



The Potential of Machine Learning Algorithms in Discriminating Chronic Obstructive Pulmonary Disease and Healthy Saliva Samples

Atefeh Goshvarpour¹ , Ateke Goshvarpour^{2,3*}

¹Department of Biomedical Engineering, Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran.

²Department of Biomedical Engineering, Imam Reza International University, Mashhad, Razavi Khorasan, Iran.

³Health Technology Research Center, Imam Reza International University, Mashhad, Razavi Khorasan, Iran.

Abstract

Background: Today, with the spread of tobacco use and increased environmental pollutions, respiratory diseases are considered important factors threatening human life. Chronic obstructive pulmonary disease (COPD) is a kind of inflammatory lung disease. Clinically, COPD is currently diagnosed and monitored by spirometry as the gold-standard technique although spirometry systems encounter some limitations. Thanks to the economical handling and sampling, practicality, and non-invasiveness of saliva biomarkers, it is promising for the testing environment. Accordingly, the current analytic observational study aimed to propose an intelligent system for COPD detection.

Materials and Methods: To this end, 40 COPD (8 females and 32 males in the age range of 71.67 ± 8.27 years) and 40 controls (17 females and 23 males within the age range of 38.23 ± 14.05 years) were considered in this study. The samples were characterized by absolute minimum value and the average value of the real and imaginary parts of saliva permittivity. Additionally, the age, gender, and smoking status of the participants were determined, and then the performance of various classifiers was evaluated by adjusting k in k -fold cross-validation (CV) and classifier parameterization.

Results: The results showed that the k -nearest neighbor outperformed other classifiers. Using both 8- and 10-fold CV, the maximum classification rates of 100% were achieved for all k values. On the other hand, increasing the k in k -fold CV improved classification performances. The positive role of parameterization was revealed as well.

Conclusions: Overall, these findings authenticated the potential of machine learning (ML) algorithms in the diagnosis of COPD using subjects' saliva features and demographic information.

Keywords: Chronic obstructive pulmonary disease, Saliva, Machine learning, Diagnosis

*Correspondence to

Ateke Goshvarpour,
Department of Biomedical
Engineering, Imam Reza
International University,
Mashhad, Razavi Khorasan,
Iran.
Tel: +98 5138041,
Ext. 3131,
Email: ak_goshvarpour@
imamreza.ac.ir



Received: July 30, 2021, Accepted: October 4, 2021, ePublished: December 30, 2021

Introduction

Respiratory diseases are now considered as important factors threatening human life with the spread of tobacco use and an increase in environmental pollutions. Chronic obstructive pulmonary disease (COPD) is a type of inflammatory lung disease regarded by long-term breathing complications and deficient airflow which is not fully reversible (1). In addition, it is a progressive disease with shortness of breath and cough with sputum production as the key symptoms (1). Although smoking tobacco is the main cause of COPD, several other factors such as environmental stimuli and genetics are also identified in this regard (2).

Generally, 25% of adults aged 40 years and above have moderate airflow obstruction (3). Currently, COPD is the fourth leading cause of death in the world, but the World Health Organization predicts that it will become the third

leading cause by 2030 (4).

COPD is now diagnosed and monitored by spirometry as the gold-standard technique conducted in primary and secondary care (5). Despite recent developments in spirometry systems, they encounter some limitations. Environmental restrictions include not being able to respond to sudden temperature changes, patient constraints not recommended for patients with heart complications or dementia, and systematic errors due to a false start of the test, cough or hesitation during exhalation, lack of system calibrations, or leakage, and the like (6).

Recent proteomics experiments have suggested that protein biomarkers can help accurately diagnose and predict disease outcomes. A recent study by Santos et al (7) has presented the progress of research in this area toward the management of COPD. Reviewing the literature,

they showed that the biomarkers of protein in the saliva and sputum are promising for the point-of-care (PoC) testing environment owing to their economical handling and sampling, practicality, and non-invasiveness. Due to the complexities of non-invasive daily sputum sampling, saliva is preferred for PoC programs with better patient compliance (8). However, accurate and inclusive diagnosis of the disease requires the provision of other information from the patient’s medical history and demographic data. The smoking status, weight, age, gender, cytokine level, and pathogen load are some important issues in this regard (9-11).

