IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

An Early Detection of Asthma using BOMLA Detector

MD. ABDUL AWAL[1], MD. SHAHADAT HOSSAIN[2], KUMAR DEBJIT[3], NAFIZ AHMED[4], RAJAN DEV NATH[5], GM MONSUR HABIB[6], MD. SALAUDDIN KHAN[7], MD. AKHTARUL ISLAM[8], AND M. A. PARVEZ MAHMUD[9]

[1]Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208, Bangladesh
[2]Department of Quantitative Sciences, International University of Business Agriculture and Technology, Dhaka-1230, Bangladesh
[3]Faculty of Health, Engineering and Sciences University of Southern Queensland, Australia
[4]Faculty of Law and Business (Peter Faber Business School), Australian Catholic University, Australia
[5]Faculty of Business, Education, Law and Arts (School of Commerce), University of Southern Queensland, Australia
[6]Global Health Academy, The University of Edinburgh, Edinburgh, Scotland, United Kingdom
[7]Statistics Discipline, Khulna University, Khulna-9208, Bangladesh
[8]Statistics Discipline, Khulna University, Khulna-9208, Bangladesh
[9]School of Engineering, Deakin University, Geelong, VIC-3216, Australia

Corresponding author: Md. Abdul Awal (e-mail: m.awal@ece.ku.ac.bd).

ABSTRACT Asthma is a chronic and airway-induced disease, causing the incidence of bronchus inflammation, breathlessness, wheezing, is drastically becoming life-threatening. Even in the worst cases, it may destroy the quality to lead. Therefore, early detection of asthma is urgently needed, and machine learning can help identify asthma accurately. In this paper, a novel machine learning framework, namely BOMLA (**B**ayesian **O**ptimisation-based **M**achine **L**earning framework for **A**sthma) detector has been proposed to detect asthma. Ten classifiers have been utilized in the BOMLA detector, where Support Vector Classifier (SVC), Random Forest (RF), Gradient Boosting Classifier (GBC), eXtreme Gradient Boosting (XGB), and Artificial Neural Network (ANN) are state-of-the-art classifiers. In contrast, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QLDA), Naive Bayes (NB), Decision Tree (DT), and K-Nearest Neighbor (KNN) are conventional popular classifiers. ADASYN algorithm has also been employed in the BOMLA detector to eradicate the issues created due to the imbalanced dataset. It has even been attempted to delineate how the ADASYN algorithm affects the classification performance. The highest accuracy (ACC) and Matthews's correlation coefficient (MCC) for an Asthma dataset provide 94.35% and 88.97%, respectively, using BOMLA detector when SVC is adapted, and it has been increased to 96.52% and 93.04%, respectively, when ensemble technique is adapted. The one-way analysis of variance (ANOVA) has also been performed in the 10-fold cross-validation to measure the statistical significance. A decision support system has been built as a potential application of the proposed system to visualize the probable outcome of the patient. Finally, it is expected that the BOMLA detector will help patients in their early diagnosis of asthma.

INDEX TERMS Classification, Clinical and Non-clinical data, Asthma, ADASYN, ANOVA.

I. INTRODUCTION

ASTHMA is one of the chronic lung diseases that inflame the airways and insists bronchi swell. Subsequently, it narrows down the airways and ultimately makes a person hard to breathe. Although asthma symptoms can vary from person to person, common symptoms may include breathlessness [1], [2], wheezing, tightening the chest or chest pain. The asthma severity can be reduced by avoiding some common triggers such as allergies [3], smoke, tobacco, wood fires, air pollution, cold air, dust mites, pollen, chemicals fumes, nasal polyps, pneumonia, sinusitis, an infection like colds and flu, the pungent odors, e.g., perfume. However, it is not always possible to avoid these triggers. On the other hand, people of all ages may be affected by asthma, especially children and age-old people may develop this disease without any permanent cure; however, adults are not out of danger. People get into trouble talking and being active for a long time. Consequently, it hampers their health and productivity, especially service holders. Therefore, proper management and detection of asthma are ultimately needed [4], [5].

It is approximated that over 300 million people are infected by asthma over the globe [6]. It is also forecasted that there will be approximately 100 million asthma patients by 2025 [6]. According to the World Health Organization (WHO), every year, the pre-mature death due to asthma is nearly 250,000. And still, now, asthma remains a poorly controlled disease [7]. Therefore, early detection of asthma is essential, especially in developing countries and rural areas, where a healthy environment is a significant concern [8], [9].

Due to the overwhelming innovations in medical science and intelligent diagnosis systems [10], a massive amount of data is being generated. This data can be processed and used to build a machine learning model and diagnose asthma more efficiently and early than conventional techniques. To get a better intuition of the mechanisms of asthma, sometimes, a more complex model can be used [11]. This is because we should take performance, complexity and interpretability into consideration to test a system.

Machine learning is a general field of artificial intelligence, which allows us to learn from available data and predicts the unknown targets [2]. A proper prediction of asthma disease progression and early identification of asthma patients creates the opportunity for better treatment and maintains this disease's stability [12]–[15]. To warn about the poor outcomes and identify asthma patients for care management, predictive models are widely exercised as the best method [16]–[18]. Many researchers have done numerous researches to identify asthma patients using different machine learning algorithms. For evidence, Dexheimer et al. [19] attempted to compare different classifiers and found that the expert-based model was better than others and: Bayesian Network (BN) achieved the highest accuracy. Prasad et al. [20] used different machine learning algorithms to identify asthma and found neural networks better than the others. To clinically trace various respiratory diseases like asthma and chronic obstructive pulmonary disease (COPD), several studies were conducted using different artificial neural network (ANN), machine learning techniques using different datasets with other features [21]–[23]. In 2017, Spathis and Vlamos [24] used some machine learning algorithm and demonstrated that the RF algorithm is the highest performing classifier in diagnosing COPD and asthma with 97.7% and 80.3% precision. In 2018, Yahyaoui and Yumusak [25] used an adaptive support vector machine algorithm to diagnose asthma and COPD with an accuracy of 98.45% for asthma and 92.63% for COPD. This study does not incorporate sophisticated optimization and data-balancing, improving the classification accuracy and improving the artificial intelligence (AI) based diagnosis system. None of the above studies built a decision support system (DSS), which can be very beneficial for the clinical staff and the end-users. Motivated by this, the state-of-the-art machine learning algorithms, along with an efficient optimization technique, have been adapted to predict asthma and provide a methodology to design a DSS.

This study examines the factors that characterize asthma diagnosis and its prediction using different machine learning algorithms such as RF, DT, SVC, etc. The major contributions and pivotal topics of our proposed study are provided as follows:

- A novel dataset has been collected for the analysis (see Section II-A).
- It has been endeavored to design a framework based on Machine Learning, known as BOMLA, to trace asthma patients, where the Bayesian optimization (BO) algorithm has been utilized in Section II.
- In the proposed detector, the ADASYN algorithm has been used to exterminate the existing imbalance between the classes of the dataset. How the ADASYN algorithm influences the entire framework's performance has also been illustrated in Section III-A.
- This paper also presents the effect of combining different optimized classifiers to enhance the classification performance through ensemble technique [see Section III-E.]
- Important features from the dataset are calculated, and the cumulative influence of features are explained (see Section III-F).
- The BOMLA optimizes the high-tech machine learning framework, and comparison has also been drawn with the conventional search algorithms, such as random search and grid-search techniques in Section III-G.
- A DSS has been built as a potential application of the BOMLA detector in Section IV-B.

The rest of the paper is organized as follows. The materials and methods are illustrated in Section II. We present the findings of the proposed study in Section III. A comparative study with other methods is drawn in Section IV along with an application of the proposed BOMLA detector. In the end, some conclusions are drawn in Section V.

II. MATERIALS AND METHODS

A. DATA COLLECTION

We have selected real-world Asthma datasets through the clinical study conducted in Khulna, Bangladesh. Since there was no systematic research-based information on maternal in the community, authors and trained research assistants used a semi-structured questionnaire after an appropriate explanation of the study's purpose and consent obtained from the respondents. This study was conducted according to the Declaration of Helsinki guidelines and approved by the "Data Acquiring Ethics Evaluation Committee (DAEEC)" of Community respiratory centre, Khulna, Bangladesh. The dataset contains 389 persons with nine attributes: AGE, SPO2, PULSE (P), FEV, FEV_percent, FVC, FVC_percent, FEV1_by_FVC, MEF2575. The explanation of the attributes is given in Supplementary Material-I. The samples are collected from patients of different ages ranging from 11 to 72. The datasets are stored in excel format for preprocessing the data grouped into two classes regarding whether a patient has asthma positive or negative recorded. There were 90.49% asthma positive and 9.61% asthma negative in the dataset.

Therefore, there is an imbalanced nature in the dataset where the asthma positive class is immensely higher than the asthma negative class. This is because normally, patients with asthma symptoms come to clinics for medical treatments and tested as asthma positive.

To get better intuition regarding the distribution of the entire dataset, the box plot is not enough because, when data morphing occurs, the box plot remains the same. To get rid of this complexity, the violin plot representation could be the best choice to visualize the statistical distribution of the dataset because the violin plot displays the mean, interquartile range and outliers of the dataset, and at the same time, the distribution of the whole dataset [26]. Besides, the wider portion of the violin represents the more significant probability, while the thinner part refers to the more negligible probability of the classes. In our proposed BOMLA detector, an illustration regarding the target variable's distribution against all the numeric variables has been given. The violin representations of both Asthma-yes and Asthma-no classes have been split into Male and Female [Figure 1]. The overall observation shows that SPO2 has long-tail distribution below the first quartile, whereas FVC (%) and MEF2575 has a long-tail above the third quartile.

1) Clinical Interpretation of Asthma and the Use of Spirometry

Asthma inflames the air duct with variable symptoms and patterns of the course of the disease [27]. Variability of the airway calibres resulting in the noisy chest (wheezing) is one of the cardinal features of asthma [27]. To diagnose asthma, both the subjective evaluation by history and objective confirmation to establish the airflow limitation's variable nature is essential [28]. Subjective evaluation is done by good clinical history and physical examination to suspect the probable diagnosis of asthma. However, the confirmation of variable airflow limitation, a cardinal feature of asthma is done by Spirometry as the gold standard [29], although other cheaper method like a peak flow meter is used as a bedside examination tool; however, it is not the substitute of Spirometry. Therefore, this study uses data recorded from Spirometry. The flow chart of the diagnosis of asthma is shown in Figure 2.

The entire working architecture of our proposed framework has been clarified in Figure 3. The first step is data collection, which has been followed by data imputation. Data was then over-sampled using the ADAptive SYNthetic (ADASYN) algorithm and proceeded to data scaling before splitting them into train and test dataset. Based on the extant literature, ten known classification algorithms, such as RF, XGB, ANN, GBC, SVC, LDA, QLDA, NB, KNN, and DT have been applied in this research. Additionally, the BO has been used to tune the hyperparameters of the classification algorithm. Shortly afterwards, the statistical analysis has been performed by applying the Boxplot analysis and Analysis of Variance (ANOVA). The performance of different classifiers has been measured by the confusion matrix and the 10-fold cross-validation.

Firstly, features are normalized using the z-score method. As this is a real-world dataset, and there are several missing values, we have to impute the dataset using the KNN imputation technique. This technique imputes the missing values from the nearest-neighbor column using the Euclidean distance. The MATLAB function *'filloutliers'* has been used to detect and fill outliers with the previous non-outlier element. The ADASYN algorithm having $K = 5$ neighbors is used to balance out the classes by generating an adequate amount of synthetic data, which helps eradicate the overfitting issue [30]. The weighted distribution for the minority classes is used according to their difficulty learning to generate synthetic data. The benefits of using ADASYN are: (1) it reduces the biasness solving class imbalance problem, and (2) it also fit the classification boundary adaptively to complex examples. It uses data density distribution ($\hat{r}_i$) of the i^{th} minority class to automatically decide the number of synthetic examples that need to be generated. In other words, $\hat{r}_i$ measures the characteristics of the data-dependent distribution of weight for the i^{th} minority class according to the level of *"harder- to-learn"*. Unlike SMOTEBoost (Synthetic Minority Over-sampling TEchnique Boost) and DataBoost-IM, the ADASYN does not require a hypothesis to generate synthetic data [30]. Appendix A provides the algorithm of ADASYN, which describes the sample generation process step-by-step procedure.

