Intoxication Detection using Audio

Aishwarya B S, Shilpitha R Shetty, Shreya Srinivas, Vedant Mantri, V R Badri Prasad
Computer Science and Engineering, PES University, Banglore, India
aishwaryabs6@gmail.com, shilpithashetty@gmail.com, shreyasri01@gmail.com, vedant02mantri01@gmail.com

***Abstract*—Intoxication detection using speech involves analyzing an individual's speech patterns to determine if they are intoxicated. Such trained models can reduce the usage of unhygienic devices like breathalyzers. This is generally done by examining numerous aspects of utterance and acoustics of speech signals used for training MFCC and CNN models. The primary objective is to develop a non- invasive, accurate approach to intoxication detection that can be deployed for law enforcement and occupational intoxication testing.**

***Index Terms*—CNN, MFCC, speech analysis, feature extraction, model training, intoxication**

I. Introduction

Alcohol consumption has resulted in a variety of undesirable consequences. These can range from short- term effects, such as impaired judgment and motor coordination, to long-term effects, such as liver damage, heart disease, and increased risk of certain types of cancer. Alcohol abuse can also lead to a range of social and psychological problems, including relationship problems, financial difficulties, and an increased risk of depression and anxiety. In addition, excessive alcohol consumption can contribute to a range of public health problems, including increased risk of traffic accidents and increased healthcare costs associated with the treatment of alcohol-related conditions. To address alcohol-related problems, a range of strategies can be employed, including education, regulation, and treatment. The goal of these strategies is to reduce the negative impact of alcohol on individuals, families, and communities, and to promote healthy and responsible alcohol use.

Alcohol-impaired driving is a significant public health problem and is responsible for a large number of deaths and injuries each year. In many countries, laws have been established to prevent individuals from driving while under the influence of alcohol, and there are strict penalties for those who are caught driving under the influence. Despite these efforts, alcohol-impaired driving remains a major problem and is a significant contributor to road traffic accidents and fatalities.

Voice analysis has been explored as a method for intoxication detection due to the potential for speech patterns to reflect changes in the individual's state of mind, including the effects of drugs and alcohol. By analyzing various aspects of the speech signal, such as speech rate, volume, and prosody, researchers have sought to determine if there are reliable markers that can be used to determine if someone is under the influence of drugs or alcohol.

Breathalyzers, which are commonly used for testing blood alcohol content (BAC), can be unhygienic and potentially harmful if not used and maintained properly.

In terms of unhygienic effects, breathalyzers can harbor bacteria and other microorganisms if not cleaned regularly. This can pose a health risk to users, especially if they have compromised immune systems or open wounds in their mouths. Additionally, if multiple people use the same Breathalyzer without proper sanitation, the risk of spreading infections or illnesses can increase,in terms of harmful effects, breathalyzers that use fuel cell technology can produce ozone, which is a respiratory irritant.

To minimize the potential unhygienic and harmful effects of breathalyzers, it is important to follow the manufacturer's instructions for cleaning and maintenance and to replace the device if it becomes damaged or worn. Additionally, individuals should be aware of the potential health risks associated with Breathalyzer use should seek medical attention if they experience any adverse symptoms.

II. BACKGROUND WORK

Though there aren't many projects available, a few approaches to detecting drunkenness have been examined in a few research studies. Intoxication detection is an important issue in various domains, including road safety, health care, and public security. Currently, traditional methods of detecting intoxication, such as Breathalyzer tests, blood tests, and field sobriety tests, are invasive and time-consuming. To overcome these limitations, non-invasive methods for detecting intoxication are in demand.

Speech analysis is a promising non-invasive method for intoxication detection. Alcohol consumption can affect speech production and result in changes in speech patterns, such as speech rate, tone, and prosody. These changes in speech patterns can be analyzed to detect alcohol intoxication.

There have been numerous research studies in this field, investigating the use of different speech analysis techniques and machine-learning algorithms for intoxication detection. Some studies have focused on acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) coefficients, while others have analyzed prosodic features, such as speech rate, energy, and pitch. There have also been studies that have used deep learning techniques, such as Convolutional Neural Networks (CNN) and Long

Short-Term Memory (LSTM) networks, for the detection of alcohol intoxication from speech.

