Scientific Reports

https://doi.org/10.1038/s41598-026-62302-y

Article in Press

Data-efficient machine learning approach for predicting asthma attack risk

Received: 5 March 2026

Accepted: 10 July 2026

Published online: 15 July 2026

Cite this article as: Jayamini W.K.D., Mirza F., Naeem M.A. *et al.* Data-efficient machine learning approach for predicting asthma attack risk. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-62302-y

Widana Kankanamge Darsha Jayamini, Farhaan Mirza, M. Asif Naeem, Andrew Tomlin, Holly Tibble, Kebede Abera Beyene & Amy Hai Yan Chan

Data-Efficient Machine Learning Approach for Predicting Asthma Attack Risk

Widana Kankanamge Darsha Jayamini[a,b*], Farhaan Mirza[a], M. Asif Naeem[c], Andrew Tomlin[d], Holly Tibble[e], Kebede Abera Beyene[d], Amy Hai Yan Chan[d]

[a]*School of Engineering, Computer and Mathematical Sciences,*
Auckland University of Technology,
Auckland,
New Zealand.
[b]*Department of Software Engineering,*
Faculty of Computing and Technology,
University of Kelaniya,
Kelaniya,
Sri Lanka.
[c]*Department of Computer Science National University of Computer & Emerging Sciences,*
National University of Computer & Emerging Sciences,
Islamabad,
Pakistan.
[d]*School of Pharmacy, Faculty of Medical and Health Sciences,*
The University of Auckland,
Auckland,
New Zealand.
[e]*Usher Institute,*
University of Edinburgh,
Edinburgh,
United Kingdom.

Widana Kankanamge Darsha Jayamini
darsha.jayamini@autuni.ac.nz
0000-0001-6472-4815
Farhaan Mirza
farhaan.mirza@aut.ac.nz
0000-0001-7684-9973
M. Asif Naeem
asif.naeem@nu.edu.pk
0000-0001-6785-7875
Andrew Tomlin
andrew.tomlin@auckland.ac.nz
0009-0004-4755-9845

Holly Tibble
holly.tibble@ed.ac.uk
0000-0001-7169-4087
Kebede Abera Beyene
k.beyene@auckland.ac.nz
0000-0002-1057-2593
Amy Hai Yan Chan
a.chan@auckland.ac.nz
0000-0002-1291-3902
***Corresponding author**

Widana Kankanamge Darsha Jayamini
Level 8, WZ Building,
Engineering, Computer and Mathematical Sciences School,
Auckland University of Technology,
6 St Paul Street,
Auckland CBD,
Auckland 1010,
New Zealand.
darsha.jayamini@autuni.ac.nz

Abstract

Asthma caused around 436,000 deaths globally in 2021. Predicting asthma attacks can save lives, reduce costs, and strengthen healthcare systems. However, risk prediction requires extensive data, which is often unavailable. This study aims to develop an approach that minimizes data requirements, reducing complexity and enabling broader applicability in data-limited settings. We employed the CRISP-DM methodology, including combinations of feature selection strategies, machine learning algorithms, and data imbalance handling techniques. Altogether, 120 models were constructed. Model performance was evaluated based on accuracy and efficiency. The top models were externally validated. The key finding is that the LR and XGB models, when combined with an under-sampling technique, perform better, given that the feature selection is conducted based on the feature importance generated by the XGB-RUS model. Even a few features, including asthma attacks in the past year and

SABA_ICS ratio, can achieve a reasonable level of performance. Through rigorous experimentation, we achieved high accuracy in multiple scenarios. We have presented 6 feature sets. Among those, FS1 achieved the best performance, while FS6 yielded a reasonable performance with only two features. This paper presents a comprehensive analysis of the data requirements for predicting the risk of asthma attacks. This will enhance the acceptability of the prediction models using simplified feature sets in clinical settings.

Keywords: asthma attack, risk prediction, feature optimization, data imbalance

1 Introduction

Asthma is a non-communicable, chronic respiratory disease that affects people worldwide. Globally, 37.86 million incident cases and 260.48 million prevalent cases of asthma were reported in 2021 [1]. Asthma attack risk prediction has become a significant area of research in recent years as researchers seek to leverage machine learning (ML) and advanced data integration strategies to improve the expected outcome. These studies explored the use of heterogeneous datasets, including biological, hospital, weather, socio-demographic and radiology image data, to gain a higher accuracy in the asthma attack risk prediction [2-4]. This presents a significant demand for data collection, and it may not be feasible to find cases where it would be realistic to obtain such a diverse range of data to perform predictive analysis for escalating asthma risk. A previous study [5] found that there were a few important predictors among the vast array of hospital and pharmaceutical data features. We believe that predicting asthma attacks is a life-saving endeavor, and it is crucial for both patients and healthcare professionals. However, setting up a prediction model that is both accurate and validated is a significant challenge, requiring vast amounts of data.

Hence, we pose our research question: Can we simplify the method for predicting asthma attack risk? This simplification is approached from two different angles. One approach is to experiment with simplifying the feature set used to develop the risk prediction model. The other approach is to use

standard ML algorithms rather than complex deep learning models, which also simplifies the development of the model. Constructing deep learning models often demands large datasets, and the "black-box" nature limits the transparency and the interpretability, which is critical in clinical settings [6].

There is an increasing availability of high-dimensional data across various domains, including bioinformatics, text mining and finance. This has created a growing need for effective feature selection or feature reduction techniques to identify the most important features, thereby achieving the best accuracy for a given task. Feature selection reduces dimensionality and helps to identify the most informative attributes, thereby enhancing predictive performance while reducing the training and prediction time [7,8]. Also, high dimensionality not only increases computational cost but also contributes to issues such as model overfitting, data sparsity and the curse of dimensionality [9-11]. Feature selection is the strategy of identifying the most informative subset of original features, while feature reduction creates new features through mathematical transformations and projections [8,11].

Feature selection methods are traditionally divided into filter, wrapper and embedded approaches. Filter methods assess individual variables using statistical measures such as correlation or mutual information, without the involvement of any predictive model, making them computationally efficient, scalable and model-agnostic, despite the limited ability to detect feature interactions [12]. However, filter methods traditionally overlook interactions between features [12]. In contrast, wrapper methods assess subsets from the original feature set by training and evaluating learning models iteratively, typically delivering higher predictive performance because they tailor the selection to the classifier's behavior. A study on EEG-based emotion recognition across two datasets found that applying the wrapper method, the genetic algorithm, generally improved the model performance [13]. However, they can become computationally intensive and are prone to overfitting [14]. Embedded methods integrate feature selection into the model training process, such as Least Absolute Shrinkage and

Selection Operator (LASSO) regression (L1 regularization) or tree-based feature importance, while balancing efficiency and accuracy by integrating feature pruning into the learning algorithm itself [13,15]. This study addresses limitations of past research and makes the following contributions.

- Implementing various feature selection strategies to identify the most important predictors while evaluating the model performance under each strategy.
- Employing a range of ML techniques to develop prediction models under each feature selection strategy to determine the model with the best performance.
- Handling data imbalance in the data sets using several Data Imbalance Handling Techniques (DIHTs) to compare which technique performs best.

The remainder of this article is structured as follows. First, the *Related Work* section presents the literature published on asthma attack risk prediction. Next, the *Methods* section outlines the methodology used in this study, including data collection, feature optimization, data preprocessing and model development steps. The *Results* section presents the models' performance in terms of accuracy and efficiency. The *Discussion* section then interprets the findings in relation to the existing literature and highlights their implications. Finally, the *Conclusion* section summarizes the main contributions of the study and suggests future research directions.

2 Related Work

2.1 Asthma Attack Risk Predictors

In predicting asthma attack risk, past studies have utilized a wide range of features from various domains [2]. The most frequently used data sources are hospital, clinical and pharmaceutical data [5,16,17]. Asthma-related hospitalizations or emergency visits have been considered as the occurrence of asthma attacks [16,18]. Further, asthma attacks can be identified based on the prescription or usage of medications [16,18,19]. Comorbidities have been accounted for using the Charlson Comorbidity Index (CCI). The history of previous asthma attacks has also been considered, as it is a key predictor

of future attacks [5,16,20,21]. In addition, smoking status and Body Mass Index (BMI) have been included [16,20]. Demographic data (e.g., age, gender and ethnicity) were included in the feature sets, as variations in these factors have been shown to influence the impact of asthma among patients [5,18,19,21-23].

The risk prediction can be made at either the short-term individual level or the long-term population level. In addition to the above, clinical factors such as asthma symptoms, peak expiratory flow rate, and inhaler use have been included in previous studies, which are generally collected on a daily basis through telemonitoring systems and smart devices [16,24-26] for short-term, individual-level risk prediction. On the other hand, environmental and meteorological data have been integrated into models for predicting asthma attack risk, such as temperature, wind speed, air quality index, air pressure, pollen count, particle matter ($PM_{2.5}$) and different types of gases (CO, NO_2, SO_2) [20,27-31] for long-term population-level risk prediction.

Past studies have identified the most important predictors. Several studies found that previous asthma attacks, asthma symptoms, medication usage and daily peak flow readings are important asthma attack risk predictors [16-20,25,27]. Additionally, the literature indicates that comorbidities and hospital stays are also significant predictors [16,18,20]. Moreover, it has been found that age, gender, race and smoking status are important in predicting impending asthma attacks [17,23,27,30,31]. Furthermore, prior studies have demonstrated the importance of weather factors - including temperature, wind speed, relative humidity and air quality - in predicting future asthma attacks [27,30].

2.2 Identifying Important Predictors

Previous studies have employed various techniques to conduct feature selection in predicting asthma attack risk. A study has applied LASSO regularization, which imposes a penalty on various features, thereby avoiding overfitting [27]. LASSO has the limitation that it tends to select only one variable from a group of highly correlated variables, while ignoring the others, which may lead to instability in

feature selection [32]. A study identified the most important features using the feature importance scores generated from the Random Forest (RF) algorithm [22]. This method provides a ranking of the features that contribute most to the prediction [33]. The ensemble nature of the algorithm reduces its sensitivity to noise, thereby generating robust outcomes. However, this method can be biased based on certain properties of the features, like feature correlation. If two features are highly correlated, their importance could be shared, or one feature may dominate, making it difficult to determine which feature is truly important [34]. Another study employed the Recursive Feature Elimination (RFE) method [35]. After the model is trained on the full feature set, it ranks the features by their importance and subsequently removes the least important features. It iteratively removes less informative features while validating the performance to identify the optimal subset of features [36]. Another study employed feature elimination through manual observation, followed by RFE with RF, and then generated feature importance using the Extreme Gradient Boosting (XGB) algorithm [5]. As the RFE technique utilizes the actual model intended for prediction, the selected features are more likely to perform well with that model. This process continues until the key features are identified and does not affect predictive performance. However, it can be computationally costly due to iterative training. But this is only until the most important features are identified and will not affect the prediction performance.

2.3 Handling Data Imbalance

Datasets for predicting asthma attack risk are often highly imbalanced because attacks occur relatively rarely compared to the total number of instances - i.e., there are more days without attacks when patients are well, rather than days with attacks. Most studies did not address the target class imbalance, although a few accounted for it. A study that detected asthma exacerbations using daily home monitoring data employed three different techniques: Random Under-sampling (RUS), Random Over-sampling and Synthetic Minority Oversampling Technique (SMOTE) to handle the data imbalance

with a ratio of about 1:1265 (exacerbation:no exacerbation) [35]. The RUS technique performed best. Another study, which had a data imbalance of 1:2631 (exacerbation:no exacerbation), has shown that sequential RUS and weighting achieved the best performance [16]. A study reported a similar finding, identifying pediatric asthma patients at risk of hospital revisits, with a one-time to multiple-time visits ratio of about 1:4 [23]. SMOTE has been used in a study for asthma risk prediction, having a proportion of nearly 1:8 (present:absent) in the target class [27]. A key limitation of this study is the application of SMOTE on the full dataset, which leads to data leakage. Therefore, any of these sampling techniques should only be applied to the training data.

