Analysis of Regression and Neural Network Models in Predicting Patient Visit Volume

Harizahahyu[1], Friendly[2], Muhammad Fathoni[3], Yuyun Yusnida Lase[4], Santi Prayudani[5], Nur Laily Harfita[6]
[1,2,4,5]Politeknik Negeri Medan, Medan, Indonesia
[3]Politeknik Ganesha Medan, Indonesia
[6]Universitas Deztron Indonesia, Medan, Indonesia
Email: harizahayu@gmail.com

Abstract

Predicting patient visit volume plays a crucial role in supporting decision-making and resource allocation in healthcare services. This study aims to compare the performance of Multiple Linear Regression and an Artificial Neural Network (ANN) in forecasting patient visits at a dental clinic, using daily patient visit data and predictor variables such as holidays and promotional activities. Multiple regression was used to capture the linear relationship between the predictor and response variables, while ANN was applied to explore potential non-linear relationships. The results indicate that multiple regression outperformed the ANN, demonstrated by lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values, and provided clearer interpretability, making it more beneficial for healthcare practitioners, particularly in the context of a limited dataset. In contrast, the ANN tended to produce overestimates and was less responsive to short-term variations. Therefore, multiple regression can still be considered a reliable, efficient, and interpretable prediction method for clinical data with a moderate sample size, while future research is recommended to use larger datasets and test other machine learning algorithms to improve the accuracy and generalizability of the results.

***Keywords**: Patient Visit Prediction, Multiple Linear Regression, Artificial Neural Network, Healthcare Management.*

A. INTRODUCTION

Effective healthcare service planning relies heavily on the ability to accurately predict patient visit volume. Accurate forecasting enables more efficient resource management, including the scheduling of medical staff, inventory of medicines, and enhancement of service quality (Li et al., 2019). In the context of a dental clinic, fluctuations in patient visit volume are influenced by various factors, such as time trends, holidays, and promotional programs. Therefore, a prediction method capable of accommodating these variables is essential for improving service quality (Bai et al., 2023).

Various approaches have been employed to predict patient volume in dental clinics, ranging from simple regression models to artificial intelligence-based methods. While simple regression models are effective for capturing linear relationships between variables, they have limitations in handling data with more complex patterns (Montgomery et al., 2021). On the other hand, multiple regression offers an advantage by incorporating relevant additional variables, such as

promotional programs and holidays, making it more effective at modeling complex data (Gujarati & Porter, 2020). Meanwhile, the Artificial Neural Network (ANN) method has become a popular choice in health research due to its ability to recognize non-linear patterns and more complex relationships between variables (Al-Taie et al., 2021).

This study aims to compare the performance of three primary prediction methods, namely simple regression, multiple regression, and ANN, in forecasting patient visit volume at a dental clinic. The results are expected to provide a clearer picture of which method is most effective in addressing the challenge of predicting patient volume, as well as to offer practical recommendations for the clinic's management in making decisions related to service scheduling and resource allocation (Ju et al., 2014).

The results of this analysis will provide crucial insights into the strengths and limitations of each method in predicting patient visit volume. While simple regression may still be useful in scenarios with simpler data structures, both multiple regression and ANN can deliver higher accuracy when dealing with data involving more complex variables (Sari Rochman et al., 2018). Therefore, the selection of an appropriate method will heavily depend on the complexity of the data and the specific analytical objectives of the clinic. This study will evaluate the effectiveness of simple regression, multiple regression, and ANN in predicting patient visit volume at a dental clinic. Based on the findings, the clinic management can select the most suitable approach to enhance service quality and optimize operational management.

B. METHOD

This study employs three types of models to predict patient visit volume: simple regression, multiple regression, and Artificial Neural Network (ANN). Each model was selected based on its capability to handle relationships between predictor variables and patient visit volume at different complexity levels.

1. Simple Linear Regression

The first model developed was a simple regression model, where the time variable (in days) was used as the sole predictor. The purpose of this model is to test whether a linear relationship exists between time and patient visit volume. While this model can provide an initial understanding of the relationship between the time variable and patient visits, it has limitations in addressing more complex relationships. Formally, this model is expressed by the following equation (Angelini, 2025):

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

Where:

Y is the response variable representing patient visit volume.

