2022 IEEE 19th India Council International Conference (INDICON) | 978-1-6654-7350-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/INDICON56171.2022.10040129

Text Based Diagnosis of COVID-19 Using Data Mining Techniques: A Comparative Study

Aadarsh Gupta[1*] Aastha Valecha[2*] Sapna Mishra[3] Tapan Gandhi[4]

[1, 3, 4]*Dept. of Electrical Engineering*
IIT Delhi
New Delhi, India

[2]*Dept. of Mechanical Engineering*
IIT Delhi
New Delhi, India

[1]aadarsh.iitd@gmail.com [2]aasthavalecha9@gmail.com [3]eez208443@iitd.ac.in [4]tgandhi@ee.iitd.ac.in

***Abstract*— In the course of the recent pandemic, we have witnessed non-clinical approaches such as data mining and artificial intelligence techniques being exceedingly utilized to restrain and combat the increase of COVID-19 across the globe. The emergence of artificial intelligence in the medical field has helped in reducing the immense burden on medical systems by providing the best means for diagnosis and prognosis of COVID-19. This work attempts to analyze & evaluate superlative models on robust data resources on symptoms of COVID-19, consisting of age, gender, demographic information, pre-existing medical conditions, and symptoms experienced by patients. This study establishes paradigmatic pipeline of supervised learning algorithms coupled with feature extraction techniques and surpassing the current state-of-the-art results by achieving an accuracy of 93.360. The optimal score was found by performing feature extraction on the data using principal component analysis (PCA) followed by binary classification using the AdaBoost classifier. In addition, the present study also establishes the contribution of various symptoms in the diagnosis of the malady.**

Keywords—COVID-19, symptoms, feature reduction, supervised learning, classification, diagnosis

I. Introduction

COVID-19 (COrona VIrus Disease of 2019), a new coronavirus infection disease brought on by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2) virus, was first detected in Wuhan City of China and quickly spread to the majority of countries across the globe. The World Health Organization (WHO) announced COVID-19 a global pandemic on January 30, 2020, owing to rapid person-to-person dissemination and susceptibility of individuals to the disease[9]. The malady has posed a slew of problems for the world today at various levels and the certain losses incurred on human civilization are irrecoverable. According to the WHO, fever, dry cough and breathing difficulties are some of the most frequent signs of COVID-19. However, the presence of the COVID-19 virus is also indicated by other symptoms such as headache, exhaustion, sore throat, bowel irritation, and nausea [1].

SARS-CoV-2 is prone to mutation and therefore, more varieties of the virus could have a significant impact on humans. Hence, the requirement for a faster, more accurate, and efficient COVID-19 diagnosis and detection mechanism arises. Devising a model to predict COVID-19 based on symptoms of a person is advantageous because it is computationally very efficient when compared to other techniques such as X-ray scans, audio, MRIs (Magnetic Resonance Imaging), etc., as they involve images and audio systems that are computationally quite demanding. In addition, audio-based diagnostic methods are prone to inaccuracies due to variation in recorder systems and noisy environments.

The complete understanding of the relationship between SARS-CoV-2 symptoms and virus infection is still a work in progress. Human symptoms based diagnosis frequently relies on previous similar presentations of diseases; nevertheless, this might introduce biases for expected symptoms and possibilities of carelessness, resulting in incorrect or late interpretations of symptoms, causes and preventions. With the moderation in the severity of symptoms, the demand for computationally low-resource-based diagnostic techniques that are easily accessible to the general population and have a short testing time emerges. Although biological testing methods such as nucleic acid-based detection remain the most common and trustworthy, AI-based detection approaches possess the potential to be employed as a precursor due to non-contact testing, minimal resource requirements and cost-effectiveness.

