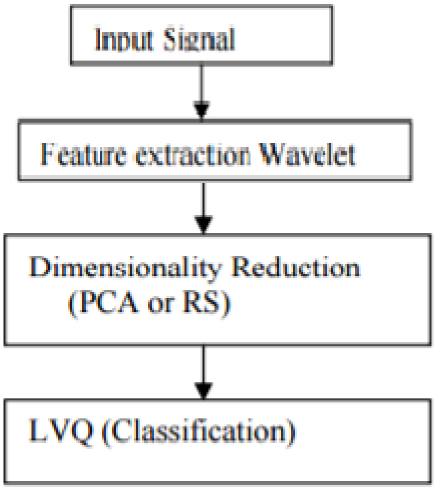

Input Signal
Feature extraction Wavelet
Dimensionality Reduction (PCA or RS)
LVQ (Classification)

Fig. 2. Steps adopted by Pari Jahankhani [5].

Folkers et al. [6] proposed a versatile framework for the classification of these signals by introducing the Neuro – Fuzzy systems – which combines the capabilities of Fuzzy Logic and Neural Networks.

B. Using Relative Power Level

Sangtae Ahn et al [7] researched the effect of sleep deprivation on driving and used simultaneous EEG and ECG recordings to give insights. The EEG data (30 minutes) collected was divided into 180 samples of each 10 seconds of data. The most informative RPL (relative power level) features between the two classes (sleep deprived driving and rested driving) were obtained by a power spectral density using EEGLAB library. A group of training and test data were obtained by repeating the process 30 times with different sliding steps. The classifier model was build using the Fisher Linear Discriminant Analysis. [7] To classify the mental state of the objects, RR-peak interval measures were extracted.

C. Using Dimensionality Reduction/ Feature Selection

The emotional state of a patient with Parkinson's was classified in [8]. The Bispectrum, power spectrum, wavelet packet feature and other non-linear measures of the signal were obtained. Principal Component analysis (PCA), Independent Component Analysis (ICA) and Correlation-based feature selector (CFS) were implemented on these features to increase the performance on the model and remove least significant features.

The feature vector was then randomly divided into 10 sets and trained repeatedly for 10 times. The classification of the emotional state was achieved by using Support Vector Machine classifiers and Fuzzy K- nearest neighbor. SVM can be easily adapted to nonlinearly separable data using kernel functions to map the data to a much Emotional state classification in PD 21 higher dimensional space where the data becomes more separable. FKNN assigns a class based on the predominant class among the k nearest neighbors. [8]

D. Using Eigen Vector Methods

Elif Derya Übeyli's et al [9] paper presents some eigen vector methods for selecting relevant features of ECG and EEG signals. Pisarenko, multiple signal classification (MUSIC) and Minimum Norm are the 3 eigen vector methods for feature extraction. The importance of each method can be summarized as:

- Pisarenko: useful for estimating PSD which contains sharp peaks at expected frequencies.
- MUSIC: uses average spectra of all eigen vectors, hence eliminates effect of spurious zeroes
- Minimum Norm: uses linear combination of all eigen vectors.

IV. Data Description

To analyze the change in drivers' behavior due to distraction, we collected the EEG and ECG signals of a driver, first while driving without any distraction (Normal Data) and then with distraction (Phone Data). To take the drivers' mental status and other physiological factors into consideration, we analyzed the signals from the same subject with and without distraction again (Session 2). Furthermore, to take every individual's differences in the behavior into consideration, we analyzed the signals from three different subjects. The following figure describes the data.