2.4 ML Algorithms

Recent studies have widely employed ML approaches in this scope. Common algorithms include Logistic Regression (LR) [5,18-20,27,30,37], RF [5,18-20,27,30,37] and XGB [5,18-20,27,30]. Across many studies, XGB achieved the best performance. XGB is an ensemble learning algorithm that sequentially constructs decision trees, where each new tree is developed to correct the residual errors in the preceding trees, thereby building a strong predictive model [38]. In some studies, LR performed slightly better. LR showcases the importance of individual predictors through the estimated coefficients, providing interpretability. Despite not being the top performer, RF was included in numerous studies. It is an ensemble algorithm that combines multiple decision trees to enhance predictive accuracy [39]. It has several advantages, including robustness to outliers and missing values, as well as the ability to handle a larger number of features [40]. Using bootstrap samples and random feature subsets reduces variance and improves generalization over a single decision tree. A few studies applied Decision Trees [27,30], Support Vector Machines (SVM) [27,29], LightGBM [37], Naïve Bayes [41] and Neural Networks [28,29].

The literature shows that several types of predictors can be used for predicting asthma attack risk. However, the addition of more features complicates model development. Despite several studies

employing feature selection methods, they have not compared multiple methods to determine the optimal feature set, which would simplify model development, making it generalizable and user-friendly. Hence, we propose systematically eliminating features while determining the most important ones using different strategies. This enhances the simplicity and generalizability of the asthma attack risk prediction models.

3 Methodology

The CRISP-DM methodology was employed in this study, as illustrated in Figure 1 and discussed in the following subsections.

3.1 Data Collection

Hospital and pharmaceutical datasets were obtained from national datasets of the New Zealand (NZ) Ministry of Health. These datasets comprise publicly funded hospital admissions in NZ, including acute admissions through emergency departments, as well as data on the national dispensing of subsidized medications. The ethics approval for data collection was obtained from the Auckland Health Research Ethics Committee (Reference AH23069).

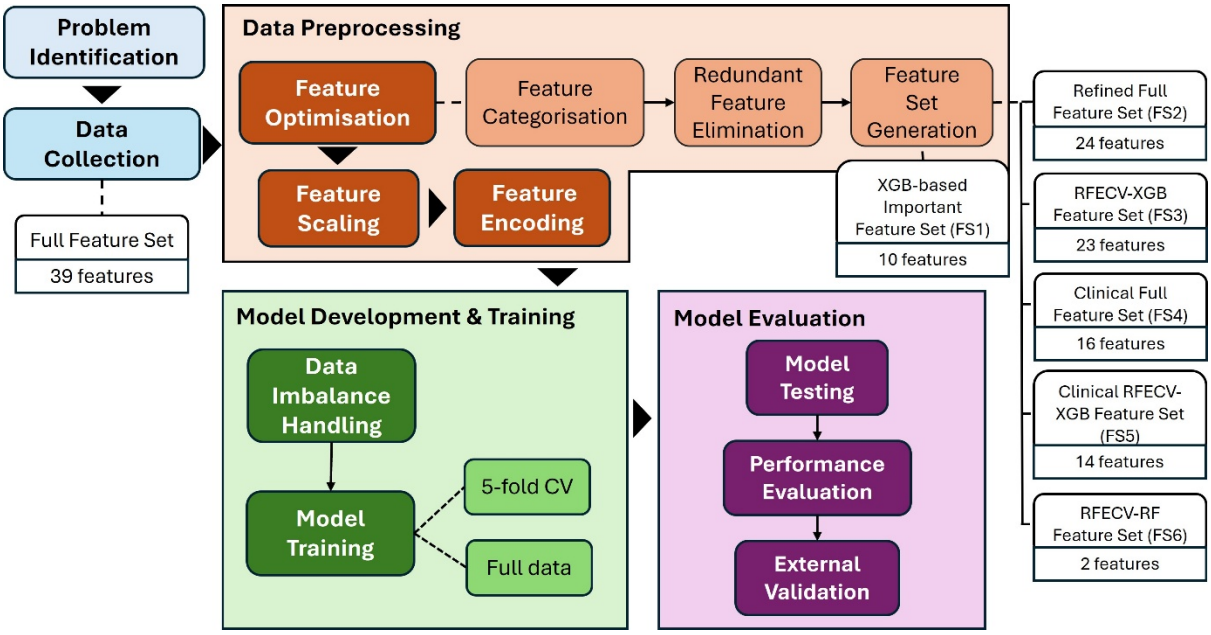

Figure 1 Overview of the CRISP-DM methodology followed in the study.

The data used for analysis spanned from January 1, 2008, to December 31, 2017. A feature named patient-quarter was derived to predict impending asthma attacks in the next 3 months. This represents a 3-month period for each patient. As some yearly-based features were included in the feature set, we extracted data for the fifth patient-quarter so that it could include data for the previous year. There were 355,113 patient records with 39 features.

The target variable represents the presence or absence of an asthma attack within a 3-month period. Asthma hospital admission and/or oral corticosteroid prescription were used to define an asthma attack [42,43]. For external validation, similar data were obtained for another patient-quarter (the ninth patient-quarter), which included 1,093,645 instances. It is worth noting that during the model training for external validation, the dataset was expanded to a total of 1,142,322 instances.

3.2 Data Preprocessing

Before developing the prediction models, several preprocessing steps were followed to refine the feature set. The following subsections will discuss each of those steps in detail.

3.3 Feature Optimization

3.3.1 Feature Categorization and Redundant Feature Elimination

Table 1 presents the newly derived features *HistoryOfAllergies* and *CardiovascularDiseases*, by categorizing the same group of medical conditions together. This approach was used to reduce the high number of categories. As all the original features are binary, the derived features were assigned 1 if one of the corresponding original features is 1; otherwise, 0. Furthermore, there were some features that appeared to exhibit higher collinearity among them. Features were grouped by category, and a single representative from each was retained to avoid redundancy (see Table 2).

Previous asthma attacks were represented by the number of asthma attacks in the past 12 months, as indicated by the feature *P12MNoOfAsthAttacks*. Similarly, to represent the asthma medications, we chose *SABA_ICS_Ratio*. It is the ratio between the Inhaled Corticosteroids (ICS) and Short-Acting Beta-Agonist (SABA) inhaler usage, as a parameter to reflect both SABA and ICS inhaler use. In addition, *CharlsonComorbidityScore* was selected by excluding *DementiaAlzheimers* and *RheumatologicalDisease,* as these are already included in the score calculation.

3.3.2 Feature Set Generation

The study included six feature sets. The workflow of feature set generation is presented in Figure 2. The first feature set (FS1) was extracted from the results of a previous study [5], which highlights the most important features for predicting asthma attack risk based on the optimal prediction model, XGB-RUS. In this study, the top 10 features from that feature importance list were considered as the first feature set. Secondly, the refined full feature set (FS2) was considered to compare with the other strategies. Next, the Recursive Feature Elimination with Cross-Validation (RFECV) method was used to determine the most important features based on the two classifiers, RF and XGB, separately on FS2. This generated the feature sets FS3 and FS6, respectively. Additionally, with the clinical expertise knowledge, a full clinical feature set (FS4) was defined, which were considered the most important

features in the original feature set for predicting impending asthma attacks. Similarly, RFECV was applied with XGB and RF on FS4 to extract the feature sets FS5 and FS6, respectively.

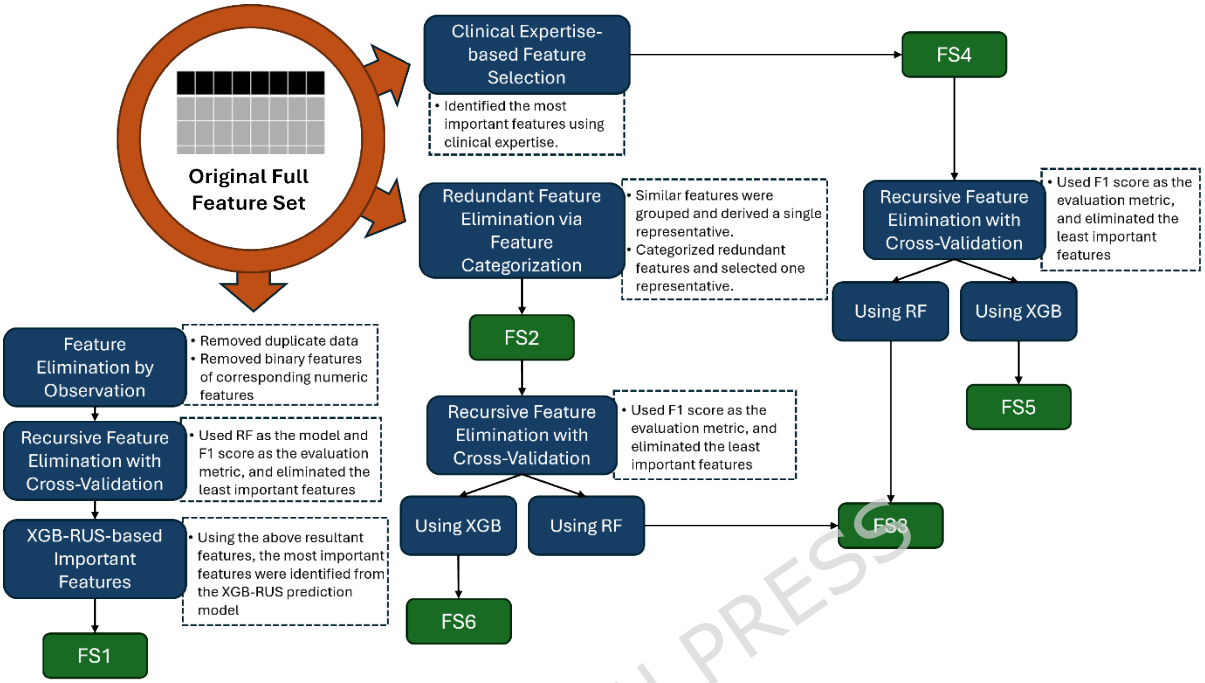

Figure 2 Feature sets generation workflow.

Table 1 New features derived by grouping existing features.

Original features	**Derived Feature**
Rhinitis	HistoryOfAllergies
Nasal Polyps	
Atopic Dermatitis	
Anaphylaxis	
Pulmonary Eosinophilia	
Cardiovascular Cerebrovascular Disease	CardiovascularDiseases
Ischaemic Heart Disease	
Heart Failure	

Table 3 displays the division of the feature sets according to different strategies. The first column shows the top 10 features extracted from the previously developed XGB model (FS1). It consists of information on past asthma attacks, exposure to winter, SABA and ICS medication usage and some demographic factors, such as gender, ethnic group and region (District Health Board (DHB)). The FS2 is the full

feature set generated after refining the original feature set (24 features). The FS3 has 23 features, excluding *Psoriasis*.

Table 2 Features selected from different feature categories to exclude redundant features.

Category	Features	Selected Feature
Previous asthma attacks	P12MNoOfAsthAttacks	P12MNoOfAsthAttacks
	P6MNoOfAsthAttacks	
	P3MNoOfAsthAttacks	
Asthma medication	NoOfICSInhalers	SABA-ICS ratio
	SABA_ICS_Ratio	
	NoOfSABAInhalers	
	NoOfAsthmaControllerMeds	
	Nebulised SABA	
	Asthma severity step	
Charlson Comorbidity Index	CharlsonComorbidityScore	CharlsonComorbidity Score
	DementiaAlzheimers	
	RheumatologicalDisease	

Figure A1 in Appendix A shows the selection of the optimal number of features using this strategy. Notably, the plot indicates that the optimal number of features is 48, as it has counted the one-hot encoded features separately. After analyzing, the distinct feature count was 23. The FS4 has 16 features, excluding *Psoriasis, DHB, Obesity, Lower Respiratory Tract Infection (LRTI), Anxiety Depression, Diabetes_NoMetformin, BetaBlockers*. FS5 excluded only the two features *NoOfHospitalizations* and *NumberOfMetforminRx,* from FS4 (see Figure A2). Significantly, the FS6 comprises only two features: *SABA_ICS_Ratio* and *P12MNoOfAsthAttacks*. It is worth noting that the RFECV-RF on FS2 and FS4 yielded the same feature set (FS6) as the most important features, as illustrated in Figures A3 and A4.

