Final Project Report

Likelihood Detection of PCOS from Symptoms using IoT

Submitted by

Balubhai Sukani
Soumi Ray
Rohit Andani

Program:

Master of Science in Applied Artificial Intelligence

Course:

Data Analytics and Internet of Things (AAI-530-A1)

University:

University of Sandiego

Professor:

Dr. Premkumar Chithaluru

GitHub Repository:

www.Github.comAlmighty13579

Date: February 23, 2025

Abstract:

PCOS affects a significant number of women globally, impacting their reproductive health and overall well-being. Traditional diagnosis often involves potentially more invasive procedures. Our project aims to create a user-friendly, non-invasive alternative. We have developed an IoT system that uses wearable devices to collect health data. This data is then analyzed using machine learning to predict the likelihood of PCOS symptoms, all visualized in an easy-to-understand dashboard. Our system identifies key factors related to PCOS and processes them using a carefully refined dataset. Through rigorous testing of various machine learning models, we have integrated the best one into our IoT framework. We believe this project shows how wearable tech can be a valuable tool in supporting PCOS detection and promoting proactive health management.

Introduction:

Addressing a Critical Need

PCOS can manifest in various ways, from irregular periods and excessive hair growth to acne and weight gain. Early detection can significantly improve management and outcomes. Our project explores how we can leverage the power of IoT and readily available wearable devices to monitor health non-invasively, making early detection of possible PCOS symptoms more accessible. We are focusing on how smart technology can predict the likelihood of symptoms based on common physiological parameters. We hope to aid early PCOS risk identification based on common physiological markers and patient self-reports.

Components of the Device:

We have proposed an IoT System for Proactive Health Monitoring to identify likelihood of PCOS without pathological tests. Our system has three core components working together seamlessly:

1. Wearable Devices: Think smartwatches and fitness trackers constantly gathering real-time data about your body.
2. Machine Learning Model: A smart "brain" that learns from the data to predict the likelihood of PCOS symptoms.
3. Visualization Dashboard: A clear and intuitive dashboard powered by Tableau, showing real-time analytics and insights for both users and healthcare professionals.

Methodology:

The application is designed to collect data from wearable devices, combine it with user-provided information, and then use machine learning to provide a risk assessment. Here's a more detailed look:

Data Collection Layer: Wearable sensors such as smartwatches, fitness bands, and smart scales capturing parameters like pulse rate, respiratory rate, blood pressure, and weight.

Data Processing Layer: Cloud-based machine learning models processing real-time inputs.
User Interface Layer: A Tableau dashboard visualizing risk scores and trends.

System Architecture:

