Predictive Modeling for Asthma Disease Detection: A Comparative Study of Machine Learning Algorithms

Arka Mukherjee, Junali Jasmine Jena, Mahendra Kumar Gourisaria

Predictive Modeling for Asthma Disease Detection: A Comparative Study of Machine Learning Algorithms

Arka Mukherjee
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, Odisha, India
2328078@kiit.ac.in

Junali Jasmine Jena
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, Odisha, India
junali.jenafcs@kiit.ac.in

Mahendra Kumar Gourisaria
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, Odisha, India
mkgourisaria2010@gmail.com

***Abstract*— Artificial Intelligence (AI) and Machine learning (ML) models have proven to be scalable approaches for handling several biomedical problems. Recent availability of high-quality datasets which captures various factors contributing to the respiratory disease, have enabled the development of robust models that deliver high accuracy and precision scores in early detection of respiratory diseases. This paper focuses on asthma disease detection. It makes two primary contributions: (1) an empirical evaluation of their performance on an asthma disease detection dataset with data mining and pre-processing techniques, and (2) the identification of the most effective approach for asthma disease detection based on rigorous evaluation using metrics such as precision, accuracy, recall, F-1, and F-beta scores. Cat Boost Classifier was found to be the best model which predicted asthma disease with a 96.04% accuracy. Our code can be found here:** https://github.com/ArkaMukherjee0/AsthmaDetection*

Keywords—Asthma prediction, Machine Learning, Data mining

I. INTRODUCTION

Asthma, a pervasive chronic inflammatory disease of the airways, affects over three hundred million individuals globally. Epidemiological studies indicate a significant portion of the population, approximately 10% of pediatric and 6-7% of adult patients, are affected by the condition [1]. It leads to impaired lung function with severe exacerbations causing hospitalizations, thereby impacting the quality of life for the affected. Effective and accurate prediction of the disease is crucial to or in reducing healthcare costs, significantly decreasing the frequency and severity of asthma attacks, preventing long-term lung damage, and identifying potential risk factors and triggers. Interventions of AI and ML models are remarkably efficient at analyzing diverse patient data — spanning medical history, environmental and allergic influences, symptomatic presentations, and clinical measurements — to predict the likelihood of asthma onset [2].

A. The Respiratory System and Asthma

The human respiratory process occurs in two phases: inhalation, when air is drawn into the body, and exhalation, whereby carbon dioxide, a metabolic byproduct, is expelled. Its primary function is to facilitate gas exchange, a bidirectional flow of oxygen into the body and carbon dioxide out. Asthma's hallmark is airway hyperresponsiveness — characterized by airway inflammation, bronchoconstriction, and mucus hypersecretion. []. Thus, hypersensitivity renders the airways overly sensitive to various triggers, leading to sudden and often severe episodes of airflow limitation. Asthma attacks, marked by symptoms like wheezing, coughing, and shortness of breath, can be induced by multiple factors. These include allergens such as dust mites and pollen, respiratory infections, weather changes, emotional stress, and irritants like tobacco smoke and air pollutants [3].

B. Risk Factors for Asthma

The development of asthma can be attributed to a series of lifestyle and environmental factors. Exposure to allergens and irritants can significantly increase the risk of developing the condition. Additionally, lifestyle choices such as low physical activity levels, obesity, and poor diet choices like consumption of processed foods increase asthma susceptibility. Studies have shown that local environment and lifestyle factors influence the development of asthma. An analysis of data collected from Vancouver, Canada, and select spots in China shows significant differences in the prevalence of the disease. Healthcare professionals typically rely on medical history, clinical measurements, existing symptoms, and allergy factors to determine the condition [4].

C. Role of Technology

Data-driven approaches using machine ML techniques can efficiently process vast corpora of data to make nearly-accurate predictions. Innovation in the field of supervised learning augmented with data pre-processing techniques has pushed the state of the art significantly on a variety of tasks. Rapid advancements have been made in applying this technology in healthcare, such as in pneumonia detection [5], arrhythmia detection [6], tuberculosis detection [7], malaria detection [8] etc. Asthma detection, similarly, can be efficiently approached with ML. In this study, we probe a variety of existing technologies and evaluate their performance. We try to answer which models perform the best and which are better avoided. While there has been extensive research on asthma detection using various data modalities, the application of ML to numerical datasets remains relatively unexplored. To the best of our knowledge, this is the first study to comprehensively evaluate the performance of a diverse set of machine learning models to the newly released "Asthma Disease Dataset" comprising solely of numerical features.