B. CLASSIFIERS USED IN THE FRAMEWORK

Ten classifiers have been applied in the BOMLA detector, where LDA, QLDA, NB, KNN, and DT are popular classifiers. In contrast, RF, XGB, ANN, GB, and SVC (also known as SVM) are state-of-the-art classifiers, which, in general, provide outperforming accuracy in many classification problems, such as COVID-19 patients classification. All the classifiers except LDA and QLDA have some hyperparameters, which improve classification accuracy. Note that the variants of LDA and QLDA, such as Regularized LDA, Reduced-rank LDA have hyperparameters; however, we have not used the variant of LDA and QLDA. The intention of adding these two classifiers is that they are straightforward and easy to construct. However, due to linear classifier, these two classifiers do not provide excellent accuracy. We have used them to show how the complex algorithm with hyperparameters can outperform the linear classifiers, such as LDA and QLDA. And these hyperparameters have been optimized to solve our proposed optimization problem by proposing the BO-based framework.

C. REQUIREMENT OF OPTIMIZATION

As mentioned above, all the classifiers except LDA and QLDA have several hyperparameters that control the classification performance. A short description of the hyperparameters and their role in the proposed BOMLA detector is given in the following paragraph. Note that only important

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3073086, IEEE Access

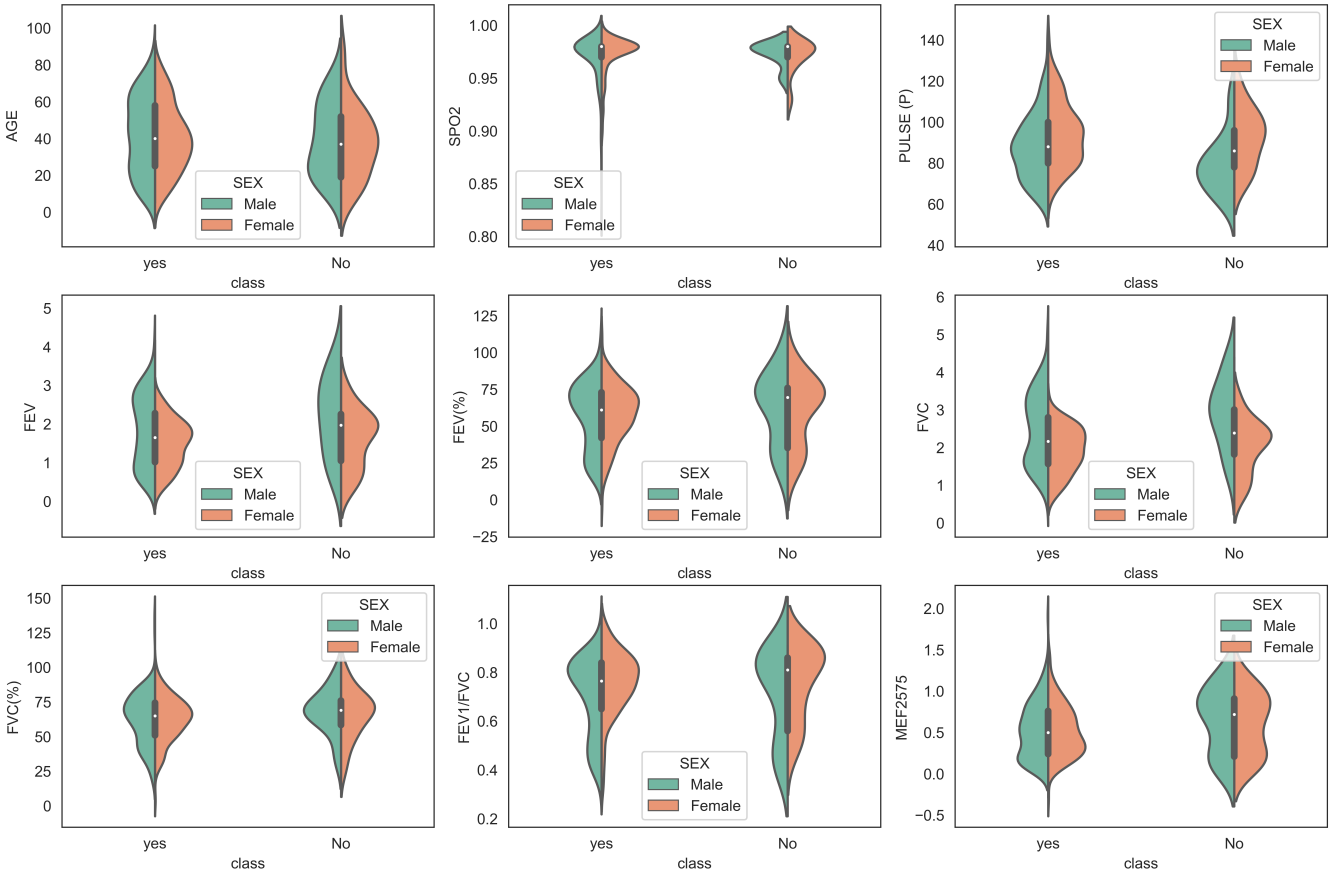


Figure 1: The violin plot representation of the entire dataset.

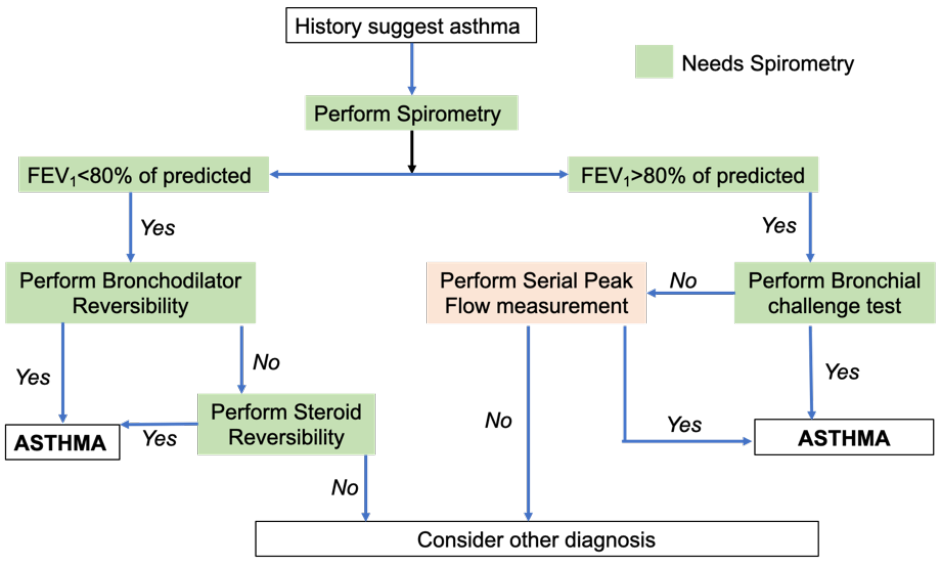


Figure 2: Flowchart of the diagnosis of asthma and the use of spirometry.

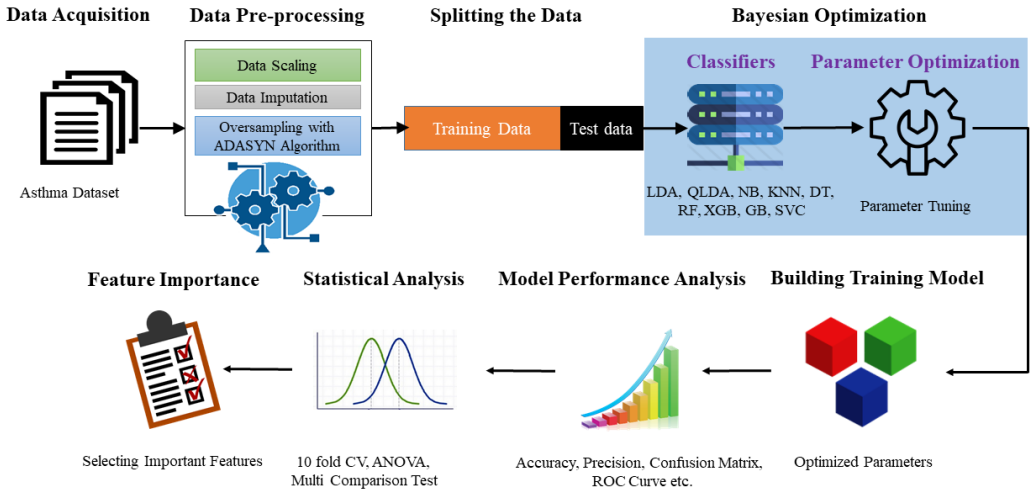


Figure 3: The overall workflow of the study.

hyperparameters among several controllable parameters have been considered to tune by the BOMLA detector.

To commence, two hyperparameters have been tuned while using RBF-SVC in the proposed framework: *gamma* (γ) is the coefficient for RBF kernel and *Cost* (C) is the regularization parameter. The RF and DT are tree-based classifiers. In the case of RF, four hyperparameters have been used: *criterion* is tree-specific, which is used to find the quality of the split, *max_depth* represents the maximum number of levels, *max_features* represent the maximum number of features used to find the best split, and *n_estimators* specify the number of trees. The important hyperparameters of DT are *criterion, max_depth, max_features*. Furthermore, GBC and XGB are ensemble-based classification techniques. In the case of GBC, four important hyperparameters have been used: *learning_rate* is used to shrink the contribution of each tree, *loss* refers to deviance (logistic regression), *n_estimators* are the total boosting stages, *max_depth* represents the maximum depth of the estimators, and *max_features* refer to the total features to be used for classification. There are seven important hyperparameters of XGB have been optimized using BOMLA, and they are: *n_estimators* define the total boosting stages, *max_depth* represents the maximum depth of the estimators, *learning_rate* is used to shrink the contribution of each tree, *Gamma* refers to the minimum loss to further split of the leaf, the minimum sum of instance weight is denoted as *min_child_weight*, subsampling of columns is marked as *colsample_by_tree*, and *n_jobs* represent the number of parallel threads. Furthermore, in ANN algorithm using MLP (Multilayer Perceptron), the *neurons* in the i^{th} hidden layer transforms the weighted values to the next layer, *solver*, such as Stochastic Gradient Descent (SGD), ADAM, L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) algorithm is used to update the gradient of the lost function, *alpha* is the regularization term, and the *learning_rate* is used to schedule the weight updates.

Besides, one hyperparameter has been tuned while using Naive Bayes (Alpha) and KNN (*number_of_neighbors*) classifiers. The optimal values of the hyperparameter mentioned above can improve the classification performance, e.g., accuracy (ACC), error, specificity (SP), sensitivity (SE), and these values are mainly depending on the classification data and classification problem. The general framework of the proposed BOMLA detector, comprised of the hyperparameters mentioned above, can be written as:

$$\arg\min_{p \in P} J(Clf(P); P) \tag{1}$$

where $p \in P$ represents the hyperparameters of the classifiers, such as $p_1, p_2, p_3, ...p_n \in P$, and Clf characterizes the machine learning classifiers, e.g., GBC, RF, XGB, etc. The Eq. (1) can be described as the minimization of the cost function $J(.)$ by selecting the proper value of P. For an imbalanced dataset, e.g., the dataset used in this study, the main goal is to maximize recall (=sensitivity) without sacrificing PPV (=precision) [31]. The F1Score takes both of these measures into account and weights equally as F1Score is the harmonic mean of recall and PPV; see Eq. (10). As our proposed BOMLA works on the minimization problem, we have used the average of $K = 10$ -fold cross-validation F1Score loss $(F1loss_{CV})$ calculated from the training dataset as the cost function, $J(.)$ in this study, which can be expressed as:

$$F1loss_{CV} = \frac{1}{K}\sum_{k=1}^{K}(1 - F1Score_k(Clf(P))) \tag{2}$$

In the next section, we will discuss the BO technique, on which the BOMLA detector is built up [32].

D. BAYESIAN OPTIMIZATION

In general, the BO is a global optimization technique superior to grid search, manual search, exhaustive search, and random research [33]. In this study, we have used $F1loss_{CV}$ as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3073086, IEEE Access

the cost function, which is time-consuming. In that case, BO is a suitable technique rather than other meta-heuristic algorithms, such as Harris Hawks optimization [34], slime mould optimization [35]. Considering these rationales, we have used the BO technique for the proposed BOMLA detector in this study. The advantage of BO is that it can memorize the previous evaluation and tune the hyperparameter in the probabilistic gaussian fashion. Our proposed framework has used the Hyperparameter Optimization toolbox (in short, HyperOpt) developed by Bergstra et al. [36]. The fundamental steps of adapting HyperOpt in the BOMLA detector are as follows:

- Cost function minimization,
- Search Space,
- Iterations numbers and
- The search algorithm to use.

A brief explanation of the steps, as mentioned above, is added below:

1) Step-1: Define a Cost Function

Hyperopt is a convenient open-source Python library to minimize the objective function. As mentioned earlier, the mean of the 10-fold cross-validation F1Score loss, $F1loss_{CV}$ calculated from the training dataset, is utilized as the cost function. To illustrate, in the case of XGB, seven important hyperparameters have been optimized using BO. The detail of the hyperparameters has already been discussed in Section II-C. The BO algorithm selects the optimal value of these seven XGB hyperparameters for which the cost function $J(XGB; (XGB\ Hyperparameters))$ provides the minimum value.