Compared to traditional methods, speech analysis-based intoxication detection methods have the advantage of being non-invasive, fast, and less prone to errors. However, there are also some challenges in this field, such as the variability of speech patterns between Individuals and the need for large amounts of annotated speech data for training machine learning algorithms.

In one study, the authors present a comparative analysis of different machine-learning algorithms for alcohol intoxication detection using speech analysis. The authors found that Support Vector Machine (SVM) with a radial basis function (RBF) kernel showed the best performance. There are other strategies such as a speech-based method for detecting alcohol intoxication by analyzing acoustic and prosodic features such as speech rate, energy, and pitch. The authors also evaluated different machine learning algorithms and found that decision trees and random forests performed well for this task. In another study, the authors propose a speech-based method for detecting alcohol intoxication using recurrent neural networks (RNNs) to analyze the speech signal. They found that the proposed RNN-based method outperforms traditional feature-based methods of detecting alcohol intoxication.

Despite these challenges, the research in this field has shown promising results and has the potential to make a significant impact on the detection of alcohol intoxication.

III. III. DATA AND PRE-PROCESSING

A. *Dataset collection*

There is no existing free available data set that contains audio clips of drunk people. And there is only one data set which is in German and must be paid for. Only data sets of people with different emotions are found. So, we are creating our data set by taking inputs from real people and audio clips from television shows, movies, and videos where drunken behavior can be seen. We will be taking both sober and drunken voice clips to perform pre-processing and extract speech features. This will help in training the model better so that the estimation metric we use gives a good result.

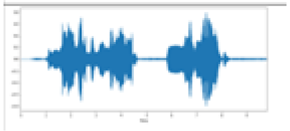

Fig. 1. Drunk audio waveform

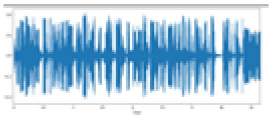

Fig. 2. Sober audio waveform

TABLE I
DATASET CLASSIFICATION

Class	Original	TV/Series	Augmented	Total
Drunk	50	10	540	600
Sober	600	NA	NA	600

B. *Data Augmentation*

All the data collected from multiple sources have been converted to .wav files. Since the amount of drunk audio files was not enough, different augmentation techniques were performed to generate more data.

1) Noise Injection and whitespace injection: Noise injection in speech refers to the intentional introduction of background noise or other types of interference into speech signals. This can be done to assess the robustness of our model or to simulate real-world conditions where speech signals are often degraded by noise

2) Time shifting: Time shifting in audio signals refers to the process of changing the timing of an audio signal relative to some reference point, such as shifting all audio samples ahead or behind in time. This can be useful for synchronizing audio signals from different sources, correcting for time delays in recording or transmission systems, or for creating special effects.

3) Time stretching: Time stretching in audio signals refers to the process of changing the duration of an audio signal without affecting its pitch. This allows for the manipulation of the speed at which the audio is played without changing its perceived pitch

4) Pitch: Pitch in audio augmentation refers to the process of changing the perceived pitch of an audio signal while preserving its duration. This can be used to create variations of a recording, to transpose the pitch of a recording to a different key, or to manipulate the perceived pitch.

5) Changing Speed: Changing the speed of an audio signal in the audio augmentation refers to the process of altering the playback rate of an audio signal, resulting in changes to its duration and pitch. Increasing the speed of the audio will result in a higher pitch and a shorter duration while decreasing the speed will result in a lower pitch and a longer duration.

C. *Preprocessing*

All the audio files, which have been converted to one kind of file i.e., .wav files, are being visualized to see the spectrographs and waveforms or time series plots of the audio data with the help of Librosa, matplotlib, and Scipy modules.

Inside this class, we will have a get _item () function where we do all the pre-processing steps and it will return the signal and the label. We will also define different pre- processing functions like

- cut_if_necessary () Cutting the audio length refers to the process of reducing the duration of an audio signal by removing some portion of its beginning or end. This is often done to keep the signal length uniform.

- right_pad_if_necessary () in this function we will apply padding to the audio signal.
- re-sample_if_necessary () here we will re-sample the signal to the desired sampling rate.
- mix_down_if_necessary () here we will convert the audio signals to keep it uniform to a mono channel.

