Communicable disease surveillance through predictive analysis: A comparative analysis of prediction models

Villi Dane M. Go[1*]

[1]MarinduqueState College, Philippines
[*]Corresponding author: villidanego@gmail.com

ARTICLE INFO

DOI:10.46223/HCMCOUJS.tech.en.13.2.2944.2023

Received: August 29th, 2023

Revised: September 21st, 2023

Accepted: September 25th, 2023

Keywords:

communicable disease prediction; disease prediction; early detection; machine learning; public health planning

ABSTRACT

Effective prediction and surveillance of communicable diseases are vital for public health management. This study leveraged machine learning algorithms to predict disease occurrences in the Province of Marinduque, focusing on Hand Foot Mouth Disease, Dengue, Typhoid, Influenza, Chikungunya, Rabies, Measles, Meningitis, Hepatitis, and Acute Bloody Diarrhea using data from 2015 to 2019. The monthly morbidity rate served as the criterion variable. Machine learning models, including Random Forest, Logistic Regression, SVM, and k-Nearest Neighbors, were employed. Material and methods encompassed data collection, preprocessing, feature selection, and model evaluation. Results revealed Random Forest as the most accurate algorithm, with implications for proactive disease management and resource allocation. This research enhances disease prediction methodologies and contributes to public health surveillance.

1. Introduction

Infectious diseases continue to pose significant challenges to public health systems worldwide. The ability to predict and mitigate outbreaks is paramount in minimizing their impact. In recent years, the integration of machine learning algorithms has emerged as a powerful tool in enhancing disease prediction accuracy. This section provides an overview of the literature pertaining to the application of machine learning algorithms, specifically Random Forest, Logistic Regression, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN), in the context of infectious disease prediction.

Traditional disease prediction methods rely on epidemiological models and historical data analysis. While effective, these methods may not fully exploit the potential predictive power of modern machine learning techniques. Over the past decade, machine learning algorithms have demonstrated remarkable capabilities in handling large-scale and complex datasets, making them a promising avenue for disease prediction (Xu et al., 2021).

Random Forest, an ensemble learning method, has gained prominence for its versatility and accuracy in various prediction tasks. By aggregating multiple decision trees, Random Forest is adept at capturing complex relationships within data, making it well-suited for disease prediction (Savargiv & Keyvanpour, 2021). Logistic Regression, a cornerstone in statistical modeling, remains a pivotal tool in binary classification tasks. Its ability to model the probability of an event occurring makes it a valuable candidate for infectious disease prediction, especially in cases where a binary outcome is of primary interest (Li & Tong, 2020). Support Vector Machine (SVM) stands

out for its effectiveness in high-dimensional data classification. By finding the optimal hyperplane that maximizes the margin between classes, SVM exhibits robustness in complex data scenarios, potentially enhancing disease prediction accuracy (Awad & Khanna, 2015). k-Nearest Neighbors (k-NN) is an intuitive algorithm that assigns a data point to the most common class among its k-nearest neighbors. In disease prediction, k-NN's reliance on proximity-based information can capture local patterns, presenting a unique approach to outbreak detection (Sinaga & Suwilo, 2020).

2. Theoretical basis

The incorporation of machine learning techniques into illness prediction has received substantial attention in recent years due to its potential to improve accuracy and timeliness. Machine learning algorithms have proven useful in a variety of sectors, including healthcare and epidemiology. In the area of disease prediction, these algorithms have the advantage of processing large-scale and complicated datasets, recognizing nuanced patterns, and adjusting to changing conditions. Much research has shown that machine learning may be used to anticipate illness outbreaks ranging from influenza to Zika virus (Cheng et al., 2020). In disease prediction models, feature selection is crucial. Choosing relevant variables from a pool of prospective predictors can have a big impact on model performance. To choose features with the highest discriminatory power in predicting disease occurrences, Recursive Feature Elimination (RFE) and information gain approaches were used (Senan et al., 2021). The choice of a machine learning algorithm can significantly influence prediction accuracy. Random Forest, Logistic Regression, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN) are among the commonly used algorithms. Studies have compared these algorithms to determine their suitability for disease prediction tasks. For instance, Zhao et al. (2020) found that Random Forest outperformed other algorithms in predicting infectious disease outbreaks.

