Purpose of BMI Analysis System Based on Linear Regression With Implementation

Alexander Park

Korea International School

Abstract

Despite advancements in the medical field, there are still many challenges to be overcome, notably heart disease. To prevent and treat these heart diseases, the medical sector has resorted to newly created technology for the early detection of prospective disorders. This paper will use a series of sensors and Arduino technologies to determine whether correlations between specific characteristics such as age, BMI, or pulse can be compared to data obtained from the sensors, including oxygen percent, volumetric variations of blood circulation, and heart's bathymetric and electrical activity, to identify and analyze possible correlations between such variables. This study introduces linear regression to analyze the data and predict other samples based on the acquired data to achieve this comparison of data points.

I. Introduction

Despite the advancements made in the medical field, there are still many challenges, notably heart disease. More than 697,000 people die annually in the United States alone [1]. due to heart disease. In order to prevent and treat these heart disorders, the medical sector has resorted to newly created technology to diagnose potential problems in advance.

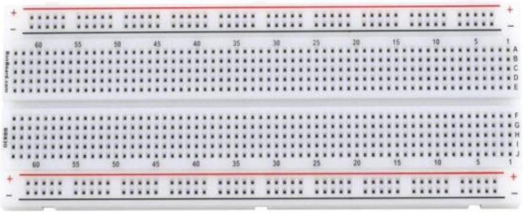
This paper utilizes a series of sensors and Arduino technologies to determine whether correlations between certain characteristics such as age, BMI, or pulse can be compared to data obtained from the sensors, including oxygen percent, volumetric variations of blood circulation, and heart's bathymetric and electrical activity, to identify and analyze possible correlations between such variables. This study introduces linear regression to analyze the data and forecast other samples based on the acquired data in order to carry out this comparison of data points.

The following study is far-reaching since it will add to the field's growing focus; consequently, cardiovascular problems may be able to be diagnosed and treated in a patient before the disease becomes terminal.

II. Embedded part

A. Uno

The Arduino controller/board is one of the most fundamental components of an Arduino circuit. This board may be connected to the Arduino through USB.

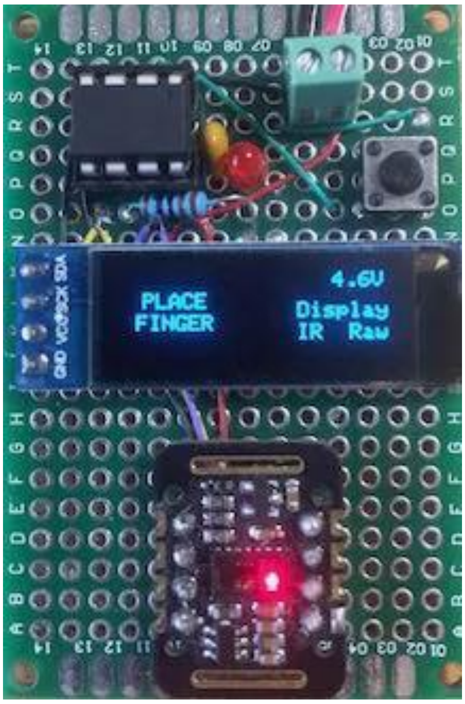


In addition to the controller, the breadboard is required for the construction of a circuit. With Arduino, the breadboard is used to connect electrical components to create the desired circuit. All the hole sockets on the breadboard are connected by a metal wire that runs underneath the holes, which is an intriguing feature. The +/- holes on the top and bottom of the breadboards are joined horizontally. The center holes, on the other hand, are joined by a vertical line. Wires and other things can be plugged into a breadboard to create electric circuits.

