National College of Ireland

Hybrid predictive model for Asthma Diagnosis Using Environmental and lifestyle factors

MSc Research Project
MSc in Data Analytics

Anna Joy
Student ID: x23238241

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Research Submission Sheet

School of Computing

Student Name: …….ANNA JOY……

Student ID: ……x23238241……..……

Programme: ……MSc in Data Analytics…… **Year:** ……1……..

Module: ……MSc Research project…….……

Supervisor: ……Jorge Basilio……..……

Submission Due Date: ……12/12/2024……..……

Research Title: …….. Hybrid predictive model for Asthma diagnosis using Environmental and lifestyle factors……..……

Word Count: …… **18**……..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this research. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the research.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Anna Joy……

Date: ……11/12/2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each research (including multiple copies)	□
Attach a Moodle submission receipt of the online research submission, to each research (including multiple copies).	□
You must ensure that you retain a HARD COPY of the research, both for your own reference and in case a research is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hybrid predictive model for Asthma diagnosis using Environmental and lifestyle factors

Anna Joy
x23238241

Abstract

Asthma diagnosis contains various challenging problems and issues which emerge due to its complexity of the disease by the influence of various environmental, lifestyle and historical factors that included in imbalanced datasets. To ensure timely interventions, improving patient outcomes and reducing the economic burden, which is related with untreated or misdiagnosed asthma, early detection is very crucial. Traditional diagnostic methods such as Logistic Regression, SVM, decision tree and Random Forest often struggle to accurately predicting asthma cases that particularly in imbalanced datasets where non-asthma cases dominate which result high accuracy. This research study mainly focus the limitations of traditional diagnostic models in handling such complexities by proposing a hybrid machine learning model that combines Gradient Boosting and Neural Networks. In order to handle the issue of class imbalance, we have used two different oversampling technique such as synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN). While these methods improved data balance which shows a high overall accuracy of 95% for predicting non-asthma cases using the hybrid model, even though fails to detect asthma cases, with a recall of 0.00 ,that were marginal which results in the need for more advanced methods to optimize hybrid architectures. We are analysing the stacked model that integrating neural networks and gradient boosting which has a poor performance with an AUC of 0.4864 when using oversampling technique ADASYN .
The research also explores the potential for various diagnostic tools and proposes meaningful future work that including advanced ensemble learning methods.

Keywords- Hybrid model, SMOTE, ADASYN, Asthma Diagnosis, class imbalance

1 Introduction

1.1 Research Background

About 5–10% of people have asthma, which which makes it the most prevalent chronic non-communicable disease in the world. It still confronts a significant challenge in the diagnosing process due to its intricate molecular mechanics. There are many predictive and diagnostic models that being developed rely on “bio signals” and “phenotypic symptoms”, but relatively few of these models are centred on environmental variables (Huang, 2024). Since asthma has many different contributing elements and factors, many of which are intricate and

interrelated, so thereby the diagnosing the condition is getting more difficult nowadays. Treatment may be delayed or unsuccessful if current diagnostic techniques overlook instances or give erroneous evaluations. Therefore, in order to increase the accuracy of predictions, the researcher intends to develop a “hybrid predictive model” which incorporates Gradient Boosting and Neural Network techniques. This model expected to be specifically target the identification of actual positive instances of asthma (high recall) when compared to various conventional approaches, the model will enhance early detection and deliver a more tailored diagnosis by integrating several variables such as patient information which including lifestyle and environmental variables and medical records.

1.2 Research Significance

This research deals with the critical need in asthma diagnostics, since it constructs a hybrid predictive model built on Gradient Boosting combined with Neural Networks. It could use different kinds of patient data and incorporate factors like environmental and lifestyle conditions to build a better recall and possibly accuracy in asthma diagnosis (Lee *et al.*, 2024). For that reason, early and specifically tailored asthma detection might reduce instances of delayed treatment or ineffective interventions and therefore, would bring about better results for patients.

1.3 Research Objectives and Questions

- To build a hybrid predictive model by using two traditional models such as Gradient Boosting and Neural Networks to enhance the efficiency and accuracy of asthma diagnosis by using various factors.
- To integrate diverse data sources to identify a comprehensive and personalized asthma diagnosis approach by comparing different models.

Question

How does the integration of neural networks and gradient boosting techniques improve the predictive accuracy and efficiency of asthma diagnosis when compared to traditional diagnostic methods?

1.4 Research Rationale

Asthma's multiperature causes, which is also interdependent and has all contributing components, that challenges the traditional diagnostic approach leading to the diagnosis being incorrect (Kavanagh et al., 2019). Hence, this attempt applies Gradient Boosting alongside with Neural Networks to improve a measurable diagnostic accuracy and recalls towards early detection and well-curated treatment as demanded and accepted in modern health management.

1.5 Research Limitations

It will also introduce noise into the model due to the complexity and interdependence of the contributing factors for asthma. There would also be variability in the data quality, especially

about the environmental and lifestyle factors affecting predictions. Class imbalance persists despite the "Synthetic Minority One-sampling Technique" (SMOTE) and "Adaptive Synthetic Sampling" (ADASYN). The scalability of the model across diverse populations also presents an issue.

This study may face the challenges of relying on the quality and completeness of the dataset where accuracy shall be impacted, class imbalance may still be imperfect even though SMOTE and ADASYN are used. Hybrid models are difficult and may require much computation; generalizability may be limited to populations and datasets with similar characteristics and variables.

1.6 Ethical considerations

The researcher has kept patient data confidential and adheres to ethical standards such as GDPR. Data anonymization has been protected with sensitive information, and the researcher also took informed consent to ensure ethical use. Modelling has been designed to be unbiased and does not produce outcomes that are discriminatory; thus, it is fair and provides diagnostic accuracy equally to all demographics.

