Proposed Methodology

The implementation of machine learning algorithms for paediatric asthma classification required a comprehensive data preprocessing framework addressing the complexities of questionnaire-derived healthcare data. Our methodology was designed around three principles: preservation of clinical validity, maintenance of statistical integrity, and optimization for machine learning performance. The approach employed a five-stage preprocessing pipeline motivated by the heterogeneous nature of questionnaire-oriented attributes, object-datatype biased dataset, dual asthma-related target variables and missing data patterns requiring sophisticated handling strategies. Figure 2. depicts the proposed methodology and the subsequent sections delves into the implemented methodology.

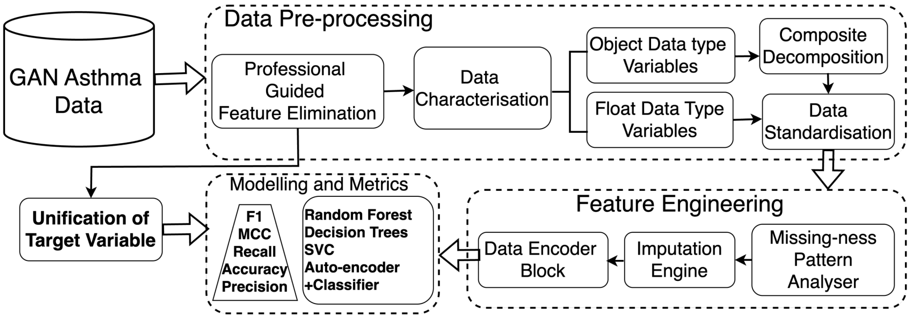

Fig. 2. Proposed Methodology.

Expert guided feature elimination.

Under supervised guidance of Dr. PA Mahesh, a professional contributor to the data acquisition process, 38 asthma irrelevant attributes were systematically removed from the dataset. The eliminated features encompassed administrative metadata (form identifiers, version numbers, geographical markers), data collection logistics (serial numbers, centre codes, school identifiers), measurement unit specifications, date-related variables with limited predictive value, and highly specific clinical variables deemed non-essential for core asthma classification objectives. This evidence-based reduction eliminated 20.9% of the original feature space while maintaining comprehensive coverage across all major asthma-related domains, resulting in 143 variables preserving critical information spanning respiratory symptoms, environmental exposures, family history, and treatment patterns.

Composite Feature Decomposition.

Several questionnaire items encoded multiple temporal dimensions within single categorical variables, requiring systematic decomposition to preserve temporal granularity. These composite variables typically represented four temporal categories: never experienced, currently experiencing, experienced during first year, and ever experienced. The decomposition process employed vectorized string parsing to transform each composite variable into four binary indicators, expanding the feature space from 143 to 156 variables while maintaining complete temporal information.

Data Standardization.

Medication-related variables presented unique challenges due to sparsity, inconsistent naming conventions, and logical contradictions. The standardization framework addressed these through a sequential approach: Standardized 47 unique medication entries into four therapeutic classes (bronchodilators, anti-inflammatory agents, antihistamines, and antibiotics) using clinical pharmacology principles; Geographic birth location variables exhibited significant heterogeneity and inconsistent entries, requiring systematic feature engineering. The standardization process implemented conditional logic frameworks producing seven distinct geographic categories: in_india, outside_india, uae, nepal.

2.5.1 Unification of Target Variable

The dataset contained two primary asthma-related variables: doctor-diagnosed asthma (asthdoc) representing professional medical diagnosis indicators, and self-reported asthma events (asthmaev) capturing patient-reported experiences. Initial statistical analysis revealed significant disparities in data completeness, necessitating a systematic approach to target variable unification that would maximize sensitivity while maintaining clinical interpretability.

The doctor-diagnosed asthma variable exhibited extreme data sparsity with 98.27% missing values, 1.3% positive cases, and 0.2% negative cases. In contrast, the self-reported asthma events variable demonstrated superior data completeness with no missing values, 1.72% positive cases, and 98.27% negative cases. The substantial missing data rate in the professional diagnosis variable rendered it unsuitable for direct utilization as a target variable.

To address this incompatibility, introducing statistical analysis revealed the patterns in the co-occurrence of asthmatic cases which is explained in Table 2. These occurrence patterns inspired the usage of logical OR operation combining strengths of both sources (asthdoc, asthmaev) for potential asthma case detection. A comprehensive conflict resolution analysis revealed zero conflicts between the two target variables, indicating perfect concordance when both data points were available, validating the feasibility of creating a unified target variable without introducing inconsistencies. An unified target variable (asthma_combined) was created with logical strengths of dual target variables with null co-occurrence conflicts.