3.4 Feature Scaling and Encoding

The feature set has several numeric features with varying value ranges. Therefore, to bring them onto a similar scale, standardization was applied before feeding them into the prediction models, ensuring

that no single feature would disproportionately influence the models. The previous study [5] found that standardization performed better than normalization. There were a few categorical features, including ethnic group and DHB. These features were encoded using the one-hot encoding technique, a method commonly employed for categorical feature encoding. Additionally, the *Age* was grouped into separate bins as 12-14, 15-44, 45-64 and >=65 years, to best represent the age groups rather than individual ages. These categories were chosen in line with previous literature [44-47].

3.5 Model Development and Training

3.5.1 Handling Data Imbalance

The dataset was highly imbalanced in terms of the target class, with a 1:12 ratio where the asthma attacks positive-class was the minority. To address this data imbalance, four different DIHTs were employed. They included two down-sampling techniques and two oversampling techniques. The down-sampling techniques included RUS and Edited Nearest Neighbor (ENN), while SMOTE and KMeans-SMOTE were used as oversampling techniques. For comparison, models were developed without applying any DIHTs, which are denoted as No-sample.

3.5.2 Model Training

Asthma attack risk prediction models were generated using three classifiers, LR, RF and XGB, as they performed better in similar studies. In each case, the dataset was split into a training set (70%) and a test set (30%). Initially, each model was built and tested with 5-fold cross-validation (CV) using the training set. For comparison, Multi-Layer Perceptron (MLP) neural network models (128-64-32) were also developed on the same dataset. Later, models were rebuilt on the full training dataset. Full specification of the hyperparameters of all the models is given in Table C1 in Appendix C. The models were implemented using the Python programming language and executed on a system equipped with a 3.50 GHz processor and 128 GB of RAM. The python code used to develop these models are available at https://github.com/DarshaWK/asthma-prediction-feature-optimisation.

Table 3 Details of the feature sets generated based on feature selection strategies.

ased eature Set	FS2-Refined Full Feature Set	FS3-RFECV-XGB Feature Set	FS4-Clinical Full Feature Set	FS5-RFECV-XGB Clinical Feature Set	FS6-RF Set
hAttacks[a]	Q5_WeeksDuringWinter	Q5_WeeksDuringWinter	Q5_WeeksDuringWinter	Q5_WeeksDuringWinter	SABA_
uringWinter	Age	Age	Age	Age	P12MN
alers	Gender	Gender	Gender	Gender	
_Ratio[b]	EthnicGroup	EthnicGroup	EthnicGroup	EthnicGroup	
Attacks[c]	DeprivationQuintile	DeprivationQuintile	DeprivationQuintile	DeprivationQuintile	
Attacks[d]	DHB[e]	DHB[e]	SmokingStatus	SmokingStatus	
Inhalaers	Psoriasis	Obesity	CharlsonComorbidityScore	CharlsonComorbidityScore	
	Obesity	LRTI[f]	NumberOfMetforminRx	NoOfOPEDVisits	
	LRTI[f]	AnxietyDepression	NoOfHospitalizations	Paracetamol	
	AnxietyDepression	Diabetes	NoOfOPEDVisits	NSAIDs[g]	
	Diabetes	SmokingStatus	Paracetamol	SABA_ICS_Ratio[b]	
	SmokingStatus	CharlsonComorbidityScore	NSAIDs[g]	P12MNoOfAsthAttacks[a]	
	CharlsonComorbidityScore	NumberOfMetforminRx	SABA_ICS_Ratio[b]	HistoryOfAllergies	
	NumberOfMetforminRx	Diabetes_NoMetformin	P12MNoOfAsthAttacks[a]	CardiovascularDiseases	
	Diabetes_NoMetformin	NoOfHospitalizations	HistoryOfAllergies		
	NoOfHospitalizations	NoOfOPEDVisits	CardiovascularDiseases		
	NoOfOPEDVisits	BetaBlockers			
	BetaBlockers	Paracetamol			
	Paracetamol	NSAIDs[g]			
	NSAIDs[g]	SABA_ICS_Ratio[b]			
	SABA_ICS_Ratio[b]	P12MNoOfAsthAttacks[a]			
	P12MNoOfAsthAttacks[a]	HistoryOfAllergies			
	HistoryOfAllergies	CardiovascularDiseases			
	CardiovascularDiseases				

[a]Number of asthma attacks in the previous 12 months, [b]Ratio between short-acting beta$_2$-agonist and inhaled corticosteroids inhaler count,[c]Number of asthma attacks in the previous 6 months, [d]Number of asthma attacks in the previous 3 months, [e]District health board, [f]Lower respiratory tract infection, [g]Non-steroidal anti-inflammatory drugs

3.6 Model Evaluation

3.6.1 Model Testing and Performance Evaluation

Initially, each model was analyzed using 5-fold CV, then trained on the full training set and tested on the test set. In each of these steps, model performance was evaluated using the Area Under the Receiver Operating Characteristic (AUROC) and F1-score. AUROC measures how well a model distinguishes between the classes, while the F1-score produces a balanced value between precision and recall. In addition, we calculated Area Under the Precision-Recall Curve (AUPRC) and the calibration metric,

brier score. AUPRC measures how well a model identifies the positive class by balancing precision and recall, particularly when the outcome is rare or imbalanced. Brier score measures the accuracy of predicted probabilities by calculating the average squared difference between predicted probabilities and actual outcomes, where lower values indicate better performance. Additionally, identifying the top-performing models, prediction time (in seconds) and model size (in megabytes (MB)) were compared as efficiency metrics. These efficiency factors are critical in clinical settings for making decisions faster without minor delays.

3.6.2 External Validation

To validate the performance of the top models, they were tested on an additional dataset, known as the validation data. Here, the model performance was evaluated based on the prediction accuracy. The best model combinations selected from the above were trained on the training data and then tested on the validation data. The validation data had around a 1:22 imbalance in the target class. To quantify potential dataset shift between the test and external validation cohorts, the Population Stability Index (PSI) was calculated for each predictor variable as presented in Table B1 in Appendix B.

4 Results

4.1 Comparison of Models Based on Accuracy

Mainly there are 90 different models generated by combining various ML algorithms, feature selection strategies and DIHTs. The performance of all these models is graphically represented in Figure 3. The columns of the facet grid plot show the specific feature sets used to develop the models, while each row shows the evaluation metric. The x-axis of each subplot represents the ML algorithm used to build the prediction model. Different colored bars indicate the corresponding DIHT. The first row shows the mean AUROC values for each model in 5-fold CV, with the error bars representing the 95% confidence interval. Similarly, the second row shows the mean F1 scores.

Table D1 in Appendix D presents the models' performance on the 5-fold CV and the test set separately. Based on the results, it is challenging to select the best model, as some models that achieve higher AUROC scores have lower F1 scores. Overall, the best performance was achieved by a few of the LR and XGB classifiers built on the FS1. In comparison, MLP models were developed on the training set and tested on the test set. The performance of these models are given in Table D2.

Table 4 presents the top-performing models based on the evaluation metrics. The models LR-RUS, LR-ENN and LR-SMOTE achieved the highest AUROC of 0.75 and F1 score of 0.28 on the FS1 test set. Similar results were shown by the XGB-ENN model on FS1 (see Figure 3). However, with a marginal drop in the AUROC values to 0.74, the LR-KMeans-SMOTE model on FS1 achieved the highest F1 score of 0.30. Despite the overall best performance of LR models, they have wider confidence intervals during the training phase, especially for F1 scores of the LR-KMeans-SMOTE models. Additionally, the F1 score uncertainty gap during the training phase is generally higher for models developed with ENN and KMeans-SMOTE using FS6. KMeans-SMOTE technique produces the highest gap in the confidence intervals for RF (0.12-0.29) and XGB (0.12-0.30) with FS6 (see Table B1).

Among the data-balanced models, the lowest AUROCs of 0.55 and 0.52 were achieved on the test data by RF-ENN and XGB-ENN, respectively. For FS6, the lowest F1 score of 0.13 was observed for both models on the test set. However, in the 5-fold CV, the models show an AUROC of around 0.70 and an F1 score of 0.23. Overall, most models achieved an AUROC above 0.70. The best F1 score of 0.30 is achieved by the LR-KMeans-SMOTE models developed based on FS2 and FS3. However, the AUROCs of those two models are 0.73 each, slightly less than the best. Further, a very few MLP models developed on FS1, FS4 and FS5 without data sampling showed AUROCs between 0.72 and 0.74; however, their F1 scores were very low.

Overall, for most models built on the first five feature sets, RF models have slightly lower AUROC values. However, there are fluctuations in the F1 scores of these models for different ML algorithms and DIHTs. As shown in Figure 3, all the models without a data balancing technique (No-sample) have the lowest F1 scores, although the AUROC values remain higher. Comparing the performance of the models, especially the F1 scores, for different DIHTs, the LR algorithms show more robust behavior regardless of the DIHT applied.

The findings clearly show that under-sampled data improves the performance of the RF and XGB algorithms regardless of the feature set, except for ENN on FS6. Among the two data under-sampling techniques, RUS and ENN, ENN performs slightly better with RF and XGB algorithms except for FS6. The oversampling technique, KMeans-SMOTE, shows marginally higher F1 scores for the LR models on all the feature sets. The SMOTE technique has marginally lowered the AUROC values in the RF models, except for FS6.

4.2 Comparison of Models Based on Model Efficiency

In addition to the evaluation metrics, the efficiency of the models was considered to determine the best-performing models. Prediction time and model size were the two key metrics used for this analysis. The efficiency of all the models developed is represented in Figure E1 in Appendix E.

In comparison to all models, RF was the lowest performer with the highest prediction time. Overall, the RF-SMOTE models, ranked at the lowest efficiency, with a model size exceeding 540 MB and a prediction time exceeding 2.5 seconds. These values are not considered as any standards for a prediction model. These could change based on different settings. The sole idea was to identify a threshold to make a comparison between the other models developed.

Based on the ML algorithms, FS2-LR-ENN (prediction time=0.13 s, model size=90.00 MB), RF-SMOTE with FS1, FS2 and FS3 with average measures of prediction time=2.67 s and model

size=579.37 MB and XGB-ENN with FS3 and FS2 with average prediction time=0.14 s and model size=89.48 MB were the lowest performers. Overall, MLP models showed average model sizes and prediction times compared to other models. The lowest model sizes and prediction times were recorded by MLP-RUS models for FS4, FS5 and FS6.

Based on all the values, models with a prediction time of less than 1 second and a model size of less than 1 megabyte were chosen as the top-performing models. It should be noted that, although a few models with the No-sample DIHT were among the top-ranked models based on efficiency, they were eliminated due to lower F1 scores. The primary objective was to identify the best-performing models based on all performance measures.

Figure 4 displays the comparison of the 12 top-performing models based on the efficiency metrics. Among these, a significant observation is that models with KMeans-SMOTE were larger in size (~0.90 MB), while the FS6-XGB-SMOTE took a similar amount of space (0.93 MB). But it was one of the models with the lowest prediction time (0.01 s). Notably, all LR-RUS models in this subset yield the lowest model size (0.25 MB), independent of the feature set. However, there are slight variations in the prediction time for the LR-RUS models using different feature sets, with a marginally lower time for FS6 and a higher time for FS2. The shortest prediction time ($\leq$ 0.01 s) is reported by the models using FS6, despite their varying model sizes. Oversampling techniques have increased the model size. Within this group of top models, approximately two-thirds are based on the RUS technique, whereas none incorporates ENN.

DeLong's test revealed that LR-KMeans-SMOTE achieved a significantly lower AUROC than the other models (all $p < 0.001$). However, the absolute difference in AUROC was small (0.74 vs. 0.75), indicating that the practical significance of this finding is limited. No significant differences were observed among LR-RUS, LR-SMOTE, and XGB-ENN, suggesting comparable discriminative performance among these models. The test results are given in Table 5.

