IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 27, 2020, accepted September 10, 2020, date of publication September 14, 2020,
date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023971

In-Vitro Classification of Saliva Samples of COPD Patients and Healthy Controls Using Machine Learning Tools

POUYA SOLTANI ZARRIN[1], NIELS ROECKENDORF[2,3], AND CHRISTIAN WENGER[1,4]

[1]Leibniz-Institut für Innovative Mikroelektronik (IHP), 15236 Frankfurt, Germany
[2]Division of Mucosal Immunology and Diagnostics, Priority Area Asthma and Allergy, Research Center Borstel–Leibniz Lung Center, Leibniz Research Alliance Health Technologies, 23845 Borstel, Germany
[3]German Center for Lung Research, 23845 Borstel, Germany
[4]BTU Cottbus-Senftenberg, 01968 Cottbus, Germany

Corresponding author: Pouya Soltani Zarrin (soltani@ihp-microelectronics.com)

This work was supported in part by the Federal Ministry for Education and Research (BMBF) of Germany for the EXASENS Project under Grant 13U13862, and in part by the Open Access Fund of the Leibniz Association.

ABSTRACT Chronic Obstructive Pulmonary Disease (COPD) is a life-threatening lung disease and a major cause of morbidity and mortality worldwide. Although a curative therapy has yet to be found, permanent monitoring of biomarkers that reflect the disease progression plays a pivotal role for the effective management of COPD. The accurate examination of respiratory tract fluids like saliva is a promising approach for staging disease and predicting its upcoming exacerbations in a Point-of-Care (PoC) environment. However, the concurrent consideration of patients' demographic and medical parameters is necessary for achieving accurate outcomes. Therefore, Machine Learning (ML) tools can play an important role for analyzing patient data and providing comprehensive results for the recognition of COPD in a PoC setting. As a result, the objective of this research work was to implement ML tools on data acquired from characterizing saliva samples of COPD patients and healthy controls as well as their demographic information for PoC recognition of the disease. For this purpose, a permittivity biosensor was used to characterize dielectric properties of saliva samples and, subsequently, ML tools were applied on the acquired data for classification. The XGBoost gradient boosting algorithm provided a high classification accuracy and sensitivity of 91.25% and 100%, respectively, making it a promising model for COPD evaluation. Integration of this model on a neuromorphic chip, in the future, will enable the real-time assessment of COPD in PoC, with low cost, low energy consumption, and high patient privacy. In addition, constant monitoring of COPD in a near-patient setup will enable the better management of the disease exacerbations.

INDEX TERMS COPD classification, AI in medicine, personalized healthcare, permittivity spectroscopy, precision diagnostic, saliva characterization, medical machine learning, XGBoost.

I. INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is a life-threatening lung disease, causing breathing difficulties in patients due to airflow constraints in lungs [1]. It is a progressive disease, developing slowly over time, while its symptoms often worsen. COPD is one of the main leading causes of death worldwide, affecting millions of people and causing a considerable economical burden on healthcare systems [2]. The major cause of COPD is the long-term exposure of subjects to either tobacco smoke (being an active–secondhand smoker) or other lung-irritants such as air pollution, chemical fumes, or industrial dust. In some scarce cases, however, a genetic condition called alpha-1 antitrypsin deficiency may also contribute to lung damages and COPD [1]. The main symptoms of COPD are shortness of breath, chronic coughs, wheezing, chest tightness, and abnormal sputum (mucus) production. Although an absolute cure for reversing caused lung damages has yet to be found, an early-stage diagnosis has shown to have a pivotal role for the effective management of COPD [3]. COPD, as one of the most prevalent lung diseases worldwide, runs a perfidious

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Kamrul Hasan.

course with an often long-lasting undiagnosed initial phase. Clinical treatment approaches for COPD result in repeated clinical visits and extended hospitalization for patients. This fact, apart from being an economical burden for healthcare infrastructures, drastically impacts patients' life quality. To address this issue, contemporary healthcare systems have encouraged the development of personalized solutions, through which patients can receive appropriate medical assistance in an outpatient clinic or a home-care environment [4].

The clinical ground truth methodology for diagnosing COPD is comprehensively reported within the GOLD guidelines [5]. Among available screening and detection methods, spirometry pulmonary function test is the most rudimentary and systematic method in primary care for the diagnosis of COPD. Along this test, the lung capacity of patients is measured during breathing in–out cycles [6]. According to a study by Haroon *et al.*, COPD is widely under-diagnosed due to the limited sensitivity of the spirometry test in the range of 64.5–79.9% [7]. As a result, examining mucin, present in sputum or saliva samples, provides more reliable information on the course of the disease which can be affected by bacterial infections [6]. Sputum and saliva are both mucosal secretions, their composition is affected by changes in health conditions of individuals suffering from inflammatory lung diseases such as COPD [4], [8]. Alterations in the mucin production during the course of COPD, which impacts the viscosity of mucosal secretions, has long been studied [9]. In addition, the expression of aquaporin-5 has found to be decreased in some COPD patients, affecting dielectric properties of their respiratory tract fluids [10]. Main contents of sputum, produced by lungs, are mucin, water, epithelial cells of the airway mucosa, and salt (in physiological concentrations). Salt concentrations in epithelial lining fluid was investigated by Effros *et al.*, indicating its effect on dielectric properties of the fluid [11]. However, a direct correlation of dielectric properties with COPD was not stated in their work [11]. Therefore, investigating dielectric and supramolecular properties of sputum can provide useful information for staging COPD [12]. In other words, water content variations, at different stages of the disease, affects the supramolecular properties of mucin gels [12]. Upon the entry of water into mucin's gels matrix, there is a considerable amount of proton release resulting from cations exchange (particularly Ca and Na), which drastically changes the dielectric properties of mucin samples [12]. In other words, sputum samples collected from COPD patients are expected to have different permittivity characteristics compared to samples of Healthy Controls (HC), which could be used as a biomarker for the assessment of the disease in a Point-of-Care (PoC) environment [13], [14]. However, due to complexities of obtaining sputum samples non-invasively on a daily-basis, saliva could be an alternative with better patient compliance for PoC applications [15].