In this study, we employed an integration of machine learning algorithms and feature reduction techniques to establish a paradigmatic pipeline to determine the methods for predicting COVID-19 results based on the symptoms experienced by a person. We aim to discriminate amongst COVID-19 subjects, using information such as age and gender in addition to various symptoms including but not limited to, cold, cough, diarrhea, fever, loss of smell, muscle pain, and fever for preliminary COVID-19 detection test with a primary focus on subcontinental subjects due to availability of datasets for the region. The paper implements diverse sets of supervised algorithms to perform binary classification task and coupling them with deployed feature reduction methods such as Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA) and Isometric Mapping (ISOMAP) followed by

[*] These authors contributed equally

978-1-6654-7350-7/22/$31.00 ©2022 IEEE

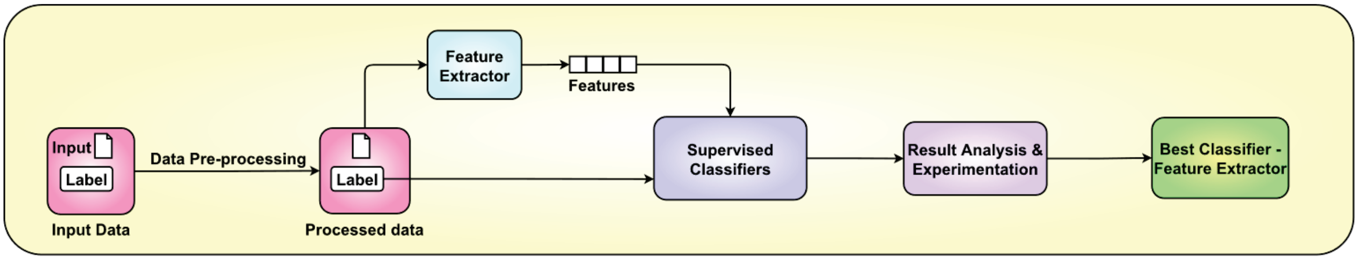

Fig. 1. Schematic pipeline for data mining techniques on the datasets

classification using antecedent algorithms in order to decrease the dimensionality of the input data. In this study, we also show that dimensionality reduction techniques help in optimizing computational time while also boosting model correctness and efficiency.

II. LITERATURE SURVEY

There have been multitudinous positive efforts in the past to identify links between symptoms of a disease and the likelihood or severity of the syndrome in the subjects. The ailments have ranged from SARS-CoV to MERS-CoV (Middle East respiratory syndrome-related coronavirus) and comparative analyses of MERS-CoV and SARS-CoV over spike glycoprotein show that the two viruses are comparable [8]. Extensive surveys conducted on the performance of various algorithms pertaining to MERS-CoV reveal several challenges that need to be addressed while deploying classification strategies for SARS-CoV-2. Text-based data mining techniques on MERS-CoV data reflect the superior performance of decision tree classifiers for binary classification over Naive Bayes and K-nearest neighbors algorithm (K-NN) with an accuracy of 90% [6]. COVID-19 detection techniques have been tried in a variety of ways [53] spanning from detection from computed tomography (CT) images [10] giving an accuracy of 96.23%, X-ray images providing results accurate upto 96.30% [11] and AI techniques on audio signals [3] including the multi-temporal Convolutional Neural Network (CNN) architectures [2].

Other significant CT (Computed Tomography) image-based research uses deep learning approaches [14], Support Vector Machine (SVM), and Random forest classifiers on a variety of datasets, including the COVID-19 chest X-ray data set [16] and automatic detection using X-Ray images [17]. The highest performance is reached by lung crop stage + Resnet-50, which displayed an accuracy of 99.6% [15]. However since the forenamed approaches require either CT or X-ray imaging data, these can only be employed at medical centers equipped with necessary infrastructure. Prior research has attempted to use spectrogram data from healthy and infected coughs to test a variety of machine learning algorithms, including ResNet-50 (Residual neural Network), Long short-term memory (LSTMs), Logistic Regression (LR), KNN, CNN, etc., [3] where Resnet50 architecture was implemented to achieve Area under the ROC curve (AUC-ROC) of 0.98 and LSTM to achieve an AUC-ROC of 0.94. The lack of a large and reliable dataset challenges our capabilities to develop robust methods which can recognize viral trends and features. Notable datasets available presently are: Coswara [12], Korea Centers for Disease Control & Prevention[1] (KCDC) and IATos [13]. Previous research has also shown the crucial role played by age and gender in the diagnosis of the infection [4]. Multi-modal Point-of-Care Diagnostic Methods for COVID-19 Based on Acoustics and Symptoms (MuDiCoV) [5] presented a formulation of multi-modal integration of the acoustic and symptoms classifiers, incorporating the principle that the diagnosis result is a culmination of the two classifiers and achieves an AUC of 0.924 with an accuracy of 0.74 over symptoms data. Logistic Regression and Decision Tree classifiers are utilized for the acoustic and symptoms data; however, a fusion approach based on mean of classifier scores presents a mathematically simpler, yet analytically uninterpretable approach to combination of multiple modalities. COVID-19 sensor-based detection [9] is a stride forward in the development of easily accessible, time and cost-effective systems that use superior algorithms such as transfer learning approaches.