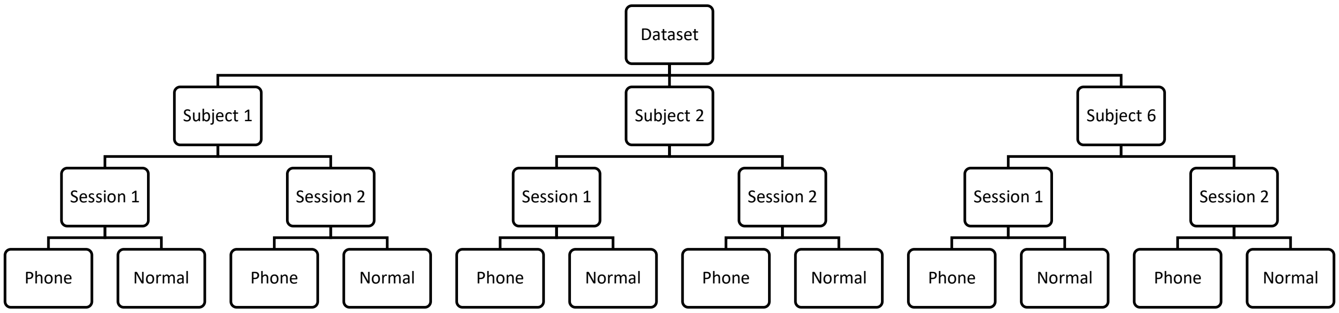


Figure3: Data Used in Project

V. Methodology

The steps involved to extract and run measures on the recorded time series are mentioned below:

1. The time series for normal data and distracted data are loaded onto MATS.
2. Both the time series are first segmented with a window size of 500 with an overlap window size of 250.
3. Measures are then selected by setting relevant changes in the parameters and the selected measures are run.
4. The result of each of the implemented measures are then viewed and saved for analysis.
5. The number of feature per channel is large, so, the dimensions of the feature matrix need to be reduced. For that, the relevance of each feature is obtained by applying different feature selection and ranking methods.
6. The best 100 and 200 features are selected and then only those features are used to form the training and test data set.
7. Using the classification learner App in the MATLAB, the model is trained using the training data (Session 1) and tested on the Test data (Session 2).

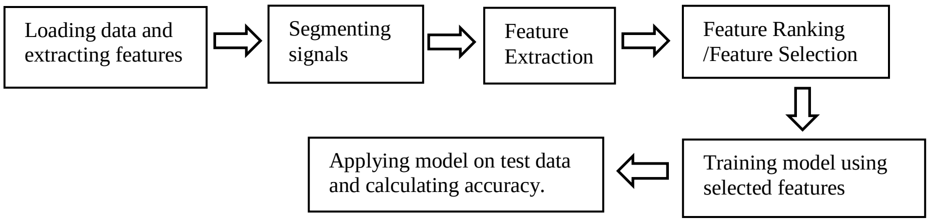

Loading data and extracting features
Segmenting signals
Feature Extraction
Feature Ranking /Feature Selection
Training model using selected features
Applying model on test data and calculating accuracy.

Figure 4: Project Work Flow

A. Step 1: Feature Extraction

The Measures of Analysis of Time Series (MATS) MATLAB toolkit is designed to handle a large set of scalar time series and compute a large variety of measures on them, allowing for the specification of varying measure parameters as well. The purpose of MATS is not only to implement well-known and well tested measures, but also to facilitate some sort of pre-processing.

The MATS Toolbox enables us to do all the following:

1. Pre-Processing of the Time Series (Segmentation and Transformation).
2. Extraction of Measures.
3. Viewing extracted measures in different ways (Tables and Plots).

