2021 IEEE 7th Information Technology International Seminar (ITIS)
Surabaya, Indonesia, October 6 - 8, 2021

Identification of Acute Respiratory Infections in Toddlers Based on the Chi-Square And Naive Bayes Methods

Devie Rosa Anamisa
Faculty of Engineering
University of Trunojoyo Madura
Bangkalan-Madura, Indonesia
devros_gress@trunojoyo.ac.id

Muhammad Yusuf
Faculty of Engineering
University of Trunojoyo Madura
Bangkalan-Madura, Indonesia
muhammadyusuf@trunojoyo.ac.id

Wahyudi Agustiono
Faculty of Engineering
University of Trunojoyo Madura
Bangkalan-Madura, Indonesia
wahyudi.agustiono@trunojoyo.ac.id

Mohammad Syarief
Faculty of Engineering
University of Trunojoyo Madura
Bangkalan-Madura, Indonesia
mohammad.syarief@trunojoyo.ac.id

Muhammad Ali Syakur
Faculty of Engineering
University of Trunojoyo Madura
Bangkalan-Madura, Indonesia
alisyakur@trunojoyo.ac.id

Husna
Faculty of Engineering
University of Trunojoyo Madura
Bangkalan-Madura, Indonesia

2021 IEEE 7th Information Technology International Seminar (ITIS) | 978-1-6654-0807-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ITIS53497.2021.9791614

***Abstract*— Acute Respiratory Infections usually attack the respiratory tract of toddlers, both the upper and lower respiratory tracts, because the body's defense system against viruses that cause infection has not yet been formed. And usually, parents will know if the baby's condition is very chronic so that the baby experiences complications. This causes the need for a system that can assist in the early detection of respiratory tract infections. This study proposes the Chi-Square and Naive Bayes (NB) method. The Chi-Square method is a feature selection method to reduce features that have no effect. At the same time, the NB method is a prediction method that performs a simple probability-based identification process based on the application of the Bayes theorem with the assumption of strong independence. The contribution of this study is to determine respiratory tract disease in infants using the chi-square feature selection and the NB method, which can assist parents in detecting respiratory tract infections. From the tests that have been carried out using 120 datasets with 90 as training data and 30 as test data, the accuracy is 75.833%. This proves that the Chi-Square and NB methods are able to identify respiratory tract infections.**

Keywords*— *Identification; Acute Respiratory Infections; Method; Chi-Square; Naive Bayes;

I. INTRODUCTION

Infection is one of the factors that cause death in children under five years of age, especially Acute Respiratory Infections (ARI)[1]. ARI is the most common infection in humans in all[2]. Children and infants are the most vulnerable and most affected by ARI because their immune system has not yet formed as an antidote to the virus[3]. The classification of acute respiratory infections is divided into two parts, including upper, such as rhinitis, pharyngitis, tonsillitis, rhinosinusitis, and otitis media. At the same time, the lower part includes epiglottitis, croup, bronchitis, bronchiolitis, and pneumonia[4]. This infection is a disease caused by various kinds of microorganisms and can cause infection. Approximately four million people die from ARI every year. Therefore, this study developed an ARI disease identification system for toddlers with the aim of facilitating early identification.

Several methods have been used for decision-making, one of which is Naive Bayes. The Naive Bayes method is a simple probability-based identification method based on the Bayes theorem approach. The advantages of the Naive Bayes method, including being: easy to implement and giving good results for many cases. This is evidenced in research conducted by [5] for early identification of soybean disease, where NB can identify input based on the rule and has a high accuracy value. In addition, in terms of performance measures, NB is very good and shows that considering NB as a classifier is the optimal choice in measuring students' internal aspects with psychometric measurements[6]. Then in 2017, a primary headache early detection design using NB was developed, which is able to produce high accuracy in making a diagnosis decision between migraine, cluster, and TTH[7]. The identification method uses all the features contained in the data to build a model, even though not all of these features match the identification results[8][9]. Therefore, in this study, the chi-square method was used in the feature selection stage. Feature selection is a technique to select important and relevant features from the data and reduce irrelevant features [10]. Feature selection aims to select the best feature from a feature data set. The chi-square method is a hypothesis testing method for the difference between two or more proportions using the chi-square distribution[11][12]. The chi-square distribution is comparing the observed frequencies with H_0 as all proportions are equal to the criteria if the test will accept H_0 if the calculated X^2 value is less than the critical value and vice versa.[13][14].