1) Spectrogram: A spectrogram is a graphic representation of a signal's frequency spectrum as it varies over time. The intensity is typically shown as a heat map with varying color gradients. The time of the clip is indicated by the horizontal axis, while the intensity of the audio wave is indicated by the color variation on the vertical axis The short-term Fourier transform is created using a stft() function. STFT[r] converts signals so that we can determine the frequency's amplitude at a given time, so STFT[r] can be used to calculate the amplitude of different frequencies playing at a specific time in an audio signal and display the spectrogram

2) Spectral Centroid: The spectral centroid indicates the frequency at which a spectrum's energy is centered for a sound. This is like a weighted mean:

$$\mathrm{f}_c = \frac{\sum_k S(k) f(k)}{\sum_k S(k)}$$

where S(k) is the spectral magnitude at frequency bin k, f(k) is the frequency at bin k. The spectral centroid for each frame in a signal is calculated using librosa which will give you an array whose columns correspond to the number of frames in your sample.The spectral centroid is a measure of the amplitude at the centre of the spectrum of the signal distribution over a window calculated from the Fourier transform frequency and amplitude information.

3) Spectral Rolloff: It is a measurement of the signal's shape. It shows how frequently high frequencies drop to zero. To find it, we must compute the percentage of power spectrum bins where 85%
of the power is at lower frequencies. The roll-off frequency for each frame in a signal is calculated using librosa.

4) Spectral Bandwidth: The difference between the higher and lower frequencies in a continuous band of frequencies is known as the bandwidth. The spectral bandwidth is represented by the two vertical red lines SB, on the wavelength axis, which is defined as the size of the light band at 50%
of its maximum. Librosa helps compute the order-p spectral bandwidth:

$$(\sum_k S(k)(f(k) - f_c)^p)^1/p$$

where S(k) is the spectral magnitude at frequency bin k, f(k) is the frequency at bin kk, and fc is the spectral centroid.

5) Zero Crossing Rate: Calculating the number of zero crossings within a segment of a signal is a very straightforward method for gauging its smoothness. The rate at which signal shifts from positive to zero to negative or from negative to zero to positive is known as the zero-crossing rate (ZCR). As a crucial feature to classify percussive sounds, its value has been widely applied in both speech recognition and music information retrieval. Rock, metal, emo, and punk music are examples of highly percussive genres that frequently have higher zero-crossing rate values.

$$zcr = \frac{1}{1-T} \nVdash \sum_{t=1}^{T-1} \{s_t s_{t-1} < 0\}$$

where,
s_t is signal of length t
1{X} is the indicator function(=1 if X is True,else 0)

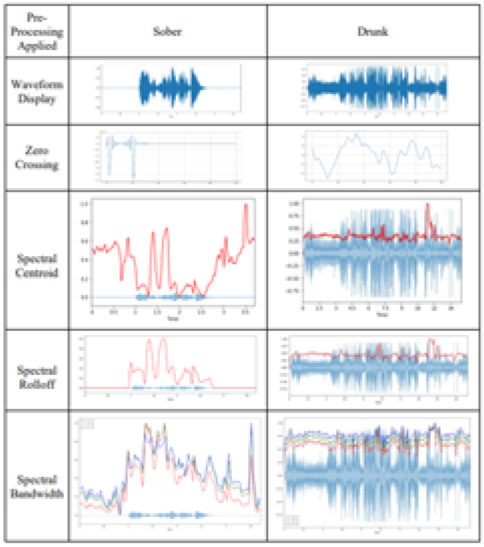

Fig. 3. Preprocessing - Drunk Vs Sober.

IV. IMPLEMENTATION

A. MFCC

Mel-Frequency Cepstral Coefficients (MFCCs) are a class of features used in audio processing tasks like speech recognition and speaker identification. They are intended to capture important audio signals characteristics such as pitch and rhythm while remaining relatively insensitive to background noise and other variations in the audio signal.

1) A/D Conversion: In this step, we will convert our audio signal from analog to digital format Converting the audio signal that we have taken should be converted to digital format from the analog format

2) Pre-emphasis: Pre-emphasis amplifies the energy at a higher frequency by a significant amount. The energy at a higher frequency is significantly smaller than the energy at a lower frequency when we examine the frequency domain of the audio signal for certain vocal elements like vowels. The performance of the model tends to increase when we increase

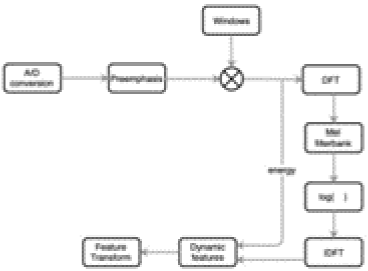

Fig. 4. MFCC Roadmap

the energy available at a high frequency. This can be done by enhancing accuracy.

