https://doi.org/10.31436/iiumej.v23i2.2164

FORECASTING OF INFECTION PREVALENCE OF *HELICOBACTER PYLORI* USING REGRESSION ANALYSIS

KOMILJON USAROV[1], ANVARJON AHMEDOV[1], MUSTAFA FATIH ABASIYANIK[2] AND KU MUHAMMAD NA'IM KU KHALIF[1]

[1]*Centre for Mathematical Sciences College of Computing & Applied Sciences of Computer Science Universiti Malaysia Pahang, 26300 Kuantan, Pahang, Malaysia.*
[2]*Pritzker School of Molecular Engineering, The University of Chicago, Edward H. Levi Hall, 5801 South Ellis Avenue Chicago, Illinois 60637, USA.*

[*]*Corresponding author: ukomiljon@gmail.com*

(Received: 27th July 2021; Accepted: 16th November 2021; Published on-line: 4th July 2022)

ABSTRACT: Global warming may have a significant impact on human health because of the growth of the population of harmful bacteria such as *Helicobacter pylori* infection. It is crucial to predict the prevalence of a pathogen in a society in a faster and more cost-effective way in order to manage caused disease. In this research, we have done predictive analysis of *H. pylori* infection spread behavior with respect to weather parameters (e.g., humidity, dew point, temperature, pressure, and wind speed) of Istanbul based on a database from Istanbul Samatya Hospital. We developed a forecasting model to predict *H. pylori* infection prevalence. The goal is to develop a machine learning model to predict *H. pylori* (Hp) related infection diseases (e.g., gastric ulcer diseases, gastritis) based on climate variables. The dataset for this study covered years from 1999 to 2003 and contained a total of 7014 rows from the Samatya Hospital in Istanbul. The weather information related to those years and location, including humidity (H), dew point (D), temperature (T), pressure (P) and wind speed (W), were collected from the following website: https://www.wunderground.com. In this paper we analyzed the forecasting model, which was used to predict *H. pylori* infection prevalence, by non-linear multivariate linear regression model (MLRM). We applied the non-linear least square method of minimization for the sum of squares to find optimal parameters of MLRM. Multiple Regression Method was used to determine the correlation between a criterion variable and a combination of predictor variables. It was established that the Hp infection disease is most influenced by humidity. Hp prevalence is modelled using the Multiple Regression Method equation, the average H, D, T, P, and W were the most important parameters to deviation of the datasets (testing dataset was 17% and 18% for training dataset). This showed that the statistical model predicts the Hp prevalence with about 83% accuracy of the testing data set (11 months) and 87% accuracy of the training data set (42 months). Based on the proposed model, monthly infection can be predicted early for medical services to take preventative measures and for government to prepare against the bacteria. In addition, drug producers can adjust their drug production rates based on forecasting results.

ABSTRAK: Pemanasan global mungkin mempunyai kesan langsung terhadap kesihatan manusia kerana pertambahan populasi bakteria merbahaya seperti infeksi *H. pylori*. Adalah penting bagi mengesan kehadiran patogen dalam masyarakat bagi mengawal penularan penyakit dengan cepat, dan melalui kaedah kurang mahal. Kajian ini berkaitan analisis ramalan penularan infeksi *H. pylori* secara langsung terhadap parameter cuaca (cth: kelembapan, titik embun, suhu, tekanan, kelajuan angin) di Istanbul berdasarkan data dari Hospital Samatya Istanbul. Kajian ini membentuk model ramalan bagi menjangka

penyebaran infeksi *H. pylori*. Matlamat adalah bagi mencipta model pembelajaran mesin bagi mengjangka penyakit berkaitan infeksi *H. pylori* (Hp) (cth: penyakit ulser gastrik, gastrik) berdasarkan pembolehubah cuaca. Dari tahun 1999 ke 2003, set data telah digunakan bagi mempelajari di mana sejumlah 7014 baris dari Hospital Samatya di Istanbul. Informasi berkaitan tahun-tahun tersebut dan lokasi mengenai kelembapan (H), titik embun (D), suhu (T), tekanan (P) dan kelajuan angin (W) dikumpul dari laman sesawang https://www.wunderground.com. Kajian ini mengguna pakai model ramalan bagi meramal kelaziman infeksi *H. pylori*, melalui model regresi berkadaran multivariat tidak-berkadaran (MLRM). Kaedah Kuasa Dua Terkecil tidak linear digunakan bagi pengurangan jumlah ganda dua bagi mencapai parameter optimum MLRM. Kaedah Regresi Gandaan digunakan bagi mencari persamaan antara kriteria pembolehubah dan gabungan pembolehubah ramalan. Dapatan menunjukkan infeksi penyakit Hp adalah disebabkan oleh faktor kelembapan. Penyebaran Hp dimodel menggunakan persamaan Kaedah Regresi Gandaan, purata H, D, T, P dan W adalah parameter terpenting bagi sisihan data latihan iaitu sebanyak 17% dan 18% bagi set data latihan. Ini menunjukkan model statistik menjangkakan penyebaran Hp adalah sebanyak 83% adalah tepat pada set data yang diuji (selama 11 bulan) dan 87% tepat pada set data latihan (selama 42 bulan). Berdasarkan model yang dicadangkan ini, infeksi bulanan dapat di jangka lebih awal bagi membendung servis kepada perubatan dan kerajaan bersiap-sedia memerangi bakteria ini. Tambahan, prosedur jumlah ubatan dapat dihasilkan lebih atau kurang daripada jumlah ubatan berdasarkan dapatan ramalan.