4.3 External Validation

Table 6 presents the performance of the models on the validation data. The results show that the top selected models yield AUROC values of 0.74, which is slightly lower than the test results. However, reductions in the F1 scores were observed. Compared with the smallest feature set (FS6), the two models with higher efficiency and better accuracy were included in this validation. However, their overall performance dropped.

Table 4 Top performing models for FS1 feature set.

ML Algorithm	DIHT	AUROC		F1 Score		AUPRC Test set	Brier Score Test set
		CV Mean (95% CI)	Test set	CV Mean (95% CI)	Test set		
LR	RUS	0.75 (0.74-0.75)	0.75	0.28 (0.28-0.29)	0.28	0.29	0.06
	ENN	0.75 (0.74-0.76)	0.75	0.28 (0.27-0.29)	0.28	0.29	0.06
	SMOTE	0.75 (0.74-0.75)	0.75	0.28 (0.27-0.29)	0.28	0.28	0.19
	KMeansSMOTE	0.74 (0.73-0.76)	0.74	0.29 (0.27-0.31)	0.30	0.27	0.09
XGB	ENN	0.75 (0.74-0.75)	0.75	0.28 (0.27-0.30)	0.28	0.29	0.06

Table 5 Pairwise comparison of the top models' AUROC values using the DeLong test.

Model 1	Model 2	p-value
LR-RUS	LR-ENN	< 0.001
LR-RUS	LR-SMOTE	0.705
LR-RUS	LR-KMeansSMOTE	< 0.001
LR-RUS	XGB-ENN	0.443
LR-ENN	LR-SMOTE	< 0.001
LR-ENN	LR-KMeansSMOTE	< 0.001
LR-ENN	XGB-ENN	0.795
LR-SMOTE	LR-KMeansSMOTE	< 0.001
LR-SMOTE	XGB-ENN	0.366
LR-KMeansSMOTE	XGB-ENN	< 0.001

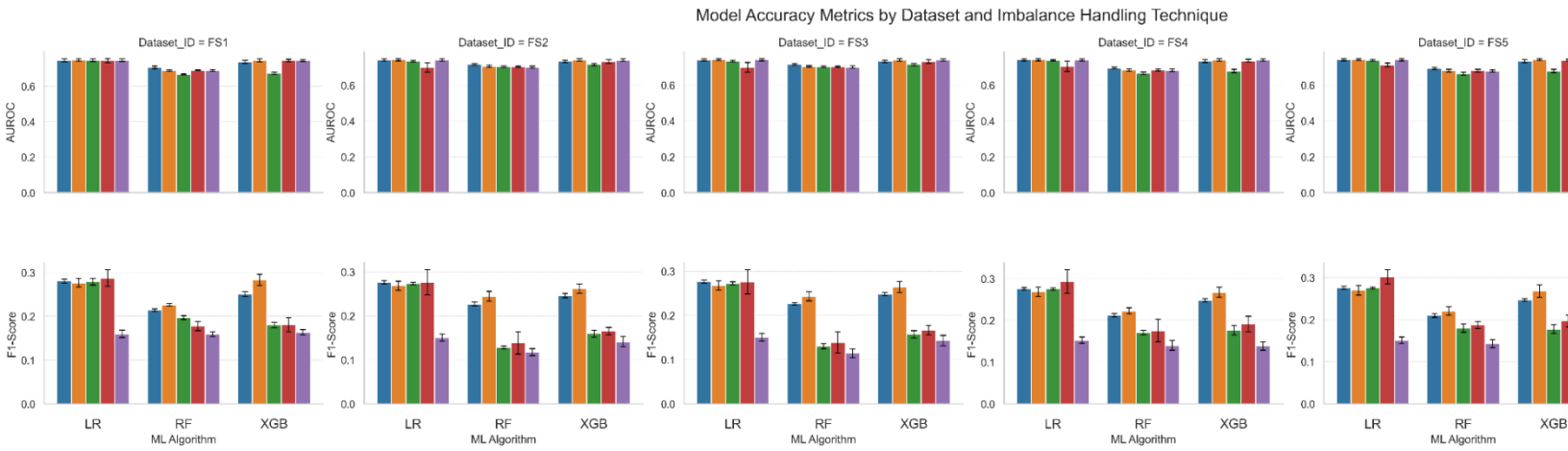

Figure 1 Performance of the ML algorithms for different feature sets and data imbalance handling techniques.

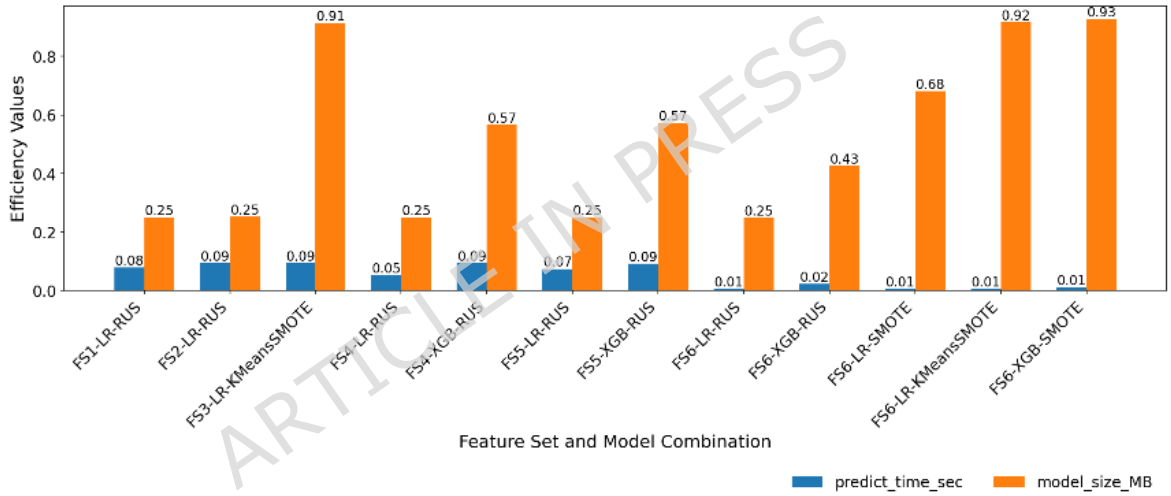

Figure 2 Comparison of the top-performing models with respect to both efficiency metrics, prediction time (s) and model size (MB).

Table 6 Performance of the top models on the validation data

Feature Set	ML Algorithm	DIHT	AUROC	F1 Score	AUPRC	Brier Score
FS1	LR	RUS	0.74	0.22	0.23	0.18
	LR	ENN	0.74	0.24	0.23	0.03
	LR	SMOTE	0.74	0.22	0.23	0.18
	XGB	ENN	0.74	0.24	0.24	0.03
	LR	KMeansSMOTE	0.71	0.27	0.22	0.06
FS6	LR	RUS	0.71	0.23	0.21	0.19

	XGB	RUS	0.71	0.19	0.20	0.19

5 Discussion

This is an important study that aimed to evaluate the performance of a range of ML algorithms and traditional statistical models in predicting impending asthma attacks using national health datasets with different feature optimization strategies. These strategies generated several feature sets by excluding some of the less important features in making predictions of asthma attack risk. The FS1 was defined based on the previous study [5] which developed several prediction models for predicting asthma attack risk. Based on the best-performing XGB model, the study identified the most important features for predicting asthma attack risk.

Several common features are included in the feature sets generated in this study. One of them is the number of past asthma attacks. This has been identified as a key predictor of future asthma attacks in several previous studies [2,48-50]. Demographic factors such as Gender and Ethnic Group were common among the first five feature sets. A study reveals the impact of asthma on different ethnic groups in NZ [51]. There is evidence that shows some ethnic groups have a greater influence from asthma, specifically, Māori, the indigenous population in NZ [51]. Another feature was the *SABA_ICS_Ratio*, which shows the ratio between preventer and reliver medication. This has been considered a predictor of asthma attack risk in previous studies [52,53]. However, the method we used to calculate this ratio differs from that employed in previous studies. In our study, we calculated the SABA:ICS ratio as the number of SABA inhalers divided by the number of ICS dispensed to a patient over the last year. It has been calculated as the average daily dose of SABA/ICS in studies [52,53]. However, this reflects key aspects of asthma control. The external validation results (Table 6) show a slight reduction in the models' performance. The PSI calculated (Table B1) supports this performance

reduction by indicating a significant and moderate shift in the key predictors *SABA_ICS_Ratio* and *P12MNoOfAsthAttacks,* respectively, across the test and validation cohorts.

RFECV-XGB on FS2 resulted in FS3, excluding only *Psoriasis*. Psoriasis is a long-term, immune-driven inflammatory condition that primarily affects the skin and/or joints [54]. The lower importance of this feature is evident in the values in Table E1 in Appendix E, which indicates that 99.98% of instances involve the absence of Psoriasis, regardless of whether asthma attacks occur or not. Therefore, Psoriasis has had no significant impact on differentiating the occurrence or non-occurrence of an asthma attack within the prediction models developed in the study. However, the literature shows that individuals diagnosed with psoriasis are more likely to be susceptible to asthma [55,56]. In this study, only 0.02% of patients had this health condition; therefore, the lack of data may have led the models to overlook it as an important predictor of asthma attacks.

Although previous asthma hospitalizations are a key predictor for determining future asthma attacks, in this cohort, we considered hospitalizations due to all causes; therefore, they might have become a less important factor in differentiating between the target class. This is evident from the descriptive statistics presented in Table E2. Although the mean number of hospitalizations was slightly higher among those who experienced asthma attacks (mean=0.53) compared to those who did not (mean=0.32), the median and first quartile (Q1) values were both zero for both groups, indicating that most individuals had no prior hospitalizations. In fact, 80.6% of attack cases and 80.7% of non-attack cases resulted in zero hospitalizations in the previous 12 months. This limited variability might have reduced the predictive value of this feature. Furthermore, regarding *NumberOfMetforminRx,* although the mean number of metformin prescriptions was slightly higher among individuals with asthma attacks (mean=0.34) compared to those without (mean=0.32), the distribution is highly skewed, with the median and all quartiles equal to zero in both groups since metformin use is uncommon amongst asthma patients (refer to Table E3). Additionally, 95.4% of attack cases and 95.7% of non-attack cases

had no metformin use. This suggests that the variation in this feature is limited and its contribution to distinguishing between the two groups is minimal.

Generally, the LR and XGB models performed marginally better than the RF and MLP models. The best AUROC was given by the LR classifier with both under-sampling and oversampling techniques on FS1, while the highest F1 scores were given by the KMeans-SMOTE with the LR classifier on FS1 although the AUROC was marginally lower than others. Pairwise comparison of AUROC values using DeLong's test indicated that LR-KMeansSMOTE performed significantly differently from the other models. However, the observed AUROC difference was small (0.74 versus 0.75), suggesting limited practical significance despite statistical significance. This finding highlights the importance of considering effect size in addition to statistical significance when evaluating predictive models. The comparable AUROC values observed for LR-RUS, LR-SMOTE, and XGB-ENN indicate that these approaches achieved similar discriminative performance. The FS1 feature set includes some redundant features, such as previous asthma attacks in different time windows and medication usage, such as SABA and ICS inhalers. Therefore, the multicollinearity between these features might have impacted the prediction models, especially on the LR classifier [57,58]. However, all LR models were developed using ridge regularization, which constrains coefficient magnitude and reduces overfitting.

Although the AUROC values for the No-sample models were sufficient, all these models exhibited the lowest F1 scores (see Figure 3). This could be because AUROC summarizes the model’s ability to discriminate between the positive and negative classes at all possible thresholds. But not the number of correctly classified instances under each class. In highly imbalanced datasets, specificity (true negative rate) can appear very high because the negative class is much larger. This can make the ROC curve appear to indicate that the model performs well overall, even if it does not accurately identify the minority (positive) class. Therefore, using ROC curves alone in these scenarios may lead to an overestimation of the model's actual performance [59]. The F1 score is the harmonic mean of precision

and recall, providing a more informative evaluation of the classifier's performance, especially for the minority class [59,60]. Therefore, it is essential to consider the F1 score in model evaluation, particularly when there is a significant data imbalance. Thus, in highly imbalanced datasets, DIH improves overall model performance.