3) Windowing: We'll take 39 features out of each chunk. Additionally, if we chop the signal off at its edges directly during signal splitting, the sudden decrease in amplitude at the margins ultimately results in noise in the more frequency range.

4) DFT: Applying the DFT transform will move the signal from the time domain to the frequency domain. Analyzing in the frequency domain rather than the time domain is simpler for audio signals.

5) MEL filter bank: The method by which we hear and the method by which the machines hear is completely different. Low frequencies are more sensitive to humans than higher frequencies. Human ears are more sensitive. Hence the addition of human hearing and its characteristic in the process involving feature extraction will improve the machine's performance.
The MEL formula is:

$$mel(f) = 1127ln(1 + f/700)$$

6) Log Application: Compared to lower energy levels, humans are less sensitive to changes in auditory signal energy at higher energy levels. One of the characteristic which can be compared regarding the log function is how the gradient increases when the input value given decreases. So, for getting close to the human hearing system, we have to take the logarithmic function on the output, that is MEL filter that we got in the previous step

7) IDFT: The logarithmic output we got in the last step is transformed back in this step. We do this to understand the anatomical part of how speech is formed in humans. The glottis is what makes the sound in humans. The sound is created by the air in the glottis vibrating. The vibrations will take the form of harmonics, with the fundamental frequency being the lowest of all the produced frequencies. All other frequencies are multiples of the fundamental frequency and reveal information about pitch, while frequencies to the right will reveal details about phones. Since the fundamental frequency doesn't reveal anything about phones, we shall disregard it.

After this procedure is completed, MFCC makes use of the starting 12 coefficients along with the other 12 and this helps in classification. So MFCC technique will produce 39 features from one audio signal sample. These are used as input for the model. There are 39 features of MFCC:

- 12 MFCC features
- 12 Delta MFCC features
- 12 Delta Delta MFCC features
- 1 (log) frame energy
- 1 Delta (log) frame energy
- 1 Delta Delta (log) frame energy

TABLE II
MFCC ENERGY VALUES

Class	Drunk	Sober
MFCC Feature Value	151.03284	-94.681946
Frame energy	-0.8791891	-0.6417764
Delta(log) frame energy	0.6423699	1.7095648
Delta delta log frame energy	3.2135344	0.32022324

B. CNN

Convolutional Neural Networks (CNNs) are deep learning neural networks that are designed to process data with a grid-like topology, such as images. They have an input layer, several hidden layers, and an output layer. Convolutional layers, pooling layers, and normalization layers are examples of hidden layers. Convolutional layers oversee detecting features such as edges and shapes in the input data. The convolutional layer applies a set of filters to the input, each of which generates a feature map by convolution with the input. Pooling layers are employed to reduce the spatial dimensionality of feature maps while retaining critical information. They usually use maximum or average pooling. Normalization layers are used to improve CNN's stability and performance. They typically employ batch or layer normalization. Finally, based on the feature maps generated by the hidden layers, the output layer generates the final prediction.

Convolutional Neural Networks (CNNs) are used to analyze audio data like speech or music signals. The idea is to treat the audio signal as a one-dimensional input, similar to how an image is treated in traditional CNNs as a two-dimensional input.

One method is to use the Spectrogram technique, which converts the audio signal into a spectrogram, which is a 2D representation of the audio signal with time on the x-axis and frequency on the y-axis. The CNN can then be applied to this spectrogram, employing convolutional, pooling, and normalization layers in the same way that they are applied to image data.

Another method is to process the audio data using a 1D convolutional layer, which can detect patterns along the temporal dimension of the audio signal. CNNs can be used in audio processing tasks such as speech recognition, speaker identification, and music classification.

It is worth noting that CNNs are not the only neural network architectures used to process audio data; due to the sequential nature of audio data, Recurrent Neural Networks (RNNs) and their variants (GRU, LSTM) are also popular in audio processing tasks.

1) Adam's optimiser: One simple way to think about Adam is in terms of coefficient of variation (CV or simply uncertainty), which is often used instead of SD (Standard Deviation) to compare the spread of data sets with different units of measurement or with the same units but varying substantially in size. Then we can say that we have a data set of gradients history for each weight in our model that we want to compare. We define the rule that the greater the spread in the gradient's history dataset (higher CV/uncertainty), the smaller the individual learning rate for each weight.

