*Article*

# An IoT-based Bi-Cluster Forecasting using Automated ML-Model Optimization for COVID19

**Hasan Tariq[1,*], Farid Touati[1], Damiano Crescini[2], and Adel Ben Mnaouer[3]**

1 Department of Electrical Engineering, College of Engineering, Qatar University, 2713, Doha, Qatar; hasan.tariq@qu.edu.qa (H.T.); touatif@qu.edu.qa (F.T.)
2 Dipartimento di Ingegneria dell'Informazione, Brescia University, 25121 Brescia, Italy; damiano.crescini@unibs.it (D. C.)
3 Dept. of Computer Engineering and Computational Sciences, Canadian University Dubai, Dubai, UAE; adel@cud.ac.ae (A. B. M.)
* Correspondence: hasan.tariq@qu.edu.qa

**Abstract:** The current COVID19 pandemic has raised huge concerns for outdoor air quality due to expected lungs deterioration. These concerns include the challenges in the scalable prediction of harmful gases like carbon dioxide, iterative/repetitive inhaling due to mask and environmental temperature harshness. Even in the presence of air quality sensing devices, these challenges lead to failed planning and strategy against respiratory diseases, epidemics, and pandemics in severe cases. In this work, a dual time-series with bi-cluster sensor data-stream-based novel optimized regression algorithm was proposed with optimization predictors and optimization responses that use automated iterative optimization of the model based on the similarity coefficient index. The algorithm was implemented over SeReNoV2 sensor nodes data, i.e. multi-variate dual time-series of environmental and US Environmental Protection Agency standard sensor variables for air quality index measured from air quality sensors with geospatial profiling. The SeReNoV2 systems were placed at four locations that were 3 km apart to monitor air quality and their data was collected at Ubidots IoT platform over GSM. Results have shown that the proposed technique achieved a root mean square error (RMSE) of 1.0042 with a training time of 469.28 seconds for normal and RMSE of 1.646 in the training time of 28.53 seconds for optimization. The estimated R-Squared error of 0.03 with Mean-Square Error for temperature 1.0084 $^{o}$C and 293.98 ppm for CO2 was observed. Furthermore, the Mean-Absolute Error (MAE) for temperature 0.66226 $^{o}$C and 10.252 ppm for CO2 at a prediction speed of ~5100 observations/second for temperature 45000 observations/second for CO2 due to iterative optimization of the training time 469.28 seconds for temperature and 28.53 seconds for CO2 was very promising in forecasting COVID19 countermeasures before time.

**Keywords:** Indoor air quality; forecasting; machine learning; IoT; Covid-19; environmental mapping; pandemic.

## 1. Introduction

According to the WHO and US Environmental Protection Agency (EPA) guidelines, the future air quality and climatic conditions are a signature of life security for healthy respiration. The quality of respiratory life processes is directly related to air quality specifically dependent on oxygen (O2) and carbon dioxide (CO2) at a particular geo-location under tolerable temperature (WHO, 2021) [1]. The National Ambient Air Quality Standards (NAAQS) report that the gradual deterioration in urban air quality is ambient each year due to the increasing population, chemical emissions from machinery, and depreciation in green ecology [2]. Several studies have concluded that a poor air quality index (AQI), a higher concentration of CO2, and temperature extremities are more disastrous and fatal when inhaled/exhaled and its real-time monitoring is the key to public safety [3]. This mandated measurement methods like geo-spatially sensed outdoor gases to be a critical decision source in forecasting the COVID-19 threat intensity at lower temperatures

 doi:10.20944/preprints202212.0534.v1



and higher CO2 [4]. The flu-based pandemics and COVID-19 being a scalable problem needed a scalable research approach that is an ambient gap.

Globally, CO2 sensing time-series analysis and forecasting are the most widely capitalized approach in respiratory research i.e., an ensemble time-series model with machine learning approaches as the projection benchmark have shown that China's carbon peak will be achieved by 2021–2026 with > 80% probability [5] from the logged dataset with a gap in real-time CO2 sensing at regional temperatures. The Long Short-Term Memory (LSTM) networks, DeepLMS resulted in average testing Root Mean Square Error (RMSE) <0.009, and average correlation coefficient between ground truth and predicted values $r \geq 0.97$ ($p<0.05$) when tested on logged data from one database pre-COVID19 and two during-Covid-19 pandemic years [6]. This study [6] had gaps in data interpretation and collective forecasting from multiple real-time CO2 and temperature sensing units due to data structuring challenges. The first step has been structuring the dual-series sensor data into decomposed time-series as mentioned in the reviews as either additive or multiplicative by valued researches [7-9].

The second challenge was sorting of time-series to prepare for the next stage called time-series trend assessment by using Theil-Sen's Slope (TSS), Mann-Kendall (MK), Modified Mann-Kendall (MMK), and Kendall Rank Correlation (KRC) tests need the incorporation of improved trending for seasonality tests [8, 9]. Several studies used the above-mentioned tests very useful for logged data but they had gaps in real-time sensor data. The real-time processing, time-series decomposition methods REG and GAM based on OLS; FFT, FFT, AVG, LOESS, and LHM based on Backfitting [10] had gaps in stationarity assessment. Various time-series hypothesis tests, Durbin-Watson test, Box-Pierce, and Ljung-Box tests, Breusch-Godfrey test, Jarque-Bera test, and Augmented Dickey-Fuller test were used for stationarity and seasonality assessment and are useful in auto-regressive moving average (ARIMA); the advanced Seasonal autoregressive integrated moving average model (SARIMA) [11] needed clustered approach for real-time forecasting of multi-variate sensor data. The statistical techniques and machine learning approaches mentioned above were found to have the sensors' dependent and wireless sensor network-based anomalies' dependent results.

The third major gap was the absence of an optimized and adaptive real-time forecasting approach for networked CO2 and temperature measurement sensor nodes. For this many air quality, sensing systems were studied. The top contributions AirNut, PA-I and PA-II, Egg, PATS+, and S-500, CairClip, Portable ASLUNG, AirSensEUR, Met One, AQY v0.5, Vaisala AQT410, 2B Tech, and AQMesh V3.0 systems had measurement capabilities in specific pollutants and gases [12] impacted by real-time health monitoring systems [13] and infrastructure and architectures of specialized platforms [14]. FIS SP-61 by FIS, O3-3E1F by CityTechnology, AirSensEUR v.2 by LiberaIntentio, and S-500 by Aeroqual, and AirSensEUR used a built-in AlphaSense OX-A431 limited to $O_3$ and likewise the PMS1003 and PMS3003 by Plantower; DC1100 PRO and DC1700 by Dylos; OPC-N2 by AlphaSense had only sensing support for particulate matter (PM) [15] from multi-agent perspective had gaps in clinical biomarker space of COVID-19 using feature selection and prognosis classification for time-series forecasting problem [16]. The networked sensing errors found in the previously mentioned works using CO-3E300 by City Technology; CO-B4 by Alphasense, MICS-4515 by SGX Sensortech, Smart Citizen Kit by Acrobotic, and the RAMP had wireless sensor network errors that can be corrected by the multi-objective prediction monitoring algorithm [17-19]. All the above-mentioned research were suited for a fixed network of sensing systems but had gaps in every updating threshold that gave errors in forecasting spatially placed sensing nodes clusters [20].

In clustered sensing for CO2 and temperature, there was a pressing need for a concurrent forecasting chain in addition to dimensionality reduction using matrix factorization (MF) [21] for the air quality nodes with parametric ML model deployment support on embedded systems like SeReNoV2 [22]. Considering the recent studies conducted at European Commission, Joint Research Centre (JRC) [23] on the impact of masks on CO2

 doi:10.20944/preprints202212.0534.v1



concentration zones in the breathing zones have concluded that the increase in the CO2 due to breathing exhaled air temperature as well [24].

The existing O-AQNs, Urban AirQ, Smart Citizen Kit, Air SensUR 4.0, SeReNo V1, and AirQ Mesh needed improvement in AQI dependent principal component approach [25] in the scope of automated optimization of forecasting. The multi-time series-based forecasting required novel melioration in linear regression and tree-based time-series learning, regression, and forecasting tree that was an innovative step in the SeReNoV2 AQM systems.

The literature review and addressed gaps have shown that COVID-19 and pandemics being a scalable problem and diseases that needed a novel scalable and ubiquitous solution. The present and future safety of the entire populous and assisting the EPAs and state health agencies in precise decision-making through well-informed forecasting is the key motivation. The innovative aspects and novel contributions of this research are: a) bi-cluster regression, i.e., real-time CO2 and temperature sensing systems placed at different locations with different surroundings (different CO2 and temperature curves) to be used for evaluation as a bi-cluster time-series interpolated with air quality index (AQI) from principle pollutants; b) networked assessment to have multiple sensing sources based on the dual-redundancy and resilience in real-time forecasting and machine learning model (MLM) training; c) the automated optimization to tune the scalable spatial gradients with different thresholds; d) automated iteration to achieve the minimum RMSE and MAE for trackable similarity coefficient index (SCI) for accurate forecasting. The research work is organized as:

1. The Real-time Gradient Aware Multi-Variable Sensing Model (GAM-VSM)
2. The Optimized Bi-Cluster Regression Machine Learning Model (OBR-MLM)
3. Case Study: Urban Scale IoT-based AQI Monitoring System.

## 2. Materials and Methods

The materials in this work comprise of a real-time air quality monitoring system and methods consist of GAM-VSM and OBRM-MLM. The results section gives further insights into this contribution.