Although the dielectric characterization of saliva samples can potentially shine a spotlight onto the detection of COPD in a PoC setting, the comprehensive diagnosis of the disease requires a sophisticated algorithm by concurrent consideration of all essential parameters related to a patient's personal and medical backgrounds [14]. These demographic parameters include, but are not limited to, age, gender, weight, cytokine level, pathogen load, and the smoking background of subjects [14]. Therefore, without analytical insight, information obtained on one specific parameter has a low clinical value for the disease diagnosis [16]. As a result, implementation of Machine Learning (ML) tools is crucial for the conversion of collected raw data from subjects into meaningful clinical–diagnostic information [17]–[19]. Furthermore, advanced ML analytics could make the management of COPD in PoC applications more efficient. Therefore, the novel hypothesis of this work was to scrutinize whether dielectric properties of saliva change upon the development of a COPD; and whether ML tools, applied on this information together with demographic parameters, can identify the diagnostic status of patients.

Among various ML classifiers, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), principal component analysis, Logistic Regression (LR), eXtreme Gradient Boosting (XGBoost) algorithm, and Naïve Bayes (NB) are among the most common models used for the classification of medical data [20]–[23]. Although ANNs generally provide acceptable performance for classifying data, their sensitivity to outliers— especially in small datasets—causes overfitting issues, thus degrading their accuracy [24]. On the other hand, non-perceptron classifiers such as XGBoost or SVM are less prone to overfitting and less sensitive to outliers, thus performing notably better in applications with a small-sized dataset. In addition, unlike ANNs, non-perceptron classifiers are computationally more efficient since their computational complexity does not depend on the dimension of the input space, making them an appropriate tool for edge computing applications [24]. Therefore, energy–computation efficiency of non-perceptron classifiers make them a suitable choice for medical data classifications in PoC applications.

The objective of this work was to apply machine learning tools on data obtained from characterizing saliva samples of COPD patients and HC for diagnostic classifications. This study is the extension of our previous work, which introduced a neuromorphic-compatible ANN for COPD pattern recognition using synthesized data [25]. However, the current study deals with real data collected from COPD patients and HC in a clinical setting. The high performance of the XGBoost algorithm for classifying saliva samples, with relatively a small number of data points, and its less susceptibility to overfitting made it an adequate tool for clinical analytics in this work. Although the presented research in this work targets the PoC detection of COPD in a personalized care scheme, introduced ML techniques can be used in the future for the enhancement of conventional clinical-based standard of care methods available for diagnosing COPD.

II. METHODS AND MATERIALS

Two groups of saliva samples, 160 for HC and 79 for COPD patients, were collected in the frame of a joint research project Exasens at the Research Center Borstel, BioMaterialBank Nord (Borstel, Germany). Patient materials were collected between November 2016 and February 2018 and were anonymized prior to accessibility. The sampling procedure of saliva samples was approved by the local ethics committee of the University of Luebeck under the approval number AZ-16-167 and a written informed consent was obtained from all patients. COPD subjects of the study were patients who had been previously hospitalized in the pulmonary clinic Borstel (Borstel, Germany) and several outpatients. Therefore, the inclusion criterion for enrolling patients into the COPD group was a diagnosed COPD without acute respiratory infection, with respect to the GOLD guidelines [5]. Inclusion criteria for the healthy group were the absence of a diagnosed COPD or asthma affections. Demographic information—including gender, age, smoking status (smoker, ex-smoker, and non-smoker), the date of probing, sampling conditions, and special notes regarding the contamination of saliva with blood—were collected at the recruitment, based on patients' self-declarations. Saliva sampling (5 ml) after mouth wash was induced using a chewing gum (GC Corporation, Leuven, Belgium). Collected samples were aliquoted and snap frozen in liquid nitrogen immediately after receipt and were stored at −80 °C. To avoid frequent freeze–thaw cycles, samples were thawed and transferred onto the sensor immediately before dielectric measurements, as recommended in standard operating procedures for keeping the integrity of human biospecimens such as saliva [26]. Although effects of freezing and de-freezing of saliva samples on their dielectric properties have yet to be investigated, all characterized samples in this work have exactly gone through a one-freeze–one-thaw cycle. As a result, possible effects of freezing samples have not been considered as a model variable for our ML models. Measurements on dielectric properties of saliva samples were conducted *in-vitro* at the Research Center Borstel, Leibniz lung center.