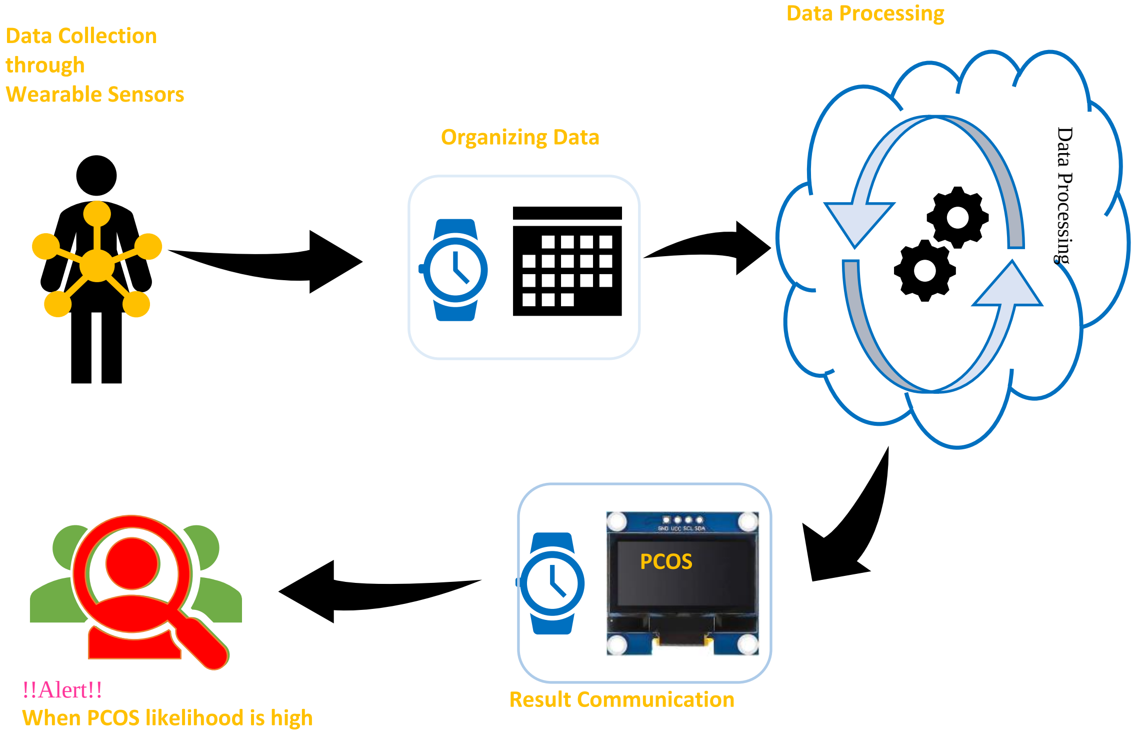


Figure 1: Proposed IoT system

The proposed IoT system architecture is shown in figure 1. The wearable devices with required sensors are the interface between physiological signals and digital world. The collected data is then preprocessed to be organized locally before transmitted to the cloud. The cloud computing is beneficial for real time modification of the created model for better performance by real time adaptation of changes. The cloud-based application will compute data and return the prediction to the local device. In case of alarming situation, the local device will create an alert to initiate the diagnosis process.

Table 1: Key Components of the System Architecture

Component	**Description**
Data Collection Layer	Through the installed application in a wearable device (smartwatches, fitness bands). The IoT sensors of that device will be used for real-time health monitoring (heart rate, temperature, BMI, etc.)

Data Transmission Layer	Collected data will the transmitted through IoT Gateway (Raspberry Pi, Arduino, or cloud-based IoT hub) using compatible Communication protocols (Wi-Fi, Bluetooth, MQTT)
Data Processing & Storage	Data will be processed in Cloud (Google Firebase, AWS, Azure) where the trained model is stored. The prediction with feedback will be stored to upgrade the model in real time.
Machine Learning & AI Analysis	Data preprocessing (feature extraction, normalization) is the part of Model training (Random Forest, CNN, or any deep learning model). The trained model will be used for Prediction and classification of PCOS risk levels
User Interface Layer	This is a Mobile application (Android, iOS) where User dashboard for visualization and doctor consultation will show the prediction output.
Decision & Alert System	The application will Generates alerts and recommendations based on ML model predictions. It will be capable of sending notifications via mobile apps, emails, or SMS

Dataset and Preprocessing:

Cleaning and Preparing the Data:

The dataset was well structured and had no duplicate row. Missing value count was also very low, i.e. 2.

Total missing values in each column with missing data:
Marraige Status (Yrs)   1
Fast food (Y/N)   1

These values are predicted using regression model of their respective feature and the values with equations are given below.

Equation for Marraige Status (Yrs):
Marraige Status (Yrs) = 0.5876 * Age + -10.7838
Column = Marraige Status (Yrs), Imputed Value = 10.4

Equation for Fast food (Y/N):
Fast food (Y/N) = -0.0035 * Age + 0.6243
Column = Fast food (Y/N), Imputed Value = 1

Next we checked the correlation between the features in the dataset. The 41 initial features reduced to 38 by removing highly correlated features like BMI, FSH/LH ratio, and waist circumference. The details are given below:

Weight (Kg) ↔ BMI | Correlation: 0.90
FSH(mIU/mL) ↔ FSH/LH | Correlation: 0.97
Hip(inch) ↔ Waist(inch) | Correlation: 0.87

Feature Selection:

To avoid overfitting and identify potential features for real-time sensing, only the suitable features are filtered out. Correlation between each of the 38 features and PCOS is taken as reference. The details information is shared in table 1.