Therefore, in this study, the NB method was applied to identify ARI diseases based on the Chi-Square feature selection system. The goal is that the system can reduce features that have no effect on the identification stage of ARI disease so that it can help someone in identifying ARI disease easily and quickly with high accuracy.

II. LITERATURE REVIEW

ARI is one of the main causes of high mortality and morbidity rates in infants and toddlers in Indonesia[15]. The death rate increases every year due to ARI disease. This

978-1-6654-0807-3/21/$31.00 ©2021 IEEE

condition creates many obstacles experienced by health workers. In addition to the obstacles in the medical section, most people are still unfamiliar with medical matters, so that if they experience symptoms of the illness, they are not necessarily able to understand the ways to overcome them[16]. This is because the lack of public knowledge about the symptoms and how to treat ARI disease is one of the factors causing the high mortality rate due to ARI[17]. An expert system is a system that adopts human knowledge to computers so that computers can solve problems as is usually done by experts [18]. The Naive Bayes Classifier method that can help the community identify ARI disease based on the symptoms they face is like consulting a doctor. Previously, someone had done research on this ARI disease expert system, including [19] discussing the expert system for diagnosing ARI in children using the web-based Naive Bayes Classifier method. Then according to [20] discusses the web-based ARI diagnosis expert system with the Forward Chaining method. However, in making the identification, it is still not optimal because it has used all the existing features. Therefore, in this study, we developed an identification method for ARI disease based on feature selection. Feature selection can help select features that are relevant to the identification process[21]. In a previous study, feature selection was carried out to select somatic depression symptoms as an important indicator of depression, especially among female acute coronary syndrome (ACS) patients. [22]. According to [23] for the fittest of the distribution of a population within the limiting distribution of the test-statistics derived by Kolmogorov for the cumulative distribution of theoretical contributions.

In this case, we propose a model to improve accuracy in the identification of ARI disease using NB based on feature selection as a decision-making solution in the early treatment of ARI disease. This problem is how to identify high probability ARI diseases making quick decisions based on feature selection. As a result of the identification of ARI diseases to achieve expert system performance in diagnosis comparable to optimization procedures.

III. Research Method

Data is a model that is represented in the form of pictures, words, and numbers and processed into important information [24]. The systematic procedure used to collect data is quantitative research methods. The dataset used is data on six types of ARI disease, 120 patient data with the spread of ARI disease from Syarifah Ambami Rato Ebu Hospital, Bangkalan, and 16 data on symptoms of ARI disease. From these data, a feature selection process was carried out using chi-square and NB for the identification process of ARI disease.

A. Feature Selection of Chi-Square

Feature selection is made to reduce irrelevant features in the identification process by the NB method. There are several methods for feature selection, namely Gain Information (GI), chi-square (X^2), and the most commonly used are frequency-based[25]. The of the purpose of using feature selection is to remove confounding features in classification or identification. Chi-Square feature selection uses statistical theory to test the independence of a term with its category. Based on statistical theory, two events include the occurrence of features and circumstance of categories, which then each term value is sorted from the highest based on equation 1 [26].

$$X^2(D,t,c) = \sum_{e_t \epsilon \{0,1\}} \sum_{e_c \epsilon \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \quad (1)$$

Chi-Square feature selection is made by sorting each feature based on the Chi-Square feature selection from the largest value to the smallest value, which can be seen in Fig. 1. The Chi-Square feature selection value that is greater than the significant value indicates the rejection of the independence hypothesis. Meanwhile, if two events show dependents, then the feature resembles or is the same as the appropriate category label in the category.