The rest of the paper is organized as follows. In Section, 2 we conduct a thorough literature review of existing techniques for solving asthma disease detection. Section 3 describes the machine learning models used in this study. In Section 4, the empirical results obtained in our experiments are presented. Subsequently, Section 5 distils the findings into a conclusion and highlights future directions to improve asthma disease detection with AI and ML techniques followed by References.

II. RELATED WORK

Yahyaoui and Yumuşak (2021) [9] used a private dataset collected from patients in Diyarbakir hospital, Turkey for asthma and pneumonia disease detection. The study utilized K-Nearest Neighbors and Deep Neural Network models. They reported 95% and 94.3% accuracies respectively. Similarly,

*Preprint. Accepted to AESPC 2024.

Spathis and Vlamos (2019) [10] conducted another study on asthma and chronic obstructive pulmonary disease (COPD) detection with data collected by a pulmonologist in Thessalonik, Greece. Multiple ML models, encompassing Naïve Bayes, Logistic Regression, Neural Network, K-Nearest Neighbors, Decision Trees, and Random Forest was employed. The best performing technique was Naïve Bayes classifier, which delivered a precision score of 82%.

Zhang et al (2020) [11] predicted asthma exacerbations with data from the AstraZeneca SAKURA records. ML models such as Logistic Regression, Decision Trees, Naïve Bayes classifier, and Perceptron were applied. The best technique was experimentally verified to be Logistic Regression with Principal Component Analysis (PCA). It achieved sensitivity, specificity, and AUC scores of 90%, 83%, and 85% respectively. Also, Zhan et al (2020) [12] used Mahalanobis–Taguchi system (MTS) and SVM to identify asthma with routine blood biomarkers. The data was collected from patients admitted to Wuxi People's Hospital, China. The study recorded sensitivities of 94.15% and 93.55% with MTS and SVM respectively. Amaral et al. (2020) [13] ran tests on a private dataset created with ninety-seven volunteers with asthma and restrictive respiratory disease conditions. ML models such as K-Nearest Neighbors, Random Forests, AdaBoost with Decision Trees (ADAB), Support Vector Machines with radial basis function kernel (SVMR), and Neural Fuzzy Classifier (NFC). SVMR reported the best sensitivity of 99.1%, while ADAB reported the best specificity of 96.5%.

Awal et al (2021) [14] developed a system for early detection of asthma leveraging multiple ML models. Data from a clinical study conducted in Khulna, Bangladesh was utilized. The best model was Support Vector Classifier (SVC) with an accuracy of 94.35%. Similarly, Xie and Xu (2024) [15] utilized machine learning models to predict asthma disease in youth with National Health Interview Survey (NIHS) data. The best reported model is linear SVM with an accuracy of 89.06%. In terms of precision, Random Forest is the best with a score of 47.06%. Also, Barua et al (2022) [16] contributed a 1D-ARCSLBP model that can synthesize coughing sounds to predict asthma in patients. A novel dataset was proposed for the study with over a thousand subjects. The model delivered accuracy and precision scores of 98.24% and 98.49% respectively. This method, however, doesn't take statistical data into account. Table 1 below provides a summary of the literature review.