On the other hand, the results showed that LR-RUS, LR-ENN, LR-SMOTE and XGB-ENN achieved the highest AUROCs and F1 scores. Further analysis on the performance using AUPRC and brier score values confirmed the better performance of the LR-RUS, LR-ENN and XGB-ENN models on FS1. A similar finding that LR-RUS performs better has been reported in another study [35]. In the RUS technique, data from the majority class is randomly removed. In contrast, ENN eliminates most samples that differ significantly from most of their k-nearest neighbors, thereby improving class separability and reducing noise in the data. This sharpens the decision boundary, guiding classifiers to prioritize learning from more informative data. A study proposed a combined approach of SMOTE-ENN on diabetes prediction and found that RF with this approach achieved the best performance compared to SVM [61]. However, overall, our findings showed that the RF models had marginally lower performance compared to the other two classifiers. LR has outperformed XGB in predicting asthma exacerbations in a study, although it has not considered the data imbalance [24], whereas the opposite is true in another study [20]. A similar observation was found in another work on developing asthma attack risk prediction models [5]. These findings suggest that LR and XGB classifiers generally perform better on highly imbalanced data. Despite the better performance of XGB, the XGB-ENN also showed the lowest AUROC for FS6. This could be due to the lack of features in discriminating between the two predictive classes, as too few features might not provide enough information to the models [62]. These algorithms attempt to construct multiple decision trees with various combinations, and a lack of features may reduce their predictive power. Fewer features will result in fewer possible splits

in the decision trees. This is further supported by the validation results, which show that limited features reduce model performance, especially with larger and highly imbalanced datasets.

Comparatively, the findings showed that the LR-RUS models have shorter prediction times for FS6 and longer for FS2. A parallel observation was there for the RF and XGB algorithms. Similarly, MLP models showed a reduction in prediction time and model size for FS4, FS5 and FS6. This could be mainly due to the difference in the number of features the algorithms need to consider during a prediction. Overall, RF-SMOTE was the least performer for the first five feature sets compared to the other prediction models developed in the study. Accordingly, Figure 5 proposes a generalized workflow for obtaining better results in asthma attack risk prediction. Considering both the factors, accuracy and efficiency, and feature selection strategies experimented within the study, using XGB-RUS to determine the most important features is identified as a better option. While it is critical to handle data imbalance, down-sampling data generates marginally better results. Overall, as discussed above, LR and XGB generated better results. Hence, the study findings helped in generating the proposed workflow which would be a hybrid model that combines two ML algorithms to make a better prediction with improved accuracy and efficiency.

Several factors strengthen this study. Most researchers will typically stick to one feature selection method in their individual work. However, in this study, several feature selection strategies were employed simultaneously. Additionally, the study focused on addressing data imbalance. Various DIHTs were applied to the dataset to handle the data imbalance before the actual prediction models were constructed. Further, it utilized a few conventional ML algorithms to develop the prediction model. As a result, a total of 90 model combinations were to be compared. Moreover, the performance of the models was evaluated, not only based on accuracy but also on efficiency. Another key strength was the external validation of the models.

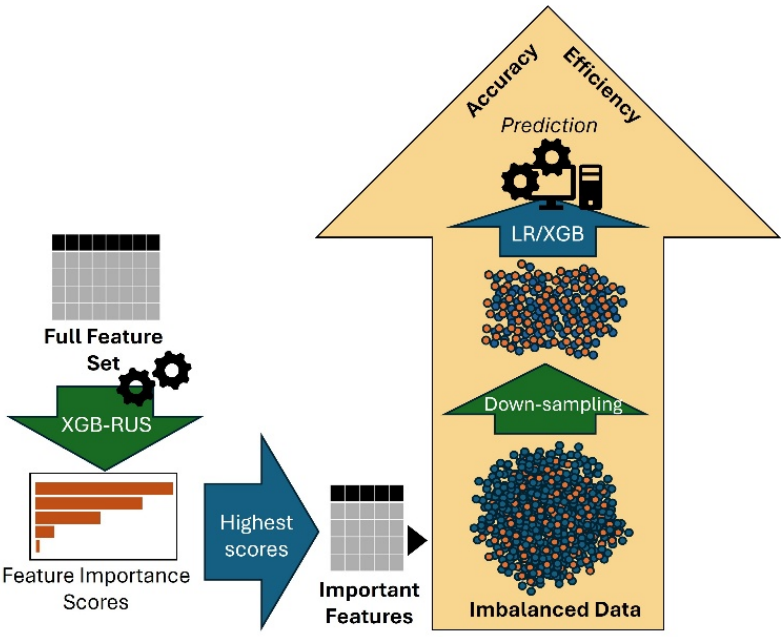

Figure 3 Generalized workflow for asthma attack risk prediction.

Apart from the significant strengths of the study, several limitations should be noted. The study only applied feature selection to a dataset that has hospital and pharmaceutical data. The findings might be different if the data parameters were broader to include weather, peak flow, inhalation data, etc. Including further data sources might help in identifying further key predictors for asthma attack risk prediction models. Also, there are several other feature selection strategies that could be applied to confirm the most important predictor identification. In the current study, it only applied RFECV and did not compare the other feature selection methods. Therefore, it is worth to experiment and compare the other strategies/techniques for feature selection in the future work.

6 Conclusion

This study explores a method to minimize the complexity of predicting asthma attack risk while determining the least amount of data required. This study contributes to asthma attack risk prediction by experimenting with variant feature selection methods together with different data resampling techniques and ML algorithms, creating 90 combinations for comparison. Further, 30 MLP models were developed to make a comparison of deep learning models. By paying attention not only to

accuracy but also to efficiency, the study offers a balanced perspective on model performance evaluation. The results show that the number of asthma attacks in the past year and medication use, especially the SABA_ICS ratio, are strong predictors of impending attacks. A hybrid model that combines XGB-RUS-based feature selection with a down-sampling technique (RUS or ENN) for LR and XGB ML algorithms performed slightly better than others. Also, the findings reveal that a very simple model using only two features with LR-RUS can also give promising results, but with a higher variation. However, this requires further exploration. Overall, the findings suggest that accurate predictions can be achieved even with fewer features, thereby enhancing the simplicity and generalizability of the models. Future research can be conducted to utilize multiple data sources with various feature selection methods to identify the most important features, thereby enhancing the performance and simplicity of models. Additionally, researchers could focus on further enhancing the external validity of the models, particularly by conducting a pilot study to apply these models in hospitals and communities that lack access to a wide range of data features.

List of Abbreviations

AUROC - Area Under the Receiver Operating Characteristic

AUPRC - Area Under the Precision Recall Curve

BMI - Body Mass Index

CCI - Charlson Comorbidity Index

CV - Cross-Validation

DHB - District Health Board

DIHT - Data Imbalance Handling Technique

ENN - Edited Nearest Neighbor

FS - Feature Set

ICS - Inhaled Corticosteroids

LASSO - Least Absolute Shrinkage and Selection Operator

LR - Logistic Regression

LRTI - Lower Respiratory Tract Infection

ML - Machine Learning

MLP – Multi-Layer Perceptron

NZ - New Zealand

PM - Particulate Matter

PSI - Population Stability Index

RF - Random Forest

RFECV - Recursive Feature Elimination with Cross-Validation

RFE - Recursive Feature Elimination

RUS - Random Under-Sampling

SABA - Short-Acting Beta-Agonist

SMOTE - Synthetic Minority Oversampling Technique

SVM - Support Vector Machines

XGB - Extreme Gradient Boosting

Declarations

Ethical Approval and Consent to Participate

Ethics approval for this study was obtained from the Auckland Human Research Ethics Committee (Reference AH23069) to access and analyze hospital data provided by the New Zealand Ministry of Health. The requirement for informed consent to participate was waived by the Auckland Human Research Ethics

Committee, as the study involved secondary analysis of routinely collected, de-identified hospital data and posed minimal risk to individuals. All procedures involving human data were conducted in accordance with the ethical standards of the relevant institutional and national research committees and with the principles of the Declaration of Helsinki.

Consent for publication

Not applicable

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available, as participants did not consent to the sharing of data beyond the research team, but are available from the author Amy Hai Yan Chan at a.chan@auckland.ac.nz on reasonable request.

Code Availability

Python codes used for the model development are at https://github.com/DarshaWK/asthma-prediction-feature-optimisation.

Competing Interests

The authors declare that they have no competing interests.

Funding

No funding was received for conducting this study.

Authors' Contribution

W.K.D.J., F.M., M.A.N., and A.H.Y.C. contributed to the conceptualization and methodology. A.H.Y.C., A.T., H.T., and K.A.B. performed the data curation. Formal analysis was conducted by A.T., and W.K.D.J. Investigation, visualization, software development, validation and writing the original manuscript were conducted by W.K.D.J. The study was supervised and administered by F.M., M.A.N., and A.H.Y.C.

Resources were provided by W.K.D.J., F.M., and A.H.Y.C. All authors participated in the review and editing of the manuscript and have read and approved the final version.

Acknowledgment

This work was supported by the Health Research Council of New Zealand [grant number HRC 21/867].

References

1 Mao, Z. *et al.* Global, regional, and national burden of asthma from 1990 to 2021: A systematic analysis of the global burden of disease study 2021. *Chinese Medical Journal Pulmonary and Critical Care Medicine* **3**, 50-59 (2025). https://doi.org:10.1016/j.pccm.2025.02.005

2 Jayamini, W. K. D., Mirza, F., Asif Naeem, M. & Chan, A. H. Y. Investigating Machine Learning Techniques for Predicting Risk of Asthma Exacerbations: A Systematic Review. *Journal of Medical Systems* **48**, 49 (2024). https://doi.org:10.1007/s10916-024-02061-3

3 Budiarto, A., Tsang, K. C., Wilson, A. M., Sheikh, A. & Shah, S. A. Machine learning–based asthma attack prediction models from routinely collected electronic health records: systematic scoping review. *Jmir Ai* **2**, e46717 (2023). https://doi.org:10.2196/46717

4 Alharbi, E. T., Nadeem, F. & Cherif, A. Predictive models for personalized asthma attacks based on patient's biosignals and environmental factors: a systematic review. *BMC Medical Informatics and Decision Making* **21**, 345 (2021). https://doi.org:10.1186/s12911-021-01704-6

5 Jayamini, W. K. D. *et al.* in *58th Hawaii International Conference on System Sciences* 3731-3741 (Hawaii, USA, 2025).