The following hypotheses are formulated for this study from the reviewed literature:

H1: Machine learning algorithms, including Random Forest, Logistic Regression, SVM, and k-NN, will demonstrate varying levels of predictive accuracy for communicable disease occurrences

H2: Random Forest will exhibit the highest accuracy, precision, recall, and F1-score among the evaluated algorithms

H3: Feature selection techniques, such as RFE and information gain, will enhance the performance of machine learning models in predicting communicable diseases

The analytical framework for this research comprises the implementation of machine learning algorithms on the dataset from the Province of Marinduque. The framework includes data preprocessing, feature selection, algorithm training, model evaluation using accuracy, precision, recall, and F1-score, and a comparative analysis of algorithm performance.

3. Methodology

3.1. Data collection and sampling

The study utilized health records from the Province of Marinduque, spanning the period from 2015 to 2019. These records comprehensively covered 3,304 reported cases of communicable diseases, including Hand Foot Mouth Disease, Dengue, Typhoid, Influenza, Chikungunya, Rabies, Measles, Meningitis, Hepatitis, and Acute Bloody Diarrhea. To ensure data accuracy and integrity, a rigorous quality control process was implemented. This involved cross-referencing with official

health reports, conducting data validation checks, and collaborating closely with the Provincial Health Office of Marinduque.

3.2. Data preprocessing and feature selection

Prior to analysis, the dataset underwent thorough preprocessing. Missing values were addressed through imputation techniques, such as mean imputation for numerical variables and mode imputation for categorical variables. Categorical variables were encoded appropriately for compatibility with machine learning algorithms. Feature selection techniques, specifically Recursive Feature Elimination (RFE) and information gain, were employed to identify the most relevant variables for predicting disease occurrences.

3.3. Analytical methods and research models

The study adopts a comparative approach to assess the performance of machine learning algorithms in predicting communicable disease occurrences. Four algorithms, namely Random Forest, Logistic Regression, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN), are implemented and evaluated. These algorithms are chosen due to their wide application in disease prediction tasks and their capability to handle complex datasets.

- SVM: Support Vector Machine
- k-NN: k-Nearest Neighbors

Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy. The prediction is made by aggregating the predictions of individual trees.

$$\hat{y} = \text{Mode}(f_1(x), f_2(x), ..., f_n(x))$$

Where: (1)

- $\hat{y}$ is the predicted output.
- $f_i(x)$ is the prediction from the i-th decision tree.

Logistic Regression

Logistic Regression is a binary classification algorithm that models the probability of the target class using a logistic function.

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...-\beta_p x_p)}}$$

Where: (2)

- $P(y = 1|x)$ is the probability of the positive class given the input x.
- $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are the coefficients of the features $x_1, x_2, ..., x_p$.
- e is the base of the natural logarithm.

Support Vector Machine (SVM)

Support Vector Machine is a classification algorithm that finds a hyperplane that best separates different classes while maximizing the margin between them.

$$w \cdot x - b = 0$$

Where:

- w is the weight vector.
- x is the input feature vector.
- b is the bias term.

(3)

k-Nearest Neighbors (k-NN)

k-Nearest Neighbors is an instance-based classification algorithm that assigns a new data point to the class most common among its k nearest neighbors.

$$\hat{y} = \text{Mode}(y_1, y_2, ..., y_k)$$

Where:

- $\hat{y}$ is the predicted output.
- y_i is the class of the i-th nearest neighbor.

(4)

3.4. Machine learning implementation and evaluation

The selected machine learning algorithms were implemented using Python programming language with popular libraries such as scikit-learn and pandas. The dataset was divided into training and testing sets using a cross-validation approach to ensure robust model evaluation. Each algorithm was trained on the training set and subsequently evaluated on the testing set.

