World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Check for updates

Evaluation of the performance of machine learning algorithms applied to voice parameters in prioritizing candidate for office laryngoscopy: Automated triaging in focus

Sanyaolu Alani Ameye [1], *, Mike Okuwe Ikoko [1], Michael Olusesan Awoleye [2] and Josephine Adetinuola Eziyi [1]

[1] *Department of Otorhinolaryngology, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria.*

[2] *African Institute for Science Policy and Innovation, Obafemi Awolowo University, Ile-Ife, Nigeria.*

World Journal of Advanced Research and Reviews, 2021, 11(01), 169–175

Publication history: Received on 18 May 2021; revised on 19 July 2021; accepted on 21 July 2021

Article DOI: https://doi.org/10.30574/wjarr.2021.11.1.0286

Abstract

Background: We examined the performance of different Machine Learning Algorithms while also comparing two methods of voice assessment to look at the workability of automating triage of patient that will need prompt office laryngoscopy.

Methods: We recruited consecutive adult subjects excluding those with a history of being regular singers or choristers in the past one year and those with the previous history of laryngeal trauma. We then carried out the perceptual voice assessments on the GRBAS Scale and also obtained the basic acoustic parameters of the voice samples. Laryngeal examinations with 70-degree Hopkins' Rod were then carried out by another examiner for all the participants to identify gross laryngeal changes or lesions. We then evaluated each machine learning algorithm comparing the perceptual and acoustic parameters in determining how well each algorithm predicts the presence of those categorized with having lesion or not by the laryngeal examination.

Results: One hundred and twenty respondents were analyzed out of which 89(74.2%) were females. The mean age was 46.5 ± 9.2 years. The perceptual evaluation generally outperformed the acoustic evaluation. Also, the Naïve Bayes Classifier (NBC) outperformed other algorithms with a F1 score of 0.55 followed by Artificial Neural Network (ANN) with the score of 0.53. However, the ANN outperformed the other with regards to the Area-under-the-curve (AUC).

Conclusion: When these metrics are taken together, the ANN still remains the best algorithm for this dataset. We are however cognisance of the needed improvement to the various aspects of this work including a larger dataset more scientific sampling.

Keywords: Machine Learning; Voice; Office Laryngoscopy; Larynx; Metrics; Triage

1. Introduction

This work though inspired by the ongoing Covid-19 pandemic sought to generally look at how the workflow of a typical otorhinolaryngology service can be automated in the area of triaging to identify patients those whose laryngeal examination need prioritization by using Machine Learning (ML)[1-3]. Voice changes is critically important because it could indicate the first clinical sign of laryngeal disease. Given the presence of voice changes, the main concern of an otorhinolaryngologist is often the screening for possible laryngeal cancer through laryngeal examination. Office

* Corresponding author: Sanyaolu Alani Ameye
Department of Otorhinolaryngology, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria.

laryngoscopy is an essential component in assessing these patients and expectedly, this procedure carries the potential risk of transmission of contagion which can be readily appreciated in the background of the ongoing pandemic[4].

We are not unaware of useful recommendations and options regarding the use of higher levels of Personal Protective Equipment(PPE) or screening of patients for COVID-19 before this examination is carried out to reduce this risk[5]. However, this work tries to look at an automated means of triaging so that the risk to a healthcare provider can be further minimized with regards to this pandemic and possible future similar situations[6].

Deploying ML in screening has been shown to engenders early diagnosis which expectedly should lead to early treatment and improved outcomes. [7] To classify subjects into those who have laryngeal lesions from those without, we used well-established Machine Learning (ML) algorithms[8] for the classification of voice samples. The algorithms considered were Random Forest Classifier (RFC), Logistic Regression (LR), Support Vector Machines (SVM), Artificial Neural Network (ANN), and Naïve Bayes Classifier (NBC). The theories behind these algorithms and other details like the underlying assumptions are beyond the scope of this work. However there are excellent resources such as the work by Novaković et al [9] in this regards.

The application of ML in medicine is not new and examples of usage include but not limited to Support Vector Machines (SVM) for breast cancer diagnosis[10] and lung cancer survivability prediction[11], Random Forest in the healthcare monitoring system and evaluation of safety culture, Naïve Bayes Classifiers, and C4.5 [11, 12] among others. With regards to voice, works on the automated detection of pathological voices using Machine and Deep Learning algorithms have been explored with successes [13, 14] .