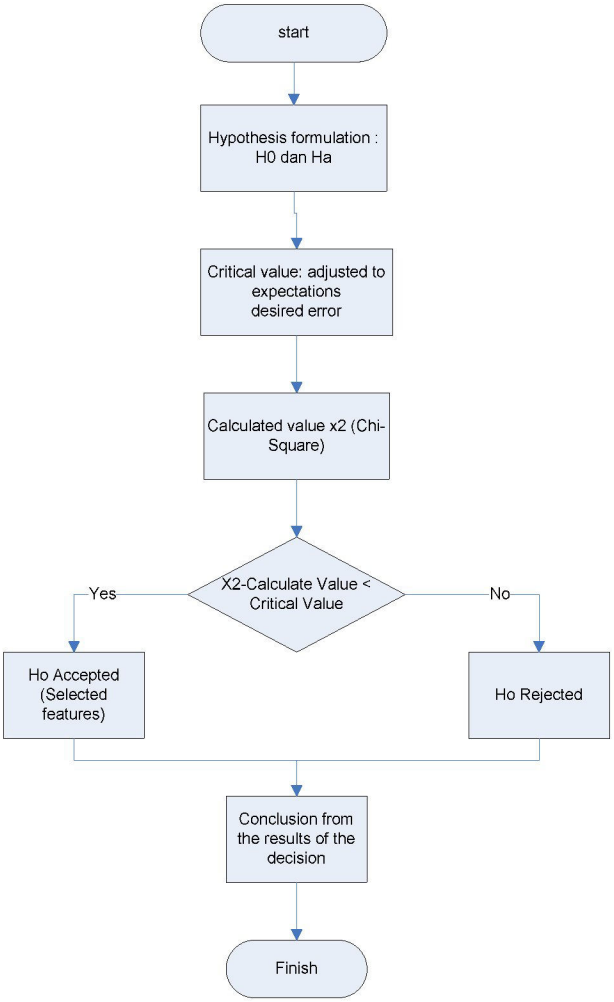

Fig. 1 Diagram of Chi-Square Feature Selection

B. Naive Bayes Method

The Naive Bayes method is one of the methods contained in the classification technique with the probability method known as Bayes' Theorem[27][28]. The Naive Bayes method is a statistical approach, conditional probability. The advantages of the NB method include: it can be used for both quantitative and qualitative data, does not require a large amount of data, does not need to do a lot of training data, is calculated quickly and efficiently[29]. Meanwhile, the steps for the NB method in the identification process include[30]:

a. Read training data
b. Calculate the amount and probability as in equation 2, but if the data is numeric, then look for the mean and standard

deviation of each parameter which is numeric data. The process diagram of the NB method can be seen in Fig. 2.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (2)$$

P(H|X) is the probability of hypothesis H based on the condition X (posterior probability), P(H) is Hypothesis probability H (prior probability), P(X|H) is The probability X based on the conditions in the hypothesis H and P(X) is the Probability X.

C. *Accuracy*

To calculate the accuracy value using equation 3, which uses the correct amount of data compared to the test data.

$$accuracy = \frac{\sum correct\ test\ data}{\sum total\ test\ data}\ x\ 100\% \quad (3)$$

IV. RESULTS AND DISCUSSIONS

In this section, we report the experimental results of identification performance assessment by conducting feature selection for 16 feature data using chi-square. The results can be seen in Table I. The calculations show that of the 16 features, which have H_o < Critical Value, there are five features, which can be seen in Fig. 3. The next step is the process of identifying ARI diseases based on the selected features using the NB method, which can be seen in Table II for class probability values and Table III for a table of prediction results in identifying ARI diseases using all features. Based on the calculation of accuracy with equation 3, this system produces an accuracy of 75.833%. Based on the results of the analysis shows that the Naive Bayes method based on the chi-square feature selection is better because this method can inspect a number of attributes that are not relevant to the chi-square method[31][32]. And eliminate a number of these attributes and perform data substitution so that the data can be classified in data mining using Naive Bayes so that it can identify the risk of ARI disease. This is evidenced by the test results using the confusion matrix, as shown in Table IV.