1.7 Structure of the research

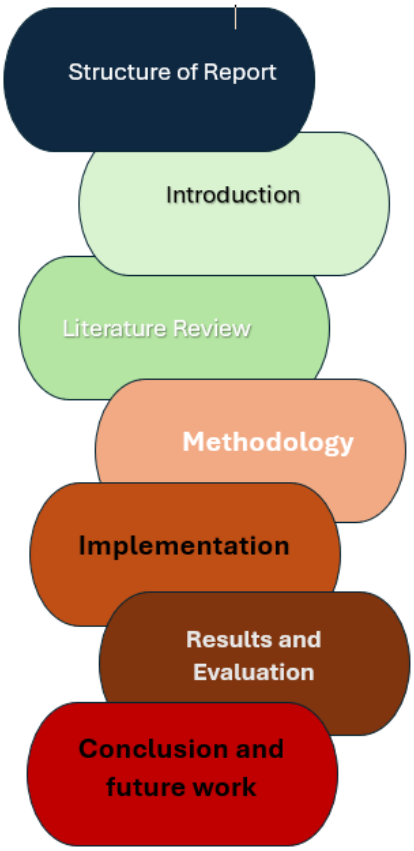

Fig: Structure of the Research

2 Literature Review

2.1 Concept of predictive model for asthma diagnosis

A large number of asthma diagnostic prediction models have been created for implementation in hospital environments which need data that is normally only accessible through medical institutions and these models frequently make utilization of comprehensive clinical data and patient populations. In order to forecast adult asthma diagnoses, Tomita et al., (2019), created a model that combines “Deep Neural Network” (DNN), “Support Vector Machine” (SVM) and “Logistic Regression”. Their study achieved a 98% DNN accuracy rate using clinical data from 566 “adult outpatients” at “Kinki University Hospital”, including symptoms, physical indicators, biochemical outcomes, respiratory function tests, as well as “bronchial challenge test results”. Similarly, Spathis and Vlamos, (2019), proposed an asthma predictive model for patients with 132 patients' medical history, demographics, habits, lung measurements, and symptoms. The “Random Forest” performed better as compared to other approaches such as “Naive Bayes”, “Logistic Regression” and SVM. This study is successful in demonstrating the usability of machine learning models, though it also raises a number of issues that depend upon clinical data and therefore not usable everywhere.

Loymans *et al.,* (2018) say that various prediction models were developed, which help to evaluate the chances of exacerbation among asthma patients. Having all these clinical and patient-level characteristics, several predictors have been established for asthma attacks such as past healthcare use, presence of symptoms and outcome of respiration test. Most models actually applied in clinical practice to the task of an asthma attack forecast depend upon older methods by logistic regression. These later methods failed at the highly imbalanced outcomes problem exposed by King and Zeng in 2001. Thus, while more precise estimates, further analytical sophistication enabling complex, non-linear connections and relationships of different variables is emerging over time Lisspers *et al.,* 2021. While a diagnostic instrument for “short-term risk prediction” of asthma exacerbations may offer a chance to respond and treat a transient and impending heightened risk, “long-term risk predictions” are helpful for establishing a long-term management plan. Both better patient outcomes and lower healthcare use costs might result from prompt intervention (Lisspers *et al.,* 2021).

2.2 Comparison between predictive model and traditional model for asthma diagnosis

The technological development brings tremendous potential into the area of medical diagnosis, particularly concerning such conditions as asthma, that are complicated in aetiology. According to Johnson *et al.,* (2019), traditionally, asthmatic diagnostic methods that are employed through spirometry and clinical assessment of symptoms are lacking because they fail to capture the many dimensions involved, such as demographics, the environment, and lifestyle, not to mention the medical background and symptomatology. Limitations to this will most likely lead to misdiagnosis or delayed treatment; hence more robust approaches to diagnosis are required. More hybrid predictive models that combine different machine learning algorithms into a better alternative have emerged. Whereas Chen

and Guestrin, (2016), state that "Gradient Boosting Machines" (GBMs) were more appropriate for analyzing structured data such as demographics and medical histories, which offer great interpretability with efficiency in ranking features based on importance. At the same time, "Neural Networks" (NNs) demonstrated an exceptional ability to understand nonlinear relationships and complex patterns and are very appropriate for high-dimensional continuous data like lung function metrics, or symptom clusters (Goodfellow, 2016).

Hybrid models integrate the best of GBMs and NNs to reach higher diagnostic precision through ensemble learning techniques. As per Devarajan *et al.,* (2021), hybrid models have been observed to be more sensitive than traditional methods with respect to critical metrics like recall in medical diagnostics where false negatives are to be minimized. The stacking approach in hybrid models enhances performance by aggregation of predictions from multiple classifiers, and hence it achieves better generalization across diverse datasets (Wolpert, 1992). With a hybrid predictive model for asthma, can utilize multivariate input data, resulting in an individualized diagnosis that would also occur earlier on. In other words, it surpasses traditional methods as these solely rely on single, isolated variables; and is part of the emerging push towards precision medicine within the medical community (Ginsburg and Phillips, 2018).

2.3 Effectiveness of collaborating Gradient Boosting and Neural Network

GBMs and NNs have recently gained popularity as hybrid models because they complement one another in terms of strength in predictive modelling, more so in healthcare applications. According to Chen and Guestrin, (2016), GBMs are a form of ensemble learning that lends itself well to structured data because it can handle categorical and numeric in addition to determining feature importance while handling missing data with minor preprocessing. Whereas, in comparison, NNs successfully capture complex, non-linear relationships and analyze unstructured or high-dimensional data such as physiological signals or imaging datasets (LeCun *et al.,* 2015). Adding these models together allows for a holistic look at each type of data that enhances the capability of generalization.