A. DIELECTRIC CHARACTERIZATION OF SALIVA SAMPLES

Prior to measurements, saliva samples of COPD and HC (40 samples for each group) were defrozen and centrifuged for removing insoluble matter. The centrifugation process was conducted using a commercialized centrifuge (Eppendorf centrifuge 5415R, Eppendorf Inc., Hamburg, Germany) at 4 °C and 4000 RPM for a duration of 5 minutes. As shown in Fig. 1, a previously developed permittivity biosensor (IHP Microelectronics, Frankfurt Oder, Germany) was used for the dielectric characterization of saliva samples [13], [14]. The output of the biosensor was extracted into an Excel file using a user-friendly data acquisition (PLX-DAQ) interface, as demonstrated in Fig. 1. It is noteworthy that the calibration inconsistency and the performance degradation, caused by frequent cleaning cycles, impairs the long-term functioning

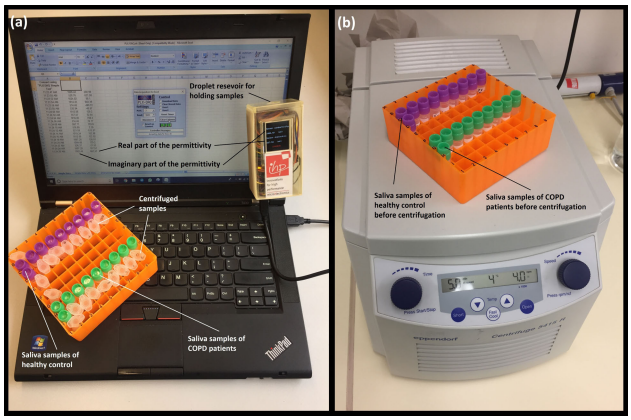

FIGURE 1. (a) Measurement setup demonstrating the biosensor output for real–imaginary parts of the permittivity of saliva samples; (b) the centrifugation process for saliva samples collected from healthy controls and COPD patients.

of the developed biosensor, thus restricting the sample population size which can be characterized in stable and reliable circumstances. Consequently, only 80 samples out of the available 239 were dielectrically characterized in this study. As extensively reported in the previous work, the developed biosensor is capable of measuring both real and imaginary parts of the permittivity of a material-under-test [13], [14]. From a physical point of view, the real part of the permittivity represents a material's energy absorption (dielectric properties) in an interaction with an electromagnetic field; while the imaginary part of it is an indicator of a material's energy loss (conductivity properties). After the sample preparation process, a droplet of 5 μL was taken and located over the sensing area of the device. Upon the presence of a sample droplet, the output voltage of the biosensor notably drops from its calibration level, depending on permittivity properties of the introduced sample. All measurements were conducted in a lab environment with a controlled room temperature following a primary cleaning procedure using ethanol and compressed air for the removal of extraneous particles from the sensor surface. It is noteworthy that temperature fluctuations can possibly impair the biosensor performance in real-world applications as part of measurement uncertainties associated with this system [13], [14]. As a solution, the ambient temperature can be introduced as an input variable into the ML model [25]. Nonetheless, considering the consistency of the ambient temperature throughout experiments in this work, this parameter was not considered in our ML approach. Furthermore, uncertainties associated with the calibration and cleaning of the biosensor need to be addressed in the future for performance enhancements in long-term applications. To obtain reliable results, experiments were repeated three times for every sample and the average (for the duration of the sample's presence over the sensing area) and minimum values of observed results for each experiment were recorded. Subsequently, the absolute minimum and overall average of all three trials were reported for final results of real and imaginary parts of the permittivity of every sample. While the

average value of observations represents the overall dielectric characteristic of a sample, the minimum value could be an indicator of the presence of some specific suspending particles inside a sample. Further information on the working principle of the dielectric biosensor is presented in details in previous studies [13], [14].

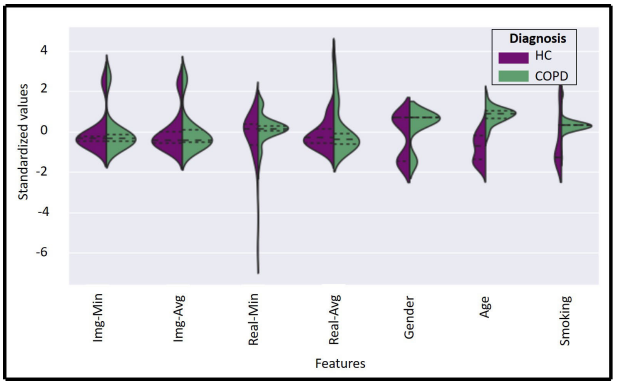

FIGURE 2. Violin plot of dataset attributes used for classifications, representing the Gaussian standard normal distribution of features with zero mean and unit variance.

B. MACHINE LEARNING IMPLEMENTATION FOR CLASSIFICATIONS

1) DATA PREPARATION

As discussed in the previous section, dielectric characterizations were conducted on only 80 samples out of the available 239, due to the limited life-cycle of the biosensor. However, to highlight the important role of demographic features in COPD detection, analysis were performed on both datasets with and without dielectric properties. The first dataset includes information on 80 characterized saliva samples (40 samples for each group of COPD and HC). Attributes of this dataset include both demographic features—or more specifically gender, age, and the smoking status of patients—and the permittivity properties of saliva samples obtained through measurements, as shown in Fig. 2. On the other hand, the second dataset includes only the demographic information of all 239 saliva samples with following attributes: gender, age, and the smoking status of patients. For computational purposes, non-quantitative attributes—diagnosis, gender, and smoking status—were converted into numerical values using following labels: diagnosis (COPD (1)–HC (0)), gender (male (1)–female (0)), smoking status (smoker (3)–ex-smoker (2)–non-smoker (1)). Sections of the data, used for analytics in this work, are publicly available at http://ieee-dataport.org/2361 [27]. To improve the performance of ML models, the first and second datasets were normalized and standardized to the Gaussian standard normal distribution with zero mean and unit variance, respectively, as presented in Fig. 2. Standardization of datasets is significantly important for improving the performance of many machine learning classifiers, since the objective function of their learning algorithm considers all attributes of a dataset to be centered around zero with a variance in a similar order of magnitude. Data preparations and ML implementations were performed on the JupyterLab environment using Keras 2.2.5 and Scikit-learn 0.22 libraries of Python [23].

