

Electrocardiogram (ECG) data mining with Arduino and MATLAB

Author: Daniel (Jihong) Park

Affiliation: Korea International School

E-mail: djpark22@student.kis.or.kr

ABSTRACT

The development of multiple platform medical services has increased rapidly recently, thanks to advances in technology. However, these services have not yet to be fully commercialized due to its high price. In this study, we chose to examine and analyze the bio signals from human bodies. More specifically, we analyzed the waveform from the electrocardiogram (ECG) measurement and determined the normal range to predict and diagnose an individual. We used two devices that are essential to convert this signal into a digital form that can be analyzed: Arduino and MATLAB. We analyzed the data from the ECGs with Arduino, MATLAB, and linear regression.

Keywords: electrocardiogram, Arduino, MATLAB

1. INTRODUCTION

In recent years, medical services have developed rapidly. Due to the corona crisis, interest in medical services has increased, and medical device patent applications are showing an explosive upsurge. While patent applications increased at an average annual rate of 2.2% over the past ten years (2011~2020), the number of patent applications in the medical device field increased sharply at an average yearly rate of 8.0%. In particular, in the case of last year, the number of medical device patent applications increased by 15.8% compared

to the previous year due to the impact of COVID-19. Korea has 29.1 pieces of magnetic resonance imaging (MRI) equipment and 38.2 computed tomography (CT scanner) per 1 million people; expensive diagnostic equipment is installed at about 1.5 times the OECD average. [4]

Due to the rapid development of medical services, the ageing phenomenon is occurring more frequently than ever. The rapid ageing of the population means medical services and frequency of service use will increase in the future. As of 2008, about 68% of under 65s used at least one medical service, while 90% of those aged 65 and over used at least one medical service; service use is relatively high in the elderly population (Statistics Office, 2008b). [6]

After examining the distribution by matching the proportion of the elderly population aged 65 and over in the OECD member countries with the number of outpatient visits per person, Korea has a relatively large number of cases compared to the proportion of the elderly population. [5]

However, medical services are expensive. Frequent hospital visits not only result in huge bills but also high overall health insurance costs. Considering that the population ageing is progressing rapidly, it is necessary to devise a policy plan to reduce the number of outpatient treatments by ensuring that the prevention and management of chronic diseases

are conducted through primary medical institutions. It is also possible to commercialize medical services to make a diagnosis at home before visiting a hospital.

However, the technology has not yet developed enough for medical services to be fully commercialized. Nevertheless, some important medical services can be done at home. Researchers are continuously investigating and developing medical services that can be conveniently used in daily life. For example, they are developing smart clothing that can measure various bio signals, including electrocardiogram, temperature, and movement, by attaching sensors to clothes. [1]

Among the numerous important medical services, we selected a service that analyzes bio signals.

The advantages of digitally analyzing bio signals are that the initial cost is low and time-efficient. Medical services that minimize the intervention of experts and receive appropriate measures for the patient's disease symptoms are pursued worldwide. Using bio signals makes it possible to obtain more medical information and accurately determine the patient's health status and whether a specific disease has occurred or not by analyzing the pattern of signals from the human body. [2] Currently, the solution that collects and analyzes bio signals on its own seems to be the one that people prefer the most due to its usefulness. In the future, it will be effective to conduct categorization studies with these digital data. Therefore, the general public can take care of their own health with ease and familiarity through these studies.

Hence, our goal is to analyze waveforms measured from Arduino ECG sensors to interpret human body signals. Another purpose of this paper is to establish the normal range for the

electrocardiogram and analyze the electrocardiogram of an individual to diagnose normality and abnormality (arrhythmia).

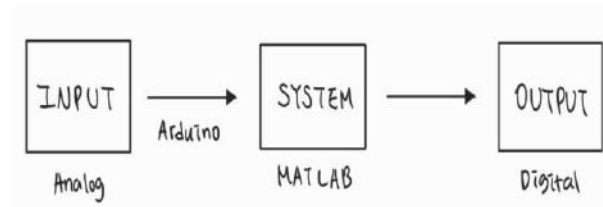


Figure 1. Block Diagram of Digitization of Analog Vital Sign

The input of the analogue signal comes from the vital signals from the human body. We collect those signals using Arduino and transfer them into MATLAB. Then the output is the digital signal that can be analyzed.