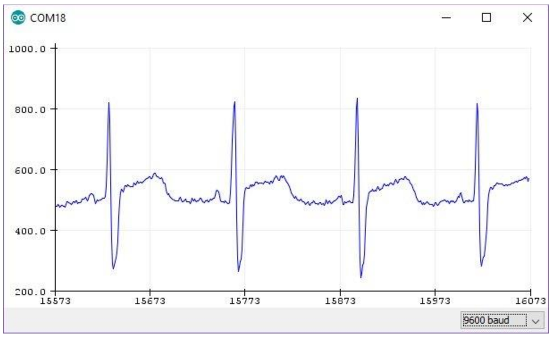
Resistors are utilized for the making of Arduino to limit the amount of current/voltage entering specific circuit components. The voltage supplied by the Arduino controller may be too high and strong for the circuit, particularly the LEDs. Therefore, it is the resistor's responsibility to offer resistance to the LED's high voltage intensity. Each resistor has distinct bands of color. The color of the resistor corresponds to a specific value. The first two bands of a four-band resistor will represent the corresponding digits of a two-digit number. The multiplier value will be applied to this number. Thus, the final band will show a deviation from the established value.

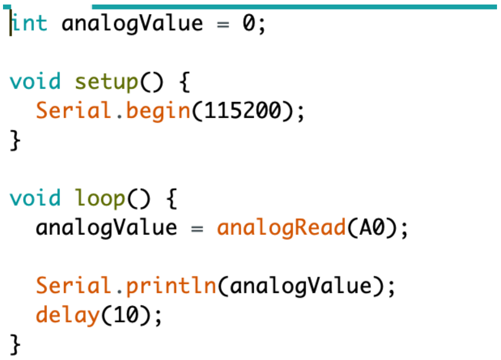
PPG stands for moving photoplethysmogram. The PPG sensor is a low-cost and non-invasive technology for measuring the skin's surface, which provides information and a diagnosis regarding the cardiovascular system.

A typical PPG sensor has both a light source and a photodetector. The PPG's light source emits light to a tissue, and the photodetector measures the light reflected from the tissue. The reflected light is proportional to the fluctuations in blood volume.



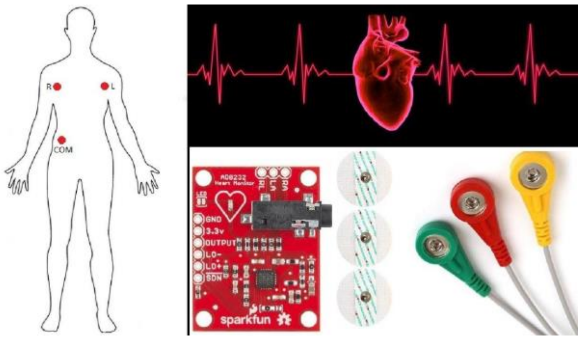
The code for the PPG sensor begins by setting the signal's default analog value, which can take on any value, to 0. We set the data rate on the sensor to 115200 bits per second. The code continuously receives the newly measured analog value through the PPG and prints it with a 10-microsecond delay. This analog value can be visualized using a serial plotter.

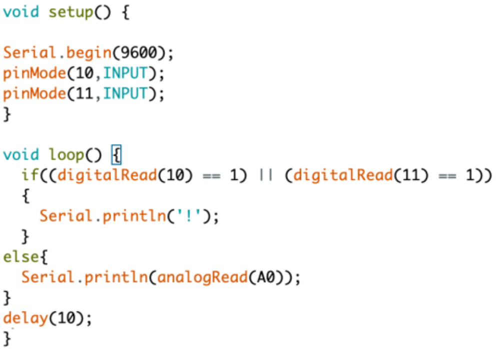




```
int analogValue = 0;

void setup() {
  Serial.begin(115200);
}

void loop() {
  analogValue = analogRead(A0);

  Serial.println(analogValue);
  delay(10);
}
```

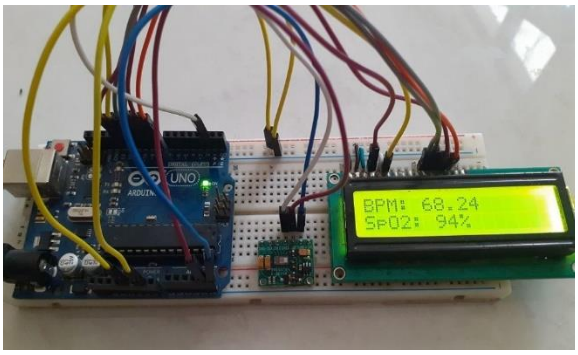
The ECG stands for electrocardiogram. The following sensor can record the passage of electrical impulses via the heart muscle. AD8232 ECG Sensor, for instance, is an Arduino-specific ECG sensor board used to measure the electrical activity of the heart. The electrical activity can be represented graphically as an ECG and produced as an analog readout.