Table 2: Features sorted by correlation with 'PCOS (Y/N)':

1.	Follicle No. (R)	0.65	20.	Reg.Exercise(Y/N)	0.07
2.	Follicle No. (L)	0.6	21.	LH(mIU/mL)	0.06
3.	Skin darkening (Y/N)	0.48	22.	RBS(mg/dl)	0.05
4.	hair growth(Y/N)	0.46	23.	Blood Group	0.04
5.	Weight gain(Y/N)	0.44	24.	RR (breaths/min)	0.04
6.	Cycle(R/I)	0.4	25.	BP _Diastolic (mmHg)	0.04
7.	Fast food (Y/N)	0.38	26.	II beta-HCG(mIU/mL)	0.01
8.	Pimples(Y/N)	0.29	27.	Waist:Hip Ratio	0.01
9.	AMH(ng/mL)	0.26	28.	PRL(ng/mL)	0.01
10.	Weight (Kg)	0.21	29.	BP _Systolic (mmHg)	0.01
11.	Hair loss(Y/N)	0.17	30.	TSH (mIU/L)	-0.01
12.	Hip(inch)	0.16	31.	Pregnant(Y/N)	-0.03
13.	Avg. F size (L) (mm)	0.13	32.	I beta-HCG(mIU/mL)	-0.03
14.	Endometrium (mm)	0.11	33.	FSH(mIU/mL)	-0.03
15.	Avg. F size (R) (mm)	0.1	34.	PRG(ng/mL)	-0.04
16.	Pulse rate(bpm)	0.09	35.	No. of aborptions	-0.06
17.	Hb(g/dl)	0.09	36.	Marraige Status (Yrs)	-0.11
18.	Vit D3 (ng/mL)	0.09	37.	Age (yrs)	-0.17
19.	Height(Cm)	0.07	38.	Cycle length(days)	-0.18

The following features showed the significant correlation with the presence of PCOS:

1. Follicle No. (Right Ovary)
2. Follicle No. (Left Ovary)
3. Skin Darkening (Yes/No)
4. Hair Growth (Yes/No)
5. Weight Gain (Yes/No)
6. Menstrual Cycle Regularity (Regular/Irregular)
7. Frequency of Fast Food Consumption (Yes/No)
8. Marraige Status (Yrs)
9. Age (yrs)
10. Cycle length(days)

The last three has significant inverse correlation with PCOS. ***The bottleneck of this project is that the above features are mostly non-traceable by common wearables in present scenario.***
To design an IoT system, our prime target is identifying the potential features among these which can be tracked using non-invasive sensors which are also mountable in a wearable device. Externally Measurable Features which have impact on PCOS as well as are available in our dataset are listed in table 2.

Table 3: Features measurable through wearable devices

Sl No	Features	Wearable Devices Available
1	Pulse rate (bpm)	Apple Watch, Fitbit, Empatica E4, Biovotion, GENEActiv, FreeStyle Libre, Dexcom G6, Muse EEG headband
2	RR (breaths/min)	Fitbit, Samsung Charm, Withings ScanWatch, WristOx2, Sense-Wear armband, Health Tags
3	BP _Systolic (mmHg)	HeartGuide, Smart Wear
4	BP _Diastolic (mmHg)	HeartGuide, Smart Wear
5	Weight (Kg)	Fitbit, Samsung Charm

Using the best possible combination we tried to reach a solution. The features used for our prediction system are **Cycle(R/I), Pulse rate(bpm), RR (breaths/min), BP_Systolic (mmHg), BP _Diastolic (mmHg), Waist:Hip Ratio, Weight (Kg)**.

Model Selection:

To find the best possible model the features mentioned in last segment are used along with few self-reporting parameters. The final feature list is –

Measurable:
'Pulse rate(bpm)',
'RR (breaths/min)',
'BP _Systolic (mmHg)',
'BP _Diastolic (mmHg)',
'Waist:Hip Ratio',
'Weight (Kg)',
'Weight gain(Y/N)'
'Skin darkening (Y/N)',
'hair growth(Y/N)',
'Cycle(R/I)',
'Age (yrs)',
'Cycle length (days)

Self-reported:

Initially we started with the basic models and the performance was not good. Hence, we modified the models through hyperparameter tuning and trained to identify the best performing model.

We tested three machine-learning models to predict the likelihood of PCOS, Logistic Regression, Random Forest Classifier and SVM. The results with their hyperparameter details are as below in figure 2.