Human saliva involves a composition of water and important substances (e.g., electrolytes), and organic molecules (12). These substances originate from the serum. On the account of the intracellular passive diffusion in the capillary bed and the osmotic gradient, they come into the saliva. Consequently, most of them are representative of systemic diseases. Blicharz et al (13) evaluated the saliva samples of COPD and asthmatic volunteers and highlighted that interleukin 8 (IL-8) and IL-6 were significantly higher in COPD patients compared to the controls. The results of the study by Dillon et al. (14) on healthy subjects showed no correlation between blood and saliva C-reactive protein (CRP) levels in healthy subjects. This finding is not consistent with the results of Patel et al (15), demonstrating a strong relationship between salivary serum counterparts and CRP. This conclusion was supported by the experimental study performed by Bhavsar et al (16), confirming a positive association between saliva CRP and serum in COPD participants. Ji et al. (17) also reported that depending on disease severity, the salivary matrix metalloproteinase-9 and IL-8 activity altered in COPD patients with a negative correlation between them and lung function. Concentration changes in other salivary substances such as tumor necrosis factor- α and norepinephrine have been previously studied as well (15, 17).

In the field of health care, the digital diagnosis

of a disease is possible using machine learning (ML) algorithms, which can alert physicians of any abnormalities by detecting specific disease patterns in the electronic patient health record. It is noteworthy that the computational speed increases while some human errors reduce by utilizing these intelligent systems based on ML algorithms. A limited number of studies have so far focused on this issue. For instance, Zarrin et al (10) evaluated some algorithms including the artificial neural network (ANN), Gaussian Naïve Bayes, support vector machine (SVM), linear regression (LR), and a decision tree (DT) algorithm (XGBoost) for the classification of saliva dielectric characteristics of COPD and healthy controls. A hidden layer with four neurons and a sigmoid activation function were implemented for ANN. The SVM was realized using a radial basis function (RBF) kernel, and LR was optimized by implementing a limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm. Their results revealed a maximum classification accuracy of 91.25% and sensitivity of 100% using the XGBoost gradient boosting algorithm (10). Although satisfactory results were presented by studying some classification approaches in this research, more studies are still needed in this field.

In this respect, the current study sought to evaluate the performance of several ML algorithms in the classification of the saliva samples of COPD and healthy volunteers. For this purpose, the works by Zarrin et al (10, 11) were extended by implementing different structures of the DT algorithm, Feed-forward neural network (FNN), probabilistic neural network (PNN), layer recurrent neural network (LRNN), Elman neural network (ENNN), a generalized regression neural network (GRNN), different ensemble methods (EMs), k-nearest neighbor (kNN), SVM, and Naive Bayes (NB). Figure 1 shows the proposed diagnostic framework. The organization of the paper is as follows.

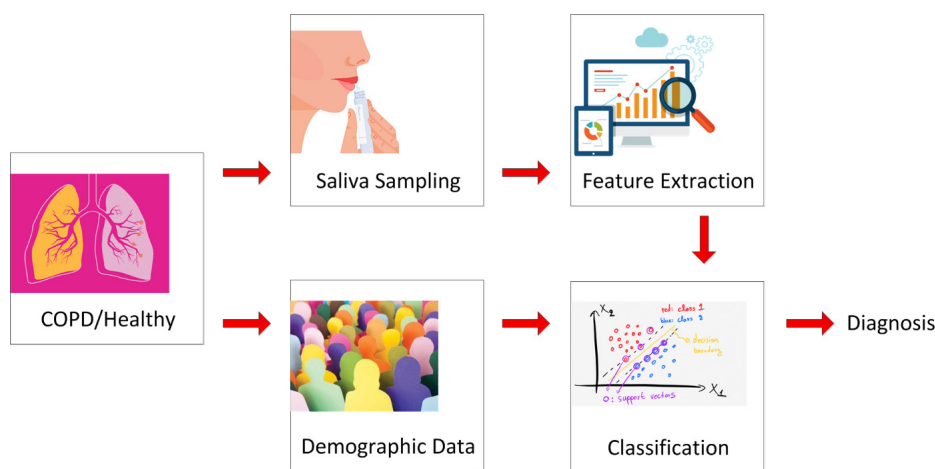


Figure 1. The Suggested Framework. Note. COPD: Chronic obstructive pulmonary disease.