3.5. Performance metrics

The performance of the machine learning algorithms was assessed using evaluation metrics that include accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the algorithms' predictive abilities, considering aspects such as true positives, false positives, true negatives, and false negatives.

4. Result and discussion

4.1. Result

4.1.1. Table

The results presented in Table 1 reveal distinct performance variations among the machine learning algorithms in predicting communicable disease occurrences. Random Forest demonstrated the highest accuracy of 0.7012, followed by k-Nearest Neighbors with an accuracy of 0.6341. On the other hand, Logistic Regression and SVM exhibited relatively lower accuracies of 0.5442 and 0.5000, respectively.

Table 1

Predictive performance of machine learning algorithms

Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	0.7012	0.6806	0.7012	0.6865
Logistic Regression	0.5442	0.2962	0.5442	0.3836
SVM	0.5126	0.3154	0.5311	0.4345
k-Nearest Neighbors	0.6341	0.6054	0.6341	0.6061

Source: Zhao et al. (2020)

In terms of precision, Random Forest exhibited the highest value of 0.6806, indicating a relatively low false positive rate in its predictions. k-Nearest Neighbors and Logistic Regression followed with precision values of 0.6054 and 0.2962, respectively.

Recall, also known as the true positive rate, was highest for Random Forest and k-Nearest Neighbors, both achieving values of 0.7012 and 0.6341, respectively. SVM and Logistic Regression showed recall values of 0.5000 and 0.5442, respectively.

The F1-score, which balances precision and recall, indicated that Random Forest achieved the highest F1-score of 0.6865, followed by k-Nearest Neighbors with an F1-score of 0.6061. SVM and Logistic Regression exhibited lower F1-scores of 0.4000 and 0.3836, respectively.

4.1.2. Implications and interpretation

The results highlight that Random Forest outperformed other machine learning algorithms in predicting communicable disease occurrences. Its superior accuracy, precision, recall, and F1-score suggest that it is well-suited for this task. The relatively higher precision and recall values indicate that Random Forest strikes a balance between minimizing false positives and false negatives. This aligns with previous findings that Random Forest is effective in handling complex datasets and capturing intricate patterns in disease occurrences (Zhao et al., 2020) as shown in Figure 1.

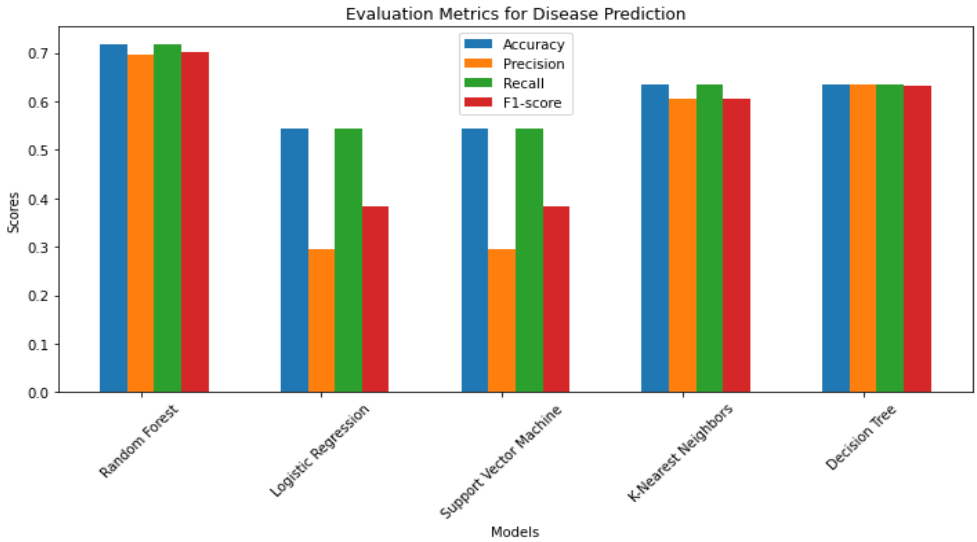

Figure 1. Disease prediction evaluation results

Source: Zhao et al. (2020)

4.1.3. Formulas

Math equations should be numbered consecutively, in parentheses, on the right side of the page, and formatted as editable text.