===== Tuned Logistic Regression =====

Best Parameters: {'C': 1, 'class_weight': None, 'penalty': 'l2', 'solver': 'lbfgs'}

Accuracy: 0.84

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.86	0.89	59
1	0.60	0.75	0.67	16

===== Tuned Random Forest =====

Best Parameters: {'bootstrap': True, 'class_weight': 'balanced', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}

Accuracy: 0.79

Classification Report:

	precision	recall	f1-score	support
0	0.88	0.85	0.86	59
1	0.50	0.56	0.53	16

===== Tuned SVM =====

Best Parameters (SVM): {'C': 10, 'class_weight': None, 'gamma': 'scale', 'kernel': 'rbf'}

Accuracy: 0.8

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.88	0.87	59
1	0.53	0.50	0.52	16

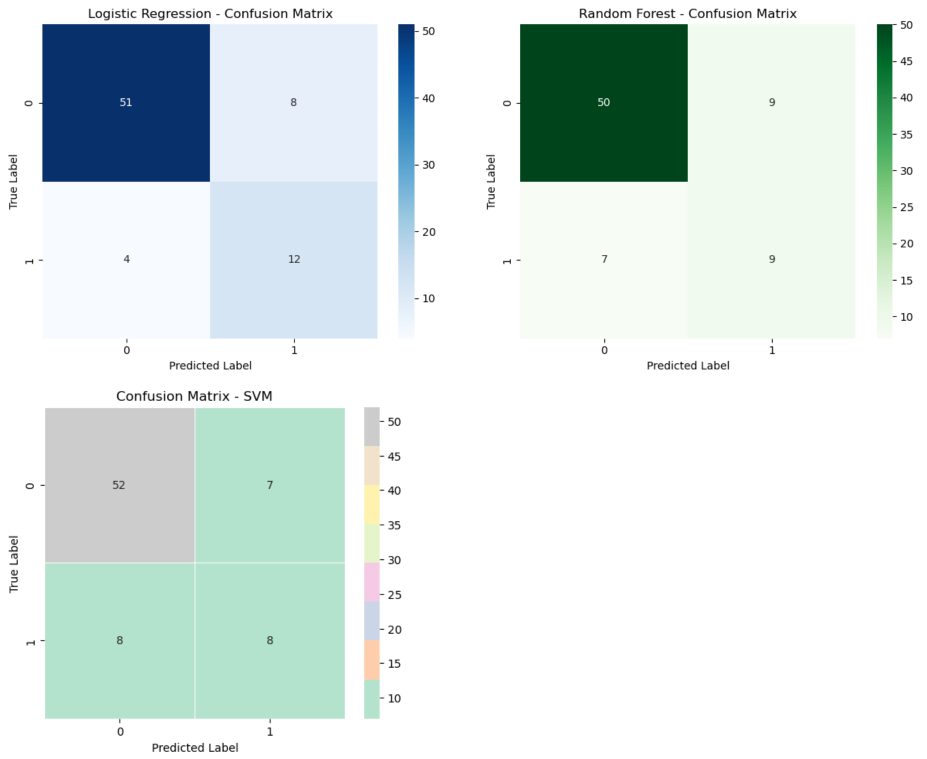


Figure 2: Parameter details and performance matrices of Logistic Regression, Random Forest Classifier and SVM models

Results and Discussion:

The models are evaluated using the quantitative parameters like accuracy, precision, recall, and F1-score to identify the best performer. The Logistic Regression Classifier showed the best performance, achieving an accuracy of 84%. SVM followed closely at 80%, and Logistic Regression at 79%.

Logistic Regression outperform by turning a better F1 score than others though in practice it is quite low, only 67%. To improve the performance and design a better model, more diverse and larger dataset is required.

Conclusion and Future Scope:

This project successfully demonstrates the feasibility of using an IoT system to predict PCOS symptoms non-invasively. The further improvement of prediction accuracy depends on availability of better dataset of higher volume, better geographical diversity. The final model must able to stream real-time data for continuous monitoring when required.

References:

ESHRE, T. R., & ASRM-Sponsored PCOS Consensus Workshop Group. (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertility and sterility, 81(1), 19-25.

Tehrani, F. R., & Amiri, M. (2019). Polycystic ovary syndrome in adolescents: challenges in diagnosis and treatment. International journal of endocrinology and metabolism, 17(3), e91554.