TABLE I. RESULTS OF FEATURE SELECTION FOR ARI DISEASE

Feature	Result	
	x^2hitung	x^2tabel
Cough	60.71	69.93
Out of breath	61.60	69.93
Chest pain	43.04	69.93
Nausea and Vomiting	46.96	69.93
Fever	52.52	69.93
Headache	76.64	69.93
Heartburn	80.78	69.93
Shivering	85.14	69.93
Cold Sweat	70.59	69.93
Insomnia	75.38	69.93
Flue	97.93	69.93
Exit secretions/snot	71.34	69.93
Easily Tired	74.95	69.93
Joint Stiffness	77.96	69.93
Moldy Tongue	72.46	69.93
Pain in the Nose	105.26	69.93

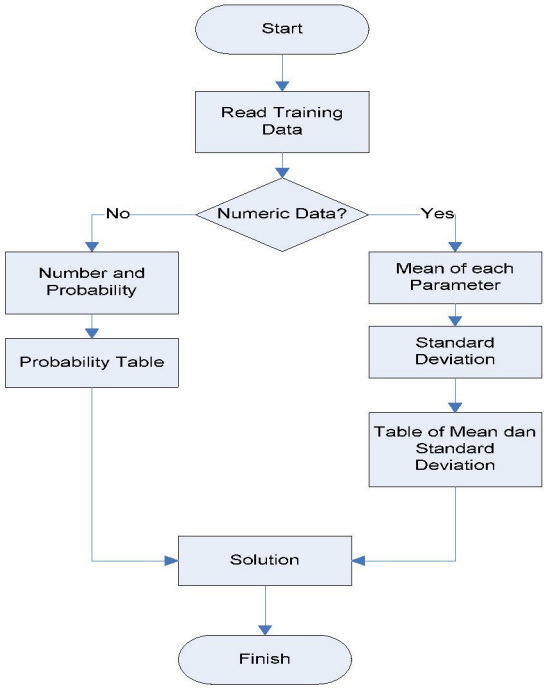

Fig. 2 Diagram of NB Method

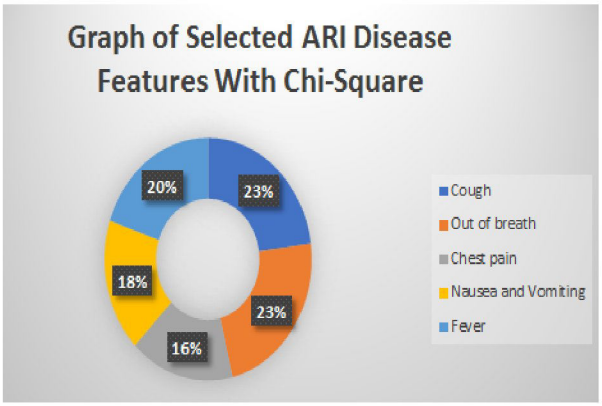

Fig. 3 Feature Selection of Chi-Square To ARI DiseaseS

TABLE II. VALUE OF CLASS PROBABILITY

Class			
Yes		*No*	
Total class "Yes."	90	**Total class "No."**	30
P(Yes)	0,75	**P(No)**	0,25

TABLE III. POSTERIOR OF PROBABILITY

Attribute	Attribute values	P(X | H)		P(H|X)	
		Yes	No	Yes	No
Cough	Y	53,33	0	0,021	0
	N	13,33	0,33		
Out of breath	Y	18,66	0	0,007	0
	N	48	0,33		

