Machine learning approach for Acute Respiratory Infections (ISPA) prediction: Case study Indonesia

N L W S R Ginantra*, I G A D Indradewi and E Hartono
STMIK STIKOM Indonesia, Informatic Engineering, Denpasar, Bali, Indonesia

*wiiwik@stiki-indonesia.ac.id

Abstract. The progress of human civilization is growing rapidly in all areas of life. Science and technology are hereby inseparable from the life of modern society and capital. Can’t be denied the results of the life of modern society and capital spawned industries that have a big impact to the cause of the Acute Respiratory Infections (ISPA). Acute Respiratory Infections (ISPA) is a disease that attacks the breath through the nose, throat including adnexa network like sinuses with symptoms such as fever, sore throat and cough. In Indonesia ISPA ranks first cause of death in infant group and toddlers with percentages 38.80% of all toddler mortality. ISPA is a major cause of patient visits in health services, 40-60% visit at community health clinic and 15-30% visit treatment at hospital inpatient or outpatient. In this research, we collect the dataset we get from Community Health Clinic around Denpasar and Tabanan, Bali, Indonesia. The dataset consists of 14 attribute and 150 rows of data. In this study we conducted a study to predict and diagnose disease in a person whether positive suffering from ISPA or negative in accordance with the symptoms caused by using the machine approach. In this research we will make comparisons from some algorithm is KNN, SVM, Naïve Bayes and with Neural Network. The result this research the best performance is SVM with value sensitivity 100%, specificity 100% and accuracy 100%.

1. Introduction
The progress of human civilization is growing rapidly in all areas of life. Science and technology are hereby inseparable from the life of modern society and capital. Can’t be denied the results of the life of modern society and capital spawned industries that have a big impact to the cause of the Acute Respiratory Infections (ISPA). ISPA is a disease that attacks the breath through the nose, throat including adnexa tissue such as sinuses with symptoms such as fever, sore throat, and cough.

Anatomically ISPA grouped into ISPA over such as cough, colds, pharyngitis and ISPA such as bronchiolitis, pneumonia. Mortality rate (the size number of deaths in a population) cause by ISPA in infants, children and the elderly are high especially in countries with low per capita incomes and medium. According to data at United Nations International Children’s Emergency Fund (UNICEF) and World Health Organization (WHO) in 2011 ISPA is first dangerous toddler disease in the world compared to other diseases. According to a survey conducted in Indonesia ISPA ranks first cause of death in the infant group with a percentage of 38.80% of all toddler mortality [1]. ISPA is a major cause of patient visits in health services, 40-60% visit at community health clinic and 15-30% visit treatment at hospital inpatient or outpatient [1].
Early detection technology to diagnose illness ISPA of the symptoms of the disease ISPA currently needed to prevent delay in handling the disease ISPA, thereby reducing the adverse impact on toddlers and death.

In medicine, the application of machine learning has been tested in terms of prediction. Sangita Khare et al. apply machine learning to Investigation of Nutritional Status of Children using the dataset Indian Demographic and Health Survey Data [2]. Konstantina Kourou et al. do research by using machine learning approach for cancer prognosis. In his research did review on the research using which using machine learning for disease prediction [3]. Nooritawati Md. and Hany Hazfiza Manap did deep research Parkinson Disease Gait Classification. They do the classification using algorithm ANN and SVM. Maryam Zare et al. Conduct research using Chou’s Pseudo Amino Acid Composition and Machine Learning method to predict the antiviral Peptides. using several algorithms such as RBF, Naïve Bayes, J48, Decision Stump REPTree [4]. Hugo G. Schnack do research by applying machine learning for detecting and attacking heterogeneity in schizophrenia [5]. Joseph R. Imbus conduct research to identify multigland disease in primary hyperparathyroidism. In their research they apply the method RandomTree + AdaBoost compared with JRip + cost-sensitive learning [6].

From the problems above then we do the study to predict and diagnose disease in a person whether positive suffering ISPA or negative on accordance with the symptoms inflicted. By using machine learning approach in the prediction we will try some algorithm to get the best accuracy.

2. Method

2.1. Proposed method
In this research we will make comparisons from some algorithm i.e. KNN, SVM, Naïve Bayes and with Neural Network. Our first step is to do data collection for our dataset, then we filling the blank data, because if there is empty data then greatly affects the outcome of the prediction, the next step is to normalize the data. By using K-Fold validation, we do the dataset division for training and validation. The next step we apply machine learning algorithm which we tested for predictions. Our end result computes the results of the performance of each algorithm. Our proposed method explains on figure 1.