Materials and Methods

Data

The present system was evaluated on the publicly available data, namely, the Exasens dataset (9-11). The database includes some attributes of 4 sample groups. The first group (I) contains 40 samples of outpatients and hospitalized patients with COPD without acute respiratory infection, and the second one (II) encompasses 10 samples of outpatients, along with hospitalized patients with asthma without acute respiratory infections (Asthma). In addition, the third group (III) involves 10 samples of patients with respiratory infections while without COPD or asthma (Infected). The last group (IV) includes 40 samples of healthy controls without COPD, asthma, or any respiratory infection (HC). The attributes of COPD and HC groups were used in this study.

The saliva samples were characterized by four measures including the absolute minimum value (Min. (Δ)) and the average value (Avg. (Δ)) of real and imaginary parts of saliva permittivity. Additionally, the demographic information of the subjects was obtained and provided, including the age, gender, and smoking status of the participants.

Figure 2 illustrates the frequency of the saliva samples of COPD and controls concerning the demographic information and Figure 3 depicts the distribution of saliva permittivity features in the two groups.

Classification

Feed-Forward Neural Network

The most classic kind of the developed ANN was the FNN, in which three layers are incorporated, including the input, hidden, and output layers. Without any loop or cycle, the information moves in the only forward direction from the first to the last layer.

In the current study, the ANN was trained using the Levenberg-Marquardt back-propagation algorithm. Further, the performance of the classifier was tested using different numbers of neurons (i.e. 2-10) in the hidden layer.

Elman Neural Network

ENN is an extended version of FNN. A context layer is added to the hidden layer of the FNN as a time delay operator. This structure helps memorize and thus makes a time-varying characteristic. This network belongs to the family of recurrent neural networks (RNN).

Similar to FNN, Levenberg-Marquardt back-propagation function was used to train the classifier in this study. Furthermore, the classifier was implemented by applying different numbers of neurons (i.e., 2-10) in the hidden layer with tap delays of 1:2.

Layer Recurrent Neural Network

LRNN is also a kind of FNN, in which each layer has a

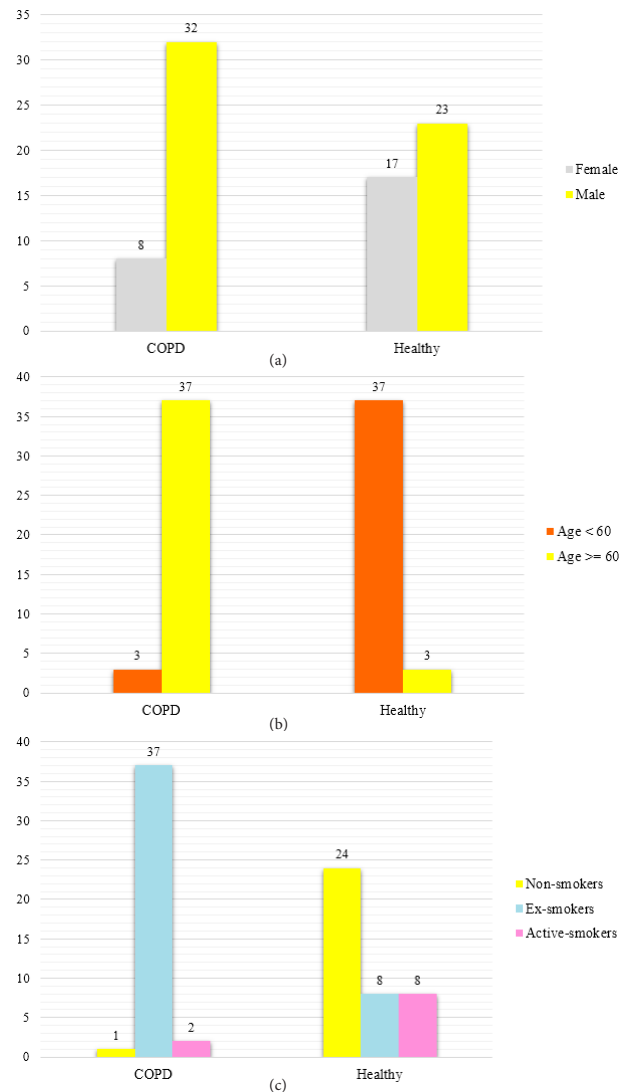


Figure 2. Frequency of the saliva samples of COPD and controls with respect to demographic information: (a) Gender, (b) Age, and (c) Smoking status. Note. COPD: Chronic obstructive pulmonary disease.