KEY WORDS: *H. pylori; infectious disease prediction; multivariate linear regression*

1. INTRODUCTION

Helicobacter pylori is highly prevalent in approximately 50% of the world's population [1], causes inflammation in the stomach and leads to chronic gastritis, peptic ulcer diseases (PUD), gastric ulcers (GU), duodenal ulcers (DU), and eventually gastric cancer in the human stomach [1,2,3]. In the United States, about 10% of the population will develop a duodenal ulcer at some point in their lives. Peptic ulcer disease affects about 4 million people annually in the world [4]. The occurrence of peptic ulcer disease is similar in men and women. Approximately 11%-14% of men and 8%-11% of women will develop peptic ulcer disease in their lifetime. The mortality rate for peptic ulcer disease is approximately one death per 10,000 cases. The mortality rate due to ulcer hemorrhage is approximately 5%. According to GLOBOCAN 2018 data, stomach cancer is the 3rd most deadly cancer with an estimated 783,000 deaths in 2018 [5]. It causes high cost to society and even brings on high risk to human lives.

The prediction of *H. pylori* prevalence has essential impact on the minimizing of the transmission of *H. pylori* related infectious diseases, which is a core function of public health law. Laws can be affected that prevent prevalence of the infection, but it is crucial to know early about the infection prevalence using forecasting models. The literature consists of some interesting research work related to forecasting models for infections such as malaria, scarlet fever, chickenpox [5] combining Big Data and Neural Networks. Moreover, there are some articles about predictions based on environmental factors which have a great impact on the prevalence of the infections. For instance, Song et al. built a time series model based on eight climate variables to predict hand, foot, and mouth disease [6]. In addition, Lu et al. showed that average daily sunshine time correlated positively with *H. pylori* infection [1]. Previous studies showed that using climate variables can be more accurate and efficient to predict infection prevalence.

Since there is no prediction model for *H. pylori* disease prevalence, we set a goal to design forecasting model for *H. pylori* prevalence based on the following environmental factors: humidity, dew point, temperature, and wind speed by using outstanding machine learning tools such as multivariate linear regression model (MVLR). This model will enable market players (e.g., doctors, government, pharmaceutical firms, etc.) to take sufficient precautions before outbreaks.

Ultimately, by building the forecasting model, we have proven that it is possible to predict the *H. pylori* infection prevalence and to know early information about the spread of the infection. It gives a chance to act for prevention procedures against the infection, which leads not only to reducing the prevalence of the infection, but it also minimizes social costs for the public and saves many people's lives. Furthermore, it can increase hospital services for patients and drug producers can develop drugs based on the demand of the patients.

2. MATERIALS AND METHODS

2.1 Research Data

From the original dataset from 7014 patients, only non-null values of CLO attribute were selected, leaving 4388 patients between 1999 to 2003 in the Samatya hospital in Istanbul, Turkey, which includes 48 attributes such as visiting date, gastritis cancer, DU, GU, gastritis, abdominal pain, stomachache, and CLO results. The ages were divided into below 20, 20-30, 30-40, 40-50, 50-60, and above 60 years old, representing 2%, 13%, 19%, 24%, 21% and 30% of the total dataset, respectively. More than half of patients were above 50 years old. Cases were 16% of DU, 18% of deformative pylorus, 19% of (peptic ulcer) (PU), 27% of deformative bulbus, 41% of erosive duodenitis, 46% of feel pain, 58% Hp infected, 22% stomachache, 93% of pangastritis and 99% of gastritis of the patients. In addition, there were 51% male and 49% female patients. The bacterial infection of each patient was detected by a special test called CLO and patients with a positive CLO test were assumed to be infected (Table 1).