2) Cross Entropy: Entropy of a discrete probability distribution is given by

$$p(x)logp(x)$$

It is defined as "the number of bits required to efficiently encode a message assuming the message's characters were generated at random using the provided distribution." And that is how many characters would be required to encode the same message if, instead of the best encoding for p, the optimal encoding for q was used to encode the characters being created from q. Now, in machine learning, if you have a "neural network" model with an expected output of 0 or 1, While its real output is a binomial distribution (probability value of getting output 1 or 0), cross-entropy is frequently employed as the loss function for this network.

V. Results and conclusion

Developers have various challenges when it comes to audio signal processing. Using libraries like Librosa, on the other hand makes it much easier to grasp. To acquire the best classification for each ML model, we tested different audio classifiers and employed Spectrograms, Important features were found using the Mel frequency co-efficient and tested using various ML models.

A. MFCC

For modeling and classification, we have MFCC, which is widely utilized in speech and audio recognition. An MFC is extracted from the cepstral representation from an audio clip and they comprise coefficients of Mel-frequency cepstral coefficients (MFCCs). The MCs are equally spaced and available on the Mel scale which mimics and is similar to the human Audio system unlike the linearly-spaced bands of frequency used in any other normal spectrum, which can potentially lower transmission bandwidth and storage requirements of audio. We have repeatedly done our training and testing on the MFCC models to train the model to note the accuracies

TABLE III
MFCC Statistics table

Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
96.6	97.2	94.8	96.5	97.5	94.3
97.08	97.0	95.6	96.4	97.5	94.6
96.02	97.4	94.6	97.3	97.6	95.6
97.05	97.3	96.7	96.6	97.6	94.6
96.4	97.3	94.6	96.4	97.5	96.7

B. CNN

CNN models can be utilized for audio classification as strong baseline networks. Transfer learning assumptions remain strong. We investigated what rate of pertained weight is effective for the spectrograms to leam. We demonstrated it on a model, pertained weight outperformed randomly initialized weights in value of quality results of what is learned by CNNs from spectrograms through visualization of the gradients. Furthermore, we know that when pertained model weights are used for instantiation, there are variations seen during the performance. This change in the performance of the model is attributable to the randomization of the classification of linear laver initialization and randomizes the mini-batch order in numerous instances. This adds to variance in the model and builds it with much more accuracy.

TABLE IV
CNN Statistics table

Accuracy	Precision	Recall	F1 Score
96.1	94.8	97.4	96.8
95.8	94.6	96.8	96.4

Our work explores how ML models like CNN and MFCC can be used in the Audio classification pipeline for the intoxication detection task. Our work establishes a solid The baseline for further experimentation is intoxication detection. Extracting features from the audio samples and comparing these different features in order to classify the audio as drunk or sober is extensively scrutinized and analyzed thoroughly, yielding an accuracy of 96.6%
by the MFCC model and 96%
by the CNN model. The high accuracy is a result of the lack of dataset available, which we had to collect from scratch and then performed augmentations on these samples.

VI. Future work

In the future, the project can be made more precise by increasing the intoxicated audio samples in the data set. This can be included as software or a mobile application, making it easily accessible by traffic cops, event entry checking, and various other circumstances where there is a restriction on the consumption of alcohol. Other models like transformers or hybrid-RNN can be used to compare the accuracies. Overall, this project can be advanced to a full-fledged working intoxication detection system that can be integrated with hardware devices or as a software application on mobile phones.

VII. Acknowledgment

The authors would like to extend their sincere gratitude to the following individuals and institutions for their invaluable support and contribution toward the successful completion of this research.

The authors would like to express their deepest appreciation to their supervisor, Mr. VR Badri Prasad, for his guidance, support, and encouragement throughout the research process. His expertise and insights have been invaluable in shaping the direction and outcome of this work.

The authors are grateful to the Chairperson of the Department of Computer Science and Engineering, Mrs. Shylaja, for her support and encouragement in pursuing this research.

Finally, the authors would like to acknowledge the support and facilities provided by PES University, where this research was conducted.