2) Step-2: Search Space

Hyperopt is allowed to search over the configuration space, and in this step, both the upper limit and the lower limits of the hyperparameter should be predefined (for instance, $[0, 1]$). In this paper, we have used several classifiers, and the upper limit and the lower limit have been defined based on our previous experience.

3) Step-3 and 4: Number of iteration and Choose a Search Algorithm

The BO, generally, explores the appropriate set of parameters using the previously gained information. In this algorithm, a pair of variants are considered: the first relies on the Gaussian process and another on the Tree Parzen Estimator (TPE). In our research, the TPE algorithm is used by the HyperOpt package to accomplish the optimization process (choosing a search algorithm). The search algorithms are treated as global functions, having extra keyword arguments to control their operation and iterate 50 times. The step-wise process of hyperparameter optimization is provided in Appendix B.

E. STATISTICAL EVALUATION OF CLASSIFICATION MEASURES

The performance of all the classifiers are measured by different measurement factors, such as ACC, SE, SP, confusion matrix, ANOVA, ROC curve, recall vs decision boundary, and the 10-fold cross-validation.

1) Confusion Matrix

The Confusion matrix or error matrix is one of the most useful techniques [37], which has been used to visualize the classifiers' overall performance [38]. The confusion matrix comprises two rows and two columns representing the number of false-negative, false-positive, true positive, true negative, which can be shortly denoted as FN, FP, TP, TN. The common as well as robust measures for classification, calculated from the confusion matrix, which can be expressed as follows [equations (3) to (11)]:

$$ACC(\%) = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (3)$$

$$Error : E(\%) = (1 - ACC) \times 100\% \quad (4)$$

$$Sensitivity : SE(\%) = \frac{TP}{TP + FN} \times 100\% \quad (5)$$

$$Specificity : SP(\%) = \frac{TN}{TN + FP} \times 100\% \quad (6)$$

$$PPV(\%) = \frac{TP}{TP + FP} \times 100\% \quad (7)$$

$$NPV(\%) = \frac{TN}{FN + TN} \times 100\% \quad (8)$$

MCC(%)=

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100\% \quad (9)$$

$$F_{1\_}score(\%) = \frac{2(PPV \times SE)}{PPV + SE} \times 100\% \quad (10)$$

Here, PPV and NPV represent Positive Predictive Value (also called Precision) and Negative Predictive Value, respectively. We have also added the Kappa index to differentiate between observed accuracy and expected accuracy, and it is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \quad (11)$$

where p_o and p_e denote the observed agreement and the expected agreement, respectively. Another worth-noting point is that a higher value of ACC, F1_score, SE, MCC, SP, and Kappa index, and the lower value of error indicate a better model.

2) Box-plot and Analysis of Variance (ANOVA)

Box plot is a graphical delineation of groups of numerical data according to its quartiles (or percentiles), enabling us to get the intuition of our data better [39] through summarizing five numbers: maximum, third quartile, median, first quartile and minimum. The median value divides the box into two parts, where the extreme lines represent the highest and the lowest value, excluding outliers.

ANOVA is conceptually the most straightforward technique to determine statistical differences between the means of multiple datasets. The procedure of ANOVA includes estimating the effects of various experimental factors on the experimental outcomes and how these factors interact with each other. Afterward, the significance of the outcomes and their effects is determined, and the F-test is used in measurements that are distributed with equal variance for various experimental conditions [40].

3) ROC and AUC Value

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the true positive rate (TPR) and false-positive rate (FPR) at different threshold settings. In this study, the ROC is considered to ascertain different classifiers' performance under different threshold values. It also represents the relative tradeoffs between asthma and non-asthma classes. Moreover, the Area Under ROC curve (AUC), a metric to ascertain the accuracy, is calculated as the area bounded by the ROC curve [41]. The higher value of AUC indicates a better classification model.

4) Recall rate Versus decision boundary

For asthma prediction, the recall rate can be interpreted as the number of asthma patients identified over the dataset. Recall rate is the function of the decision boundary. In this study, the decision threshold of 0.5 has been used to give equal emphasis on Asthma-yes and Asthma-no classes.

5) The 10-fold cross-validation

Cross-validation (CV) is used to generalize a model and test the whole dataset [42]. There are several CV techniques, such as K-fold, five-times two folds etc. In this study, 10-fold CV has been used to partition the total dataset into ten equal subsets. Each time one fold is used to test while the other nine folds are used in the training phase to build a model, and this is repeated ten times to test the entire dataset. After that, we have used the ANOVA test on each fold of the 10-fold dataset. Note that we have used the same index for each classifier for a fair justification in the ANOVA test.

6) Feature importance and cumulative feature importance

There could be several features in the original feature set with zero importance (or near to zero due to rounding). We can get rid of these unimportant features without impacting the overall performance of the classifier. The negligible features can be erased by feature importance, and in this study, we have presented a feature importance graph on which the X-axis represents the features, and the Y-axis represents the importance of the features. On the other hand, the cumulative feature importance is a graph that shows the contribution of each feature to the overall importance or overall accuracy. In this way, we can select distinguishable features that are much lower than the primitive feature set, and it saves the overall cost of training time and affords related to obtaining features. To determine important features, the gain is used for the optimal node split during training in the tree-based classifiers such as RF, XGB, and GBC by Eq. (12) [43]:

$$gain = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda}+\frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda}-\frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i+\lambda}\right]-\gamma \tag{12}$$

where $g_i, h_i, I_L, I_R, \lambda$ and γ represent the first-order gradient, second-order gradient, left nodes and right nodes after segmentation, penalty parameter and regularization parameter, respectively. $I = I_L\ U\ I_R$. The gain symbolizes the information gain of each tree split. After that, the average gain is calculated from the ratio of gain of all the trees to the total number of splits for each feature. The final feature importance score is calculated, generally in descending order from the average gain.

F. ENSEMBLING (HARD VOTING AND SOFT VOTING)

The combination of different optimized classifiers is often used to improve the classification through the ensemble technique. In the ensemble technique, both hard voting and soft voting schemes are used. To briefly discuss, hard voting is the simplest version of majority voting, where the number of each class label is enumerated and assigned to a class that is voted by a majority of the classifiers [44]. If C_j represents each classifier and $\hat{y}$ be the class label, then the hard voting is accomplished using the following formula:

$$\hat{y} = mode\{c_1(x), c_2(x), c_3(x).....c_m(x)\} \tag{13}$$

For instance, if the class labels are counted as $0, 1, 1, 0, 1$, it is mostly voted because 1 occurs in most class labels. $\hat{y} = mode\{0, 1, 1, 0, 1\} = 1$. One drawback of the normal hard voting ensemble is that it can only be applied for binary classification. In this situation, the weighted majority vote can be alternatively chosen [45]. If w_j be a weight with classifier C_j, then the weighted majority vote can be computed using the following formula.

$$\hat{y} = \arg\max_i \sum_{j=1}^{m} w_j \chi_A\left(C_j\left(x\right) = i\right) \tag{14}$$

where χ_A represents the characteristic function, and A is the set of class labels.

On the other hand, in soft voting, the predicted probabilities, for example, scores, are summed for each class label and predict the class having the largest probability. If

p is the predicted probability of the classifier and w is the corresponding weight, then the class label will be defined using the following equation [45].

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{m} w_j p_{ij} \tag{15}$$

III. FINDINGS

In this paper, both the ADASYN-balanced dataset and primary test data have been used to evaluate the proposed BOMLA detector. The existing imbalance between the majority and minority classes has been erased by generating adequate simulated data. The effect of ADASYN has also been shown in Figure 4 under Section III-A. Again, in Section III-B under the actual test data, the balanced ADASYN model was examined. To determine the significance of statistical evaluation of cross-validation, ANOVA and Box-plot appeared in Section III-C, followed by the decision boundary vs recall rate curve. The bootstrapping and feature importance have been discussed after that. Finally, a comparative study regarding the performance of BOMLA with other search techniques has been presented. Table 1 shows the optimal value of hyperparameters calculated using the BOMLA detector and used in this study.

Table 1: The best performing classifiers with tuned hyperparameters.

Classifiers	Best Hyperparameters using BOMLA
KNN	*n_neighbors* = 5
NB	*alpha* = 1.7024
DT	*max_depth* = 11 *max_features* = 4 *criterion* = 'gini'
RF	*max_depth* = 20 *max_features* = 5 *n_estimators* = 29 *criterion* = 'gini'
XGB	*n_estimators* = 250 *learning_rate* = 0.05 *n_jobs* = 4 *max_depth* = 14 *gamma* = 0.1048 *colsample_by_tree* = 0.9810
GBC	*loss* = 'exponential' *learning_rate* = 0.7136 *max_features* = 1 *n_estimators* = 9 *max_depth* = 14
ANN	*neurons_in_hidden_layer* = 42 *activation* = "ReLU" (Rectified Linear Unit) *solver* = "L-BFGS" *alpha* = 0.1731 *learning_rate* = "constant"
SVC	*C* = 2.6033 *gamma* = 5.5778 *kernel* = 'RBF' (Radial Basis Function) *probability* = True

A. OPTIMIZATION RESULTS WITH ADASYN AND WITHOUT ADASYN

The entire dataset, balanced by ADASYN, has been split to get training, validation and test set. Two-third of the total data is used for training and validation, and one-third of the whole dataset is used for the test. Afterwards, our proposed BOMLA algorithm has been applied to ten state-of-the-art classifiers. The proposed BOMLA detector has been evaluated using several metrics, followed by delineating the effect of ADASYN in this subsection.

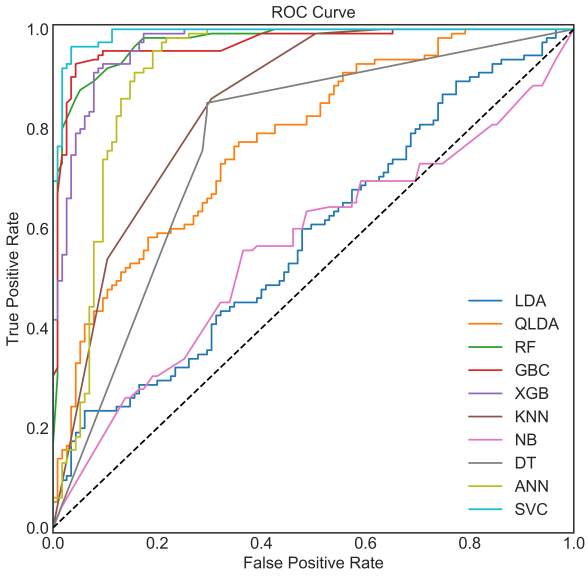


Figure 4: ROC curve for Asthma with ADASYN.

To initiate, Table 2 represents the effect of ADASYN, where the upper portion provides the result of the ADASYN-balanced asthma dataset by utilizing equations (3) to (11). Notably, SVC yields the utmost ACC, where GBC occupies the second-highest position in this case. Conversely, LDA and NB show the lowermost ACC among multiple classifiers Table 2. It is also noticeable that the SVC occupies the highest position in the case of AUC calculation, whereas RF, XGB, and GBC are very close to borderline [Figure 4].

The effect of ADASYN has also been clarified by evaluating the optimized model using original test data exhibited in the lower portion of Table 2. Note that without using ADASYN, the unbalanced dataset affects the classification performance. It is evidenced that XGB provides us with the highest ACC of 88.37%; however, its AUC is just below 50% [Figure 5]. The higher ACC and lower AUC reveal that the imbalanced model can show higher ACC but cannot accurately classify asthma-yes and asthma-no classes. Besides, due to the classification model's imbalanced nature, its $TN = 0$, meaning that the model cannot detect any true negative class, and hence its specificity or true negative rate becomes zero. It degrades the performance in the AUC and Kappa index, consequently loses its reliability. These results indicate that the ADASYN algorithm's usefulness to balance the model as a balanced model can provide better results.

Table 2: Asthma Classification (%).