### *2.1. The Real-time Multi-Variable Geospatial Gradient-aware AQI Sensing Model (GAM-VSM)*

For the precise impact of CO2 and temperature for COVID19, a real-time multi-variable structured data time-series vector was needed to proceed with geospatial profiling of gradient awareness as per our past work [25, 26]. Let us consider an EPA standard outdoor air quality index (O-AQI) real-time variables as temperature T in centigrade, pressure P in pascals, humidity H in %, volatile organic compounds VoC (ppm), particulate matter as PM (ppm), Ozone as O3 (ppm), Nitrogen Dioxide as NO2 (ppm), Carbon Monoxide as CO2 (ppm), and Sulphur Dioxide as SO2(ppm). The real-time O-AQI data was proposed as a commutative time series multi-variable vector $V_{O\text{-}AQI}$ of two non-linear time-series with $t_1$ and $t_2$ of environmental E and gas G sensors data at a particular geo-location L given as:

$$V_{O\text{-}AQI}(t) = [E(t_1), G(t_2)]: L(t) \quad (1)$$

where t = (0, 1, 2, 3, …}

The practicality of response time of heterogeneous sensors were taken into account for non-linear time-series decomposition t, gas sensor response time $t_2$ is greater than the response time of environmental sensors $t_1$ with relationship $t_2 > t_1$ given as:

$$t_2 = 3t_1 \quad (2)$$

where $[t_1, t_2] \in t$

The environmental sensor variables function E for sensor array $A_E$ (T, P, H, VoC, PM) as E ($A_E$, $t_1$); and for gas sensors array $A_G$ (O3, NO2, SO2, CO) as G($A_G$, $t_2$) and position

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



vector L as reference function GPS using GSM network cell locations (using AT+CIPGSMLOC=1,1) for $L_{GPRS}$ and GPS module as $L_{GPS}$ (using AT+CGPSINF). For precise AQM the $L_{GPS}$ must belong to the slope of $L_{GPRS1}$ and $L_{GPRS2}$ in a particular slope format by NEMEA specifier for consecutive cells is given as:

$$L_{GPS}(X, Y) \in [L_{GPRS1}(X2, Y2), L_{GPRS2}(X1, Y1)] \quad (3)$$

The agreed $L_{GPS}$ was termed as L(t) where condition (3) was satisfied. From equations (1), (2), and (3) the finalized AQM vector of $V_{O\text{-}AQI}$ was derived as:

$$V_{O\text{-}AQI}(t) = [E(A_E(T, P, H, VoC, PM), t_1), G(A_G(O3, NO2, SO2, CO), t_2)] : L(t) \quad (4)$$

Three bounded value conditions we applied on GAM programmed in the SeReNo V2 firmware are presented in Figure 1:

a) The mandatory gradient unit $\Delta_1CO2$ to monitor the CO2 gradient from inhaled air at temperature $\Delta_1T$.
b) The role of the gradient of the temperature of exhaled air $\Delta_2T$ with $\Delta_2CO2$ recycled in the breathing zone due to mask.

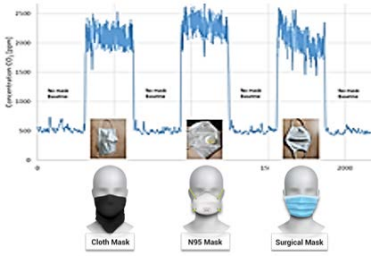



**Figure 1.** The Model Optimization is based on gradients in Temperature and CO2

The optimization scalar is presented as (CO2 is in ppm):

$$Mask(\Delta CO2) = \Delta_1CO2 \times \Delta_1T + \Delta_2CO2 \times \Delta_2T \quad (5)$$

*2.2. The Optimized Bi-Cluster Regression Machine Learning Model (OBR-MLM)*

The GAM reduced the bulk time-series curation operations needed for forecasting. The dual time-series data was queued to OBRM with (AQI, CO2) and (AQI, Temperature) vectors at the same time with $t_1$ and $t_2$ time-series. The iterative regression parameter setting was performed based on default parameters (RMSE, RSS, and MAE). On every cycle, these parameters were optimized as the KPI requirements. The two simultaneous regression models were trained for $A_E(t_1)$ and $A_G(t_2)$ vectors. The root-mean-square error (RMSE), and mean absolute error (MAE) were the common KPIs that were analyzed before model approval. The approved model was set for forecasting from SeReNoV2 test data and disapproved was fed to an optimizer that used a configurable tree-based machine learning approach by variable iterations based on the similarity coefficient index (SCI). The generic regression model Y for

$$Yt = \sum_{m=1}^{i} \beta_m X_{m,t} + \varepsilon = \beta 1 X_{1,t} + \beta 2 X_{2,t} + \cdots + \beta i X_{i,t} + \varepsilon \quad (6)$$

The flowchart of OBRM is presented in figure 2 below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



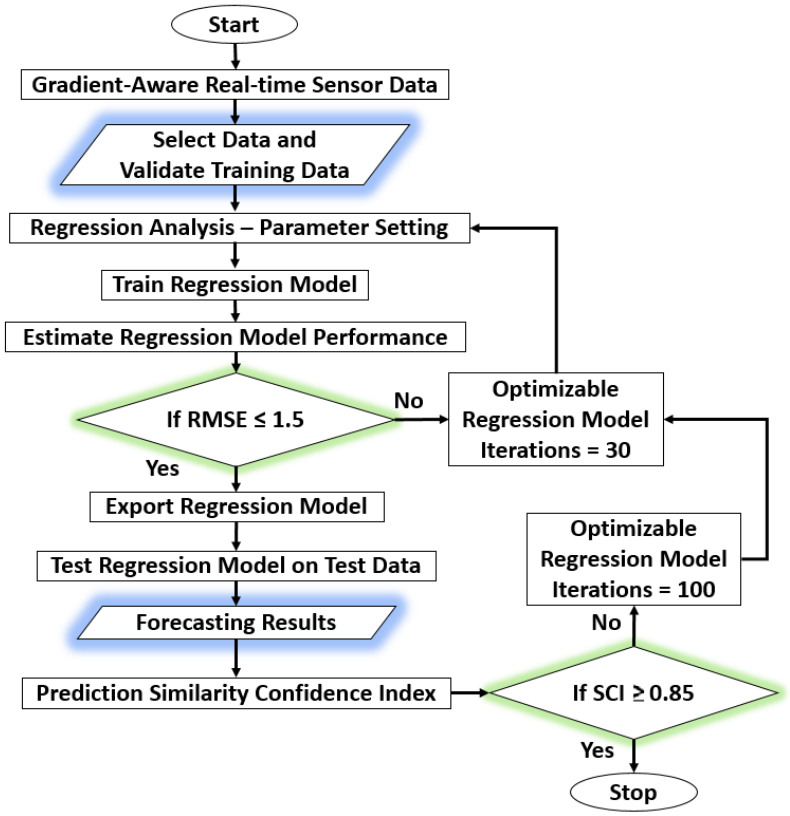



**Figure 2.** The Optimized Bi-Cluster Regression Algorithm

As per the proposed bi-cluster networked forecasting of the regression models $Y_{t\text{-}CO2}$ (AQI, CO2) and $Y_{t\text{-}Temperature}$ (AQI, Temperature), the regression models must have an acceptable similarity > 85%. If the forecasted time-series curves from two similar sensors installed at two different locations have curve-similarity concerning their AQI curves termed as similarity coefficient index (SCI) is less than 85%, the iterations will keep running automatically. For RMSE < 1.5, statistically the SCI < 0.85 conditions should be satisfied in real-time. The US EPA AQI standard for outdoor air quality is presented in figure 3 below:

| Breakpoints | | | | | | | AQI | Category |
|---|---|---|---|---|---|---|---|---|
| $O_3$ (ppm) 8-hour | $O_3$ (ppm) 8-hour[1] | $PM_{10}$ (µg/m³) | $PM_{2.5}$ (µg/m³) | CO (ppm) | $SO_2$ (ppm) | $NO_2$ (ppm) | | |
| 0-0.064 | – | 0-54 | 0-15.4 | 0-4.4 | 0-0.034 | ([2]) | 0-50 | Good |
| 0.065 0.084 | – | 55 154 | 15.5 40.4 | 4.5 9.4 | 0.035 0.144 | ([2]) | 51 100 | Moderate |
| 0.085-0.104 | 0.125-0.164 | 155-254 | 40.5-65.4 | 9.5-12.4 | 0.145-0.224 | ([2]) | 101-150 | Unhealthy for sensitive groups |
| 0.105-0.124 | 0.165-204 | 255-354 | 65.5-150.4 | 12.5-15.4 | 0.225-0.304 | ([2]) | 151-200 | Unhealthy |
| 0.125-0.374 (0.155-0.404)[4] | 0.205-0.404 | 355-424 | 150.5-250.4 | 15.5-30.4 | 0.305-0.604 | 0.65-1.24 | 201-300 | Very unhealthy |
| ([3]) | 0.405-504 | 425-504 | 250.5-350.4 | 30.5-40.4 | 0.605-0.804 | 1.25-1.64 | 301-400 | Hazardous |
| ([3]) | 0.505-0.604 | 505-604 | 350.5-500.4 | 40.5-50.4 | 0.805-1.004 | 1.65-2.04 | 401-500 | Hazardous |

**Figure 3.** The US EPA AQI Standard for Outdoor Air Quality

The AQI is generically estimated as:

$$I_P = [(I_{high} - I_{low})/(B_{P\text{-}high} - B_{P\text{-}low})] \times (C_P - B_{P-low}) + I_{low} \quad (7)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



Every pollutant was formulated using equation (7) and given by in equations (8 to 12).