Table 1: Baseline and outcome clinical characteristics of *H. pylori* patients

Parameters	4388 of patients (%) percentage from total patients
Female	2247 (51%)
Male	2141 (49%)
Pain	2015 (46%)
Hp	2547 (58%)
Pangastritis	4087 (93%)
Erosive duodenitis	1811 (41%)
Gastritis	4347 (99%)

Weather data (WD), (Humidity - the concentration of water vapor present in the air, dew point - the temperature to which air must be cooled to become saturated with water vapor, temperature - a degree of heat or cold the can be measured using a thermometer in degrees on the Fahrenheit, Celsius, and Kelvin scales, pressure or air pressure - the force per unit of area exerted on the Earth's surface by the weight of the air above the surface, wind speed or wind flow speed - a fundamental atmospheric quantity caused by air moving

from high to low pressure, usually due to changes in temperature) including humidity (%), dew point (°F) (https://en.wikipedia.org/wiki/Dew_point), temperature (°F), pressure (Hg), wind speed (mph) was obtained from historical data by average daily information in the https://www.wunderground.com/history website, joined with the visitor date attribute. The joined data was transformed into a monthly dataset using sum of Hp and mean of humidity, dew point, temperature, pressure, wind speed aggregate functions. The final data contains V (number of visitors), Hp (sum of Hp), H (mean of humidity), D (mean of dew point), T (mean of temperature), P (mean of pressure), W (mean of wind speed) attributes and 53 months of observed rows (see Table 2). All of this was performed using Google Colab Notebook (https://colab.research.google.com/) and Python3 (https://www.python.org/) machine learning libraries (https://scipy.org/) on Google's Cloud TPU Server.

The dataset above was divided into a train subset and a test subset with a ratio of 80% and 20%, respectively. It means that all 53 months of rows were split into 42 months of rows of training data subset and 11 months of rows of testing data subset.

2.2 Method

The following multivariate linear regression model was used to forecast Hp based on Table 2 data:

$$y = f(H, D, T, W) = \beta_1 sin^2(\rho_1 H + \rho_2) sin^2(\rho_3 T + \rho_4) + \beta_2 sin^2(\rho_5 H + \rho_6) sin^2(\rho_7 T + \rho_8) + \beta_3 sin^4(\rho_9 D + \rho_{10}) sin^4(\rho_{11} W + \rho_{12}) + \beta_4 sin^4(\rho_{13} D + \rho_{14}) sin^4(\rho_{15} W + \rho_{16}) + \beta_5 sin(\rho_{17} H\, D + \rho_{18}) + \beta_6 \quad (1)$$

Where:

y - dependent variable (Hp – the number patients who had positive CLO infection test),
H - the average of humidity (%)
D - the average of dew point (°F)
T - the average of temperature (°F)
W - the average of wind speed (MPH)
ρ_{1-18} - non-linear regression coefficients,
β_{1-5} - regression coefficients,
β_6 - constant.

2.3 The Algorithms

In order to determine which $\rho_{1-18}$and β_{1-6} vector paraments give the best fit to the data, the sum of squares of the residuals is minimized. The residuals are defined for each observed data-point as

$$\varepsilon_i = y_i - f(H_i, D_i, T_i, W_i) \quad (2)$$

Where y_i is the number of the total infected per month by *H. pylori* in the given region. We perform the leastsq command (nonlinear least square solver) in SciPy in python from scipy.org.

2.4 MVLR Assumptions

In order to achieve validity of the tests of hypothesis (like t-test, F-test) and to enhance that OLS estimators are the Best Linear Unbiased Estimator (BLUE), it needs to follow four base assumptions:

1. The relationship between the dependent variable and the independent variables is linear.
2. The residuals are independent.
3. Homoscedasticity.
4. Normality of residuals with mean equals to zero.

The Durbin-Watson statistic (DW) was used to check that residuals are independent. If DW is between 1.65 and 2.35, there is no autocorrelation. If DW is between 1.21 and 1.65 or between 2.35 and 2.79, the test is inconclusive [7]. Homoscedasticity is a word used for the "constant variance" assumption. The regression model assumes that the residuals have the same variance throughout. When this assumption is violated, the problem is called "heteroscedasticity," or changing variance. We used the Breusch – Pegan and White test to check it. Errors need to be a normal probability distribution. This makes no difference to the estimates of the coefficients, or the ability of the model to forecast. But it does affect the F- and t-tests and confidence intervals. We used more testing algorithms such as Jarque–Bera Test (JB), Shapiro-Wilk Test, D'Agostino's K-squared Test, Anderson-Darling Test because it is a very important assumption to rate the model.

3. RESULTS AND DISCUSSION

3.1 Statistical Analysis

In Table 2, the four-year hospital statistics show the average monthly infection rate was 58% among average monthly visitors (78 ± 44) of 45.1 patients among average 78.4 visitors per month in four years. It is shown that more than half of visitors were infected by *H. pylori* infection.