With ADASYN										
Classification Algorithms	**ACC**	**AUC**	**MCC**	**Error**	**SE**	**F1_score**	**FPR**	**Kappa**	**PPV**	**SP**
LDA	54.35	58.40	8.79	45.65	46.96	50.70	38.26	8.70	55.10	61.74
QLDA	61.30	77.80	34.57	38.70	23.48	37.76	0.87	22.61	96.43	99.13
KNN	74.35	84.80	56.07	25.65	49.57	65.90	**0.87**	48.70	**98.28**	**99.13**
NB	57.39	56.20	14.89	42.61	51.30	54.63	36.52	14.78	58.42	63.48
DT	76.96	76.90	54.38	23.04	70.43	75.35	16.52	53.91	81.00	83.48
RF	90.43	97.30	81.48	9.57	84.35	89.81	3.48	80.87	96.04	96.52
XGB	90.00	96.80	80.69	10.00	83.48	89.30	3.48	80.00	96.00	96.52
GBC	92.61	97.00	85.38	7.39	89.57	92.38	4.35	85.22	95.37	95.65
ANN	88.26	91.30	77.37	11.74	80.87	87.32	4.35	76.52	94.90	95.65
SVC	**94.35**	**99.20**	**88.97**	**5.65**	**98.26**	**94.56**	9.57	**88.70**	91.13	90.43
Without ADASYN										
LDA	87.60	55.70	3.21	12.40	99.12	93.39	100.00	1.45	88.28	0.00
QLDA	46.51	54.30	0.68	53.49	45.61	60.12	46.67	0.40	88.14	53.33
KNN	87.60	46.50	3.21	12.40	99.12	93.39	100.00	1.45	88.28	0.00
NB	87.60	51.80	3.21	12.40	99.12	93.39	100.00	1.45	88.28	0.00
DT	82.95	52.40	6.24	17.05	92.11	90.52	86.67	6.15	88.98	13.33
RF	85.27	49.80	6.49	14.73	96.49	92.05	100.00	4.90	88.00	0.00
XGB	88.37	49.10	15.02	11.63	99.12	93.78	93.33	9.28	88.98	6.67
GBC	82.17	42.80	10.67	17.83	90.35	89.96	80.00	10.66	89.57	20.00
ANN	82.95	61.30	8.69	17.05	93.86	90.68	100.00	7.40	87.70	0.00
SVC	87.60	60.60	3.21	12.40	99.12	93.39	100.00	1.45	88.28	0.00

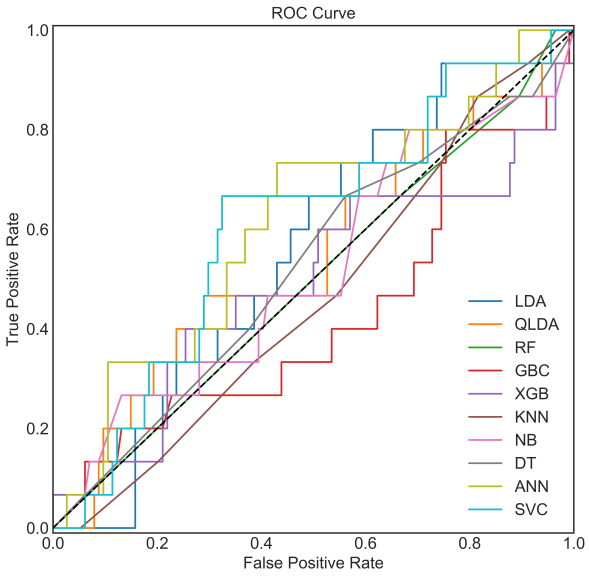


Figure 5: ROC curve for Asthma excluding ADASYN (i.e., using the original imbalanced dataset).

B. EFFECT OF BALANCED MODEL ON ORIGINAL TEST DATASET

In the previous subsection, we have used the ADASYN algorithm to balance the dataset, and we have seen how the balanced model helps improve classification performance. So, what is the balanced model's effect on enhancing the classification task (i) when no synthetic data is used and (ii) on the original test data only?

For the sake of providing the answer to the given question, we have again used Bayesian-optimized models to justify its performance on the ADASYN-balanced dataset. After that, the ACC, SE, SP, and ROC have been enumerated, as shown in Table 3 and Figure 6, which provide that SVC, again, grabbed the topmost position in terms of every classification metric. The worth-noting point should be that both the ACC and AUC obtained from RF, XGB, GBC, and KNN are slightly lower than SVC.

Interestingly, the results presented here represent the trajectory of the upper portion of Table 2. Therefore, to culminate, the ADASYN-balanced and Bayesian-optimized model can also help to classify the original test data only. Furthermore, the above argument can also be evidenced in Figure 7, where we have presented the high-performing confusion matrix using SVC on both Asthma-yes and Asthma-no classes.

C. 10-FOLD CROSS-VALIDATION

This paper has applied 10-fold cross-validation to test the whole dataset [Figure 8]. The ACC of each fold has been presented in Table 4, from where it is evidenced that LDA yields the most petite average ACC, whereas SVC touched the mountain point. We have also tested the statistical significance using ANOVA, which provided the p-value as 8.75034×10^{-33} for the Asthma dataset with ADASYN, and this value is statistically significant. Besides, we have added an interactive plot of the multi-comparison test [Figure 9].

The recall rate is calculated using a certain threshold. For instance, in Figure 10, representing the decision boundary vs recall rate curve, and 0.5 has been taken as the decision boundary threshold (T) for Asthma-yes class. The "Asthma-yes" with the ADASYN dataset found QLDA having the best results with a 0.99 Recall rate, which means that 99% times QLDA can truly classify the "Asthma-yes" class. In comparison, LDA showed the worst recall rate (around 0.59). On the other hand, for the "Asthma-no" class (Figure 10),

Table 3: Asthma classification (%) tested on original data.

Classification Algorithms	ACC	AUC	MCC	Error	SE	F1_score	FPR	Kappa	PPV	SP
LDA	53.49	48.80	6.82	46.51	56.14	68.09	66.67	4.65	86.49	33.33
QLDA	34.11	67.00	19.53	65.89	25.44	40.56	0.00	7.35	100.00	100.00
KNN	66.67	93.80	40.13	33.33	62.28	76.76	0.00	27.75	100.00	100.00
NB	50.39	52.70	5.85	49.61	49.12	63.64	40.00	3.64	90.32	60.00
DT	86.82	91.20	57.16	13.18	86.84	92.09	13.33	53.41	98.02	86.67
RF	93.02	98.70	75.87	6.98	92.11	95.89	0.00	73.07	100.00	100.00
XGB	93.02	98.40	73.57	6.98	92.98	95.93	6.67	71.77	99.07	93.33
GBC	97.67	99.60	90.08	2.33	97.37	98.67	0.00	89.59	100.00	100.00
ANN	94.57	97.70	80.00	5.43	93.86	96.83	0.00	78.05	100.00	100.00
SVC	**100.00**	**100.00**	**100.00**	**0.00**	**100.00**	**100.00**	**0.00**	**100.00**	**100.00**	**100.00**

Table 4: 10-fold cross-validation Accuracy.

	LDA	QLDA	KNN	NB	DT	RF	XGB	GBC	ANN	SVC
Fold-1	60.00	65.71	85.71	58.57	77.14	90.00	91.43	82.86	85.71	97.14
Fold-2	47.14	52.86	80.00	41.43	78.57	80.00	82.86	81.43	90.00	88.57
Fold-3	44.29	54.29	71.43	41.43	84.29	88.57	85.71	85.71	85.71	87.14
Fold-4	55.71	51.43	72.86	58.57	80.00	82.86	81.43	80.00	84.29	81.43
Fold-5	44.29	58.57	82.86	54.29	87.14	88.57	94.29	91.43	91.43	84.29
Fold-6	54.29	60.00	80.00	67.14	80.00	97.14	94.29	90.00	95.71	95.71
Fold-7	65.22	57.97	81.16	78.26	79.71	91.30	86.96	79.71	81.16	86.96
Fold-8	47.83	59.42	71.01	46.38	75.36	86.96	84.06	86.96	78.26	91.30
Fold-9	49.28	63.77	76.81	63.77	91.30	98.55	95.65	95.65	88.41	95.65
Fold-10	62.32	60.87	78.26	57.97	85.51	91.30	92.75	88.41	88.41	92.75
Average	**53.04**	**58.49**	**78.01**	**56.78**	**81.90**	**89.53**	**88.94**	**86.22**	**86.91**	**90.10**

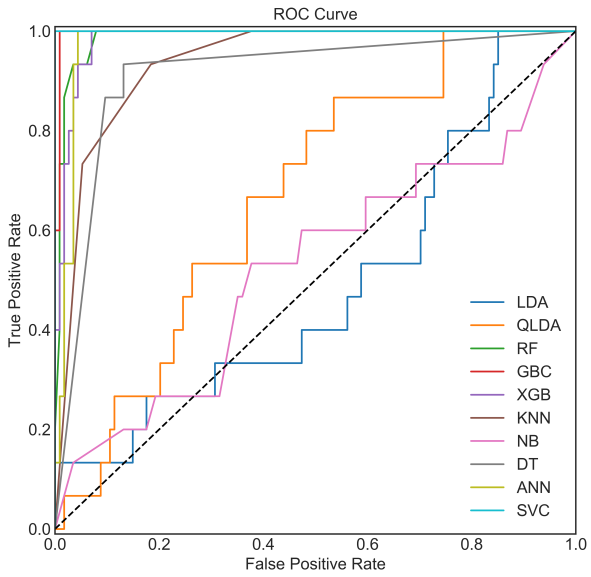


Figure 6: ROC Analysis when the BOMLA detector is tested on original test data. Note that a balanced model using ADASYN has only been used during model building.

SVC displayed the highest recall rate (almost 0.97), which means that 97% of the time we classify a non-asthma patient that truly non-asthma, whilst QLDA displayed the worst recall rate (approximately 0.23).

D. BOOTSTRAPPING OF THE ROC

In order to test the model's biasness and test that the training phase is extremely biased to the training dataset or not. We have applied the bootstrapping $N\_boot = 100$ on the SVC model and found that the mean AUC of 98% reveals that the training phase is not biased to the training dataset [Figure 11]. We have also calculated the 90% confidence interval (CI). The most crucial point is the CI of 90% is very close to the upper limit and lower limit of the mean AUC, which is another indication that the training phase is rarely biased in the training dataset.

E. EFFECT OF ENSEMBLING

The best performing classifiers, such as SVC, RF, XGB, and GBC, have been combined through weighted hard voting and soft voting approach to enhance the classification performance. For weighted hard voting, with the weight of $[5, 1, 1, 1]$ has been used for SVC, RF, XGB, GBC; and for soft voting, with the weight of $[5, 1, 1, 2]$ has been used for SVC, RF, XGB, GBC, respectively. In both voting schemes, it can be seen that we have added a higher weight of 5 to SVC to emphasize the importance of SVC, as it individually provides the best performance. Table 5 expresses a comparison among the performance obtained from SVC, hard voting, and soft voting techniques while applied on ADASYN-balanced dataset, unbalanced dataset, and the original test data. It is evidenced that the ACC of weighted soft voting far outweighs both weighted hard voting and the proposed SVC technique. Note that the classification performance of SVC is added from Table 2 and Table 3 to show how the ensemble technique enhances the classification performance compared to the best individual performance using SVC.

Table 5: Effect of ensembling.

With ADASYN (%)									
Classification Techniques	ACC	MCC	Error	SE	F1_score	FPR	Kappa	PPV	SP
SVC [from Table 2]	94.35	88.97	5.65	98.26	94.56	9.57	88.70	91.13	90.43
Weighted Hard Voting	94.78	89.58	5.22	93.91	94.74	4.35	89.57	95.58	95.65
Weighted Soft Voting	96.52	93.04	3.48	96.52	96.52	3.48	93.04	96.52	96.52
Without ADASYN (%)									
SVC [from Table 2]	87.60	3.21	12.40	99.12	93.39	100.00	1.45	88.28	0.00
Weighted Hard Voting	89.15	24.37	10.85	100.00	94.21	93.33	11.21	89.06	6.67
Weighted Soft Voting	89.92	35.36	10.08	99.12	94.56	80.00	28.06	90.40	20.00
With the original test data (%)									
SVC [from Table 3]	100	100	0	100	100	0	100	100	100
Weighted Hard Voting	100	100	0	100	100	0	100	100	100
Weighted Soft Voting	100	100	0	100	100	0	100	100	100

Figure 12 visualizes that the AUC of the SVC is 99.2%, while the weighted soft voting provided us with a better AUC of 99.6%. Therefore, the ensemble technique significantly magnifies the classification performance compared to the traditional way of using classification algorithms.

F. FEATURES IMPORTANCE AND CUMULATIVE FEATURE IMPORTANCE

During the visualization of the feature importance, the most influential variables are categorized in decreasing manner (a feature from higher importance to lower importance), which means that the feature having the highest important value will be presented first. Other features are presented one by one, following their importance. The variables having higher importance convey higher predictive power. To provide an instance, Figure 13(a) showed the feature importance, where it has been received that 'FEV1_by_FVC' got the highest importance, 'FVC_percent' was second-highest, and 'age', 'PULSE (P)' were also noteworthy features. Besides, Figure 13(b) showed the cumulative feature importance for the Asthma dataset using the RF algorithm, where it provided nearly 60% accuracy after adding the five most important features, and 90% importance retains, where all the features were added cumulatively. As the same result found from Figure 13(c) and Figure 13(d).

