Journal of Soft Computing Exploration

Restricted boltzmann machine and softmax regression for acute respiratory infections disease identification

Afrizal Rizqi Pranata[1], Alamsyah[2], Budi Prasetiyo[3], Hilda Vember[4]
[1,2,3]Department of Computer Science, Universitas Negeri Semarang, Indonesia
[4]Department of Nursing Science, Cape Peninsula University of Technology, South Africa

Article Info

Article history:

Received Sep 22, 2022
Revised Sep 27, 2022
Accepted Sep 29, 2022

Keywords:

Restricted boltzmann machine
Artificial neural network
Deep learning
ARI identification

ABSTRACT

Restricted boltzmann machines (RBM) have attracted much attention lately after being proposed as building blocks of deep learning blocks. RBM is an algorithm that belongs to the artificial neural network (ANN) algorithm. Deep learning models can be used in the health field to identify diseases using medical data records. Acute Respiratory Infection (ARI) is a disease that infects the respiratory tract. A patient infected by ARI diseases is high. To identify ARI can use the symptoms that the patient had experienced. Based on this background, this study aims to help identify ARI disease using its symptoms. The method used for identification is the deep learning model, which was built using the RBM and softmax regression. Three steps were used in this research, which are training, testing, and implementation. The trained deep learning model will be implemented to identify ARI disease. This research will use ARI data from Puskemas Warungasem, Indonesia. From the research result, the deep learning model can get an accuracy of 96%. The deep learning configuration used in this research has 4 RBM layers, 1 Softmax layer as the output layer, and a learning rate value of 0.01 and 1000 iterations. This research can be used as a reference so that the next researcher can add other algorithms to Deep learning to improve accuracy.

Corresponding Author:

Afrizal Rizqi Pranata,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: afrizalrizqi@gmail.com

1. INTRODUCTION

In the subdivision of information technology, several methods can be used to help everyday life. One of which is for the identification of a disease. Deep learning is one of several methods that can be used for disease identification. Deep learning models have been successfully applied in classification tasks, regression, dimensional reduction, and information retrieval [1].

The deep learning model consists of several processing layers [2]. The deep learning model is better than the Shallow Model because the Shallow Model cannot get the expected knowledge [3]. To build a deep learning model can use the Artificial Neural Network (ANN) algorithm. ANN algorithm is a part of artificial intelligence that adopts the work of the human nervous system [4]. ANN imitates neural networks that exist in humans [5], [6]. The relation between nodes in the ANN is usually called network architecture. Nodes collected in the same layers are called layers [7]. RBM has become one ANN algorithm that can be used to build deep learning models. RBM has recently attracted much attention after being proposed as a building block of deep learning [8]–[10]. In this research, the proposed deep learning model comprises several layers

of RBM with a different number of visible and hidden nodes and an output layer. RBM layers will be used for unsupervised learning. To perform better training, the setting in full node and learning rate is essential [11], [12]. After that, the model will be used for supervised learning using softmax regression in the output layer. A study has been conducted to analyze RNA-seq data from Huntington's disease using Deep learning, which was built using the RBM model. The study's results produced an AUC value of 0.167 [13].

In this research, deep learning, built from some RBM layer and a softmax regression, will be used to identify ARI disease. ARI disease is currently causing nearly 4 million people to die each year, 98% of which are caused by lower respiratory tract infections [14]. Infants, children, and the elderly are the most vulnerable to ARI, especially in developing countries.

Based on the problem of high cases of ARI, it is necessary to identify ARI disease. A deep learning model will be built using RBM and softmax regression for the output layer to identify ARI disease. The ARI symptoms can be used as input for the deep learning model. The patient's symptoms can detect disease more quickly [15]–[17]. The deep learning model can study these symptoms unsupervised and supervised manner. Training must be done to make a deep learning model that best classifies ARI disease because accuracy is essential when classification [18]. So, implementing the deep learning model can give an accurate output of ARI disease suffered. This study will use medical record data of patients with ARI disease obtained from Puskemas Warungasem, Indonesia. The data was taken in the form of the results of medical records of Acute Respiratory Infection (ARI) patients at Puskemas Warungasem, Indonesia which amounted to 143 data. The data consists of patient that were diagnosed with common cold, sinusitis, pharyngitis, lung embolism and tuberculosis. Data taken in the form of symptoms experienced by the patient and the results of diagnosis.

