Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Respiratory disease classification using selected data mining techniques

Abrahem P. Anqui *

College of Technology, Cebu Technological University-Naga Campus, City of Naga, Cebu, Philippines

ARTICLE INFO

Article history:
Received 25 February 2023
Received in revised form
11 June 2023
Accepted 12 June 2023

Keywords:
Lung cancer
Early detection
Linear discriminant analysis
Predictive accuracy
Medical diagnostics

ABSTRACT

Lung cancer, known for its high mortality rate, continues to claim numerous lives worldwide. Early detection has proven to offer significant advantages, substantially improving the prospects for successful treatment, medication, and the healing process. Despite various classification methods used to identify certain illnesses, their accuracy has often been suboptimal. In this paper, we employ Linear Discriminant Analysis (LDA) as a classifier and dimensionality reduction model to enhance the predictive accuracy of lung cancer presence. This study aims to predict the occurrence of lung cancer by utilizing a set of predictor variables, including gender, age, allergy, swallowing difficulty, coughing, fatigue, alcohol consumption, wheezing, shortness of breath, yellowish finger, chronic disease, smoking, chest pain, anxiety, and peer pressure. The goal is to enable early diagnosis, leading to timely and effective interventions. The results of our investigation demonstrate that LDA achieves an impressive accuracy rate of 92.2% in predicting lung cancer presence, surpassing the performance of the C4.5 and Naïve Bayes classifiers. This finding underscores the potential of LDA as a valuable tool for the early detection of lung cancer, ultimately contributing to improved patient outcomes. Through the utilization of LDA, we hope to advance the field of medical diagnostics and enhance the prospects for successful lung cancer management and treatment.

1. Introduction

Respiratory diseases, particularly lung cancer, have emerged as one of the predominant contributors to mortality on a global scale (Sung et al., 2021). Accurate and early diagnosis of certain illnesses in their early stage greatly prevents the worse symptoms and even prevents the disease from getting severe (Senturk, 2020). Prediction of future occurrences is a great step for any organization or individual to be prepared for whatever risk may arise (Petelin et al., 2023). Prediction is gaining a name in diverse areas such as medical and mental health diagnoses, agriculture, environmental fields (Babu et al., 2019), intrusion detection systems (Subba et al., 2015; Saranya et al., 2020), distributed denial of service attacks (DDoS) (Thapngam et al., 2012), and remote sensing (Cui et al., 2011). Data mining is widely used in prediction, classification, and clustering generally. It is the process of extracting information from a large data set and using that data set the algorithm will be able to produce useful outputs (Delima, 2019).

Discriminant Analysis, Regression, neural networks, and self-organizing maps are only a few of the many data mining methods that can be used to extract relevant information from a large database (Şuşnea, 2011). One of the most widely employed data mining techniques, primarily utilized for classification and prediction, is discriminant analysis. This method constitutes a supervised learning approach that aims to derive the discriminant function through regression analysis. The resulting discriminant function is subsequently employed to compute the anticipated value, thereby assigning it to a specific group. In discriminant analysis, the predicted variable (dependent variable) is consistently categorical, while the predictor (independent variable) exhibits inherent continuity. Linear discriminant analysis (LDA), predicated on linear regression, is a frequently adopted variant of discriminant analysis (Şuşnea, 2011). To ascertain the efficacy of the aforementioned classification techniques, the present study proceeds to compare the classification performance of LDA with that of the Naïve Bayes algorithm and C4.5. These two algorithms have been acknowledged for delivering

* Corresponding Author.
Email Address: abrahem.anqui@ctu.edu.ph
https://doi.org/10.21833/ijaas.2023.07.024
Corresponding author's ORCID profile:
https://orcid.org/0009-0000-1231-569X

optimal outcomes across various classification problems (Phoenix et al., 2021; Ponciano et al., 2015).

2. Review of related literature

Data mining also known as Knowledge Discovery in Database (KDD) is a method of unleashing novel (Osuna-Galán et al., 2022) and potentially valuable information from a huge amount of data (Han et al., 2011). Data mining techniques such as LDA, neural network, k-means clustering, and decision tree to name some contributory to extracting essential information from large data sets (Yadav and Pal, 2012). One field of data mining application is the field of Medical Mining, it is greatly essential and significant to medical practitioners to assist them in their decision-making, especially in critical situations such as in diagnosing a certain illness (Anguera et al., 2016). This emerging field is concerned with developing new methods that discover knowledge from data originating from the medical environment which eventually will have a great impact not only on the medical practitioners but, on the patients more importantly (Prather et al., 1997).