Table 2. Unified Target Variable Decision Matrix

asthdoc	asthmaev	asthma_combined	Decision Logic
Yes	No	Yes	Professional diagnosis precedence
No	Yes	Yes	Self-reported inclusion

NaN	No	No	Clear negative classification
NaN	NaN	None	Insufficient information

asthma_combined results in 2,682 with non-asthmatic targets, 47 with Asthmatic targets and 1 as Un-confirmed target. This Un-confirmed target is decided to be excluded from the dataset by consideration of non-specific knowledge of origin. This final data pre-processing step resulted a comprehensive dataset with 2,729 total number of samples where 2,682 are non-asthmatic samples and 47 are asthmatic samples. Dataset pre-processed, provides a platform for further feature engineering strategies and modelling.

2.6 Feature Engineering

The pre-processed dataset requires systematic feature engineering to optimize machine learning compatibility, addressing strategic imputation of missing values while preserving clinical relationships and systematic encoding of diverse data types ensuring algorithmic compatibility.

Data Imputation.

The comprehensive analysis of missing data patterns revealed complex heterogeneous missingness mechanisms requiring appropriate imputation strategies. Our novel wrapper framework systematically addresses missing value challenges through multi-dimensional characterization of feature properties including: distribution normality, skewness, multimodality, outlier presence, sparsity levels, and co-relation analysis.

The implementation employed a comprehensive analytical pipeline evaluating each variable across eight critical dimensions including: Missing mechanism assessment categorizes patterns as Missing Completely at Random (MCAR), Missing at Random (MAR) or Missing Not at Random (MNAR) through correlation analysis, skewness of the variable is detected by employing Fisher-Pearson's co-efficient, outlier analysis is carried out using the Inter-Quartile Range (IQR) where datapoints outside the Q1 and Q3 are considered to be outliers, distribution of the data is estimated using D'Agostino-Pearson test and Sparsity of the variables are calculated by a ratio of Number of Unique values to the ratio of Total Unique values. Attributes with ratio<0.01 is considered to be sparse. A decision matrix integrates these multiple assessment criteria to recommend optimal imputation techniques. Table 3. portray missingness patterns and handling methods.

Table 3. Data Imputation Rationale.

Missing Pattern	Skew-ness	Multi-modal	Outliers	Sparsity	Missing-ness Pattern	Features Relation-ship	Distribution	Technique
MCAR	No	No	No	Yes	Independent	Low Correlation	Normal	Mean/Median Imputation

Missing Pattern	Skew-ness	Multimodal	Outliers	Sparsity	Missingness Pattern	Features Relationship	Distribution	Technique
MCAR	No	No	Yes	Yes	Independent	Low Correlation	Normal	Median/ Winsorized Mean
MCAR	No	Yes	No	Yes	Independent	Low Correlation	Multimodal	Mode or Multiple Imputation
MCAR	Yes	No	Yes	Yes	Independent	Low Correlation	Skewed	Winsorized Mean
MCAR	Yes	Yes	Yes	No	Independent	High Correlation	Multimodal	Predictive Imputation
MAR	No	No	No	No	Dependent	High Correlation	Normal	Median, Predictive Imputation
MAR	Yes	Yes	Yes	Yes	Dependent	High Correlation	Multimodal	KNN Imputation, Multiple Imputation
MNAR	Any	Any	Any	Any	Any	Any	Any	Advanced Methods (MICE, GANs)

Data Encoding.

Following data imputation, we introduce a novel automated encoding pipeline (Global-Asthma-Network Pipeline) later referred as GAN pipeline that revolutionizes categorical variable transformation through intelligent encoding strategy selection. This framework addresses optimal encoding method selection by implementing a sophisticated multi-dimensional analytical approach that systematically evaluates variable characteristics and automatically recommends appropriate encoding strategies.

The comprehensive framework introduces an innovative decision matrix that automatically categorizes variables into optimal encoding strategies. Variables with limited unique values (≤10 categories) undergo automated ordinality assessment through integrated domain knowledge and statistical analysis. Ordinal variables with natural ordering relationships receive label encoding to preserve hierarchical structure, while nominal variables benefit from one-hot encoding to maintain categorical independence. High-cardinality variables (>50 unique values) trigger advanced evaluation protocols for target encoding or binary encoding strategies.