References

[1] Muralidhar, H. K., Muralidhar, K. H., Krishna, M. N. (2013). Alcohol Intoxication Detection from Speech Signal. In 2013 Third International Conference on Advances in Computing and Communication Engineering (pp. 51-55). IEEE.

[2] Garg, V., Prasad, R. (2019). Alcohol Intoxication Detection from Speech using Prosodic Features. arXiv preprint arXiv:1907.06683.

[3] Rana, S., Garg, V., Prasad, R. (2018). Alcohol Intoxication Detection from Speech using Machine Learning Algorithms. arXiv preprint arXiv:1809.064

[4] Parsaei, E., Wibowo, A. (2018, June). Alcohol Intoxication Detection Using Speech Analysis and Machine Learning Algorithms. In 2018 International Conference on Computer and Information Sciences (ICCOINS) (pp. 615- 620). IEEE.

[5] Yilmaz, A., Ekenel, H. K., Erzin, E. (2018, May). Speech Intoxication Detection using Acoustic and Prosodic Features. In 2018 25th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[6] Hu, J., Fan, Y., Lu, J. (2020, May). Speech-based Intoxication Detection using Recurrent Neural Networks. In 2020 27th International Conference on Computer and Information Sciences (ICCOINS) (pp. 50-54). IEEE.

[7] Garg, V., Rana, S., Prasad, R. (2019). Alcohol Intoxication Detection: A Review. In Handbook of Research on Speech Processing and Voice Assisted System Design (pp. 1-36). IGI Global.

[8] Kapoor, A., Tiwari, P., Shukla, V. (2016). Alcohol Intoxication Detection from Speech Signal Using Spectral and Prosodic Features. In 2016 International Conference on Emerging Trends in Engineering and Technology (ICETET) (pp. 1-6). IEEE.

[9] Kaur, J., Kaur, M. (2020). Alcohol Intoxication Detection using LSTM-based Deep Neural Network. In 2020 IEEE 7th International Conference on Control, Decision and Information Technologies (CoDIT) (pp. 437-442). IEEE.

[10] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, ”An overview of noise-robust automatic speech recognition”, IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 22, no. 4, pp. 745-777, Apr 2014.

[11] B. Carnahan, V. E. Byrne, J. Legrand, B. Davis, R. E. Grace, J. J. Staszewski, et al., ”An Alcohol Detection system for heavy vehicles”, Proc. Digital Avionics Systems Conference, vol. 2, pp. I36/1-I36/8, 1998.

[12] D. Garrett, D.A. Peterson, C.W. Anderson and M.H. Thaut, ”Comparison of linear nonlinear and feature selection methods for EEG signal classification”, IEEE Trans. Neural Syst, vol. 11, pp. 141-144, 2003.

[13] U. Svensson, ”Blink behavior based drowsiness detection – method development and validation”, Master’s thesis Linkoping University, 2004.

[14] M. Hahn and Y. K. Kim, ”System to detect driver fatigue based on the Driver’s Response Pattern and the Front View Environment of an Automobile”, Second International Symposium on Universal Communication Osaka, pp. 237-240, 2008.

[15] P. Xie, Y. Wu, X. wei ji and Y. Xia, ”The design of an Automotive Anti-Drunk Driving System”, International Conference on Information Engineering and Computer Science, Dec 19-20, 2009.

[16] B.G. Lee and W.Y. Chung, ”Driver alertness monitoring using fusion of facial features and bio-signals”, IEEE Sens, pp. 2416-2422, 12.

[17] A. Furdea, C.A. Ruf, S. Halder, D.D. Massari, M. Bogdan, W. Rosenstiel, et al., ”A new (semantic) reflexive brain-computer interface: In search for a suitable classifier”, Journal of Neuroscience. Methods, vol. 203, pp. 233-240, 2012.

[18] G. Li and W.Y. Chung, ”Detection of Driver Drowsiness Using Wavelet Analysis of Heart Rate Variability and a Support Vector Machine Classifier”, Sensors, vol. 13, pp. 16494-16511, 2013.

[19] Z. Zhao, L. Yang, D. Chen and Y. Luo, ”A Human ECG Identification System Based on Ensemble Empirical Mode Decomposition”, Sensors, vol. 13, pp. 6832-6864, 2013.

[20] X. Zhao, X. Zhang and J. Rong, ”Study of the effects of alcohol on drivers and driving performance on straight road”, Mathematical Problems in Engineering, vol. 2014, pp. 1-9, 2014.