Discriminant analysis is one of the most commonly used techniques in data mining. It is gaining more and more attention in this big data era (Osuna-Galán et al., 2022). Discriminant analysis is generally used for determining the group membership of a particular subject. In general, the discriminant analysis started from formulating the equation using the regression concept, and this equation is then used as the discriminant function to determine the actual value of the prediction which is then classified into a particular group using the cut-off score computed using the data set (Taylor et al., 2022). The study of Al-Nasa'h et al. (2021) used discriminant analysis to examine the online learning efficacy of the students in relation to generalized anxiety and fear caused by the covid-19 virus. Two discriminant functions were subsequently involved one for academic engagement and the other one for the classification of fear of the covid-19 virus. The study found that those with high learning satisfaction have a moderate level of general anxiety and low fear of covid-19, and those with a low level of learning satisfaction have higher fear of covid-19 and a high level of general anxiety issue.

Moreover, Goyal and Mehta (2012) applied Naïve Bayes Algorithm and C4.5 Algorithm to assist or predict the student's performance in the board examinations using stanine, GWA, the honorary title received, scholarships grants, review center admission, and Licensure Examinations for Teachers (LET) variables. By comparing the results of these two algorithms, it showed an accuracy rate of 85.14% and 86.13%, respectively.

With the findings of this literature, it will be very useful and advantageous to accurately diagnose a certain type of cancer. This will enable the medical practitioner to prevent cancer from metastasizing to other parts of the body that become more difficult to cure and even the primary cause of cancer death (Vasudha Rani et al., 2022). In addition, to furtherly validate the accuracy of the algorithm's performance, it would be best to compare it to other algorithms.

3. Methodology

3.1. Data set

The data set for cardiovascular disease which was used in this study was from the Kaggle website (https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection) with 309 instances. The dataset has 16 variables where variables 1 to 12 are input features while variable 13 is output. The data set is divided into 70% as a training set and the remaining 30% was allocated for the testing set. The index description of the data set is shown in Table 1.

Table 1: Data set

No.	Variables	Type
1	Gender	Categorical; 1: No, 2: Yes
2	Age	Int-Continuous
3	Smoking	Categorical; 1: No, 2: Yes
4	Yellow fingers	Categorical; 1: No, 2: Yes
5	Anxiety	Categorical; 1: No, 2: Yes
6	Peer pressure	Categorical; 1: No, 2: Yes
7	Chronic disease	Categorical; 1: No, 2: Yes
8	Fatigue	Categorical; 1: No, 2: Yes
9	Allergy	Categorical; 1: No, 2: Yes
10	Wheezing	Categorical; 1: No, 2: Yes
11	Alcohol intake	Categorical; 1: No, 2: Yes
12	Coughing	Categorical; 1: No, 2: Yes
13	Shortness of breath	Categorical; 1: No, 2: Yes
14	Swallowing difficulty	Categorical; 1: No, 2: Yes
15	Chest pain	Categorical; 1: No, 2: Yes
16	Presence (or Absence) of lung cancer	Binary

3.2. Data processing

Prior to using the dataset as input to the LDA, several data preprocessing was conducted. The following steps were undertaken:

a. Remove insignificant parameter (ID)
b. Converting age from days to years
c. Variable recoding of the categorical independent variables and binary dependent variables

3.3. LDA

LDA is a variant of discriminant analysis that discriminate the membership of a particular subject based on the linear equation to a certain group.

3.3.1. Steps of LDA

A. Intercept equation: To get the intercept value, this equation must be used:

$$b_0 = \tilde{y} - b_1 \hat{X}$$

where, b_0 is the intercept; $\tilde{y}$ is the mean of the dependent variable; b1 is the slope value; $\hat{X}$ is the mean of the independent variable.

B. Slope equation: To get the slope value here's the slope equation:

$$\boldsymbol{b_1} = \frac{\sum(\boldsymbol{x_i} - \hat{X})(\boldsymbol{y_i} - \tilde{y})}{\sum(\boldsymbol{x_i} - \hat{X})^2}$$

where, b_1 is the slope value; x_i is the value of the independent variable; $\hat{X}$ is the mean of the independent variable; y_i is the value of the dependent variable; $\tilde{y}$ is the mean of the dependent.