Figure 1. Our proposed method.
2.2. Algorithm comparison

2.2.1. K-Nearest Neighbours (KNN). KNN is one simple algorithm and straightforward data mining techniques [7]. Calculations on KNN use Euclidean distance. If the first distance is \((a_1, a_2, a_3, \ldots, a_n)\) and the second distance is \((b_1, b_2, b_3, \ldots, b_n)\) then the distance between them is calculated by the formula:

\[
\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}
\]

(1)

KNN usually deals with continuous attribute, although also deal with discrete attribute. When deal with discrete attribute, if the value of attribute value between two instances \(a_2, b_2\), is different, then the difference between them is worth one or is worth with zero [8].

2.2.2. Naïve Bayes. ISPA on patient. Naïve Bayes it perform arithmetical prediction [9]. Naive Bayes is a simple probabilistic classification which calculates a set of probabilities by summing the frequency and value combination of the given dataset. The algorithm uses Bayes's theorem and assume all attributes are independent or not the interdependence given by the value on the class variable [10, 11]. Naive Bayes is a classification by method probability and the statistics it finds by British scientist Thomas Bayes, predicting future opportunities based on past experience [11].

2.2.3. Support Vector Machine (SVM). Support Vector Machine (SVM) is a learning system that uses space hypothesis in the form of linear functions in a feature space high dimension [12, 13], trained with learning algorithms which is based on optimization theory by implementing learning bias derived from the theory of statistics learning. In this study we will test SVM in making predictions for illness of ISPA.

2.2.4. Neural network. Artificial Neural Network inspired by the consciousness of the complex learning systems in the brain consisting of closely related sets of neurons [14]. Neural networks are capable of performing tasks which are very complex such as classification and understanding patterns. ANN can be used for the same problem with multi-variable statistical problems such as multiple regression, discriminant analysis, and cluster analysis. In many cases the results obtained with ANN comparable to multi-variant statistics.

In this neural network we use Backpropagation algorithm. The backpropagation algorithm learning formula is explained in the following formula:

\[
\Delta w_{ij} = \eta x_i + \alpha \frac{\delta E_j}{\delta w_{ij}}
\]

(2)

Where \(w_{ij}\)is weight, \(\eta\) is learning rate, \(E\) is an error value and \(\alpha\) adalah factor moment. The parameters determine how big the effect of the old parameters against the direction of the new parameter changes.

In this research the network of neural networks is composed of 15 Input Neuron one hidden layer with 9 hidden neuron which was chosen carefully because too many neurons makes the network over specialized and leading to loss of generalizing capacity. The output of this neural network is 2 where it determines the positive and negative values.

2.3. K-Fold validation
Cross-validation (CV) is a statistical method that can be used to evaluate model performance or algorithm where data is separated into two subsets that is data of learning process and validation data / evaluation. The model or algorithm is trained by a learning subset and validated by a subset of validation [5]. Furthermore, the selection of CV types can be based on dataset size. Usually CV K-fold is used because it can reduce computational time while maintaining the accuracy of the estimation.
2.4. **ISPA data collection**

The dataset we use in this research we got from Community Health Clinic around Denpasar and Tabanan, Bali, Indonesia. The dataset we tested consists of 150 rows where the dataset consists of 14 attribute. We normalize the datasets we collect, value 1 for a valuable attribute “Yes” and 0 for a valuable attribute “No”. The attribute dataset is shown in table 1.

| Attributes      | Description                  |
|-----------------|------------------------------|
| Fever           | Yes = 1, No = 0              |
| Headache        | Yes = 1, No = 0              |
| Flu             | Yes = 1, No = 0              |
| Blurry View     | Yes = 1, No = 0              |
| Red Rash        | Yes = 1, No = 0              |
| Diarrhea        | Yes = 1, No = 0              |
| Cough           | Yes = 1, No = 0              |
| Gag             | Yes = 1, No = 0              |
| Hard to Swallow | Yes = 1, No = 0              |
| Nausea          | Yes = 1, No = 0              |
| Sore throat     | Yes = 1, No = 0              |
| Out of breath   | Yes = 1, No = 0              |
| Itchy           | Yes = 1, No = 0              |
| Limp            | Yes = 1, No = 0              |

Results: Positive or Negative

2.5. **Tools**

In this research we use computer with specification Mac OS X El Capitan, Intel Core i5 (1.5 GHz), 8 GB memory. We use python library in testing algorithm.