This study will first introduce the process of conversion from an analogue signal to a digital signal, the principles of the system, and linear regression. Then, we will present the implementation using Arduino and MATLAB. Finally, we use linear regression to interpret the data and predict other samples using the collected data.

2. RELATED IDEAS

2.1 The Conversion from Analog to Digital

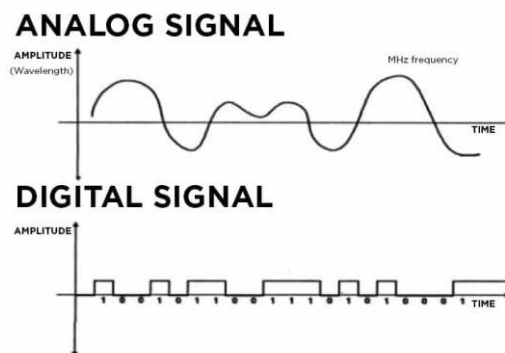


Figure 2. Analog signal vs. Digital signal

The main distinction between the two signals is that analogue signals have continuous electrical signals, but digital signals do not. Analogue signals are represented by sine waves and include signals like the human voice and natural sounds. In contrast, square waves represent digital signals and include the signals of electronic devices such as computers and optical drives.

MATLAB is a program that converts analogue signals to digital signals. MATLAB makes it easier for signal processing, a technique essential in data science and real-time embedded systems, to analyze time-series data. With signal processing produced by MATLAB, we can interpret signals from many sources, design streaming algorithms, and prototype and implement DSP algorithms on PCs.

To analyze bio signals, it is necessary to mathematicise and digitalize natural analogue signals measured through sensors. If digitized, the signal can be easily tested and analyzed. The most important thing in this process is to generalize, that is, to mathematicise the laws of nature. There are three basic formulas for forces in physics: $F_1=ma$, $F_2=kx$, and $F_3=\mu v$. Here, x , v , and a have a differential relationship. Differential relationships are when the function presents the position as a function of time, the first derivative results in its velocity, and the second derivative gives its acceleration. You differentiate position for velocity, and you differentiate velocity for acceleration; they are all interrelated. Therefore, all systems can be expressed as the sum of F_1 , F_2 , and F_3 , including x , v , and a . In other words, when expressed as an expression for x , the equation can be defined as $m\ddot{x}+\mu\dot{x}+kx=F$. This is a generalization (mathematization) of natural laws that can be applied to all systems.

In addition to this equation, all systems can be expressed as sine and cosine through the Euler equation. The formula is $e^{i\theta}=\cos\theta+i\sin\theta$. If the previous equation is a generalized equation that converts all physical signals in the world into a system, the second equation, the Euler equation, is an equation that can be used to express all signals in the world.

Using the Euler equation is necessary to convert a signal to the generalized expression (the first equation) when a certain signal is received. However, using the Euler equation is highly onerous and complicated. Thus, it is necessary to do a Laplace transformation. The Laplace transformation is an integral transform that converts a function of a real variable to a function of a complex variable. Due to its ability to solve differential equations, it applies to many fields in science and engineering. Laplace transformations can also be done through MATLAB, which will be introduced later in the paper.

In addition to the Laplace transformations, there is the Fourier transform. The Fourier transform decomposes functions that are dependent on space or time into functions that are dependent on spatial or temporal frequency, such as a musical chord's representation in terms of the loudness and frequencies of its constituent notes. The frequency-domain representation and the mathematical operation that connects the frequency domain representation with a function of space or time are both referred to as the Fourier transform.

Signals can be classified into continuous or discrete signals. Time can be expressed in the form of $t=0:0.1:2\pi$. It points to the starting point, the interval, and the ending point, respectively. By sampling, the accuracy and the amount of data can be increased. Signals can be sampled and

quantized. All basic signals consist of three signals: Impulse signal (delta) $\delta(t)$, Step signal $u(t)$, and Ramp signal $r(t)$. These three types of signals all have a differential relationship.