recurrent connection. Similar to ENN, the Levenberg-Marquardt back-propagation training function and tap delays of 1:2 were applied, and then the classifier was assessed using different numbers of neurons (i.e., 2-10) in the hidden layer.

Generalized Regression Neural Network

GRNN incorporates a radial basis layer and a specific linear layer. The former calculates weighted inputs with “Euclidean distance weight function” and net input by combining its weighted inputs and biases. The latter has a linear transfer function, which calculates weighted input with “normalized dot product weight function”, and net inputs by combining its weighted inputs and biases. It is noted that only the first layer has biases in this classifier. The first layer biases are all set to 0.8326/spread. Different spread values in the range of 0.3-1 with the step size of 0.05 were examined in the current study, and the best results were obtained in this regard.

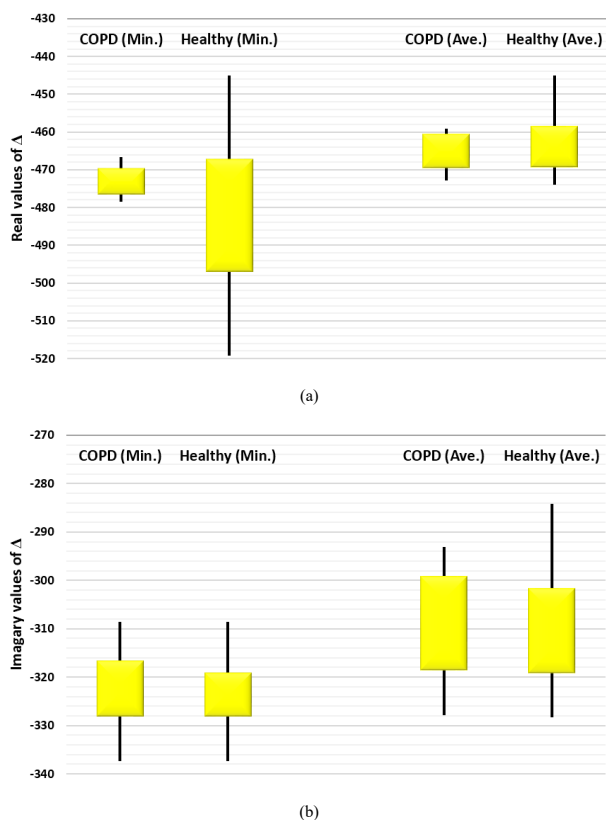


Figure 3. The saliva permittivity features in the two groups of COPD and healthy: (a): Real and (b) Imagery values
 Note. COPD: Chronic obstructive pulmonary disease.

k-Nearest Neighbor

kNN belongs to the family of non-parametric and supervised classifiers. This procedure incorporated a well-defined distance calculation. Likewise, the new sample query outcome is classified relying upon the contiguity of the *k*-nearest samples easily reached in the feature space. The exceptional performance is obtained by allocating a short period to training (18, 19). Different *k* values (*k*) were tested in the classification procedure, where *k* represents the number of neighbors in the classification model for calculating the classification outcomes. The performance of the 2NN, 3NN, 4NN, 5NN, 6NN, 7NN, 8NN, 9NN, 10NN, 11NN, and 12NN were precisely reported in this study. In addition, the nearest neighbors search was performed using an exhaustive search and the Minkowski distance.

Support Vector Machine

SVM is driven by taking on a kernel function, which is a nonlinear one. Using SVM, the input features were transformed into a high-dimensional space. Compared to the original input features, a transformed one contributes a more trivial task for separating data. Contingent upon the input, an iterative operation of learning provides an optimum hyperplane with the maximum margin between the groups in a high dimensional feature space. Eventually, the maximum-margin hyper-planes will

outline the decision borders over the data clusters. The higher distance between hyper-planes and data points in miscellaneous categories leads to higher classification rates. The current study employed an RBF as a kernel function. A sub-sampling method was applied to select the kernel scale value.