C. Linear regression equation: Linear regression is the first step of discriminant analysis. It is used to predict the value of a variable (dependent) based on the value of another variable (independent/predictor). To do this we need to use the following equation:

$$Y = a + b_1X_i + b_2X_i \ldots b_iX_i$$

where, Y is the predicted value; a is the intercept; b_i is the slope; X_i is the respondent's score on the variable; i is the number of predictor variables.

D. Cut-off score: After the discriminant function is determined and the predicted values are computed, the cut-off score must be calculated to determine the membership of the values to a certain group by the following equation:

$$\boldsymbol{Z_c} = \frac{\boldsymbol{n_a z_b + n_b z_a}}{\boldsymbol{n_a + n_b}}$$

where, n_a=Number in Group $A(1)$; n_b=Number in Group $B(0)$; z_a=Centroid of Group $A(d1)$; z_b=Centroid of Group $B(d0)$.

E. Significance of regression coefficients: When the discriminant function has more than one independent variable, it is natural to determine whether each independent variable contributes significantly to the regression after the effects of other variables are taken into account.

3.4. Naïve Bayes algorithm

The naive Bayes Algorithm is a supervised learning algorithm based on Bayes Theorem which is primarily used for solving classification problems. Naïve Bayes Classifier predicts based on the probability of an object using the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where, $P(A|B)$ is the posterior probability; $P(B|A)$ is the likelihood probability; $P(A)$ is the prior probability; $P(B)$ is the marginal probability.

3.5. C4.5 classifier

The C4.5 Algorithm is primarily used for representing a valuable set of methods (Mapa et al., 2019). J48 classifier in Weka is used for generating the tree dependent on C4.5 calculations. This computation was built by Quinlan (1993) (Hussain et al., 2018), which is known as the statistical classifier because it generates decision trees that can be used for classification (Goyal and Mehta, 2012).

4. Results and discussion

In this analysis, the classification of lung cancer was undertaken using LDA and compared to other classification algorithms to test which algorithm performs well in classifying lung cancer based on given predictor variables. 10-Folds cross-validation of the data has been performed to verify each classifier.

Based on the data, 162 or 52.4% were males and the remaining 147, or 47.5% were females. Amongst the data 56.3% are smokers and 43.7% are non-smokers. Moreover, 55.7% of the data are said to consume alcohol while 44.3% do not take alcoholic beverages. These 2 factors are traditionally believed to be primarily contributory to lung cancer. These data metrics, in symptoms and causes features, identify that the dataset has a relatively good variation to be used in the lung cancer classification. Using 309 instances, the author observed, and examined the accuracy of the LDA in predicting lung cancer and compared the result to Naïve Bayes and C4.5 to determine and prove that LDA is a better performing algorithm in terms of disease classification. Fig. 1 shows that LDA correctly classified 92.6% of the test data, and results in 92.2% on cross-validation using Statistical Package for the Social Sciences (SPSS) software.

Classification Results

		Lung Cancer	Predicted Group Membership		Total
			0	1	
Original	Count	0	32	7	39
		1	16	254	270
	%	0	82.1	17.9	100.0
		1	5.9	94.1	100.0
Cross-validated[a]	Count	0	31	8	39
		1	16	254	270
	%	0	79.5	20.5	100.0
		1	5.9	94.1	100.0

a: Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case

Fig. 1: Classification result of LDA

Through the stepwise statistics in Fig. 2, the author identified the key features that greatly affect the diagnosis of lung cancer to wit: Allergy, swallowing difficulty, coughing, fatigue, alcohol consumption, yellowish fingers, chronic disease smoking, and peer pressure. LDA is also used in the

dimensionality reduction process to determine which variables are statistically significant to the prediction or classification (Li et al., 2021). This capability of LDA, helps the classification accuracy result become higher than other classifiers without a feature optimizer (Hossain et al., 2022).

Table 2 shows the results of the Naïve Bayes Algorithm with an accuracy rate of 89.32% which 276 instances were correctly classified and 33 are incorrectly classified while the C4.5 algorithm yields a 90.29% accuracy results of which 279 were correctly classified and 30 were not.