2.2 System

Systems can be divided into deterministic and stochastic. A deterministic system is a system that can identify a specific value within a specific time. A stochastic system is a more random and mathematically expressed system that uses probability. Among them, the deterministic system can be divided into Lumped and Distribution. Lumped can be subdivided into linear and non-linear. Among these two, we will use and learn more about the linear system.

A linear system is a system that moves in parallel with time. Its main characteristic is that it changes with time. A linear system has the Addition property and Homogeneity. It can be defined as a linear system only when both of these properties are satisfied. This is called the Principle of superposition: The superposition principle is a property that states for all linear systems, the net response caused by two or more stimuli is the sum of the responses that each stimulus would have caused. So that if input A produces response X and input B produces response Y , then input $(A + B)$ produces a response $(X + Y)$.

2.3 Linear Regression

Linear regression is when one variable increases at a constant rate to another variable. Linear regression analysis is to identify and predict the relationship that depends linearly. Some of the examples of the real-life linear models are sales of houses, economic forecasts, sales.

$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$, where ε is the difference between the predicted y value (Y_p) and the observed y value (Y_o).

Simple linear regression has only one explanatory variable. However, in real life, we often have multiple explanatory variables rather than just one.

The predicted y value (Y_p) can be written as $Y_p = \beta_0 + \beta_1 X_i$.

$\sum_{i=1}^n \varepsilon_i^2$, where n is the number of data points.

$$\hat{y}_i = \beta_1 x_i + \beta_0$$

$$\sum e_i^2 = \sum (y_i - \beta_1 x_i - \beta_0)^2$$

Partial derivative about β_0

$$\frac{\partial \sum e_i^2}{\partial \beta_0} = 2 \sum (-1)(y_i - \beta_1 x_i - \beta_0)$$

$$= -2 \{ \sum y_i - \beta_1 \sum x_i - \sum \beta_0 \}$$

$$= -2 \{ \sum y_i - \beta_1 \sum x_i - n \beta_0 \}$$

Since we are considering when this equation is 0

$$0 = \sum y_i - \beta_1 \sum x_i - n \beta_0$$

$$n \beta_0 = \sum y_i - \beta_1 \sum x_i$$

$$\beta_0 = \frac{\sum y_i - \beta_1 \sum x_i}{n}$$

Let the mean of Y_i be \bar{y}

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Partial derivative about β_1

$$\frac{\delta \Sigma e_i^2}{\delta \beta_1} = (-1)2\Sigma(y_i - \beta_1 x_i - \beta_0)(x_i)$$

$$0 = -2\{\Sigma y_i \cdot x_i - \beta_1 \Sigma(x_i^2) - \beta_0 \Sigma x_i\}$$

$$\beta_1 \Sigma(x_i^2) = \{\Sigma y_i \cdot x_i - \beta_0 \Sigma x_i\}$$

$$\beta_1 \{\Sigma(x_i^2) - \underline{x} \Sigma x_i\} = \Sigma y_i \cdot x_i - \underline{y} \Sigma x_i$$

$$\beta_1 = \frac{\Sigma y_i \cdot x_i - \underline{y} \Sigma x_i}{\Sigma(x_i^2) - \underline{x} \Sigma x_i}$$

We can plot the data with regression software that uses the data to find the parameter estimate with mathematical formulas. Generally, a linear scatter plot includes a straight-line fit to the data. Regression provides us with the best line, which is also called the trendline. The least-square criterion is to minimize the error or residual that inevitably occurs with multiple data points. It minimizes the sum of squared errors from a mathematical function.

$$R^2 = \frac{\text{explained variability}}{\text{total variability}}$$

However, the question is whether this regression line is accurate and reliable enough to predict the new y values. To answer this question, researchers use the R-square value. R-square is the ratio of the explained variation, which measures how much variability is defined by the differences in x values and total variation, which is the total variability in the y value. Thus, the closer the R-square value is closer to 1 or 100 (depending on whether the number is in decimals or percentages), the closer the data points are to the trend line, and vice versa.