X is the predictor variable representing time.

β_0 is the intercept parameter that estimates the baseline visit volume while $X_t = 0$.

β_1 is the slope parameter that quantifies the average change in Y_t for each one-

unit increase in time (one day).

ε_t is the random error term at time t that satisfies the *i.i.d*. (independent and identically distributed) assumption with zero mean and constant variance, $\varepsilon_t \sim N(0, \sigma^2)$ (Harizahayu et al., 2023).

The purpose of developing this model is to test the statistical significance of the linear relationship between time and visit volume, which is formally evaluated through the following hypothesis test:

H_0: $\beta_1 = 0$ (no linear relationship)

H_1: $\beta_1 \neq 0$ (a linear relationship exists)

Although this model provides valuable initial insights through the strength and direction of the relationship indicated by the estimated β_1 value and coefficient of determination dan (R^2), it possesses inherent limitations. The model assumes a constant linear relationship, thus being unable to capture non-linear patterns, seasonality, or more complex stochastic fluctuations often present in patient visit time series data. These limitations can lead to model misspecification and inaccurate forecasting results.

2. Multiple Linear Regression

The second model is multiple regression, which utilizes three predictors: day, promotion, and holiday. By incorporating additional variables such as promotional programs and holidays, this model aims to analyze how these three factors collectively influence patient visit volume. This approach provides a more comprehensive perspective compared to the simple regression model, as it can account for interactions among multiple independent variables (Gujarati & Porter, 2020).

Multiple regression expands the simple regression concept of$Y = \alpha + \beta X$ by incorporating more than one independent variable. The mathematical model for three predictors is expressed as follows:

Population Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad (2)$$

Where:

Y: the dependent variable is defined as the number of patient visits on the i-th day

$\hat{Y}$: the predicted number of patient visits.

X_1: Day (can be in numerical form, such as 1, 2, 3, ..., or as a dummy variable for a specific day, e.g., Monday).

X_2: Promo (Dummy Variable: 1 = a promotional program is active, 0 = no promotional program)

X_3: Holiday (Dummy Variable: 1 = national holiday, 0 = workday).

β_0: the population intercept. This is the predicted value of Y when all independent variables (X_1, X_2, X_3) sama dengan nol.

b_0: The sample estimate of β_0, calculated from the sample data

β_1, β_2, β_3: The population regression coefficient, which describes the slope or the change in Y for every one-unit change in X, assuming the other variables are held constant (Basri, 2018).

3. Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) is a computational model inspired by the structure and function of the human brain, designed to handle problems with non-linear relationships between inputs and outputs (Lo Brano, 2014). An ANN consists of three primary layers: the input layer, one or more hidden layers, and the output layer. Each neuron in a hidden layer calculates the weighted sum of its inputs, adds a bias term, and then processes this value using an activation function (Zhang, 2016; Isha et al., 2020). The fundamental formula for computing the output of a neuron i in a hidden layer is:

$$Z_i = \sum_{j=1}^{n} w_{ij} x_j + b_i \qquad (3)$$

Where:

w_{ij} is the weight connecting input neuron j to hidden neuron i.
x_j is the input value from neuron j.
b_i is bias.

The output of the neuron is then computed by applying an activation function, $f(z_i)$:

$$y_i = f(z_i) \qquad (4)$$

Commonly used activation functions include sigmoid, ReLU, or tanh, which transform input values into desired outputs. The ANN is employed to predict non-linear relationships in data, such as patient visit volume. This model can capture complex interactions among various factors like time, weather, and historical data. By using this method, more accurate predictions can be obtained for clinic service planning and management. ANN excels in handling data with more complex structures compared to traditional linear models (Isha et al., 2020).

C. RESULTS AND DISCUSSION

Simulated data was developed to reflect patient visit fluctuation patterns by accounting for factors that influence visit volume, namely time, promotional programs, and holidays. The data were generated using a simulation model that incorporated the effects of these variables to produce realistic and representative patterns. The dataset consists of 1000 data points over a specified time period, enabling a more comprehensive model evaluation.