TABLE I. REVIEW OF LITRATURE

Paper	Method/Model	Dataset	Results
[9] Yahyaoui and Yumuşak (2021)	K-Nearest Neighbors, Deep Neural Networks	dataset created with patient data from Diyarbakir hospital, Turkey	Accuracy scores: KNN – 95%, DNN – 94.3%
[10] Spathis and Vlamos (2019)	Naïve Bayes, Logistic Regression, Neural Network, SVM, K-Nearest Neighbors,	dataset recorded by pulmonologist in a suburb of Thessaloniki, Greece	Precision scores: Naïve Bayes – 82%,
[11] Zhang et al (2020)	Logistic Regression, Decision Tree, Naïve Bayes, Perceptron	AstraZeneca SAKURA dataset	Sensitivity – 90%, Specificity – 83%, AUC – 85% by LR with PCA
[12] Zhan et al (2020)	Mahalanobis–Taguchi system (MTS), SVM	dataset recorded at Wuxi People's Hospital, China	MTS: Sensitivity – 94.15%, Specificity – 97.20%; SVM: Sensitivity – 93.55%, Specificity – 96.80%
[13] Amaral et al. (2020)	Best FOT Parameter (BFP), AdaBoost with Decision Trees (ADAB), Support Vector Machines (SVMR), Neural Fuzzy Classifier (NFC)	Private dataset created with 97 volunteers	SVMR: Sensitivity – 99.1%, Specificity – 80.7%;
[14] Awal et al (2021)	K-Nearest Neighbors (KNN), XGBoost (XGB), (ANN), Support Vector Classifier (SVC)	Data collected in a clinical study conducted in Khulna, Bangladesh	Accuracy: SVC – 94.35%
[15] Xie and Xu (2024)	XGBoost (XGB), Neural Networks (NN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR)	National Health Interview Survey (NIHS) data	Accuracy: SVM – 89.09%
[16] Barua et al (2022)	One-dimensional Attractive-and-Repulsive Center-Symmetric Local Binary Pattern (1D-ARCSLBP)	Novel cough sound dataset with >1,000 subjects	Accuracy – 98.24%, Sensitivity – 98.19%, Specificity – 98.30%, Precision – 98.49%, Geometric mean – 98.24%, F1-score – 98.34%

III. MATERIALS AND METHODS

This section is a comprehensive exposition of the various technologies and methods used for the study.

A. Dataset Description

For empirical results, we used a statistical asthma dataset from Kaggle [17] titled "Asthma Disease Dataset" with diagnosis information for the disease. The dataset encompasses a variety of numerical parameters resulting in the condition, including age, gender, ethnicity, education level, lifestyle factors like Body Mass Index (BMI), smoking habits, levels of physical activity, diet quality, and sleep quality; environments and allergy factors like exposure to pollution, pollen, dust, pet allergies; medical histories like family history of the condition, eczema, hay fever, history of other allergies, and gastroesophageal reflux; clinical measurements like lung function FEV-1 and FVC; and existing symptoms like wheezing, shortness of breath, chest tightness, coughing, night-time symptoms, and exercise-induced exacerbations. Fig.1. shows distribution of data as per asthmatic and non-asthmatic patients.

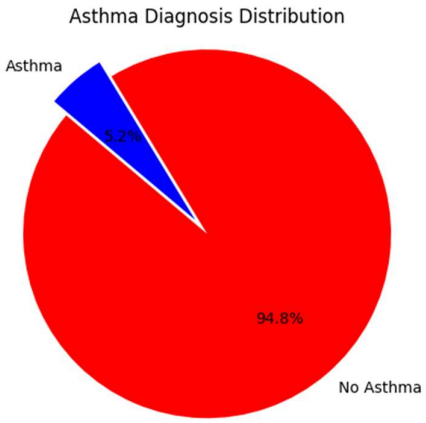

Fig. 1. Distribution of asthma disease diagnosis in the dataset

B. Data exploration and pre-processing techniques

The dataset used for this study includes patient ID and confidential information of the doctor in charge. These features were dropped. The ground truth values were derived as follows: 124 cases of asthma and 2,268 healthy individuals. Subsequently, we probed the dataset for any missing or duplicate values — none were found.

Before applying any data pre-processing techniques, we plotted distribution graphs for these features — age, BMI, physical activity, diet quality, sleep quality, pollution exposure, pollen exposure, dust exposure, lung function FEV1, and lung function FVC. A correlation graph was plotted to check for excessive dependence between any features. However, the dataset is extremely well-behaved and doesn't require any further processing. Additionally, data augmentation techniques like Principal Component Analysis (PCA) and Synthetic Minority Oversampling Technique (SMOTE) were utilized to further enhance the effectiveness of certain machine learning models. These methods enhanced performance in certain cases.