In evaluating the success of ML, it is expedient to have well-defined and appropriate metrics because of their domain-specificity. [15, 16] [17] In the context of machine learning algorithms used in classification, a good number of the measures have been used extensively in the literature, some of which have been standardized [18] [11] [7] .

We evaluated our ML algorithms using Classification Accuracy (CA), Receiver Operator Characteristics/Area Under the Curve (AUC), F1-Score, Precision, and Recall. These metrics are derivatives of the Confusion Matrix (CM) generated by the classification results i.e. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [9].

2. Material and methods

The data used in this paper were collected pre-pandemic. After obtaining the approval of the Ethical Committee of the parent institution and obtaining the consent of subjects, consecutive subjects between 18 and 60 years of age were recruited into this study. We excluded subjects who have been regular singers or choristers in the past year and patients with a previous history of laryngeal trauma. Two independent assessments of the voice scoring on the GRBAS (Grade, Roughness, Breathiness, Asthenia, Scale) Scale[19] were carried out in a quiet room (the departmental audiology section) with ambient noise less than 45dB. With the subjects comfortably seated, the speech therapist and one of the authors obtained independently the Perceptual Parameters (PP). These scores were later compared for inter-rater agreement using the Fleiss Kappa (κ) test.

Following this, an acoustic examination was performed in the same quiet room. The subject is asked to sustain an /a/ for at least 3 seconds and this is recorded (sampling frequency =44100Hz) with the low impedance head-mounted microphone at 45^0 and positioned 8cm from the angle of the mouth and connected to a computer laptop with PRAAT© software for the recording and extraction of the Acoustic Parameters (AP) from voice sample. The AP extracted were Fundamental Frequency (F0), Shimmer, Jitter, and Harmonic-to-Noise Ratio (HNR).

Finally, a laryngeal examination was carried out using a 70-degree Hopkins rod to examine the larynx looking out for gross changes such as mucosa edema, erythema, singers' nodules, polyps, and contact ulcers. The presence of one or more of these lesions is scored as positive for laryngeal lesions for the purpose of classifying whether a subject had laryngeal lesion or otherwise.

For each of the methods of voice assessment, the ML algorithms presented below were applied with five-fold cross-validation to avoid overfitting. Variables entered for the acoustic parameters' assessments were the F_0, Shimmer, Jitter and HNR while for the perceptual parameters' assessment, the individual components of the GBRAS score were used. Other details of the parameters which are peculiar to some algorithms are presented in Table 1.

Table 1 Specific parameters applied for some of the ML Algorithms used

ML Algorithm	Parameters
Logistic Regression	Regularization Type: Ridge(L2)
Neural Network	Neurons in Hidden Layers: 100
	Activation: Rectified Linear Unit (ReLU)
	Solver: Adam
	Regularization: α=0.0001
	Maximum number of Iterations: 200
Random Forest	Number of Trees: 10
Support Vector Machine	Cost(C): 1.00
	Regression Loss Epsilon(ε): 0.10
	Kernel: Sigmoid
	Numerical Tolerance: 0.0015
	Iteration Limit: 100

For each application of the ML algorithms, we evaluated the performance by obtaining the CA, Precision, Recall from the CM. The ROC was also plotted and AUC calculated. The results were presented in tables and charts.

All the analysis and ML algorithms were run in the Python© Data Science Environment with requisite packages loaded.

3. Results and discussion

Out of the one hundred and twenty respondents analyzed in the study, eighty-nine (74.2%) were females while the remaining 31(25.8%) were males. Table 2 below shows the other characteristics of the dataset.

The assessments of perceptual parameters showed very good interrater agreement (κ=0.821) which shows good interrater agreement.

Table 2 Summary of the Demographic data, clinical data, and voice parameters

Variables	Subject N =120
Gender	
Male	31(25.8%)
Female	89 (74.2%)
Mean (SD) Age (years)	46.5(9.2)
Age range	Frequency
20-29	12
30-39	16
40-49	38
50-59	54
Presence of Laryngeal lesion	Yes: 21(17.5%), No: 99 (82.5%)
Reinke's Oedema	8
Nodules	5
Hyperaemia	6
Polyps	2