2) ANALYTICAL TOOLS

Non-perceptron machine learning classifiers including Gaussian NB (GNB), SVM, and LR—provided by the Scikit-learn 0.22 library [23]—as well as the powerful decision tree algorithm, XGBoost [21], were used for the classification of saliva samples of COPD and HC. In addition, a dense ANN with one hidden-layer and one read-out layer was developed for the classification of COPD and HC samples. To replicate the intrinsic structure of a neuromorphic platform, a hidden layer with 4 neurons and a sigmoid activation function was modeled. The read-out layer, with a sigmoid activation function, consisted of two neurons for two possible classes of COPD and HC. A dropout with 20% probability was applied to the hidden-layer for the overfitting prevention. Adam optimization algorithm, with 0.0001 learning rate, and a cross entropy error function were used for training network in the backend using Google Colab GPU platform. The simple architecture of the developed ANN was chosen for the integration compatibility with the intended neuromorphic hardware. For the proposed SVM model, a radial basis function kernel was chosen with a gamma and cost parameters of 0.1 and 1000, respectively. For the LR algorithm, a limited-memory Broyden–Fletcher–Goldfarb–Shanno optimization algorithm was used for the parameter estimation with a multinomial loss fit across the entire probability distribution. The XGBoost model was fine-tuned with a learning rate and random state values of 0.01 and 1, respectively. In addition, its number of trees in the forest was chosen as 100 with a maximum depth of 3 for every tree. A multiclass log loss function and minimum weighted leaf fraction of 0 were chosen as recommended in its default instruction [21]. XGBoost is an optimized distributed gradient boosting decision tree framework, providing high efficiency and flexibility for portable applications. The parallelization of tree construction in its algorithm leads to efficiency of compute time and memory resources, thus making XGBoost an adequate tool for edge computing applications such as PoC diagnostic devices. All metrics and models used in this study are available in details at https://github.com/Pouya-SZ/HCOPD.

Considering the small size of the investigated COPD data set, k-fold cross-validation method was implemented for the evaluation of models, thus preventing overfitting circumstances. Hence, relevant tools provided at the Scikit-learn 0.22 library were used for the 5-fold cross-validation of models [23]. The average of five cross-validation iterations was reported as the 5-fold accuracy (5-fold Acc.) performance of models, as shown in Tables 1 and 2. Sensitivity (recall), specificity, and precision measures for models were calculated on a single-fold iteration with the best accuracy. Since for every cross-validation iteration, the dataset was split into test–train subsets with a ratio of 20–80%, the test-fraction, with unseen

TABLE 1. Performance of ML models for the classification of the first dataset with 80 saliva samples (64 training and 16 test data).

Classifier	5-fold Acc.	Sensitivity	Specificity	Precision
XGBoost	91.25%	100%	88.89%	87.5%
SVM	91.25%	100%	77.78%	77.78%
GNB	87.5%	85.71%	100%	100%
LR	90%	100%	11.11%	46.66%
ANN	73.75%	100%	44.45%	58.33%

TABLE 2. Performance of ML models for the classification of the second dataset with 239 saliva samples (191 training and 48 test data).

Classifier	5-fold Acc.	Sensitivity	Specificity	Precision
XGBoost	92.05%	95.24%	100%	100%
SVM	92.05%	90.47%	100%	100%
GNB	92.89%	95.24%	100%	100%
LR	91.65%	95.24%	96.3%	95.24%
ANN	93.33%	90.47%	96.3%	95%

data points during model training, was considered as an external validation dataset for the evaluation of models. The sensitivity (recall) of a model was calculated as the proportion of true positives out of all diseased cases; while the specificity value shows the number of true negatives over number of true negatives and false positives. Precision criterion shows the ratio of true positives over true plus false positives.

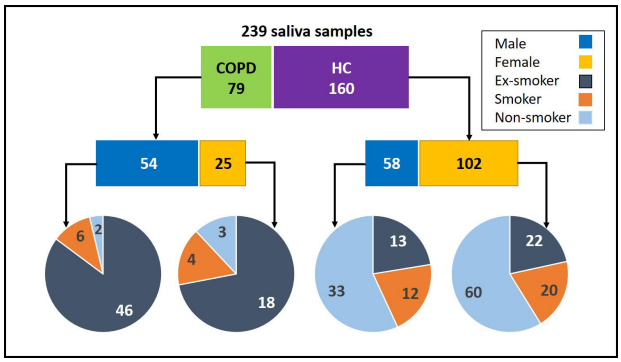

FIGURE 3. Hierarchical categorization of collected saliva samples into extended subgroups with respect to their diagnosis, gender, and smoking status.