6 Xi, H. *et al.* Forecasting Hospitalization for Adult Asthma Patients in Emergency Departments Based on Multiple Environmental and Clinical Factors. *Journal of Asthma and Allergy*, 861-876 (2025). https://doi.org:10.2147/JAA.S512405

7 Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research* **3**, 1157-1182 (2003). https://doi.org:10.5555/944919.944968

8 Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **40**, 16-28 (2014). https://doi.org:10.1016/j.compeleceng.2013.11.024

9 Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-based systems* **86**, 33-45 (2015). https://doi.org:10.1016/j.knosys.2015.05.014

10 Taha, Z. Y., Abdullah, A. A. & Rashid, T. A. Optimizing feature selection with genetic algorithms: a review of methods and applications. *arXiv preprint arXiv:2409.14563* (2024). https://doi.org:10.48550/arXiv.2409.14563

11 Kraev, E., Koseoglu, B., Traverso, L. & Topiwalla, M. Shap-Select: Lightweight Feature Selection Using SHAP Values and Regression. *arXiv preprint arXiv:2410.06815* (2024). https://doi.org:10.48550/arXiv.2410.06815

12 Li, Y., Mansmann, U., Du, S. & Hornung, R. Benchmark study of feature selection strategies for multi-omics data. *BMC Bioinformatics* **23**, 412 (2022). https://doi.org:10.1186/s12859-022-04962-x

13 Abdumalikov, S., Kim, J. & Yoon, Y. Performance analysis and improvement of machine learning with various feature selection methods for EEG-based emotion classification. *Applied Sciences* **14**, 10511 (2024). https://doi.org:10.3390/app142210511

14 Ramos-Pérez, I., Barbero-Aparicio, J. A., Canepa-Oneto, A., Arnaiz-González, Á. & Maudes-Raedo, J. An extensive performance comparison between feature reduction and feature selection preprocessing algorithms on imbalanced wide data. *Information* **15**, 223 (2024). https://doi.org:10.3390/info15040223

15 Barbieri, M. C., Grisci, B. I. & Dorn, M. Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems with Applications* **249**, 123667 (2024). https://doi.org:10.1016/j.eswa.2024.123667

16 Lisspers, K. *et al.* 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*, 106483 (2021). https://doi.org:10.1016/j.rmed.2021.106483
17 Xiang, Y. *et al.* Asthma Exacerbation Prediction and Risk Factor Analysis Based on a Time-Sensitive, Attentive Neural Network: Retrospective Cohort Study. *J Med Internet Res* **22**, e16981 (2020). https://doi.org:10.2196/16981
18 Lugogo, N. L. *et al.* A Predictive Machine Learning Tool for Asthma Exacerbations: Results from a 12-Week, Open-Label Study Using an Electronic Multi-Dose Dry Powder Inhaler with Integrated Sensors. *J Asthma Allergy* **15**, 1623-1637 (2022). https://doi.org:10.2147/jaa.S377631
19 Jiao, T., Schnitzer, M. E., Forget, A. & Blais, L. Identifying asthma patients at high risk of exacerbation in a routine visit: A machine learning model. *Respir Med* **198**, 106866 (2022). https://doi.org:10.1016/j.rmed.2022.106866
20 Hurst, J. H., Zhao, C., Hostetler, H. P., Ghiasi Gorveh, M., Lang, J. E. & Goldstein, B. A. Environmental and clinical data utility in pediatric asthma exacerbation risk prediction models. *BMC Medical Informatics and Decision Making* **22** (2022). https://doi.org:10.1186/s12911-022-01847-0
21 Inselman, J. W. *et al.* A prediction model for asthma exacerbations after stopping asthma biologics. *Ann. Allergy Asthma Immunol.* **130**, 305-311 (2023). https://doi.org:10.1016/j.anai.2022.11.025
22 Xu, M. *et al.* Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. *BMC Med. Genet.* **12** (2011). https://doi.org:10.1186/1471-2350-12-90
23 Shin, E. K., Mahajan, R., Akbilgic, O. & Shaban-Nejad, A. Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. *NPJ Digit Med* **1**, 50 (2018). https://doi.org:10.1038/s41746-018-0056-y
24 de Hond, A. A. H., Kant, I. M. J., Honkoop, P. J., Smith, A. D., Steyerberg, E. W. & Sont, J. K. Machine learning did not beat logistic regression in time series prediction for severe asthma exacerbations. *Sci Rep* **12**, 20363 (2022). https://doi.org:10.1038/s41598-022-24909-9
25 Finkelstein, J. & Jeong, I. C. in *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON).* 1-5.
26 Finkelstein, J. & Jeong, I. C. Machine learning approaches to personalize early prediction of asthma exacerbations. *Annals of the New York Academy of Sciences* **1387**, 153-165 (2017). https://doi.org:10.1111/nyas.13218
27 Alharbi, E., Cherif, A., Nadeem, F. & Mirza, T. in *IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA).* 1-7.
28 Haque, R. *et al.* in *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices.* (eds Hamido Fujita, Ali and Selamat, Jerry Chun-Wei and Lin, & Moonis and Ali) 297-308 (Springer International Publishing).
29 Priya, C. K., Sudhakar, M., Lingampalli, J. & Basha, C. Z. in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC).* 1138-1145.
30 Patel, S. J., Chamberlain, D. B. & Chamberlain, J. M. A Machine Learning Approach to Predicting Need for Hospitalization for Pediatric Asthma Exacerbation at the Time of Emergency Department Triage. *Acad Emerg Med* **25**, 1463-1470 (2018). https://doi.org:10.1111/acem.13655
31 Do, Q., Tran, S. & Doig, A. in *2019 IEEE International Conference on Big Data (Big Data).* 4829-4838.
32 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301-320 (2005). https://doi.org:10.1111/j.1467-9868.2005.00503.x
33 Rothacher, Y. & Strobl, C. Identifying Informative Predictor Variables With Random Forests. *Journal of Educational and Behavioral Statistics* **49**, 595-629 (2024). https://doi.org:10.3102/10769986231193327
34 Gerstorfer, Y., Krieg, L. & Hahn-Klimroth, M. A notion of feature importance by decorrelation and detection of trends by random forest regression. *Data Science Journal* **22**, 42 (2023). https://doi.org:10.5334/dsj-2023-042
35 Zhang, O., Minku, L. L. & Gonem, S. Detecting asthma exacerbations using daily home monitoring and machine learning. *Journal of Asthma* **58**, 1518-1527 (2021). https://doi.org:10.1080/02770903.2020.1802746
36 Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine learning* **46**, 389-422 (2002). https://doi.org:10.1023/A:1012487302797
37 Zein, J. G., Wu, C. P., Attaway, A. H., Zhang, P. & Nazha, A. Novel Machine Learning Can Predict Acute Asthma Exacerbation. *Chest* **159**, 1747-1757 (2021). https://doi.org:10.1016/j.chest.2020.12.051
38 Chen, J., Zhao, F., Sun, Y. & Yin, Y. Improved XGBoost model based on genetic algorithm. *International Journal of Computer Applications in Technology* **62**, 240-245 (2020). https://doi.org:10.1504/IJCAT.2020.106571

39 Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001). https://doi.org:10.1023/A:1010933404324
40 Rigatti, S. J. Random forest. *Journal of insurance medicine* **47**, 31-39 (2017). https://doi.org:10.17849/insm-47-01-31-39.1
41 Farion, K. J., Wilk, S., Michalowski, W., O'Sullivan, D. & Sayyad-Shirabad, J. Comparing predictions made by a prediction model, clinical score, and physicians: Pediatric asthma exacerbations in the emergency department. *Appl. Clin. Informatics* **4**, 376-391 (2013). https://doi.org:10.4338/ACI-2013-04-RA-0029
42 Reddel, H. K. *et al.* An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *American journal of respiratory and critical care medicine* **180**, 59-99 (2009).
43 Suruki, R. Y., Daugherty, J. B., Boudiaf, N. & Albers, F. C. The frequency of asthma exacerbations and healthcare utilization in patients with asthma from the UK and USA. *BMC pulmonary medicine* **17**, 1-11 (2017).
44 Chan, A. H. Y., Tomlin, A., Beyene, K. & Harrison, J. Asthma exacerbations in New Zealand 2010-2019: A national population-based study. *Respiratory Medicine* **217**, 107365 (2023). https://doi.org:10.1016/j.rmed.2023.107365
45 Selroos, O. *et al.* National and regional asthma programmes in Europe. *Eur Respir Rev* **24**, 474-483 (2015). https://doi.org:10.1183/16000617.00008114
46 Bloom, C. I., Nissen, F., Douglas, I. J., Smeeth, L., Cullinan, P. & Quint, J. K. Exacerbation risk and characterisation of the UK's asthma population from infants to old age. *Thorax* **73**, 313-320 (2018). https://doi.org:10.1136/thoraxjnl-2017-210650
47 American Lung Association. Asthma Trends and Burden. (2022). <https://www.lung.org/research/trends-in-lung-disease/asthma-trends-brief/trends-and-burden>.
48 Miller, M. K., Lee, J. H., Miller, D. P. & Wenzel, S. E. Recent asthma exacerbations: a key predictor of future exacerbations. *Respiratory medicine* **101**, 481-489 (2007). https://doi.org:10.1016/j.rmed.2006.07.005
49 Loymans, R. J. *et al.* Exacerbations in adults with asthma: a systematic review and external validation of prediction models. *The Journal of Allergy and Clinical Immunology: In Practice* **6**, 1942-1952. e1915 (2018).
50 Lowden, R. & Turner, S. Past asthma exacerbation in children predicting future exacerbation: a systematic review. *ERJ Open Research* **8**, 00174-02022 (2022). https://doi.org:10.1183/23120541.00174-2022
51 Jayamini, W. K. D., Mirza, F., Naeem, M. A. & Chan, A. H. Y. State of Asthma-Related Hospital Admissions in New Zealand and Predicting Length of Stay Using Machine Learning. *Applied Sciences* **12**, 9890 (2022). https://doi.org:10.3390/app12199890
52 Hussain, Z., Shah, S. A., Mukherjee, M. & Sheikh, A. Predicting the risk of asthma attacks in children, adolescents and adults: Protocol for a machine learning algorithm derived from a primary care-based retrospective cohort. *BMJ Open* **10**, e036099 (2020). https://doi.org:10.1136/bmjopen-2019-036099
53 Blakey, J. D. *et al.* Identifying Risk of Future Asthma Attacks Using UK Medical Record Data: A Respiratory Effectiveness Group Initiative. *The Journal of Allergy and Clinical Immunology: In Practice* **5**, 1015-1024.e1018 (2017). https://doi.org:10.1016/j.jaip.2016.11.007
54 Boehncke, W.-H., Boehncke, S. & Schön, M. P. Managing comorbid disease in patients with psoriasis. *BMJ* **340** (2010). https://doi.org:10.1136/bmj.b5666
55 Wang, J. *et al.* in *Allergy and asthma proceedings.* 103-109.
56 Fang, H. Y., Liao, W. C., Lin, C. L., Chen, C. H. & Kao, C. H. Association between psoriasis and asthma: a population-based retrospective cohort analysis. *British Journal of Dermatology* **172**, 1066-1071 (2015). https://doi.org:10.1111/bjd.13518
57 Bayman, E. O. & Dexter, F. Vol. 133 362-365 (LWW, Anesthesia & Analgesia, 2021).
58 Zeng, G. & Zeng, E. On the relationship between multicollinearity and separation in logistic regression. *Communications in Statistics - Simulation and Computation* **50**, 1989-1997 (2021). https://doi.org:10.1080/03610918.2019.1589511
59 Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* **10**, e0118432 (2015). https://doi.org:10.1371/journal.pone.0118432
60 Jeni, L. A., Cohn, J. F. & Torre, F. D. L. in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction.* 245-251.
61 Hairani, H. & Priyanto, D. A new approach of hybrid sampling SMOTE and ENN to the accuracy of machine learning methods on unbalanced diabetes disease data. *International Journal of Advanced Computer Science and Applications* **14** (2023). https://doi.org:10.14569/IJACSA.2023.0140864
62 Wang, Y. & Ni, X. S. A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433* (2019). https://doi.org:10.48550/arXiv.1901.08433

Appendix A

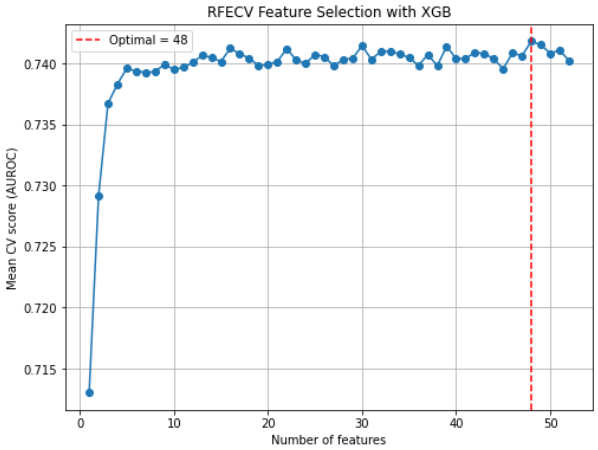

Figure A1 Recursive Feature Elimination with Cross-Validation (RFECV) using the XGB classifier on refined full feature set (FS2), showing mean Area Under the Receiver Operating Characteristic curve (AUROC) across different number of selected features. It is noted that the optimal is shown as 48, as it has counted the one-hot encoded features, but it has only 23 distinct features.