G. COMPARISON AMONG DIFFERENT SEARCH TECHNIQUES

The BO technique has been implemented in our proposed framework. Therefore, to illustrate the proposed method's superiority, it is obvious to add a comparative delineation of the BO algorithm with other hyperparameter search algorithms, such as grid and random search. In the grid search algorithm, the hyperparameters are evaluated on the search grid defined by the users and assessed on each grid. Whereas in the random search optimization algorithm, the hyperparameters are randomly selected on the search boundaries. In Table 6, we have compared our BO algorithm with a grid search and random search technique. We have evaluated our algorithm with a core i9 computer, which has 64GB RAM. It has been evidenced that the SVC model performed better, previously shown; therefore, this model has been used for the comparison task. To compare our proposed BO algorithm, we have used four performance indices: parameter evaluated, the time required to complete the task, CV score, and test score. It can be seen from Table 6 that BO takes nearly 58 seconds, which is much lower than grid search, but six times greater than the random search. Despite the above time consumption of the BO technique, its CV score and test score are higher than that of both grid and random search. Therefore, comparing all these aspects, it can be said that BO provides us with better results in terms of time consumption, CV score, and test score.

The graphical representation [Figure 14] helps us compare different search techniques. It is evidenced that the initial ACC of the BO and random search technique were 60% and 75%, respectively. Following a couple of iterations, the ACC of both techniques showed a steep increase. Following the fourth iteration, the ACC of random search (82%) showed an unflattering condition, which was almost unchanged until 46 iterations. By way of comparison, the proposed BO technique steeply increases, grabbing the approximate ACC of 83%, which was almost static before completing 36 iterations. Afterwards, the ACC changed its pattern a little bit, and the most eye-catching point is that the ACC hits the mountain point before accomplishing 50 iterations.

IV. DISCUSSION

In this study, a new machine learning framework has been proposed, applying different state-of-the-art classifiers, called BOMLA, in order to detect asthma patients from their clinical and demographic data. The proposed algorithm has been evaluated with numerous classification metrics, namely ACC, SE, SP, Kappa index, MCC, etc. Besides, 10-fold CV, ANOVA, and recall vs decision boundary have been applied to analyze the proposed model. In the end, the most dominating features have been traced through the discussion of feature importance and cumulative feature importance.

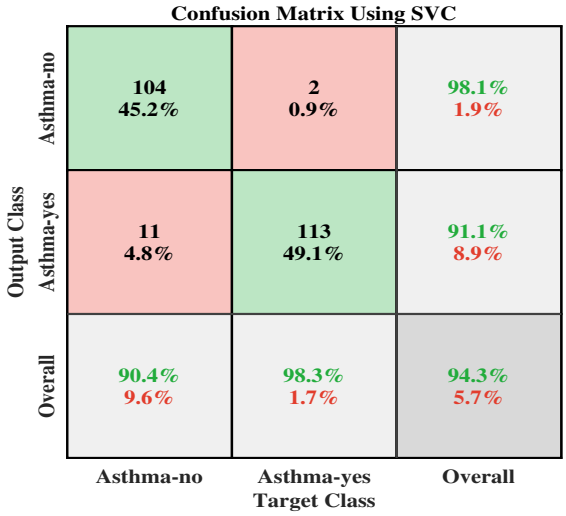
It can be seen that the tree-based classifiers, such as XGB, RF, and GBC, provide about 90% ACC, while SVC delivers the highest performance in terms of ACC, SE, AUC, etc. [see Table 2]. Moreover, the ADASYN-balanced SVC model also provided the foremost ACC in the ANOVA test [Table 4] when applied to the original test data. It is conspicuous that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3073086, IEEE Access

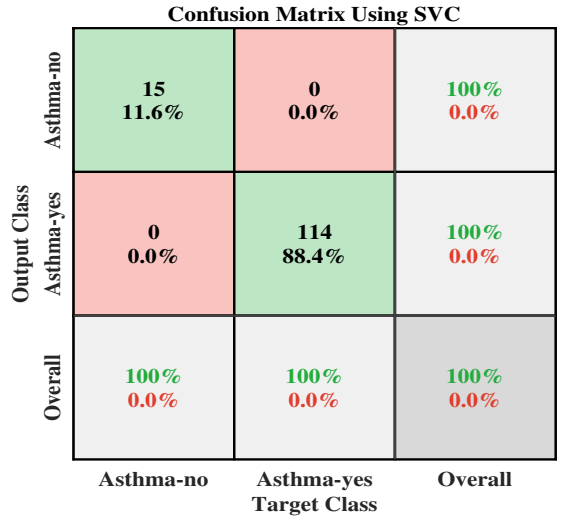
Table 6: Comparison with other hyperparameter Optimization techniques.

Optimization Algorithms	Number of evaluated parameters	Total time to complete (sec)*	CV Accuracy	Test Accuracy
Grid Search	6561	571.49	0.80	0.86
Random Search	**50**	**9.01**	0.82	0.83
The BO	**50**	57.94	**0.86**	**0.91**

* In our proposed BO and Random search algorithm, the time required to calculate 50 iterations is presented. In Grid search, the time required to evaluate all the parameters has been taken into account.



(a)



(b)

Figure 7: Confusion matrix evaluated in Asthma dataset with ADASYN (Figure 7a), without ADASYN, i.e., test data only (Figure 7b). The first two diagonal cells of the confusion matrix represent the total number of correctly classified Non-Asthma and Asthma patients. In contrast, the third diagonal cell shows the overall accuracy and misclassification rate (in %). The third row and third column represent the overall results in row-wise and column-wise, respectively.

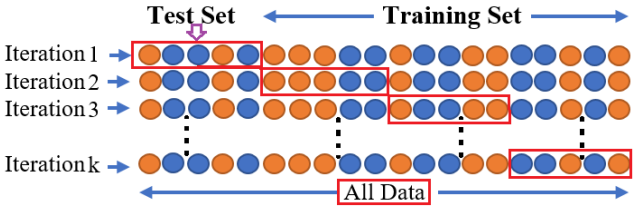


Figure 8: K(k=10)-Fold Cross-Validation.

the ADASYN-balanced and optimized SVC could be the best choice in the BOMLA detector framework to detect asthma.

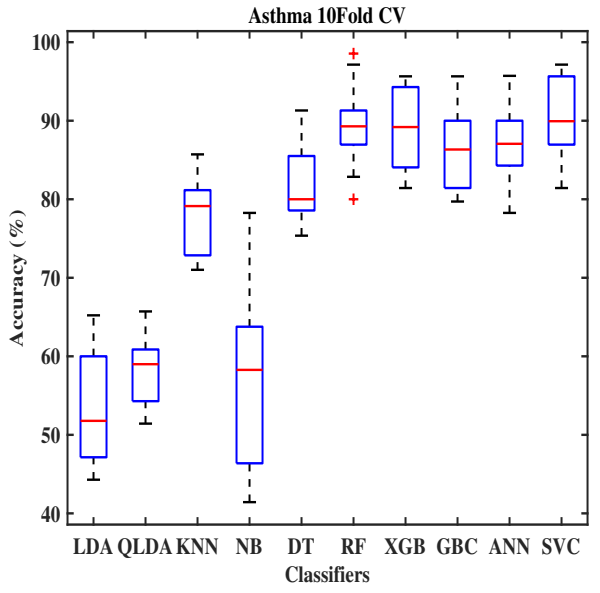
As the dataset is not linearly separable (see the violin plot shown in Figure 1), the kernel-trick used in the RBF-SVC helps separate linearly in a high-dimensional space. Furthermore, by selecting a proper value of cost (C) using BO, an optimal margin between asthma-yes and asthma-no class is achieved. The optimal value of C controls the influence of the misclassification rate and hence, improves the classification accuracy. On the other hand, the RBF kernel's optimal γ parameter using BO helps control the distance of influence of a single training example and shapes the decision boundary. These two factors help improve classification performance.

Regarding the ADASYN algorithm, it can be characterized that ADASYN can adaptively generate a sufficient amount of synthetic data to balance the model and improve the classification performance [30] (lower part of Table 2). Concerning optimization, the BO has been applied in the BOMLA framework, which outperforms the commonly-used machine learning hyperparameter optimization technique, namely random search and grid search. The recent study shows no statistically significant difference using BO, and other meta-heuristic optimization algorithms, such as Harris Hawk Optimization [34]. Furthermore, the meta-heuristic algorithm needs a longer time to evaluate the objective function and finish the program. Considering this, the BO is adapted to the BOMLA detector.

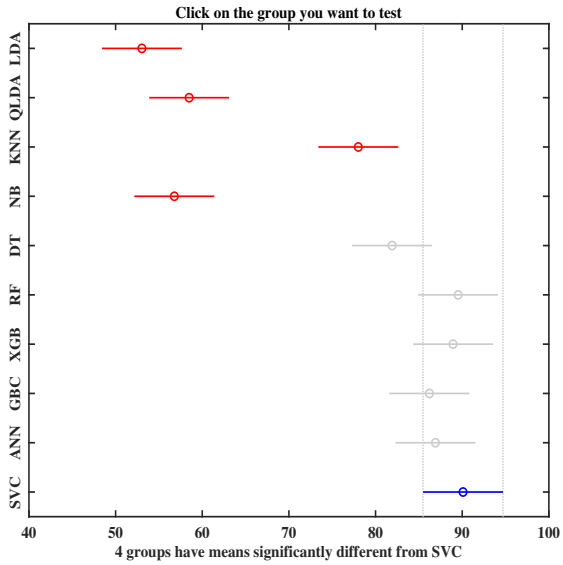
From the above discussion, we can write some of the salient features of the BOMLA detector:

- the proposed BOMLA detector can also be applied to detect the COVID-19 patients, diabetes prediction, hypertension patient classification etc.
- Hyperparameters can easily be tuned by using the BO algorithm.
- The developed DSS can be helpful for the end-users and clinical staff.
- Although ten state-of-the-art classifiers have been used in the BOMLA detector, some other recent classifiers,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3073086, IEEE Access



(a)



(b)

Figure 9: Box-plot for asthma dataset with ADASYN (Figure 9a) and multi-comparison test (Figure 9b). Notably, the right figure is a GUI tool, which helps us ascertain any classifier's statistical significance. In this diagram, the importance of SVC has been analyzed only. Similarly, the significance of other classifiers can be tested.

such as Light Gradient Boosting Machine (LightGBM), CatBoost can also be used for the BOMLA framework.

While describing the salient features, we can also write some of the weaknesses of this manuscript. The dataset used in this paper is small, which needs to be validated on a larger dataset. This is relevant to the data collection.

A. COMPARATIVE STUDY WITH BENCHMARK

Our proposed algorithm has also been compared with other studies. Here we briefly described some state-of-the-art algorithms applied for asthma detection [Table 7]. To exemplify, Chatzimichail et al. [46], Finkelstein and Jeong [47], and Amaral et al. [48] used SVC, along with other well-known classifiers, where ACC obtained from [46] and [47] were 95.54% and 80%, respectively. Chatzimichail et al. [46] used the PCA (Principal Component Analysis) method and least square SVC; however, they did not apply the original features in their work. They have projected the features into Principal Components and then used them in SVC. In that case, they have lost the original features' performance, which is so much crucial in this context. In contrast, we have used the original features and showed the feature importance and how the important features cumulatively improve the performance to explain the machine learning algorithm. Besides, the ensemble result (e.g., ACC, SE) presented in our study outperform the result of [46]. Furthermore, Xu et al. [49] and Krautenbacher et al. [50] used RF, obtaining AUC of 66% and 81% respectively. On top of that, Luo et al. [51] and Patel et al. [52] developed their methods with a different Boosting algorithm, obtained with the AUC values of 76.1% and 85% respectively, while Brasier and Ju [55] and Kuo et al. [58] reported 94.00% and 86.60%, respectively. Therefore, it can be decided that in terms of AUC, ACC, SE, and SP, our proposed method outperforms the conventional techniques. It can also be noted that, SVC classifier wins in most of the studies [47], [48], [56], [57], [59] including our proposed BOMLA. So, it is noticeable that optimized SVC can detect asthma patients more accurately than other classifiers.

B. DEVELOPMENT OF A DECISION SUPPORT SYSTEM

A DSS could be beneficial to support clinical staff for screening asthma patients from clinical data. The DSS is a graphical representation of the decision to visualize the probable state of an asthma patient. A possible outcome of asthma patient is presented in Figure 15, in terms of the posterior probability calculated from SVC. A probabilistic result is more intuitive to the clinical staff and, therefore, used in this DSS. Note that 50 patients are used from the test database for illustration purposes. The patient is sorted in ascending order so that patients with "Asthma-no" labelled appear first, and then patients with "Asthma-yes" appear.