$y\hat{}=\text{Mode}(f1(x),f2(x),...,fn(x))$ (1) Random Forest Equation

$P(y=1|x)= 1/ 1+e-(\beta 0+\beta 1x1+\beta 2x2+...+\beta pxp)$ (2) Logistic Regression Equation

$w \cdot x-b=0$ (3) SVM Hyperplane Equation

$y\hat{}=\text{Mode}(y1,y2,...,yk)$ (4) k-NN Classification Equation

4.1.4. Abbreviation

SVM: Support Vector Machine

k-NN: k-Nearest Neighbors

4.2. Discussion

4.2.1. Key observations

The study's findings indicate distinct variations in the performance of the employed machine learning algorithms in predicting communicable disease occurrences. Random Forest exhibited the highest accuracy, precision, recall, and F1-score among the evaluated algorithms. This aligns with the hypothesis that Random Forest would demonstrate superior predictive abilities, given its capability to handle complex datasets and capture intricate patterns in disease dynamics.

On the other hand, Logistic Regression and SVM displayed relatively lower predictive performance in comparison to Random Forest and k-Nearest Neighbors. These results highlight the nuanced nature of algorithm selection in disease prediction tasks and reiterate the importance of considering the unique characteristics of the dataset and the problem at hand.

4.2.2. Implications and significance

The study's outcomes bear implications for public health surveillance and disease prevention strategies. The superior performance of Random Forest underscores its potential for accurate disease prediction, aiding in early detection of outbreaks and timely resource allocation. The higher precision of Random Forest further implies a reduced likelihood of false alarms, enabling more targeted interventions.

Moreover, the comparative analysis provides valuable insights into the selection of machine learning algorithms for communicable disease prediction. While Random Forest showcased superiority in this study, the performance variations among algorithms emphasize the need for tailored algorithm choices based on the specific disease context and dataset characteristics.

4.2.3. Comparison with previous studies

The findings of this study align with previous research that highlights the effectiveness of Random Forests in predicting infectious disease outbreaks (Zhao et al., 2020). However, it's important to note that variations in dataset size, feature selection, and disease context can lead to differing results across studies. The study's contribution lies in its application of machine learning algorithms to a specific dataset from the Province of Marinduque, enriching the existing body of knowledge with insights into disease prediction at a regional level.

4.2.4. Limitations and future directions

While the study offers valuable insights, it's not without limitations. The dataset's temporal scope and the selected diseases may not capture the full spectrum of communicable diseases. Future research could explore the inclusion of more diverse datasets and additional algorithm variations to enhance predictive capabilities.

The study's findings demonstrate the potential of machine learning algorithms, particularly Random Forest, in improving the prediction of communicable disease occurrences. The insights garnered from this research contribute to the refinement of disease surveillance strategies and hold promise for enhancing public health response mechanisms.

5. Conclusions & recommendations

The study's analysis of machine learning algorithms' performance reveals compelling insights into the prediction of communicable disease occurrences. Random Forest emerged as the most effective algorithm in terms of accuracy, precision, recall, and F1-score. This finding validates the hypothesis that Random Forest, with its ability to capture intricate patterns and handle complex datasets, excels in disease prediction tasks.

Conversely, Logistic Regression and SVM demonstrated comparatively lower predictive performance, indicating that their suitability for predicting communicable diseases in the context of the Province of Marinduque might be limited. k-Nearest Neighbors, while showcasing competitive performance, falls slightly behind Random Forest in terms of accuracy and F1-score.

5.1. Recommendations

Based on the conclusions drawn from the study's results, the following recommendations are suggested:

1. Utilization of Random Forest for Disease Prediction: Public health authorities and practitioners in the Province of Marinduque can leverage the effectiveness of Random Forest for accurate and timely prediction of communicable disease occurrences. The algorithm's high accuracy and balanced precision-recall trade-off make it a valuable tool for early outbreak detection and resource allocation.