III. RESULTS AND DISCUSSIONS

Fig. 3 demonstrates a hierarchy chart, categorizing collected saliva samples into extended subgroups with respect to their diagnosis, gender, and smoking status. As reported, more than two-thirds of COPD diagnosed subjects are male patients. Although this phenomenon could be explained considering the fact that smoking tobacco, and consequently COPD, is more prevalent among men, some studies suggest a more complex interpretation by taking into account various gender-specific factors such as differential susceptibility to tobacco, anatomic and hormonal differences, behavioral differences, and differences in response to available therapeutic modalities [28]. In addition, according to observations reported in Fig. 3, 81% of COPD diagnosed patients

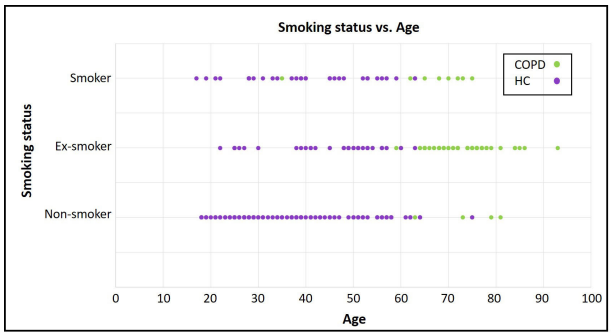

FIGURE 4. Distribution of saliva samples of COPD and HC with respect to age and the smoking status of subjects.

are holding an ex-smoker status, while only 13% are active smokers. This could possibly be due to the reason that some of ex-smoker subjects have already reached a severe stage of COPD before making a decision to quit smoking. This point is also noticeable in Fig. 4, which presents the distribution of saliva samples with respect to age, diagnosis, and the smoking status of subjects. Furthermore, as shown in Fig. 4, in most cases, COPD diagnosed patients are middle-aged or older adults. This observation complies with the fact that, increasing age means a longer exposure of subjects to risk factors and, consequently, further damages to their lungs. Moreover, as the body ages, the recovery process of damaged lung cells becomes more difficult, making a subject more susceptible to COPD [29].

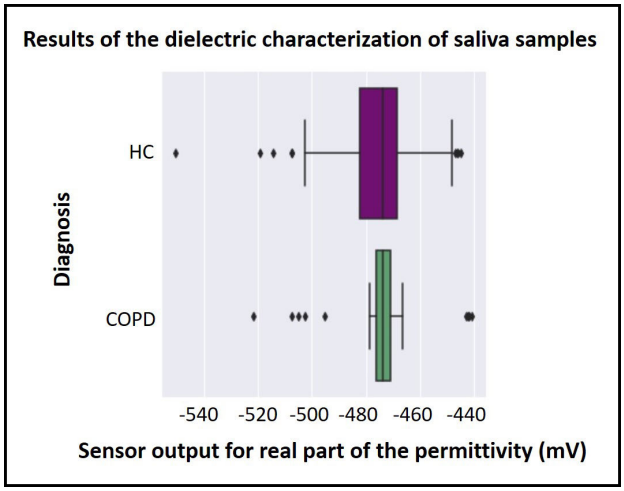

FIGURE 5. Results of the biosensor output for the dielectric characterization (minimum values of the real part of the permittivity) of saliva samples, providing distributional information including minimum–maximum, median, and the first–third quartile values.

Fig. 5 presents the results of the biosensor output for the dielectric characterization of saliva samples (the minimum values of the real part of the permittivity). The presented box-plot provides distributional information including minimum–maximum, median, and the first–third quartile values for the obtained results on both categories of COPD

and HC. Reported results in this figure represent the output voltage drop of the biosensor, after calibration with respect to dielectric properties of the surrounding air, representing dielectric properties of tested samples, as explained in previous studies [13], [14]. As illustrated in Fig. 5, minimum values of the real part of the permittivity (dielectric features) for HC samples has a greater standard deviation value (21.86) compared to the COPD group (16.77), making it a useful feature for clustering data points using ML classifiers. In contrast, the imaginary part of the permittivity of saliva has a symmetric distribution for both HC and COPD patients, as shown in Fig. 2, thus lacking valuable information for data segregation.

Due to the dependent nature of aforementioned attributes, implementation of ML methods was crucial for the realistic classification of samples by concurrent consideration of all parameters. Tables 1 and 2 present the performance of the proposed ML models including XGBoost, SVM, GNB, LR, and ANN for the classification of first and second datasets with 80 and 239 saliva samples, respectively. These results indicate the high performance of ML-based analytical tools for the classification of saliva samples of COPD and HC. Especially, among introduced methods, the XGBoost decision tree algorithm provided the best performance in terms of accuracy, sensitivity, specificity, and precision, thus making it a suitable model for this work. As reported in Table 1, XGBoost classifier has exceeded other models by providing accuracy, sensitivity, specificity, and precision values of 91.25%, 100%, 88.89%, and 87.5%, respectively, for the classification of saliva samples with respect to their dielectric and demographic properties. Acquired results illustrate the practicality of the concept of applying ML tools for classifying saliva samples of COPD and HC, which was proposed as a hypothesis for this work. The current study is an important cornerstone, presenting the fact that dielectric properties of saliva together with the demographic information of respective patients can be analyzed using ML tools for the discrimination of patients affected by COPD from healthy controls. Moreover, results presented in Table 2 indicate the superiority of XGBoost performance for the classification of samples based on merely demographic information with accuracy, sensitivity, specificity, and precision values of 92.05%, 95.24%, 100%, and 100%, respectively. The remarkable performance of classifiers based on only demographic attributes—age, gender, and smoking status—indicates the significant role of demographic information for the detection of COPD. In contrast to non-perceptron classifiers, the proposed ANN provided a poor performance due to its sensitivity to outliers and overfitting in small-sized datasets.