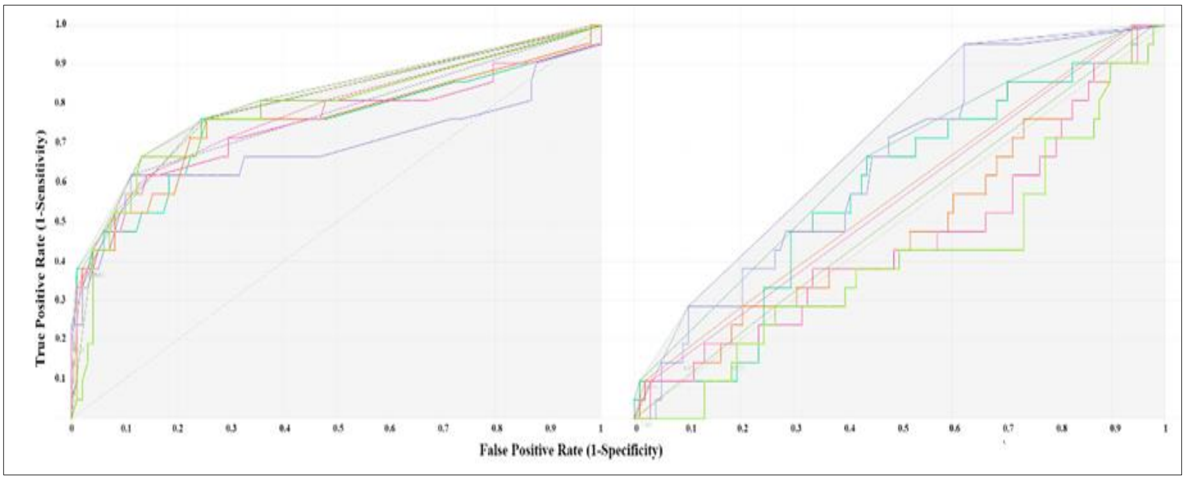

Figure 1 The ROC curves of Perceptual Parameters (left) and Acoustic Parameters (right) with the more convex shape for the Perceptual Parameters' ML Classifications Outputs

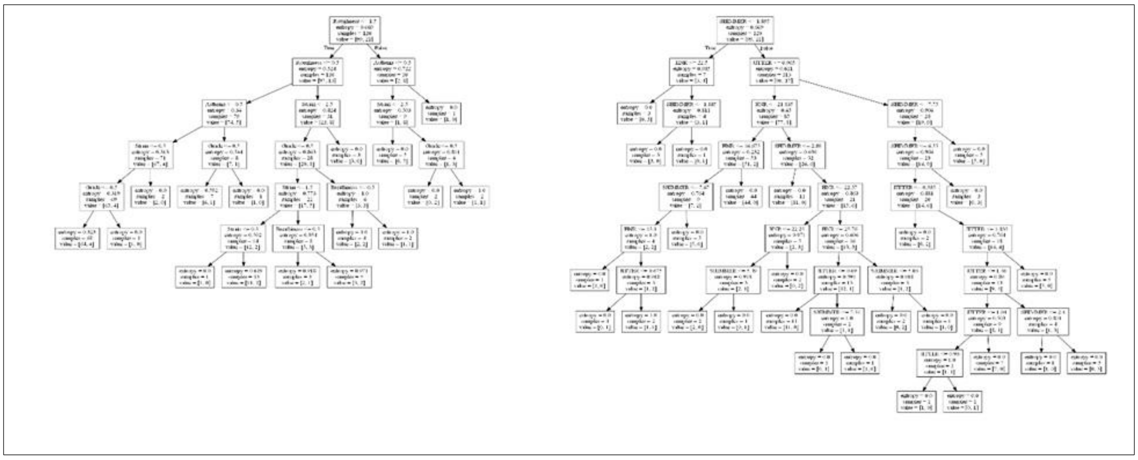

Figure 2 The sample decision trees from the Random Forest Classifiers for the Perceptual Parameters-PP (left) and Acoustic Parameters- AC (right) depicting a shallower depth for the PP ML decision tree graphical output

The essence of this paper is to see the performance of the ML algorithms in automating the triaging process of the patients whose need for laryngeal visualization should be prioritized. Attempt to use gender as a parameter lead to the failure of convergence in some of the algorithms and hence it was dropped as a parameter. This is understandable since our data is skewed towards the female gender. An obvious improvement in recruitment will be to have equal numbers of gender categories.