This study demonstrates an efficient pipeline for conducting investigations involving data processing and modality selection to achieve the desired findings. In this work, we present an effective strategy for combining feature extraction and classification algorithms for low-resource databases of COVID-19 constituting widely accessible and computationally efficient symptoms data.

III. METHODOLOGY

The data mining workflow (as in Fig. 1) can be divided into three distinct phases: the preliminary stage involves gathering and cleaning of available data, the second phase incorporates experimentation with feature reduction techniques to extract key information, and the final phase involves the identification of workable classifier modalities and utilization of cross-validation to assess classification metrics. Furthermore, extracting meaningful characteristics from the modalities can serve as a useful basis for streamlining future diagnostic process.

A. Data

This study is primarily based on two datasets for our experimental analysis, namely the Coswara dataset [12] and Symptoms and COVID Presence[2].

1) Coswara Data: The Coswara project is aimed to create a COVID-19 diagnostic tool based on speech, cough, and breathe sounds. [12]. Acoustic and text data was collected from public participants and the text data utilized comprised age,

[1] https://www.kaggle.com/kimjihoo/coronavirusdataset/data

[2] https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence

present health status, geographic location, gender, and health issues that already exist.

2) Symptoms and COVID Presence Data: The dataset was created as a synthetic substitute based on a WHO report with an aim to study viral patterns in COVID-19 patients from the sub-continental region. The dataset, predominantly comprising of Indian subjects, incorporates the symptoms responses in binary form for 5434 subjects.

B. Dataset Preparation

Relevant attributes were retrieved from the original dataset after processing and cleaning independently for each data. This phase involves two stages to create the dataset: gathering and cleaning data. Coswara data contains 820 labelled samples, and 12 salient features were adopted based on past findings. The data was passed through an encoder layer to extract binary labels for features, while a ten-fold categorization has been used for age. In the case of the Symptoms and COVID Presence Data, 14 critical features were selected after analysis of feature importance from Random Forest classifier attributes. Moreover, data was passed through an encoder layer of subjects.

C. Feature Reduction Analysis

The importance of various features in the diagnosis of COVID-19 is not pre-determined and can be established by learning patterns in the data. Moreover, these can vary significantly with mutations in the biological structure of viruses. In this study, three feature reduction methods were utilized to reduce the feature space of high-dimensional data and to extract meaningful interpretations of patterns within the data. The feature reduction techniques used are Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) and Isometric Mapping (ISOMAP).

1) PCA: It is a statistical technique that optimizes dimensionality of data by reducing loss in variance in data [18]. Consider a given n-dimensional random vector (x) representing training set as:

$$\boldsymbol{x} = \{x_1, x_2, x_3 \dots x_n\} \tag{1}$$

Subsequently, we get a set of n n-dimensional orthogonal unit vectors, say $(\boldsymbol{u_1}, \boldsymbol{u_2}, \boldsymbol{u_3} \dots \boldsymbol{u_n})$ and form projections of x to construct, $\boldsymbol{a} = \{a_1, a_2, a_3 \dots a_n\}$ where, $a_i = \boldsymbol{x^T u_i}$ .The selection of unit vectors $(\boldsymbol{u_1}, \boldsymbol{u_2}, \boldsymbol{u_3} \dots \boldsymbol{u_n})$ is made in a manner so that projections, a_i contain decreasing variance. It can be shown [18] that the variance vector ($\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{\lambda_3} \dots \boldsymbol{\lambda_n}$) corresponds to the eigenvalues of the data covariance matrix $\boldsymbol{R}$ arranged in descending order, and $(\boldsymbol{u_1}, \boldsymbol{u_2}, \boldsymbol{u_3} \dots \boldsymbol{u_n})$ are the corresponding eigenvectors of $\boldsymbol{R}$. We reduce the dimensionality of the data from *n* to *p* by selecting the first p dimensions with the largest variance to generate data of reduced dimensionality [18]. PCA analysis is carried out by varying the number of principal components chosen, *n_component* which represents the final dimension of the extracted data. The parameter is varied for each dataset from 2 to n-1 where n represents the number of features in the data. The analysis is executed for various sets of classifiers and *n_component* corresponding to the superlative sets of classifiers is considered for subsequent execution.

2) UMAP: It creates a topological representation of high-dimensional data by utilizing local manifold approximations and combines together their local fuzzy simplicial set representations [19]. UMAP uses exponential probability distribution in high dimensions.

$$p_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}} \tag{2}$$

In a nutshell, UMAP generates a high-dimensional graph representation of the data before optimizing smaller dimensional graph to be structurally similar. The dimension of feature redacted data obtained after applying UMAP is 2.

3) ISOMAP: This technique transforms the dimensionality of data as a graph problem by extending the PCA and multidimensional scaling (MDS) to a class of nonlinear manifolds [29]. Analyses similar to PCA is performed for ISOMAP, however extensive computational and temporal requirements on the Symptoms and COVID Presence fails to serve the purpose well besides being outperformed by PCA.

D. Binary Classification System Design & Implementation

We implemented the following classifiers for binary classification: Logistic Regression, K-Nearest Neighbor (K-NN), Decision Tree, Random Forest, Gaussian Naive Bayes (GNB), Support Vector Machines (SVM), Bagging Classifier, AdaBoost Classifier, Gradient Boosting Classifier and Ensemble method of Stacking Classifier type for the afore-mentioned algorithms. The datasets being analyzed are split into train-test sets of 4:1, i.e., 80% taken as training set and 20% as the test set.

1) Logistic regression: The logistic regression is performed to establish the link between categorical dependent variables against the independent variables [7]. The logisctic regression model generates the prediction score (p) as,

$$p = \sigma\,(\boldsymbol{w^T x} + b\,) \tag{3}$$

where, $\boldsymbol{w}$ and b represent weight vector and the bias of the model respectively. Here, σ is the logistic function, $\sigma(a) = (1 + e^{-a})^{-1}$. The logistic regression model is trained by minimizing the cost function $E(.)$ [5] defined as

$$E\,(\boldsymbol{w}, b) = -\,[\,c\ log\,(p) + (1-c)\ log\,(1-p)\,] + \\ -\,[\,c\,log\,(p) + (1-c)\,log\,(1-p)] + \lambda\|\boldsymbol{w}\|_2^2 \tag{4}$$

where c denotes the class label of the feature vector x.

2) K-Nearest Neighbor: K-NN is a non-parametric and supervised classifier utilized for regression and classification tasks. The algorithm learns patterns from labeled data in order to give output class when fed with unlabelled data by a plurality vote of its neighbors [20].

3) Random Forest: This method is an ensemble learning technique which creates a large number of decision trees during training time and combines predictions from all trees for classification and regression tasks [7].

4) Gaussian Naive Bayes: It is a probabilistic classification algorithm that assumes a Gaussian distribution for real-valued data and uses the Bayes theorem for various classification tasks [25].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (5)$$

5) Support vector machine: The linear SVM (LinSVM) model generates prediction scores as,

$$p = f\ (\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + b) \quad (6)$$

where, $\boldsymbol{w}$ represents the weight vector and b, the bias of the model; and $f(\,)$ denotes the Platt scaling-based calibration [26]. The model is learned by minimizing the soft-margin cost function $E(.)$ defined as

$$E\ (\boldsymbol{w}, b) = \max\ (0, 1 - c\ (\mathbf{w}^{\mathbf{T}}\mathbf{x} + b)) + \lambda\ \|w\|^2 \quad (7)$$

where, c denotes the class label of the feature vector x.