The simulated data in this study were systematically developed to replicate patient visit fluctuation patterns approximating real-world conditions. This model development considered three main exogenous factors that have been theoretically and empirically proven to significantly influence visit volume, namely: temporal variables (time) capturing seasonal trends and long-term tendencies, the presence of health promotion programs (such as free vaccinations or mass screenings) that can stimulate demand increases, and indicators of national holidays which often cause drastic visit reductions. To generate realistic and representative data, an agent-based simulation model or time series model was used, integrating the stochastic (random) influence of each variable. This simulation produced a series of 1000 synthetic data points (observations). This extensive data scope was selected to ensure model stability,

enhance statistical power, and enable more comprehensive and robust evaluation of fluctuation patterns before model implementation in actual situations.

Before being used to build the models, the data underwent several processing stages:

a. Data Splitting: The data was divided into two parts: training data (75%) and test data (25%). The training data was used to train the models, while the test data was used to evaluate the performance of the built models. This split aims to prevent overfitting and ensure that the models generalize well to previously unseen data.
b. Data Normalization**:** Numerical variables such as 'day' and patient visit volume were normalized to ensure all variables were on a comparable scale, which helps enhance the performance of deep learning models.
c. Dummy Variable Encoding: Categorical variables promo, and holiday are encoded using one-hot encoding to convert them into numerical variables suitable for model input.

This study's results compare three predictive approaches for patient visit volume: simple regression, multiple regression, and Artificial Neural Network (ANN). The evaluation was conducted using Mean Absolute Error (MAE) on the test data. A summary of the results is presented in Table 1.

Table 1. Mean Absolute Error (MAE) Values for Each Model

Model	MAE
Simple Linear Regression	9.94
Multiple Linear Regression	6.63
Artificial Neural Network	9.53

Based on the table, the multiple regression model yields the smallest MAE value (6.63), indicating its superior accuracy compared to simple regression (9.94) and ANN (9.53). These results suggest that additional variables such as promotions and holidays play a significant role in explaining variations in patient visit volume. Simple regression, which relies solely on the time variable (day), consequently misses relevant information

Meanwhile, the ANN, which is theoretically capable of capturing non-linear relationships, did not demonstrate better performance compared to multiple regression. This can be attributed to the relatively small dataset size (n = 200) and the simple ANN structure with a single hidden layer containing 5 neurons. ANNs generally require larger datasets to achieve effective generalization.

Overall, the research findings demonstrate that multiple regression delivers the best performance in predicting patient visit volume compared to both simple regression and ANN. This finding aligns with previous studies confirming that multiple regression remains a reliable method when predictor variables exhibit a sufficiently strong linear relationship with the response variable [1]. In the context of clinical data with limited sample size, this model is not only computationally more efficient but also easier to interpret compared to ANN, which tends to be complex and

requires large amounts of data to achieve optimal performance. The superiority of multiple regression in this study is further supported by the visualization in Figure 1, where its prediction line appears closest to the actual data pattern compared to the other two methods. This visual alignment further strengthens the quantitative evaluation (MAE) results, confirming that multiple regression can capture patient visit variations more accurately, particularly when additional variables such as promotions and holidays are considered.

To clarify the performance differences among the three methods used, namely simple regression, multiple regression, and Artificial Neural Network (ANN), a visualization comparing the prediction results with the actual data was created. This visualization aims to demonstrate the extent to which each model can follow the fluctuation patterns of patient visit volume during the observation period. Thus, the following graph not only provides a quantitative picture of accuracy but also reveals the models' tendencies in capturing the actual data trends.

