Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 4 (2023) 751 - 757 ISSN Media Elektronik: 2580-0760

Classification of Hearing Loss Degrees with Naive Bayes Algorithm

Okky Putra Barus[1], Romindo[2], Jefri Junifer Pangaribuan[3]
[1,2,3] Information Systems, School of Information Science and Technology, Universitas Pelita Harapan
[1]okky.barus@uph.edu, [2]romindo@uph.edu*, [3]jefri.pangaribuan@uph.edu

Abstract

According to the World Health Organization (WHO), hearing loss is one of the fourth highest causes of disability. The number of people with hearing loss continues to increase yearly. This increase occurred due to delays in recognizing the hearing loss, leading to delays in providing treatment. To solve this problem one solution to deal with this is early identification to detect the degree of hearing loss. This research will use machine learning to classify the degree of hearing loss. The algorithm implemented in this study is naive Bayes. This study uses a dataset from the open-access repository Zenodo with 3105 raw data and 19 features. This study evaluates the performance of overall accuracy, precision, recall, and f1-score and classified four classes: mild, moderate, moderately severe, and severe. The methodology classification stages in this study include data pre-processing, data training, data testing to evaluation. From evaluating the performance of the Naive Bayes algorithm, the classification results obtained the highest impacts in the form of 94% overall accuracy, 100% precision, 100% recall, and 97% f1-score in classifying the degree of hearing loss.

Keywords: *classification; hearing loss degrees; naive bayes*

1. Introduction

Hearing loss is one of the most common health problems in the world. The disorder is characterized by a decrease in individual hearing ability, which is identified through an increase in the hearing threshold starting at 20 dB and above based on the results of pure tone audiometry examination through air conduction in the essential frequency range for human hearing (250 Hz – 8000 Hz)[1].

Globally, the World Health Organization (WHO) reports hearing loss as the fourth highest factor in disability, with an estimated number of sufferers reaching 466 million in 2018. This figure is predicted to continue to grow rapidly to 630 million in 2030 to more than 900 million in 2050 without preventive measure[2].

Several studies have shown that individuals with hearing loss may experience limitations in terms of cognitive development, social interaction, verbal communication, and the implementation of daily activities independently, which makes the sufferer vulnerable to social isolation, psychological pressure, and difficulty in pursuing a particular field or profession. [3], [4].

From this exposure, it is known that hearing loss is a health problem that requires serious treatment. So, to be able to suppress the increase in the number of individuals with hearing loss in the future, appropriate interventions are needed to identify the presence of hearing loss from the start. One of the interventions that can be done is to determine or classify hearing loss based on its degree accurately with machine learning, according to the recommended standards of the GBD Expert Group on Hearing Loss.

Machine learning is the development of algorithms with the ability to learn independently from a given data sample without direct programming to find data patterns in making predictions or classifications[5],[6]. There are several algorithms in machine learning. This study will use Naive Bayes. There are several reasons for selecting Naive Bayes to solve the problem of classifying hearing loss. First, naive bayes requires a relatively small amount of training data to produce a good model, so it can be used even if the available dataset is limited. Second, Naive Bayes can also overcome the multicollinearity problem when the features used have a high relationship or correlation. This is important in classifying the degree of hearing loss because the parts used can be interrelated. Finally, Naive Bayes performs reasonably well in classifying data with high dimensions. This is important because

Accepted: 17-11-2022 | Received in revised: 05-06-2023 | Published: 12-08-2023

many features must be considered when categorizing the degree of hearing loss. [7],[8]. The acoustic features used in this study include sound intensity, duration, and the number of frequencies in the sound signal. The Naive Bayes algorithm uses these features to classify the degree of hearing loss with high accuracy.

With the advantages of being able to determine the parameters for classification with a small amount of training data, it is easy to implement and requires a short computational time for training[9]. The naive Bayes algorithm has been used in various fields, including medicine, to classify disease data[10], [11].

Previous research in machine learning using the Naive Bayes algorithm to classify colon cancer data obtained an overall accuracy rate of 95.24%, 100% precision, and 94% recall[12]. While other studies that classify cerebral infarction data using the Naive Bayes algorithm get an overall accuracy rate of 92.43%, a precision of 92%, and a recall of 92%[12].