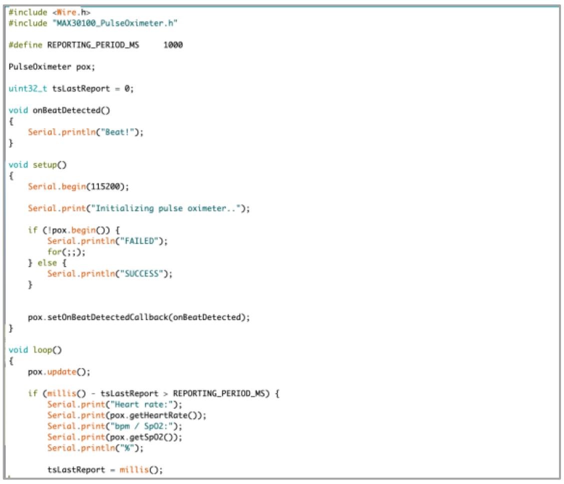
```
void setup() {

Serial.begin(9600);
pinMode(10,INPUT);
pinMode(11,INPUT);
}

void loop() {
  if((digitalRead(10) == 1) || (digitalRead(11) == 1))
  {
    Serial.println('!');
  }
else{
  Serial.println(analogRead(A0));
}
delay(10);
}
```

The code sets the ECG sensor's data rate to 9,600 bits per second. If any of the pin values is 1, the resulting value is nullified/inexpressible. If the pin value is not 1, the analog value is read immediately and can also be displayed graphically.



The oxygen concentration within red blood cells can be measured by the SPO2 sensor, also known as the pulse oximeter sensor.

To determine this value, the sensor works by shining infrared light into the capillaries, where it then detects the amount of light that is reflected off of the gases.



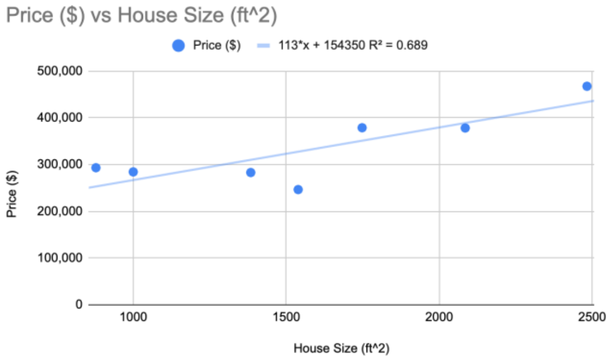
```
#include <Wire.h>
#include "MAX30100_PulseOximeter.h"

#define REPORTING_PERIOD_MS     1000

PulseOximeter pox;

uint32_t tsLastReport = 0;

void onBeatDetected()
{
    Serial.println("Beat!");
}

void setup()
{
    Serial.begin(115200);

    Serial.print("Initializing pulse oximeter..");

    if (!pox.begin()) {
        Serial.println("FAILED");
        for(;;);
    } else {
        Serial.println("SUCCESS");
    }


    pox.setOnBeatDetectedCallback(onBeatDetected);
}

void loop()
{
    pox.update();

    if (millis() - tsLastReport > REPORTING_PERIOD_MS) {
        Serial.print("Heart rate:");
        Serial.print(pox.getHeartRate());
        Serial.print("bpm / SpO2:");
        Serial.print(pox.getSpO2());
        Serial.println("%");

        tsLastReport = millis();
```

The following code uses the Wire. H library. The data rate is set to 115200 bits per second where we were using the PulseOximeter pox and pox. setOnBeatDetectedCallback, the values are derived with the sensors. Once these values are derived, pox.update is used to continuously recover this data where the data is displayed on the LCD through the loop portion of the code.