To solve the imbalance in asthma diagnosed and healthy patients (which stands at a ratio of 1:18.23), SMOTE was used widely across all models to boost performance. We also used PCA with each technique to empirically verify which model delivers the best results.

C. Technology Used

Since asthma disease detection is a binary classification problem, standard methods known to work well in the problem were utilized. These models include Logistic Regression, Random Forest, Support Vector Classifier (SVC), Decision Tree, Cat Boost, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Method (LGBM), Naive Bayes classifier, Neural Networks, and Multi-Layer Perceptron.

Logistic Regression (LR) is a statistical model used for predicting binary outcomes. It relies on the logistic function to calculate log-odds which is then translated into probabilities. Despite being simple, logistic regression is an effective technique that can handle complex input information such as patient data. The output probabilities were tested with multiple parameters such as accuracy, precision, recall, F1-score, ROC-AUC, and log loss scores to analyze the model's capabilities in the task.

Neural Networks (NN) are another widely used machine learning models for advanced predictive tasks. This technique is particularly potent in classification workloads such as asthma network. For this study, we have used a deep network with two hidden dense layers. The total trainable parameters were 14,337.

Decision Trees (DT) are another widely used tree-like hierarchical classifier. These models take into account consequences of events in the form of conditional control statements and make decisions based on the event outcomes of each. However, this makes them sensitive to the underlying data distributions.

Naïve Bayes Classifier (NBC) is a simple yet effective linear classification tool based on Bayesian statistics. It leverages the Bayes' theorem of conditional probabilities, making them extremely scalable. However, performance suffers in problems with multiple dimensions.

Random Forest (RF) is an ensemble learning technique involving aggregation of multiple decision trees that combines the results of all to give a single output. For classification tasks, such networks take the class chosen by the maximum number of trees. This way, they effectively solve overfitting issues prevalent with Decision Tree algorithms.

K-Nearest Neighbors (KNN) is another common supervised learning technique that relies on proximity of closely related data points for both classification and regression workloads. It computes the Euclidean distance between all data points to map out relationships in the training data. Learned patterns are then used to compute predictions in a fully non-parametric way. However, the simple underlying mechanism translates to poor performance in high-dimensional data.

Support Vector Classifier (SVC) is a specific implementation of a larger family of Support Vector Machine algorithms that focuses on classification tasks. Given a corpus of multi-dimensional data, the technique aims to compute the hyper-plane that results in the best segregation of the input features. However, these techniques can be computationally intensive given the quadratic calculations involved with the process.

Multi-Layer Perceptron (MLP) is a key machine learning technique driving innovation in the fields of Natural Language Processing and Computer Vision. It is a modernized feedforward neural network with mathematically modeled neurons that adapt to classification tasks through iterations of gradient manipulation during training. Much like SVMs, it can be computationally intensive at scale — with the main drawback being the multitude of matrix operations in both forward and backward steps. This study utilizes a simple five-layer MLP network with three dense layers and two dropout layers. The optimizer used is Adam and loss was computed with the binary cross entropy function. Training was carried out for 500 epochs.

Gradient Boosting (GB) [18] is an ensemble learning technique that involves aggregating multiple weak models — typically DTs — into one strong learner that results in high precision and accuracy scores. The convergence is carried out using a loss function that penalizes the trees based on their accuracy on the train set.

Ada Boost (AB) [18] is another ensemble learning algorithm that utilizes the strong and weak learner paradigm for high performance. The key difference between AB and XGBoost is in the added focus on misclassified instances, the

exponential loss function, and a stage-wise additive optimization function.

Extreme Gradient Boosting (XGBoost) [19] is another popular sequential ensemble learning technique that focuses on improving the original gradient boosting algorithm. The underlying principle remains largely intact — XGBoost employs several weak models (typically DTs) and iteratively reduces the error to achieve a single strong learner. However, this technique features several key improvements such as L1 and L2 regularization, sparsity-aware split finding, caching, and others. These improvements allow XGBoost to beat traditional GB in a wide range of tasks.

