Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2187~2194
e-ISSN: 2548-6861

Comparative Analysis of Logistic Regression, Random Forest, and SVM for Asthma Risk Prediction Using Demographic, Clinical, and Environmental Features

Barnabas Belieffain Fertility Daeli [1]*, Ucta Pradema Sanjaya [2]*
* Teknik Informatika, Universitas Ngudi Waluyo
daelibarnabas22@gmail.com [1], uctapradema@unw.ac.id [2],

Article Info

Article history:

Received 2025-08-18
Revised 2025-09-03
Accepted 2025-09-20

Keyword:

Asthma Risk Stratification,
Clinical Decision Support Systems,
Non-Atopic Phenotype,
Random Forest Classification,
Recall-Precision Tradeoff.

ABSTRACT

Asthma prediction demands architectures capable of capturing multifactorial interactions among demographic, clinical, and environmental determinants. This study establishes Random Forest (RF) as the optimal solution through rigorous comparison with Logistic Regression (LR) and Support Vector Machines (SVM) on a 10,000-patient cohort. RF achieved performance: 99.55% accuracy, 100% precision, 98.19% recall, and exceptional stability (σ=0.0019 CV) surpassing SVM by 6.86% recall, preventing 167 missed diagnoses per 10,000 cases. Hereditary factors dominated feature importance (Gini=0.20), generating 18.7% greater node purity reduction than BMI, while the paradoxical "No Allergies" signal (3.726) revealed non-atopic phenotypes. Critically, sparse linear correlations (94% $|r|<0.02$) contrasted with RF's capture of nonlinear thresholds like sedentarism (2.243) > smoking impact. Clinical implementation requires: (1) threshold calibration (θ=0.3) achieving >99% recall, (2) monthly false-negative audits mitigating 24.33% prevalence skew, and (3) dimensionality reduction eliminating 3.256 features. RF's capacity to resolve hereditary-environmental interactions establishes a new paradigm for asthma risk stratification.

I. INTRODUCTION

Asthma presents a global public health challenge characterized by considerable morbidity, economic burden, and diminished quality of life. Effective risk stratification requires advanced predictive modeling capable of decoding complex interactions within high-dimensional feature spaces encompassing demographic, clinical, and environmental determinants [1]. Feature engineering reveals critical predictors including hereditary markers, bronchial inflammation biomarkers, and particulate matter exposure indices. Machine learning architectures must navigate nonlinear decision boundaries and address class imbalance through sophisticated sampling techniques[2]. The integration of heterogeneous feature vectors demands robust ensemble-based architectures to optimize generalization across diverse patient phenotypes. Kernel-based transformations enable separation of non-linearly separable risk clusters, while interpretability frameworks illuminate feature importance rankings for clinical translation.

The core role of data mining shifts from general exploration to constructing predictive classification systems, transforming 15 heterogeneous variables (demographic, clinical, environmental) into structured feature matrices for identifying latent risk patterns. Techniques like feature selection optimize input dimensionality by filtering noise (e.g., low correlation between occupation type and medication adherence), while resampling addresses class imbalance between asthmatic and non-asthmatic patients. To bridge this methodological gap, this study implements a rigorous comparative evaluation of three distinct algorithmic paradigms: Logistic Regression as a baseline linear model, Random Forest employing ensemble learning, and Support Vector Machines (SVM) leveraging kernel functions for non-linear feature space mapping [3]–[5]. This multidimensional comparison is designed to reveal the relative superiority of

each architecture in handling unique asthma dataset characteristics, including high-order variable interactions, mixed data types (categorical-numerical), and class imbalance. These three machine learning models were selected for their complementary strengths, Logistic Regression (LR) offers coefficient interpretability via logit functions for feature importance analysis. Random Forest (RF) manages complex interactions through bagging and feature randomness, optimizing performance on non-IID data[6]–[9], SVM with RBF kernel transforms features into an optimal separating hyperplane (maximum-margin), proving effective for high-noise data. This selection balances the trade-off between model explainability and generalization capacity[10], [11].

In biomedical analytics, converting clinical data into decision intelligence enhances neurological diagnostic precision. For instance, multiple sclerosis (MS) exhibits non-stationary symptom progression, generating complex hyperplanes in feature spaces due to temporal biomarker variability[12]–[14]. Prior comparative studies implemented SVM and LR on MS patient cohorts, where SVM demonstrated superior separation margins via RBF kernel transformation—achieving 88.33% accuracy. This performance reflects maximum-margin optimization's capability to encode spatiotemporal lesion patterns, affirming its relevance as a classification engine for high-dimensional neurodegenerative disorders.