Chest pain	**Y**	56	0,33	0	0
	N	10,66	0		
Nausea and Vomiting	**Y**	21,33	0,16	0,004	0
	N	45,33	0,16		
Fever	**Y**	18,66	0,03	0,052	0
	N	48	0,3		
Headache	**Y**	16	0,1	0	0,063
	N	50,66	0,23		
Heartburn	**Y**	60	0,33	0,005	0
	N	6,66	0		
Shivering	**Y**	53,33	0	0,074	0
	N	13,33	0,33		
Cold Sweat	**Y**	18,66	0	0	0,063
	N	48	0,33		
Insomnia	**Y**	56	0,33	0,008	0
	N	10,66	0		
Flue	**Y**	21,33	0,16	0	0
	N	45,33	0,16		
Exit secretions/s not	**Y**	18,66	0,03	0	0,087
	N	48	0,3		

TABLE IV. CONFUSION MATRIX TEST

Confusion Matrix	Real Data		
	True		False
Data Prediction	True	82	0
	False	38	38

V. CONCLUSION

The conclusion from the results of this study regarding the identification of ARI disease using chi-square feature selection with Naïve Bayes has resulted in an accuracy value of 75.833% with five selected features, namely Cough, Out of breath, Chest pain, Nausea and Vomiting, Fever. This proves that the Naive Bayes method based on chi-square feature selection is better for identifying the risk of ARI disease.

REFERENCES

[1] S. Tomczyk *et al.*, "Factors associated with fatal cases of acute respiratory infection (ARI) among hospitalized patients in Guatemala," *BMC Public Health*, vol. 19, no. 1, pp. 1–11, 2019.

[2] A. B. Anggraini and S. Wirasmi, "Treatment patterns of acute respiratory tract infection in children under-fives in Bogor , Indonesia," vol. 11, no. 1, pp. 9–14, 2020.

[3] Infection prevention and control of epidemic-and pandemic-prone acute respiratory diseases in health care", WHO Interim Guidelines, pp. 12.

[4] B. Trollfors, "Acute respiratory infections in children," *Curr. Opin. Infect. Dis.*, vol. 7, no. 2, pp. 157–161, 1994.

[5] I. P. Astuti, I. Hermadi, A. Buono, and K. H. Mutaqin, "Soybean Disease With Naïve Bayes Approach," Indonesian Librarian Journal, vol. 14, no. 2, 2005.

[6] S. Mulyati, "IDENTIFYING STUDENTS ' ACADEMIC ACHIEVEMENT AND," pp. 64–68, 2014.

[7] Z. A. Azzahra, E. Purwanti, and H. B. Hidayati, "Design of Expert System As a Support Tool for Early Diagnosis of Primary Headache," *MNJ (Malang Neurol. Journal)*, vol. 3, no. 2, pp. 78–87, 2017.

[8] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. January 2003, pp. 1157–1182, 2003.

[9] D. Oreski and T. Novosel, "Comparison of Feature Selection Techniques in Knowledge Discovery Process," *TEM J.*, vol. 3, no. 4, pp. 285–290, 2014.

[10] R. Nair and A. Bhagat, "Feature selection method to improve the accuracy of the classification algorithm," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6, pp. 124–127, 2019.

[11] Herwin and Sophak Phonn, "The Application of The Generalized Lord's Chi-Square Method In Identifying Biased Items", Journal of Educational Research and Evaluation, vol 23, no.1, pp. 57-67, June 2019.

[12] Jian Sun, Xiang Zhang, Dan Liao, Victor Chang, "Efficient method for feature selection in text classification", Engineering and Technology (ICET) 2017 International Conference on, pp. 1-6, 2017

[13] N. S. Turhan, "Karl Pearson ' s chi-square tests," vol. 15, no. 9, pp. 575–580, 2020.

[14] C. Series, "A study of Hepatitis B virus infection using chi-square statistic," 2021.

[15] S. T. Zulaikhah, P. Soegeng, and T. Sumarawati, "Risk factors of acute respiratory infections in the training area for the community of medical students in Semarang," *Kesmas*, vol. 11, no. 4, pp. 192–197, 2017.

[16] K. Otomo, T. Ishibashi, H. Kataoka, Y. Hatakeyama, and Y. Okuhara, "Identifying diseases asscіated with a high risk for acute kidney injury using a hospital information system database," 6th Int. Conf. Soft Comput. Intell. Syst. 13th Int. Symp. Adv. Intell. Syst. SCIS/ISIS 2012, pp. 560–563, 2012.