3. IMPLMENTATION

3.1 Arduino

Our ultimate goal is to analyze ECG and PCG data, and Arduino, given our restraints, is the most reasonable method to meet this goal. For data analysis, Arduino is the easiest and most useful device.

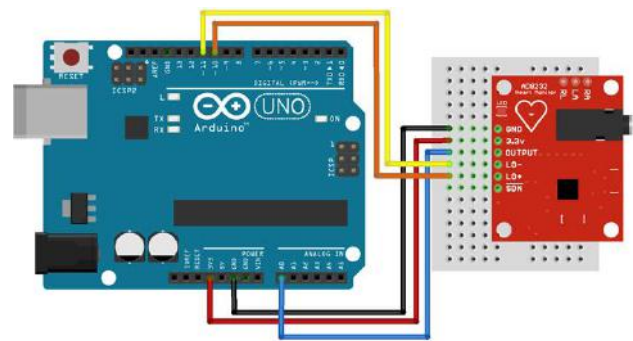


Figure 3. Circuit Diagram of ECG Measurement Toolkit

The output connects to A0, 3.3v connects to 3.3v pin, and LO- and LO+ each connect to pin 11 and pin 10.

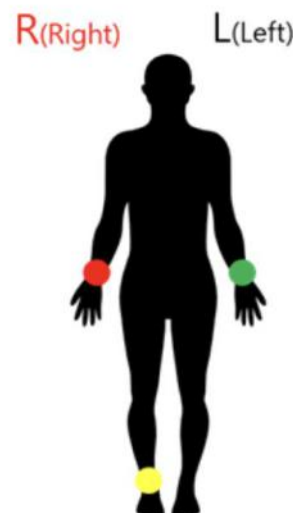


Figure 4. Diagram of the position of each ECG Pads on body parts

The red pad goes on the right wrist, the green pad goes onto the left wrist, and the yellow pad goes onto the right ankle bone.

3.1.1 Code

```
pinMode(10, INPUT); // Setup for leads off  
detection LO +
```

```
pinMode(11, INPUT); // Setup for leads off  
detection LO -
```

LO- and LO+ are terminals that detect whether the pad is well attached to the body. This is a function that expresses how these terminals are connected to pins 10 and 11.

```
if((digitalRead(10) == 1)|| (digitalRead(11) == 1)){  
    Serial.println(!);
```

Since digital has two states, values are returned as LOW and HIGH. In order to use digitalRead, the corresponding pin must already be set to INPUT by pinMode, and the status of other electronic components connected to the pin can be known through this function. When the terminals connected to pins 10 and 11 are well attached to the body, the value is converted to 1. Thus, the value converts to HIGH and current flows. The electrocardiogram functions and records the data through this process.

3.2 MATLAB

MATLAB is a program that shows signals digitally and makes bio signals accurate and easy to analyze, as mentioned earlier. MATLAB is even more useful because it has functions and features that are advantageous for analyzing bio signals. For example, one of the features of MATLAB that is especially helpful in interpreting ECG is the peak analysis feature.

Engineers and scientists use MATLAB to construct, organize, and analyze complex data sets in fields as diverse as climatology, predictive health, medical research, and finance. MATLAB provides the following features:

- Datatypes and preprocessing capabilities designed for engineering and scientific data
- Widely customizable, interactive data visualization
- Numerous pre-built functions for statistical analysis, machine learning, and signal processing
- Extensive, expertly written documentation
- Performance accelerated with simple code changes and additional hardware
- Extend analytics to big data without making major code changes
- Free distributable software without manually recording algorithms
- Automated analysis packaging as components or embeddable source code
- Shareable reports automatically generated from analytics [3]

3.2.1 Organize and Explore Data

Researchers can organize their data into tabular, time series, categorical, and data types designed for text data. Using the MATLAB language, they can write programs based on numerous algorithms in a variety of disciplines. After customizing visualizations interactively, users can automatically generate MATLAB code to reproduce with new data.