Naive Bayes

NB belongs to statistical/probabilistic ML algorithms. Using NB, the class membership probabilities (e.g., the probability that a particular tuple is included in a certain class) can be predicted based on the Bayes theorem. In NB, the effect of a feature/predictor on a precise group is assumed to be independent of the other features/predictors and known as “*class conditional independence*.” This hypothesis simplifies the intricate calculations and is considered naïve. High accuracy and speed are achieved by applying NB to large databases. In this experiment, a Gaussian distribution was selected as the model of the distributions. Additionally, the bandwidth of the kernel smoothing window was automatically selected for each combination of the feature and class using a value that is optimal for the above-mentioned distribution.

Probabilistic Neural Network

PNN employs the RBF in a set of feed-forward networks. The RBF is regulated by varying the sigma (σ) parameter. In this study, the PNN was executed for 15 different σ values in the range of 0.3-1 with a step size of 0.05. The execution of PNN can be briefly described as follows.

The first layer states the closeness of the features to the training vector and holds the answers in a distinct vector. In the second layer, the process of adding the contributions is computed for each class and the outcomes are saved in a separate probability vector. Finally, the competing transfer function votes for the highest probabilities and keeps “1” for that group and a “0” for other categories (20).

Decision Tree

DT is a decision tool that shows the possible outcome of a decision such as the consequence of the event, groups, or group distributions, and resource expenditures using a tree-like model (18). A set of hierarchical decisions is adopted on the features for classification.

Ensemble Methods

EM is an ML algorithm, where multiple “weak learners” are used to solve the same problem and fused to catch better outcomes compared to any of the constituent learning procedures alone. Bagging and boosting are the two main subgroups of EM.

Bagging is a shortened form of the “bootstrap aggregating” and implicates having an apiece model in the ensemble vote with one and the same weight. To uphold model variance, this method trains each learning

procedure in the ensemble using an arbitrary subset of the training set. Boosting implicates incrementally constructing an ensemble through training each new instance to emphasize the training cases misclassified by the former learning procedures. Adaboost is the most common boosting procedure. The “random subspace” learning algorithm is comparable to bagging excluding that the features are randomly sampled with replacement for each learner.

In the current experiment, the number of ensemble learning cycles was set to 2 with a “Tree” learner for bagging. For Adaboost, the learning rate for shrinkage was set to 1 with 10 learning cycles and a “Tree” learner.

All classification approaches were implemented by three cross-validation (CV) schemes, including 5-, 8-, and 10-fold CVs. Further, accuracy (AC), sensitivity (SE), specificity (SP), and F1-score (F1) were designated to evaluate network performances. These measures are formulated as follows:

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$SE = \frac{TP}{TP + FN} \quad (2)$$

$$SP = \frac{TN}{TN + FP} \quad (3)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (4)$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. All simulations were accomplished using MATLAB software on a VAIO laptop series SR.

Results

In this work, saliva characteristics and participants' demographic information were used to discriminate between COPD and control groups. Four groups of information were defined as the inputs of the classification module, including four saliva measures, age, gender, and smoking status.

The classification module incorporated several classifiers including different EMs, SVM, kNN, NB, PNN, DT, FNN, ENN, LRNN, and GRNN. Furthermore, the performance of the classifiers was evaluated in three different modes including (a) 5-, (b) 8-, and (c) 10-fold CV scheme. The results are depicted in Figures 4, 5, and 6.

Based on Figure 4, FNN reached a maximum performance of 100% with 5 and 9 neurons in the hidden layer using a 5-fold CV. ENN also reached the maximum performances of 100% with the 3, 4, and 9 neurons in the hidden layer. The highest performances of 100% were achieved using 3-NN, 5-NN, 9-NN, Adaboost, DT, NB, and SVM. However, the maximum accuracy of 93.75% was obtained for $\sigma = 0.6, 0.75,$ and 0.8 using PNN. The highest accuracy rate of classification using LRNN was

98.75% with 4 neurons in the hidden layer.