In Figure 15, 0 represents a subject with Asthma-no, whereas 1 represents a subject with Asthma-yes. The lower figure portrays a probabilistic outcome of the subject affected by Asthma, where the red line defines the threshold level. When the probability exceeds this threshold level (0.5), the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3073086, IEEE Access

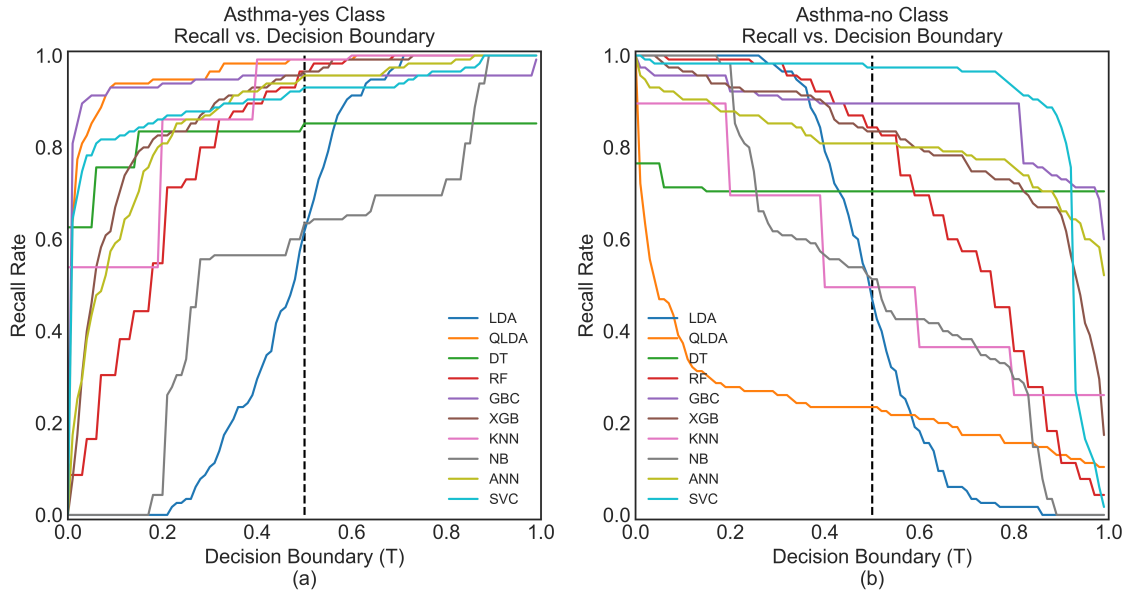


Figure 10: Decision boundary vs. recall rate curve for asthma class (Figure 10(a)) and non-asthma class (Figure 10(b)).

Table 7: Comparison with other studies.

Other studies	**Classification algorithms**	**Dataset used**	**Sample size**	**Performance indices**			
				ACC	**SE**	**SP**	**AUC**
Chatzimichail *et al.* [46]	PCA and Least square SVC	Clinical data	148	95.54%	95.45%	95.59%	
Finkelstein and Jeong [47]	SVC	Telemonitoring data	7001	80.00%	84.00%	80.00%	
Amaral *et al.* [48]	LBNC, SVC, KNN	Forced oscillation data	150		>87.00%	>94.00%	>95.00%
Xu *et al.* [49]	RF	Clinical data	417				66.00%
Krautenbacher *et al.* [50]	LASSO and stochastic gradient boosting	Clinical and genomics data	260				81.00%
Luo *et al.* [51]	Multiboost with decision stumps	Child asthma data	310	71.80%	73.80%	71.40%	76.10%
Patel *et al.* [52]	DT, LASSO logistic regression, RF, and gradient boosting machines	Clinical data	29392		99.10%	14.60%	85.00%
Li *et al.* [53]	Classification Tree	Clinical data	310		94.00%	68.00%	
Swern *et al.* [54]	Post hoc analyses	2-5 years patients data	689		66.80%	85.80%	
Brasier and Ju [55]	Multivariate Adaptive Regression Splines (MARS)	Multidimensional data	84	90.00%	88.00%	73.00%	94.00%
Lanclus *et al.* [56]	SVC	Functional CT imaging	62	80.65%	82.35%		
Wu *et al.* [57]	SVC	Clinical data	346	81.00%	62.00%	87.00%	
Kuo *et al.* [58]	DT	Clinical data	107	82.40%			86.60%
Wu *et al.* [59]	SVC	Clinical data	378	93.00%			
Messinger *et al.* [60]	ANN	Clinical data	128	80.00%			
Proposed	**SVC**	**Clinical data**	389	94.35%	**98.26%**	90.43%	99.20%
	Soft voting ensembling			**96.52%**	96.52%	**96.52%**	**99.60%**

subject will be considered asthma-yes, whereas the probability lower than 0.5 will be regarded as Asthma-no. In either way, we can point out that the probability of 0.5 is the chance that a person is affected by Asthma.

V. CONCLUSION

This study designs and optimizes a novel machine learning framework named BOMLA detector to detect asthma patients. The entire research has been accomplished based on the asthma dataset collected from Khulna, Bangladesh. The BOMLA detector's performance is examined and assessed in different machine learning perspectives, such as using other performance indices, e.g., ACC, SE, kappa index, MCC, etc., along with ROC analysis. The balanced model's effect using the ADASYN algorithm has also been presented and shown outperforming than without a balanced model. The results show that the BOMLA detector can detect asthma efficiently, where the highest classification accuracy by using ADASYN provided a value of 94.35%, which has been increased to 96.52% through ensemble technique. Thus, the proposed BOMLA detector offers a low-cost and user-friendly tool for the early detection and classification of asthma. A potential application, i.e., a DSS developed from this study, could benefit the clinical staff and end-users. It can help to build a recommender system that can easily be integrated into the mobile application. One of the limitations of this study is the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3073086, IEEE Access

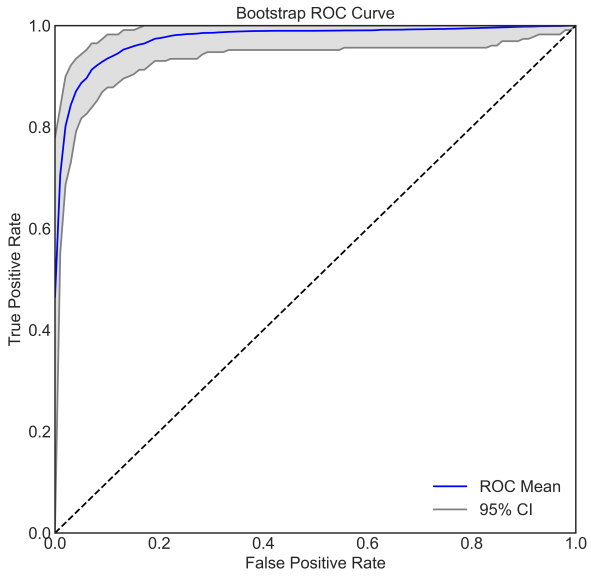


Figure 11: Bootstrapping of ROC using SVC with 95% confidence interval.

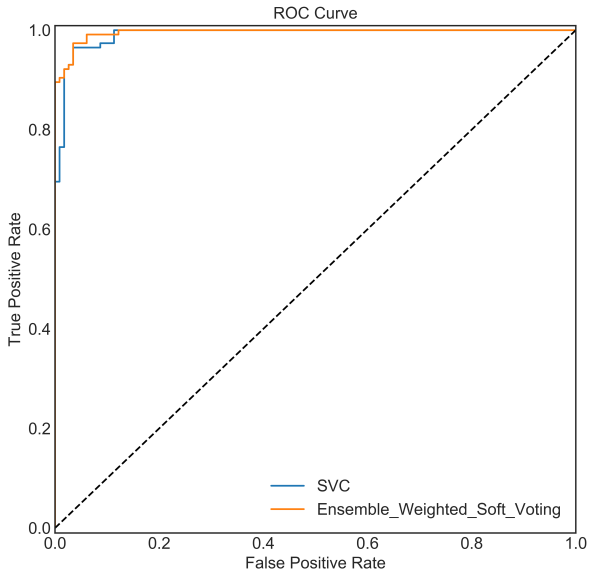


Figure 12: Effect of ensembling on classification performance.

Algorithm-1: The steps of applying ADASYN to balance the dataset

Input: The original n-dimensional unbalanced Asthma Data
Output: Balanced dataset using ADASYN algorithm

1. If n_r, n_j denote the total number of minority class (M_r) and majority class (M_j), respectively, then calculate the degree of imbalance using $d = n_r/n_j$, where $d \in (0, 1]$
2. If $d < d_j$ (d_j is the preset threshold for maximum tolerated imbalance)
   - **a.** Calculate the number of synthetic data to be generated for M_r using: $G = (n_j - n_r)\,\beta$, where $\beta \in [0, 1]$
   - **b.** Determine K for KNN for each $x_i \in M_r$, and calculate the ratio $r_i = \zeta_i/K, i = 1, ..., n_r$. Here ζ_i denotes the number of KNN example belongs to M_j
   - **c.** Calculate the synthetic data for each minority x_i, using $g_i = \hat{r_i} \times G$, where $\hat{r}$ denotes the density distribution $\hat{r} = r_i / \sum_i r_i$ and $r_i \in [0, 1]$
   - **d.** Use **loop** from 1 to g_i, and generate the synthetic data using $s_i = (x_u - x_i) \times \lambda$, where x_u is the randomly chosen minority data for K neighbors, and $\lambda \in [0, 1]$

Algorithm-2: Hyperparameter optimization using BOMLA detector.

Input: Classifiers with initial hyperparameters (P), training dataset.
Output: Classifiers with optimal hyperparameters.

1. Set the cost function $J(Clf(P))$ that need to minimize.
   - **a.** Calculate the $K = 10$ fold $F1loss_{CV}$ from: $F1loss_{CV} = \frac{1}{K}\sum_{k=1}^{K}(1 - F1Score_k(Clf(P)))$
   - **b.** Assign $F1loss_{CV}$ as cost function i.e., $J(Clf(P)) \leftarrow F1loss_{CV}$.
2. Define search space for each hyperparameter $p_1, p_2, p_3, ..., p_n \in P$ by defining the lower bound and upper bound value of P including categorical parameter(s) of the classifiers.
4. Using Tree-structured Parzen Estimators (TPE), search for optimal P for which $J(.)$ provides the best (lowest) value.
   - **a.** On each iteration, for every hyperparameter $p_1, p_2, p_3, .., p_n \in P$, TPE divides the observation into two sets.
   - **b.** Calculate the first set of p i.e., $l(p)$ for which $J(Clf(p))$ provides the best (lowest) value.
   - **c.** Calculate the second set of p i.e., $g(p)$ for which $J(Clf(p))$ provides the worst (highest) value.
   - **d.** Calculate the expected improvement $EI_p = l(p)/g(p)$.
   - **e.** Select p that maximizes the EI_p.
   - **f.** Repeat Step 4.a to 4.e for each hyper-parameter, p and store in the *Trials*.
   - **g.** Go to Step 2 for the next *Trials*.
5. Find the optimal hyperparameters from the best *Trials* and then use these hyperparameters to build an optimal classifier.

limited number of subjects, and testing on a larger dataset would be beneficial. However, our primary objective is to show the possibility of detecting asthma with high accuracy using the BOMLA detector. Finally, our findings suggest that early detection of asthma using the BOMLA detector could be employed for real-time asthma detection.

APPENDIX A

see Algorithm 1.

APPENDIX B

see Algorithm 2.