6) AdaBoost Classifier: This classifier [22] is an ensemble classifier comprising multitude of weak classifiers [23], where each classifier performs classification based on only one-dimensionality of the input vector and the result can be expressed as:

$$H\ (\boldsymbol{x}) = sign\ \{\Sigma_{t=1}^{T}\beta_t h_t(x)\} \quad (8)$$

7) Gradient Boosting Classifier: This is a boosting technique that uses ensembles of weak prediction models to generate classification and regression models [26]. After training, each tree predicts a label, and the final prediction is as follows: [27]

$$y_{pred} = y_1 + (\eta * r_1) + (\eta * r_2) + \ldots\ldots. + (\eta * r_N) \quad (9)$$

8) Ensemble Method: Ensemble learning is a broader meta approach to machine learning by pooling the predictions from different models to boost the overall predictive performance. On evaluating the performance of each classifier before applying feature reduction techniques, we eliminated classifiers that performed poorly for each dataset separately. Subsequently, we performed the analysis by applying feature reduction techniques for superlative classifiers mentioned above on the datasets.

E. Performance Evaluation of Model

Accuracy, specificity, sensitivity, F1-score and AUC-ROC (Area under curve [28]) are among the most prevalent evaluation metrics used to evaluate the performance of any data mining model. However, in this study, accuracy has been used as the key criterion for evaluation, since the datasets taken exhibit relatively lower biasedness towards either class and therefore, all classes are assumed to have equal importance which makes accuracy a better metric to scale, interpret and summarize the capability of a model effectively.

For any given model, accuracy represents the number of correct predictions for all classes overall predictions. Meanwhile, precision and recall are measures of correct positive predictions overall positive predictions and all positive data values respectively, therefore providing an offset for class biases in data. F1 Score is taken as the harmonic mean of precision and recall as given in Fig. 2. and reduces when either the precision or recall is too low. AUC-ROC curve quantifies the degree of separation between two classes for a specific classifier. The performance of the classifier is proportional to the area under the curve.

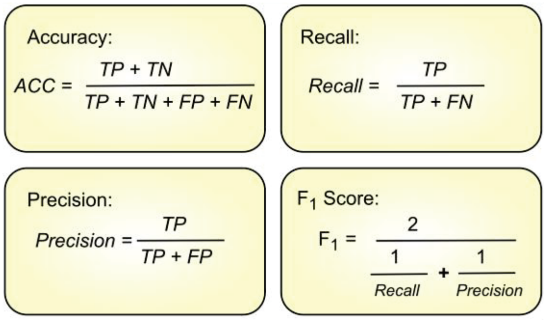

Fig. 2. Evaluation of common metrics from confusion matrix values

IV. FINDINGS AND DISCUSSIONS

A. Results on Different Classifiers

This work analyzes the accuracy of 8 superlative classifiers on the two datasets - Logistic Regression, k-NN, Random Forest, Gaussian Naïve Bayes, SVM, AdaBoost, GBMs and Ensemble classifiers gave noteworthy results on both the datasets as tabulated in Table 1 and Table 2.

TABLE 1. COSWARA DATA RESULTS

	Accuracy	F1 Score	Precision	Recall	AUC-ROC
LR	71.580	74.020	62.970	67.830	79.070
K-NN	69.880	72.500	60.150	65.500	74.870
RF	70.730	70.410	68.360	68.970	74.470
GNB	70.970	74.190	63.220	67.530	**80.220**
SVM	70.970	74.010	63.220	67.440	78.960
AdaBoost	70.850	**74.080**	60.660	66.450	78.070
GBM	**72.310**	72.350	**69.900**	**70.700**	77.630
Ensemble	71.940	73.100	67.590	69.630	78.600