3.2.2 Analyze Data with Less Code

MATLAB apps allow users to interactively perform repetitive tasks such as training machine learning models and labelling data. MATLAB apps

generate the MATLAB code needed to programmatically reproduce tasks performed repeatedly by users.

Use the pre-built family of functions to identify sensor drift, signal outliers, missing data, and noise. Users can combine separate data sets by linking tables and synchronizing time series data. Live editor actions allow users to solve these problems interactively within a live script, and the code is automatically generated.

3.2.3 *Extend Analytics with Just a Few Changes*

Accelerate parallel analysis with minimal code changes using parfor loops and multiprocessor hardware. Creating a GPU array allows the target algorithm to take full advantage of GPU acceleration. You can handle out-of-memory data sets using tall arrays that overload hundreds of functions from start to finish in your data analysis workflow to operate on out-of-memory data.

3.2.4 *Share Results*

MATLAB allows packaging analysis in freely shareable software components such as executables, C/C++ libraries, .NET assemblies, Java® libraries, and Python® packages. Users can automatically convert MATLAB code to C and C++ code for deployment to embedded targets. They can use the MATLAB Live Editor to record their work and export the results to PDF, Microsoft® Word, Latex, and HTML.

To use Arduino and MATLAB together, we first connect Arduino and Excel. Then, the data obtained through Arduino is saved as an Excel file. We open the Excel data in MATLAB and analyze it.

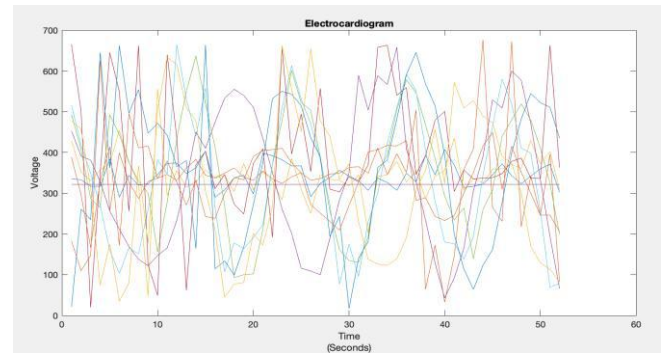


Figure 5. Plotting the data with mean

3.2.5 *Code*

```
ylines(mean(A(:,9)), 'b-', 'Mean concentration')  
%r2018b or later  
  
plot([min(xlim()), max(xlim())], mean(A(:,9))*[1,1])  
% any matlab release
```

The graph represents the aggregated data of the ECG of people of various age groups. Coding in MATLAB allows us to find and plot the mean value among the data.

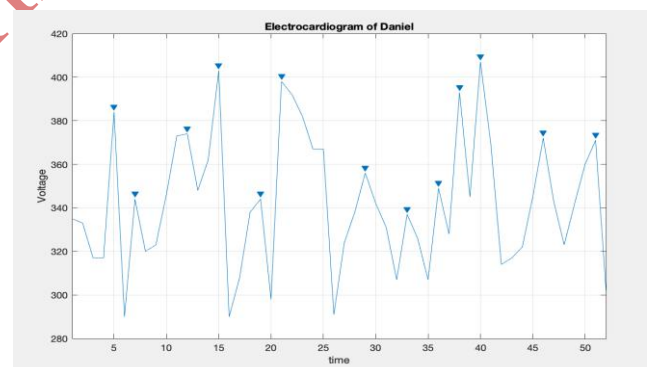


Figure 6. Peak Analysis

```
findpeaks(Voltage, time)
```

Another helpful feature in MATLAB is the ability to plot the graph and code to find the peaks of the graph. For example, it allows the users to easily identify the peaks and analyze them to see the interval. The Zurich Sunspot Relative Number measures both the number and size of sunspots.

findpeaks function can find the positions and values of peaks.