2. METHOD

This section will explain the deep learning built from some RBM layer and a softmax regression to identify ARI disease.

2.1 Restricted Boltzmann Machine

A restricted boltzmann machine is an energy-based model that uses a layer from a remote unit to model a probabilistic distribution from a visible unit [19]. A joint configuration (***v***, ***h***) of the visible and hidden unit has energy given by formula (1):

$$E(v,h) = -\sum_{i \in visible} a_i v_i - \sum_{j \in hidden} b_j h_j - \sum_{ij} v_i h_j w_{ij} \quad (1)$$

Where v_i is a binary unit of visible layer, i and h_j is a binary unit of visible layer j. a_i and b_j are their biases, and w_{ij} is the weight between them. RBM network assigns a probability to every possible pair of a visible and a hidden unit via this energy function, which is shown in formula (2):

$$p(v,h) = \frac{1}{Z} e^{-E(v,h)} \quad (2)$$

Where the “partition function", Z, is given by summing over all possible pairs of visible and hidden vectors, which is shown in formula (3):

$$Z = \sum_{v,h} e^{-E(v,h)} \quad (3)$$

Because RBM does not have a connection between neurons in the same layer, every event is independent. So, the probability calculation is, which is shown in formulas (4) and (5):

$$p(h|v) = \prod_j p(h_j|v) \quad (4)$$

$$p(v|h) = \prod_i p(v_i|h) \quad (5)$$

In everyday learning, the generally binary unit is given. So, probability calculation is which is shown in formulas (6) and (7):

$$p(h_j = 1|v) = sigm(b_j + \sum_{i=1}^{m} v_i w_{ij}) \quad (6)$$

$$p(v_i = 1|h) = sigm(a_j + \sum_{j=1}^{n} h_j w_{ij}) \quad (7)$$

Where *sigm()* is the logistic sigmoid function 1/(1+exp(-*x*)). Vihj is then an unbiased sample. RBM is usually trained using Contrastive Divergence (CD), which minimizes kullback-leiber divergence [20]. So, the procedure to update weight and bias is, which is shown in formulas (8), (9), and (10):

$$\Delta W_{ij}^k = \epsilon(< v_i^k h_j >_{data} - < v_i^k h_j >_T) \quad (8)$$

$$\Delta a_i^k = \epsilon(< v_i^k >_{data} - < v_i^k >_T) \quad (9)$$

$$\Delta b_j^k = \epsilon(< h_j >_{data} - < h_j >_T) \quad (10)$$

2.2 Softmax Regression

Softmax regression is generated from logistic regression for the multi-classification problem. Softmax regression is composed of input, classifier, and output. The input training set for softmax regression with *v* number of data vector {(x_1, y_1), (x_2, y_2), …(x_v, yv) [21], [22]. In the softmax regression-based classifier, the probability P (Y =j | X) of X belonging to each class from a set of k classes is given as shown in the formula (11).

$$P(y_i = j|x_i; \theta) = \frac{e^{\theta_j^T x_i}}{\sum_{l=1}^{k} e^{\theta_l^T x_i}} \quad (11)$$

Where $j = 1, \ldots, k$ and $Y = [y_1, y_2, \ldots, y_k]$ is output class. Input variable to this probability function is feature vector $X = [x_1, x_2, \ldots, x_v]$, and the parameter weight $\theta = [\theta_0, \theta_1, \ldots \theta_k \in R^n]$ from softmax regression model. The cost function in softmax regression can describe as, which is shown in formula (12):

$$J(\theta) = -\frac{1}{v}[\sum_{i=1}^{v}\sum_{j=0}^{k} 1(y_i = j) \log P(y_i = j|x_i; \theta)] \quad (12)$$

This softmax regression cost function has no closed form way to minimize the cost value, so the iterative algorithm, gradient descent, is used, which is shown in formula (13):

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{v}\sum_{j=0}^{k}[x_i(1(y_i = j) - P(y_i = j|x_i; \theta))] \quad (13)$$

Weight parameter can be updated using $\theta_j = \theta_j - \alpha \nabla_{\theta j} J(\theta)$ for $j = 1, \ldots, k$. The weight θ initialization of softmax regression can be done using the rando number, and the weight can be updated every vector training $x_l^{(i)}$.