TABLE 2. SYMPTOMS AND COVID PRESENCE DATA RESULTS

	Accuracy	F1 Score	Precision	Recall	AUC-ROC
LR	90.080	**96.020**	91.440	93.580	94.050
K-NN	91.040	93.320	95.850	94.480	85.010
RF	**93.300**	95.400	**96.370**	**95.870**	93.540
GNB	65.400	64.123	57.470	72.520	94.810
SVM	92.360	94.970	95.620	95.290	94.060
AdaBoost	88.630	94.310	91.330	92.790	93.920
GBM	92.640	95.400	95.460	95.420	**95.020**
Ensemble	90.610	93.870	94.480	94.130	93.590

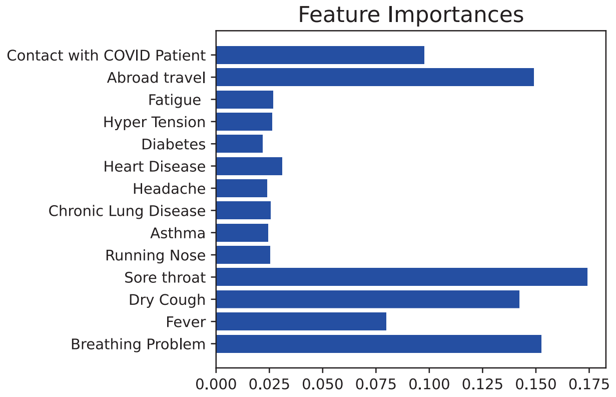

Fig. 3. Bar graph for importance of various features

Note: Calculation of feature importance from decision tree and random forest (shown in Fig. 3): The probability of a node can be determined by dividing the number of samples that reach the node by the total number of samples. The greater the value, higher is the importance of the feature.

B. Feature Reduction Methods

We used three feature reduction methods: PCA, UMAP and ISOMAP. It is concluded that PCA outperforms the viable feature reduction strategies implemented while also being the most efficient in contrast to the drop in performance in UMAP. ISOMAP did perform admirably well, however computational limitations for the technique result in PCA being chosen as the superlative technique in this experiment. Optimal results for PCA are observed when 7 principal components are chosen, i.e., *n_components* = 7 (as in Fig. 4). Inspection of Table 3 confirms aforementioned conclusion of superlative performance of PCA on Symptoms and COVID Presence Data compared to UMAP and ISOMAP. Fig. 5 shows that PCA when coupled with Adaboost classifier gives the best performance with an accuracy of 93.630.

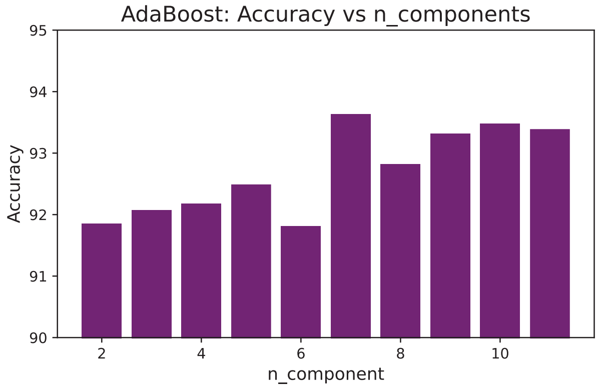

Fig. 4. Accuracy for number of principal components: AdaBoost Classifier

TABLE 3. ACCURACY OF MODELS ON SYMPTOMS AND COVID PRESENCE DATA USING FEATURE REDUCTION TECHNIQUES

	Without FR	**PCA**	**UMAP**	**ISOMAP**
LR	90.080	91.310	80.180	89.930
K-NN	91.040	89.510	86.910	88.590
RF	**92.910**	92.640	**88.830**	91.150
GNB	65.400	92.400	81.970	90.650
SVM	92.360	90.830	83.270	90.030
AdaBoost	88.630	**93.630**	84.930	92.030
GBM	92.340	93.020	88.040	**92.510**
Ensemble	91.080	92.360	87.670	89.860