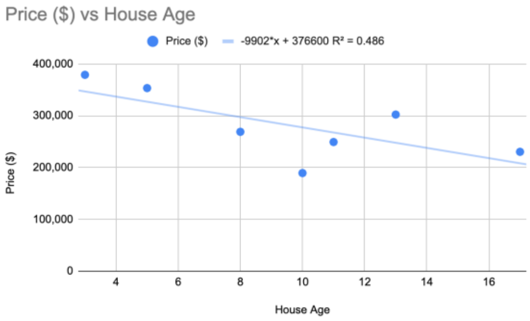
III. Linear Regression

In the past few decades, linear regression, a technique used to model the connection between two scalar variables using a linear approach, has played a central role in statistical analysis. The procedures involved in linear regression begin with the collecting of data pertinent to a particular subject of research. Once this data has been obtained, the data to be examined must be gathered through a process known as data mining, which involves identifying correlations and anomalies among the required and necessary data.

When attempting to determine the characteristics that influence the price of housing, linear regression is used as an example. Certain quantitative characteristics, such as the age or size of the home, can be regarded. By creating a scatter graph that compares the price of the house to these respective variables, when we construct a linear regression, we will be able to generate an equation that helps us compute and predict the price of the property based on characteristics such as when the house was built or if the house was a certain size.

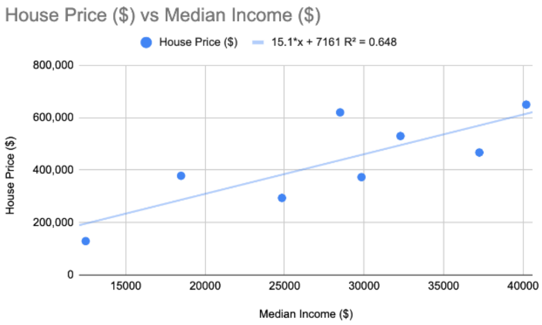


House Size (ft^2)	Price ($)
1,000	284,383
1.384	283,038
878	293,293
2,483	467,585
1,748	378,848
2,085	378,283
1,539	246,575



House Age	Price ($)
13	302,292
8	268,989
11	249,272
10	189,393
17	230,392
3	379,029
5	353,404

Nevertheless, there are other elements that may not be quantifiable, such as the community in which the house is located. In such scenarios, we must be able to quantify qualitative observations in some way. For instance, if we are seeking to estimate the price of a house that may be built in a particular neighborhood, we can obtain the median household income of such a community, which represents how wealthy a neighborhood may be, and design a linear regression based on that average value.



Therefore, if the median household income in New York City is $68,648 we may determine the price of a home in the city by putting the median household income into the x variable of the above-mentioned linear equation.

Following are instances of single linear regression graphs in which predictions are based on a single variable. In addition, there are instances of multilinear regression models. Such a model employs multiple explanatory variables to predict a result.

A. Finding Coefficients

As its name suggests, linear regression is the process of building a linear fit that helps forecast the quantity of one variable in relation to the other. This linear fit is derived from the equation y = mx + n, where m and n must be determined.

The m and n in the equation were obtained using derivatives, which allowed us to determine the instantaneous slope for each given point. The least-squares criterion is applied to obtain the most precise linear fit using derivatives, as the regression's linear fit cannot traverse all data points. By selecting the correct values for m and b in the linear regression equation, this approach ensures that the squared number of the error values for the extremes, which is the squared value of the difference between the original and the predicted values, is minimized. By reducing the distance between the original and the predicted values for each given number, we are able cane values of M and N that most accurately represent the trend of the presented data points.

For the above example in which we graphed the price of the house as a function of the variable price, we were able to construct the linear equation y = 113x + 154,350, where m and n are 113 and 154,350, respectively.