Table 2: Statistics of monthly data

	count	mean	std	min	25%	50%	75%	max
Total Visitors	53	78.4	44	1	49	81	100	190
Total Hp	53	45.1	25	0	26	47	60	113
Humidity (°F)	53	72	8.7	54	68	72	78	92
Wind Speed (mph)	53	9.8	2.2	3.5	8.7	10	11	17
Dew Point (°F)	53	50.6	12	22	44	50	60	73
Temperature (°F)	53	61.2	13	36	52	61	72	82

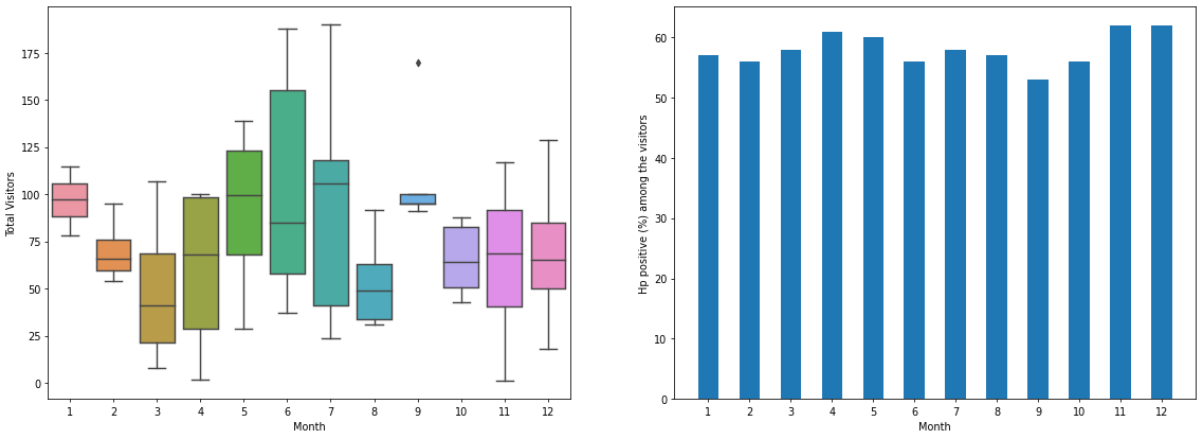
n=53, which is the total number of months in the study

Correlations of attributes for both data sets can be seen in Table 3. H attribute negative correlates to total Hp. D and T were correlated positive with 0.10 and 0.15 respectively.

Table 3: Correlations between given variables for monthly transformed data

	Total Hp	H	W	D	T
Total Hp	1.00	-0.25	-0.02	0.10	0.15
H	-0.25	1.00	0.14	-0.30	-0.53
W	-0.02	0.14	1.00	-0.14	-0.12
D	0.10	-0.30	-0.14	1.00	0.96
T	0.15	-0.53	-0.12	0.96	1.00

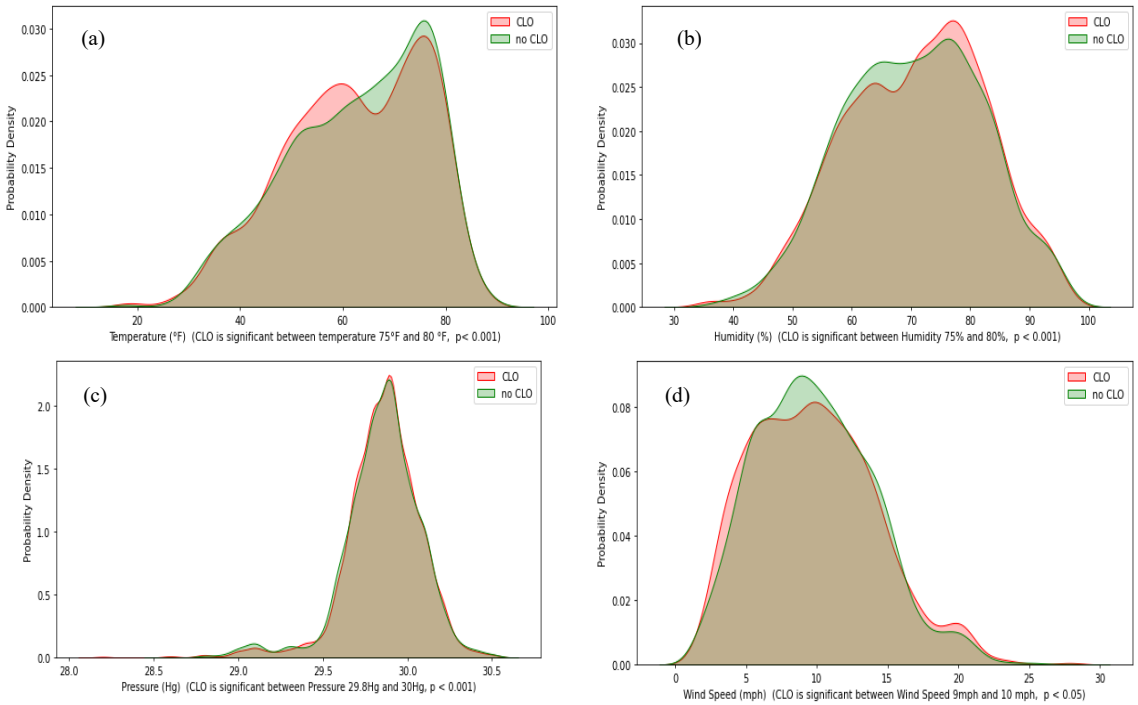
The box plot graph shows that for 5 years the number of visitors was close each month of year in January, February and September. However, in June and July there was very high difference between the number of visitors for each year. While the infected number of patients was very close in January, February, and August, there was a wide spread of numbers and the highest number of the infected patients with Hp in April, June, and July (Fig. 1).