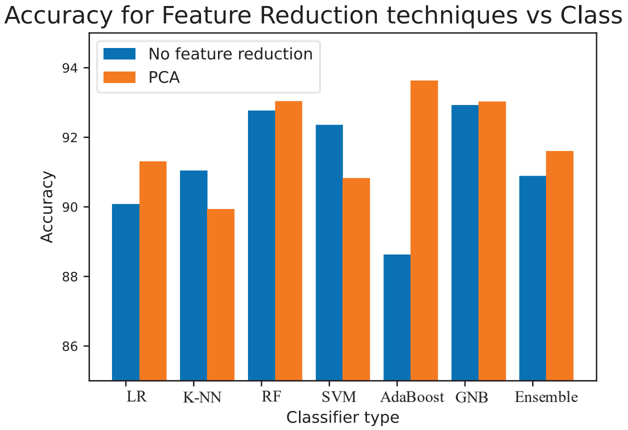

Fig. 5. Accuracy for efficient classifiers: PCA vs No feature reduction

V. CONCLUSION

We have performed extensive experimentation with feature reduction techniques and classification methods for COVID-19 diagnosis on different text datasets. The paradigmatic strategy for classification of symptoms data for diagnostic tasks of COVID-19 presents superior results and we infer that utilizing AdaBoost classifier in conjunction with PCA surpasses state-of-the-art performance on the task. This study serves as a benchmark for further enhancements to the diagnosis of COVID-19 and also aids for development of preliminary and portable diagnostic tools for the disease.

REFERENCES

[1] "Weekly epidemiological update on COVID-19 - 15 March 2022." https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---15-march-2022 (accessed Oct. 05, 2022).

[2] T. V. Kumar, R. S. Sundar, T. Purohit, and V. Ramasubramanian, "End-to-end audio-scene classification from raw audio: Multi time-frequency resolution CNN architecture for efficient representation learning," in *International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, pp. 1–5, Jul. 2020, doi: 10.1109/SPCOM50965.2020.9179600.

[3] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *Computers in Biology and Medicine*, vol. 135, p. 104572, Aug. 2021, doi: 10.1016/j.compbiomed.2021.104572.

[4] F. Asci *et al.*, "Machine-Learning Analysis of Voice Samples Recorded through Smartphones: The Combined Effect of Ageing and Gender," *Sensors*, vol. 20, no. 18, p. 5022, Jan. 2020, doi: 10.3390/s20185022.

[5] S. R. Chetupalli *et al.*, “Multi-modal Point-of-Care Diagnostics for COVID-19 Based On Acoustics and Symptoms.” arXiv, Jun. 05, 2021. doi: 10.48550/arXiv.2106.00639.

[6] H. Kurdia and N. AlMansour, “Identifying accurate classifier models for a text-based MERS-CoV dataset,” in *Proc. of Intelligent Systems Conference (IntelliSys)*, London, pp. 430–435, Sep. 2017. doi: 10.1109/IntelliSys.2017.8324330.

[7] L. J. Muhammad, Md. M. Islam, S. S. Usman, and S. I. Ayon, “Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients’ Recovery,” in *SN COMPUT. SCI.*, vol. 1, no. 4, p. 206, Jul. 2020, doi: 10.1007/s42979-020-00216-w.

[8] J. E. Huh, S. Han, and T. Yoon, “Data mining of coronavirus: SARS-CoV-2, SARS-CoV and MERS-CoV,” *BMC Research Notes*, vol. 14, no. 1, p. 150, Apr. 2021, doi: 10.1186/s13104-021-05561-4.

[9] H. S. Maghdid, K. Z. Ghafoor, A. S. Sadiq, K. Curran, D. B. Rawat, and K. Rabie, “A Novel AI-enabled Framework to Diagnose Coronavirus COVID 19 using Smartphone Embedded Sensors: Design Study.” arXiv, May 30, 2020. doi: 10.48550/arXiv.2003.07434.

[10] S. Walvekar and D. S. Shinde, “Detection of COVID-19 from CT Images Using resnet50.” Rochester, NY, May 30, 2020. doi: 10.2139/ssrn.3648863.