Figures 6(a) and (b) demonstrate confusion matrices of the XGBoost algorithm for predicting diagnostic status of unseen test samples. As shown in these figures, the XGBoost model, at its best performance (with threshold values of 0.66 and 0.53), was capable to predict the status of unseen test subjects for the first and second datasets with only one false-positive

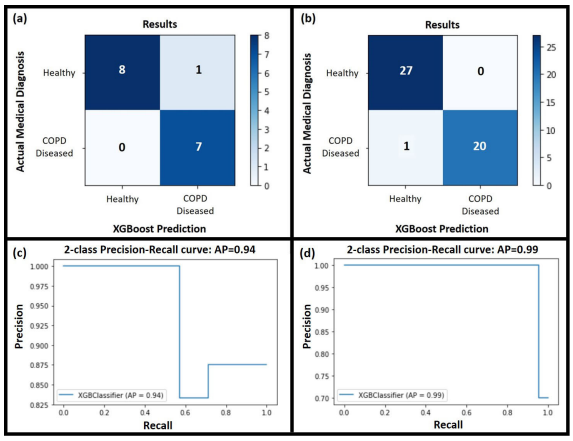

FIGURE 6. Confusion matrices of the XGBoost algorithm, presenting the prediction performance of the model on unseen test samples and the precision-recall curve, demonstrating the trade-off between precision and recall for different thresholds: (a) and (c) first dataset with 80 saliva samples (64 training and 16 test data); (b) and (d) second dataset with 239 saliva samples (191 training and 48 test data).

and one false-negative and accuracies of 93.75% and 97.92%, respectively. The precision-recall curves for these confusion matrices are shown in figures 6(c) and (d), demonstrating the trade-off between precision and recall (sensitivity) for different thresholds. The XGBoost classifier provided the best performance for the classification of first and second datasets at threshold values of 0.66 and 0.53, respectively. The high accuracy, sensitivity, specificity, and precision of the XGBoost algorithm make it an adequate model for COPD classifications using edge devices [30].

The high accuracy of the proposed ML models compared to the ground truth spirometry method, with a sensitivity range of 64.5–79.9% [7], make them a promising tool for the management of COPD in PoC environments. However, this work has only investigated the practicality of COPD classification in *in-vitro* circumstances, whereas, for real-world predictions on the progression of COPD and its exacerbations, real-time analysis of dielectric properties of saliva on a daily basis is required.

Generalizability to a larger population, as the main limitation with any ML study on a small-sized novel dataset, is a fundamental concern for this study, demanding extensive investigation. Nonetheless, to the best of our knowledge, there is no other comprehensive dataset for the COPD classification available up to date, which can be used for training and evaluating introduced ML models in this work. Therefore, we consider our study as a stepping stone to future studies in the field, while endorsing the necessity for further data collections and advanced-analytical implementations for the COPD management.

Results of this work imply the capability of ML tools for enhancing the quality of personalized healthcare solutions by facilitating the management of chronic diseases through performing complex diagnosis. The scope of ML tools goes far beyond classical statistical analyses performed

in healthcare. Therefore, ML methods, or the Artificial Intelligence (AI) from a broader scope, is expected to revolutionize healthcare in the near future by providing accurate and real-time predictions on the health status of patients or the progress of their diseases. Especially, AI will play a pivotal role in the future for assisting patients in remote locations with the management of chronic and degenerative conditions, monitoring their rehabilitation progress, and predicting critical–emergency health conditions. In addition, availability of numerous health-related data, thanks to advancements in wearable technologies and biosensors, will facilitate the better integration of AI with healthcare devices in PoC environments. However, all the astonishing capabilities of ML tools come at the cost of immense energy consumption and enormous computational power. In addition, complexities associated with cloud communications such as robustness against interference, wide bandwidth requirements, low latency, and data security limit the application of AI in sensitive fields such as medicine. As a result, the trade-off between mentioned benefits and risks related to securing sensitive medical data is, still, an on-going challenge. To address this concern, low-power neuromorphic platforms could be integrated into medical devices for locally processing of computations required for ML algorithms [31]. Neuromorphic chips have been successfully implemented in different studies for matrix-multiplications required for non-perceptron and perceptron-based ML methods [32], [33]. By bringing the data post-processing from backend onto a chip, real-time analysis of data in a less time consuming manner with a smaller time delay is feasible. Furthermore, sensitive medical data are better protected by being processed locally on a chip without external communications. In addition, the energy-efficient neuromorphic platforms offer a large fault tolerance for sensitive applications such as in healthcare [34]. Therefore, implementation of presented ML models on a neuromorphic platform, for the on-chip classification of saliva, is the next goal of this work.

Although the introduced ML model was capable to accurately classify saliva samples based on a few attributes, further demographic information on the medical-personal background of patients' (such cytokine level, blood pressure history, or pathogen) could possibly improve the performance of the model in terms of accuracy and generalizability. However, accessing such a sensitive medical information is highly restricted through governmental data protection policies and, thus, requires an appropriate approval from the ethics committee, prior to acquisitions. In addition, investigating novel ML algorithms, such as a few-shot learning, with better performance on small-sized datasets could pave the way towards more accurate and generalizable models for medical applications with limited data availability [35].