(a) (b)

Fig. 1: The box plot for total visitors (a) and the Hp infected patients in percentage among visitors (b) are demonstrated by months.

We found various behaviors of the independent variables, such as the number of visitors and the number of Hp infected patients, in different climate conditions. According to the given dataset, the behavior of the independent variables was maximized, when weather temperature was 75-80 ºF, the humidity was 75%-80%, the pressure was between 29.8 mmHg and 30 mmHg, the wind speed was between 9 MPH and 10 MPH, and the dew point was between 45 ºF and 50 ºF, 60 ºF and 65 ºF. All of those factors were significant since null hypothesis were rejected ($p<0.05$) (Fig. 2).



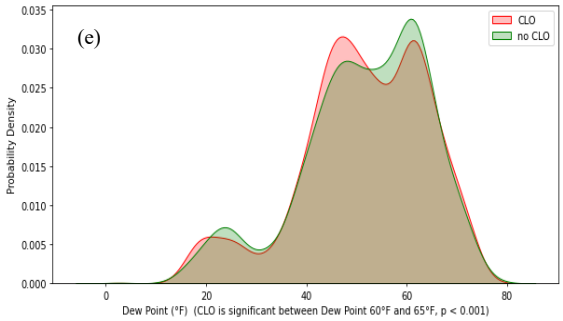


Fig. 2: For monthly average data, Hp positive and negative factors are represented by forecast weather variables: temperature (a), humidity (b), pressure (c), wind speed (d), dew point (e).

In addition, the behavior of the independent variables was highly impacted for the 45-50 years old patients (p-value<0.05) (Fig. 3). Moreover, the number of CLO infected patients grew until 50 years old, and it began to drop after that.

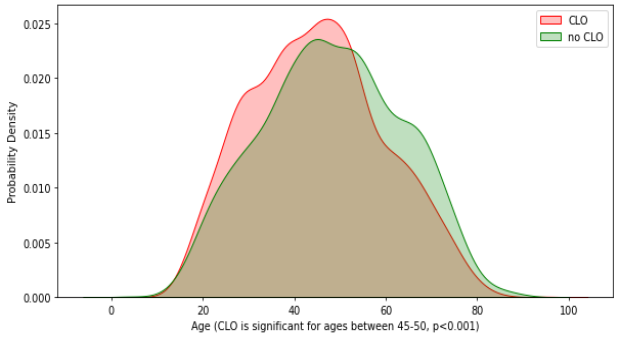


Fig. 3: The positive and negative CLO factors are represented by age of the patients.

Here we studied monthly and yearly statistics for NV and NC. NV and NC were more in June and September (p-value<0.05 each) than other months. Also, it is significant for March and August where there were p-values < 0.05 (Fig. 4a). There was strong growth of the number of visitors and CLO patients between 1999 and 2002 and it reached a peak in 2002 (it is significant for 2002, p-value<0.05) and dropped significantly in 2003. Note that there were no records for last two months in 2003 (November and December of data in 2003 (Fig. 4b).