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	ALLERGY	.893	1	1	307.000	36.951	1	307.000	.000
2	SWALLOWING DIFFICULTY	.814	2	1	307.000	34.976	2	306.000	.000
3	COUGHING	.759	3	1	307.000	32.272	3	305.000	.000
4	FATIGUE	.734	4	1	307.000	27.579	4	304.000	.000
5	ALCOHOL CONSUMING	.697	5	1	307.000	26.324	5	303.000	.000
6	YELLOW FINGERS	.650	6	1	307.000	27.050	6	302.000	.000
7	CHRONIC DISEASE	.638	7	1	307.000	24.400	7	301.000	.000
8	SMOKING	.624	8	1	307.000	22.580	8	300.000	.000
9	PEERPRESSURE	.616	9	1	307.000	20.700	9	299.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.
a. Maximum number of steps is 30; b. Minimum partial F to enter is 3.84; c. Maximum partial F to remove is 2.71; d. F level, tolerance, or VIN insufficient for further computation

Fig. 2: Stepwise statistics

Table 2: Naïve Bayes and C4.5 results

Criteria	Naïve Bayes	C4.5
Accuracy	89.30%	90.29%
Correctly classified instances	276	279
Incorrectly classified instances	33	30

To test the efficacy of the LDA model for lung cancer prediction, the comparison of LDA to Naïve Bayes and C4.5 classifiers was undertaken. Fig. 3 shows that LDA performs better than the other two classifiers. Moreover, LDA has been found to effectively predict lung cancer occurrence using the enormous patient dataset (Pradeep and Naveen, 2018).

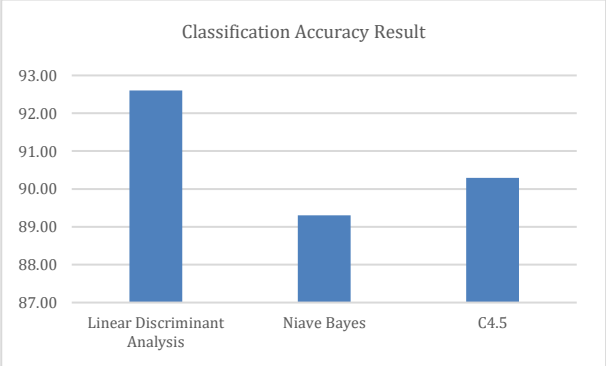

Fig. 3: Comparative classification accuracy

5. Conclusion and recommendation

The primary objective of this research is to employ LDA as a classifier to predict the occurrence of lung cancer based on multiple predictor variables. Concurrently, the study aims to identify the statistically significant variables that exert a significant influence on the prediction accuracy. The findings of this investigation demonstrate that LDA outperforms Naïve Bayes and C4.5 classifiers, achieving a remarkable prediction accuracy of 92.2%, compared to 89.32% and 90.29%, respectively, for the latter classifiers. These results underscore the efficacy of LDA as a potent tool for disease prediction. As a prospect for future research endeavors, it is recommended to explore the applicability of LDA in diverse datasets beyond the scope of this study. Furthermore, the incorporation of optimization algorithms, such as genetic algorithm and ant colony optimizer, could potentially enhance the performance and robustness of the predictive model. Such future investigations hold the potential to further advance the field of disease prediction and classification using sophisticated computational methodologies.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Al-Nasa'h M, Awwad FMA, and Ahmad I (2021). Estimating students' online learning satisfaction during COVID-19: A discriminant analysis. Heliyon, 7(12): e08544. https://doi.org/10.1016/j.heliyon.2021.e08544 **PMid:34909480 PMCid:PMC8662340**

Anguera A, Barreiro JM, Lara JA, and Lizcano D (2016). Applying data mining techniques to medical time series: An empirical case study in electroencephalography and stabilometry. Computational and Structural Biotechnology Journal, 14: 185-199. https://doi.org/10.1016/j.csbj.2016.05.002 **PMid:27293535 PMCid:PMC4887593**

Babu I, Balan RS, and Mathai PP (2019). Machine learning approaches used for prediction in diverse fields. International Journal of Recent Technology and Engineering, 8(2S4): 762-768. https://doi.org/10.35940/ijrte.B1154.0782S419

Cui M, Prasad S, Mahrooghy M, Bruce LM, and Aanstoos J (2011). Genetic algorithms and linear discriminant analysis based dimensionality reduction for remotely sensed image analysis. In the IEEE International Geoscience and Remote Sensing Symposium, IEEE, Vancouver, Canada: 2373-2376. https://doi.org/10.1109/IGARSS.2011.6049687

Delima AJP (2019). Predicting scholarship grants using data mining techniques. International Journal of Machine Learning and Computing, 9(4): 513-519. https://doi.org/10.18178/ijmlc.2019.9.4.834

Goyal A and Mehta R (2012). Performance comparison of Naïve Bayes and J48 classification algorithms. International Journal of Applied Engineering Research, 7(11): 1389-1393.