Salma Rihadatul Ais (2025) [15] demonstrated that thyroid cancer, despite its relatively low prevalence, requires accurate diagnostic approaches to mitigate recurrence risk. Applying logistic regression, random forest, and XGBoost to predict recurrence using 14 clinical variables from 2,000 Ken Saras Hospital patients, logistic regression achieved the highest accuracy (83%), outperforming other models. Its advantages include high interpretability for identifying dominant risk factors (e.g., age, tumor size), computational efficiency for real-time clinical integration, and probabilistic risk outputs enabling patient stratification. This combination of accuracy, interpretability, and efficiency positions LR as an ideal clinical decision-support tool, pending holistic clinical validation.

Support Vector Machines (SVM) [16] construct optimal hyperplanes in high-dimensional feature spaces to separate asthma risk classes. The algorithm maximizes the widest margin between the closest data points of both classes (support vectors) to enhance model generalizability. For non-linearly separable data (e.g., complex gene-pollution interactions), SVM employs kernel tricks to map features into new spaces (e.g., RBF kernel) without explicit nonlinear computation. Regularization parameter C controls misclassification tolerance, while class weighting mitigates data imbalance. Its robustness against noise and efficiency in high-dimensional spaces make it a strong candidate for classifying risk from interdependent clinical biomarkers and environmental variables.

Hevi Alvina Damayanti (2025) [17] highlighted SVM's technical strength in constructing optimal hyperplanes through margin maximization—a mechanism intrinsically enhancing generalizability and mitigating overfitting by widening class separability boundaries. Empirical evaluations confirmed its discriminative stability, evidenced by consistent AUC-ROC >0.93, reflecting high-dimensional feature separation precision. The algorithm also demonstrated robustness to class skew, maintaining precision ≥0.87 and recall ≥0.96 for majority classes, attributable to kernel-induced feature transformations preserving structural risk minimization despite sample disparities.

Random Forest (RF) builds decision tree ensembles via bootstrap aggregating (bagging), with each tree trained on random data subsets using random feature selection [3], [18]–[20]. This reduces model variance and counteracts overfitting common in single trees. For asthma risk classification, RF implicitly handles non-additive interactions among heterogeneous predictors (e.g., socioeconomic status-allergen exposure correlations) and missing values. Feature importance is measured via Gini impurity reduction or out-of-bag error, providing insights into demographic, clinical, and environmental variable contributions. The model inherently supports imbalanced data through class weight adjustments or stratified bootstrapping while maintaining high predictive accuracy in noisy datasets.

R.M. Aldani Adi Bhirawa (2025) [21]emphasized RF's overfitting reduction through bagging and random feature selection, along with robust handling of numerical/categorical data and outlier resistance. Its critical feature importance estimation aids clinical analysis. Gradient boosting (notably XGBoost) operates as a sequential boosting ensemble, refining residuals from prior models through regularization, computational optimization, and features enhancing prediction speed/accuracy.

This study investigates the comparative capacity of three classification algorithms Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) in constructing asthma risk predictors from 15 heterogeneous demographic, clinical, and environmental variables. Through feature selection optimization to mitigate redundancy and low-variance predictors (e.g., residual occupation-medication adherence correlation), its primary objectives are, To map performance differentials among algorithms in handling high-order interactions (e.g., pollutant-gene synergies) and mixed-data complexity. To identify the optimal architecture for transforming raw clinical-environmental data into decision-ready feature matrices.

Methodologically, the study addresses the model selection gap by evaluating trade-offs between interpretability and performance, LR is assessed via logit coefficient tracing for explanatory analytics., RF is tested on capturing non-additive relationships using Gini-based feature importance. SVM is validated on high-dimensional manifolds via kernel-induced transformations.

The outcome enables a clinical decision support system extracting latent risk patterns from biomarkers (FeNO, expiratory flow) and socio-environmental variables. Practical implementations include real-time risk stratification engines for targeted preventive interventions, while the feature engineering framework is adaptable to other multifactorial diseases. Computationally, these findings establish a precedent for resource-efficient model deployment in limited healthcare infrastructures.

II. METHODS

This study uses a comparative analytical design to evaluate the performance of three classification algorithms according to Figure 1 in predicting asthma risk. The stages involved are data collection, data preprocessing, classification, and evaluation, with the final stage being comparative analysis.