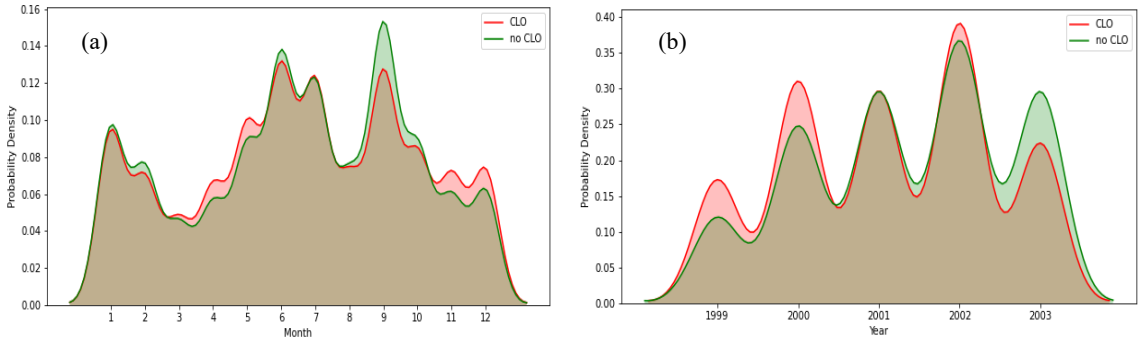


Fig. 4: The positive and negative CLO factors are represented by month (a) and year (b).

3.2 Model

Using Eq. (2) we obtained optimal parameters of Eq. (1) the proposed model described for Hp infection prevalence by the below mathematical formula:

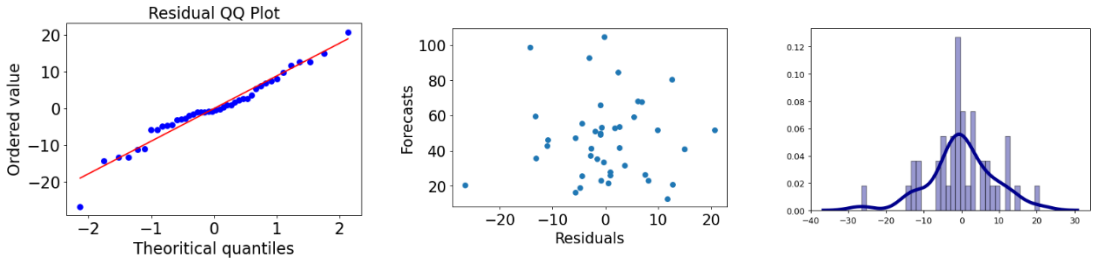
$$y = f(H, D, T, W) = 31.1237sin^2(3.7370H - 121.6778)sin^2(2.9952T+112.3183) +$$
$$68.2460sin^2(7.1261\ H + 0.2939)sin^2(8.8473\ T+17.8187) +$$
$$72.6059sin^4(12.4216\ D\text{-}15.3039)sin^4(12.4320\ W+27.8774)+$$
$$70.2747sin^4(16.9445\ D+20.2464)sin^4(18.7481\ W+21.9338) +$$
$$19.8196sin(22.0029\ HD+12.7998) + 1.6339 \quad (3)$$

The MVRM was obtained with a training subset of data. This formula predicts the accuracy with coefficient of determination (R^2) equal to 87% and 83% for train data (42 months) and test data (11 months), respectively (Fig. 4). And adjusted R^2 is 85% which is high. It means that the correlation coefficient between the observed value of the dependent variable and the forecast value based on the regression model was high.

ANOVA table showed that the value of F statistic was 48.36 and the significance of F was zero which is less than the critical value ($p<0.001$). The null hypothesis was rejected. It means that the model is significant.

3.3 Regression Assumptions

The proposed model was linear by β_{1-6} coefficients. It gives us the first assumption true. DW test was 1.998 which lies between 1.65 and 2.35. Therefore, there is no autocorrelation between residuals and predicted values (Fig. 5b). Thus, the model held 2 assumptions. The Breusch–Pegan Test showed that the null hypothesis was not rejected ($p>0.05$), meaning that the model holds 3 assumptions. It can be seen in Fig. 5a. by the QQ plot, which easily proves that it is homoscedasticity. The last assumption is also true for the given model and the mean of residuals is zero. In addition, Jarque–Bera(JB), Shapiro-Wilk, D'Agostino's K-squared, Anderson-Darling tests rejected null hypothesis ($p>0.05$) which means the residuals are normally distributed (Fig. 5c).



(a) QQ plot. (b) Independence of residuals. (c) Distribution of residuals.

Fig. 5: Residuals analysis.

3.1 Forecast Results

The forecasts result of training and testing data were represented by Fig. 6, where it was separated by a grey vertical line. By date (month, year) and the number of CLO are represented by x-axis and y-axis, respectively. Actual data is in blue color, training data is in green and testing data is in red color. The forecasting data started from November, 2002 until October 2003, which means that almost all 1-year forecasts are highly accurate.