The approach used in this research is a supervised learning method, specifically classification using the naive Bayes algorithm. The dataset used in this study was obtained from an open-access repository and pre-processed to ensure data quality. The data were then split into training and testing sets to train and evaluate the model's performance. The model was assessed by calculating overall accuracy, precision, recall, and f1-score. The classification was done into four classes: mild, moderate, moderately severe, and severe. The results obtained from this study show that the naive Bayes algorithm is effective in classifying the degree of hearing loss, with high overall accuracy, precision, recall, and f1-score. Overall, this study demonstrates the potential of machine learning in aiding the early identification and classification of hearing loss.

2. Research Methods

This research was conducted in several stages as follows:

2.1. Identification of Problems

In this research, the formulation of the problem to be solved is how to classify the degree of hearing loss using the Naive Bayes algorithm. In addition, this study will measure the performance of the Naive Bayes algorithm in classifying degrees of hearing loss based on the calculation results of overall accuracy, recall, precision, and fl-score according to the global burden of disease expert group recommendation standards.

2.2. Study of Literature

The second stage is conducting a literature study, namely collecting reference sources, and studying theories related to the research topic, especially regarding hearing loss, classification, machine learning, and the Naive Bayes algorithm from theses, books, and journals in the last five years.

2.3 Data Collection

In previous studies, there were various implementations of the amount of data and the number of features used for classification with naive Bayes. In the study of cerebral disease data classification using the Naive Bayes algorithm using 156 data and seven features and producing an overall accuracy rate of 92.43%, 92% precision, and 92% recall[12]. Another study for the classification of colon cancer with naive Bayes used 209 data and seven features to get performance results with an accuracy rate of 95.24%, 100% precision, and 94% recall[13]. Other research used to predict the recovery of Covid-19 patients in Indonesia through therapy using the Naive Bayes method using 367 data and producing an accuracy rate of 96.51%[14],[15].

The data collected in this study is hearing threshold data from the Zenodo website. Zenodo is a repository developed by OpenAIRE (Open Access Infrastructure for Research in Europe) and managed by CERN (European Organization for Nuclear Research). Zenodo provides open access to download various types of research papers and datasets.

The raw data collected is hearing threshold data for adults from 18 – 97 years, totaling 3105 data and 19 features described in Table 1.

Table 1. Dataset Features and their Descriptions

Dataset Features	Descriptions
Gender	Patient gender. (1: male, 2: female).
HearAids	Use of hearing aids (1: yes, 2: no).
Age	Patient age.
L_250_HTL	The hearing threshold of the left ear is at 250 Hz.
L_500_HTL	The hearing threshold of the left ear is at 500 Hz.
L_1000_HTL	The hearing threshold of the left ear is at 1000 Hz.
L_2000_HTL	The hearing threshold of the left ear is at 2000 Hz.
L_3000_HTL	The hearing threshold of the left ear is at 3000 Hz.
L_4000_HTL	The hearing threshold of the left ear is at 4000 Hz.
L_6000_HTL	The hearing threshold of the left ear is at 6000 Hz.
L_8000_HTL	The hearing threshold of the left ear is at 8000 Hz.
R_250_HTL	The hearing threshold of the right ear is at 250 Hz.
R_500_HTL	The hearing threshold of the right ear is at 500 Hz.
R_1000_HTL	The hearing threshold of the right ear is at 1000 Hz.
R_2000_HTL	The hearing threshold of the right ear is at 2000 Hz.
R_3000_HTL	The hearing threshold of the right ear is at 3000 Hz.
R_4000_HTL	The hearing threshold of the right ear is at 4000 Hz.

DOI: https://doi.org/10.29207/resti.v7i4.4683

Dataset Features	Descriptions
R_6000_HTL	The hearing threshold of the right ear is at 6000 Hz.
R_8000_HTL	The hearing threshold of the right ear is at 8000 Hz.