The hybrid model has its strength in merging the various capabilities of different algorithms to outmaneuver individual weaknesses. Shwartz-Ziv and Armon, (2022) consider GBMs to be an efficient feature engineering and selection technique that reinforces pattern recognition for noisy or high-dimensional datasets. As a result, this finds applications in medical diagnostics where data heterogeneity becomes a huge problem. For instance, although GBMs predict the disease outcome from clinical and demographic data quite well, there is still a possibility that NNs pick out very subtle patterns which exist in symptoms or imaging results (Bharati *et al.,* 2020). These types of hybrid approaches based on empirical evidence are indeed efficacious. For example, a cancer diagnosis paper that utilizes GBMs in conjunction with NNs reported an accuracy rate of 94%, on average about 7% better than the best models that were created independently of each other (Shahbazi *et al.,* 2020). Similar results were established in the respiratory disease prediction where hybrid models achieved better recall and F1 score as compared to other conventional methodologies and showed better ability towards dealing with unbalanced data sets (Alam *et al.,* 2023). A combination of GBMs with NNs is an excellent mechanism through which both these approaches may be optimized. This shall benefit healthcare sectors with comprehensive, accurate, and personal predictions.

3 Research Methodology

3.1 Data collection

The patient data has been collected from Kaggle and these data are structured and consist of demographic, environmental, lifestyle, and medical history data. The sources may be from health care databases, public repositories or clinical records. All relevant variables concerning the diagnosis of asthma are covered. The primary focus of the collection method is on accuracy and significance to the goals of the research.

3.2 Sampling method

A random sampling method was conducted to ensure the dataset represented the target population. There were three divisions of sets such as a training set which accounting for 70%, then a validation set at 15% and the testing set, which constituted 15%. Random sampling was made to avoid selection bias, guaranteeing the model's reliability and generality. In terms of handling the imbalanced dataset, especially when asthma-positive cases are underrepresented, oversampling techniques such as SMOTE and ADASYN can be used.

3.3 Data analysis method

For the preprocessing of the data analysis to ensure that the data are good and homogeneous, standardization was used for continuous variables such as BMI and pollution, one-hot or label encoding was applied for categorical variables such as gender and ethnicity. The imbalanced class of asthma versus non-asthma used SMOTE and ADASYN to treat the issue of unequal distribution between the two classes in the dataset. Then the code has been run in python Jupiter notebook to analyse it to develop the model which can be validated with machine learning. Gradient Boosting model and neural network model was applied to structured data such as like demographics and medical history, which contain selected features and are well interpretable which including non-linear relationships in lung function and symptom data. The models were combined together to create a hybrid predictive model using the stacking method. Finally, the performance of the model was tested with metrics such as recall, accuracy, F1 score, and ROC-AUC, specifically, recall maximization was focused on.

4 Research design

This research has been employed a quantitative approach to design a hybrid predictive model for asthma diagnosis, based on processing various patient data with a better accuracy to diagnose by using the machine learning techniques by means of gradient boosting and neural networks. These models were further enhanced through ensemble methods, with a stacking approach by integrating neural networks and gradient boosting as base learners and logistic regression considered as the meta-model. This hybrid model by combining two other models

demonstrated improved diagnostic accuracy during validation. Performance evaluation was conducted by using metrics such as confusion matrices, classification reports, ROC-AUC scores and ROC curves. This multi-model approach which including ensemble learning and advanced sampling techniques, it enabled a comparative analysis of predictive capabilities for asthma diagnosis. The design will integrates both theoretical insights and computational methods by aligning with a data-driven methodology to improve asthma diagnosis accuracy.

5 Implementation

This Research is implemented in python Jupyter notebook that consisting various libraries such as panda, sicikit learn, tensor flow and many other functions . It preprocesses the data by standardization, encoding categorical variables as well as dealing with over samples using SMOTE and the ensemble method ADASYN with handling class imbalance. Different versions traditional models are built such as Logistic Regression, Random Forest, Decision tree, gradient boosting, neural network and SVM after applying oversampling technique called SMOTE . Ensemble learning techniques were employed to enhance predictive accuracy, with stacking used to combine neural networks, gradient boosting. The ensemble approach involved training a meta-classifier on predictions from the base models. Model evaluation was conducted by using various metrics such as confusion matrices, classification reports and ROC-AUC scores. ROC curves mainly visualized to compare sensitivity and specificity in the model.

To further procedures, a hybrid model which integrating neural networks and gradient boosting predictions was developed by using stacking method, with logistic regression serving as the final estimator. In order to improve the class imbalance, ADASYN technique also performed in the following procedure. This structured leverages advanced machine learning techniques and ensemble strategies for medical diagnostics in asthma cases.

6 Evaluation

Different models that including traditional machine learning models such as Logistic Regression, Random Forest, Gradient Boosting, Decision Tree and Support Vector Machine were performed and in order to address imbalance, which is a critical issue in health-related classifications, Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling Approach (ADASYN) and a hybrid model that combines different algorithms such as gradient boosting model and neural network model through stacking. When it comes to identifying asthma cases, the results clearly indicates that all the models including the hybrid model, after with oversampling techniques such as SMOTE and ADASYN also suffer from poor predictive performance. In healthcare industry, the missed opportunities to identify asthma patients, potentially delaying critical interventions and impacting patient outcomes. In addition to that, inaccurate predictions can lead to operational inefficiencies, misallocation of resources and reduced trust in automated diagnostic tools.