Using the equation, several predictions can be drawn. For example, if we wish to discover the price of a house for a previously unseen size, we can enter the variable house size into x to determine the y value, which is ostensibly the price of the house. Consequently, if we intend to figure out the price of a 3,000 ft^2 home, we may enter this number into the x variable of the linear regression equation (y = 113 x 3,000 + 154,350) and solve for the price. This yields a y-value of 493,350, indicating that the estimated price of a 3000 ft^2 home is approximately $493,350.

B. RMSE

We may test how well a model can use the values of the predictor values to predict the value of the response variable by using the RMSE (Root Mean Squared Error) and MSE (Mean Squared Error) approaches.

As its name implies, the MSE (Mean squared error) provides a squared value of the difference between the predicted and actual values in a particular data set. Thus, this is represented by the equation ($MSE = \Sigma(\hat{y}i - yi)2 / n$).

The RMSE (Root mean squared error) performs essentially the same function as the MSE in that it provides the difference between the predicted and actual values. The distinction, however, is that this value is square-rooted. Therefore, it appears that the two are nearly the same equation, with the exception that RMSE ($RMSE = \sqrt{}\Sigma(\hat{y}i - yi)2 / n$) is the square root of the MSE equation.

In general, the RMSE is more preferred and accepted than the MSE approach. This is because the RMSE is substantially better at handling large error values, such as when one of the disparities between an actual value and a predicted value is considerably bigger than the others. If the theoretical difference between a projected point and its actual value is 100, then the MSE would be 10,000 as opposed to the RMSE's value of 100. This indicates that the interpretation of the RMSE is significantly simpler than that of the MSE, as well as more accurate in the event that an outlier data point exists.

Overfitting happens when a function is too tightly aligned to a limited set of data points as opposed to a wide array of data points. When a model is very long, the data set may contain "noise" or irrelevant information. If the algorithm recognizes this and first fits the data too close to the training set, the data becomes overfitted, resulting in a model that is unable to generalize successfully, implying that it will not be able to predict information in the future.

C. Fitness Test

The R-square value, also known as the coefficient of determination, can be used to determine the accuracy of the linear regression to assess its reliability. This value, denoted by the symbol R2, ranges from 0 to 1 and shows the degree of variability that can be generated by another factor on a variable. The closer this value is to 1 when interpreting it, the stronger the correlation between variables, whereas a value closer to 0 indicates that there is not much of a correlation between variables. Once a linear regression model has been created, it may be used to forecast the impact a change in one variable will have on another. This is vital because it allows us to predict consequences when one of the variables is adjusted.

The code for utilizing python to create a linear regression can be explained as followed:

Code	Explanation

data = {'x1': a, 'x2': a2, 'y': b}
data = pd.DataFrame(data)
x = data[['x1','x2']]
y = data['y']
	Utilizes dictionaries that store values in key-value pairs, allowing users to designate x and y values as the factor for x and forecast for y, respectively.
	Create float lists with the provided x and y values, where a float list is labeled as a (with subscript numbers depending on the number of variables) and the b value determines the float list of y values.
	The pandas, or PD, is a python-based data analysis toolkit function that may be imported to convert a previously produced dictionary created from a data frame into a table.

linear_regression = linear_model.LinearRegression()
linear_regression.fit(X=pd.DataFrame(x), y=y)
prediction = linear_regression.predict(X=pd.DataFrame(x))
print("a value: ", linear_regression.coef_, "b value: ", linear_regression.intercept_)
	Creation of linear regression
	Systematically executing a code that calculates the a and b values for a linear regression equation using derivatives.
	After locating these coefficients, the code generates and prints an equation.

residuals = y-prediction
residuals.describe()
print(residuals)
SSE = (residuals**2).sum()
SST = ((y-y.mean())**2).sum()
R_squared = 1 - (SSE/SST)
print("R^2:\n", R_squared)
	R2 is calculated using the linear regression equation described in Body 2. The code uses error to find the R2 value, which also determines the SSE and SST values.
	Once the R2 value is found, it is displayed to the user.