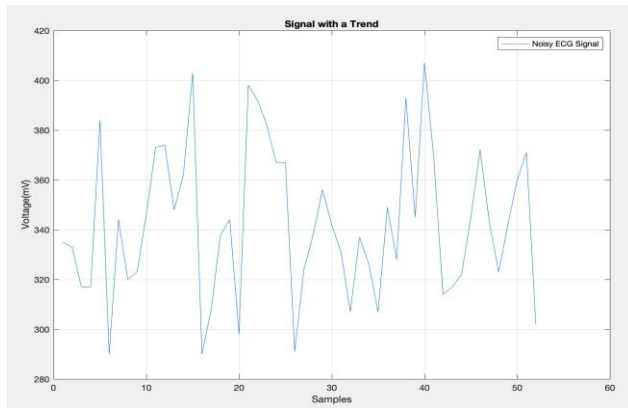


Figure 7. Amplitude Analysis

```
t = 1:length(ecg);
```

```
legend('Noisy ECG Signal')
```

```
grid on
```

This example shows peak analysis of an electrocardiogram (ECG) signal. ECG signal is measured through electrodes attached to the skin and is sensitive to disturbances such as noise and power interference due to motion artefacts.

3.3 Linear Regression

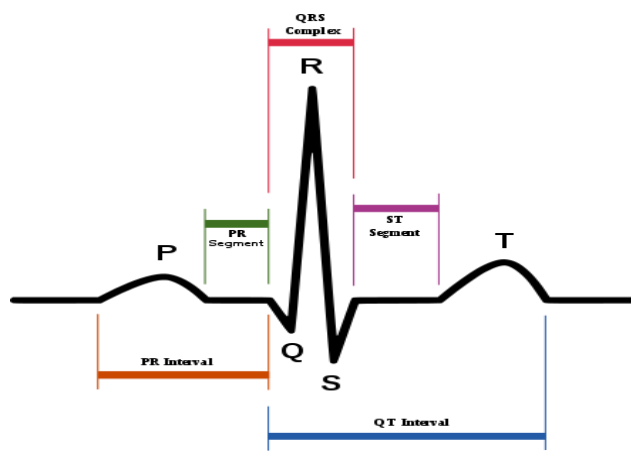


Figure 8. Representation of Normal ECG including QRS complex

The QRS interval is the combination of the three graphical deflections in ECG. The interval is often obvious because it is at the center and has the main spike. Children may have a shorter QRS interval. The QRS interval can be measured from the end of the PR interval to the end of the S wave.

Recent research from P W Macfarlane, S C McLaughlin, B Devine, and T F Yang showed the relationship between age and the QRS interval. They studied three separate populations to determine the effects of age, sex, and race on the ECG. They found that QRS duration increases linearly from about one year old to adolescence, men have longer QRS duration, and no significant correlation between race and ECG. [7]

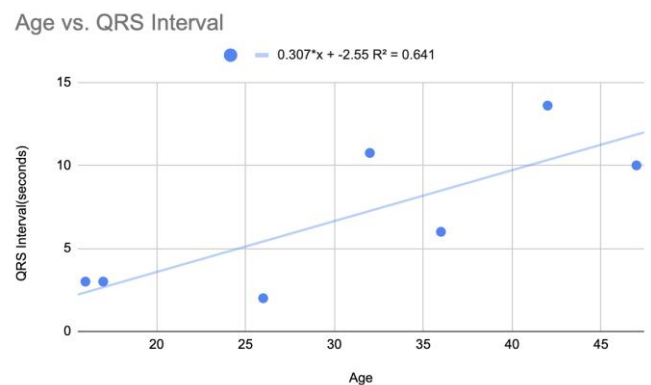


Figure 8. Age vs. QRS Interval from Google Sheets

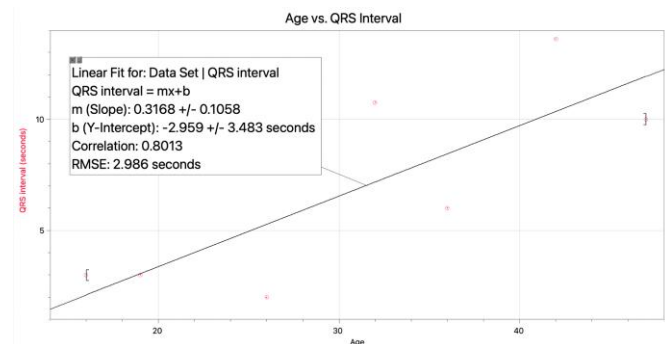


Figure 9. Age vs QRS Interval from Logger Pro

The graphs show that the age and the QRS interval have relatively a linear relationship since the R square value is 0.641, which is over 0.5 and closer to 1, and the correlation from Logger Pro is 0.8013, which is close to 1.