We calculated lower and upper prediction intervals using

Upper Prediction Interval $\mathbf{UPI_i} = \hat{y}_i + z\sqrt{\text{MSE}}$

Lower Prediction Interval $\mathbf{LPI_i} = \hat{y}_i - z\sqrt{\text{MSE}}$ (4)

https://doi.org/10.31436/iiumej.v23i2.2164

Where:

$$\text{Mean Squared Error (MSE)} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \quad (5)$$

z =1.645, the forecasted data with a 90% prediction interval (6)

Prediction intervals for train data and test data $\mathbf{UPI_i}$ and $\mathbf{LPI_i}$ are represented by green dot line and red dot line, accordingly with 90% probability (z = 1.645). (Fig. 7a and Fig. 7b).

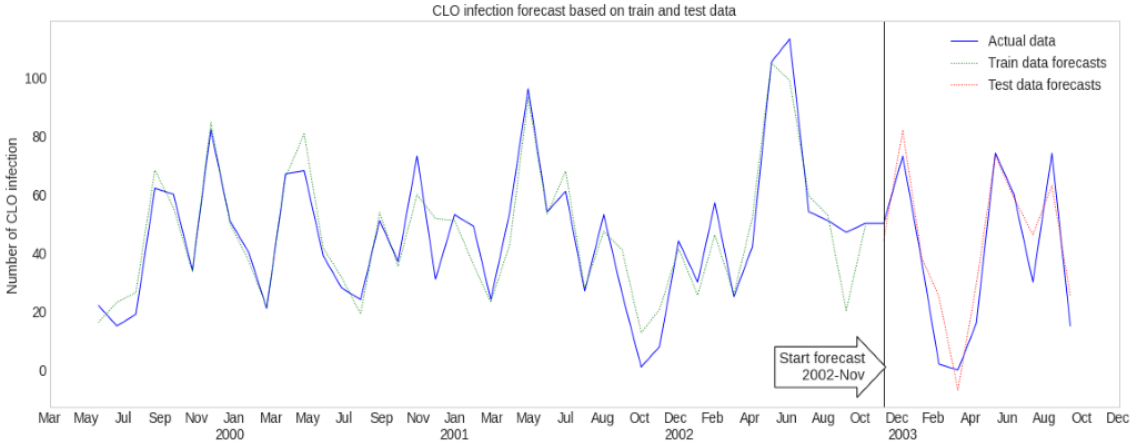


Fig. 6: True train and test data and its forecasts. (The grey line is separator between train and test data).

