International Journal of Science and Healthcare Research
Vol.9; Issue: 2; April-June 2024
Website: ijshr.com
ISSN: 2455-7587

Research Article

A Hybrid Algorithm for Multiple Disease Prediction: Radial Basis Function and Logistic Regression

Fadhillah Azmi[1], Amir Saleh[2]

[1]Department of Electrical Engineering, Universitas Medan Area, Medan, Indonesia.
[2]Department of Informatics Engineering, Universitas Prima Indonesia, Medan, Indonesia.

Corresponding Author: Fadhillah Azmi

DOI: https://doi.org/10.52403/ijshr.20240246

ABSTRACT

Disease prediction is an important aspect of modern medicine, which aims to diagnose disease early and provide appropriate treatment to patients. This research uses a hybrid approach that combines the RBF (Radial Basis Function) kernel algorithm with logistic regression to predict various diseases in medical datasets. This method is intended to improve prediction performance by exploiting the advantages of each algorithm. This research uses a dataset containing medical information about several diseases collected from the Kaggle dataset. First, the RBF kernel is applied to transform the data features into a more informative, non-linear representation. Then, the logistic regression model is used to make predictions based on the features that have been processed by the RBF kernel. In this research, the hybrid RBF (Radial Basis Function) method was proven to be superior in predicting multiple diseases. This method shows the highest accuracy of 0.9460, as well as excellent precision, recall, and F1-score values of 0.8680, 0.8097, and 0.8294, respectively. The advantage of the hybrid RBF method lies in its ability to capture complex patterns in data that other methods often cannot identify, as well as its ability to handle non-linear decision boundaries, which are a common characteristic in medical datasets.

Keywords: *Disease prediction, Hybrid approach, RBF kernel algorithm, Logistic regression, medical datasets*

INTRODUCTION

Disease prediction is an important aspect of modern medicine, which aims to diagnose disease early and provide appropriate treatment to patients. With advances in information technology and artificial intelligence, data-driven approaches have become a major focus in the development of effective and accurate disease prediction systems. Analysing data to make predictions in the health sector is a complex challenge. Nonetheless, if done well, predictive analytics can be a very useful tool for healthcare practitioners in making timely decisions regarding patient health and treatment based on huge amounts of data (1). Machine learning algorithms are a popular approach to medical data analysis. The application of machine learning techniques has revolutionized the healthcare field by enabling accurate and timely disease predictions (2). The ability to predict multiple diseases simultaneously can significantly improve early diagnosis and treatment, which in turn improves patient outcomes and reduces health care costs (3). Machine learning algorithms can exploit hidden patterns in medical data to predict the occurrence of disease in patients.

Various studies have provided satisfactory results in the development of disease predictions. The SVM algorithm was used to achieve 78% accuracy in diabetes prediction research. Likewise, for Parkinson's disease prediction, it achieved 89% accuracy with SVM. Logistic regression is used for heart disease prediction, yielding 85% accuracy (4). Other research uses supervised machine learning techniques, where implementation is carried out by applying decision trees, random forests, Naïve Bayes, and KNN algorithms, which will help in timely disease predictions and better patient care. The research results confirm that the system is functional and user-oriented for timely diagnosis of diseases in patients (5). In this research, we use two commonly used algorithms, such as logistic regression and kernel RBF (radial basis function).

Logistic regression is a machine learning method used to model the relationship between independent variables and dependent variables by estimating the probability of an event occurring (6)(7). Meanwhile, kernel RBF is an effective method for non-linearly transforming data features into higher-dimensional feature spaces (8). Although both have been shown to be effective in disease prediction separately, the combined use of both, especially in a hybrid approach, can improve prediction performance. This approach makes it possible to exploit the advantages of each algorithm so as to produce a more powerful prediction model.

This research aims to combine kernel RBF and logistic regression algorithms in a hybrid approach for disease prediction on medical datasets. The medical data used in this research will be processed and adapted to the two proposed algorithms, namely kernel RBF and logistic regression, to ensure optimal integration. Then, the system will be trained using the processed dataset and tested using separate test data to evaluate its prediction performance. Thus, the results of this study are expected to provide valuable insights into the effectiveness of hybrid approaches in disease prediction, as well as provide a basis for the development of more sophisticated and reliable prediction systems in the future.

This hybrid approach is expected to produce more accurate predictions and could be an important contribution to the development of effective disease prediction systems. In addition, this research also aims to evaluate the performance of this hybrid approach with other machine learning methods using standard evaluation metrics such as accuracy, precision, recall, and F1-score. This was done to measure the relative superiority of the hybrid approach over other machine learning methods in the context of multi-disease prediction.