ACKNOWLEDGMENT

The research team acknowledges Mehedi Hasan Masud, Fahad Ahmed Nafiz, all patients and medical assistants for their supports. They also thank the ethical approval committee.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3073086, IEEE Access

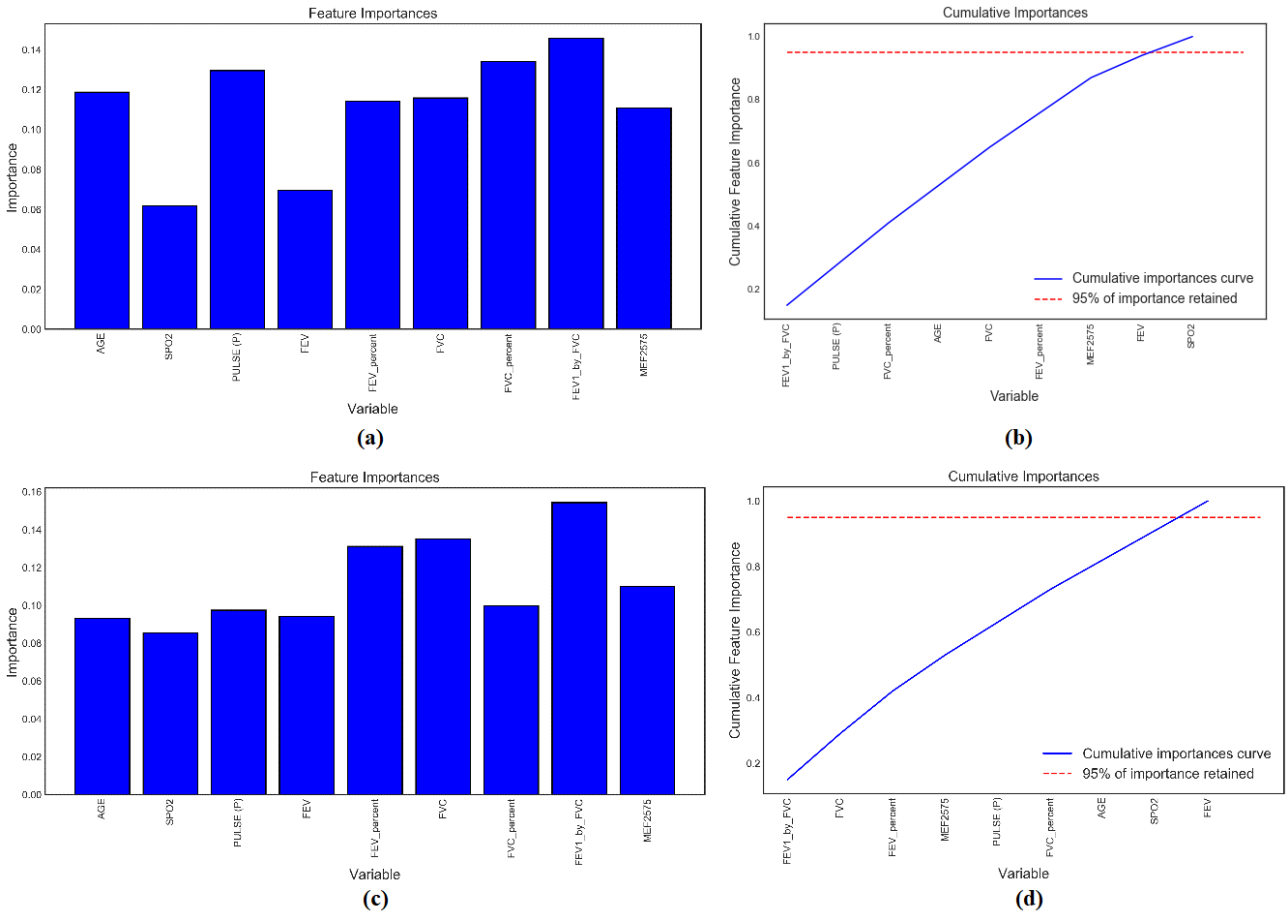


Figure 13: (a) Feature importance and (b) Cumulative feature importance using RF, (c) Feature importance and (d) Cumulative feature importance using XGB.

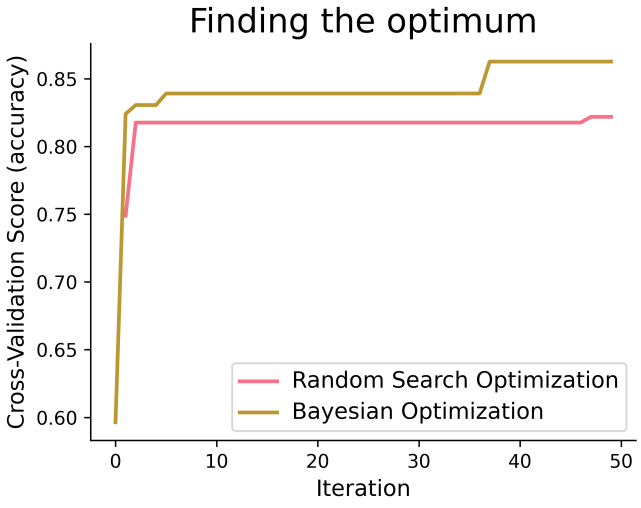


Figure 14: Cross-validation accuracy vs iteration curve for random search and BO.

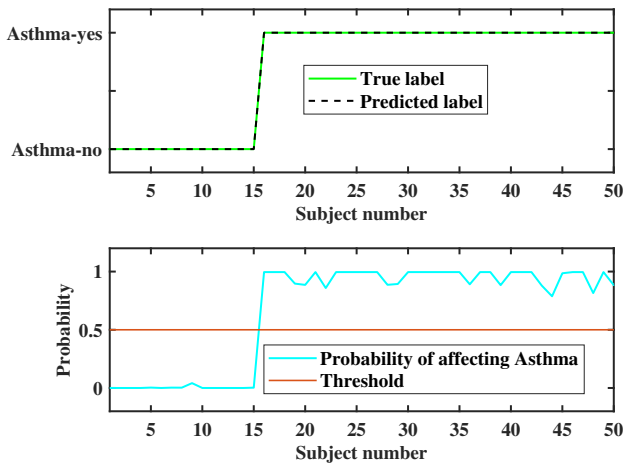


Figure 15: Probabilistic output for the DSS using SVC.

References

[1] "GINA report 2018, global strategy for asthma management and prevention," *Global Initiative for Asthma*, 2018.

[2] D. Spathis and P. Vlamos, "Diagnosing asthma and chronic obstructive

pulmonary disease with machine learning," *Health Informatics Journal*, vol. 25, no. 3, pp. 811-827, 2019.

[3] E. Bolat, "A comprehensive comparison of machine learning algorithms on diagnosing asthma disease and COPD," *International Journal*, vol. 76, no. 3/1, 2020.

[4] K. Tsang, H. Pinnock, A. Wilson, and S. A. Shar, "Application of machine learning to support self-management of asthma with mHealth," in *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2020.

[5] H.S. Zahran, C.J. Person, C. Bailey, and J.E. Moorman, "Predictors of asthma self-management education among children and adults—2006–2007 behavioral risk factor surveillance system asthma callback survey," *Journal of Asthma*, vol. 49, no. 1, pp. 98-106, 2012.

[6] "Global Initiative for Asthma (GINA): Global strategy for asthma management and prevention," 2018. [Online]. Available: http://www.ginasthma.org.

[7] M. Masoli, D. Fabian, S. Holt, R. Beasley, and Global Initiative for Asthma (GINA) Program, "The global burden of asthma: executive summary of the GINA dissemination committee report," *Global Initiative for Asthma Program: Allergy*, vol. 59, no. 5, pp. 469-478, 2004.

[8] K.R. Chapman, L.P. Boulet, R.M. Rea, and E. Franssen, "Suboptimal asthma control: prevalence, detection and consequences in general practice," *European Respiratory Journal*, vol. 31, no. 2, pp. 320-325, 2008.

[9] W. H. Organization, "Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach," in *Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach*, pp. vii. 146-vii. 146, 2007.

[10] J. Zhan, W. Chen, L. Cheng, Q. Wang, F. Han, and Y. Cui, "Diagnosis of Asthma Based on Routine Blood Biomarkers Using Machine Learning," *Computational Intelligence and Neuroscience*, Hindawi, vol. 2020, pp. 8841002, 2020.

[11] M.C. Prosperi, S. Marinho, A. Simpson, A. Custovic, and I.E. Buchan, "Predicting phenotypes of asthma and eczema with machine learning," *BMC Medical Genomics*, vol. 7, no. S1, pp. S7, 2014.

[12] F. Bacopoulou, A. Veltsista, I. Vassi, A. Gika, V. Lekea, K. Priftis, and C. Bakoula, "Can we be optimistic about asthma in childhood? A Greek cohort study," *Journal of Asthma*, vol. 46, no. 2, pp. 171-174, 2009.

[13] W.A. Balemans, C.K. van der Ent, A.G. Schilder, E.A. Sanders, G.A. Zielhuis, and M.M. Rovers, "Prediction of asthma in young adults using childhood characteristics: Development of a prediction rule," *Journal of Clinical Epidemiology*, vol. 59, no. 11, pp. 1207-1212, 2006.

[14] K. Porpodis, D. Papakosta, K. Manika, T. Kontakiotis, M. Gaga, L. Sichletidis, and D. Gioulekas, "Long-term prognosis of asthma is good—a 12-year follow-up study. Influence of treatment," *Journal of Asthma*, vol. 46, no. 6, pp. 625-631, 2009.

[15] B.G. Toelle, W. Xuan, J.K. Peat, and G.B. Marks, "Childhood factors that predict asthma in young adulthood," *European Respiratory Journal*, vol. 23, no. 1, pp. 66-70, 2004.

[16] J.P. Caloyeras, H. Liu, E. Exum, M. Broderick, and S. Mattke, "Managing manifest diseases, but not health risks, saved PepsiCo money over seven years," *Health Affairs*, vol. 33, no. 1, pp. 124-131, 2014.

[17] G. Luo, C.L. Nau, W.W. Crawford, M. Schatz, R.S. Zeiger, E. Rozema, and C. Koebnick, "Developing a predictive model for asthma-related hospital encounters in patients with asthma in a large, integrated health care system: secondary analysis," *JMIR Medical Informatics*, vol. 8, no. 11, p. e22689, 2020.

[18] G. Luo, M.D. Johnson, F. LNkoy, S. He, and B. L Stone, "Automatically explaining machine learning prediction results on asthma hospital visits in patients with asthma: secondary analysis," *JMIR Medical Informatics*, vol. 8, no. 12, p. e21965, 2020.

[19] J.W. Dexheimer, L.E. Brown, J. Leegon, and A. Dominik, "Comparing decision support methodologies for identifying asthma exacerbations," in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, 2007: IOS Press, pp. 880.

[20] B.D.C.N. Prasad, P.K. Prasad, and Y. Sagar, "A comparative study of machine learning algorithms as expert systems in medical diagnosis (Asthma)," in *International Conference on Computer Science and Information Technology*, Springer, pp. 570-576, 2011.

[21] A. Badnjević, L. Gurbeta, M. Cifrek, and D. Marjanovic, "Classification of asthma using artificial neural network," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, *IEEE*, pp. 387-390, 2016.

[22] E. Granulo, L. Bećar, L. Gurbeta, and A. Badnjević, "Telemetry system for diagnosis of asthma and chronical obstructive pulmonary disease (COPD)," in *International Conference on IoT Technologies for HealthCare*, Springer, pp. 113-118, 2016.

[23] L. Gurbeta, A. Badnjevic, M. Maksimovic, E. Omanovic-Miklicanin, and E. Sejdic, "A telehealth system for automated diagnosis of asthma and chronical obstructive pulmonary disease," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1213-1217, 2018.

[24] D. Spathis and P. Vlamos, "Diagnosing asthma and chronic obstructive pulmonary disease with machine learning," *Health Informatics Journal*, vol. 25, p. 1460458217723169, 2017, doi: 10.1177/1460458217723169.

[25] A. Yahyaoui and N. Yumuşak, "Decision support system based on the support vector machines and the adaptive support vector machines algorithm for solving chest disease diagnosis problems," 2018.

[26] J.L. Hintze and R.D. Nelson, "Violin plots: a box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, pp. 181-184, 1998.

[27] J.W. Mims, "Asthma: definitions and pathophysiology," in *International Forum of Allergy & Rhinology*, vol. 5, no. S1: Wiley Online Library, pp. S2-S6, 2015.

[28] H.K. Reddel, "GINA recommendations in adults with symptomatic mild asthma and a smoking history," *European Respiratory Journal*, vol. 55, no. 2, 2020.

[29] G.I. F.A.E. Committee, "Global strategy for asthma management and prevention (revised 2002)," in *NHLBI/WHO Workshop Report*; 2006, GINA, 2006.

[30] H. He, Y. Bai, E.A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, IEEE, pp. 1322-1328, 2008.

[31] H. He and Y. Ma, Imbalanced Learning: Foundations Algorithms and Applications, Hoboken, NJ, USA:Wiley, 2013.

[32] M.A. Rahman, S.M. Shoaib, M. Al Amin, R.N. Toma, M.A. Moni, and M.A. Awal, "A Bayesian optimization framework for the prediction of diabetes mellitus," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE), IEEE*, pp. 357-362, 2019.

[33] M. Pelikan, D.E. Goldberg, and E. Cantú-Paz, "BOA: The Bayesian optimization algorithm," in *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-99*, vol. 1: Citeseer, pp. 525-532, 1999.

[34] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Future generation computer systems*, vol. 97, pp. 849-872, 2019.

[35] S. Li, H. Chen, M. Wang, A.A. Heidari, and S. Mirjalili, "Slime mould algorithm: A new method for stochastic optimization," *Future Generation Computer Systems*, vol. 111, pp. 300-323, 2020.