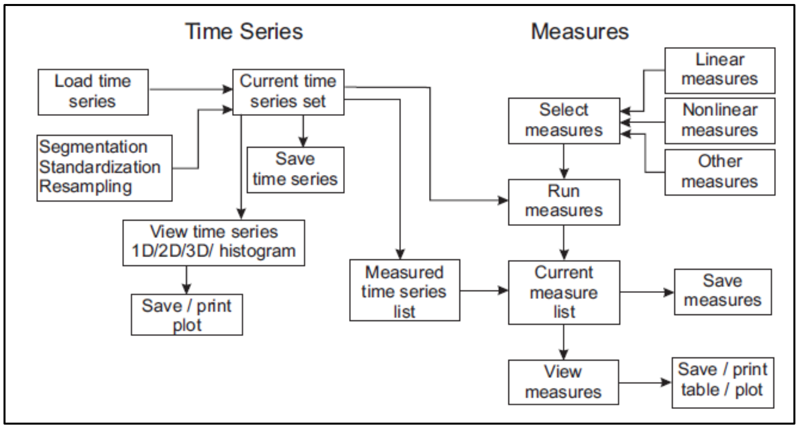


Figure 5: The flow diagram for possible operations in MATS Toolbox

i. Segmentation

As a pre-processing step, the obtained signal was segmented using a window size of 500 and sliding step of 250. Therefore, after segmentation, each segment is a 2-second-long segment of the entire signal. Segmentation of signals enables us to analyze the signals more minutely. The number of segments was different for different subjects and sessions of the same subject.

ii. Measure extraction

Out of the different measures offered by MATS Toolbox, for our project, we mainly extracted the following categories of measures.

1. Linear & Non-Linear Correlation
2. Linear Frequency
3. Linear Modelling
4. Feature Statistics
5. Long Range Correlation

Again, each main category is divided in subgroups. Most of the measures require one or more parameters to be specified and the default values are determined based on the type of time series (e.g., from discrete or continuous systems) and require different parameter settings e.g., the default value of delay for all correlation measures is 10. Measures were extracted from all three subjects for both the sessions. We could extract 109 measures for 2 subjects and 105 measures for the third subject. This task of extracting measures alone required around 80% of the whole effort.

MATS offer different ways to visualize the extracted measures. We used the tabular format for classification and the free plot for analysis purpose. So, for every session, the resulting table of measures looked something like the following figure:

		Channel 1	**Channel 2**	**…… Channel 19**	**Class Label**
		Measure 1 …… Measure 109	Measure 1 …… Measure 109	……..Measure 1…… Measure 109	
Normal	Segment 1				0
	Segment 2				0
	…….				0
	Segment n				0
Phone	Segment 1				1
	Segment 2				1
	………				1
	Segment n				1

Table 2: Table of Measures Format

The detailed summary of the feature extraction process is as follows:

	Subject 1				Subject 2				Subject 6			
	Session 1		Session 2		Session 1		Session 2		Session 1		Session 2	
	Normal	Phone	Normal	Phone	Normal	Phone	Normal	Phone	Normal	Phone	Normal	Phone
Number of Segments	108	134	110	228	109	340	131	478	109	222	113	141
Number of measures Extracted per channel	109	109	109	109	105	105	105	105	109	109	109	109
Dimension of Feature Matrix	242*2071		338*2071		449*1995		609*1995		331*2071		254*2071	

Table 3: Feature Extraction Summary

B. *Step 2: Feature Selection/ Feature Ranking*

Feature Selection is the process of selecting a subset of relevant features for model construction. Feature Selection is required for the following reasons:

1. Simplification of models to make them easier to interpret
2. Shorter training times
3. Enhance generalization by reducing overfitting

Using the Feature Selection Library Toolbox, we have obtained the rankings of each attribute, using the Relief-F and CFS (Correlation based Feature Selection) algorithms, which are filter methods. These methods, calculate the ranking sequentially.

Feature ranking gives the relevance of each attribute without considering the correlation between them. Rankfeatures function from the Bioinformatics toolbox is used to obtain the feature ranks.