Despite achieving high accuracy for non-asthma cases, the failure to predict asthma patients indicates the need for improved strategies in handling imbalanced datasets. The evaluation of each model's performance is detailed based on several key metrics such as confusion matrices, classification reports and ROC-AUC scores is given below:

6.1 Logistic Regression

The Logistic Regression model which is considered as a go-to method for classification tasks, yielded a **Confusion Matrix** in **(**table.1), that reflected the issues posed by the class imbalance. This model has a high **True Negative** value of **422**, that means it correctly classifies non-asthma patients most of the time and along with this, the false positive value is also quite large, even though it indicates that it often misclassifies non-asthma patients as asthma patients which can lead to unnecessary testing, patient anxiety and increased healthcare costs. However, the **16 True Positives (TP)** suggests that the model performs poorly at detecting asthma cases and **21 False Negatives (FN)** means that it misses some asthma patients which highlighted the model's failure to identify asthma patients, cause delay timely medical intervention, risking patient health and potentially leading to severe complications.

The classification report for Logistic Regression describes that the precision for non-asthma patients is high, has a value of 0.95, but meanwhile for asthma patients it is alarmingly low value of 0.06, which means that when the model predicts asthma cases are often wrong. When we consider the case of recall, it was 0.43, which suggests that it missed a large number of cases while it identified a portion of asthma patients which is a critical concern, particularly in a medical diagnosis context and this was further corroborated by analyzing the F1-score of value 0.10, that emphasizes poor predictive ability for asthma classification. The ROC-AUC score of 0.5406 justifies these findings that a score close to 0.5 indicates that the model is barely distinguishing between the two classes, which provides a weak model for practical use in classification tasks in this analysis.

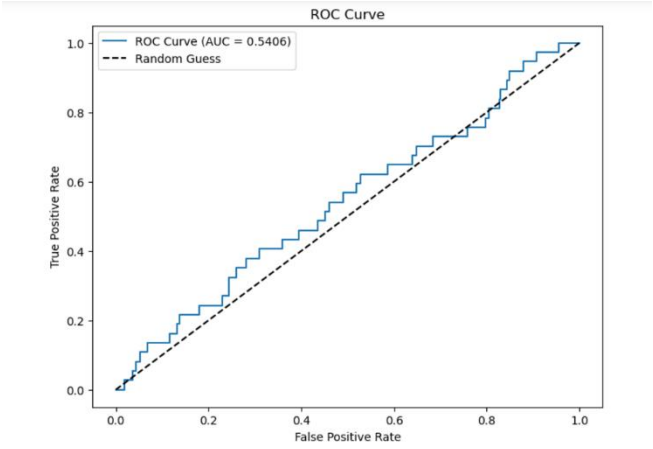

Fig.1

The ROC curve which is only slightly above the random guess line that can be reflected in the AUC value of 0.5406 which indicates that the logistic regression model for analyzing asthma diagnosis performs only marginally better than random guessing for this classification task . It is unable to detect asthma cases which consider as a critical problem particularly in the case of medical applications where predicting rare conditions is essential.

6.2 Random forest

The **Random Forest** model for predicting asthma diagnosis performed much better in classifying non-asthma cases by achieving an impressive accuracy of **95%**, but in the case of detecting asthma cases, it is skewed by the model's inability to predict asthma cases with **0 True Positives** values and **37 False Negatives** values in confusion matrix are critical in healthcare, cause undiagnosed asthma patients may miss timely treatment which lead to deteriorating health and increased risk of severe outcomes. Random Forest had almost no ability to distinguish between asthma and non-asthma which can be confirmed by the **ROC-AUC** score of **0.4602,** precision (0.00) and recall(0.00).

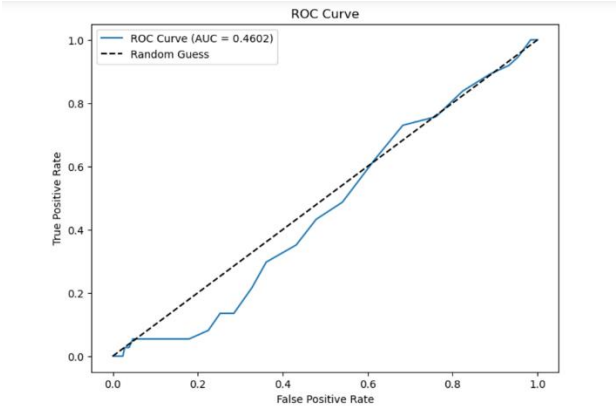

Fig.2

The ROC curve mainly fluctuates close to the random guess line with no significant rise in the TPR for increasing FPR which indicates that the random forest model fails to balance true positives and false positives effectively in the prediction of asthma diagnosis. The class imbalance which results a significantly higher number of non-asthma cases than asthma cases might have led to the model being biased towards the majority class. Deploying such a model would lead to undiagnosed asthma cases, which could increase emergency medical interventions, patient dissatisfaction and potential liability risks for healthcare institutions.

6.3 Gradient Boosting

The Gradient Boosting model performs slightly better than the Random Forest model while performing asthma diagnosis after applying oversampling technique, with **678 True Negatives** and **3 False Positives**. When it comes to the minority class, Random Forest model completely fails to identify asthma cases which results in **37 False Negatives** and **0 True Positives**. This results leads to missed opportunities for timely medical interventions and creates health complications for undiagnosed patients which negatively affecting patient outcomes and quality of care. The performance of Gradient Boosting mirrors when compared to Random Forest in terms of its ability which shows good classification performance for non-asthma patients but completely misses detecting asthma cases.