[11] R. Shree Charran and R. K. Dubey, “Chapter 1 - Deep learning-based hybrid models for prediction of COVID-19 using chest X-ray,” in *Novel AI and Data Science Advancements for Sustainability in the Era of COVID-19*, V. Chang, M. Abdel-Basset, M. Ramachandran, N. G. Green, and G. Wills, Eds. Academic Press, pp. 1–20, 2022. doi: 10.1016/B978-0-323-90054-6.00001-5.

[12] N. Sharma *et al.*, “Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis,” in *Interspeech*, pp. 4811–4815, Oct. 2022. doi: 10.21437/Interspeech.2020-2768.

[13] D. T. Pizzo and S. Esteban, “IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples.” arXiv, Dec. 09, 2021. doi: 10.48550/arXiv.2104.13247.

[14] F. Shan *et al.*, “Lung Infection Quantification of COVID-19 in CT Images with Deep Learning,” *Med. Phys.*, vol. 48, no. 4, pp. 1633–1645, Apr. 2021, doi: 10.1002/mp.14609.

[15] O. Gozes *et al.*, “Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis.” arXiv, Mar. 24, 2020. doi: 10.48550/arXiv.2003.05037.

[16] L. Wang and A. Wong, “COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images.” arXiv, May 11, 2020. Accessed: Oct. 05, 2022. [Online]. Available: http://arxiv.org/abs/2003.09871

[17] I. D. Apostolopoulos, S. I. Aznaouridis, and M. A. Tzani, “Extracting Possibly Representative COVID-19 Biomarkers from X-ray Images with Deep Learning Approach and Image Data Related to Pulmonary Diseases,” *J. Med. Biol. Eng.*, vol. 40, no. 3, pp. 462–469, Jun. 2020, doi: 10.1007/s40846-020-00529-4.

[18] J. Möcks, “The Influence of Latency Jitter in Principal Component Analysis of Event-Related Potentials,” in *Psychophysiology*, vol. 23, no. 4, pp. 480–484, Jul. 1986, doi: 10.1111/j.1469-8986.1986.tb00659.x.

[19] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” arXiv, Sep. 17, 2020. doi: 10.48550/arXiv.1802.03426.

[20] D. Kumar, T. B. Singh, P. Ghanghoria, and V. Ghanghoria, “Pregnancy related complications and its association with socio-demographic factors in Central India: A logistic regression hypothesis,” *JCHM*, vol. 6, no. 3, pp. 72–76, Oct. 2019, doi: 10.18231/j.jchm.2019.017.

[21] O. Harrison, “Machine Learning Basics with the K-Nearest Neighbors Algorithm,” *Medium*, Jul. 14, 2019. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (accessed Oct. 05, 2022).

[22] Y. Freund and R. E. Schapire, “Experiments with a New Boosting Algorithm.” Proceedings of the 13th Conference on Machine Learning, 1996.

[23] W.-C. Cheng and D.-M. Jhan, “A cascade classifier using Adaboost algorithm and support vector machine for pedestrian detection,” in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1430–1435, Oct. 2011. doi: 10.1109/ICSMC.2011.6083870.

[24] R. Aler, I. M. Galván, J. A. Ruiz-Arias, and C. A. Gueymard, “Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting,” *Solar Energy*, vol. 150, pp. 558–569, Jul. 2017, doi: 10.1016/j.solener.2017.05.018.

[25] R. Gandhi, “Naive Bayes Classifier,” *Medium*, May 17, 2018. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (accessed Oct. 05, 2022).

[26] J. C. Platt, “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods,” in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.

[27] M. Telgarsky, “Margins, Shrinkage, and Boosting.” arXiv, Mar. 18, 2013. doi: 10.48550/arXiv.1303.4172.

[28] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing Imbalanced Data Recommendations for the Use of Performance Metrics,” in *Proc. of Int Conf Affect Comput Intell Interact Workshops*, vol. 2013, pp. 245–251, 2013, doi: 10.1109/ACII.2013.47.

[29] O. Samko, A. D. Marshall, and P. L. Rosin, “Selection of the optimal parameter value for the Isomap algorithm,” Pattern Recognition Letters, vol. 27, no. 9, pp. 968–979, Jul. 2006, doi: 10.1016/j.patrec.2005.11.017.