2.4. Preparation and Processing Data

The fourth stage is data preparation and processing, namely, changing raw data to be used as needed. The process includes the following:

Import libraries: Import the required Python libraries for this research, such as Numpy, Pandas, and Sklearn; Import datasets: The dataset file is transferred to a data frame to be displayed and processed in tabular form; Data cleaning: The data in the data frame is cleaned by removing duplicate data, error values, and new features; Labelling data: After the data cleaning process, the fourth stage is data labelling which begins with calculating the modified PTA6 (average hearing threshold at six frequencies: 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, 6000 Hz, 8000 Hz)[16] and then labeled the degree of hearing loss based on the standard recommendations of the GBD expert group, [17] which are listed in Table 2:

Table 2. Recommended Standard Degree of Hearing Loss GBD Expert Group.

Hearing loss	Threshold hear
Normal	-10.0 – 4.9 dB (*excellent hearing*) 5.0 – 19.9 dB (*good hearing*)
Mild	20.0 – 34.9 dB
Moderate	35.0 – 49.9 dB
Moderately severe	50.0 – 64.9 dB
Severe	65.0 – 79.9 dB
Profound	80.0 – 94.9 dB
Complete	≥ 95.0 dB
Unilateral	≤ 20.0 dB (*better ear*) ≥ 35.0 dB (*worse ear*)

Undersampling data: Perform data undersampling to equalize the amount of data from other classes with the number of minority class data; Data splitting: After the undersampling process, the data is separated from the labels and divided by 75% for training and 25% for testing.

2.5. Naive Bayes Classification

Naive Bayes is a classification algorithm in the supervised learning category. Naive Bayes works by finding or predicting opportunities for future events (posterior probability), namely the probability of a class of data that is not yet known based on information on past events (prior probability, likelihood, and evidence)[18].

In Naive Bayes, the probability calculation for classifying non-numeric data is done by Equation 1.

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \quad (1)$$

Where X is the feature value, *Ci* is the i-th data class, P(*Ci*|X) is the *posterior probability*, P(X|*Ci*) is the *likelihood*, P(*Ci*) is the prior probability, and P(X) is the *evidence*.

While calculations for numerical data are carried out using the Gaussian density in Equation 2.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X-\mu)^2}{2(\sigma)^2}} \quad (2)$$

Where σ is the standard deviation, μ is the mean (average), X is the feature value, π (pi) is 3.14, and e (Euler's number) is 2.718282.

2.6. Algorithm Performance Evaluation

Evaluation is carried out to determine the algorithm's performance in classifying data based on calculations of overall accuracy, recall, precision, and f1-score. The results of these calculations are displayed in the confusion matrix, a table containing the data classified as true positive, false positive, true negative, and false negative, as shown in Table 3.

Table 3. *Confusion Matrix*

	Predicted Values		
	Class	Positive	Negative
Actual Values	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

True Positive (TP) is the number of positive class data correctly predicted as a positive class. False Positive (FP) is the number of harmful class data incorrectly predicted as a positive class. True Negative (TN) is the number of negative class data correctly predicted as a negative class. Meanwhile, False Negative (FN) is the number of positive class data incorrectly predicted as a negative class.

Following are the details and equations for calculating overall accuracy, precision, recall, and f1-score. The overall accuracy in Equation 3 shows how accurate the algorithm is in correctly predicting the class of data.

$$\text{Accuracy} = \frac{\text{Correct classifications}}{\text{Total Data}} \quad (3)$$

Precision in Equation 4 shows the percentage of positive data that is correctly classified from all predicted data, either true positive or false positive.

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

Recall in Equation 5 shows the percentage of positive data that is correctly classified from all data that is positive.

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

DOI: https://doi.org/10.29207/resti.v7i4.4683

f1-score in Equation 6 shows the average harmonic value of recall and precision.

$$\text{f1}-\text{score}=\frac{2(\text{Precision}*\text{Recall})}{\text{Precision}+\text{Recall}} \quad (6)$$

3. Results and Discussion

3.1 Data Cleaning

The raw data collected is selected for its features according to research needs that classify the degree of hearing loss in the left ear. Therefore, features that are not needed, such as gender, use of hearing aids, the hearing threshold of the left ear at two frequencies (250 Hz, 3000 Hz) and the right ear at eight frequencies (250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 3000 Hz, 4000 Hz, 6000 Hz, and 8000 Hz) removed.