The precision from classification report for non-asthma patients is **0.95** meanwhile it is **0.00** for asthma patients. Following that the recall for non-asthma patients is perfect (**1.00**), but for asthma patients is **0.00** which means that no asthma cases are detected by the gradient boosting model. The **ROC-AUC** of **0.5222**, which slightly better than Random Forest but still suggested limited discriminatory ability between the two classes such as asthma and non-asthma classes.

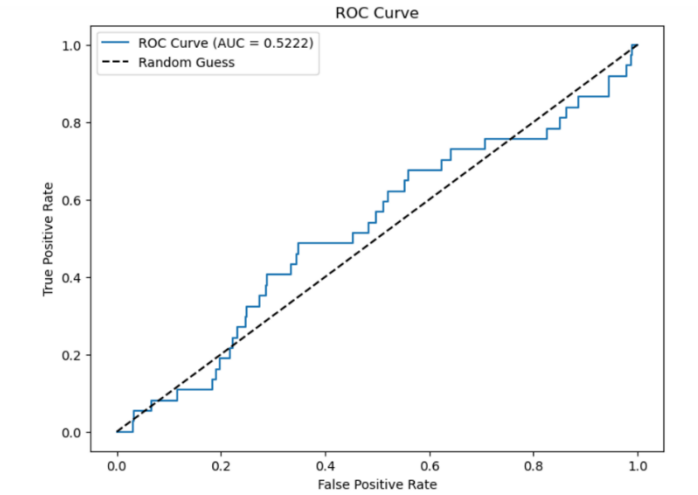

Fig.3

The ROC curve is more close to the diagonal line of random guessing and has small fluctuations above the baseline which indicates that the gradient boosting model may lacks strong discriminatory power between classes. This creates an imbalance in service delivery, where the needs of non-asthma patients are met, but those of asthma patients are entirely overlooked, which undermining the credibility of the solution in a clinical or business application.

6.4 Decision Tree

In contrast to other models, **Decision Tree** classifier performed somewhat better than the previous models which results in a more reasonable confusion matrix and It achieved 4 **True Positives** for asthma patients which demonstrates some ability to detect asthma patients, although the recall is still quite low and the **False Negatives** are **33**. The decision tree model correctly classifies **638** non-asthma patients and misclassifies **43** non-asthma patients as asthma. This class Imbalance generates a strategic challenge in meeting healthcare objectives, cause the solution skews heavily toward one class, which undermining its utility in practical applications.

The precision for non-asthma patients is **0.95** and for asthma patients is **0.09** when performing the decision tree model. The model's inability to correctly identify asthma cases consistently reflects in Recall for asthma patients, which is low (**0.11**), meanwhile the recall for non-asthma cases has a very high value of 0.94. The **ROC-AUC** score of decision tree model is **0.5225** which is similar to that of Random Forest and Gradient Boosting concluding that Decision Trees models are more interpretable, but still struggled with the class imbalance problem.

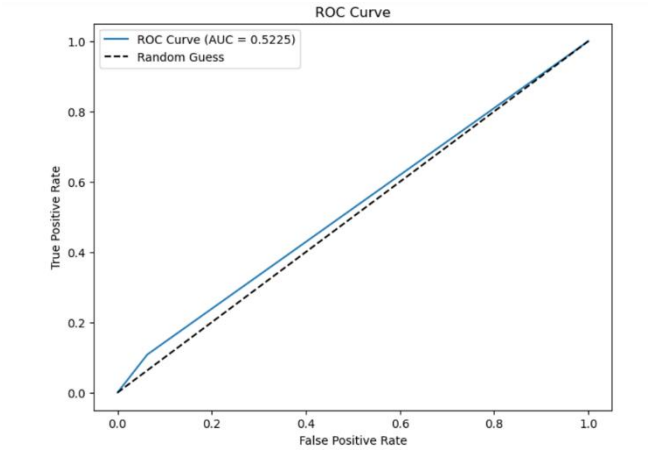

Fig.4

By detecting a few asthma cases, The decision tree model shows some improvement over Random Forest and Gradient Boosting, but it still fails to reliably classify asthma patients and it cannot be justify the operational costs of deploying this model, especially when considering the potential downstream expenses associated with undiagnosed conditions. This indicates that Decision Trees require additional modifications to better handle imbalanced data by using various oversampling techniques.

6.5 Support vector machine

The **SVM** model showed identical results to Random Forest and Gradient Boosting that it classifies all non-asthma patients correctly patients with **681 True Negatives (TN)** and **37 False Negatives (FN**), but fails to predict any asthma cases **which results in** a **0 True Positives**. Misdiagnosed asthma patients may remain untreated, which resulting in worsened health outcomes and potential liability risks for healthcare providers. The model is biased toward the majority class which is clearly evident from the value of precision for asthma with 0.00 value while for non-asthma cases its 0.95 value. The **ROC-AUC score** has a value of **0.4624** that confirms this model has poor performance in distinguishing between asthma and non-asthma cases, which undermines the business viability of deploying the model, as it does not provide meaningful insights for clinical or operational decision-making.