According to the obtained data (Figure 5), kNN outperformed the other classifiers using an 8-fold CV. The highest performance rates of 100% were obtained for all k values. ENN reached the maximum performances of 100% with 6, 7, and 10 neurons in the hidden layer. LRNN also attained a maximum performance of 100% with 3, 4, 5, and 7 neurons in the hidden layer. Using PNN, the maximum accuracy of 100% was obtained for $\sigma = 0.75$. Moreover, the highest performances of 100% were achieved using bagging, Adaboost, DT, NB, and SVM. However, FNN reached a maximum performance of 97.5% with 3, 6, 7, and 10 neurons in the hidden layer. Using LRNN, the highest accuracy rate of classification was 98.75% with 4 neurons in the hidden layer.

The results (Figure 6) further revealed that kNN outperformed other classifiers by applying a 10-fold CV. The highest performance rates of 100% were achieved for all k values. Additionally, ENN obtained the maximum performances of 100% with the 3, 4, 6, 7, 8, and 9 neurons in the hidden layer. LRNN also conquered a maximum performance of 100% with 2, 3, 6, and 9 neurons in the hidden layer. Similarly, FNN reached a maximum performance of 100% with 6 neurons in the hidden layer. Using PNN, the maximum accuracy of 100% was found for $\sigma = 0.75$. The highest performances of 100% were achieved using bagging, Adaboost, DT, NB, and SVM. Ultimately, the highest accuracy rate of classification was 100% with 2, 3, 6, and 9 neurons in the hidden layer by employing LRNN.

Discussion

In the present study, the performances of several ML tools were compared for the discrimination of healthy and COPD subjects. The ML algorithms were SVM, kNN, NB, PNN, DT, FNN, ENN, LRNN, GRNN, and different EMs. Of all ML algorithms, kNN was the best since the classification rates of 100% were obtained for different k values (in kNN). Totally, our results showed that 12, 24, and 32 classifiers attained the highest performances of 100% using 5, 8, and 10-fold CV, respectively. These results represent that increasing the value of k in the k-fold CV strategy can increase the performance rates of classification algorithms. Nonetheless, Zarrin et al (10, 11) failed to investigate this issue in their studies.

The results of this study also demonstrated that reforming the structure of classification can strongly affect its performance rate. For example, the number of neurons in the hidden layer of the neural network fulfilled a highly important role in classifying the two groups. As an instance, FNN reached a maximum performance of 100% with 5 and 9 neurons in the hidden layer in a 5-fold CV, which was not considered by Zarrin et al in their studies (10, 11).

In terms of classifier performances, some ML algorithms



Figure 4. The Performance of Different ML Schemes in Discrimination Between COPD and Controls Using 5-Fold CV: (a) FNN, (b) PNN, (c) ENN, (d) LRNN, (e) kNN, and (f) other remaining approaches including DT, EMs, GRNN, SVM, and NB. *Note.* ML: Machine learning; COPD: Chronic obstructive pulmonary disease; CV: Cross-validation; FNN: Feed-forward neural network; PNN: Probabilistic neural network; ENN: Elman neural network; LRNN: Layer recurrent neural network; kNN: k-nearest neighbor; DT: Decision tree; EM: Ensemble method; GRNN: Generalized regression neural network; SVM: Support vector machine; NB: Naive Bayes.

reached the highest rate of 100% for accuracy, sensitivity, specificity, and F1-score, which contradicts the results of Zarrin et al (10, 11). In the former, the highest accuracy of 91.25% and the sensitivity of 100% were reported using the XGBoost gradient boosting procedure. The latter achieved the maximum accuracy and sensitivity of 89% and 86% using ANN, respectively.

Astonishing COPD diagnostic rates were obtained in this study by examining several common classification algorithms based on a limited number of salivary biomarkers and some demographic information. Regardless of the excellent accuracy of the classifiers, it should be noted that the framework was appraised on a limited number of samples. To warrant the efficiency of the approach, it should be evaluated on a more considerable number of samples in the future. Additionally, the

saliva samples were associated with COPD cases. The effectiveness of the scheme in classifying different respiratory diseases such as asthma, as well as different levels of the disease should be inspected in the future. Traditional ML algorithms were used in this study. Most of these algorithms (e.g., SVM and kNN) are sensitive to parameter settings (21, 22). For example, K computation, nearest neighbor choice/search, and classification rules are some challenges for kNN (22). Considering that the model parameter adjustment can impact classification results, future works should carefully think through this subject.