The Session 1 feature matrix of each subject is used to generate the feature subset. The first 100 and 200 best features are used to generate the final feature matrix for classification. The subset with the first ten best features are shown below for the 3 methods:

Relief-F			CFS		
Subject 1	Subject 2	Subject 6	Subject 1	Subject 2	Subject 6
1936	1236	1410	778	396	1033
1937	1237	1409	815	1041	1251
1934	1252	1393	1142	1084	1360
1935	1253	1390	1360	1528	1402
1939	1232	1389	1469	1621	1469
1938	1233	1394	1622	1848	1578
1953	602	1192	594	1427	738
1954	603	1176	1046	575	2065
1955	622	1191	832	1701	1071
1918	623	1172	290	1711	1916

Table 4 Ten best features from all subjects using feature selection

rankfeatures		
Subject 1	Subject 2	Subject 6
1936	1972	1393
1918	602	1394
1919	1987	1410
1953	1971	1409
1920	1988	1390
1937	603	1389
1917	622	1192
1934	1967	1172
1921	1652	1176
1954	623	1175

Table 5. Ten best features from all subjects using feature ranking

C. Step 3: Classification

After obtaining the best 100 and 200 feature subsets, we generated the training and test data. For this, we wrote a MATLAB code to generate a matrix with only those selected features. The training data was generated from the Session 1 of each subject. For the test data subset, the subset was generated from Session 2 of the same subject. Thus, we need to construct a model that trains on the Session1 of a subject and predicts the class labels (0-normal and 1- distracted) for the Session 2.

We selected the Complex Tree classifier as the model for classification. We used 5-fold cross validation to train the model. The dimensions of the feature matrix before and after feature selection are listed in the table below:

Subject 1		**Subject 2**		**Subject 3**		
Train	**Test**	**Train**	**Test**	**Train**	**Test**	
242x2072	338x2072	449x1996	609x1996	331x2072	254x2072	→ Before Feature Selection
242x100	338x100	449x100	609x100	331x100	254x100	→ After Feature Selection- 100 features
242x200	338x200	449x200	609x200	331x200	254x200	→ After Feature Selection- 200 features

Table 6. Dimension Reduction after Feature Selection

D. Extended Analysis (Take – Home Assignment)

Using different criterions of the rankfeatures function- “ttest”, “entropy”,” bhattacharyya”,”roc” and Wilcoxon, in the Bioinformatics Toolbox, we generated the best 10 features for three cases:

1. Session 1 of Subject 2 and 6
2. Session 2 of Subject 2 and 6
3. Session 1 and 2 mixed for Subject 2 and 6

The main objective is to analyze the features selected for each case and look for similarities in the selected feature set. The results of this analysis are discussed in the F. section.

E. Results

In table 7 and table 8, the results of accuracy for train and test data for all subjects with different Feature selection and ranking method are shown. In table 9 and table 10, Best feature obtained from all feature selection and ranking methods is listed.

Feature selected	**Baseline**		**ReliefF**				**CFS**			
	All		**100**		**200**		**100**		**200**	
Subject	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
1	99.90	32.54%	98.80%	32.50%	98.80%	32.50%	98.80%	32.50%	98.80%	32.54%
2	100%	21.50%	99.80%	21.50%	100%	21.50%	98.90%	21.50%	100%	21.50%
6	100%	44.40%	100%	44.50%	100%	44.50%	81.30%	57.08%	79.50%	56.69%

Table 7. Results of Accuracy for all subjects with relief-F and CFS method.

Rank Feature				
	100		**200**	
Subject	Train	**Test**	**Train**	**Test**
1	99.20%	32.54%	99.20%	32.54%
2	100%	21.50%	100%	21.50%
6	100%	44.48%	100%	44.48%

Table 8. Results of Accuracy for all subjects with rankfeatures method.

Subject	ReliefF		CFS	
1	LocalMaximSDa1w1	channel 18	PearsCAutot5	channel8
2	LocalMinimMEANa1w1	channel12	LocalMinimMEANa1w1	channel4
6	MedianTimS	channel13	MedianFreql5u480	channel10

Table 9. Best Feature per feature selection method for all subjects.

Subject	Rank Feature	
1	LocalMaximSDa1w1	channel18
2	LocalMinimMEDIANa1w1	channel19
6	LocalMinimMEANa1w1	channe13

Table 10. Best Feature with rankfeatures for all subjects.