[36] J. Bergstra, D. Yamins, and D.D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proceedings of the 12th Python in Science Conference*, vol. 13: Citeseer, pp. 20, 2013.

[37] A. Hay, "The derivation of global estimates from a confusion matrix," *International Journal of Remote Sensing*, vol. 9, no. 8, pp. 1395-1398, 1988.

[38] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M.J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics Decision Making*, vol. 10, no. 1, pp. 16, 2010.

[39] M. Frigge, D.C. Hoaglin, and B. Iglewicz, "Some implementations of the boxplot," *The American Statistician*, vol. 43, no. 1, pp. 50-54, 1989.

[40] D.J. Benjamin, J.O. Berger, M. Johannesson, B.A. Nosek, E.J. Wagenmakers, R. Berk, K.A. Bollen, B. Brembs, L. Brown, C. Camerer, and D. Cesarini, "Redefine statistical significance," *Nature Human Behaviour*, vol. 2, no. 1, pp. 6, 2018.

[41] M.C. Hemmsen, T. Lange, A.H. Brandt, M.B. Nielsen, and J.A. Jensen, "A methodology for anatomic ultrasound image diagnostic quality assessment," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 64, no. 1, pp. 206-217, 2017.

[42] T.G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895-1923, 1998.

[43] C. Chen, Q. Zhang, B. Yu, Z. Yu, P.J. Lawrence, Q. Ma, and Y. Zhang, "Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier," *Computers in Biology and Medicine*, vol. 123, 2020.

[44] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516-76531, 2020.

[45] S. Raschka and V. Mirjalili, Python Machine Learning, Birmingham, UK, Packt Publishing; 2nd Ed., 2017.

[46] E. Chatzimichail, E. Paraskakis, M. Sitzimi, and A. Rigas, "An intelligent system approach for asthma prediction in symptomatic preschool children," *Computational Mathematical Methods in Medicine*, vol. 2013, 2013.

[47] J. Finkelstein and I. cheol Jeong, "Machine learning approaches to personalize early prediction of asthma exacerbations," *Annals of the New York Academy of Sciences*, vol. 1387, no. 1, pp. 153, 2017.

[48] J.L.M. Amaral, A.J. Lopes, J.M. Jansen, A.C.D. Faria, and P.L. Melo, "Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease," *Computer Methods Programs in Biomedicine*, vol. 105, no. 3, pp. 183-193, 2012.

[49] M. Xu, K.G. Tantisira, A. Wu, A.A. Litonjua, J.H. Chu, B.E. Himes, A. Damask, and S.T. Weiss, "Genome wide association study to predict severe asthma exacerbations in children using random forests classifiers," *BMC Medical Genetics*, vol. 12, no. 1, pp. 1-8, 2011.

[50] N. Krautenbacher, N. Flach, A. Böck, K. Laubhahn, M. Laimighofer, F.J. Theis, D.P. Ankerst, C. Fuchs, and B. Schaub, "A strategy for high-dimensional multivariable analysis classifies childhood asthma phenotypes from genetic, immunological, and environmental factors," *Allergy*, vol. 74, no. 7, pp. 1364-1373, 2019.

[51] G. Luo, B.L. Stone, B. Fassl, C.G. Maloney, P.H. Gesteland, S.R. Yerram, and F.L. Nkoy, "Predicting asthma control deterioration in children," *BMC Medical Informatics Decision Making*, vol. 15, no. 1, pp. 1-8, 2015.

[52] S.J. Patel, D.B. Chamberlain, and J.M. Chamberlain, "A machine learning approach to predicting need for hospitalization for pediatric asthma exacerbation at the time of emergency department triage," *Academic Emergency Medicine*, vol. 25, no. 12, pp. 1463-1470, 2018.

[53] D. Li, D. German, S. Lulla, R.G. Thomas, and S.R. Wilson, "Prospective study of hospitalization for asthma: a preliminary risk factor model," *American Journal of Respiratory Critical Care Medicine*, vol. 151, no. 3_pt_1, pp. 647-655, 1995.

[54] A.S. Swern, C.A. Tozzi, B. Knorr, and H. Bisgaard, "Predicting an asthma exacerbation in children 2 to 5 years of age," *Annals of Allergy, Asthma Immunology*, vol. 101, no. 6, pp. 626-630, 2008.

[55] A. R. Brasier and H. Ju, "Analysis and predictive modeling of asthma phenotypes," in *Heterogeneity in Asthma: Springer*, 2014, pp. 273-288.

[56] M. Lanclus, J. Clukers, C. Van Holsbeke, W. Vos, G. Leemans, B. Holbrechts, K. Barboza, W. De Backer, and J. De Backer, "Machine learning algorithms utilizing functional respiratory imaging may predict COPD exacerbations," *Academic Radiology*, vol. 26, no. 9, pp. 1191-1199, 2019.

[57] W. Wu, S. Bang, E.R. Bleecker, M. Castro, L. Denlinger, S.C. Erzurum, J.V. Fahy, A.M. Fitzpatrick, B.M. Gaston, A.T. Hastie, and E. Israel, "Multiview cluster analysis identifies variable corticosteroid response phenotypes in severe asthma," *American journal of respiratory and critical care medicine*, vo. 199, no. 11, pp. 1358-1367, 2019.

[58] C.H.S. Kuo, S. Pavlidis, M. Loza, F. Baribaud, A. Rowe, I. Pandis, U. Hoda, C. Rossios, A. Sousa, S.J. Wilson, and P. Howarth, "A transcriptome-driven analysis of epithelial brushings and bronchial biopsies to define asthma phenotypes in U-BIOPRED," *American journal of respiratory and critical care medicine*, vol. 195, no. 4, pp. 443-455, 2017.

[59] W. Wu, E. Bleecker, W. Moore, W.W. Busse, M. Castro, K.F. Chung, W.J. Calhoun, S. Erzurum, B. Gaston, E. Israel, and D. Curran-Everett, "Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data," *Journal of Allergy and Clinical Immunology*, vol. 133, no. 5, pp. 1280-1288, 2014.

[60] A.I. Messinger, N. Bui, B.D. Wagner, S.J. Szefler, T. Vu, and R.R. Deterding, "Novel pediatric-automated respiratory score using physiologic data and machine learning in asthma," *Pediatric pulmonology*, vol. 54, no. 8, pp. 1149-1155, 2019.

MD. ABDUL AWAL has completed his B.Sc. in Electronics and Communication Engineering (ECE) from ECE Discipline, Khulna University in 2009. Later on, he has finished his M.Sc. in Biomedical Engineering from Khulna University of Engineering and Technology in 2011. He completed his PhD in Biomedical Engineering from The University of Queensland, Australia, in 2018. His research interests are Signal Processing, especially Biomedical Signal Processing, Big Data Analysis, Image Processing, Time-Frequency Analysis, Machine Learning Algorithms, Deep Learning, Optimization, and Computational Intelligence Biomedical Engineering. He has more than 35 papers published in internationally accredited journals and conferences. He is currently working as an Associate Professor at ECE Discipline, Khulna University, Khulna, Bangladesh. He is now investigating some projects as principal investigator and co-investigator and supervising several undergraduate and post-graduate students.

MD. SHAHADAT HOSSAIN has completed his B.Sc. (Honours) in Mathematics from Mathematics Discipline, Khulna University in 2014 and M.Sc. in Applied Mathematics in 2016 from the same institution. His research interest comprises Machine Learning, Deep Learning, Data Science, Optimization, Image Processing, Signal Processing, Applied Mathematics etc. He has six research articles published in international journals. He is currently working as a lecturer of Mathematics at the Department of Quantitative Sciences, International University of Business Agriculture and Technology (IUBAT), Dhaka-1230, Bangladesh. Currently, he is working on several research works with collaboration in the field of Machine Learning, Optimization, and Data Science.

KUMAR DEBJIT received M.Sc. in Computing Technology in 2019 from University of Southern Queensland, Australia. As a precursor, he completed his B.Tech degree in Information Technology from KIIT University, Bhubaneswar, India. He is now working as a programmer. His research interest includes big data analysis, IoT, artificial intelligence and robotics, Software engineering and programming, Role of human-computer interaction, computer-assisted education.

NAFIZ AHMED received M. Com in Business Information System in 2019 from Australian Catholic University, Australia. He also completed his BBA degree in Marketing from North South University, Bangladesh. He is now working as an ICT Business Analyst. His research interest includes public health informatics, telehealth system, big data analysis, artificial intelligence and robotics, the internet of things, consumer behavior, and business research.

RAJAN DEV NATH received Masters of Professional Accounting in 2020 from University of Southern Queensland, Australia. He also completed his BBA degree in Finance and Accounting from North South University, Bangladesh. He is now working as an accountant. His research interest includes public health informatics, big data in health care, health economics, accounting information systems, corporate finance and capital market efficiency.

DR GM MONSUR HABIB received MBBS from the University of Dhaka in 1982, Comm. Paediatrics from the University of Edinburgh in 1992, Diploma in Asthma from the Open University, UK, COPD Diploma module from the Open University, UK in 2014. Furthermore, he was a primary care physician by the Government of Bangladesh in a rural setting (1983-1984), primary care physician by the Government of Iran in the rural and Hospital setting (1985-1992). Besides, he is President, Bangladesh Primary Care Respiratory Society since 2016, Founding and Senate member of International Primary Care Respiratory Group (IPCRG) since 2002, Life member, Royal Medical Society, Edinburgh, UK since 1992, Member, Primary Care Respiratory Society, UK, Life Member, Bangladesh Medical Association since 1996, Life member, Bangladesh Asthma Association since 1998, Founding and board member, Bangladesh Lung Foundation (BLF) since 2007, and EC member, South Asian Association of Respiratory Physician since 2009. Currently, he is pursuing his PhD at the University of Edinburgh, based at Bangladesh Primary Care Respiratory Society (BPCRS), aiming to develop and pilot a PR programme scaled up in Bangladesh. He has 11 articles on medical science published in internationally accredited journals. Most importantly, he is an expert in Lung Diseases, Spirometry, Airway Obstruction, and Asthma Management.

MD. SALAUDDIN KHAN received B.Sc. and M.S. degree in Statistics from Jahangirnagar University, Savar, Dhaka, Bangladesh in 2011 and 2013, respectively securing the 3rd position of his department. He is now serving as an Assistant Professor in the Statistics Discipline, Khulna University, Khulna, Bangladesh. His research interest includes Data Science, especially multivariate analysis, health science, data mining, big data analysis, machine learning algorithms, time series analysis with applications, computational intelligence in bioinformatics and biostatistics.

MD. AKHTARUL ISLAM received B.Sc and M.S. degree in Statistics Biostatistics & Informatics from Dhaka University, Dhaka, Bangladesh in 2012 and 2013 respectively. His research interest includes Biostatistics, Epidemiology, Public Health, Infectious disease, Meta-analysis, Statistical computing, and Multivariate Analysis. He has authored and co-authored around 12 publications in different peer-review journals. He is now serving as an Assistant Professor in the Statistics Discipline, Khulna University, Khulna 9820, Bangladesh.

DR. M. A. PARVEZ MAHMUD received his B.Sc. degree in Electrical and Electronic Engineering and Master of Engineering degree in Mechatronics Engineering. After the successful completion of his PhD degree with multiple awards, he worked as a Postdoctoral Research Associate and Academic in the School of Engineering at Macquarie University, Sydney. He is currently an Alfred Deakin Postdoctoral Research Fellow at Deakin University. He worked at World University of Bangladesh (WUB) as a 'Lecturer' for more than 2 years and at the Korea Institute of Machinery and Materials (KIMM) as a 'Researcher' for about 3 years. His research is focused on Energy Sustainability, Secure Energy Trading, Microgrid Control and Economic Optimization, Machine Learning, Data Science, and Micro/nanoscaled Technologies for Sensing and Energy Harvesting. He accumulated experience and expertise in machine learning, life cycle assessment, sustainability and economic analysis, materials engineering, microfabrication, and nanostructured energy materials to facilitate technological translation from the lab to real-world applications for a better society. He has produced over 50 publications, including 1 authored book, 3 Book Chapters, 29 Journal Papers, and 21 fully refereed Conference Papers. He received several awards, including "Macquarie University Highly Commended Excellence in Higher Degree Research Award 2019". He was involved in teaching engineering subjects in the electrical, biomedical and mechatronics engineering courses at the School of Engineering, Macquarie University, for more than 2 years. Currently, he is involved in the supervision of 7 PhD students at Deakin University. He is a key member of Deakin University's Advanced Integrated Microsystems (AIM) research group. Apart from this, he is actively involved with different professional organizations, including Engineers Australia and IEEE.