Cat Boost (CB) [20], short for Categorical Boosting, is one of the most powerful machine learning algorithms for classification and regression tasks. It relies on the same weak and strong learner framework as AB and XGBoost where new trees are trained to reduce the errors of previous iterations. However, the key strength is its ability to handle categorical features without explicitly using techniques such as one-hot encoding. For this study, CB was run for 1,000 iterations with learning rate set to 0.017861.

Light Gradient Boosting Machine (LGBM) [21] is a powerful histogram-based variant of GB. It leverages an ensemble of DTs with residual-correcting mechanism, Exclusive Feature Binding (EFB) to tackle high-dimensional data and Gradient-based One-size Sampling (GOSS) to speed up training. For this study, a learning rate of 0.1 was used for LGBM.

D. Methodology

Following the application of the machine learning algorithms, we used ensemble learning techniques such as *bagging*. Also known as Bootstrap Aggregating, bagging is a popular technique used to reduce high variance or overfitting and improve the stability of the model. It works by creating multiple instances of the originally trained base model with smaller subsets of the training data. Some of these samples are then drawn randomly with replacement and their performance is evaluated. In the backdrop of classification tasks such as this study, the method outputs an average prediction based on the principle of majority voting.

The dataset underwent rigorous pre-processing and exploratory analysis prior to an 80-20 train-test split. A suite of machine learning classifiers was systematically applied to address the binary classification problem of asthma disease detection. Each model was imported from the scikit-learn library, instantiated, and then fitted on the corresponding training data. Subsequently, we carried out output sampling, evaluation score calculation, confusion matrix and ROC-AUC curve computation. To further enhance predictive capabilities, the integration of dimensionality reduction techniques, specifically PCA, and ensemble techniques, notably bagging were explored. These methods were empirically probed and the best results are listed in Section IV. Fig.2. shows the workflow of the methodology used.

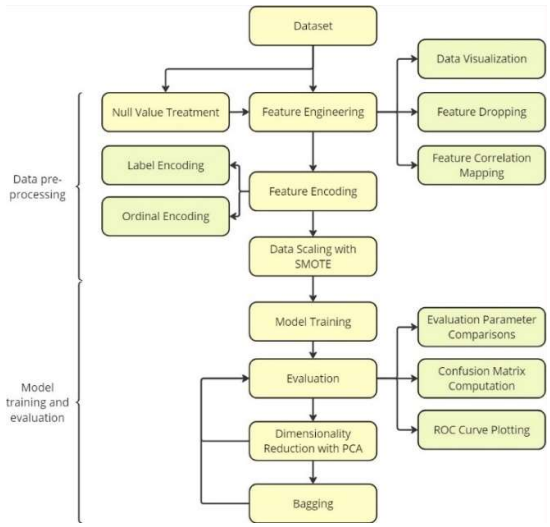

Fig. 2. Workflow diagram for asthma detection with ML algorithms

IV. RESULTS AND DISCISSION

This section outlines the results recorded with various machine learning algorithms on the dataset. All ML models were imported from the scikit-learn library and evaluation was carried out locally in Visual Studio Code with Python 3.12.1 in an Anaconda environment. The workstation is powered by an Intel® Core™ i5-12450H CPU (4P+4E cores), 16 GB of DDR5-4800 memory, Nvidia® RTX™ 4060 8 GB GPU, and 1 TB Gigabyte PCIe Gen 4 NVMe storage.

For the remainder of the paper, we have used some standard abbreviations. They are summarized in Table II.