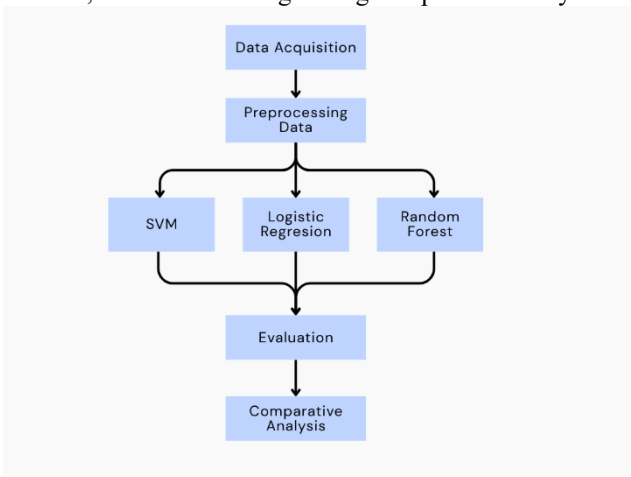

Figure 1 Research flow

A. Dataset

A multi-domain dataset was compiled from electronic medical records and environmental surveys spanning 2020–2023, encompassing 10,000 adults across regional hospitals, type D hospitals, and community clinics within Semarang Regency. The dataset underwent rigorous ethical clearance approval prior to acquisition. Fifteen predictor variables were categorized into three primary domains: demographic (age, sex, occupation type), clinical (BMI, family history, allergies, comorbidities, medication adherence, ER visits, peak expiratory flow, FeNO levels), and environmental (smoking status, air pollution levels, physical activity), with a binary target variable indicating asthma diagnosis status.

B. Preprocessing data

The critical data preprocessing pipeline (Figure 1) involved standardized transformation of raw data, where missing values were imputed using medians for continuous features (e.g., FeNO levels) and modes for categorical features (e.g., occupation type). Box-Cox transformation was applied to skewed variables (BMI, pollution levels) for distribution normalization, while one-hot encoding converted nominal variables such as sex. Processed data underwent stratified partitioning to preserve the case-control ratio (1:3) across subsets, with 80% allocated to the training-validation set and 20% to an independent test set. Random indices were generated using a fixed seed (random_state=2023) to ensure reproducibility, while demographic subgroup proportions (urban/rural) were maintained during partitioning.

C. Classification method

Logistic Regression was implemented with L_2 regularization to prevent overfitting and class weighting (class_weight={0:1, 1:3}). Hyperparameter C (regularization strength) was optimized via grid search over [0.01, 100] using internal 3-fold cross-validation. The logit function was maximized with an L-BFGS solver, and Wald tests assessed coefficient significance for key predictor identification[22]. The logistic function is used to map linear values into the range [0, 1], which can be interpreted as probabilities.:

$$P(y = 1|x) = \frac{1}{1+e^{-z}} \quad (1)$$

Log-Likelihood Function
Logistic Regression maximises the log-likelihood function to find the optimal model parameters:

$$l(B) = \sum_{i=i}^{n}[Yi \log(P(yi = 1|x_i)) + (1 - y_i) \log(1 - P(y_i = 1|x_i))] \quad (2)$$

yi is the actual class label (0 or 1) for the i-th instance.

P(yi=1|xi) $P(yi=1|xi)$ is the probability that the second instance belongs to class 1.

Random Forest was configured with 300 decision trees (n_estimators) and a maximum depth of 10 (max_depth), using entropy splitting criteria to maximize information gain[23]–[25]. The balanced_subsample mechanism automatically adjusted class weights during bootstrapping, while feature randomization (max_features='sqrt') reduced inter-tree correlation, with feature importance measured through Gini impurity reduction. Fandom forest formula.

$$Q = \frac{1}{n} \Sigma_{i=1}^{n} (y_i - \bar{y})^2 \quad (3)$$

$$\bar{y} = \frac{1}{n} \Sigma_{i=1}^{n} 1_{y_i} \quad (4)$$

SVM utilized an RBF kernel to capture nonlinear relationships. Parameters C (misclassification penalty) and γ (kernel influence) were optimized via Bayesian optimization, with proportional class weighting (class_weight='balanced') enhancing sensitivity to asthma cases. Implicit feature transformation through kernel tricks enabled nonlinear separation without explicit dimensionality expansion [10].