The sample we used was an individual of the age of 24. Using linear regression, we can predict the QRS interval of a person with only their age. Therefore, the x value would be 24. Plugging the x-value into the equation derived above from the graph,

$$y=0.3168*24-2.959=4.644$$

According to our data, we can predict that a person of 24 can have a QRS interval of 4.644 seconds.

The measurement we took was in seconds intervals, while the actual ECG measurement is 0.04 seconds. Therefore, the individual ECG graph for each person did not look exactly the same as the average ECG record. Yet, we still saw the pattern and the interval, though it may not necessarily be the QRS interval. We cannot determine whether the predicted value of the sample above is in the normal range of the standard QRS interval because the interval time is different.

To implement the analysis in Python, we used the module sklearn, numpy, pandas, and matplotlib. We coded for the linear regression that will print the intercept and the coefficient of the line. The linear regression also allows us to code for determining the R-square value.

The result showed -55.48175182481753 [7.32846715] and 0.9753156179610034, representing intercept, coefficient, and R-square

value, respectively. The R-square value is very close to 1, meaning that the data points relatively correlate with the trend line well.

4. CONCLUSION AND FUTURE WORKS

The paper introduces how to analyze analogue signals by converting them into digital signals. Using Arduino and MATLAB, we first analyzed ECG signals by obtaining them as analogue signals and converting them into analyzable digital data. The paper presents the principles of signal and system. It also introduces the effectiveness of using Arduino and MATLAB. Using the principle of linear regression, we organized the obtained data and predicted the ECG for an individual with only their age.

MATLAB is becoming a more frequently used software by numerous researchers and engineers. There are new features constantly added on. Its friendly language makes it easier to use and learn than the other prominent software, Python. In fact, MATLAB is backed by a single, well-known organization that has strong sales.

However, MATLAB is not freeware. Unless the price of MATLAB lowers, there might be a risk that Python or other freeware will replace MATLAB in the future despite its convenience. Freeware makes the software accessible to everyone, including less developed countries, where health issues are the most severe. MATLAB is useful software but making it freeware will enhance its popularity as it will have a better chance to be commercialized in the near future.

There are many methods for mining the data, pre and post processing. We used regression in this paper, which require discrete form of data.

However, the signal formed data can be used to mine other critical signals.

5. REFERENCES

[1] Joo, M.-I., Ko, D.-H., & Kim, H.-C. (2016, May). *Development of Smart Healthcare Wear System for Acquiring Vital Signs and Monitoring Personal Health*. koreascience.or.kr.
<https://www.koreascience.or.kr/article/JAKO201620240561718.pdf>.

[2] Yong Tak Jo. (2019, March 1). *[economist] Analysis of the patient's condition and prognosis with bio-signals*. The JOONANG.
<https://news.joins.com/article/23399547>.

[3] Matlab official website.
<https://kr.mathworks.com/solutions/data-analysis.html>.

[4] Yang, S. (2021, July 1). *The growth rate of patent applications for Korean Intellectual*

Property Office and medical devices is 3.6 times higher than the overall average. The etnews.
<https://m.etnews.com/20210701000171>.

[5] Kim, Ju-Kyeong. (2020, February 2nd). Current status and implications of the Korean people's use of medical services. National Assembly Legislative Research Office.

[6] Jeon, Haesook., Sangkyeong Kang. (2012). Age difference between predictors of the use of medical services between the elderly and the elderly: Implications for medical services in an aging society. *Health and Social Welfare Review*.

[7] Macfarlane, P W et al. "Effects of age, sex, and race on ECG interval measurements." *Journal of electrocardiology* vol. 27 Suppl (1994): 14-9.
doi:10.1016/s0022-0736(94)80039-1

*i*Journals