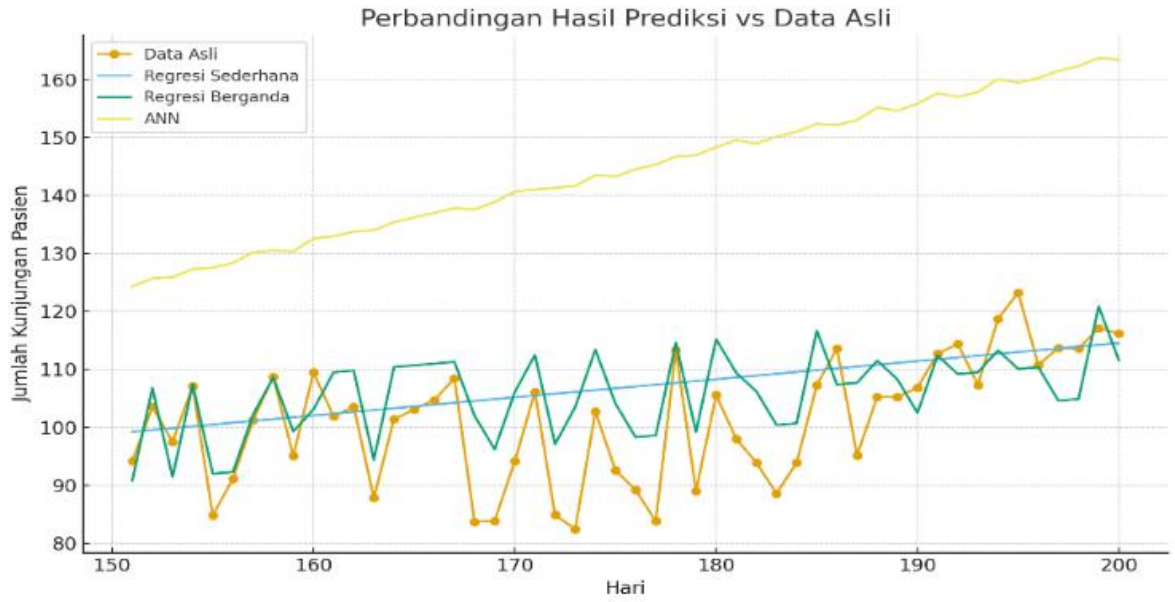

Figure 1. Comparison of the Predictive Performance of Simple Regression, Multiple Regression, and ANN Against Actual Data

Figure 1 shows a comparison between the actual data and the prediction results of three models: Simple Regression, Multiple Regression, and Artificial Neural Network (ANN).

a. Actual Data (yellow line with dots) displays patient visit fluctuations influenced by time trends (day), promotions, and holidays.
b. Simple Regression (blue line) uses only the day variable, thus tending to follow the general linear trend but lacking the ability to capture sudden fluctuations. This is reflected in its MAE value of 9.94, which is relatively high.
c. Multiple Regression (green line) considers the day, promotion, and holiday variables. This model is more accurate in capturing visit fluctuation patterns, with an MAE of 6.63. This indicates that promotional and holiday factors significantly influence patient visit volume.

d. ANN (light yellow line) attempts to model non-linear relationships among variables. However, in this study, ANN did not demonstrate better performance compared to multiple regression, with an MAE of 9.53. This may be attributed to the relatively small dataset size, preventing the neural network from learning patterns optimally.

As shown in Figure 1, multiple regression produces a prediction pattern that is relatively closer to the actual data compared to ANN. This confirms that the inclusion of several predictor variables contributes significantly to explaining variations in patient visit volume. The multiple regression prediction line follows the fluctuations of the actual data with relatively small deviations, although some extreme points are not fully captured. Meanwhile, ANN shows a tendency to consistently generate higher predictions, even during periods when actual patient visits decrease. This pattern indicates that ANN is more sensitive to long-term trends but less flexible in capturing seasonal variations or sudden external factors.

These findings underscore that model selection should not rely solely on numerical accuracy metrics such as RMSE or MAE, but also on practical needs in the field. In the context of clinics with limited data, multiple regression can be considered more efficient, interpretable, and relevant for supporting managerial decision-making. Conversely, the use of ANN may be more suitable for contexts with larger data volumes and high pattern complexity, such as in real-time monitoring-based prediction systems. Thus, this analysis opens opportunities for further research to explore hybrid models capable of combining the strength of multiple regression in interpretability with the advantage of ANN in capturing non-linear patterns.