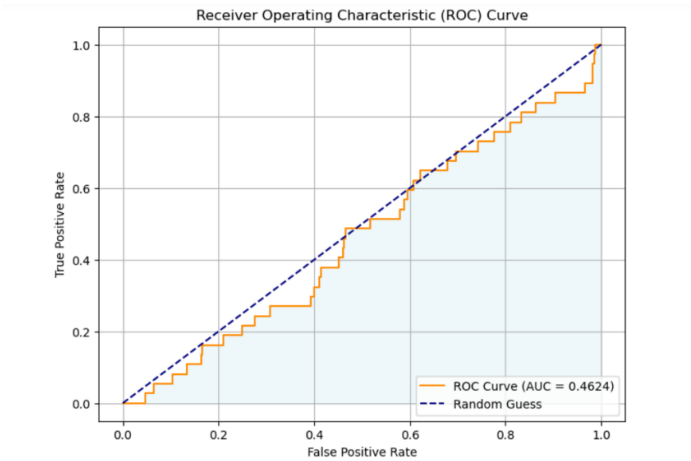

Fig.5

SVM is a robust algorithm which is used for classification tasks but can be highly sensitive to class imbalance. Without appropriate adjustments such as oversampling techniques SVM will struggles in this scenario, which can taken as evidence by its failure to detect asthma cases using lifestyle and environmental factors.

6.6 Neural network model

The neural network model after performing oversampling technique like smote describes a better performance with an evident of nearly perfect accuracy 92% on the training data and up to 100% on the validation data which can be clearly shows in fig.7. The following analysis shows a remarkably low loss values such as 0.0186 for training and 0.0045 for test data and these metrics indicate that the model has learned to represent the data effectively while avoiding overfitting. This efficiency mainly suggests reliability in predicting non-asthma cases, even with unseen data, which is critical for scalable deployment in healthcare systems. By emerging this level of performance of the neural network model suggests that it is able to capture even the most nuanced patterns in the imbalanced dataset, by integrated into workflows, reducing manual diagnostic errors and improving decision support for clinicians.

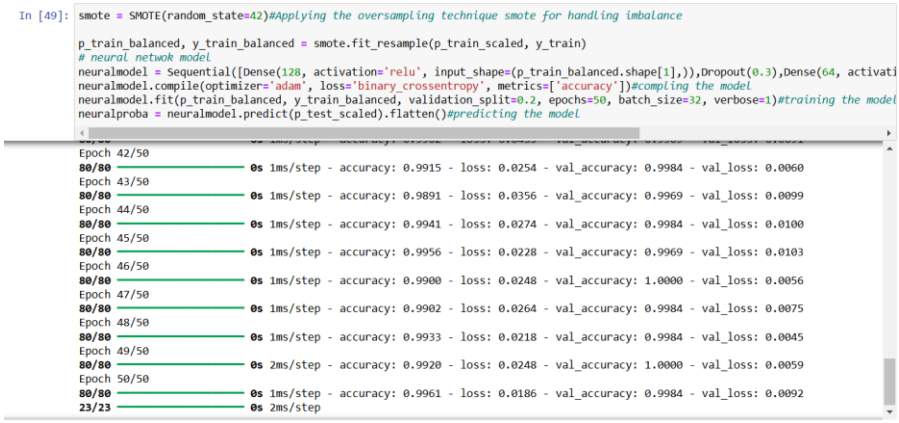

Fig.6

The rapid growth which can be observed during training where achieving high accuracy within the first few epochs shows the model's efficiency in learning meaningful representations that can be useful in prediction asthma diagnosis. The model has a strong generalization capabilities which making it reliable when applied to unseen data. So that's why this model will be chosen as good efficient model that can be used to create hybrid model by combining its strengths with other algorithms may further enhance accuracy and robustness, particularly in scenarios requiring both interpretability and predictive power.

6.7 Hybrid model with SMOTE

By generating synthetic samples for the minority class (asthma), the **hybrid model** after performing **SMOTE** aimed to address class imbalance, unfortunately the outputs of the hybrid model which is done by using stacking method did not show significant improvement, that results in the confusion matrix with a value of **681 True Negatives (TN)** and **37 False Negatives (FN).** Along with **0 True Positives (TP)** and **0 False Positives (FP)** values were emerged as a result of this hybrid model which is a combination of gradient boosting and neural networking. Following that, the **ROC-AUC score** of **0.5364** suggested a slight

improvement over previous models, even though it still failed to identify any asthma cases by using various factors. This results can undermine clinical decision-making, which resulting in either missed diagnoses or unnecessary treatments or left untreated, that leads to poor patient outcomes and increased healthcare costs due to complications arising from delayed treatment in asthma diagnosis.

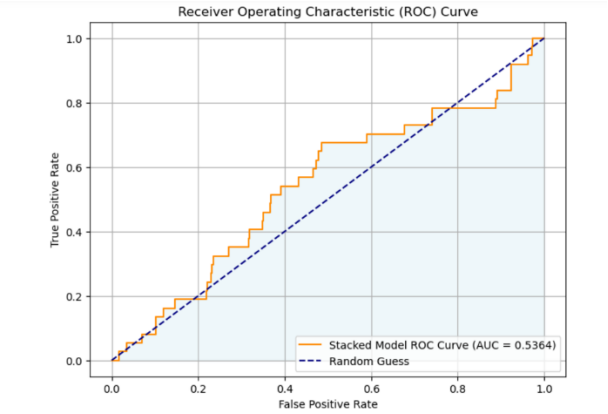

Fig.7

The ROC curve for the stacked hybrid model demonstrates a weak performance where the Area Under the Curve (AUC) score is 0.5364, that cannot predict some rare conditions such as asthma which means that the model will be of limited use in real-world applications, where detecting rare or underrepresented conditions is crucial for patient safety and effective healthcare delivery.

This approach alone may not sufficient to improve model performance in detecting rare cases like asthma. The model continues to be biased toward the majority class which may due to the impact of gradient boosting and neural network, base models which cannot predict the minority classes. So implementing another technique might have improve the efficiency of the model.