IV. CONCLUSION AND FUTURE WORK

This work investigated the *in-vitro* classification of saliva samples of COPD and HC using machine learning techniques. Saliva samples were initially collected from different subjects in a clinical setting and their demographic information on the age, gender, and smoking status of patients were recorded. In addition, dielectric characteristics of a smaller subset of collected samples were measured using a permittivity biosensor. Various ML tools including XGBoost, SVM, NB, LR, and ANN were applied for classifying collected samples into COPD and HC categories. The XGBoost algorithm provided the best performance, among other methods, for classifying and predicting saliva samples of COPD and HC with respect to their dielectric and demographic properties. Although implementation of ML tools enables the fast and efficient diagnosis of COPD, their existing shortcomings in terms of data availability, data safety, and computation cost limit their application in real-world. Therefore, further data collection is necessary in the future for enhancing the performance of proposed models. Moreover, as a future work, deployment of the introduced ML models on hardware-based neuromorphic platforms will enable the on-chip recognition of COPD with low energy consumption and high patient privacy.

ACKNOWLEDGMENT

The authors thank the BioMaterialBank Nord (BMB Nord), popgen 2.0 network (P2N), and the German Center for Lung Research for the collection of saliva samples and the staff at IHP and FZ Borstel-Leibniz Lung Center for their precious support with this work, especially Andreas Frey for providing access to facilities.

REFERENCES

[1] P. J. Barnes, "Mechanisms in COPD: Differences from asthma," *J. Chest*, vol. 117, no. 2, p. 10S–14S, 2000.

[2] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Med.*, vol. 3, no. 11, p. e442, Nov. 2006.

[3] N. G. Csikesz and E. J. Gartman, "New developments in the assessment of COPD: Early diagnosis is key," *Int. J. Chronic Obstructive Pulmonary Disease*, vol. 9, pp. 277–286, 2014.

[4] T. Dong, S. Santos, Z. Yang, S. Yang, and N. E. Kirkhus, "Sputum and salivary protein biomarkers and point-of-care biosensors for the management of COPD," *Analyst*, vol. 145, no. 5, pp. 1583–1604, 2020.

[5] S. Mirza, R. D. Clay, M. A. Koslow, and P. D. Scanlon, "COPD guidelines: A review of the 2018 GOLD report," *Mayo Clinic Proc.*, vol. 93, no. 10, pp. 1488–1502, Oct. 2018.

[6] D. Price, A. Crockett, M. Arne, B. Garbe, R. Jones, A. Kaplan, A. Langhammer, S. Williams, and B. Yawn, "Spirometry in primary care case-identification, diagnosis and management of COPD," *Primary Care Respiratory J.*, vol. 18, no. 3, pp. 216–223, Aug. 2009.

[7] S. Haroon, R. Jordan, Y. Takwoingi, and P. Adab, "Diagnostic accuracy of screening tests for COPD: A systematic review and meta-analysis," *BMJ Open*, vol. 5, no. 10, Oct. 2015, Art. no. e008133.

[8] S. Chiappin, G. Antonelli, R. Gatti, and F. Elio, "Saliva specimen: A new laboratory tool for diagnostic and basic investigation," *Clinica Chim. acta*, vol. 383, no. 1, pp. 30–40, 2007.

[9] M. C. Rose and J. A. Voynow, "Respiratory tract mucin genes and mucin glycoproteins in health and disease," *Physiological Rev.*, vol. 86, no. 1, pp. 245–278, Jan. 2006.

[10] K. Wang, Y.-L. Feng, F.-Q. Wen, X.-R. Chen, X.-M. Ou, D. Xu, J. Yang, and Z.-P. Deng, "Decreased expression of human aquaporin-5 correlated with mucus overproduction in airways of chronic obstructive pulmonary disease," *Acta Pharmacologica Sinica*, vol. 28, no. 8, pp. 1166–1174, Aug. 2007.