Overall, multiple regression has proven to be the best model for this case, as it provides the most accurate predictions closest to the actual data. This finding indicates that considering additional variables beyond time is crucial for improving the accuracy of patient visit volume predictions.

D. CONCLUSION

This study demonstrates that the multiple regression model exhibits superior predictive performance compared to the Artificial Neural Network (ANN) in the context of patient visit volume at a dental clinic. The multiple regression model consistently produced results that were more aligned with actual data while maintaining interpretability, which is crucial for supporting managerial decision-making. In contrast, ANN tended to generate overestimated predictions and was less responsive to seasonal fluctuations or sudden changes, making it less optimal when working with limited datasets Practically, these findings confirm that conventional statistical models remain relevant and effective for medium-sized datasets, particularly when interpretability is a priority. However, artificial intelligence-based models still hold significant potential when applied to larger and more complex data. Therefore, future research is recommended to explore hybrid approaches that combine the strength of regression in interpretation with ANN's capability to capture non-linear patterns.

ACKNOWLEDGEMENTS

The authors gratefully express their appreciation and express their gratitude for the financial support provided through the 2025 DIPA Politeknik Negeri Medan funds with contract number: B/287/PL5/PM.01.05/2025 dated 19 August 2025.

REFERENCES

1. Al-Taie, Z., Liu, D., Mitchem, J. B., Papageorgiou, C., Kaifi, J. T., Warren, W. C., & Shyu, C. R. (2021). Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential. *Journal of Biomedical Informatics*, *118*, 103792.
2. Angelini C. (2025). *Regression Analysis*. Elsevier Ltd.
3. Bai, L., Lu, K., Dong, Y., Wang, X., Gong, Y., Xia, Y., ... & Li, C. (2023). Predicting monthly hospital outpatient visits based on meteorological environmental factors using the ARIMA model. *Scientific Reports*, *13*(1), 2691.
4. Basri, H. (2018). Pemodelan Regresi Berganda Untuk Data Dalam Studi Kecerdasan Emosional. *DIDAKTIKA: Jurnal Kependidikan*, *12*(2), 103-116.
5. Gujarati, D. N., & Porter, D. C. (2020). *Basic Econometrics*. New York: McGraw-Hill Education.
6. Harizahayu, H., Hermanto, K., & Yuniarti, R. R. (2023). Analisis Viral Marketing Pada Online Customer Terhadap Minat Pembelian Melalui Tiktok Shop Dengan Regresi Linier Sederhana. *Jurnal Sains Matematika dan Statistika*, *9*(2), 31-40.
7. Isha, Chaudhary, A. S., & Chaturvedi, D. K. (2020). Effects of Activation Function and Input Function of ANN for Solar Power Forecasting. In *Advances in Data and Information Sciences: Proceedings of ICDIS 2019* (pp. 329-342). Singapore: Springer Singapore.
8. Ju, X., Brennan, D. S., & Spencer, A. J. (2014). Age, period and cohort analysis of patient dental visits in Australia. *BMC Health Services Research*, *14*(1), 13.
9. Li, P., Kong, D., Tang, T., Su, D., Yang, P., Wang, H., ... & Liu, Y. (2019). Orthodontic treatment planning based on artificial neural networks. *Scientific reports*, *9*(1), 2037.
10. Lo Brano, V., Ciulla, G., & Di Falco, M. (2014). Artificial neural networks to predict the power output of a PV panel. *International Journal of Photoenergy*, *2014*(1), 193083.
11. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.
12. Sari Rochman, E. M., Rachmad, A., Syakur, M. A., & Suzanti, I. O. (2018, January). Method extreme learning machine for forecasting number of patients' visits in dental poli (A case study: Community Health Centers Kamal Madura Indonesia). In *Journal of Physics: Conference Series* (Vol. 953, p. 012133). IOP Publishing.
13. Zhang, Z. (2016). A gentle introduction to artificial neural networks. *Annals of Translational medicine*, *4*(19), 370.