The evaluation metrics show better performance generally for all the ML algorithms when applied to perceptual parameters of voice compared with the acoustic parameters of voice. At first blush, it may seem that human subjective assessment as represented by the GBRAS evaluation is still better than the objective acoustic evaluation in identifying those with lesion. This performance difference is easily apparent from the more convex shape of the AUC figures for the two modes of evaluation. Similarly, a sample decision tree plot selected from the RFC showed less depth for perceptual compared with the acoustic analyses. However, we have to bear in mind the fact that the vocal sampling for the acoustic evaluation was just a sustained vowel /a/ whereas the sampling done for perceptual assessment was a whole sentence. Studies have shown better sampling with whole sentences, as we applied to the perceptual evaluation than the sustained vowel for the acoustic sampling[20].

The Area Under Curve (AUC) is a common and robust metrics for evaluation. The AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. The AUC is closely linked to two basic terms of classifications i.e., Sensitivity (True Positive Rate) and the False Positive Rate. So, AUC is the area under the curve of plot of the False Positive Rate versus the True Positive Rate at different points. Using the performance of the models by the area under the ROC curve (AUC) which are based on the gradation such that 0.9–1 is excellent, 0.8-0.9 is good; 0.7–0.8 is fair, 0.6–0.7 is, 0.50–0.6 is fail, [21, 22] it is only the NBC when applied to the perceptual parameters that crossed the threshold of good assessment.

Table 3 Performance scores of the different metrics when applied to the different NL algorithms and voice evaluation methods

Machine Learning Algorithm	Perceptual Parameters	Acoustic Parameters
Logistic Regression Classifier		
CA (%)	86.7	83.3
AUC (%)	76.5	60.0
F1	0.5	0.17
Precision	0.73	0.67
Recall	0.38	0.10
Random Forest Classifier		
CA (%)	85.0	80.8
AUC (%)	75.5	60.1
F1	0.47	0.26
Precision	0.61	0.40
Recall	0.38	0.19
Support Vector Machine Classifier		
CA (%)	87.5	81.7
AUC (%)	73.2	53.0
F1	0.48	0.08
Precision	0.88	0.33
Recall	0.33	0.05
Artificial Neural Network		
CA (%)	87.5	82.5
AUC (%)	76.2	49.7
F1	0.55	0.16
Precision	0.75	0.50
Recall	0.43	0.10
Naïve Bayes Classifier		
CA (%)	82.5	70.0
AUC (%)	80.4	40.6
F1	0.53	0.14
Precision	0.50	0.14
Recall	0.57	0.14

All algorithms across the two evaluation methods showed impressive Classification Accuracy (CA) above 80% except when Naïve Bayes Classifier was applied to acoustic parameters. This is however not unexpected, considering that CA tends to give deceptively high results when there are unequal numbers in the classes[23] as in this study where only

17.5% of the subjects were found to have a laryngeal lesion. If we were to base our choosing the optimal algorithm on this metric only, there will be a great problem in the cost of missing those with a laryngeal lesion.

Precision, is the ratio of *TP* to all the positives predicted by the model. When the precision is low, it means more FP are predicted by the model. The Recall (Sensitivity) is the ratio of TP to all the positives in your Dataset. Low recall implies that you there is more FN predicted by the model. The concern between precision and recall is that high precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. [24]This concern is easily handled by another metric known as the F1-Score, the Harmonic Mean between precision and recall. The range for F1 Score is from 0 to 1 with 0 being a useless model and 1 being a perfect model. [24, 25] Our assessment showed that with regards to F1, the ANN and NBC performed better than the other algorithms when applied to the PP. Again, the performance is generally dismal for AP.

We appreciate the bias towards the female gender which we attribute this to the cohort from which this data emanated i.e., the hospital staff and teachers. We are also mindful of the fact that some ML perform better with larger or smaller sample sizes. For example, a study conducted by Jollans *et al.* quantifies the performance of various machine learning for neuroimaging data. The research employed some machine learning models showed that with a dataset of at least 400 observations and 400 or more predictor variables, Elastic Net was noted to have shown the best analysis approach for ROI data. The study also found that for a smaller size, multiple regression seems to show the best result when supported by bootstrap aggregation [12].

It is also important to point out also that though this data was collected pre-pandemic, at a time which all patients enjoyed close-quarter interaction with practitioners with just minimal PPE (face mask), we hope that further works expand on this pilot and look at the utility of remotely acquired vocal samples which will not require close interaction with practitioners thus reducing the risks of transmitting contagions.