The pipeline achieved optimal encoding distribution with 5 variables for one-hot encoding and 115 variables for label encoding, accurately reflecting the predominant ordinal nature of clinical variables through automated detection algorithms. This novel GAN-Pipeline framework represents a significant contribution to automated feature engineering, establishing automated categorical variable encoding that ensures optimal representation while maintaining clinical interpretability and computational efficiency.

2.7 Modelling

Feature Engineered dataset (2,729 samples: 2,682 non-asthmatic samples, 47 asthmatic samples) underwent 70-30 train-test split with stratified sampling method, yielding 1910 training and 819 test samples.

The modelling pipeline employed Logistic Regression and Random Forest classifiers with balanced class weights, Support Vector Machines with Grid-Search-Cross-Validation optimization.

All models were evaluated using: Accuracy, AU-ROC, AU-PR, Precision, Recall, F1-score, and Matthews Correlation Coefficient (MCC), with MCC emphasized as the primary metric for imbalanced dataset assessment.

3 Results

Table 4. Provides a comprehensive analysis across different algorithms for modelling over sophisticated metrics. Where, False Negative cases (FN) and False Positive (FP) is critical in medical ML, as missing a disease diagnosis can lead to severe, potentially fatal consequences.

Table 4. "Comparative analysis."

Algorithm	Accuracy	MCC	F1	Precision	Recall	ROC-AUC	AUC-PR	FN	FP
Random Forest	99.98%	0.965	0.966	0.933	1.00	1.00	1.00	0	1
SVC	99.87%	0.965	0.966	0.933	1.00	1.00	1.00	0	1
Logistic Regression	99.98%	0.965	0.966	0. 933	1.00	1.00	1.00	1	1

Achieving the superior performance and minimal false negative case classification, the Random Forest Classifier algorithm is proposed as the optimal solution for paediatric asthma prediction applications with Figure 3. Portraying the superior performance of decision tree algorithm on the test-set of the data.

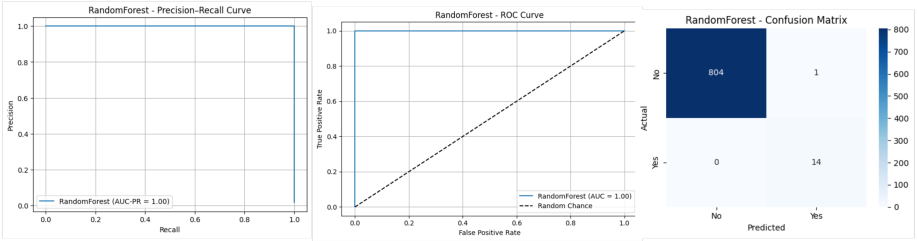

Fig. 3. Test Metrics on Decision Tree.

We propose Random Forest Classifier as deployable and efficient algorithm for real world paediatric asthma classification with our novel feature-engineering framework.

4 Conclusion

This research presents a comprehensive machine learning framework for paediatric asthmatic and non-asthmatic classification specifically targeting the critical 6-7 year age group. The study addresses significant gaps in current asthma screening methodologies through several key innovations that demonstrate both technical excellence and practical clinical applicability.

The novel GAN-Pipeline represents a substantial advancement in automated feature engineering, employing intelligent decision-making frameworks that systematically evaluate categorical variables and automatically recommend optimal encoding strategies. This framework eliminates manual encoding decisions while maintaining clinical interpretability and computational efficiency. The sophisticated multi-dimensional analytical approach successfully transformed heterogeneous questionnaire data into optimized machine learning features through evidence-based imputation strategies and automated encoding pipelines.

The Random Forest Classifier algorithm achieved superior performance with 99.98% accuracy with nill false negative case classification and single false positive case classification, demonstrating clinical reliability crucial for pediatric healthcare applications. Null false negative rate ensures that virtually no at-risk children are missed during screening, addressing a critical requirement for early intervention programs. The methodology's exclusive reliance on questionnaire-derived features eliminates the need for expensive clinical testing, laboratory procedures, and specialized equipment, offering a significantly cost-effective solution for resource-constrained healthcare settings. This approach reduces screening costs by noticeable value compared to traditional clinical diagnostic methods while maintaining practical accuracy.

The novelty of our work lies in the integration of Vast Dataset with GAN standardized protocols, domain-expert guidance, automated feature engineering and clinical-focused optimization that achieves state-of-the-art accuracy while prioritizing no false negatives cases. This cost-effective framework promises to transform pediatric asthma screening by providing an accessible, reliable, and economically viable early detection system that can significantly improve respiratory health outcomes in vulnerable populations where healthcare resources are limited.