2.3 Research Data

This study will use medical record data of patients with ARI disease obtained from Puskemas Warungasem, Indonesia. The data was taken in the form of the results of medical records of ARI patients at Puskemas Warungasem, Indonesia, which amounted to 143 data. The data consists of patients diagnosed with common cold, sinusitis, pharyngitis, lung embolism, and tuberculosis. The data taken is gender, age, and symptoms from ARI patients. The symptoms used to identify ARI is 29 symptoms. A list of symptoms used to identify ARI can be shown in Table 1.

Table 1. List of symptoms that used to identify ARI

Symptoms	ARI Disease				
	Common Cold	Acute Sinusitis	Acute Pharyngitis	Lung Embolism	Tuberculosis
Fever	v	v	v	v	v
Headache	v	v	v	v	
Cough	v	v	v	v	v
Sneezing	v		v		
Nasal Congestion	v	v			
Malaise	v	v		v	v
Muscleache	v				v
Watery eyes	v				
Tenderness over the sinus area		v			
The sense of smell deteriorated	v	v			
Tooth ache		v			
A thick yellow or green discharge from the nose		v			
The face feels painful or depressed		v			
The mucous membrane is very red			v		
Sore throat	v		v		
Reddish tonsils			v		
Lymphoid follicles swell and are filled with exudates			v		
Enlargement and tenderness of cervical lymph nodes			v		
Rasp			v		
Chest pain				v	v
Dyspnea				v	
Takipnea				v	
Takikardia				v	
Diaforesis				v	
Sinkop				v	
Weight loss					v
Night sweats					v
Cough settled					v
No appetite	v	v	v	v	v

2.4 Deep learning

This research will build deep learning using some RBM layers and a softmax regression. The relation between softmax regression and some RBM layers is connected as a soft-hybrid system. Three steps will be used in this research, that is training, testing, and implementation steps. In the training step, some RBM layers will perform training in an unsupervised way. Then, Softmax Layer, which performs as the

output layer, will perform in a supervised way. ARI data with labels/diagnose from a doctor will be used in this step. K-fold cross-validation will be used for validation. In k-fold cross-validation, which breaks data into k parts in the same size. Each time test is conducted, the fraction is used as a set of test data while the other fraction is used as a set of training data [23].

The trained deep learning model will be implemented in the implementation step to identify ARI disease. ARI data without labels/diagnose from the doctor will be used in this step. The deep learning illustration is shown in Figure 1.

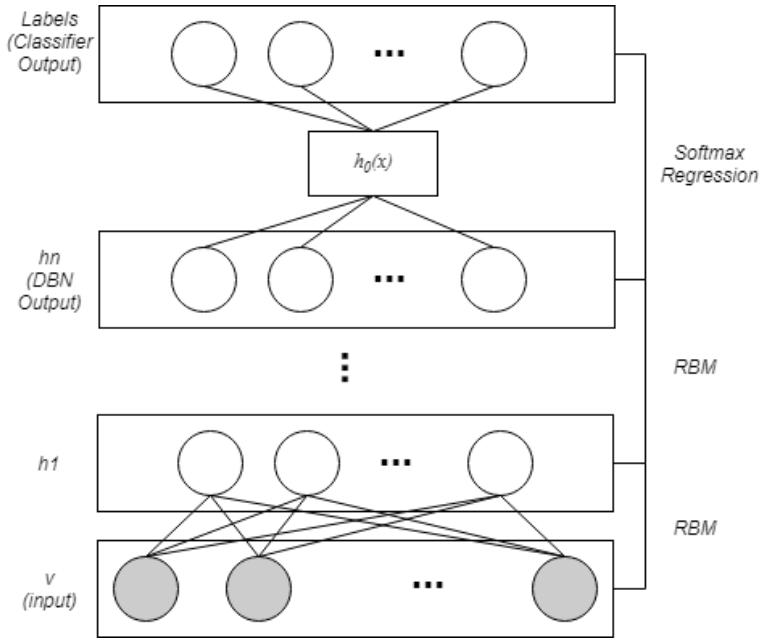

Figure 1. Deep learning illustration

3. RESULTS AND DISCUSSIONS

This research's first and second steps are training and testing the deep learning model using ARI data with a label. The ARI data used in this step is 118 data. K-fold cross-validation was used to divide the data into training and testing data. Some deep learning configurations will be used to perform this training step. Table 2 shows some deep learning configurations that are used.