$$I_{PM}= [(I_{high} - I_{low})/(B_{PM-high} - B_{PM-low})] \times (C_{PM} - B_{PM-low}) + I_{low} \quad (8)$$

$$I_{NO2}= [(I_{high} - I_{low})/(B_{NO2-high} - B_{NO2-low})] \times (C_{NO2} - B_{NO2-low}) + I_{low} \quad (9)$$

$$I_{SO2}= [(I_{high} - I_{low})/(B_{SO2-high} - B_{SO2-low})] \times (C_{SO2} - B_{SO2-low}) + I_{low} \quad (10)$$

$$I_{O3}= [(I_{high} - I_{low})/(B_{O3-high} - B_{O3-low})] \times (C_{O3} - B_{O3-low}) + I_{low} \quad (11)$$

$$I_{CO}= [(I_{high} - I_{low})/(B_{CO-high} - B_{CO-low})] \times (C_{CO} - B_{O3-low}) + I_{low} \quad (12)$$

From the AQI equations the resulting relative regression models for CO2 ($I_{m-CO2,t}$) and Temperature ($I_{m-T,t}$) are given as:

$$Y_{t\text{-}CO2} = \sum_{m-CO2=1}^{i} \beta_{m-CO2} I_{m-CO2,t} + \varepsilon_{CO2} = \beta_{1\text{-}CO2} I_{1\text{-}CO2,t} + \beta_{2\text{-}CO2} I_{2\text{-}CO2,t} + \cdots + \beta_{i\text{-}CO2} I_{i\text{-}CO2,t} + \varepsilon_{CO2} \quad (13)$$

$$Y_{t\text{-}T} = \sum_{m-T=1}^{i} \beta_{m-T} I_{m-T,t} + \varepsilon_{T} = \beta_{1\text{-}T} I_{1\text{-}T,t} + \beta_{2\text{-}T} I_{2\text{-}T,t} + \cdots + \beta_{i\text{-}T} I_{i\text{-}T,t} + \varepsilon_{T} \quad (14)$$

The function Iμ has been used to express the rate of change in AQI at the corresponding time derivative.

$$(I\mu)^n = \frac{\Delta AQI}{\Delta t} = \frac{2kre^2}{(1+e^{-rt})^2} \times t + \cdots = 0,1,2\ldots \quad (15)$$

To compare the relative influence level among the various influencing factors, the regression coefficients were normalized.

$$\beta'_m = \beta_m \times \frac{\sigma Xm}{\sigma Y} \quad (16)$$

where $\beta'_m$ is the normalized regression coefficient of the mth driving force, and $\beta_m$ is the regression coefficient of the driving force. $\sigma X_m$ is the standard deviation of the driving force, and $\sigma_Y$ is the standard deviation of the dependent variable. The RMSE will be the first step in ML model testing and optimization and given as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (M_i - RM_i)^2} \quad (17)$$

Indicates the magnitude of the error and retains the variable's unit; is sensitive to extreme values and outliers; tends to vary as a function of the standard deviation of the RM. Based on the RMSE the iteration will be performed by eliminating the temperature effect $E_T$ and CO2 $E_{CO2}$ effect respectively from equations (13) and (14) using:

$$E_T = \frac{n \sum_{i=1}^{n} Y_{i-T} C_i - \sum_{i=1}^{n} Y_{i-T} \sum_{i=1}^{n} C_i}{n \sum_{i=1}^{n} (Y_{i-T})^2 - (\sum_{i=1}^{n} Y_{i-T})^2} \quad (18)$$

$$E_{CO2} = \frac{n \sum_{i=1}^{n} Y_{i-CO2} C_i - \sum_{i=1}^{n} Y_{i-CO2} \sum_{i=1}^{n} C_i}{n \sum_{i=1}^{n} (Y_{i-CO2})^2 - (\sum_{i=1}^{n} Y_{i-CO2})^2} \quad (19)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



sensitivity coefficient C for the measurement influencing variable and finally:

$$RSS = \sum_{i=1}^{\infty} (\gamma_i - b_0 - b_1 x_i)^2 \quad (20)$$

RSS is measured as the sum of the square of residuals as the final step in the iterative optimization. In the results section (fig 9 to 14), an ambient role of magnitudes of two variables can be observed that RMSE and MAE are not enough to resolve the sensor data with different scales and orders of magnitude for this SCI was proposed. The SCI for CO2 ($SCI_{CO2}$) and temperature ($SCI_T$) will be the tacking gradient (real-time difference divided by their average) ratio of two cluster nodes 1 and 2 given as:

$$SCI_{CO2} = 2 \times \left| \frac{SCI_{CO2\text{-}1} - SCI_{CO2\text{-}2}}{SCI_{CO2\text{-}1} + SCI_{CO2\text{-}2}} \right| \quad (21)$$

$$SCI_T = 2 \times \left| \frac{SCI_{T\text{-}1} - SCI_{T\text{-}2}}{SCI_{T\text{-}1} + SCI_{T\text{-}2}} \right| \quad (22)$$

$$SCI = \beta'_m \times \left| \frac{(SCI_T \times E_{CO2}) + (SCI_T \times E_{CO2})}{2} \right| \quad (23)$$

The present probability of infection ($P_{Infection\text{-}Present}$) is based on present data and the future probability of infection ($P_{Infection\text{-}Future}$) is based on forecasted data. Based on previous research mentioning the COVID19 relationship with temperature and CO2 and Mask(ΔCO2) (cycling the CO2 into the lungs that gradually weakens the lungs) from equation (5) and relative influence level based on $\beta'_m$ (equation 16) the probability ($P_{Infection}$) of trans-respiratory pandemics and COVID19 is given by:

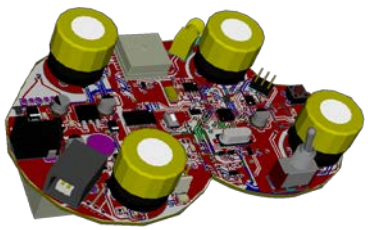
$$P_{Infection\text{-}Present} = \left| \frac{SCI}{RMSE \times I_\mu} \right| \quad (24)$$

$$P_{Infection\text{-}Future} = (P_{Infection\text{-}Present} \times \frac{1}{I_\mu}) + (Mask(\Delta CO2) \times \left| \frac{Y_{t\text{-}CO2}}{Y_{t\text{-}T}} \right|) \quad (25)$$

The proposed automated iterative optimization for COVID19 and other pandemics that are based on some sensing variables is independent of the personal immunity and the infection capability or the strength of germs being a medical science research area.

*2.3. Case Study: Urban Scale IoT-based AQI Monitoring System*

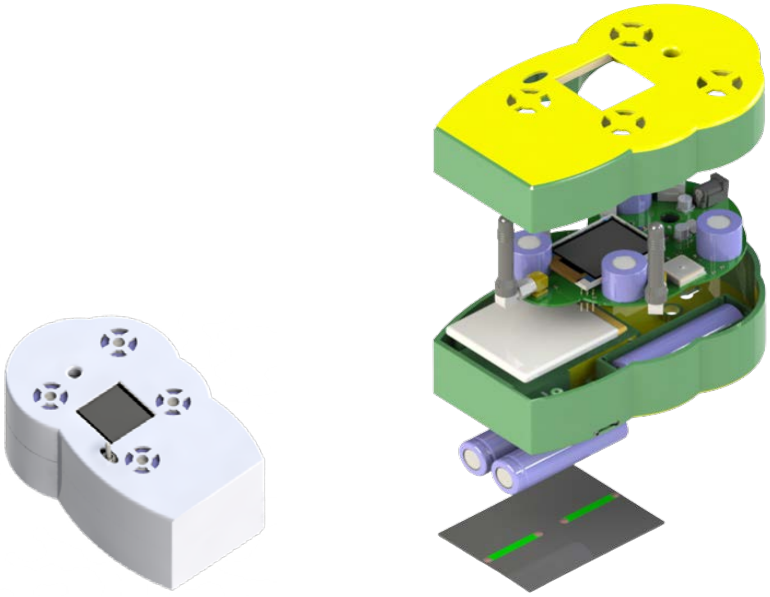
The proposed model and applied algorithm were tested and validated using our TRL7 autonomous AQI mapping system from past works [25, 26]. A 1-1 correspondence electronics and instrumentation system were designed in a single package, i.e. SeReNoV2 presented in figure 4 presented below:



(**a**)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1

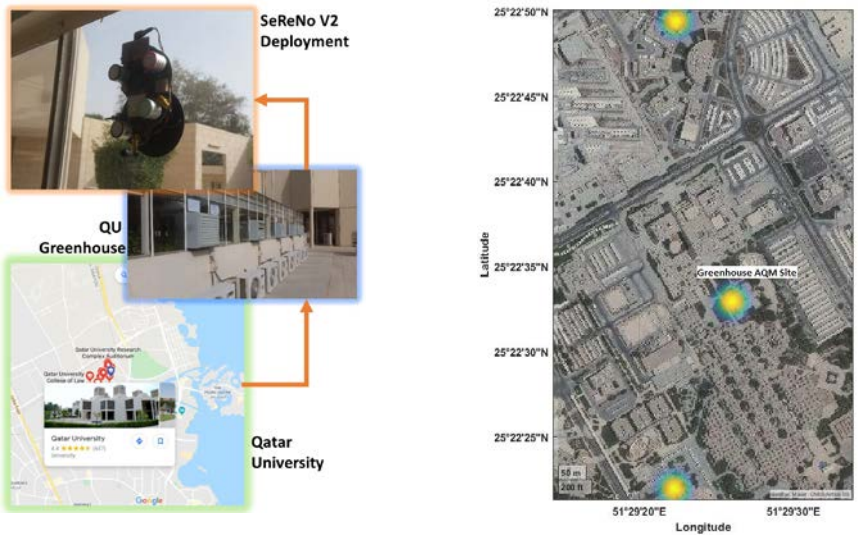




(**b**) (**c**)

**Figure 4.** Two 3D Layouts of SeReNo V2 AQM Node: (**a**) Top View; (**b**) Bottom View and (**c**) SeReNo V2 Complete Assembly

Three SeReNo V2 nodes were fabricated and deployed in QU for outdoor testing. The fabricated SeReNo V2 was deployed based on the efficient utilization of GAM, i.e. QU Greenhouse exhibited in the figure below.