The below plot shows the best feature according to relief-F method for subject 2. The red line separates the normal and distracted data.

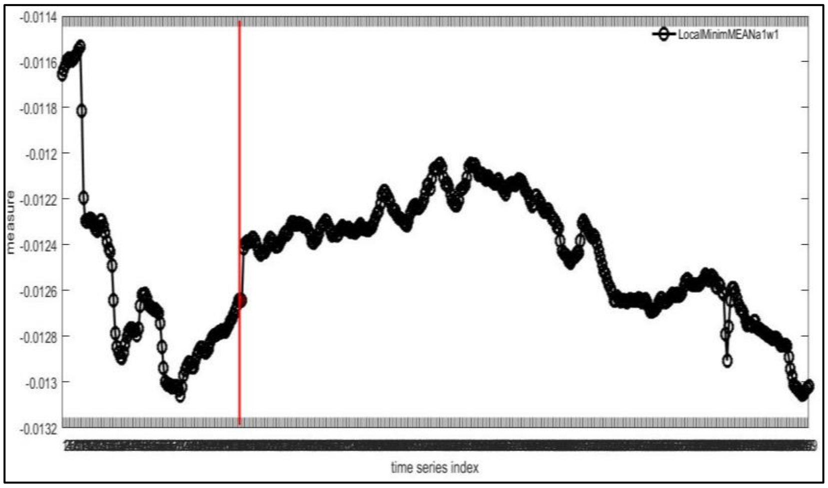


Figure 4. Best feature from relief-F for subject 2

F. *Results – Extended Analysis (Take Home):*

After generating the 10 best features using ttest, entropy, Bhattacharyya, roc and wilcoxon in rankfeatures function, the results are as shown below.

subject 2 session 1 best 10 feature					Subject 6 session 1 best 10 feature				
ttest	**entropy**	**bhattacharyya**	**roc**	**wilcoxon**	**ttest**	**entropy**	**bhattacharyya**	**roc**	**wilcoxon**
1996	191	191	367	602	2072	1943	1943	82	82
1972	**401**	**401**	**602**	**603**	**1393**	**1393**	**1838**	**86**	**86**
602	506	506	603	622	1394	1394	1948	102	102
1987	**716**	**716**	**622**	**623**	**1410**	**1409**	**1839**	**300**	**300**
1971	821	821	623	607	1409	1410	2065	304	304
1988	**1031**	**1031**	**926**	**606**	**1390**	**1390**	**2048**	**320**	**320**

603	1136	1136	1556	1653	1389	1389	2047	517	517
622	**1241**	**1241**	**1996**	**1652**	**1192**	**2065**	**2044**	**518**	**518**
1967	1346	1346	607	1672	1172	1838	2043	521	521
1652	**1556**	**1556**	**606**	**1673**	**1176**	**1737**	**2064**	**522**	**522**

Table 11. 10 Best Feature with rankfeatures for session1

subject 2 session 2 best 10 feature					**subject 6 session 2 best 10 feature**				
ttest	**entropy**	**bhattacharyya**	**roc**	**wilcoxon**	**ttest**	**entropy**	**bhattacharyya**	**roc**	**wilcoxon**
1443	1136	1136	52	1337	304	1616	1616	299	299
1548	**1241**	**1241**	**157**	**1338**	**300**	**1725**	**1725**	**300**	**300**
1442	52	52	182	1341	303	488	488	303	303
1547	**1761**	**1968**	**183**	**1342**	**320**	**924**	**924**	**304**	**304**
1463	1777	1967	186	1357	299	1142	1142	319	319
1568	**1778**	**1987**	**187**	**1358**	**319**	**1360**	**1360**	**320**	**320**
1462	1762	1988	202	1442	82	1469	1469	488	1498
1567	**1757**	**1971**	**203**	**1443**	**86**	**1578**	**1578**	**924**	**1499**
1447	1758	1972	262	1446	102	1687	1687	1142	1502
1552	**99**	**1761**	**287**	**1447**	**81**	**1520**	**1520**	**1360**	**1503**