6.8 Hybrid model with ADASYN

The hybrid model which is a combination of neural network model and gradient boosting by using stacking method with an oversampling technique ADASYN produced a slight improvement when compared to SMOTE. The confusion matrix contains 662 True Negatives values and 19 False Positives along with 36 False Negatives and 1 True Positive values. This resulted in the case of asthma prediction with a precision of 0.05 and recall of 0.03 for asthma patients, meanwhile the accuracy for non asthma cases was still high at 92%. The F1-score for asthma cases for hybrid model remaining low at 0.04 and the ROC-AUC score of 0.4864 indicated little improvement while consider the smote technique.

Confusion Matrix:

[[662 19]
[36 1]]

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.97	0.96	681
1	0.05	0.03	0.04	37
accuracy			0.92	718
macro avg	0.50	0.50	0.50	718
weighted avg	0.90	0.92	0.91	718

table.1

Stacked Model ROC-AUC: 0.4864

The healthcare organizations while predicting asthma diagnosis models without addressing these imbalances, they risk misdiagnosing asthma patients or failing to detect asthma cases. The models like Logistic Regression and SVM were particularly sensitive to class imbalance and they need more adjustment to improve their ability to predict the minority class using various environmental and lifestyle factors. When comparing the hybrid model with traditional models such as decision trees, Random Forest and Gradient Boosting can handle some level of class imbalance, but they still require advanced sampling methods to overcome the class imbalance.

6.9 DISSCUSION

The results of this research study provide various valuable insights into the challenges posed by class imbalance in an imbalanced dataset containing various features such as Age, Pollution Exposure, Sleep Quality, Diet Quality, Physical Activity, LungFunctionFEV1, Lung Function FVC, BMI, Dust Exposure when performing machine learning models and hybrid models especially in the prediction of rare disease called asthma diagnosis. Despite the use of oversampling techniques such as SMOTE and ADASYN in order to address the imbalance, the various models generally performed poorly in detecting the minority class called asthma cases. Chen and Guestrin (2016) suggest that Gradient Boosting excels with structured data like medical histories, meanwhile Neural Networks are better at detecting non-linear relationships and this combination of both models should ideally allow for better performance in detecting asthma cases. As we consider Shwartz-Ziv and Armon (2022) ideas which contains even hybrid models may struggle if the dataset is highly imbalanced, as seen in this research.

Model	Accuracy	Precision		Recall		F1-score		ROC-AUC
		0(non-asthma)	1(asthma)	0	1	0	1	
Logistic Regression	0.61	0.95	0.06	0.62	0.43	0.75	0.10	0.5406
Random Forest	0.95	0.95	0.00	1.00	0.00	0.97	0.00	0.4602
Gradient Boosting	0.94	0.95	0.00	1.00	0.00	0.97	0.00	0.5222

Decision Tree	0.89	0.95	0.09	0.94	0.11	0.94	0.10	0.5225
Support Vector Machine (SVM)	0.94	0.95	0.00	0.99	0.00	0.97	0.00	0.4624
Neural Network	0.94	0.95	0.00	0.99	0.00	0.97	0.00	0.4893
Stacked Hybrid Model using SMOTE	0.95	0.95	0.00	1.00	0.00	0.97	0.00	0.5364
ADASYN + hybrid Model	0.92	0.95	0.05	0.97	0.03	0.96	0.04	0.4864

Table.2

From table.2, we can understand that, ADASYN showed a slight improvement in hybrid model in terms of detecting a few asthma cases which has accuracy 92%. This hybrid model miss a significant number of asthma diagnoses, which could lead to delayed treatment, worsened patient outcomes and higher healthcare costs due to the need for more urgent treatments later on.

In case of **ROC-AUC scores**, the models were slightly above random chance, they did not even reach levels that would be considered acceptable for detecting rare diseases. But this research paper we are comparing the traditional models performance with the hybrid model which is a combination of neural network and gradient model. So the performance of hybrid is slightly better by oversampling technique ADASYN when compared to other models with SMOTE. This suggests that all models had some ability to distinguish between classes, even though they lacked the robustness required for rare event detection and in order to improve model reliability, we need to consider the combination of other multiple methods like cost-sensitive learning or anomaly detection.

7 CONCLUSION AND FUTURE WORKS

This research study mainly focused to improve asthma diagnosis by developing a hybrid predictive model that combines neural networks and gradient boosting techniques that leveraging diverse data that including various environmental, lifestyle and medical factors. The research question for this study was "How does the integration of neural networks and gradient boosting techniques improve the predictive accuracy of asthma diagnosis compared to traditional diagnostic methods?". In order to minimize false negatives by focusing on recall values along with the main objective to improve diagnostic accuracy, we introduced the hybrid model which is a combination of other two base models. The neural network component of the model was designed to capture various complex patterns in lung function and symptoms, meanwhile gradient boosting also addressed demographic and medical history data and by integrating both these strengths, the hybrid model sought to overcome the limitations of traditional diagnostic approaches. After performing the hybrid model, it demonstrated mixed results. The neural network also achieved better performance metrics in non-asthma cases which indicating its strength in identifying complex relationships in patient

data. Unfortunately, the stacked model integrating neural networks and gradient boosting after using smote has a poor performance with an AUC of 0.5364.

When compared to other traditional models like logistic regression, random forest, gradient boosting, decision tree and support vector machine also have low efficiency in predicting the minority section called asthma diagnosis, even though they predict high accuracy on non-asthma cases. The full hybrid model did not fully cover the promise of enhanced predictive performance. Despite of this results, this research provided valuable insights into the applicability of machine learning models in asthma diagnostics.