Conway, G., Dewailly, D., Diamanti-Kandarakis, E., Escobar-Morreale, H. F., Franks, S., Gambineri, A., ... & Yildiz, B. O. (2014). The polycystic ovary syndrome: a position statement from the European Society of Endocrinology. European journal of endocrinology, 171(4), P1-P29.

Islam, S. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K. S. (2015). The internet of things for health care: a comprehensive survey. IEEE access, 3, 678-708.

Pantelopoulos, A., & Bourbakis, N. G. (2009). A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(1), 1-12.

Heikenfeld, J., Jajack, A., Rogers, J., Gutruf, P., Tian, L., Pan, T., ... & Wang, J. (2018). Wearable sensors: modalities, challenges, and prospects. Lab on a Chip, 18(2), 217-248.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

Few, S. (2012). Show Me the Numbers: Designing Tables and Graphs to Enlighten/Stephen C.

Appendix I

Dataset Overview:

The dataset consisted of 41 features, categorized as:

4.1 Demographics and Personal Information

[1]. Age (yrs)
[2]. Blood Group
[3]. Marriage Status (Yrs)
[4]. Pregnant (Y/N)
[5]. No. of abortions

4.2 Physical Measurements

[6]. Weight (Kg)
[7]. Height (Cm)
[8]. BMI
[9]. Hip (inch)
[10]. Waist (inch)
[11]. Waist:Hip Ratio
[12]. Pulse rate (bpm)
[13]. RR (breaths/min)
[14]. Hb (g/dl)
[15]. BP Systolic (mmHg)
[16]. BP Diastolic (mmHg)

4.3 Menstrual and Cycle Information

[17]. Cycle (R/I) (Regular/Irregular)
[18]. Cycle length (days)

4.4 Hormonal Levels

[19]. I Beta-HCG (mIU/mL)
[20]. II Beta-HCG (mIU/mL)
[21]. FSH (mIU/mL) (Follicle-Stimulating Hormone)
[22]. LH (mIU/mL) (Luteinizing Hormone)
[23]. FSH/LH Ratio (Ratio of FSH to LH)
[24]. TSH (mIU/L) (Thyroid Stimulating Hormone)
[25]. AMH (ng/mL) (Anti-Müllerian Hormone)
[26]. PRL (ng/mL) (Prolactin)
[27]. PRG (ng/mL) (Progesterone)

4.5 Lifestyle Factors

[28]. Weight gain (Y/N)
[29]. Hair growth (Y/N)

[30]. Skin darkening (Y/N)
[31]. Hair loss (Y/N)
[32]. Pimples (Y/N)
[33]. Fast food (Y/N)
[34]. Regular exercise (Y/N)

4.6 Blood and Metabolic Information

[35]. RBS (mg/dl) (Random Blood Sugar)
[36]. Vit D3 (ng/mL) (Vitamin D3)