The features used in this study were obtained from this selection, namely age and left ear hearing threshold at six frequencies (500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, 6000 Hz, and 8000 Hz), as shown in Table 4.

Table 4. Research Data Features

	Age	Hearing Thresholds (Hz)					
		500	1000	2000	4000	6000	8000
0	45	0	10	20	35	20	5
1	58	105	95	110	115	110	111
2	23	120	120	115	115	110	105
3	68	45	55	65	55	75	80
4	43	15	20	45	50	70	75

Furthermore, data cleaning is carried out to remove error values and duplicate data. There are two error values in the hearing threshold data used in this study, 111 and 222, one of which is shown in Table 5.

Table 5. Data with *Error Values* 111

	Age	Hearing Thresholds (Hz)					
		500	1000	2000	4000	6000	8000
1	58	105	95	110	115	110	111
42	83	75	65	60	105	111	111
43	89	65	60	70	95	110	111
122	35	5	5	-5	-5	5	111
245	61	30	45	60	65	60	111

A thorough check is carried out before data cleaning to determine the amount of data that has these error values.

Table 6. Total Data Before Error Values Cleaning

Age	Hearing Thresholds (Hz)						
	500	1000	2000	4000	6000	8000	
F	F	F	F	F	F	F	3042
						T	57
						F	5
						T	1

From the inspection results in Table 6, there are values 111 and 222 in one or both of theL_6000_HTL and L_8000_HTL feature columns, a total of 63 data that need to be cleaned.

Table 7 is the result of re-examination, which gets a false value for all data. This means that there are no longer error values 111 and 222 after cleaning the data.

Table 7. Number of Data Error Values After Cleaning

Age	Hearing Thresholds (Hz)						
	500	1000	2000	4000	6000	8000	
F	F	F	F	F	F	F	3042

In addition to cleaning error values, duplicate cleaning data is also done. Table 8 shows that there are 24 rows of the same data, so to overcome this, only one row of information is taken from several existing duplicates, while other copies are deleted.

Table 8. Data Before Duplicate Values Cleaning

	Age	Hearing Thresholds (Hz)					
		500	1000	2000	4000	6000	8000
823	59	10	10	10	15	20	20
1208	45	5	5	5	15	15	10
1401	45	10	0	5	0	5	15
1875	59	10	10	10	15	20	20
1922	45	5	5	5	15	15	10
1954	30	5	5	5	5	10	10
1977	47	0	5	5	5	10	15
2028	42	0	5	5	5	5	20
2208	45	5	5	0	10	15	15
2227	40	5	0	5	0	10	5
2275	35	0	5	0	0	5	5
2571	40	0	5	10	20	30	45
2680	45	5	5	0	10	15	15
2687	45	0	5	10	0	15	20
2690	45	0	5	10	0	15	20
2724	40	5	0	5	0	10	5
2752	47	0	5	5	5	10	15
2785	34	0	5	0	5	5	15
2787	42	0	5	5	5	5	20
2792	35	0	5	0	0	5	5
2798	45	10	0	5	0	5	15
2920	34	0	5	0	5	5	15
2974	30	5	5	5	5	10	10
2984	40	0	5	10	20	30	45

From the cleaning results, 12 non-duplicate data were obtained, as shown in Table 9.

Table 9. Data After Duplicate Values Cleaning

	Age	Hearing Thresholds (Hz)					
		500	1000	2000	4000	6000	8000
823	59	10	10	10	15	20	20
1208	45	5	5	5	15	15	10
1401	45	10	0	5	0	5	15
1954	30	5	5	5	5	10	10
1977	47	0	5	5	5	10	15
2028	42	0	5	5	5	5	20
2208	45	5	5	0	10	15	15
2227	40	5	0	5	0	10	5
2275	35	0	5	0	0	5	5
2571	40	0	5	10	20	30	45
2687	45	0	5	10	0	15	20
2785	34	0	5	0	5	5	15

3.2 Data Labelling

After cleaning the data, data labeling is then carried out. It begins with calculating the modified PTA6 (average hearing threshold at six frequencies) and adding it as a data feature, as shown in Table 10.