(**a**) (**b**)

**Figure 5.** SeReNo V2 Deployment in QU to utilize the GAM-based OBRM: (**a**) The Greenhouse Site Details and (**b**) The Bi-cluster data-fusion at the central site.

The GAM reduced the bulk time-series curation operations needed for forecasting. The dual time-series data was queued to OBRM with (AQI, CO2) and (AQI, Temperature) vectors at the same time with $t_1$ and $t_2$ time-series. The iterative regression parameter setting was performed based on default parameters (RMSE, RSS, and MAE). On every cycle, these parameters were optimized AQI refers to a structured chart with a bio-tolerable threshold of specific pollutants and bio-hazardous gases recommended by EPA in the area

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



under a specified border agency18-24. The top 10 environmental protection agencies (EPAs) unanimously agreed on the standard of four core gases for outdoor.

## 3. Results and Discussion

After the long-haul deployment of six months, the data results obtained were displayed on the Ubidots IoT platform as shown in figure 6.

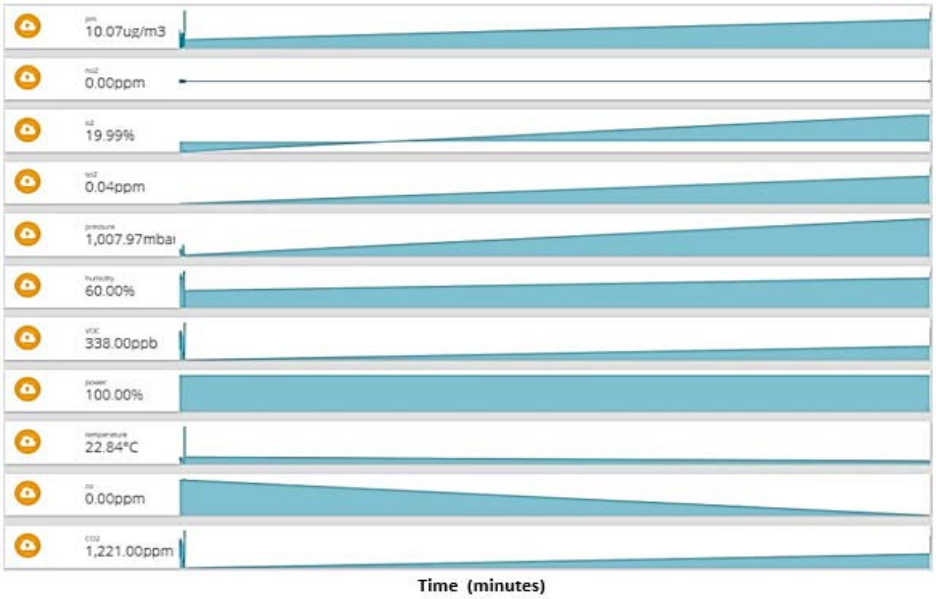
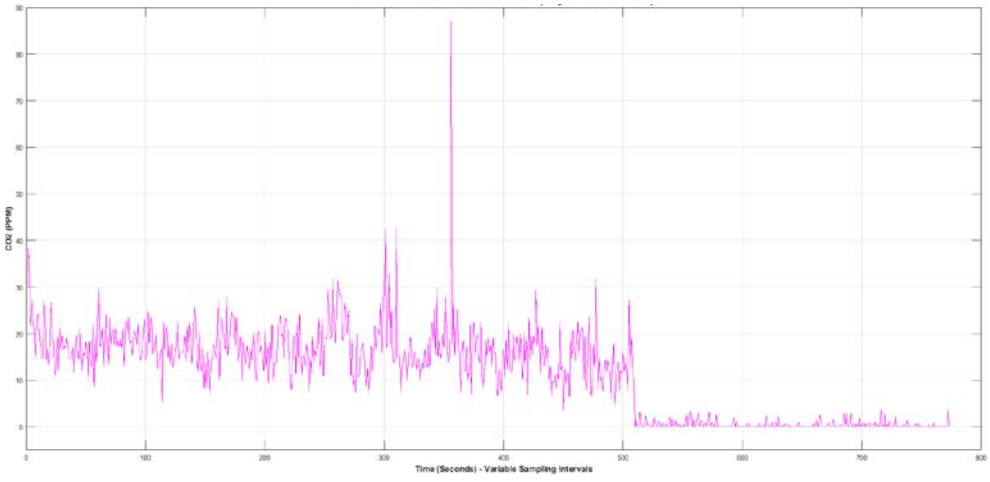



**Figure 6.** The SeReNo2 IoT Dashboard at Ubidots IoT

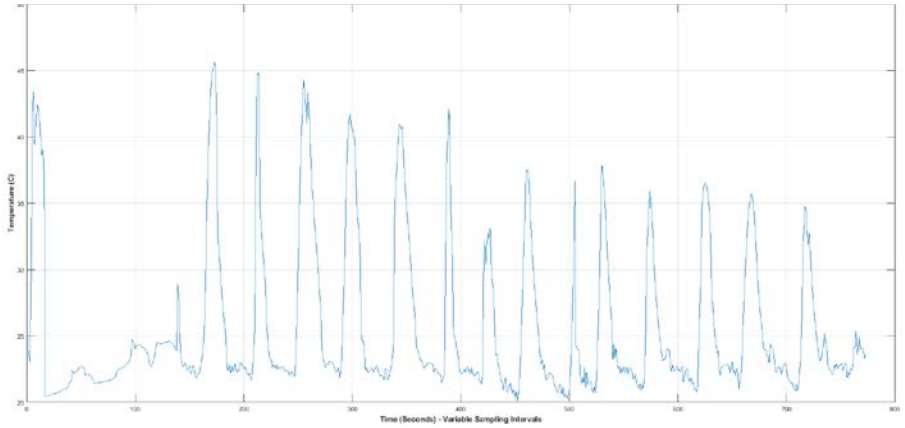
The Eleven real-time variables were exhibited in Figure 4 sending data through the GSM model QuecTel M10. The bi-cluster considered in this special data-fusion case for data collected at the central site, i.e. QU Greenhouse. The two variables CO2 and temperature were double interpolated from four sites (top: QU H10 and QU C05) and presented in the figure below from the Ubidots IoT platform.




(**a**)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1





(**b**)

**Figure 7.** The Bi-cluster formation from QU-H10 and QU-C05 data-captured was during 8 hours from CO2 and Temperature Variables: (**a**) The double-interpolated CO2 from QU-H10 and QU-C05 data-captured during 8 hours in ppm and (**b**) The double-interpolated Temperature from QU-H10 and QU-C05 data-captured during 8 hours in $^{o}$C.

Since the trans-respiratory diseases as per WHO and US EPA get more worst due to poor AQI, from (8) to (12) the Cumulative AQI of four sites is given in figure 8.

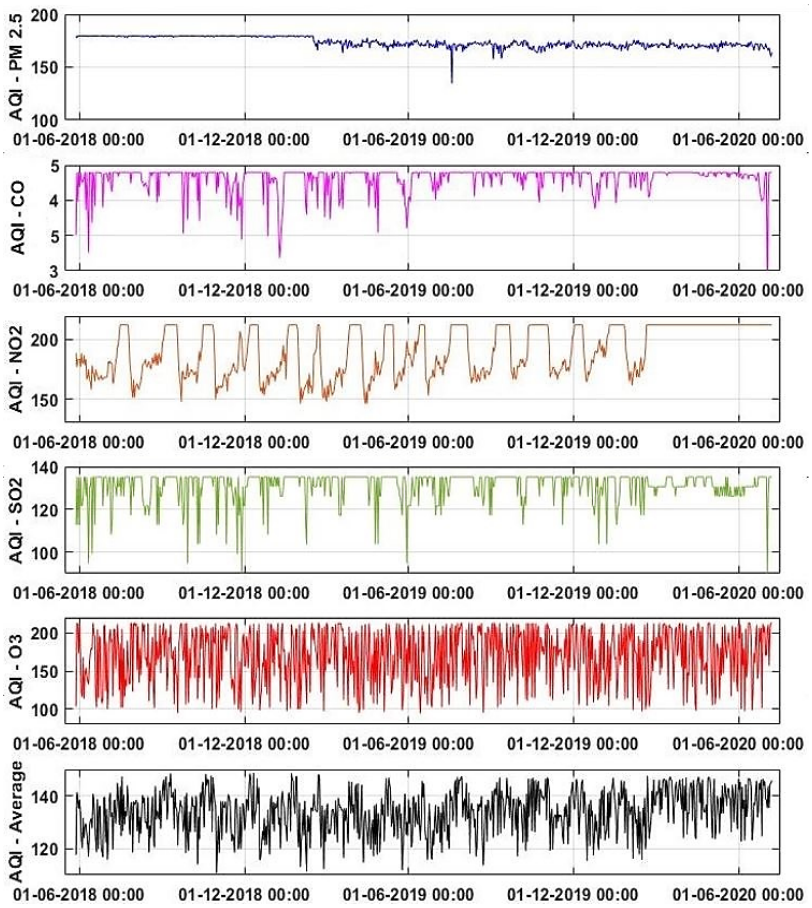



**Figure 8.** The Cumulative AQI of Four SeReNo V2 nodes using equations (8-12)

The application of GAM enabled only meaningful data to be sent to the cloud which made time-series more non-linear as only gradients impacted values were being transmitted. The accuracy of bi-clustered data measurements in terms of autonomous AQI system by applying our previous work is exhibited in figure 7. The following plots of individual

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



variables give more insights into GAM in SeReNoV2. The KPIs of GAM contributed to the accuracy and efficiency of OBRM.