Han J, Kamber M, and Pei J (2011). Data mining: Concepts and techniques. 3rd Edition, Morgan Kaufmann Publishers, Burlington, USA.

Hossain MM, Swarna RA, Mostafiz R, Shaha P, Pinky LY, Rahman MM, and Iqbal MS (2022). Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease. Machine Learning with Applications, 9: 100330. https://doi.org/10.1016/j.mlwa.2022.100330

Hussain S, Dahan NA, Ba-Alwib FM, and Ribata N (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2): 447-459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Li CN, Shao YH, Chen WJ, Wang Z, and Deng NY (2021). Generalized two-dimensional linear discriminant analysis with regularization. Neural Networks, 142: 73-91. https://doi.org/10.1016/j.neunet.2021.04.030 **PMid:33984737**

Mapa JS, Sison A, and Medina RP (2019). A modified C4.5 classification algorithm: With the discretization method in calculating the goodness score equivalent. In the IEEE 6th International Conference on Engineering Technologies and Applied Sciences, IEEE, Kuala Lumpur, Malaysia: 1-4. https://doi.org/10.1109/ICETAS48360.2019.9117309

Osuna-Galán I, Pérez-Pimentel Y, and Aviles-Cruz C (2022). A novel 2D clustering algorithm based on recursive topological data structure. Symmetry, 14(4): 781. https://doi.org/10.3390/sym14040781

Petelin G, Cenikj G, and Eftimov T (2023). Towards understanding the importance of time-series features in automated algorithm performance prediction. Expert Systems with Applications, 213: 119023. https://doi.org/10.1016/j.eswa.2022.119023

Phoenix P, Sudaryono R, and Suhartono D (2021). Classifying promotion images using optical character recognition and Naïve Bayes classifier. Procedia Computer Science, 179: 498-506. https://doi.org/10.1016/j.procs.2021.01.033

Ponciano R, Pais S, and Casal J (2015). Using accuracy analysis to find the best classifier for intelligent personal assistants. Procedia Computer Science, 52: 310-317. https://doi.org/10.1016/j.procs.2015.05.090

Pradeep KR and Naveen NC (2018). Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4.5 and Naive Bayes algorithms for healthcare analytics. Procedia Computer Science, 132: 412-420. https://doi.org/10.1016/j.procs.2018.05.162

Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, and Hammond WE (1997). Medical data mining: Knowledge discovery in a clinical data warehouse. In the AMIA Annual Fall Symposium, American Medical Informatics Association, 101-105.

Quinlan JR (1993). C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, USA.

Saranya T, Sridevi S, Deisy C, Chung TD, and Khan MA (2020). Performance analysis of machine learning algorithms in intrusion detection system: A review. Procedia Computer Science, 171: 1251-1260. https://doi.org/10.1016/j.procs.2020.04.133

Senturk ZK (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. Medical Hypotheses, 138: 109603. https://doi.org/10.1016/j.mehy.2020.109603 **PMid:32028195**

Subba B, Biswas S, and Karmakar S (2015). Intrusion detection systems using linear discriminant analysis and logistic regression. In the Annual IEEE India Conference, IEEE, New Delhi, India: 1-6. https://doi.org/10.1109/INDICON.2015.7443533

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, and Bray F (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3): 209-249. https://doi.org/10.3322/caac.21660 **PMid:33538338**

Şuşnea E (2011). Data mining techniques used in on-line military training. In the 7th International Scientific Conference E-learning and Software for Education, Bucharest, Romania: 201-205.

Taylor C, Guy J, and Bacardit J (2022). Prediction of growth in grower-finisher pigs using recurrent neural networks. Biosystems Engineering, 220: 114-134. https://doi.org/10.1016/j.biosystemseng.2022.05.016

Thapngam T, Yu S, and Zhou W (2012). DDoS discrimination by linear discriminant analysis (LDA). In the International Conference on Computing, Networking and Communications, IEEE, Maui, USA: 532-536. https://doi.org/10.1109/ICCNC.2012.6167480

Vasudha Rani V, Das S, and Kundu TK (2022). Risk prediction model for lung cancer disease using machine learning techniques. In: Saini HS, Sayal R, Govardhan A, and Buyya R (Eds.), Innovations in computer science and engineering: Proceedings of the Ninth ICICSE: 417-425. Springer, Singapore, Singapore. https://doi.org/10.1007/978-981-16-8987-1_44

Yadav SK and Pal S (2012). Data mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal, 2(2): 51-56.