3. **Result and discussion**

To evaluate the accuracy of each algorithm, in this study using confusion matrix, Sensitivity and specificity also used in statistical measure of ISPA Prediction. Accuracy value is used to determine closeness of measurement to the true value. The sensitivity is used to determine the truth of the person with the disease ISPA, specificity used to assess whether a person does not have the disease ISPA. Sensitivity, Specificity and accuracy calculated using the following formula:

\[
Sensitivity = \frac{TP}{TP + FN} \quad (3)
\]

\[
Specificity = \frac{TN}{TN + FN} \quad (4)
\]

\[
Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)
\]

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is false Negative.

Through the above measurements we get the results shown in the table 2. We also shown performance measure based on graph (figure 2).
Table 2. Performance measure each Algorithm.

| Method    | Sensitivity | Specificity | Accuracy |
|-----------|-------------|-------------|----------|
| K-NN      | 0.974       | 0.90        | 0.947    |
| Naive Bayes | 0.971      | 0.891       | 0.942    |
| SVM       | 1.0         | 1.0         | 1.0      |
| Neural Network | 0.962     | 0.802       | 0.938    |

From the table above shows that by using SVM method get the best result with value sensitivity 100%, specificity 100% and accuracy 100%. This concludes that in the case of predictions on the dataset ISPA, SVM algorithm has a very good prediction rate because it can predict with truth value up to 100%.

Figure 2. Performance measure graph each Algorithm.

4. Conclusion

In this research, we collect the dataset we get from Community Health Clinic around Denpasar and Tabanan, Bali, Indonesia. The dataset consists of 14 attribute and 150 rows of data. Using the 4 algorithms we use to make comparisons in measurements. The algorithm we use is KNN, Naïve Bayes, SVM and Neural Network. We do testing by measurement sensitivity, specificity and accuracy. Algorithm with the best performance i.e. SVM with value sensitivity 100%, specificity 100% and accuracy 100%. Machine learning can help people to diagnose illness of ISPA. The next research we will do is to test other algorithms and apply machine learning for other case predictions.

Acknowledgments

This research was partially supported Community Health Clinic around Denpasar and Tabanan, Bali, Indonesia. We thank our colleagues from LPPM STMIK STIKOM Indonesia. We would also like to show our gratitude to STMIK STIKOM Indonesia and all team research.

References

[1] World Health Organization 2007 Epidemic-prone & pandemic-prone acute respiratory diseases: Infection prevention & control in health-care facilities
[2] Khare S, Kavyashree S, Gupta D and Jyotishi A 2017 Procedia Comput. Sci. 115 338-349
[3] Kourou K, Exarchos T P, Exarchos K P, Karamouzis M V and Fotiadis D I 2015 CSBJ 13 8-17
[4] Zare M, Mohabatkar H, Faramarzi F K, Beigi M M and M Behbahani2015 Op. Bioinform. Jour.
19 13-19

[5] Schnack H G 2017 Schizophr. Res.
[6] Imbus J R, Randle R W, Pitt S C, Sippel R S, Schneider D F and Surg J 2017 Res. 219 173-179
[7] Khamis H S, Cheruiyot K W and Kimani S 2014 Inter. Jour. Of Infor. And Comm. Tech. Res. 4 121-128
[8] Akhil M, Deekshatulu B L and Chandra P 2015 Procedia Technol. 45 70-75
[9] Agarwal S, Jain N and Dholay S 2015 Procedia - Procedia Comput. Sci. 45 70-75
[10] Science C and Studies M 2015 Inter. Jour. Of Adv. Res. In Comp. Sci. And Manag. Stud. 3 9-19
[11] Chaurasia V and Pal S 2013 Inter. Jour. Of Adv. Comp. Sci. And Inform. Tech. 2 56-66
[12] Agarap A F M 2017 CoRR
[13] Meyer D, Leisch F and Hornik K 2003 Neurocomput 55 169-186
[14] Vapnik V and Izmailov R 2017 Annals. Of Math. And Artif. Intell. 81 3-19