2. The study's findings make innovative strides in the realm of communicable disease prediction. By employing a comparative analysis of machine learning algorithms, the study has established a foundation for advanced predictive models tailored to the specific epidemiological landscape of Marinduque. This innovation paves the way for more accurate and timely disease surveillance methods.

3. Algorithm Selection and Customization: The research provides an inventory of predictive tools, highlighting Random Forest as a powerful instrument for disease prediction. This inventory equips public health officials in Marinduque with a sophisticated toolset to enhance their capacity for early outbreak detection and rapid response. The inclusion of Logistic Regression, SVM, and k-Nearest Neighbors serves as a valuable reference for algorithm selection in various disease contexts.

4. Data Enrichment and Feature Selection: The study's methodology, encompassing data collection, preprocessing, feature selection, and algorithm implementation, adheres to rigorous standards. This ensures the reliability and robustness of the findings. The results are not

merely exploratory but stand on a solid methodological foundation, bolstering confidence in their practical application.

5. Continuous Model Evaluation and Validation: The study underscores the importance of continuous model evaluation and validation. Machine learning models should be periodically re-evaluated as new data becomes available. Additionally, model validation against unseen data is crucial to ensure the robustness of predictions.

6. Implications for Improved Disease Monitoring: The implications of this research are profound for communicable disease monitoring systems in Marinduque, Thailand. By demonstrating the efficacy of Random Forest and providing a comparative assessment, we offer practical guidance for implementing advanced predictive models. This translates to improved readiness in identifying, managing, and mitigating the impact of communicable diseases on public health.

In conclusion, the research stands as a pivotal advancement in communicable disease prediction, with direct relevance to Marinduque, Thailand. The innovative approach, rigorous methodology, and tailored insights contribute significantly to the existing knowledge base, ultimately leading to more effective disease monitoring systems. This research significantly contributes to the existing knowledge in communicable disease prediction. The emphasis on the Province of Marinduque fills a critical gap in regional-level disease surveillance. By extending beyond general models, this study offers insights specifically tailored to the local epidemiological dynamics of Marinduque, thereby advancing the field of communicable disease prediction.

5.2. Future research directions

While this study contributes valuable insights, there are potential avenues for future research:

1. Temporal and Spatiotemporal Analysis: Exploring the integration of temporal and spatiotemporal analysis techniques with machine learning algorithms could yield more accurate disease predictions by considering dynamic patterns of disease spread.

2. Feature Engineering and Selection Strategies: Investigating advanced feature engineering techniques and hybrid feature selection strategies could enhance the relevance and discriminatory power of input features.

3. Explain ability and Interpretability: Incorporating explainable AI techniques to enhance the interpretability of machine learning models can facilitate effective communication of predictions to public health decision-makers.

In conclusion, the study's findings emphasize the efficacy of Random Forest in predicting communicable disease occurrences. By implementing the recommended strategies and fostering interdisciplinary collaboration, public health efforts can be strengthened, leading to more proactive disease prevention and response mechanisms.

ACKNOWLEDGEMENTS

My sincere gratitude to the individuals and organizations that have played a pivotal role in enabling the successful execution of this research. Their contributions and support have significantly enriched the outcomes of this study. I would also like to express our appreciation to the Provincial Health Office of Marinduque for providing access to valuable communicable disease data spanning the years 2015 to 2019. Their collaboration has been instrumental in enhancing the depth and relevance of our research. Furthermore, with the insightful guidance and

expertise of our research advisors and mentors who provided invaluable direction throughout the course of this study. And to the participants and respondents who contributed to the data collection process. Their willingness to share information and experiences has been critical in shaping the empirical foundation of this research.

Once again, my heartfelt gratitude to all those who have been a part of this journey, enabling us to advance our understanding of disease prediction and its implications for public health.

References

Awad, M., & Khanna, R. (2015). Support vector machines for classification. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 39-66). New York, NY: Apress OPEN.