METHODS

In this study, we will explain the steps involved in predicting disease using the proposed method. The RBF kernel and logistic regression methods are very effective methods for classifying complex and non-linear data, which are often found in medical datasets. The description of each method used can be explained as follows:

1. Data Preparation

At the data preparation stage, the dataset was taken from online sources, namely the Kaggle Dataset. This dataset contains medical information obtained from various patients, which includes various medical attributes such as blood pressure, cholesterol, blood glucose, and others. The aim of this study is to predict the probability of the existence of several diseases based on these medical attributes. Information on the dataset used can be seen in Table 1 below.

Table 1. Information on Datasets

No	Name of Features	Range Value	Unit
1	Glucose	70, 140	mg/dL
2	Cholesterol	125, 200	mg/dL
3	Hemoglobin	13.5, 17.5	g/dL

4	Platelets	150000, 450000	50
5	White Blood Cells	4000, 11000	per microliter of blood
6	Red Blood Cells	4.2, 5.4	million cells per microliter of blood
7	Hematocrit	38, 52	percentage
8	Mean Corpuscular Volume	80, 100	femtoliters
9	Mean Corpuscular Hemoglobin	27, 33	picograms
10	Mean Corpuscular Hemoglobin Concentration	32, 36	grams per deciliter
11	Insulin	5, 25	microU/mL
12	BMI	18.5, 24.9	kg/m^2
13	Systolic Blood Pressure	90, 120	mmHg
14	Diastolic Blood Pressure	60, 80	mmHg
15	Triglycerides	50, 150	mg/dL
16	HbA1c	4, 6	percentage
17	LDL Cholesterol	70, 130	mg/dL
18	HDL Cholesterol	40, 60	mg/dL
19	ALT	10, 40	U/L
20	AST	10, 40	U/L
21	Heart Rate	60, 100	beats per minute
22	Creatinine	0.6, 1.2	mg/dL
23	Troponin	0, 0.04	ng/mL
24	C-reactive Protein	0, 3	mg/L

Once the dataset is collected, other data preparation steps are performed, such as checking the structure and contents of the dataset, performing data processing, such as dealing with missing or duplicate values, and converting target variables into a format suitable for analysis. Aside from that, the data is also divided into two parts: training data to train the model and test data to test the trained model's performance. This process is important to ensure that the developed model can be evaluated objectively on data not used during training.

2. RBF Kernel Modeling

The Kernel RBF (Radial Basis Function) algorithm is used to measure the distance between each pair of data in the transformed feature space. Basically, it is one of the algorithms used to measure the similarity between two points in a complex feature space (9). The mathematical approach underlying this algorithm can be calculated using the following equation 1 below (10).

$$K = \left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right) \qquad (1)$$

Where,

$K(x_i, x_j)$: RBF kernel value between two points x_i and x_j.

γ : Kernel parameter that determines how big the influence of each data point is on distance measurements.

$||x_i - x_j||^2$: Squared Euclidean distance between points x_i and x_j.

Using the RBF kernel function, we can calculate the RBF kernel matrix, where each element i,j represents the similarity between the i-th data and the j-th data in the dataset.

3. Logistic Regression Modeling

Logistic regression is one of the most commonly used methods for classification modeling where the target variable is binary. In this context, logistic regression will be trained using the RBF kernel matrix as input. The model will learn to predict the probability of disease occurrence based on patterns identified in the data. Using the following equation 2, the mathematical approach underlying logistic regression can be calculated (11):

$$p(y = 1 \mid x) = \frac{1}{1 + e^{-w^T x}} \qquad (2)$$

Where,

$p(y=1|x)$: Probability that target y is 1 (or disease occurs) based on feature x.

w : The weight vector will be learned during the model training process.

x : Feature vector representing each data sample.
e : Euler's number.

To make predictions, the probabilities generated by the sigmoid function will be converted into binary values using a threshold. If the probability is greater than the threshold, then the predicted class will be 1 (the disease occurs); if it is less than or equal to the threshold, then the predicted class will be 0 (the disease does not occur) (12). In this context, logistic regression will be used with the RBF kernel matrix as input, so that the model will learn the relationship between the features represented by the kernel matrix and the probability of disease occurrence.