Conclusion

In this experiment, it was intended to develop an intelligent algorithm for the diagnosis of COPD based



Figure 5. The Performance of Different ML Schemes in Discrimination Between COPD and Controls Using 8-Fold CV. (a) FNN, (b) PNN, (c) ENN, (d) LRNN, (e) kNN, and (f) other remaining approaches, including DT, EMs, GRNN, SVM, and NB. *Note.* ML: Machine learning; COPD: Chronic obstructive pulmonary disease; CV: Cross-validation; FNN: Feed-forward neural network; PNN: Probabilistic neural network; ENN: Elman neural network; LRNN: Layer recurrent neural network; kNN: k-nearest neighbor; DT: Decision tree; EM: Ensemble method; GRNN: Generalized regression neural network; SVM: Support vector machine; NB: Naive Bayes.

on saliva permittivity and demographic information of the participants. Different ML algorithms, including SVM, kNN, NB, PNN, DT, FNN, ENN, LRNN, GRNN, and different EMs were examined in this regard. Three strategies, three different CVs, including 5-, 8-, and 10-fold CVs, and different structures of the classifiers were examined to evaluate the potential of ML algorithms. The experimental results revealed the imperative role of parameterization for the classification module, which can increase the performance of the classifier up to 100%. Overall, kNN outperformed the other classifiers. Using both 8- and 10-fold CVs yielded the highest performance rates of 100% for all k values.

Conflict of Interest Disclosures

The authors declare that they have no conflict of interests.

Acknowledgements

This article used data from the Exasens dataset (9-11). We hereby express our gratitude to Dr. Pouya Soltani Zarrin, Dr. Niels Roekendorf, and Dr. Christian Wenger for making the data available for free.

Ethical Statement

This article examined the Exasens dataset (9-11), which is freely available in the public domain. This article contains no studies with human participants performed by any of the authors.

Authors' Contributions

Both authors were equally involved in performing all the sections of the article.

Funding/Support

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.



Figure 6. The Performance of Different ML Schemes in Discrimination Between COPD and Controls Using 10-Fold CV: (a) FNN, (b) PNN, (c) ENN, (d) LRNN, (e) kNN, and (f) other remaining approaches including DT, EMs, GRNN, SVM, and NB. *Note.* ML: Machine learning; COPD: Chronic obstructive pulmonary disease; CV: Cross-validation; FNN: Feed-forward neural network; PNN: Probabilistic neural network; ENN: Elman neural network; LRNN: Layer recurrent neural network; kNN: k-nearest neighbor; DT: Decision tree; EM: Ensemble method; GRNN: Generalized regression neural network; SVM: Support vector machine; NB: Naive Bayes.

Informed Consent

Informed consent was obtained from all individual participants of the study.

References

1. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report: GOLD executive summary. *Respirology*. 2017;22(3):575-601. doi: 10.1111/resp.13012.
2. Decramer M, Janssens W, Miravittles M. Chronic obstructive pulmonary disease. *Lancet*. 2012;379(9823):1341-51. doi: 10.1016/s0140-6736(11)60968-9.
3. Buist AS, McBurnie MA, Vollmer WM, Gillespie S, Burney P, Mannino DM, et al. International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. *Lancet*. 2007;370(9589):741-50. doi: 10.1016/s0140-6736(07)61377-4.
4. WHO. World Health Statistics 2008. <http://www.who.int/>

5. Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med*. 2013;187(4):347-65. doi: 10.1164/rccm.201204-0596PP.
6. Agnew M. Spirometry in clinical use: practical issues. *Breathe*. 2010;6(3):196-203. doi: 10.1183/18106838.0603.196.
7. Dong T, Santos S, Yang Z, Yang S, Kirkhus NE. Sputum and salivary protein biomarkers and point-of-care biosensors for the management of COPD. *Analyst*. 2020;145(5):1583-604. doi: 10.1039/c9an01704f.
8. Khan RS, Khurshid Z, Yahya Ibrahim Asiri F. Advancing point-of-care (PoC) testing using human saliva as liquid biopsy. *Diagnostics (Basel)*. 2017;7(3):39. doi: 10.3390/diagnostics7030039.
9. Soltani Zarrin P, Jamal FI, Roekendorf N, Wenger C. Development of a portable dielectric biosensor for rapid detection of viscosity variations and its in vitro evaluations whosis/whostat/EN_WHS08_Full.pdf.