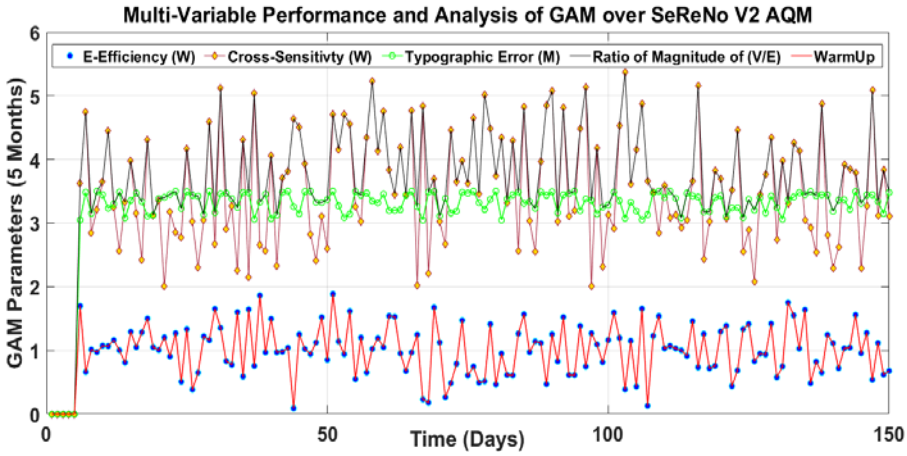
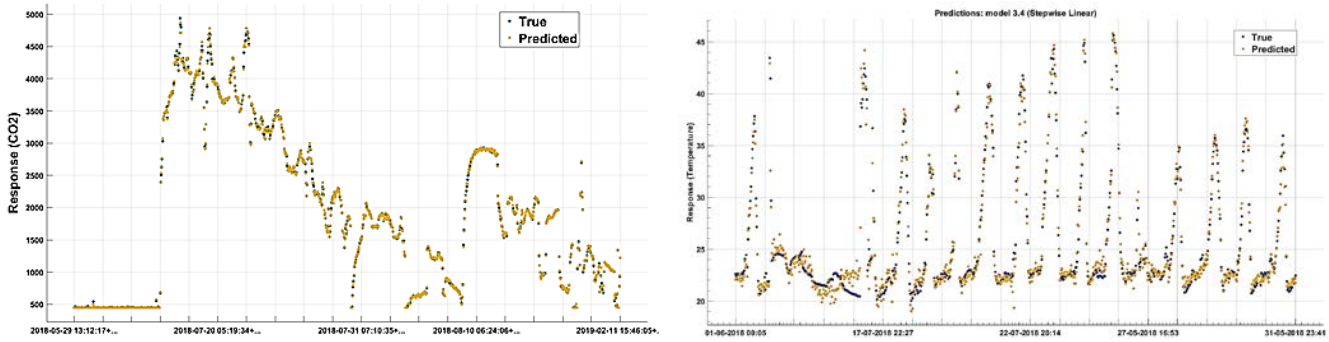



**Figure 9.** The GAM KPIs for SeReNo V2

The impact of GAM can be warm times below 1.83 seconds throughout 5 months. The reduced warm-up times reduced the boot time power spike that was reduced and resulted in stable or voltage above 3.3V needed for sensors. The typographic error observed around 3.1 to 3.4 that is also very less. A minute typographic error can be observed due to the correlation of the GPS and GPRS-assisted cell network locations scheme. The key performance indicators (KPIs) of GAM efficiency on SeReNoV2 were the major contribution that enabled all the outcomes presented in figures 8 to 18 as detailed in figure 18.

The dual-time series regression of OBRM is presented for CO2 being the top concern in Qatar. This outcome contributed to potential safety during the CoVID19 precautionary measures. A four-step procedure was followed for OBRM. First, the predicted response was assessed and ML KPIs mentioned in Table I was streamlined. Then the comparison was performed between real and predicted, at this step, the trained model residuals were estimated and finally, the optimization was performed as per conditions.



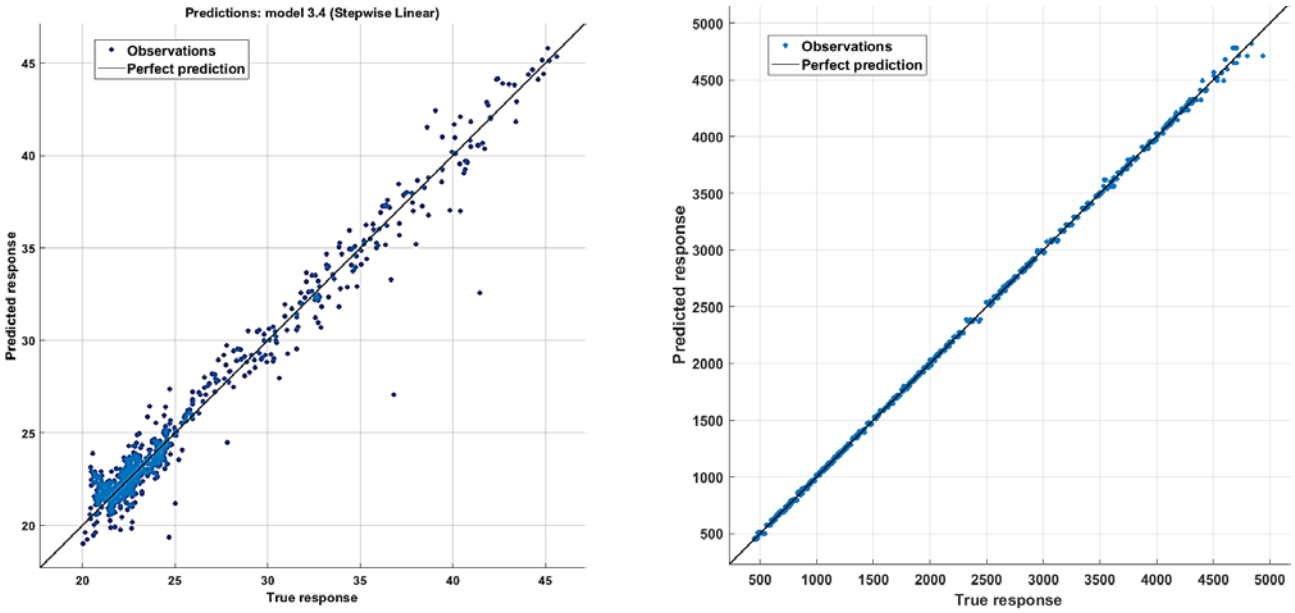

(**a**) (**b**)

**Figure 10.** The Bi-cluster formation from QU-H10 and QU-C05 data-captured during 8 hours from CO2 and Temperature Variables. (**a**) The OBRM1 Response for CO2 (ppm) and (**b**) The OBRM2 Response for Temperature (°C).

Figure 10 (a) exhibits the temperature response for model 1 termed OBRM1. The real data is in blue and the predicted is in orange. It was measured for one month. The RMSE of 1.0042 is almost ideal and needed no further tuning and verification. Figure 10 (b) is a realization of close prediction as the predicted and actual are almost overlapping with RMSE 1.7+.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



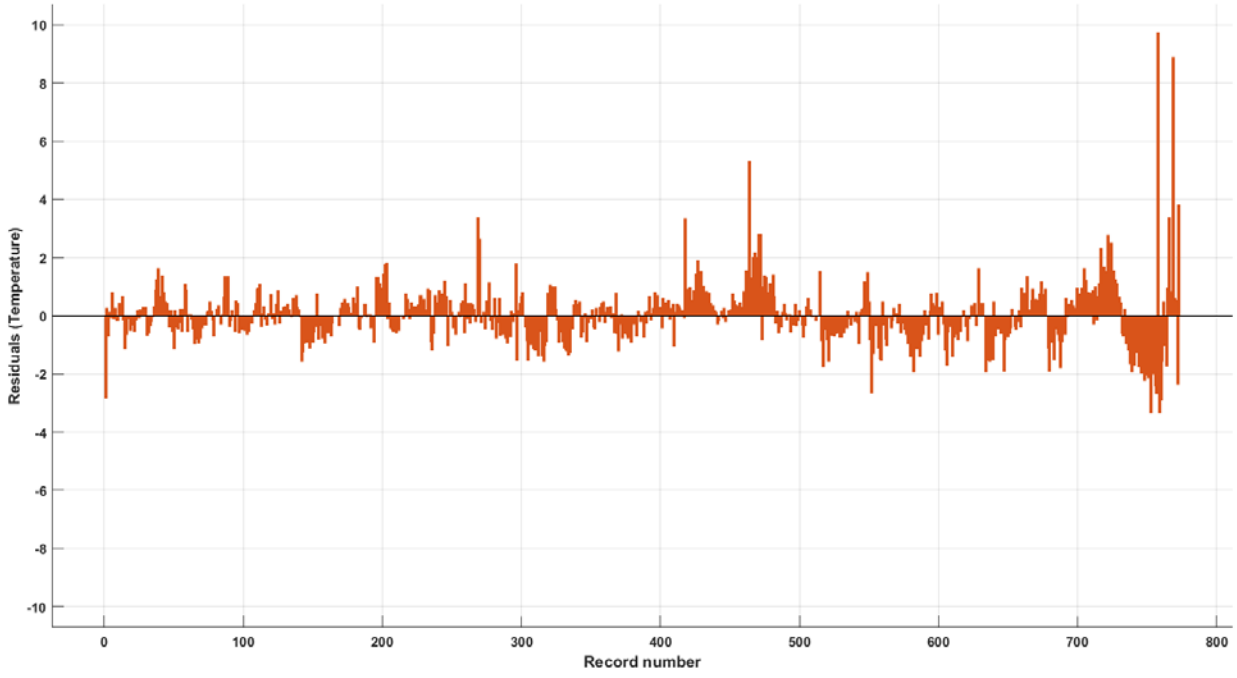



(**a**) (**b**)

**Figure 11.** SeReNo V2 Deployment in QU to utilize the GAM: (**a**) The Greenhouse Site Step-wise Linear prediction and (**b**) The Bi-cluster data-fusion at the central site.