DOI: https://doi.org/10.29207/resti.v7i4.4683

Table 10. Calculation and Addition of modified PTA6 Features

	Age	Hearing Thresholds (Hz)						PTA6
		500	1000	2000	4000	6000	8000	
0	45	0	10	20	35	20	5	15.0
2	23	120	120	115	115	110	105	114.2
3	68	45	55	65	55	75	80	62.5
4	43	15	20	45	50	70	75	45.8
5	45	70	80	95	110	110	105	95.0

Then, the data are given class labels mild, moderate, moderately severe, and severe based on the standard classification of hearing loss degrees according to the GBD expert group from the results of the modified PTA6 calculation in Table 11.

Table 11. Example of Data Labeling Results

	Age	PTA6	Class
0	45	15.0	Normal
2	23	114.2	Complete
3	68	62.5	Moderately severe
4	43	45.8	Moderate
5	45	95.0	complete

3.3 Undersampling Data

The data already labeled is checked for the amount per class.

Table 12. Number of Data per Class

Class	Data
Normal	1634
Mild	665
Moderate	423
Moderately severe	211
Severe	72
profound	14
complete	11

From Table 12, the regular class will be deleted because it is not a hearing loss. The other two categories, profound and complete, will also not be used in this study because the numbers are too small. In addition, it was found that the amount of data for each degree class of hearing loss needed to be uniform or had a class imbalance. To overcome this, data under sampling is carried out. Table 13 shows a comparison before and after under sampling the data.

Table 13. Total Data Before and After Under sampling

Before Under sampling		After Under sampling	
mild	665	mild	72
moderate	423	moderate	72
moderately severe	211	moderately severe	72
severe	72	severe	72

3.4 Data Splitting

Thus, the preprocessing stage to training and testing in this study was carried out with left ear hearing threshold data to classify the degree of hearing loss because the amount of left ear hearing threshold data (72 data) that can be used for training and testing purposes exceeds the right ear hearing threshold data (64 data) per class after checking, processing and under sampling data separately, as shown in Table 14.

Table 14. Comparison of Total Hearing Threshold Data for Left and Right Ears

Class	Data	Class	Data
mild	72	mild	64
moderate	72	moderate	64
moderately severe	72	moderately severe	64
severe	72	severe	64

So that after the data cleaning process was carried out in the form of cleaning error values (63 data), duplicate data (12 data), deletion of normal, profound, and complete classes (1659 data), undersampling data (1083 data) totaling 2817 data to a total of 3105 raw hearing threshold data left ear, a total of 288 training and testing data were obtained, which were divided into 72 data per class, namely mild, moderate, moderately severe and severe.

Classification of the degree of hearing loss was preceded by dividing the 288 data into 75% training data and 25% testing data using the train_test_split function before being processed with the Naive Bayes algorithm. In other words, there are 216 data used for training and 72 for testing. After classification, the performance evaluation of the Naive Bayes algorithm is carried out by building a confusion matrix and calculating the overall accuracy, precision, recall, and f1-score values.

3.5 Naive Bayes Performance Evaluation

Classification of hearing loss degrees from hearing threshold data with the Naive Bayes algorithm using 75% training data and 25% testing data to get the results described in the following confusion matrix table.

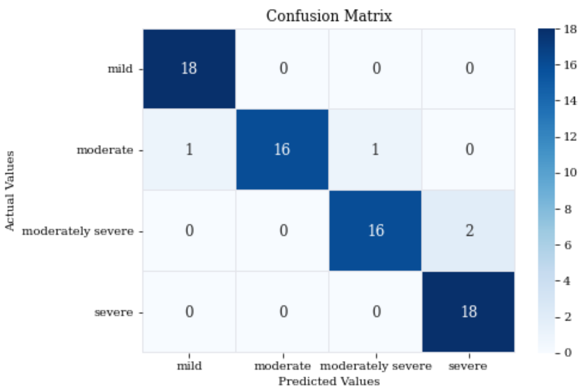

Figure 1. Confusion Matrix Results

The first row of the confusion matrix in Figure 1 shows that the Naive Bayes algorithm correctly classifies 18 mild data. In the second line, the algorithm correctly classifies 16 moderate data, and there are two misclassifications, namely 1 data classified as soft and one other as moderate. Likewise, with the third row, the algorithm classifies 16 moderately severe data correctly, and there are 2 data misclassifications as an intense class. The fourth row shows that the 18 powerful data are correctly classified.