Table 12. 10 Best Feature with rankfeatures for session2

subject 2 mixed best 10 feature					**subject 6 mixed best 10 feature**				
ttest	entropy	bhattacharya	roc	wilcoxon	ttest	entropy	bhattacharya	roc	wilcoxon
1783	1136	1136	367	183	480	1360	1360	1360	1944
406	1241	1241	262	182	371	1469	1469	1469	1948
886	367	367	577	203	370	1578	1578	1578	1835
1764	99	99	997	202	698	1520	1520	1251	1839
402	1101	1101	926	187	807	1904	1904	1142	371
1759	204	204	1417	186	372	1956	1956	1033	2048
1747	1669	1967	892	98	745	1943	1943	1289	2047
769	1459	1968	1312	77	589	1193	2047	706	480
768	1354	1987	1241	97	479	1623	2048	815	2044
770	1144	1988	1207	78	527	1514	2043	1687	2043

Table 13. 10 Best Feature with rankfeatures after mixing session 1 and session 2

VI. Conclusion

The analysis of EEG and ECG signals has been comprehensively used for diagnosing driver distraction. We could extract measures for all 3 subjects which includes Linear, Non-Linear, Correlation based, Frequency based, Modelling based, and some newly introduced statistical features. For feature selection, we executed 2 methods namely Relieff and CFS and for feature ranking we used Rank feature to get the best features for each subject. From table 9 and 10, it can be inferred that the feature “LocalMinimMEANa1w1” is a highly relevant feature, which classifies the normal and phone data very well. The accuracy obtained after feature selection and ranking is same as that of the baseline accuracy. Therefore, we can conclude that the features selected using the feature selection and ranking methods for session 1 didn’t perform well on session 2. Only the features selected using CFS for Subject 6 gave a better accuracy compared to the baseline.

After comparing different criterions that rank features offers, we observed that the 10 best features obtained from “entropy” and “bhattacharya” are very similar, and in some cases, they are even 100% same. Also, since the best features from session 1, session 2, and mixed sessions, are entirely different from each other, the accuracy is very poor.

REFERENCES

[1] Official Website for US Distraction Driving http://www.distraction.gov/stats-research-laws/facts-and-statistics.html

[2] National Center for Statistics and Analysis. (2016, April). Distracted driving 2014 (Traffic Safety Facts Research Note. Report No. DOT HS 812 260).

[3] Table 1 and Figure 1: National Center for Statistics and Analysis. (2016, April). Distracted driving 2014 (Traffic Safety Facts Research Note. Report No. DOT HS 812 260).

[4] Thesis on the Analysis of Naturalistic Driving Data by Jonas Bargman

[5] Data Mining an EEG Dataset With an Emphasis on Dimensionality Reduction - JeeEun Lee, HyeJin Kim, Byuck jin LEE, Chungki Lee, Sun K. Yoo

[6] Folkers, A., Mosch, F., Malina, T., & Hofmann, U. G. Realtime bioelectrical data acquisition and processing from 128 channels utilizing the wavelet-transformation. Neurocomputing, 52–54, 247–254, 2003

[7] Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and fNIRS Data

[8] Optimal set of EEG features for emotional state classification and trajectory visualization in Parkinson’s disease - R. Yuvaraja,*, M. Murugappana , Norlinah Mohamed Ibrahimb , Kenneth Sundaraja , Mohd Iqbal Omara , Khairiyah Mohamadb , R. Palaniappanc

[9] Eigenvector Methods for Analysis of Human PPG, ECG and EEG Signals - Elif Derya Übeyli, Dean Cvetkovic, Irena Cosic

[10] Journal of Statistical software, February 2010, volume 33, Issue 5, by Dimitris Kugiumtzis, Alkiviadis Tsimpiris

[11] Infinite Feature Selection, 2015 by G Roffo, S Melzi, M Cristani