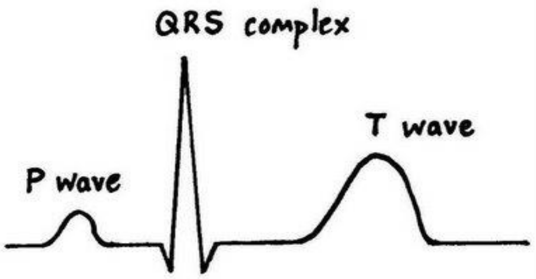
data.plot(kind="scatter",x='x1', y='y',figsize=(5, 5), color = "black")
plt.plot(data['x1'],prediction,color = 'red')
plt.show()
	Final body plots the given information into a graph, which is shown through the plt.show().

IV. Assumptions

A. Factor analysis

ECG

The ECG sensor also referred to as an electrocardiogram sensor can record the pathway of electrical impulses through the heart muscle. The ECG sensor records these electrical impulses to measure the heart rate, which can be used to determine information about the heart and its conditions, such as the heart rhythm or heart rate. To accomplish this, electrodes, which are small plastic patches attached to the skin, are implanted on certain regions of the arms and legs and connected to a sensor that can measure, interpret, and print the heart rate and strength.



To interpret how the ECG sensor works, it is necessary to understand how the normal cardiac cycle works, as the ECG displays data regarding when the 4 chambers function. The typical cardiac cycle begins when the SA node, which is also the heart's pacemaker, provides an electrical impulse to the upper heart atria, which contract, and then sends an impulse through the AV node to the lower ventricles, which contract or pump, a process that is repeated continuously. By measuring the time interval between the P wave, QRS complex, and T wave, which are the three waves that may be obtained from an ECG, and comparing it to the standard values, the status of a heart can be determined. The printed data can be utilized to aid in the diagnosis of certain heart problems, such as arrhythmia, which occurs when the heart beats excessively rapidly, slowly, or irregularly. In addition, ECG sensors can aid in the diagnosis of heart attacks, which take place when the blood supply is abruptly cut off, and cardiomyopathy, which happens when the heart wall becomes thickened or enlarged.

PPG

The PPG sensor, also known as the photoplethysmogram sensor, displays a photoplethysmogram in motion together with the pulse rate. PPG is mostly used to detect volumetric changes in peripheral blood circulation, which involves the transfer and exchange of blood and tissues. A PPG sensor emits light into the tissue, and its photodetector monitors the light reflected from the tissue.

Utilizing a PPG sensor can provide information regarding the cardiovascular system and aid in the diagnosis of potential cardiovascular diseases. Blood volume fluctuations in the microvascular bed of tissue are one of the PPG sensor's most prominent displays [2]. By observing fatty deposits and the buildup of blood pressure, we are able to diagnose prospective cardiovascular issues such as strokes.

SPO$_2$

The amount of oxygen concentration in the blood will also be taken into account. In other words, the circulation of oxygen around red blood cells is calculated. This is accomplished by comparing how much red light and infrared light an individual absorbs.

A healthy individual should have blood oxygen levels of approximately 95%. Thus, by analyzing an individual's blood oxygen level, we can diagnose specific lung, kidney, and heart conditions [3]. Blood oxygen level determines the amount of blood and red blood cells that are carried. By analyzing these results, we can spot potential illnesses for the patient, including hypoxemia due to low levels of oxygen in the blood. Hypoxemia is a sign of breathing and circulation issues, and it can result in symptoms such as shortness of breath.

Pulse Rates:

The variable that will be examined includes the clients' pulse rates. Similar to blood oxygen levels, by comparing the patient's pulse rate to the average value of their peers, we can determine potential diagnoses for the patient. A pulse rate reflects the number of times per minute that capillary contractions cause blood to flow. These abnormally high pulse rates may result in tachycardia, which can lead to weakened heart muscles or heart attacks. In contrast, a lower heart rate can aid in the diagnosis of illnesses such as bradycardia, in which the heart beats significantly fewer than 60 times per minute, which may not be able to provide enough oxygen-rich blood to the body, resulting in dizziness, fatigue, and shortness of breath.