Table 2. Deep learning configurations

	Layers	Visible Nodes	Hidden Nodes	Learning Rate	Iteration
Configuration 1	4	29	19;14;7;5	0.01	500
Configuration 2	4	29	20;10;7;5	0.01	1000
Configuration 3	4	29	20;10;7;5	0.01	750

The training step in RBM layers will be done in an unsupervised way. The output layer will use softmax regression which is composed of 5 nodes. This output layer will do training in a supervised way.

The trained deep learning model will be tested using testing data. The result was evaluated using a confusion matrix. After the test results are known, determine the accuracy using a confusion matrix [24]. Besides that, count precision and recall too. Because this research used k-fold cross-validation with default *k=10*, there will be 10 testing result. Table 3 shows the result in every fold of k-fold cross-validation.

Table 3. The result in every fold of k-fold cross validation

		Configuration 1	Configuration 2	Configuration 3
Fold 1	Accuracy	50 %	91.67 %	75 %
	Precision	46.67 %	73.33 %	65 %
	Recall	60 %	80 %	75 %
Fold 2	Accuracy	83.33 %	91.67 %	83.33 %
	Precision	83.33 %	73.33 %	85 %
	Recall	90 %	80 %	85 %
Fold 3	Accuracy	100 %	91.67 %	100 %
	Precision	100 %	93.33 %	100 %
	Recall	100 %	95 %	100 %
Fold 4	Accuracy	100%	91.67 %	100 %
	Precision	100%	90 %	100 %
	Recall	100%	95 %	100 %
Fold 5	Accuracy	100%	100 %	100 %
	Precision	100%	100 %	100 %
	Recall	100%	100 %	100 %
Fold 6	Accuracy	100 %	100 %	91.67 %
	Precision	100 %	100 %	90 %
	Recall	100 %	100 %	95 %
Fold 7	Accuracy	100%	100 %	91.67 %
	Precision	100%	100 %	75 %
	Recall	100%	100 %	80 %
Fold 8	Accuracy	91.67%	100 %	100 %
	Precision	95%	100 %	100 %
	Recall	93.33%	100 %	100 %
Fold 9	Accuracy	100%	100 %	100 %
	Precision	100%	100 %	100 %
	Recall	100%	100 %	100 %
Fold 10	Accuracy	100 %	100 %	91.67 %
	Precision	100 %	100 %	93.33 %
	Recall	100 %	100 %	90

All accuracy, precision, and recall from each fold are calculated and then count the average accuracy, precision, and recall. Table 4 shown of the result of test from each configuration.

Table 4. The result of test from each configuration

	Training Time	Accuracy	Precision	Recall
Configuration 1	10 Minutes 12 Seconds	92.5 %	93 %	94.333 %
Configuration 2	16 Minutes 10 Seconds	96.667 %	93 %	95 %
Configuration 3	12 Minutes 04 Seconds	93.333 %	90.833 %	92.5 %

The deep learning model will be implemented with the best accuracy, precision, and recall. ARI data without labels was used in this implementation step. Because the data don't have labels, the ARI data will be identified manually using symptoms data used in this research to check the result from the system. The evaluation in this step will be used a confusion matrix. And then count the accuracy, precision, and recall. Table 5 shows accuracy, precision, and recall in this model implementation.

Table 5. Accuracy, precision, and recall in this model implementation

Accuracy	Precision	Recall
96%	96.67 %	96.67 %

4. CONCLUSION

Research of acute respiratory infections disease identification using 143 ARI patient data records from Puskesmas Warungasem, Indonesia. The data consists of patients diagnosed with common cold, sinusitis, pharyngitis, lung embolism, and tuberculosis. Then, of the methods proposed in this study is the implementation of the deep learning model using a 4-layer restricted bolztmann machine, a learning rate of 0.01, and an iteration of 1000 obtains the best accuracy of 96% to identify acute respiratory infections disease.

REFERENCES

[1] N. Agarwalla, D. Panda, and M. K. Modi, "Deep learning using restricted boltzmann machines," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1552–1556, 2016.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with EMRs," in *2014 IEEE Int. Conf. Bioinform. Biomed. (BIBM)*, 2014, pp. 556–559.

[4] B. Prasetiyo, Alamsyah, M. A. Muslim, Subhan, and N. Baroroh, "Artificial neural network model for banckrupty prediction," *J. Phys. Conf. Ser.*, vol. 1567, no. 3, pp. 8–12, 2020.