TABLE II. STANDARD ABBREVIATIONS

Standard name	Abbreviation used
Cat Boost	CB
Random Forest	RF
Light Gradient Boosting Machine	LGBM
Extreme Gradient Boosting (XGBoost)	XGB
Support Vector Classifiers	SVC
Neural Networks	NN
Gradient Boosting	GB
Multi-Layer Perceptron	MLP
Decision Trees	DT
Ada Boost	AB
Logistic Regression	LR
K-Nearest Neighbors	KNN
Naïve Bayes Classifier	NBC
Accuracy Score	Acc
Precision Score	Prec
Recall Score	Rc
F1-Score	F1
F-beta Score (beta=0.5)	Fb
Jaccard Score	Jc
ROC-AUC Score	Auroc
Average Cross Validation Score	Cv
Log Loss Score	Ll

Table 3 provides the performance score of ML algorithms used on the asthma dataset without bagging. Among the evaluated models, CatBoost exhibited superior performance with an accuracy of 96.04%, an F1-score of 0.96, a Jaccard score of 0.92, an ROC-AUC score of 0.96, and a Log loss of 1.429. Notably, the SVC classifier attained the highest precision and F-beta score, reaching 99.52% and 0.9773, respectively. Furthermore, the Neural Network model

achieved the highest average cross-validation (CV) score of 0.9954, while the K-Nearest Neighbor model demonstrated perfect precision with a score of 1.00. LGBM, SVC, Random Forest, and Gradient Boosting models also displayed commendable performance, surpassing 90% accuracy. Multi-Layer Perceptron, Decision Tree, Logistic Regression, and AdaBoost achieved moderate performance with accuracy scores exceeding 80%. Overall, the models demonstrated good performance on this asthma dataset. Notably, applying SMOTE for addressing class imbalance led to a discernible improvement in performance. However, the utilization of Principal Component Analysis (PCA) for dimensionality reduction did not consistently yield enhanced results, suggesting that the efficacy of bagging, ensemble techniques, and dimensionality reduction is contingent upon the specific context and dataset characteristics. The distribution and inherent behavior of the data play a pivotal role in determining the suitability of these techniques.

TABLE III. COMPARATIVE PERFORMANCE OF ML MODELS

MODEL	Acc	Prec	Rc	F1	Fb	Jc	Auroc	Cv	Ls
CB	**96.04**	95.04	0.9714	**0.9608**	0.9545	**0.9245**	**0.9604**	0.9555	**1.4290**
XGB	95.81	94.44	0.9736	0.9588	0.9501	0.9208	0.9581	0.9533	1.5084
LGBM	95.81	94.83	0.9692	0.9586	0.9524	0.9205	0.9581	0.9555	1.5084
RF	95.81	95.81	0.9581	0.9581	0.9581	0.9197	0.9581	0.9550	1.5084
SVC	95.37	**99.52**	0.9119	0.9517	**0.9773**	0.9079	0.9537	0.9511	1.6672
NN	94.93	94.54	0.9537	0.9496	0.9471	0.9040	0.9493	**0.9954**	1.8260
GB	92.84	91.65	0.9427	0.9294	0.9216	0.8682	0.9284	0.9189	2.5802
MLP	87.67	87.34	0.8811	0.8772	0.8749	0.7812	0.8767	0.8993	4.4459
AB	87.00	84.15	0.9119	0.8753	0.8547	0.7782	0.8700	0.8477	4.6841
DT	87.33	86.61	0.8833	0.8746	0.8695	0.7771	0.8733	0.8755	4.5650
LR	86.23	85.22	0.8767	0.8643	0.8570	0.7610	0.8623	0.8488	4.9620
KNN	80.62	72.06	**1.0000**	0.8376	0.7633	0.7206	0.8062	0.8289	6.9864
NBC	80.40	77.27	0.8612	0.8146	0.7889	0.6872	0.8040	0.8062	7.0658

Models were trained extensively with and without pre-processing and bagging techniques. With bagging applied, performance improved marginally in some models while effects were adverse in some other implementation. Table III contains data before and after applying bagging to all the models. XG Boost performed best with bagging by achieving an accuracy of 96.04%.