[17] A. Mulwinda, "Intelligent Diagnosis System for Acute Respiratory Infection in Infants," pp. 558–562, 2017.

[18] A. A. Pramesti, R. Arifudin, and E. Sugiharti, "Expert System for Determination of Type Lenses Glasses Using Forward Chaining Method," *Sci. J. Informatics*, vol. 3, no. 2, pp. 177–188, 2016.

[19] M. Marlina, W. Saputra, B. Mulyadi, B. Hayati, and J. Jaroji, "Application of an expert system for diagnosing ARI disease based on speech recognition using the Naive Bayes classifier method," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 8, no. 1, pp. 58–70, 2017.

[20] T. F. Ramadhani, I. Fitri, and E. T. E. Handayani, "Web-Based ARI Disease Diagnosis Expert System With Forward Chaining Method," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 5, no. 2, p. 81, 2020.

[21] Z. Pang, D. Zhu, D. Chen, L. Li and Y. Shao, "A computer-aided diagnosis system for dynamic contrast-enhanced mr images based on level set segmentation and relieff feature selection", Computational and mathematical methods in medicine, vol. 2015, no. 2015, 2015.

[22] L. Frazier *et al.*, "Gender Differences in Self-Reported Symptoms of Depression among Patients with Acute Coronary Syndrome," *Nurs. Res. Pract.*, vol. 2012, pp. 1–5, 2012.

[23] S. Journal and A. Statistical, "The Kolmogorov-Smirnov Test for Goodness of Fit Author (s): Frank J . Massey , Jr . Published by : Taylor & Francis , Ltd . on behalf of the American Statistical Association Stable URL : https://www.jstor.org/stable/2280095," vol. 46, no. 253, pp. 68–78, 1951.

[24] A. Arman, "Information System for Processing Population Data of Nagari Tanjung Lolo, Tanjung Gadang District, Sijunjung Regency Web-Based," *Journal of Informatics Education.*, vol. 2, no. 2, pp. 163–170, 2017.

[25] O. Al-Harbi, "A Comparative Study of Feature Selection Methods for Dialectal Arabic Sentiment Classification Using Support Vector Machine," vol. 19, no. 1, pp. 167–176, 2019.

[26] M. Bhagat, "Sentiment Analysis using an ensemble of Feature Selection Algorithms," 2018.

[27] I. Technology, "Analysis of the Naïve Bayes Method in Classifying Formalized Fish Images Using GLCM Feature Extraction," vol. 1, no. 2, pp. 120–128, 2020.

[28] G. Rashmi, A. Lekha and N. Bawane, "Analysis of efficiency of classification and prediction algorithms (naive bayes) for breast cancer dataset", Emerging Research in Electronics Computer Science and

Technology (ICERECT) 2015 International Conference on, pp. 108-113, 2015.

[29] H. H. Manap, N. M. Tahir, and R. Abdullah, "Anomalous gait detection using Naive Bayes classifier," ISIEA 2012 - 2012 IEEE Symp. Ind. Electron. Appl., pp. 378–381, 2012.

[30] A. Hammouch, "Comparison of Classification Methods to Detect the parkinson disease," pp. 0–3, 2016

[31] F. Uğurlu, S. Yıldız, M. Boran, Ö. Uğurlu, and J. Wang, "Analysis of fishing vessel accidents with Bayesian network and Chi-square methods," *Ocean Eng.*, vol. 198, no. December 2019, 2020.

[32] D. Seka, B. S. Bonny, A. N. Yoboué, S. R. Sié, and B. A. Adopo-gourène, "Arti fi cial Intelligence in Agriculture Identi fi cation of maize (Zea mays L .) progeny genotypes based on two probabilistic approaches : Logistic regression and naïve Bayes," *Artif. Intell. Agric.*, vol. 1, pp. 9–13, 2019.