In figure 11 (a), the wrapping of blue markers or bubbles over ideal or accurate prediction shows the accuracy of prediction by customized linear regression. Maximum similarity can be observed in magnitudes or 21 °C to 24.5 °C. In the next process, the residual was estimated as the vertical distance between a data point and the regression line. Each data point has one residual. They are positive if they are above the regression line and negative if they are below the regression line. If the regression line passes through the point, the residual at that point is zero.

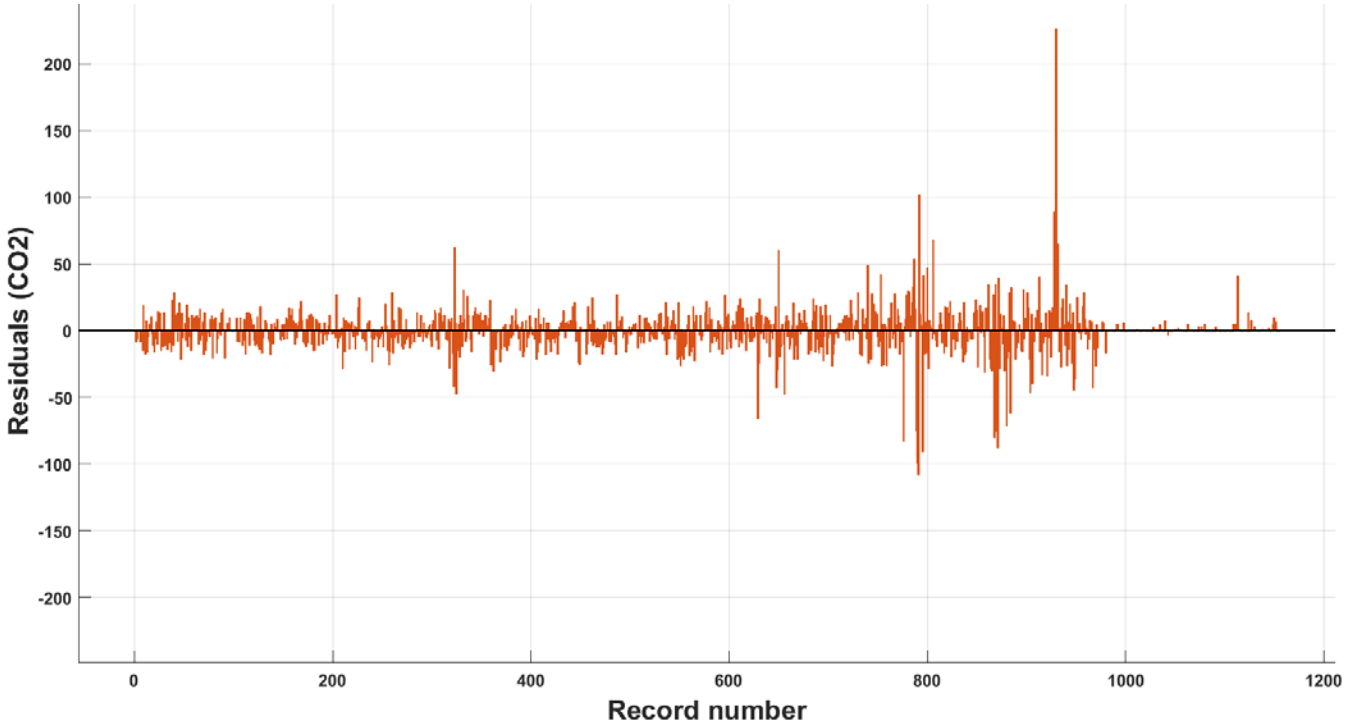
The RMSE 1.7+ is extremely small for magnitudes like 6000, thus the comparative plot for predicted and true is almost overlapping in figure 10 (b).




(**a**)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1






(**b**)

**Figure 12.** The Bi-cluster formation from QU-H10 and QU-C05 data-captured during 8 hours from CO2 and Temperature Variables: (**a**) The OBRM1 Response for Temperature (°C) and (**b**) The OBRM2 Response for CO2 (ppm).

In figure 12 (a), the magnitudes of 9+ for residuals are non-convex and impact the error in the prediction by OBRM1 for temperature. The $A_E(t_1)$ cluster was not optimized due to RMSE 1.0042. The optimization was performed for RMSE > 1.5 for $A_G(t_2)$ presented in figures 9-12.

The 200 residual magnitudes for amplitudes of PPM like 4500+ are minute, i.e. 200/4500 = 0.044 shown in figure 11 (b).

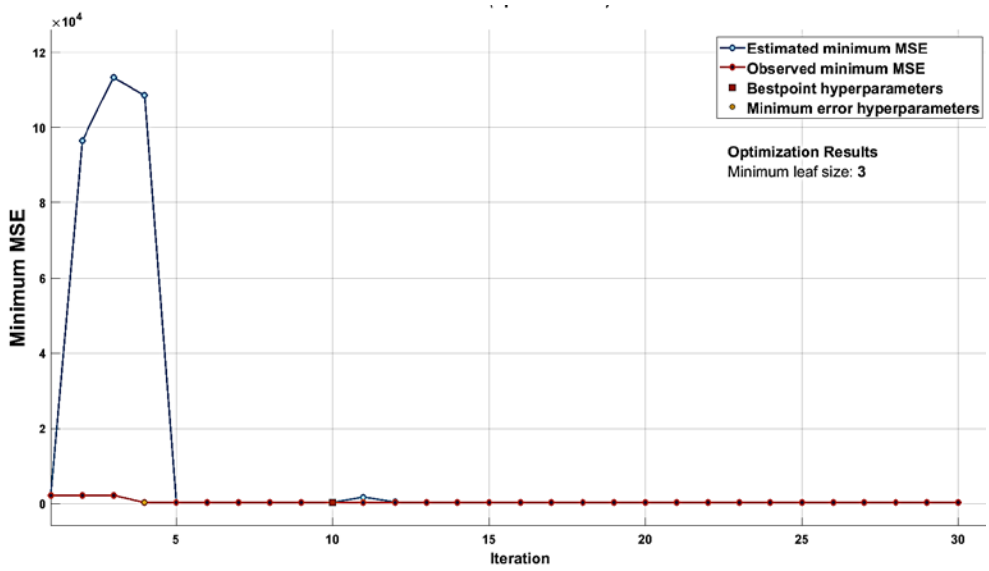
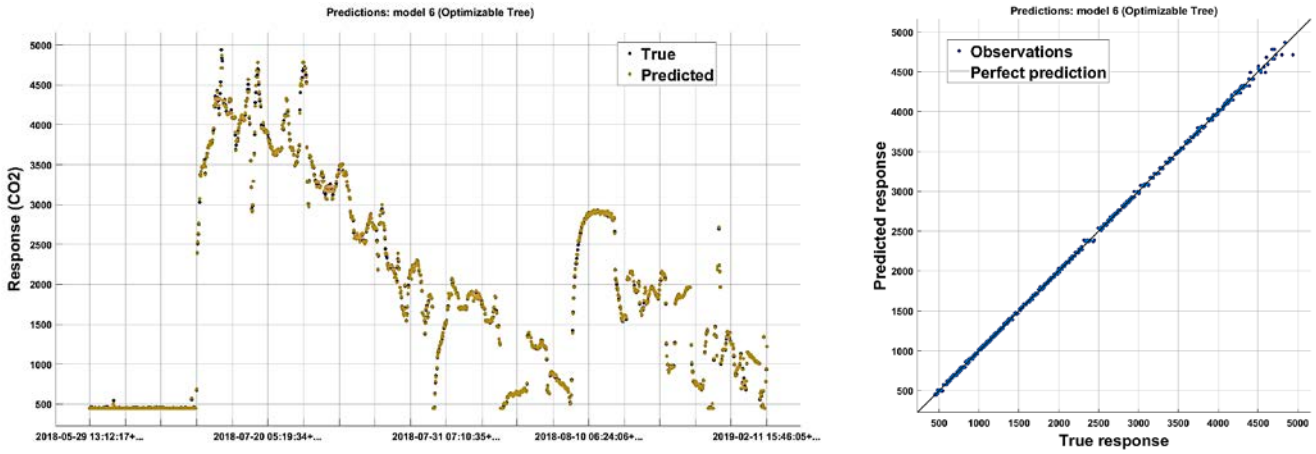



**Figure 13.** Cross-Validation using MSE of OBRM1 for CO2

The results in figure 13 lead to level 2 optimization of the OBRM1 based on the leaf size 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1






(**a**) (**b**)

**Figure 14.** SCI based Iterative Optimization of the OBRM2 for CO2: (**a**) Iterative Optimization of OBRM1 to OBRM2 and (**b**) OBRM3 Prediction Response

The tracking and alignment performed by OBRM3 for the observed and predicted CO2 (PPM) is up to 4400ppm in figure 14.