TABLE IV. PERFORMANCE OF ML MODELS WITH BAGGING APPLIED

Models	Accuracy (in %)	Accuracy with bagging (in %)
Cat Boost	**96.04**	95.93
Random Forest	95.81	95.26
LGBM	95.81	95.70
XGBoost	95.81	**96.04**
Support Vector Classifiers (SVC)	95.37	94.93
Neural Networks	94.93	N/A
Gradient Boosting	92.84	92.84
Multi-Layer Perceptron	87.67	N/A
Decision Trees	87.33	91.63
Ada Boost	87.00	86.45
Logistic Regression	86.23	85.79
K-Nearest Neighbors	80.62	80.73
Naïve Bayes Classifier	80.40	80.62

Receiver Operating Characteristic (ROC) curves offer a powerful visual and qualitative assessment of the models' discriminative capabilities on the test dataset. Steeper slopes indicate higher sensitivity, measuring the models' capability to correctly identify positive cases. Ideally, the curves should hug the top left corner of the plot. Additionally, the area under the curve (AUC) provides a qualitative measure of performance. The larger the area under the curve denotes higher performance of the model. Fig. 3 represents the ROC characteristic observed with each ML model used in this study.

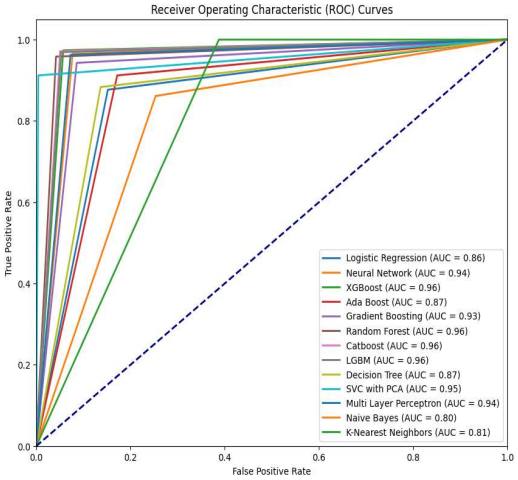

Fig. 3. ROC Curve for all ML models used

V. CONCLUSION AND FUTURE WORK

Empirically, we conclude that Cat Boost is the best machine learning algorithm for asthma disease detection with statistical data, especially in the context of the "[illegible] Asthma Disease Dataset [illegible]" from Kaggle. This model achieved accuracies of 96.04% in our tests, making it significantly better than most of the literature cited in section II. Per Table III, it delivered best results in five out of the eight metrics used for the comparison. Some notable mentions include Ada Boost and Support Vector Classifier (SVC), the latter of which was the best performing model in two of the evaluation metrics. While Ada Boost delivered similar accuracies to Cat Boost with bagging applied, the added computational cost of the ensemble technique makes it less feasible when implementing the model at scale.

In summary, this work presented three contributions: (1) a review of popular classification models, (2) all relevant recent ML-based applications in the field of asthma disease detection

and their performance as reported in literature, (3) an empirically-verified machine learning technique that best fits asthma detection. To our best knowledge, this is the first study to incorporate a comprehensive study of numerous popular machine learning models on a moderately large-scale numerical dataset. This proves the novelty of the paper.

Future work can expand on even more ML techniques for asthma detection. Better datasets with added features can be developed to address overfitting issues with some sophisticated models such as MLP and SVC. Further, multi-modal data involving both coughing sounds, medical history, and statistical information could be combined for more effective predictive models.

ACKNOWLEDGEMENTS

The author extends sincere thanks to Dr. Junali Jasmine Jena and Dr. M. K. Gourisaria for their invaluable guidance and support throughout this research. Their unwavering encouragement and insightful direction at every stage were instrumental in the successful completion of this endeavor.

REFERENCES

[1] B.Lambrecht, H. Hammad, "The immunology of asthma. Nat Immunol" 16, 45–56 ,2015

[2] C. Porsbjerg E. Melén L. Lehtimäki, D. Shaw, "Asthma". Lancet. Mar 11;401(10379):858-873. 2023

[3] N.A.Molfino, G. Turcatel, & D. Riskin, "Machine Learning Approaches to Predict Asthma Exacerbations: A Narrative Review". Adv Ther 41, 534–552,2024

[4] R. Beasley, A. Semprini & E.A.Mitchell "Risk factors for asthma: is prevention possible?" The Lancet, 386 (9998), 1075–1085,2015.

[5] H. GM, M.K. Gourisaria, S.S. Rautaray, and M. Pandey. "Pneumonia detection using CNN through chest X-ray." *Journal of Engineering Science and Technology (JESTEC)* 16, no. 1: 861-876.2021.