4. Model Evaluation

Once the model has been trained, the final step is to evaluate its performance using standard evaluation metrics such as accuracy, precision, recall, and F1-score. This evaluation will provide insight into how well the model can predict disease events. A model evaluation is conducted to determine how well the trained model can predict disease events. This research uses several metrics that can be calculated using Equations 3, 4, 5, and 6 as follows (13)(14).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad 3$$

$$Precision = \frac{TP}{TP+FP} \quad 4$$

$$Recall = \frac{TP}{TP+FN} \quad 5$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad 6$$

Where,
TP (True Positive) : The number of samples that are actually positive and predicted to be positive.
TN (True Negative) : The number of samples that are actually negative and predicted to be negative.
FP (False Positive) : The number of samples that are actually negative but predicted to be positive.
FN (False Negative) : The number of samples that are actually positive but predicted to be negative.
This evaluation metric provides a comprehensive understanding of the model's performance in predicting disease incidence by considering various aspects of the confusion matrix.

RESULT & DISCUSSION

In this section, the results obtained from the experiments carried out will be described, as well as discussions related to these results. The findings will include an evaluation of the performance of a hybrid approach combining the kernel RBF and logistic regression algorithms in disease prediction. Performance evaluation will be carried out using standard evaluation metrics such as accuracy, precision, recall, and F1-score to assess how well the system can predict diseases with the proposed approach. Additionally, the results will be compared with other machine learning methods used as benchmarks, such as Decision Tree, Random Forest, Naïve Bayes, and KNN, to gain a better understanding of the relative merits of the hybrid approach. Table 2 below displays a comparison of prediction results using this method.

Table 2. Comparison of Methods

No	Methods	*Accuracy*	*Precision*	*Recall*	*F1-score*
1	KNN	0.8732	0.7332	0.7489	0.7396
2	Naïve Bayes	0.8462	0.7946	0.7346	0.7474
3	Decision Tree	0.9354	0.8378	0.8399	0.8378
4	Logistic Regression	0.8497	0.7204	0.7258	0.7228
5	SVM	0.8192	0.7057	0.7018	0.7011
6	Random Forest	0.9448	0.8065	0.7933	0.7981
7	The Proposed Method	0.9460	0.8680	0.8097	0.8294

According to Table 2, the KNN method has an accuracy of 0.8732, with precision 0.7332, recall 0.7489, and F1-score 0.7396, indicating a fairly good balance between precision and recall. Naïve Bayes, although its accuracy is slightly lower at 0.8462, shows a higher precision of 0.7946 and recall of 0.7346, with an F1-score of 0.7474, which means it has good ability in positive prediction.

The decision tree shows excellent performance with accuracy 0.9354, precision 0.8378, recall 0.8399, and F1-score 0.8378, indicating excellent ability to detect positive cases. Logistic regression has an accuracy of 0.8497, precision of 0.7204, recall of 0.7258, and F1-score of 0.7228, which indicates a lower balance between precision and recall compared to other methods. SVM shows the lowest performance with an accuracy of 0.8192, precision of 0.7057, recall of 0.7018, and F1-score of 0.7011, indicating that many positive predictions are wrong and many positive cases are not detected.

Random Forest shows very high performance, with an accuracy of 0.9448, precision of 0.8065, recall of 0.7933, and F1-score of 0.7981, showing a good balance between precision and recall. The proposed method showed the best performance, with the highest accuracy of 0.9460, precision of 0.8680, recall of 0.8097, and F1-score of 0.8294, making it the best choice for disease prediction among all the tested methods. Decision Tree and Random Forest methods are also very good and can be considered for interpretability or model complexity, while SVM seems less suitable for this dataset. Overall, the proposed method has the best performance, followed by the decision tree and random forest, which also show excellent results. SVM seems less suitable for this dataset because of its lowest performance, while other methods show quite good performance with some advantages and disadvantages.

The proposed hybrid RBF (Radial Basis Function) method demonstrates its superiority in multi-disease prediction for a variety of reasons. First, RBF has extraordinary capabilities for capturing complex patterns in data that other methods often cannot identify, both in local and global contexts. This is reflected in the highest accuracy achieved of 0.9460, indicating that this model is very effective in predicting correctly both on training data and test data. Second, RBF is powerful in dealing with non-linear decision boundaries, which is a common characteristic in complex medical datasets. This ability is reflected in the high precision (0.8680) and recall (0.8097) values, indicating that this model is able to detect the majority of positive cases without producing many false positive predictions. Third, the 0.8294 F1-score of the hybrid RBF method shows an optimal balance between precision and recall, which is very important in a medical context because errors in detecting disease (false negatives) can have a serious impact on patient health.