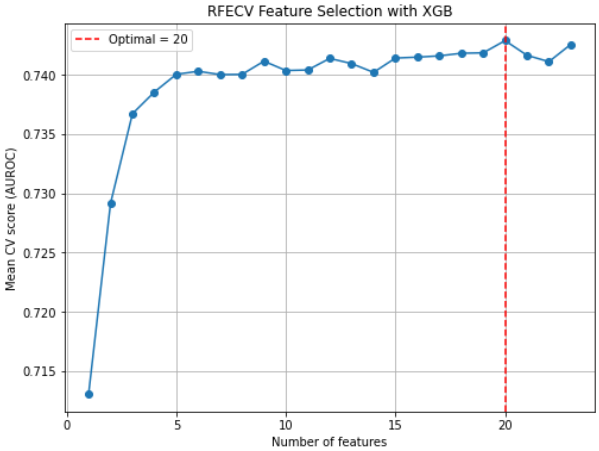

Figure A2 Recursive Feature Elimination with Cross-Validation (RFECV) using the XGB classifier on clinical full feature set (FS4), showing mean Area Under the Receiver Operating Characteristic curve (AUROC) across different number of selected features. It is noted that the optimal is shown as 20, as it has counted the one-hot encoded features, but it has only 14 distinct features.

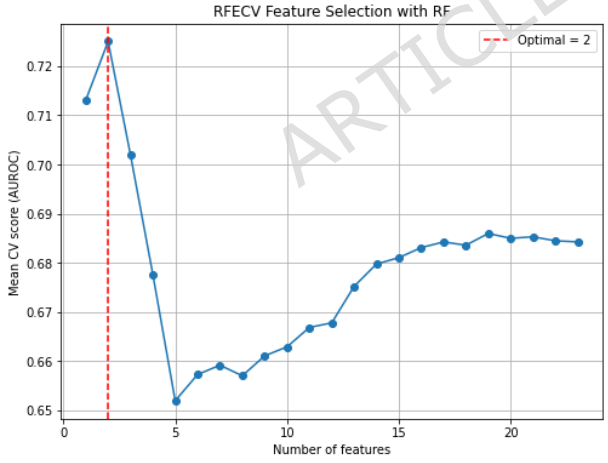

Figure A3 Recursive Feature Elimination with Cross-Validation (RFECV) using the RF classifier on refined full feature set (FS2), showing mean Area Under the Receiver Operating Characteristic curve (AUROC) across different number of selected features.

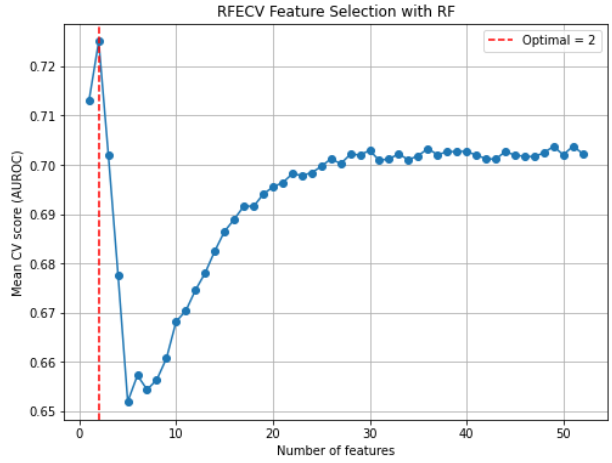

Figure A4 Recursive Feature Elimination with Cross-Validation (RFECV) using the RF classifier on refined full feature set (FS4), showing mean Area Under the Receiver Operating Characteristic curve (AUROC) across different number of selected features.

Appendix B

Table B1 Population Stability Index (PSI) calculated to quantify the dataset shift between the test and validation cohorts.

Feature	PSI	Shift
Q5_WeeksDuringWinter	0.0107	No Shift
Age	0.0267	No Shift
Gender	<0.0001	No Shift
EthnicGroup	0.0071	No Shift
DeprivationQuintile	0.0003	No Shift
DHB	0.0100	No Shift
Psoriasis	<0.0001	No Shift
Obesity	0.0007	No Shift
LRTI	0.0083	No Shift
AnxietyDepression	0.0118	No Shift
Diabetes	0.0027	No Shift
SmokingStatus	<0.0001	No Shift
CharlsonComorbidityIndexScore	0.0000	No Shift
NumberOfMetforminRx	<0.0001	No Shift
Diabetes_NoMetformin	0.0049	No Shift
NoOfHospitalisations	0.0056	No Shift
NoOfOPEDVisits	0.0068	No Shift
BetaBlockers	0.0015	No Shift
Paracetamol	0.0056	No Shift
NSAIDs	0.0011	No Shift
SABA_ICS_Ratio	1.8396	Significant Shift
P12MNoOfAsthAttacks	0.1232	Moderate Shift
HistoryOfAllergies	0.2667	Significant Shift
CardiovascularDiseases	0.0085	No Shift

Appendix C

Table C1 Hyperparameter values used for developing the prediction models.

ML Model	Hyperparamenter	Value
LR	Penalty	L2
	Regularisation Strength (C)	1.0
	Solver	lbfgs
	Maximum Iterations	400

	Tolerance	0.0001
RF	Number of Trees	100
	Split Criterion	Gini
	Maximum Tree Depth	None (unlimited)
	Maximum Features per Split	sqrt
	Minimum Samples per Split	2
	Minimum Samples per Leaf	1
	Bootstrap Sampling	True
	Out-of-Bag Evaluation	False
XGB	Objective	binary:logistic
	Number of trees	100
	Learning rate	0.30
	Maximum tree depth	6
	Minimum child weight	1
	Gamma	0
	Subsample ratio	1.0
	Column sampling by tree	1.0
	L1 regularisation (α)	0
	L2 regularisation (λ)	1
MLP	Optimizer	Adam
	Learning Rate	0.001
	Loss Function	BCEWithLogitsLoss
	Batch Size	64
	Maximum Epochs	100
	Dropout Rate	0.3
	Activation Function	ReLU

Appendix D

Table D1 Performance of the models in 5-fold CV and using test dataset.

Feature Set	**Feature Set Description**	**ML Algorithm**	**Imbalance Handling Technique**	**Mean AUROC (95% CI) - CV**	**AUROC (Test set)**	**Mean F1 Score (95% CI) - CV**
FS1	XGB-based importance 10 features	LR	ENN	0.75 (0.74-0.76)	0.75	0.28 (0.27-0.29)
FS1	XGB-based importance 10 features	LR	KMeansSMOTE	0.74 (0.73-0.76)	0.74	0.29 (0.27-0.31)
FS1	XGB-based importance 10 features	LR	No-sample	0.75 (0.74-0.75)	0.75	0.16 (0.15-0.17)

FS1	XGB-based importance 10 features	LR	RUS	0.75 (0.74-0.75)	0.75	0.28 (0.28-0.29)
FS1	XGB-based importance 10 features	LR	SMOTE	0.75 (0.74-0.75)	0.75	0.28 (0.27-0.29)
FS1	XGB-based importance 10 features	RF	ENN	0.69 (0.69-0.69)	0.69	0.23 (0.22-0.23)
FS1	XGB-based importance 10 features	RF	KMeansSMOTE	0.69 (0.69-0.69)	0.69	0.18 (0.17-0.19)
FS1	XGB-based importance 10 features	RF	No-sample	0.69 (0.68-0.69)	0.69	0.16 (0.15-0.16)
FS1	XGB-based importance 10 features	RF	RUS	0.71 (0.70-0.71)	0.71	0.21 (0.21-0.22)
FS1	XGB-based importance 10 features	RF	SMOTE	0.67 (0.66-0.67)	0.66	0.20 (0.19-0.20)
FS1	XGB-based importance 10 features	XGB	ENN	0.75 (0.74-0.75)	0.75	0.28 (0.27-0.30)
FS1	XGB-based importance 10 features	XGB	KMeansSMOTE	0.75 (0.74-0.75)	0.74	0.18 (0.16-0.20)
FS1	XGB-based importance 10 features	XGB	No-sample	0.75 (0.74-0.75)	0.75	0.16 (0.16-0.17)
FS1	XGB-based importance 10 features	XGB	RUS	0.74 (0.73-0.75)	0.74	0.25 (0.25-0.26)
FS1	XGB-based importance 10 features	XGB	SMOTE	0.67 (0.67-0.68)	0.67	0.18 (0.17-0.19)
FS2	Refined full set 24 features	LR	ENN	0.74 (0.74-0.75)	0.74	0.27 (0.26-0.28)
FS2	Refined full set 24 features	LR	KMeansSMOTE	0.70 (0.67-0.73)	0.73	0.28 (0.25-0.31)
FS2	Refined full set 24 features	LR	No-sample	0.74 (0.74-0.75)	0.73	0.15 (0.14-0.16)
FS2	Refined full set 24 features	LR	RUS	0.74 (0.74-0.75)	0.74	0.28 (0.27-0.28)
FS2	Refined full set 24 features	LR	SMOTE	0.73 (0.73-0.74)	0.73	0.27 (0.27-0.28)
FS2	Refined full set 24 features	RF	ENN	0.71 (0.70-0.71)	0.71	0.24 (0.23-0.26)
FS2	Refined full set 24 features	RF	KMeansSMOTE	0.70 (0.70-0.71)	0.70	0.14 (0.11-0.16)
FS2	Refined full set 24 features	RF	No-sample	0.70 (0.69-0.71)	0.70	0.12 (0.11-0.13)
FS2	Refined full set 24 features	RF	RUS	0.72 (0.71-0.72)	0.72	0.23 (0.22-0.23)

FS2	Refined full set 24 features	RF	SMOTE	0.70 (0.70-0.71)	0.70	0.13 (0.12-0.13)
FS2	Refined full set 24 features	XGB	ENN	0.74 (0.73-0.75)	0.75	0.26 (0.25-0.27)
FS2	Refined full set 24 features	XGB	KMeansSMOTE	0.73 (0.72-0.74)	0.74	0.17 (0.16-0.17)
FS2	Refined full set 24 features	XGB	No-sample	0.74 (0.73-0.75)	0.74	0.14 (0.13-0.15)
FS2	Refined full set 24 features	XGB	RUS	0.73 (0.73-0.74)	0.74	0.25 (0.24-0.25)
FS2	Refined full set 24 features	XGB	SMOTE	0.72 (0.71-0.72)	0.71	0.16 (0.15-0.17)
FS3	RFECV-XGB 23 features	LR	ENN	0.74 (0.74-0.75)	0.74	0.27 (0.26-0.28)
FS3	RFECV-XGB 23 features	LR	KMeansSMOTE	0.70 (0.67-0.73)	0.73	0.28 (0.25-0.30)
FS3	RFECV-XGB 23 features	LR	No-sample	0.74 (0.74-0.75)	0.74	0.15 (0.14-0.16)
FS3	RFECV-XGB 23 features	LR	RUS	0.74 (0.74-0.75)	0.74	0.28 (0.27-0.28)
FS3	RFECV-XGB 23 features	LR	SMOTE	0.73 (0.73-0.74)	0.73	0.27 (0.27-0.28)
FS3	RFECV-XGB 23 features	RF	ENN	0.70 (0.70-0.71)	0.70	0.24 (0.23-0.25)
FS3	RFECV-XGB 23 features	RF	KMeansSMOTE	0.70 (0.70-0.71)	0.70	0.14 (0.11-0.16)
FS3	RFECV-XGB 23 features	RF	No-sample	0.70 (0.69-0.71)	0.70	0.11 (0.10-0.12)
FS3	RFECV-XGB 23 features	RF	RUS	0.72 (0.71-0.72)	0.72	0.23 (0.22-0.23)
FS3	RFECV-XGB 23 features	RF	SMOTE	0.70 (0.70-0.71)	0.70	0.13 (0.12-0.14)
FS3	RFECV-XGB 23 features	XGB	ENN	0.74 (0.73-0.75)	0.74	0.26 (0.25-0.28)
FS3	RFECV-XGB 23 features	XGB	KMeansSMOTE	0.73 (0.72-0.74)	0.74	0.17 (0.16-0.18)
FS3	RFECV-XGB 23 features	XGB	No-sample	0.74 (0.74-0.75)	0.74	0.14 (0.13-0.16)
FS3	RFECV-XGB 23 features	XGB	RUS	0.73 (0.73-0.74)	0.74	0.25 (0.24-0.25)
FS3	RFECV-XGB 23 features	XGB	SMOTE	0.72 (0.71-0.72)	0.71	0.16 (0.15-0.17)