4. Conclusion

The ML algorithm shows some promise in classifying patients with laryngeal lesions and the NBC when applied to perceptual evaluation parameters generally outperformed the other algorithms especially when the two most consistent metrics (AUC and F1-score). However, the ANN slightly outperforms NBC when F1-score was considered alone. We are mindful of several improvement needed for this work which include but not limited to a better sampling of subjects.

Compliance with ethical standards

Acknowledgments

We wish to show our appreciation to Mrs G. M. Ogunniyi, the Departmental Speech Therapist.

Disclosure of conflict of interest

None.

Statement of informed consent

The researcher obtained Ethical Approval from the Institutional Review Board and other relevant authorities. Informed consent was obtained from all individual participants included in the study.

References

[1] Patterson JM, Govender R, Roe J, Clunie G, Murphy J, Brady G, et al. COVID-19 and ENT SLT services, workforce and research in the UK: A discussion paper. International journal of language & communication disorders. 2020;55(5):806-17.

[2] Liu Y, Yan L-M, Wan L, Xiang T-X, Le A, Liu J-M, et al. Viral dynamics in mild and severe cases of COVID-19. The Lancet Infectious Diseases. 2020;20(6):656-7.

[3] Noh JY, Yoon JG, Seong H, Choi WS, Sohn JW, Cheong HJ, et al. Asymptomatic infection and atypical manifestations of COVID-19: Comparison of viral shedding duration. The Journal of Infection. 2020.

[4] Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. New England Journal of Medicine. 2020;382(12):1177-9.

[5] Tysome JR, Bhutta MF. COVID-19: protecting our ENT workforce. Wiley Online Library; 2020.

[6] Wang Q, Wang X, Lin H. The role of triage in the prevention and control of COVID-19. Infection Control & Hospital Epidemiology. 2020;41(7):772-6.

[7] Kim H, Jeon J, Han YJ, Joo Y, Lee J, Lee S, et al. Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy. Journal of Clinical Medicine. 2020;9(11):3415.

[8] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007;160(1):3-24.

[9] Novaković JD, Veljović A, Ilić SS, Papić Ž, Milica T. Evaluation of classification models in machine learning. Theory and Applications of Mathematics & Computer Science. 2017;7(1):39–46-39–46.

[10] Wang H, Zheng B, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research. 2018;267(2):687-99.

[11] Pradeep K, Naveen N. 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. 2018;132:412-20.

[12] Jollans L, Boyle R, Artiges E, Banaschewski T, Desrivières S, Grigis A, et al. Quantifying performance of machine learning methods for neuroimaging data. NeuroImage. 2019;199:351-65.

[13] Fang S-H, Tsao Y, Hsiao M-J, Chen J-Y, Lai Y-H, Lin F-C, et al. Detection of pathological voice using cepstrum vectors: A deep learning approach. Journal of Voice. 2019;33(5):634-41.

[14] Chuang Z-Y, Yu X-T, Chen J-Y, Hsu Y-T, Xu Z-Z, Wang C-T, et al., editors. Dnn-based approach to detect and classify pathological voice. 2018 IEEE international conference on big data (big data); 2018: IEEE.

[15] Mishra D, Gunasekaran A, Papadopoulos T, Dubey R. Supply chain performance measures and metrics: a bibliometric study. Benchmarking: An International Journal. 2018.

[16] Awoleye OM, Ilori OM, Oyebisi TO. Sources of innovation capability and performance of ICT agglomerated MSMEs in Nigeria. International Journal of Innovation Management. 2020;24(04):2050032.

[17] Bedford DS, Bisbe J, Sweeney B. Performance measurement systems as generators of cognitive conflict in ambidextrous firms. Accounting, Organizations and Society. 2019;72:21-37.

[18] Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Computer Communication Review. 2006;36(5):5-16.

[19] Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurements. Journal of voice. 2004;18(3):299-304.

[20] Moon KR, Chung SM, Park HS, Kim HS. Materials of acoustic analysis: sustained vowel versus sentence. Journal of Voice. 2012;26(5):563-5.

[21] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering. 2005;17(3):299-310.

[22] Ling CX, Huang J, Zhang H, editors. AUC: a Statistically Consistent and more Discriminating Measure than Accuracy. IJCAI; 2003.

[23] Koyejo O, Natarajan N, Ravikumar P, Dhillon IS, editors. Consistent Binary Classification with Generalized Performance Metrics. NIPS; 2014: Citeseer.

[24] Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:201016061. 2020.

[25] Goutte C, Gaussier E, editors. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. European conference on information retrieval; 2005: Springer.