The evaluation results for the proposed method show very consistent and superior performance compared to other methods such as Decision Tree and Random Forest, which, despite having good performance, are not as high as hybrid RBF in the balance of evaluation metrics. Lastly, RBF's ability to handle outliers and noise, which are often present in medical datasets due to biological variations or measurement errors, allows it to produce cleaner and more accurate predictions. Overall, the hybrid RBF method excels due to its ability to handle data complexity, high performance in key evaluation metrics, and robustness to various challenges in medical datasets, making it the best choice for disease prediction.

CONCLUSION

In this research, the hybrid RBF (Radial Basis Function) method was proven to be superior in predicting multiple diseases. This method shows the highest accuracy of 0.9460, as well as excellent precision, recall, and F1-score values of 0.8680, 0.8097, and 0.8294, respectively. The advantage of the hybrid RBF method lies in its ability to capture complex patterns in data that other methods often cannot identify, as well as its ability to handle non-linear decision boundaries, which are a common characteristic in medical datasets. It has also been shown that the hybrid RBF method is better at dealing with overfitting and changing data than other methods like decision trees and random forests. These other methods also work well, but they're not as good as hybrid RBF in all evaluation metrics. In addition, RBF's ability to handle outliers and noise produces cleaner and more accurate predictions.

Declaration by Authors

Acknowledgement: None

Source of Funding: None

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

1. Mohit I, Kumar KS, Reddy AUK, Kumar BS. An Approach to detect multiple diseases using machine learning algorithm. J Phys Conf Ser. 2021;2089(1).
2. Shukla R, Sawant R. Multiple Disease Prediction System Using Machine Learning. IEEE Int Conf Electr Electron Commun Comput ELEXCOM 2023. 2023;3(June):88–94.
3. Chaudhari S, Deo P, Deshmukh P, Deshpande A, Shelke P, Chitre A. Multiple Disease Prediction Using Machine Learning Algorithm. 2023 7th Int Conf Comput Commun Control Autom ICCUBEA 2023. 2023;6(12):411–5.
4. Venkatesh M. Multiple Disease Prediction Using Machine Learning, Deep Learning and Stream-Lit. Int Res J Mod Eng Technol Sci. 2023;(07):37–45.
5. Kumar A, Pathak MA. A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives. Turkish J Comput Math Educ. 2021;12(6):4013–23.
6. Ciu T, Oetama RS. Logistic Regression Prediction Model for Cardiovascular Disease. IJNMT (International J New Media Technol. 2020;7(1):33–8.
7. Anshori M, Haris MS. Predicting Heart Disease using Logistic Regression. Knowl Eng Data Sci. 2022;5(2):188.
8. Rezki Purnajaya A, Jelita R, Tesvara E, Nestelrody M, Irwansyah J. Penerapan Metode Radial Basis Function (RBF) dalam Mengklasifikasikan Penyakit Demam Berdarah. J Digit Ecosyst Nat Sustain. 2023;3(1):2798–6179.
9. Wenmin Y. A modified radial basis function method for predicting debris flow mean velocity. J Eng Technol Sci. 2017;49(5):561–74.
10. Rochim AF, Widyaningrum K, Eridani D. Comparison of Kernels Function between of Linear, Radial Base and Polynomial of Support Vector Machine Method Towards COVID-19 Sentiment Analysis. 2021;224–8.
11. Panda NR, Pati JK, Mohanty JN, Bhuyan R. A Review on Logistic Regression in Medical Research. Natl J Community Med. 2022;13(4):265–70.
12. Boateng EY, Abaye DA. A Review of the Logistic Regression Model with Emphasis on Medical Research. J Data Anal Inf Process. 2019;07(04):190–207.
13. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep [Internet]. 2022;12(1):1–9. Available from: https://doi.org/10.1038/s41598-022-09954-8
14. Tasnim A, Saiduzzaman M, Rahman MA, Akhter J, Rahaman ASMM. Performance Evaluation of Multiple Classifiers for Predicting Fake News. J Comput Commun. 2022;10(09):1–21.

How to cite this article: Fadhillah Azmi, Amir Saleh. A hybrid algorithm for multiple disease prediction: radial basis function and logistic regression. *International Journal of Science & Healthcare Research.* 2024; 9(2): 363-368. DOI: *https://doi.org/10.52403/ijshr.20240246*

******