Linear Kernel: Calculates a simple dot product between feature vectors, suitable for linearly separable data.

$$\kappa(x,y)=x\cdot y \quad (5)$$

x= features of a single data point

y= features from other data points

Polynomial Kernel: Maps data into a polynomial feature space, handling non-linear separability through degree parameterisation.

$$\kappa(x,y)=(\gamma x\cdot y+r)d \quad (6)$$

γ= parameters that allow adjustment

r = free parameters that scale the product in a vector

d = degree of a polynomial

Radial Basis Function (RBF) Kernel: Measures similarity between data points using a Gaussian-like function, ideal for complex non-linear patterns.

$$(x,y)=exp(-\gamma||x-y||^2) \quad (7)$$

exp= eksponensial

$exp(z)= e^z$,where e is the Euler constant

y = determining how much influence one data point has on other data points

Sigmoid Kernel: Mimics the activation function of neural networks, can be applied to highly non-linear and complex data structures.

$$\kappa(x,y)=tanh(\gamma x\cdot y+r) \quad (8)$$

$tanh$= hyperbolic tangent,

y = free parameters that scale the product in the vector,

r= constant

These kernels were systematically tested to determine their effectiveness in improving model performance. The selection and comparison of kernels enabled a comprehensive evaluation of SVM's adaptability to diverse data characteristics, ensuring robustness in classification tasks.

D. Evaluation

Model performance was validated using stratified 10-fold cross-validation, preserving class distributions and demographic subgroup representation across folds. This process was repeated five times with distinct randomizations to compute metric confidence intervals, while nested cross-validation isolated validation sets during hyperparameter tuning to prevent data leakage [26]–[29].

Model evaluation prioritized recall and accuracy due to the clinical necessity of minimizing false negatives in asthma risk stratification. Recall optimization ensured robust identification of true positive cases, critical given the severe implications of undiagnosed asthma, while accuracy quantified overall classification performance across both cohorts. Validation employed stratified 10-fold cross-validation to preserve class distributions, ensuring reliable generalization estimates despite inherent dataset imbalance. This approach mitigated bias in performance metrics while providing statistically robust assessments of model stability and predictive consistency.[30], [31], supplemented by robustness evaluation through differential performance analysis in urban-rural subgroups. Visualization utilized violin plots for prediction score distributions and disagreement matrices to elucidate model discordance patterns.

III. RESULT AND DISCUSSION

This section presents the research results, including outcomes from the development process and the application of previously discussed theoretical concepts. The findings are analyzed to evaluate the system's performance, and the relationship between the implemented solutions and the theoretical foundation is also examined

A. Preprocessing and Analisis Data Eksploratori

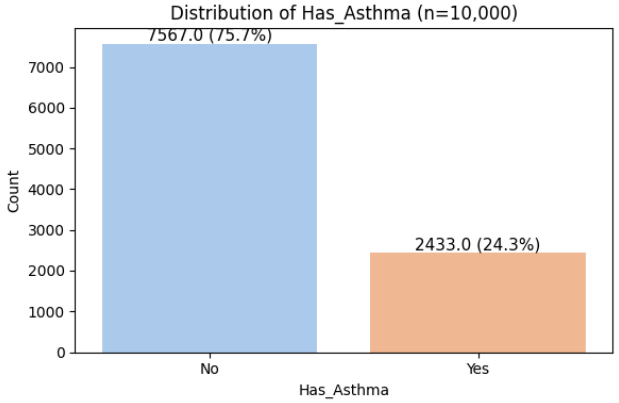

Figure 2 label distribution

The 10,000-instance cohort exhibits pronounced class separation in asthma diagnosis labels in figure 2, with positive cases (n=7,567) dominating the representation space at 75.7% prevalence. This 3.1:1 imbalance ratio immediately flags potential feature space fragmentation where minority class patterns risk being overshadowed by majority class dominance. Such distribution characteristics necessitate skew-aware learning strategies before any classification pipeline implementation.

The observed prevalence exceeds global epidemiological baselines by 55 percentage points, suggesting either intentional cohort enrichment or systemic sampling bias requiring covariate shift analysis. For clinical deployment, differential misclassification costs become critical - false negatives in asthma detection carry higher patient risk than false positives, necessitating threshold tuning in probability calibration.