(BMI)

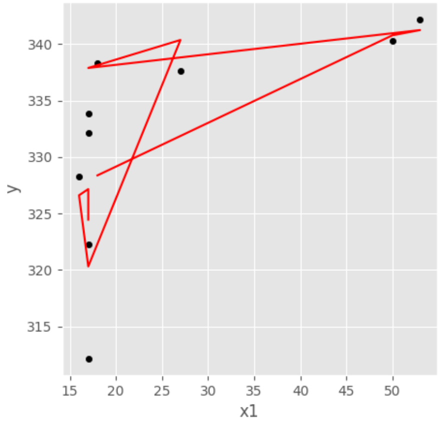
BMI, or body mass index, is a convenient measure that uses both your height and weight to calculate your total body fat, and it has significant implications. Using BMI, it is possible to identify the diseases that an individual is most prone to develop. For example, a higher BMI is associated with an increased risk of having cardiovascular disease, high blood pressure, type 2 diabetes, gallstone issues, and certain cancers. In contrast, a lower BMI may be indicative of impaired immune function, respiratory issues, digestive illnesses, osteoporosis, and even certain forms of cancer.

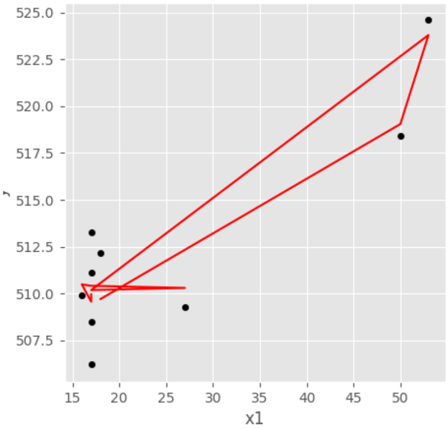
V. Results

Table 1: The collection of graph characteristics (coefficients and fitness test) for respective graphs

	A				
Figure	X_1	X_2	X_3	B	R^2
1	-0.08154603	0.27881943	1.32494409	241.84406444398	0.703297946582434
2	-0.05700138	-0.09857044	-0.08254118	106.41687328866	0.6637481884379302
3	0.42627292	0.44287794	-0.24928517	507.62908485897	0.8274878585023526
4	-1.90342306	0.19252691	N/A	634.1706906455405	0.6466009199118277

<u>Comparing the Variables to the Respective Factors</u>





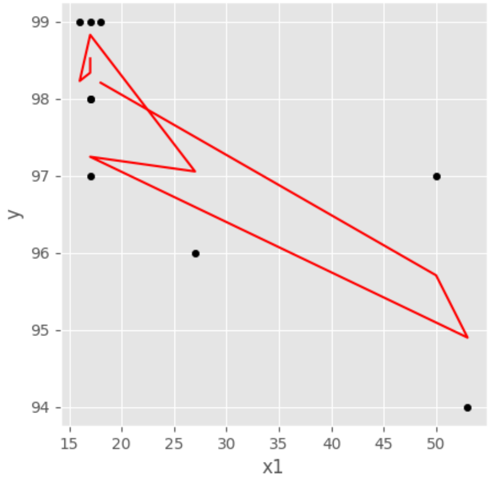


Figure 1. The regression of the 3 characteristics, age, BMI, and pulse rate in respect to the A) ECG sensor data points B) SPO_2 sensor data points C) PPG sensor data points

Identifying the Relationship Between the Data Points Obtained for the 3 Respective Sensors