The following is the number of True Positive, True Negative, False Positive, and False Negative obtained for each class: mild (TP = 18, TN = 53, FP = 1, FN = 0). Moderate (TP = 16, TN = 54, FP = 0, FN = 2). Moderately severe (TP = 16, TN = 53, FP = 1, FN = 2). Severe (TP = 18, TN = 52, FP = 2, FN = 0).

Thus, the calculation results of overall accuracy, precision, recall, and f1-score from the confusion matrix are summarized in Table 14.

Table 14 Summary of Confusion Matrix Calculation Results

Class	Overall accuracy	Precision	Recall	f1-score
mild	94%	95%	100%	97%
moderate		100%	89%	94%
moderately severe		94%	89%	91%
severe		90%	100%	95%

4. Conclusion

Based on the results of research that has been done, the conclusions obtained from this study are Classification of data on degrees of hearing loss for mild, moderate, moderately severe, and severe classes was carried out using the Naive Bayes algorithm with age features, hearing thresholds at frequencies of 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, 6000 Hz, and 8000 Hz, and PTA6 which is divided into 75% training data and 25% testing data totaling 288 data in total.

The overall accuracy results obtained by the Naive Bayes algorithm in classifying the four classes are 94%. While the precision results obtained were 95% (mild), 100% (moderate), 94% (moderately severe), and 90% (severe). Then, the recall results obtained were 100% (mild), 89% (moderate), 89% (moderately severe), and 100% (severe). Then, the f1-score results obtained were 97% (mild), 94% (moderate), 91% (moderately severe), and 95% (severe). The highest performance obtained by the Naive Bayes algorithm in classifying hearing loss degree data is 100% precision, 100% recall, and 97% f1-score.

Utilization of the results of this study can be developed into a model and implemented into a prototype for the classification of degrees of hearing loss so that it can help solve the main problems raised in this study. This research makes an essential contribution to doctors and audiologists in diagnosing and treating patients with hearing loss. Using the Naive Bayes algorithm to classify the degree of hearing loss, this study provides an accurate guide for doctors and audiologists in determining the correct type of treatment, making appropriate recommendations, monitoring patient progress, and evaluating treatment effectiveness. This improves the quality of healthcare services related to hearing loss and ensures that patients receive treatment appropriate to their condition.

Some development for this research can use hearing threshold data from frequencies of 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, 6000 Hz, and 8000 Hz. Future studies suggest using data with frequencies above 8000 Hz in classifying hearing loss for the higher frequency range. In addition, it can add the amount of hearing threshold data and increase the variations in the classification class of hearing loss degrees other than mild, moderate, moderately severe, and severe, namely profound and complete hearing loss.

Acknowledgement

This work is supported by LPPM (Lembaga Penelitian dan Pengabdian kepada Masyarakat) of Universitas Pelita Harapan.

References

[1] L. Marta, M.-S. P. A, P. E. M, M. Maximiliano, B. Riccardo, and P. Alessia, "Evaluation of Machine Learning Algorithms and Explainability Techniques to Detect Hearing Loss From a Speech-in-Noise Screening Test," *Am J Audiol*, vol. 31, no. 3S, pp. 961–979, Sep. 2022, doi: 10.1044/2022_AJA-21-00194.

[2] A. C. Davis and H. J. Hoffman, "Hearing loss: rising prevalence and impact," *Bull World Health Organ*, vol. v97(10), pp. 646-646A., 2019, Accessed: Dec. 13, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6796666/

[3] M. Zhang *et al.*, "A parsimonious approach for screening moderate-to-profound hearing loss in a community-dwelling geriatric population based on a decision tree analysis," *BMC Geriatr*, vol. 19, no. 1, pp. 1–11, 2019, doi: 10.1186/s12877-019-1232-x.