- using saliva samples of COPD patients and healthy control. *Healthcare (Basel)*. 2019;7(1):11. doi: [10.3390/healthcare7010011](https://doi.org/10.3390/healthcare7010011).
10. Soltani Zarrin P, Roeckendorf N, Wenger C. In-vitro classification of saliva samples of COPD patients and healthy controls using machine learning tools. *IEEE Access*. 2020;8:168053-60. doi: [10.1109/access.2020.3023971](https://doi.org/10.1109/access.2020.3023971).
 11. Soltani Zarrin P, Zahari F, Mahadevaiah MK, Perez E, Kohlstedt H, Wenger C. Neuromorphic on-chip recognition of saliva samples of COPD and healthy controls using memristive devices. *Sci Rep*. 2020;10(1):19742. doi: [10.1038/s41598-020-76823-7](https://doi.org/10.1038/s41598-020-76823-7).
 12. Miller CS, Foley JD, Bailey AL, Campell CL, Humphries RL, Christodoulides N, et al. Current developments in salivary diagnostics. *Biomark Med*. 2010;4(1):171-89. doi: [10.2217/bmm.09.68](https://doi.org/10.2217/bmm.09.68).
 13. Blicharz TM, Siqueira WL, Helmerhorst EJ, Oppenheim FG, Wexler PJ, Little FF, et al. Fiber-optic microsphere-based antibody array for the analysis of inflammatory cytokines in saliva. *Anal Chem*. 2009;81(6):2106-14. doi: [10.1021/ac802181j](https://doi.org/10.1021/ac802181j).
 14. Dillon MC, Opris DC, Kopanczyk R, Lickliter J, Cornwell HN, Bridges EG, et al. Detection of homocysteine and C-reactive protein in the saliva of healthy adults: comparison with blood levels. *Biomark Insights*. 2010;5:57-61. doi: [10.4137/bmi.s5305](https://doi.org/10.4137/bmi.s5305).
 15. Patel N, Belcher J, Thorpe G, Forsyth NR, Spiteri MA. Measurement of C-reactive protein, procalcitonin and neutrophil elastase in saliva of COPD patients and healthy controls: correlation to self-reported wellbeing parameters. *Respir Res*. 2015;16(1):62. doi: [10.1186/s12931-015-0219-1](https://doi.org/10.1186/s12931-015-0219-1).
 16. Bhavsar NV, Dave BD, Brahmabhatt NA, Parekh R. Periodontal status and oral health behavior in hospitalized patients with chronic obstructive pulmonary disease. *J Nat Sci Biol Med*. 2015;6(Suppl 1):S93-7. doi: [10.4103/0976-9668.166097](https://doi.org/10.4103/0976-9668.166097).
 17. Ji J, von Schéele I, Bergström J, Billing B, Dahlén B, Lantz AS, et al. Compartment differences of inflammatory activity in chronic obstructive pulmonary disease. *Respir Res*. 2014;15(1):104. doi: [10.1186/s12931-014-0104-3](https://doi.org/10.1186/s12931-014-0104-3).
 18. Han J, Pei J, Kamber M. *Data Mining: Concepts and Techniques*. 3rd ed. Elsevier; 2011.
 19. Larose DT. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons; 2014.
 20. Demuth H, Beale M. *Neural Network Toolbox*. MathWorks; 2000.
 21. Miranda PB, Prudêncio RB. Active testing for SVM parameter selection. In: *The 2013 International Joint Conference on Neural Networks (IJCNN)*. Dallas, TX: IEEE; 2013. p. 1-8. doi: [10.1109/ijcnn.2013.6706910](https://doi.org/10.1109/ijcnn.2013.6706910).
 22. Zhang S. Challenges in KNN classification. In: *IEEE Transactions on Knowledge and Data Engineering*. IEEE; 2021. doi: [10.1109/tkde.2021.3049250](https://doi.org/10.1109/tkde.2021.3049250).