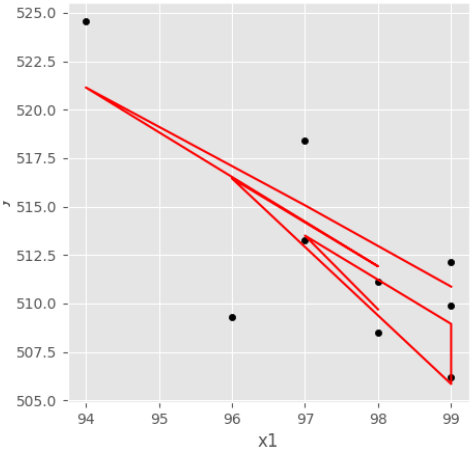


Figure 2: The regression of the 3 different data types received by the PPG, ECG, and SPO2 sensors respectively

Through regressions, it was discovered that the three individual characteristics of BMI, age, and pulse rate had an association with the yield of the respective data points received by the sensors. Observations revealed a linear correlation between the three confounding variables and the final value reported on the ECG sensor (Figure 1A). The regression revealed coefficient values and a high fitness test result of 0.703297946582434, showing a significant association between the characteristics measured from the humans to the ECG sensor data (Table 1). The same conclusions could be drawn from the regression results that compared the three human characteristics to the PPG and SPO2 sensor data, as the coefficients and R2 value were able to establish a correlation with the fitness values of 0.6637481884379302 and 0.8274878585023526, respectively (Figure 1B-C).

This experiment demonstrated that the biological relationships of human health depend on the measured BMI, age, and pulse rate. Thus, it is obvious that the three characteristics that are readily observable in persons can be used to anticipate the data that these sensors can capture, whether it be changes in blood volume, percentage of oxygen in blood vessels, or heart rate. There are a number of extreme outliers within the recorded data points, which indicates a restriction of the observed data points (Figure 1-C). This may be owing to the use of BMI, according to a possible explanation. While the BMI does reflect the optimal level of body fat for a given individual, it is not the most accurate indicator of a person's health because it does not take into account characteristics such as an individual's muscle mass, bone density, or body composition. Therefore, a more precise method of measuring body fat may be required for a more accurate diagnosis. Comparing the individual sensor data points allows us to conclude that there is a correlation between the values of the three sensors (Figure 2). When comparing the three variables, an R2 value of 0.6466009199118277 was observed, demonstrating a strong fitness test (Table 1). This further implies that one or two values can be used to anticipate the data point obtained by the other sensor, suggesting that these sensors could be used to pre-diagnose and predict specific cardiovascular diseases in individuals.

VI. Conclusion and Future works

Based on the research presented in this paper, it can be concluded that the PPG, ECG, and SPO2 sensors have the potential to be utilized in the medical field due to their high correlation and ability to anticipate one another's values. This experiment must be enhanced moving ahead.

Possible strategies to improve this experiment include replacing the PPG, ECG, and SPO2 sensors used in the experiment with more precise laboratory equipment to assure more accurate measurements. As previously indicated, it is also crucial to replace one of the features we assessed in BMI with a more accurate measurement of a person's body fat utilizing an x-ray-based DEXA scan [4]. To build upon the objective of this paper, and with the knowledge that there is the potential to identify relationships between blood volume changes, percent oxygen in blood vessels, and heart rate, data can be collected from individuals with cardiovascular diseases such as cardiomyopathy to develop an understanding of how the data points from the sensors would differ for individuals with cardiovascular diseases and compare them to those of healthy individuals. It is also crucial to build on the human characteristics measured in the experiment, such as age and pulse rate, and compare them to certain persons with cardiovascular diseases to determine which individuals are more prone to developing these conditions.

In the rapidly evolving medical business, the introduction of PPG, ECG, and SPO2 sensors may provide a low-cost option to not only diagnose but also pre-detect some cardiovascular conditions, highlighting the need for their research and application.

Reference

[1] Heart Disease in the United States, CDC

[2] Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H. A review on wearable photoplethysmography sensors and their potential future applications in health care. Int J Biosens Bioelectron. 2018

[3] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US)

[4] Shepherd JA, Ng BK, Sommer MJ, Heymsfield SB. Body composition by DXA. Bone. 2017