[6] M.K. Gourisaria, G. M. Harshvardhan, R. Agrawal, S.S. Patra, S.S. Rautaray, and M. Pandey. "Arrhythmia detection using deep belief network extracted features from ECG signals." *International Journal of E-Health and Medical Communications (IJEHMC)* 12, no. 6: 1-24.2021.

[7] V. Singh, M. K. Gourisaria, G. M. Harshvardhan, and V. Singh. "Mycobacterium tuberculosis detection using CNN ranking approach." In *Advanced Computational Paradigms and Hybrid Intelligent Computing: Proceedings of ICACCP 2021*, pp. 583-596. Singapore: Springer Singapore, 2021.

[8] M.K. Gourisaria, S. Das, R. Sharma, S. S. Rautaray, and M. Pandey. "A deep learning model for malaria disease detection and analysis using deep convolutional neural networks." *International Journal of Emerging Technologies* 11, no. 2: 699-704.2020.

[9] A. Yahyaoui and N. Yumuşak, "Deep And Machine Learning Towards Pneumonia And Asthma Detection," 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Zallaq, Bahrain, 2021, pp. 494-497, doi: 10.1109/3ICT53449.2021.9581963.2021.

[10] D. Spathis, & P. Vlamos "Diagnosing asthma and chronic obstructive pulmonary disease with machine learning". *Health Informatics Journal*, *25*(3), 811–827.2017.

[11] O. Zhang, L. L. Minku, & S. Gonem "Detecting asthma exacerbations using daily home monitoring and machine learning". *Journal of Asthma*, *58*(11), 1518–1527,2020.

[12] J. Zhan, W. Chen, L. Cheng, Q. Wang, F. Han, & Y. Cui "Diagnosis of asthma based on routine blood biomarkers using machine learning". *Computational Intelligence and Neuroscience*, *2020*, 1–8,2020.

[13] J.L.M. Amaral, A.G. Sancho, A.C.D. Faria, A. J., Lopes, & P. L. Melo, "Differential diagnosis of asthma and restrictive respiratory diseases by combining forced oscillation measurements, machine learning and neuro-fuzzy classifiers". *Medical & Biological Engineering & Computing*, *58*(10), 2455–2473,2020.

[14] M. A. Awal, M. S.Hossain, K.Debjit, N.Ahmed, R. D.Nath, G. M. M.Habib, M. S.Khan, M. A.Islam, & M. a. P. Mahmud, "An early detection of asthma using BOMLA detector". *IEEE Access*, *9*, 58403–58420,2021

[15] M. Xie, & C. Xu "Predicting the risk of asthma development in youth using machine learning models". *medRxiv (Cold Spring Harbor Laboratory)*,2024

[16] P. D.Barua, T. Keles, M. Kuluozturk, M.A.Kobat, S.Dogan, M.Baygin, T.Tuncer, R. Tan, & U.R.Acharya, "Automated asthma detection in a 1326-subject cohort using a one-dimensional attractive-and-repulsive center-symmetric local binary pattern technique with cough sounds". *Neural Computing and Applications*, *36*(27), 16857–16871,2024

[17] R. El-Kharoua, "Asthma Disease Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset. [Accessed: Sept. 09, 2024].

[18] P. Bahad, and P. Saxena. "Study of adaboost and gradient boosting algorithms for predictive analytics." In International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019, pp. 235-244. Springer Singapore, 2020.

[19] C. Wade, K. Glynn. "Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python". Packt Publishing Ltd, 2020.

[20] T. Swetha, R. Roopa, T. Sajitha., B. Vidhyashree, J. Sravani, B. Praveen. "Forecasting Online Shoppers Purchase Intentions with Cat Boost Classifier". In 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT) (pp. 1-6). IEEE. 2024.

[21] M. J. Sai, P. Chettri, R. Panigrahi, A. Garg, A.K. Bhoi, P. Barsocchi, "An ensemble of Light Gradient Boosting Machine and adaptive boosting for prediction of type-2 diabetes". International Journal of Computational Intelligence Systems, 16(1), 14, 2023.