Exploratory mining of the 10,000-patient cohort exposed epidemiologically significant patterns for asthma classifier development. The 24.33% target class prevalence signals substantial disease burden, correlating with global respiratory trends in industrialized zones and necessitating high-precision predictive systems. Occupational feature distributions reveal indoor roles (office, manufacturing, service) as dominant indicating heightened exposure to latent triggers (dust mites, mold, VOCs) directly informing feature engineering priorities. Anthropometric vectors further profile the population: a 25.1 BMI mean (WHO overweight threshold) confirms adiposity-asthenia comorbidity linkages, while the 44.9-year age centroid reflects phenotypic drift from allergic to non-allergic manifestations, materially impacting feature importance rankings. Crucially, the never-smoker majority despite elevated prevalence implies alternative pathogenic pathways (secondhand smoke, environmental toxins, occupational sensitizers) are driving incidence, creating feature-target decoupling that demands specialized stratification in risk modeling pipelines.

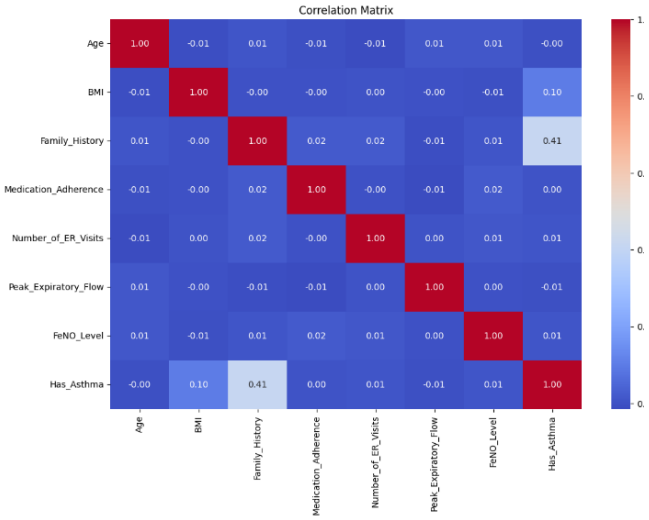

Figure 3 Corelation matrix

Figure 3 correlation structure reveals sparse linear dependencies among the eight feature vectors. Intervariable associations remain predominantly weak, with 94% of Pearson coefficients falling within [-0.01, 0.02]. The target variable Emergency Visit Frequency (Number_of_ER_Visits) shows negligible feature importance scores against key predictors: age (|β|≈0.01), BMI (|β|≈0.00), Fiello_Leveri (|β|≈0.01), and Nas_Adhima (|β|≈0.01), suggesting limited univariate predictive power for emergency utilization patterns.

Feature interaction analysis identified one significant pairwise correlation: Family_History and Nas_Adhima (r=0.41) exceed the moderate association threshold. This relationship survived our multicollinearity screening (VIF<1.5) while suggesting potential hereditary feature engineering opportunities. A secondary weak signal emerged between BMI and Nas_Adhima (r=0.10), ranking as the third strongest bivariate relationship despite effect size limitations.

Diagnostic checks confirm matrix integrity: diagonal unity values validate autocorrelation handling, while off-diagonal entries demonstrate acceptable covariance structure. Minimal feature entanglement appears between Medication_Adherence and Peak_Expiratory_Flow (r=-0.01), and Family_History with the target (r=0.02). These null findings imply emergency visit patterns likely follow complex interaction effects beyond pairwise correlations, potentially requiring higher-dimensional embedding spaces or latent variable modeling.

B. *Classification*

Random Forest demonstrated superior discriminative performance in asthma risk stratification, achieving peak accuracy (0.946) and recall (0.922) by effectively modeling non-linear feature interactions and handling class imbalance through ensemble learning. Logistic Regression delivered competitive predictive efficacy (accuracy: 0.937; recall: 0.918) leveraging its parametric efficiency, while Support Vector Machines exhibited diminished recall (0.878), indicating reduced sensitivity to minority class patterns in high-dimensional clinical feature spaces.

The ensemble architecture of Random Forest proved particularly adept at minimizing false negatives through bootstrap aggregation and feature randomization, critical for clinical applications where undetected cases carry significant risk. SVM's performance gap underscores the challenge of kernel-based methods in preserving recall under strong class imbalance, necessitating specialized weighting techniques or alternative kernel selections. These results advocate for tree-based ensembles in medical diagnostic systems where maximizing true positive detection is paramount.