[5] O. I. Abiodun *et al.*, "Comprehensive review of artificial neural network applications to pattern recognition," *IEEE Access*, vol. 7, pp. 158820–158846, 2019.

[6] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network," *Proc. Natl. Acad. Sci.*, vol. 117, no. 48, pp. 30071–30078, 2020.

[7] V. Upadhya and P. S. Sastry, "An overview of restricted Boltzmann machines," *J. Indian Inst. Sci.*, vol. 99, no. 2, pp. 225–236, 2019.

[8] C. Shang and F. You, "Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era," *Engineering*, vol. 5, no. 6, pp. 1010–1016, 2019.

[9] A. Fischer and C. Igel, "An introduction to restricted Boltzmann machines," in *Iberoam. congr. pattern recognit.*, 2012, pp. 14–36.

[10] R. Medina, R. Vasseur, and M. Serbyn, "Entanglement transitions from restricted Boltzmann machines," *Phys. Rev. B*, vol. 104, no. 10, p. 104205, 2021.

[11] S. Kassaymeh, M. Al-Laham, M. A. Al-Betar, M. Alweshah, S. Abdullah, and S. N. Makhadmeh, "Backpropagation Neural Network optimization and software defect estimation modelling using a hybrid Salp Swarm optimizer-based Simulated Annealing Algorithm," *Knowledge-Based Syst.*, vol. 244, p. 108511, 2022.

[13] X. Jiang, H. Zhang, F. Duan, and X. Quan, "Identify Huntington's disease associated genes based on restricted Boltzmann machine with RNA-seq data," *BMC Bioinform.*, vol. 18, no. 1, pp. 1–13, 2017.

[14] WHO Interim Guidelines, "Infection prevention and control of epidemic-and pandemic prone acute respiratory infections in health care," *JENEWA: WHO Interim Guidel.*, 2007. .

[15] J. Y. Nakayama, J. Ho, E. Cartwright, R. Simpson, and V. S. Hertzberg, "Predictors of progression through the cascade of care to a cure for hepatitis C patients using decision trees and random forests," *Comput. Biol. Med.*, vol. 134, no. March, p. 104461, 2021.

[16] A. M. Alfatah, R. Arifudin, and M. A. Muslim, "Implementation of decision tree and dempster shafer on expert system for lung disease diagnosis," *Sci. J. Informatics*, vol. 5, no. 1, p. 57, 2018.

[17] A. Lestari, "Increasing accuracy of C4. 5 algorithm using information gain ratio and adaboost for classification of chronic kidney disease," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 32–38, 2020.

[18] A. M. Putra and B. Pigawati, "Correlation between settlement environmental quality and Acute Respiratory Infection (ARI) disease of Gayamsari sub-district, Semarang," *Geoplanning J. Geomatics Plan.*, vol. 8, no. 1, pp. 51–60, 2021.

[19] R. Hrasko, A. G. C. Pacheco, and R. A. Krohling, "Time series prediction using restricted boltzmann machines and backpropagation," *Procedia Comput. Sci.*, vol. 55, no. Itqm, pp. 990–999, 2015.
[20] Q. Song *et al.*, "Micro-crack detection method of steel beam surface using stacked autoencoders on massive full-scale sensing strains," *Struct. Heal. Monit.*, vol. 19, no. 4, pp. 1175–1187, 2020.
[21] M. Jiang *et al.*, "Text classification based on deep belief network and softmax regression," *Neural Comput. Appl.*, vol. 29, no. 1, pp. 61–70, 2018.
[22] P. Chopra and S. K. Yadav, "Restricted Boltzmann machine and softmax regression for fault detection and classification," *Complex Intell. Syst.*, vol. 4, no. 1, pp. 67–77, 2018.
[23] A. Trihartati S. and C. K. Adi, "An identification of Tuberculosis (Tb) disease in humans using naive bayesian method," *Sci. J. Informatics*, vol. 3, no. 2, pp. 99–108, 2016.
[24] F. R. Devi, E. Sugiharti, and R. Arifudin, "The comparison combination of naïve bayes classification algorithm with fuzzy c-means and k-means for determining beef cattle quality in Semarang regency," *Sci. J. Informatics*, vol. 5, no. 2, pp. 194–204, 2018.