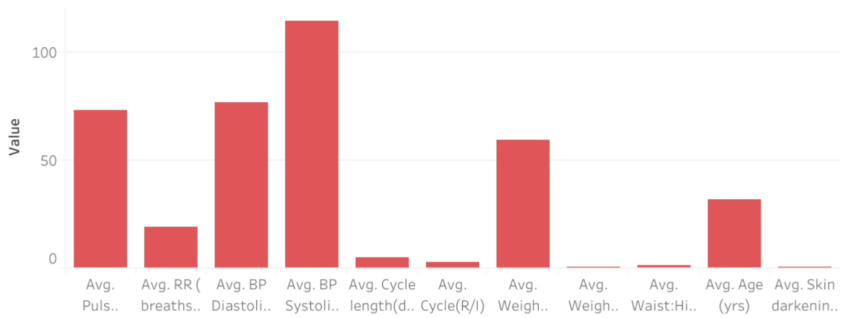
4.7 Ultrasound and Ovary Measurements

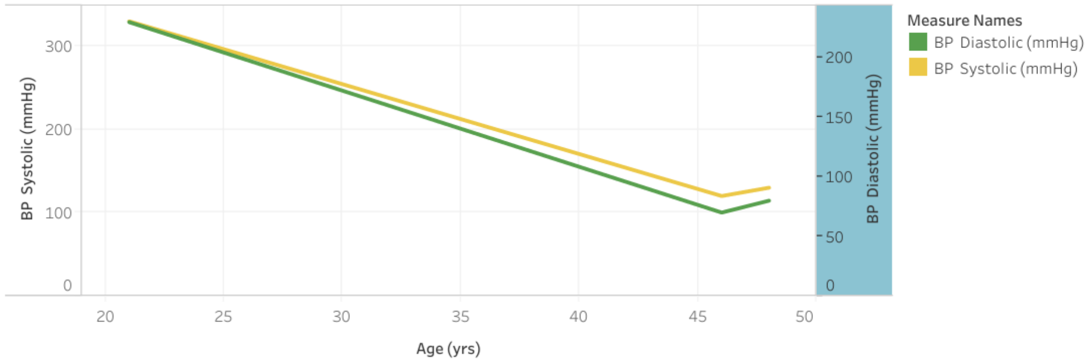
[37]. Follicle No. (L) (Follicle count on Left Ovary)
[38]. Follicle No. (R) (Follicle count on Right Ovary)
[39]. Avg. F size (L) (mm) (Average Follicle Size Left)
[40]. Avg. F size (R) (mm) (Average Follicle Size Right)
[41]. Endometrium (mm) (Endometrial Thickness)

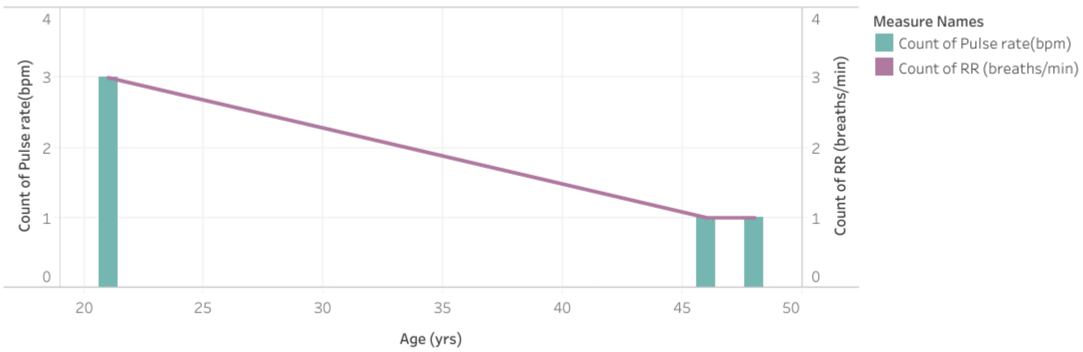
Appendix II

Tableau Dashboard:

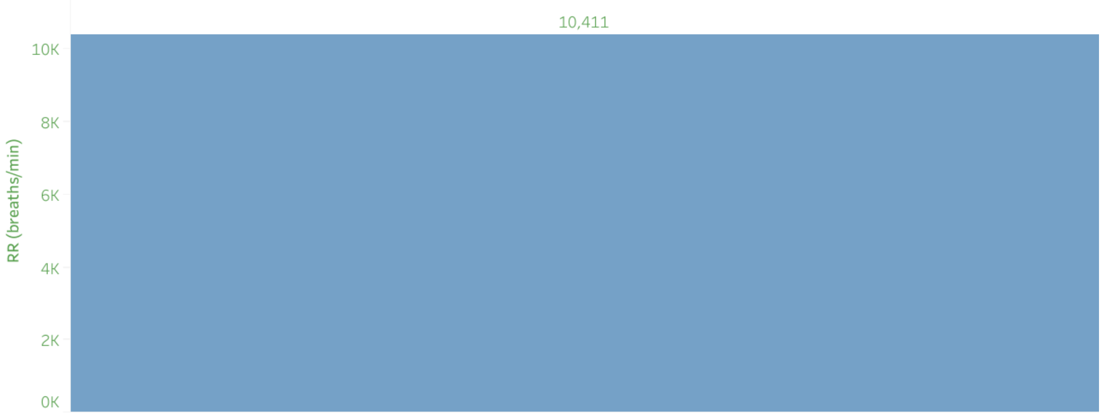
Visualizing the Insights through tableau helps a better understanding.






Measure Names
Count of Pulse rate(bpm)
Count of RR (breaths/min)
Count of Pulse rate(bpm)
Count of RR (breaths/min)
4
3
2
1
0
20
25
30
35
40
45
50
Age (yrs)

LIVE VIEW


10,411
RR (breaths/min)
10K
8K
6K
4K
2K
0K