[4] Y. Zhao *et al.*, "Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study," *Ear Hear*, vol. 40, no. 3, pp. 690–699, 2019, doi: 10.1097/AUD.0000000000000649.

[5] O. P. Barus, J. Happy, Jusin, J. J. Pangaribuan, S. Z. H, and F. Nadjar, "Liver Disease Prediction Using Support Vector Machine and Logistic Regression Model with Combination of PCA and SMOTE," in *2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA)*, 2022, pp. 1–6. doi: 10.1109/ICTIIA54654.2022.9935879.

[6] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, 2019. doi: 10.1088/1757-899X/495/1/012033.

[7] H. Zaw, N. Maneerat, and K. Win, *Brain tumor detection based on Naïve Bayes Classification*. 2019. doi: 10.1109/ICEAST.2019.8802562.

[8] K. Wabang, Oky Dwi Nurhayati, and Farikhin, "Application of The Naïve Bayes Classifier Algorithm to Classify Community Complaints," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 872–876, Nov. 2022, doi: 10.29207/resti.v6i5.4498.

[9] A. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 7, p. 91, Jun. 2019, doi: 10.3991/ijes.v7i2.10659.

[10] D. Irawan, H. Oktavianto, M. K. Anam, T. Informatika, and U. M. Jember, "Analisis Penerapan Algoritma Naive Bayes," *JASIE "Jurnal Aplikasi Sistem Informasi Dan Elektronika,"* vol. 1, no. 2, pp. 127–134, 2019.

[11] A. B. Wibisono and A. Fahrurozi, "Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung

Koroner," *Jurnal Ilmiah Teknologi dan Rekayasa*, vol. 24, no. 3, pp. 161–170, 2019, doi: 10.35760/tr.2019.v24i3.2393.

[12] S. G. Fitri, R. Selsi, Z. Rustam, and J. Pandelaki, "Naïve bayes classifier models for cerebral infarction classification," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2020. doi: 10.1088/1742-6596/1490/1/012019.

[13] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/5/052068.

[14] O. Putra Barus and A. Tehja, "Prediksi Kesembuhan Pasien Covid-19 Di Indonesia Melalui Terapi Menggunakan Metode Naïve Bayes," *Journal Information System Development (ISD)*, vol. 6, no. 2, pp. 59–66, Jul. 2021, Accessed: Jan. 19, 2023. [Online]. Available: https://ejournal-medan.uph.edu/index.php/isd/article/view/460/267

[15] O. Putra Barus and T. Sanjaya, "Prediksi Tingkat Keberhasilan Pengobatan Kanker Menggunakan Imunoterapi Dengan Metode Naive Bayes," vol. 5, no. 1, Jan. 2020, Accessed: Jan. 19, 2023. [Online]. Available: https://ejournal-medan.uph.edu/index.php/isd/article/view/406

[16] B. O. Olusanya, A. C. Davis, and H. J. Hoffman, "Hearing loss grades and the International classification of functioning, disability and health," *Bull World Health Organ*, vol. 97, no. 10, pp. 725–728, Oct. 2019, doi: 10.2471/BLT.19.230367.

[17] A. Hijra Ferdinan, A. S. Briand Osmond, and C. S. Setianingsih, "Klasifikasi Emosi Pada Lirik Lagu Menggunakan Metode K-Nearest Neighbor Emotion Classification In Song Lyrics Using K-Nearest Neighbor Method," Dec. 2018. Accessed: Dec. 13, 2022. [Online]. Available: https://core.ac.uk/download/pdf/299924955.pdf

[18] R. Chandra, "Peningkatan Kinerja Algoritma Learning Vector Quantization (LVQ) Menggunakan Nguyen Widrow," *Repositori Institusi Universitas Sumatera Utara*, 2021. https://repositori.usu.ac.id/handle/123456789/38865 (accessed Dec. 13, 2022).

DOI: https://doi.org/10.29207/resti.v7i4.4683