FS4	Clinical full set 16 features	LR	ENN	0.74 (0.74-0.75)	0.74	0.27 (0.26-0.28)
FS4	Clinical full set 16 features	LR	KMeansSMOTE	0.71 (0.68-0.74)	0.72	0.29 (0.27-0.32)
FS4	Clinical full set 16 features	LR	No-sample	0.74 (0.74-0.75)	0.74	0.15 (0.14-0.16)
FS4	Clinical full set 16 features	LR	RUS	0.74 (0.74-0.75)	0.74	0.28 (0.27-0.28)
FS4	Clinical full set 16 features	LR	SMOTE	0.74 (0.74-0.74)	0.74	0.28 (0.27-0.28)
FS4	Clinical full set 16 features	RF	ENN	0.68 (0.68-0.69)	0.68	0.22 (0.22-0.23)
FS4	Clinical full set 16 features	RF	KMeansSMOTE	0.68 (0.68-0.69)	0.68	0.18 (0.15-0.20)
FS4	Clinical full set 16 features	RF	No-sample	0.68 (0.68-0.69)	0.68	0.14 (0.13-0.15)
FS4	Clinical full set 16 features	RF	RUS	0.70 (0.69-0.70)	0.70	0.21 (0.21-0.22)
FS4	Clinical full set 16 features	RF	SMOTE	0.67 (0.66-0.67)	0.66	0.17 (0.17-0.18)
FS4	Clinical full set 16 features	XGB	ENN	0.74 (0.73-0.75)	0.75	0.27 (0.26-0.28)
FS4	Clinical full set 16 features	XGB	KMeansSMOTE	0.74 (0.73-0.74)	0.74	0.19 (0.17-0.21)
FS4	Clinical full set 16 features	XGB	No-sample	0.74 (0.73-0.75)	0.74	0.14 (0.13-0.15)
FS4	Clinical full set 16 features	XGB	RUS	0.73 (0.73-0.74)	0.74	0.25 (0.24-0.25)
FS4	Clinical full set 16 features	XGB	SMOTE	0.68 (0.67-0.69)	0.67	0.18 (0.16-0.19)
FS5	Clinical RFECV-XGBF 14 features	LR	ENN	0.74 (0.74-0.75)	0.74	0.27 (0.26-0.28)
FS5	Clinical RFECV-XGBF 14 features	LR	KMeansSMOTE	0.71 (0.70-0.72)	0.70	0.30 (0.29-0.32)
FS5	Clinical RFECV-XGBF 14 features	LR	No-sample	0.74 (0.74-0.75)	0.74	0.15 (0.14-0.16)
FS5	Clinical RFECV-XGBF 14 features	LR	RUS	0.74 (0.74-0.75)	0.74	0.28 (0.27-0.28)
FS5	Clinical RFECV-XGBF 14 features	LR	SMOTE	0.74 (0.73-0.74)	0.74	0.28 (0.27-0.28)
FS5	Clinical RFECV-XGBF 14 features	RF	ENN	0.68 (0.67-0.69)	0.71	0.22 (0.21-0.23)

FS5	Clinical RFECV-XGBF 14 features	RF	KMeansSMOTE	0.68 (0.67-0.69)	0.70	0.19 (0.18-0.20)
FS5	Clinical RFECV-XGBF 14 features	RF	No-sample	0.68 (0.67-0.68)	0.70	0.14 (0.13-0.15)
FS5	Clinical RFECV-XGBF 14 features	RF	RUS	0.69 (0.69-0.70)	0.71	0.21 (0.21-0.21)
FS5	Clinical RFECV-XGBF 14 features	RF	SMOTE	0.66 (0.66-0.67)	0.69	0.18 (0.17-0.19)
FS5	Clinical RFECV-XGBF 14 features	XGB	ENN	0.74 (0.74-0.75)	0.74	0.27 (0.25-0.28)
FS5	Clinical RFECV-XGBF 14 features	XGB	KMeansSMOTE	0.74 (0.73-0.75)	0.72	0.20 (0.18-0.21)
FS5	Clinical RFECV-XGBF 14 features	XGB	No-sample	0.74 (0.74-0.75)	0.74	0.14 (0.13-0.15)
FS5	Clinical RFECV-XGBF 14 features	XGB	RUS	0.73 (0.72-0.74)	0.74	0.25 (0.24-0.25)
FS5	Clinical RFECV-XGBF 14 features	XGB	SMOTE	0.68 (0.67-0.69)	0.72	0.18 (0.17-0.19)
FS6	RFECV-RF 2 features	LR	ENN	0.73 (0.72-0.74)	0.70	0.27 (0.24-0.31)
FS6	RFECV-RF 2 features	LR	KMeansSMOTE	0.71 (0.68-0.73)	0.70	0.30 (0.28-0.32)
FS6	RFECV-RF 2 features	LR	No-sample	0.73 (0.72-0.74)	0.73	0.16 (0.15-0.17)
FS6	RFECV-RF 2 features	LR	RUS	0.73 (0.72-0.74)	0.73	0.27 (0.26-0.27)
FS6	RFECV-RF 2 features	LR	SMOTE	0.73 (0.73-0.74)	0.73	0.27 (0.26-0.27)
FS6	RFECV-RF 2 features	RF	ENN	0.70 (0.69-0.71)	0.55	0.23 (0.20-0.25)
FS6	RFECV-RF 2 features	RF	KMeansSMOTE	0.73 (0.72-0.73)	0.72	0.21 (0.12-0.29)
FS6	RFECV-RF 2 features	RF	No-sample	0.72 (0.72-0.73)	0.72	0.13 (0.12-0.14)
FS6	RFECV-RF 2 features	RF	RUS	0.73 (0.72-0.74)	0.73	0.26 (0.26-0.27)
FS6	RFECV-RF 2 features	RF	SMOTE	0.71 (0.70-0.72)	0.71	0.26 (0.26-0.26)
FS6	RFECV-RF 2 features	XGB	ENN	0.71 (0.70-0.72)	0.52	0.23 (0.20-0.26)
FS6	RFECV-RF 2 features	XGB	KMeansSMOTE	0.73 (0.73-0.74)	0.73	0.21 (0.12-0.30)
FS6	RFECV-RF 2 features	XGB	No-sample	0.73 (0.73-0.74)	0.73	0.12 (0.11-0.13)
FS6	RFECV-RF 2 features	XGB	RUS	0.73 (0.72-0.74)	0.73	0.26 (0.26-0.27)
FS6	RFECV-RF 2 features	XGB	SMOTE	0.73 (0.72-0.74)	0.73	0.27 (0.26-0.27)

Table D2 Performance of the MLP models on the test dataset.

Feature Set	Imbalance Handling Technique	AUROC	F1-score	AUPRC	Brier Score
FS1	ENN	0.68	0.13	0.18	0.93
	KMeansSMOTE	0.65	0.13	0.20	0.90
	No-sample	0.74	0.15	0.29	0.06
	RUS	0.46	0.13	0.11	0.93
	SMOTE	0.64	0.13	0.17	0.93
FS2	ENN	0.63	0.13	0.14	0.93
	KMeansSMOTE	0.52	0.13	0.09	0.93
	No-sample	0.73	0.12	0.27	0.06
	RUS	0.68	0.13	0.17	0.93
	SMOTE	0.38	0.13	0.09	0.93
FS3	ENN	0.67	0.13	0.14	0.93
	KMeansSMOTE	0.58	0.13	0.08	0.93
	No-sample	0.73	0.11	0.26	0.06
	RUS	0.68	0.13	0.17	0.93
	SMOTE	0.68	0.13	0.16	0.93
FS4	ENN	0.64	0.13	0.17	0.93
	KMeansSMOTE	0.49	0.13	0.07	0.93
	No-sample	0.74	0.09	0.28	0.06
	RUS	0.66	0.13	0.15	0.93
	SMOTE	0.50	0.13	0.14	0.93
FS5	ENN	0.68	0.13	0.15	0.93
	KMeansSMOTE	0.39	0.13	0.06	0.93
	No-sample	0.74	0.13	0.27	0.06
	RUS	0.69	0.13	0.17	0.93
	SMOTE	0.56	0.13	0.17	0.93
FS6	ENN	0.66	0.13	0.12	0.93
	KMeansSMOTE	0.72	0.28	0.19	0.10
	No-sample	0.73	0.09	0.27	0.06
	RUS	0.66	0.13	0.13	0.93
	SMOTE	0.53	0.13	0.09	0.93

Appendix E

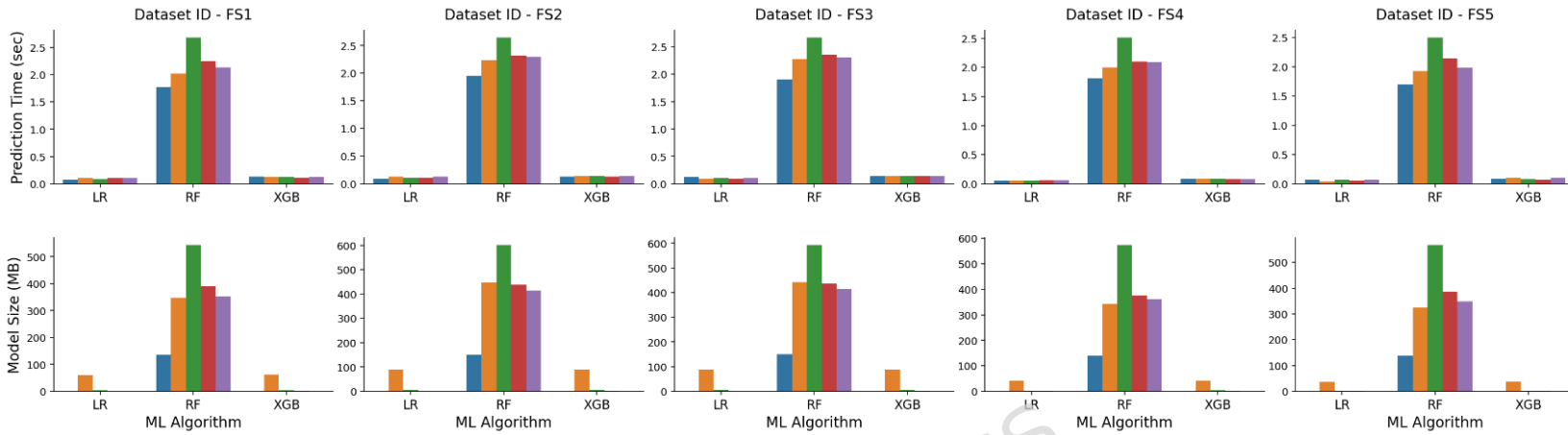

Figure E1 Comparison of the efficiency of the models based on different feature sets and data imbalance handling techniques.

Appendix E

Distribution of values in the features that were excluded from the feature sets

Table E1 Psoriasis vs asthma attacks in the training set.

Psoriasis	Asthma Attack=0	Asthma Attack=1	Total
0	207,016	16,107	223,123
1	30	5	35
Total	207,046	16,112	223,158

Table E2 Number of hospitalizations vs asthma attacks in the training set.

Metric / Category	Asthma Attack=0	Asthma Attack=1
No Hospitalizations (0)	167,127	11,784
Hospitalized (1+)	39,919	4,328
Total	207,046	16,112
Minimum	0	0
Q1	0	0
Median	0	0
Q3	0	1
Maximum	137	152
Mean	0.32	0.53

Table E3 Number of Metformin Rx vs asthma attacks in the training set.

Metric / Category	Asthma Attack=0	Asthma Attack=1
No Metformin Use (0)	198,129	15,377
Used Metformin (1+)	8,917	735
Total	207,046	16,112
Minimum	0	0
Q1	0	0
Median	0	0
Q3	0	0
Maximum	352	101
Mean	0.32	0.34