[11] R. M. Effros, B. Peterson, R. Casaburi, J. Su, M. Dunning, J. Torday, J. Biller, and R. Shaker, "Epithelial lining fluid solute concentrations in chronic obstructive lung disease patients and normal subjects," *J. Appl. Physiol.*, vol. 99, no. 4, pp. 1286–1292, Oct. 2005.
[12] P. Verdugo, "Supramolecular dynamics of mucus," *Cold Spring Harbor Perspect. Med.*, vol. 2, no. 11, 2012, Art. no. a009597.
[13] P. Soltani Zarrin, F. Jamal, S. Guha, J. Wessel, D. Kissinger, and C. Wenger, "Design and fabrication of a BiCMOS dielectric sensor for viscosity measurements: A possible solution for early detection of COPD," *Biosensors*, vol. 8, no. 3, p. 78, Aug. 2018.
[14] P. Zarrin, F. Jamal, N. Roeckendorf, and C. Wenger, "Development of a portable dielectric biosensor for rapid detection of viscosity variations and its *in vitro* evaluations using saliva samples of COPD patients and healthy control," *Healthcare*, vol. 7, no. 1, p. 11, Jan. 2019.
[15] R. Khan, Z. Khurshid, and F. Yahya Ibrahim Asiri, "Advancing point-of-care (PoC) testing using human saliva as liquid biopsy," *Diagnostics*, vol. 7, no. 3, p. 39, Jul. 2017.
[16] S. B. Baker, W. Xiang, and I. Atkinson, "Internet of Things for smart healthcare: Technologies, challenges, and opportunities," *IEEE Access*, vol. 5, pp. 26521–26544, 2017.
[17] A. L. Fogel and J. C. Kvedar, "Artificial intelligence powers digital medicine," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–4, Dec. 2018.
[18] P. S. Zarrin, A. Escoto, R. Xu, R. V. Patel, M. D. Naish, and A. L. Trejos, "Development of a 2-DOF sensorized surgical grasper for grasping and axial force measurements," *IEEE Sensors J.*, vol. 18, no. 7, pp. 2816–2826, Apr. 2018.
[19] P. S. Zarrin, A. Escoto, R. Xu, R. V. Patel, M. D. Naish, and A. L. Trejos, "Development of an optical fiber-based sensor for grasping and axial force sensing," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Singapore, May 2017, pp. 939–944.
[20] F. X. Campion, G. Carlsson, and F. Francis, *Machine Intelligence for Healthcare*. Scotts Valley, CA, USA: CreateSpace Independent Publishing Platform. 2017.
[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.
[22] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, Aug. 2001.
[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[24] R. Entezari-Maleki, A. Rezaei, and B. Minaei-Bidgoli, "Comparison of classification methods based on the type of attributes and sample size," *J. Converg. Inf. Technol.*, vol. 4, no. 3, pp. 94–102, Sep. 2009.
[25] P. S. Zarrin and C. Wenger, "Pattern recognition for COPD diagnostics using an artificial neural network and its potential integration on hardware-based neuromorphic platforms," in *Proc. ICANN*, in Lecture Notes in Computer Science, Munich, Germany: Springer, Sep. 2019, pp. 284–288.
[26] C. Ellervik and J. Vaught, "Preanalytical variables affecting the integrity of human biospecimens in biobanking," *Clin. Chem.*, vol. 61, no. 7, pp. 914–934, Jul. 2015.
[27] P. S. Zarrin, N. Roeckendorf, and C. Wenger, "Exasens: A novel dataset for the classification of saliva samples of COPD patients," in *Proc. IEEE Dataport*, 2020. Accessed: Sep. 14, 2020, doi: 10.21227/7t0z-pd65.
[28] S. Aryal, E. Diaz-Guzman, and D. M. Mannino, "COPD and gender differences: An update," *Transl. Res.*, vol. 162, no. 4, pp. 208–218, Oct. 2013.
[29] C. A. V. Fragoso, "Epidemiology of chronic obstructive pulmonary disease (COPD) in aging populations," *COPD, J. Chronic Obstructive Pulmonary Disease*, vol. 13, no. 2, pp. 125–129, Mar. 2016.
[30] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations," in *Proc. NESUG Health Care Life Sci.*, Baltimore, MD, USA, vol. 19, 2010, p. 67.
[31] P. S. Zarrin, R. Zimmer, C. Wenger, and T. Masquelier, "Epileptic seizure detection using a neuromorphic-compatible deep spiking neural network," in *Proc. Int. Work-Conf. Bioinf. Biomed. Eng. (IWBBIO)*, Granada, Spain, 2020, pp. 389–394.
[32] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, and W. D. Lu, "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations," *Nature Electron.*, vol. 2, no. 7, pp. 290–299, Jul. 2019.
[33] C. Wenger, F. Zahari, M. K. Mahadevaiah, E. Perez, I. Beckers, H. Kohlstedt, and M. Ziegler, "Inherent stochastic learning in CMOS-integrated HfO_2 arrays for neuromorphic computing," *IEEE Electron Device Lett.*, vol. 40, no. 4, pp. 639–642, Apr. 2019.
[34] D. S. Jeong, K. M. Kim, S. Kim, B. J. Choi, and C. S. Hwang, "Memristors for energy-efficient new computing paradigms," *Adv. Electron. Mater.*, vol. 2, no. 9, 2016, Art. no. 1600090.
[35] P. S. Zarrin and C. Wenger, "Implementation of Siamese-based few-shot learning algorithms for the distinction of COPD and asthma subjects," in *Proc. ICANN*, in Lecture Notes in Computer Science, Bratislava, Slovakia: Springer, 2020.

POUYA SOLTANI ZARRIN received the master's degree in biomedical engineering from Western University, Canada, in 2017. Since 2017, he has been with the IHP Microelectronics, where he is currently working as a Research Scientist on medical device development and AI integration for precision diagnostics. He has expertise in the design, development, and testing of medical mechatronic systems and biosensors and implementation of machine learning techniques for medical analytics. His research interests include medical device design, sensing systems, AI, machine learning for healthcare, and medical mechatronics and robotics.

NIELS ROECKENDORF received the Diploma degree in chemistry and the Ph.D. degree from the Christian-Albrechts University of Kiel, in 1999 and 2003, respectively. Since 2004, he has been a Postdoctoral Research Scientist with the Research Center Borstel–Leibniz Lung Center. Since 2012, he has been acting as the Deputy Head of the Division of Mucosal Immunology and Diagnostics, Research Center Borstel.

CHRISTIAN WENGER received the Diploma degree in physics from the University of Konstanz, in 1995, and the Ph.D. and Postdoctoral degrees from the Dresden University of Technology (TU Dresden), in 2000 and 2009, respectively. Since 2002, he has been with the Innovations for High Performance Microelectronics (IHP), where he is currently working in the field of functional devices for medical and space applications. He has authored or coauthored more than 200 articles and holds six patents. In 2018, he received the Professorship Microelectronics for Medical Engineering at the Brandenburg Medical School Theodor Fontane and the Professorship Semiconductor Materials at BTU Cottbus-Senftenberg, in 2020.