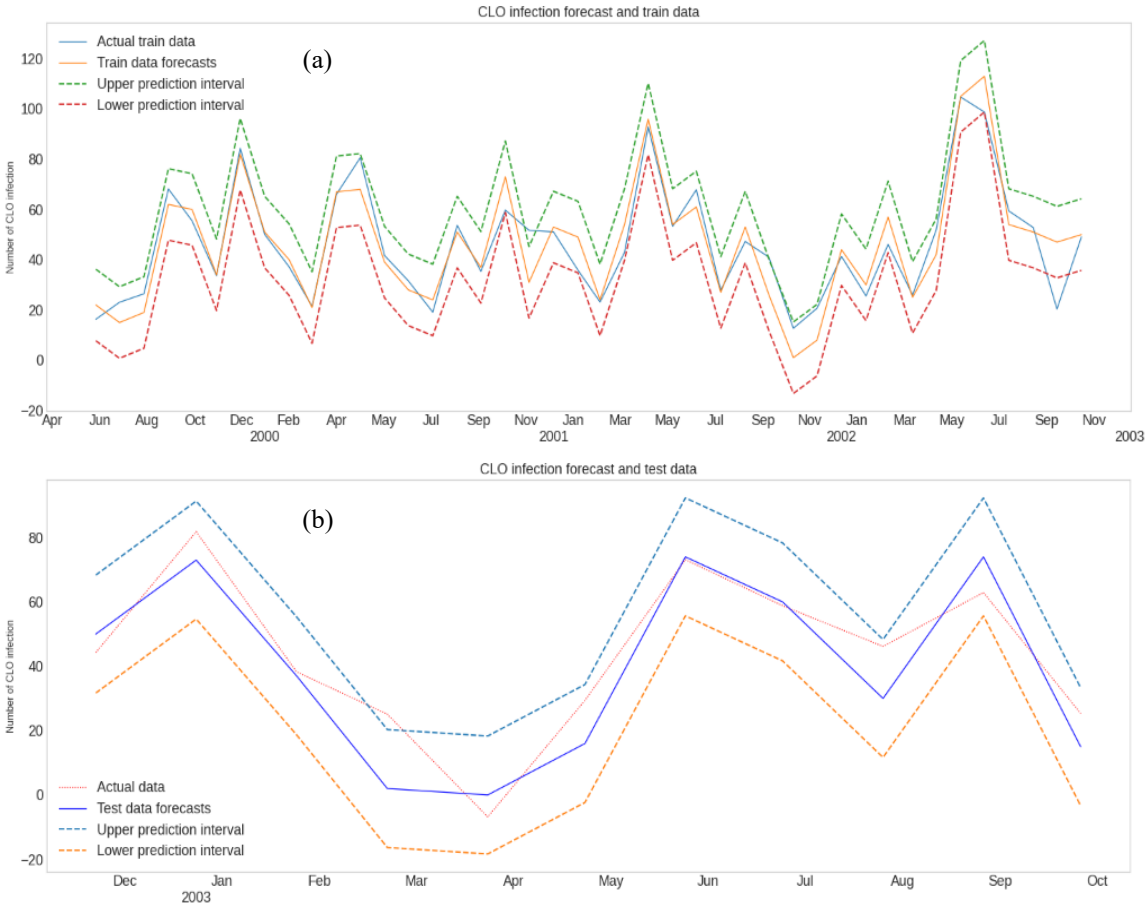


Fig. 7: Train data prediction with a 90% prediction interval (z =1.645, MSE = 79.29) for MVRM in (a), test data prediction with a 90% prediction interval (z =1.645, MSE = 124.13) for MVRM in (b.)

4. CONCLUSIONS

In this paper, we proposed non-linear MVRM to predict the prevalence of H. pylori infection prevalence based on the patients records of the hospital. If average monthly climate variables are introduced, the model predicts the number of *H. pylori* infection related to the given month's average climate variables. The proposed model uses to find patterns of H. pylori infection behavior based on the mean of humidity, dew point, temperature, and wind speed of months. Our researched showed that only the forecasting model achieves more accurate results by using the combinations of the given climate variables.

Since the infectious disease is a social problem, it can impact not only personal health, but can also cause widespread damage. Therefore, this research is being conducted to minimize social cost by predicting the prevalence of the *H. pylori* infection. The aim of this study was to design a forecasting model to predict *H. pylori* infection, which does not exist in the research papers yet, by using various input climate variable techniques based on non-linear MVRM with high accuracy. For this reason, we used non-linear Least Square method to find the regression coefficients of the model. The proposed model is significant since it holds four base assumptions of MVRM and gives 83% and 87% accuracy for training and testing dataset, respectively.

The proposed model helps to conduct precise predictive analysis of H. pylori infection prevalence for 1 year based on the dynamics of climate variables. Keeping in mind importance of climate variables in the forecast modelling of H. pylori infection prevalence we found high correlation between the climate factors and the prevalence. This model gives high accurate early forecast results which can be used by hospitals or governments to do early prevention acts against the infection prevalence, since it is critical to safe life of people and reduce cost in society. The proposed model is not only giving highly accurate results, but also it is easy to use by excel or sample calculators.

The obtained results of prediction analysis of *H. pylori* infection prevalence can be extended to the region with a similar climate condition. In further research, the model can be improved with different regions of databases and climate factors and also to check weather to possibly simplify mathematical formula of the proposed model by reducing the climate variables.

REFERENCES

[1] Lu C, Yu Y, Li L, Yu C, Xu P. (2018) Systematic review of the relationship of Helicobacter pylori infection with geographical latitude, average annual temperature and average daily sunshine. BMC gastroenterology, 18(1): 50.

[2] Tang MY, Chung PH, Chan HY, Tam PK, Wong KK. (2019) Recent trends in the prevalence of *Helicobacter pylori* in symptomatic children: A 12-year retrospective study in a tertiary centre. Journal of pediatric surgery, 54(2): 255-257.

[3] Peek Jr RM, Blasser MJ. (1997) Pathophysiology of *Helicobacter pylori*-induced gastritis and peptic ulcer disease. The American journal of medicine, 102(2): 200-207.

[4] Thorsen K, Søreide JA, Kvaløy JT, Glomsaker T, Søreide K. (2013) Epidemiology of perforated peptic ulcer: age-and gender-adjusted analysis of incidence and mortality. World Journal of Gastroenterology, 19(3): 347.

[5] Rawla P, Barsouk A. (2019) Epidemiology of gastric cancer: global trends, risk factors and prevention. Przeglad gastroenterologiczny, 14(1): 26.

[6] Song Y, Wang F, Wang B, Tao S, Zhang H, Liu S, Ramirez O, Zeng, Q. (2015) Time series analyses of hand, foot and mouth disease integrating weather variables. PloS one, 10(3): e0117296. https://doi.org/10.1371/journal.pone.0117296