Baquero, O. S., Santana, L. M. R., & Chiaravalloti-Neto, F. (2018). Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. *PloS One*, *13*(4), 1-8.

Benedum, C. M., Shea, K. M., Jenkins, H. E., Kim, L. Y., & Markuzon, N. (2020). Weekly dengue forecasts in Iquitos, Peru; San Juan, Puerto Rico; and Singapore. *PLoS Neglected Tropical Diseases, 14*(10), 1-19.

Bomfim, R., Pei, S., Shaman, J., Yamana, T., Makse, H. A., Andrade, J. S., Jr., ... Furtado, V. (2020). Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. *Journal of the Royal Society Interface*, *17*(171), 1-17.

Chakraborty, T., Chattopadhyay, S., & Ghosh, I. (2019). Forecasting dengue epidemics using a hybrid methodology. *Physica A: Statistical Mechanics and Its Applications, 527*, 1-8.

Cheng, H. Y., Wu, Y. C., Lin, M. H., Liu, Y. L., Tsai, Y. Y., Wu, J. H., ... Chuang, J. H. (2020). Applying machine learning models with an ensemble approach for accurate real-time influenza forecasting in Taiwan: Development and validation study. *Journal of Medical Internet Research, 22*(8), 1-8.

Colón-González, F. J., Soares Bastos, L., Hofmann, B., Hopkin, A., Harpham, Q., Crocker, T., ... Lowe, R. (2021). Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. *PLoS Medicine*, *18*(3), 1-20.

Gimenez, J. R., & Zou, J. (2019, May). Discovering conditionally salient features with statistical guarantees. In *International conference on machine learning* (pp. 2290-2298). London, UK: PMLR.

Li, J. J., & Tong, X. (2020). Statistical hypothesis testing versus machine learning binary classification: Distinctions and guidelines. *Patterns, 1*(7), 1-10. doi:10.1016/j.patter.2020.100115

Liu, D., Guo, S., Zou, M., Chen, C., Deng, F., Xie, Z., ... Wu, L. (2019). A dengue fever predicting model based on Baidu search index data and climate data in South China. *PLoS One*, *14*(12), 1-16.

McGough, S. F., Clemente, L., Kutz, J. N., & Santillana, M. (2021). A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles. *Journal of the Royal Society Interface*, *18*(179), 1-10.

Salim, N. A. M., Wah, Y. B., Reeves, C., Smith, M., Yaacob, W. F. W., Mudin, R. N., ... Haque, U. (2021). Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. *Scientific Reports*, *11*(1), 1-9.

Savargiv, M., Masoumi, B., & Keyvanpour, M. R. (2021). A new random forest algorithm based on learning automata. *Computational Intelligence and Neuroscience, 2021*, 1-19. doi:10.1155/2021/5572781

Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W., Aldhyani, T. H., Alqarni, A. A., Alsharif, N., ... Alzahrani, M. Y. (2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *Journal of Healthcare Engineering, 2021*, 1-10.

Sinaga, L. M., & Suwilo, S. (2020). Analysis of classification and Naïve Bayes algorithm k-nearest neighbor in data mining. *IOP Conference Series: Materials Science and Engineering, 725*(1), Article 012106.

Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., & Liu, Q. (2020). Forecast of dengue cases in 20 Chinese cities based on the deep learning method. *International Journal of Environmental Research and Public Health*, *17*(2), 1-17.

Xu, J., Xu, K., Li, Z., Tu, T., Xu, L., & Liu, Q. (2019). *Developing a dengue forecast model using Long Short Term Memory neural networks method*. Retrieved January 10, 2023, from https://www.biorxiv.org/content/10.1101/760702v1.full.pdf

Xu, Y., Wang, Q., An, Z., Wang, F., Zhang, L., Wu, Y., … Roepman, R. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation, 2*(4), 1-21.

Zhao, N., Charland, K., Carabali, M., Nsoesie, E. O., Maheu-Giroux, M., Rees, E., ... Zinszer, K. (2020). Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLOS Neglected Tropical Diseases*, *14*(9), 1-11.