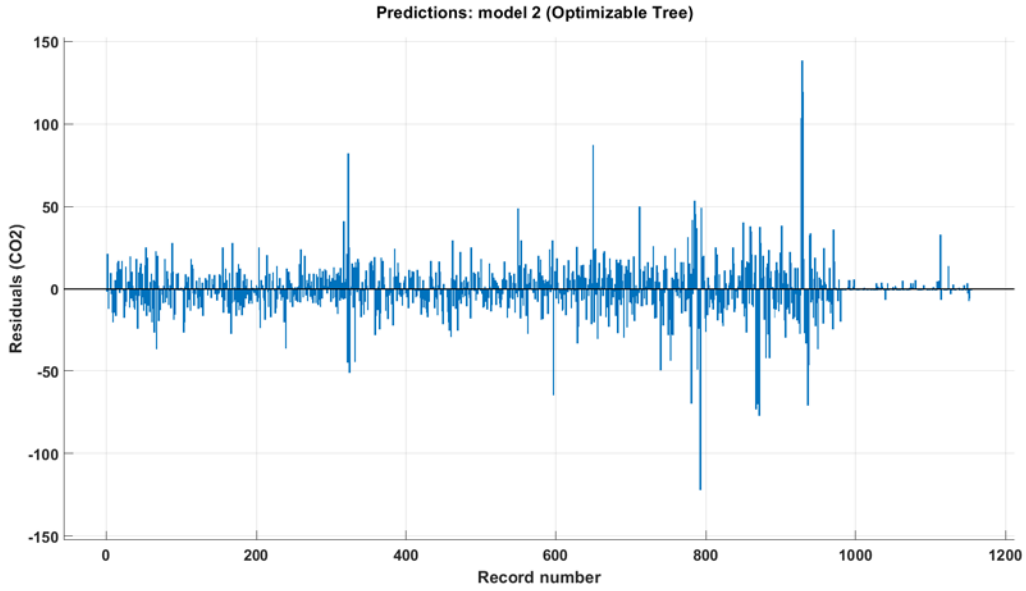



**Figure 15.** The Bi-cluster formation from QU-H10 and QU-C05 data-captured during 8 hours from CO2 and Temperature Variables. The OBRM3 Response for CO2 (ppm).

The offset or residual of 150/4800 = 0.03125ppm is almost perfect or accurate as examined in figure 14. The 200 residual magnitudes for amplitudes of PPM like 4500+ are minute, i.e. 200/4500 = 0.044 shown in figure 15 (b).

The leaf-size of 2 with 100 iterations delivered fine-tuned optimization and tracking for précised prediction observed in figure 16. Later the generated model was tested over test data for predicting the CO2 for the years 2021 and 2022 presented in figure 17.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



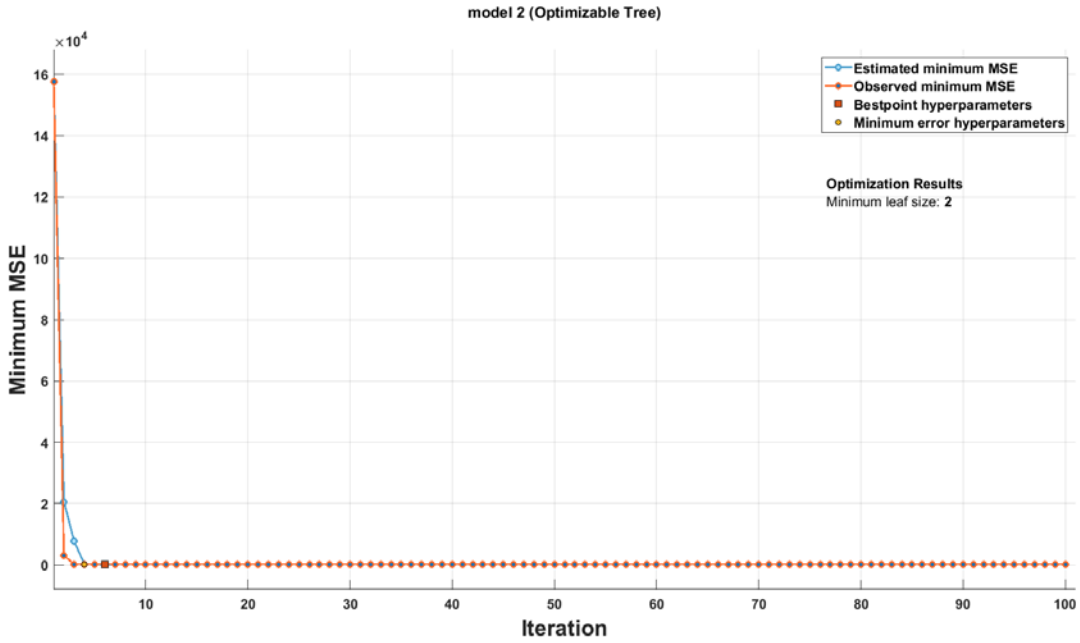
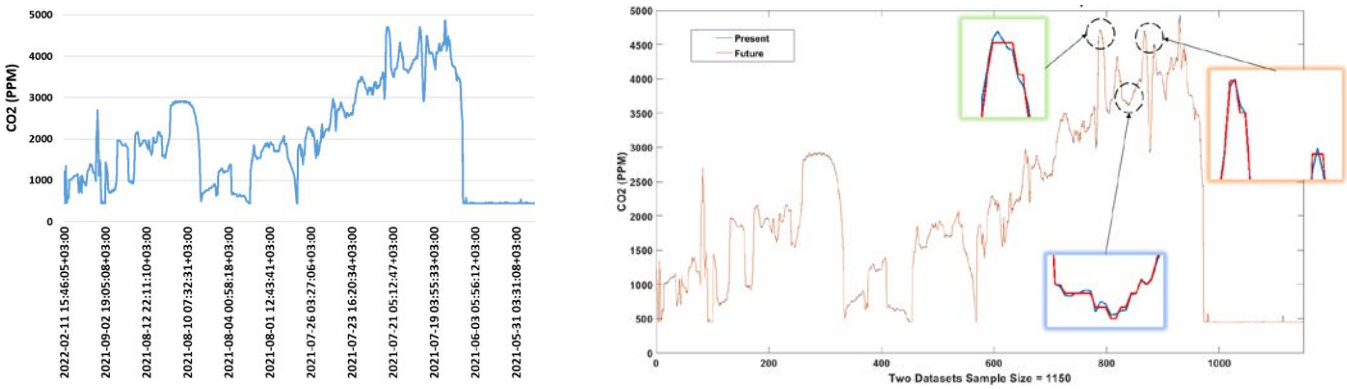



**Figure 16.** The Optimization of OBRM2 for CO2 to achieve OBRM3




**(a)** **(b)**

**Figure 17.** SeReNo V2 Deployment in QU to utilize the iterative OBRM3 optimization for $P_{Infection\text{-}Present}$ and $P_{Infection\text{-}Future}$: (**a**) The forecasted CO2 data by OBRM3 for the years 2021-22 and (**b**) The $P_{Infection\text{-}Present}$ and $P_{Infection\text{-}Future}$ from equations (24) & (25) for $Y_{t\text{-}CO2}$ and $Y_{t\text{-}T}$ for $I_{\mu}$ using SCI.

The $P_{Infection\text{-}Future}$ for CO2 by OBRM3 was almost similar, more ambient from figure 17 with numerical explanation and highlights for SCI computation. The OBRM3 had a very minute difference from the present and forecasted.

**Table 1.** Regression Parameter Setting

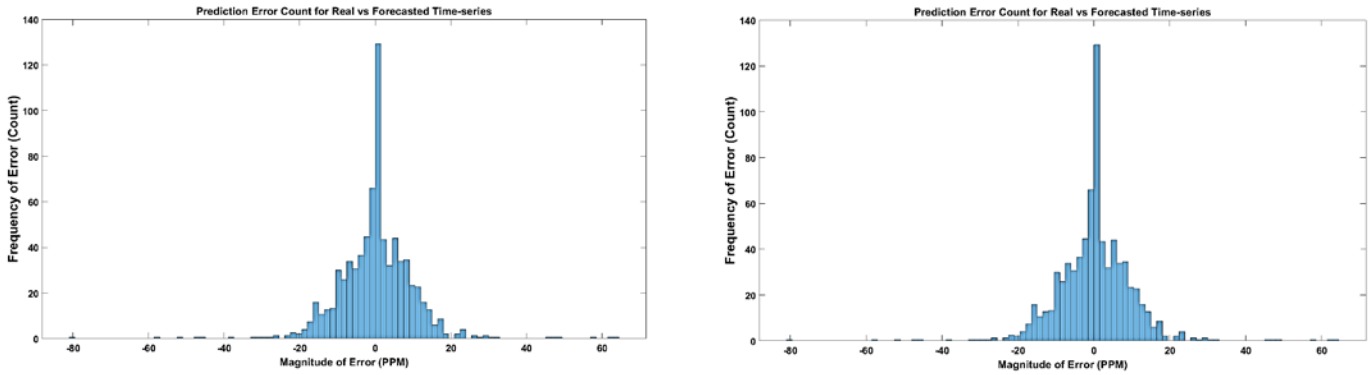
| | Optimized Bi-Cluster Regression MLT | |
|---|---|---|
| **Parameters** | **SWL (Temperature)** | **OBRM3 (CO2)** |
| Time Series Vector | [E($A_E$(T, P, H, VoC, PM),$t_1$)] | [G($A_G$(O3, NO2, SO2, CO), $t_2$)] |
| No. of Predictors | 11 | 11 |
| RMSE | 1.0042 | 1.646 |
| R-Squared | 0.97 | 1.0 |
| MSE | 1.0084 | 293.98 |
| MAE | 0.66226 | 10.252 |
| Prediction Speed | ~5100 obs.sec | ~45000 obs/sec |
| Training Time | 469.28 | 28.53 |
| Model Type | Step-wise Linear | Surrogate Split |
| Steps | 1000 | N/A |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2022 doi:10.20944/preprints202212.0534.v1



| | | |
|---|---|---|
| Iterations | N/A | 100 |
| Hyperparameter | N/A | LS (1~577) |

The parameter setting for optimized linear regression and optimized tree are presented in table 1. The training time and prediction speed are quite relatable being reciprocal of each other.