7.1 Key Findings

The neural network model showed a strong predictive capability by capturing non-linear patterns in complex medical data. Meanwhile stacked hybrid model which is a combination of neural network and gradient boosting after using ADASYN hold its limitations which shows the model's AUC of 0.4864 indicated a failure to effectively combine the strengths of the neural network and gradient boosting. The poor performance of the stacked hybrid model suggests that more sophisticated and efficient integration methods or alternative ensemble techniques are necessary for predicting asthma diagnosis.
This research mainly contributes to the growing field of asthma diagnosis that offering various valuable insights into combining predictive models for complex medical diagnoses. On the other side, the hybrid model here which is used in asthma diagnosis did not fully meet expectations and this study needs more groundwork for refining such approaches.

7.2 Future works

Future work for this research paper focuses on addressing its limitations such as imbalance, overfitting and advancing its applicability in order to predict asthma diagnosis using various models. Key proposals that can be include more sophisticated ensemble techniques, such as weighted averaging, attention-based integration or meta-learners to enhance the model efficiency between neural networks and gradient boosting and there by prediction of asthma cases in an imbalanced dataset. Improved several feature engineering that including different interaction terms, feature selection and dimensionality reduction could enhance the model's input data for better performance. Along with that, improving transfer learning through pre-trained models may improve the representation of complex medical, environmental and lifestyle factors.
The research holds significant and crucial potential for commercialization which gives an AI-driven diagnostic tool that could be integrated into electronic health record systems by providing real-time support for clinicians. In conclusion, while performing this study, it can be clearly understood that the stacked hybrid model has limitations, which provides a foundation for exploring other hybrid predictive models in asthma diagnosis and highlights the potential for creating robust, interpretable and impactful diagnostic tools in the future.

References

Alam, F.B., Podder, P. and Mondal, M.R.H., (2023). RVCNet: 'A hybrid deep neural network framework for the diagnosis of lung diseases'. *Plos one,* 18(12), p.e0293125.

Bharati, S., Podder, P. and Mondal, M.R.H., (2020). 'Hybrid deep learning for detecting lung diseases from X-ray images. Informatics in Medicine Unlocked', 20, p.100391.

Chen, T. and Guestrin, C., (2016), August. 'Xgboost: A scalable tree boosting system'. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Devarajan, J.P., Sreedharan, V.R. and Narayanamurthy, G., (2021). 'Decision making in health care diagnosis: Evidence from Parkinson's disease via hybrid machine learning'. IEEE *Transactions on Engineering Management,* 70(8), pp.2719-2731.

Ginsburg, G.S. and Phillips, K.A., 2018. 'Precision medicine: from science to value'. *Health affairs,* 37(5), pp.694-701.

Goodfellow, I., (2016). 'Deep learning'.

Huang, X., (2024), 'Establishment of a predictive model on asthma diagnosis using environmental and allergy factors'. *Theoretical and Natural Science* 50(1):68-77. DOI:10.54254/2753-8818/50/2024AU0143.

Johnson, J., Abraham, T., Sandhu, M., Jhaveri, D., Hostoffer, R. and Sher, T., (2019). 'Differential diagnosis of asthma'. *Allergy and Asthma,* p.383.

Kavanagh, J., Jackson, D.J. and Kent, B.D., (2019). 'Over-and under-diagnosis in asthma'. *Breathe,* 15(1), pp.e20-e27.

King, G. and Zeng, L., (2001). 'Logistic regression in rare events data'. *Political analysis,* 9(2), pp.137-163.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. 'Deep learning'. *nature,* 521(7553), pp.436-444.

Lee, Z.J., Yang, M.R. and Hwang, B.J., (2024). 'A Sustainable Approach to Asthma Diagnosis: Classification with Data Augmentation, Feature Selection, and Boosting Algorithm'. *Diagnostics,* 14(7), p.723.

Lisspers, K., Ställberg, B., Larsson, K., Janson, C., Müller, M., Łuczko, M., Bjerregaard, B.K., Bacher, G., Holzhauer, B., Goyal, P. and Johansson, G., (2021). 'Developing a short-term prediction model for asthma exacerbations from Swedish primary care patients' data using machine learning-Based on the ARCTIC study'. *Respiratory Medicine*, 185, p.106483.

Loymans, R.J., Debray, T.P., Honkoop, P.J., Termeer, E.H., Snoeck-Stroband, J.B., Schermer, T.R., Assendelft, W.J., Timp, M., Chung, K.F., Sousa, A.R. and Sont, J.K., (2018). 'Exacerbations in adults with asthma: a systematic review and external validation of prediction models'. *The Journal of Allergy and Clinical Immunology: In Practice*, 6(6), pp.1942-1952.

Shahbazi, Z., Hazra, D., Park, S. and Byun, Y.C., (2020). 'Toward improving the prediction accuracy of product recommendation system using extreme gradient boosting and encoding approaches'. *Symmetry*, 12(9), p.1566.

Shwartz-Ziv, R. and Armon, A., (2022). 'Tabular data: Deep learning is not all you need'. *Information Fusion*, 81, pp.84-90.

Spathis, D. and Vlamos, P., (2019). 'Diagnosing asthma and chronic obstructive pulmonary disease with machine learning'. *Health informatics journal*, 25(3), pp.811-827.

Tomita, K., Nagao, R., Touge, H., Ikeuchi, T., Sano, H., Yamasaki, A. and Tohda, Y., (2019). 'Deep learning facilitates the diagnosis of adult asthma'. *Allergology International*, 68(4), pp.456-461.

Wolpert, D.H., (1992). 'Stacked generalization'. *Neural networks*, 5(2), pp.241-259.