(**a**) (**b**)

**Figure 18.** Iterative Parametric Optimization of the Prediction Errors using Predictors Sets and Frequency of Prediction Errors: (**a**) The Prediction of Error in the count for OBRM3 and (**b**) The Probability of Prediction of Error in the count for OBRM3.

The probabilities of errors in the magnitude range set {0.06, 0.15} is 0 observed in figure 18.

## 4. Future Recommendation

This experimentation and study were based on the AQI sensing at different locations within the Qatar University. The 4 SeReNo V2 AQI sensors nodes were in two buildings with similar conditions and two buildings with different conditions. The dual installation was performed to avoid any measurement errors as per figure 9 based on previous studies (Hasan et al, 2020). Since the trans-respiratory pandemics, especially COVID19 can impact lives at place but are more impactful at gathering and populated premises so university was chosen and the equations and their respective figures provided a precise route-map of forecasting. Based on the equation (24) and (25) using SCI the countermeasures can be easily taken by raising the temperature to the un-survivable limit for COVID19 germs and using CO2 capturing units and O2 cylinders to push the fresh air into the premises.

This study will be more impactful if such AQI nodes are installed in hospitals and measured for COVID19 tested +ve and -ve patients. Our research group is looking forward to conducting this research in hospitals which was not possible during the pandemic times due to social isolation.

## 5. Conclusions

A novel similarity coefficient index-based forecasting method for COVID-19 and trans-respiratory pandemics was proposed using the SeReNoV2 nodes. A multi-time series-parallel automated iterative optimization of regression models was performed with interesting results. The presented work highlighted the practical time-series challenge of duality and multi-cluster vector forecasting for COVID19 safety with mask impact. To the best of our knowledge, this is the first real-time bi-cluster dual time-series forecasting machine learning approach for real-time multi-source sensor temporal data forecasting. The results can be summarized in three key milestones. The optimized regression methodology was able to 1) implement dual-time series analysis for non-linear composite time series vector compensating the commutative anomalies in bi-cluster sensor network; 2) the selected KPIs for data preprocessing by hardware resulted in reduced training time and

improved prediction speeds of machine learning model training; 3) the forecasted results were overlapping being a justified precision in forecasting methodology accuracy for COVID-19 infections. The proposed method can serve as a role model for dual time-series problems in COVID-19 and other complex pandemics.

**Author Contributions:** Conceptualization, H.T.; Data curation, H.T.; Formal analysis, H.T.; Funding acquisition, F.T., D.C. and A.B.M.; Investigation, H.T; Methodology, H.T.; Project administration, F.T.; Resources, F.T., D.C. and A.B.M.; Software, H.T.; Supervision, F.T.; Validation, H.T.; Visualization, H.T.; Writing—original draft, H.T.; Writing—review & editing, F.T., D.C. and A.B.M.

**Acknowledgments:** This publication was made possible by NPRP grant # 10-0102-170094 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Dorota, J. WHO Global Air Quality Guidelines 2021.
2. US EPA National Ambient Air Quality Standards (NAAQS), ***2020***.
3. Liu, Z.; Ciais, P.; Deng, Z.; et al. Near-real-time monitoring of global CO2 emissions reveals the effects of the COVID-19 pandemic. *Nature Communications*, ***2020***. Volume 11, Article number: 5172.
4. Peng, Z.; and Jose, L. Exhaled CO2 as a COVID-19 Infection Risk Proxy for Different Indoor Environments and Activities. *Environ Sci Technol Letter*, ***2021***.
5. Jiandong, C.; Chong, X.; Ming, G.; and Ding, L. Carbon peak and its mitigation implications for China in the post-pandemic era. *Scientific Reports*, ***2022***. Volume 12, Article number: 3473
6. Sofia, B., D.; Sofia, J.; Jose, H., D.; and Leontios, J. DeepLMS: a deep learning predictive model for supporting online learning in the Covid-19 era. *Scientific Reports*, ***2020***. Volume 10, Article number: 19888.
7. Zhou, Y.; Jinyan, Z.; & Shanying, H. Regression analysis and driving force model building of CO2 emissions in China. *Scientific Reports*, ***2020***. Volume 11, Article number: 6715.
8. Malik, A.; and Kumar, A. Spatio-temporal trend analysis of rainfall using parametric and non-parametric tests: case study in Uttarakhand, India. *Theoretical and Applied Climatology*, ***2020***. 140(1):183-207.
9. Abbasi, S., A. Monitoring analytical measurements in presence of two component measurement error. *Journal of Analytical Chemistry*, ***2014***.
10. Santiago, M.-C.; Eugenio, F.; and Antonio, M. Time Series Decomposition of the Daily Outdoor Air Temperature in Europe for Long-Term Energy Forecasting in the Context of Climate Change. *Energies*, ***2020***. Volume 13(7), 1569.
11. Stanislaus, S. U. Power Comparisons of Five Most Commonly Used Autocorrelation Tests. *Pakistan Journal of Statistics and Operation Research*. ***2020***. Vol.16 No. 1 2020 pp119-130.
12. Ian, F. A.; Weilien, S.; Yoegsh, S.; and Erdal, C. A Survey on Sensor Networks. *IEEE Communications Magazine*, ***2002***. Volume: 40 Issue: 8.
13. Tariq, H.; and Shafaq, S. Real-time Contactless Bio-Sensors and Systems for Smart Healthcare using IoT and E-Health Applications. *WSEAS Transactions on Biology and Biomedicine*, ***2022***. Volume 19, Pages 91-106.
14. Touati, F., et al. "IoT and IoE Prototype for Scalable Infrastructures, Architectures and Platforms", *International Robotics & Automation*, ***2018***. Vol 4, Issue 5.
15. Vinyals, M. V.; Juan, R.-A.; and Jesús, C. A Survey on Sensor Networks from a Multi-Agent Perspective. *The Computer Journal*, ***2014***.
16. Saberi-Movahed, F., et al. Decoding clinical biomarker space of COVID-19: Exploring matrix factorization-based feature selection methods. Computers in Biology and Medicine, ***2022***.
17. Tariq. H., et al, "Structural Health Monitoring and Installation Scheme deployment using Utility Computing Model", December 2018, EECS 2018, Bern, Switzerland.

18. Mehrpooya, A., et al. High dimensionality reduction by matrix factorization for systems pharmacology. Briefings in Bioinformatics, ***2022***. Volume 23, Issue 1.
19. Tariq, H.; Touati, F.; Al-Hitmi, E.; Crescini, D.; Manouer, A.B. Design and Implementation of Programmable Multi-parametric 4-Degrees of Freedom Seismic Waves Ground Motion Simulation IoT Platform. International Wireless Communications & Mobile Computing Conference, 2019.
20. Sadeghi, G., et al. A case study on copper-oxide nanofluid in a back pipe vacuum tube solar collector accompanied by data mining techniques, *Case Studies in Thermal Engineering*, ***2022***. Volume 32.
21. Najafzadeh, M.; Oliveto, G. More reliable predictions of clear-water scour depth at pile groups by robust artificial intelligence techniques while preserving physical consistency. *Soft Computing*, **2022**. Volume 25, 5723–5746.
22. Tariq, H.; Abdarazzak, A.; Farid, T.; Mohammed, A. E. A.; Damiano, C.; and Adel, B. M. An Autonomous Multi-Variable Outdoor Air Quality Mapping Wireless Sensors IoT Node for Qatar. International Wireless Communications and Mobile Computing.
23. Geiss, O. Effect of Wearing Face Masks on the Carbon Dioxide Concentration in the Breathing Zone. *COVID-19 Aerosol Drivers, Impacts and Mitigation (X)*, ***2020***.
24. Michelle, S. M.; Carin, D., L.; Matthew, T.; Amanda, C.; Jonathan, J. Y. Carbon dioxide increases with face masks but remains below short-term NIOSH limits. *BMC Infectious Diseases*, ***2021***.
25. Tariq, H.; Abdaoui, A.; Touati, F.; Al-Hitmi, E.; Crescini, D.; Manouer, A.B. Real-time Gradient-Aware Indigenous AQI Estimation IoT Platform. *Advances in Science, Technology and Engineering Systems Journal*, ***2020***. Vol. 5, No. 6, 1666-1673.
26. Tariq, H.; Abdaoui, A.; Touati, F.; Al-Hitmi, E.; Crescini, D.; Manouer, A.B. A Real-time Gradient Aware Multi-Variable Handheld Urban Scale Air Quality Mapping IoT System. IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS), 2019.