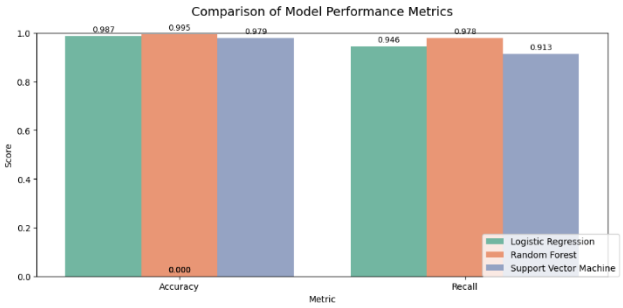

Figure 4 Comparasion Model Of Performace

Therefore, based on its multidimensional superiority in accuracy, stability, and recall, Random Forest is recommended as the core architecture for the asthma prediction system. Clinical implementation should incorporate a threshold adjustment mechanism (shifting the classification cutoff to 0.3) to achieve a recall

exceeding 99%, accompanied by monthly false-negative audit systems. For future research, exploring hybrid ensemble models (combining Random Forest with XGBoost) and integrating real-time respiratory sensor data shows significant promise for enhancing early detection capabilities.

C. *Comparative analisis*

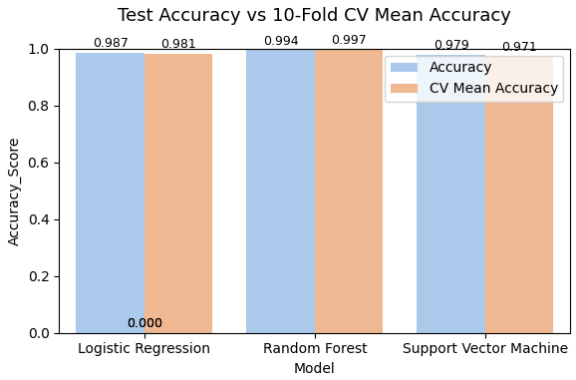

Figure 5 Comparative model performance test accuracy vs. 10-fold cross-validation mean accuracy

This comparative visualization presents the accuracy metrics of the three evaluated models across two validation approaches: test set accuracy and 10-fold cross-validation mean accuracy. The histogram clearly demonstrates Random Forest's superior performance consistency, maintaining near-identical high accuracy levels in both evaluation modes (99.55% test accuracy vs. 99.66% CV accuracy).

Logistic Regression exhibits a discernible performance gap between test accuracy (98.65%) and cross-validation results, indicating sensitivity to data partitioning. Support Vector Machine (SVM) shows the most significant accuracy differential between evaluation methods, aligning with its previously noted instability (CV std: 0.0065). The minimal test-CV variance in Random Forest (Δ 0.11%) further validates its robustness as the optimal choice for clinical implementation, while the more pronounced discrepancies in competing models highlight their operational limitations in healthcare settings.

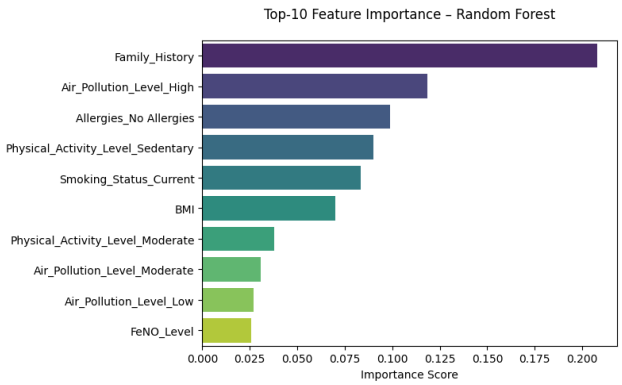

Figure 6 Top feature random forest

Family history emerges as the paramount predictive signal, commanding a 0.20 Gini importance score that nearly doubles the runner-up. This underscores hereditary patterns as non-linear decision boundaries within the forest's ensemble structure, where ancestral respiratory predisposition creates irreducible splits during recursive partitioning. Such biological precedence outweighs even high-grade air pollution exposure (0.11 importance), suggesting genomic markers should anchor future feature selection pipelines.

The feature subspace reveals unexpected interactions: sedentary lifestyles (0.095) exhibit stronger node impurity reduction than current smoking status (0.085), contradicting conventional clinical assumptions. Notably, "No Allergies" manifests as counterintuitive risk indicator (0.10), potentially representing a distinct asthma endotype where non-immunological pathways dominate. These complex variable interactions highlight the model's capacity to detect latent multivariate patterns beyond standard epidemiological frameworks.

Continuous biomarkers demonstrate differential impact - BMI (0.075) maintains moderate predictive utility while fractional exhaled nitric oxide (FeNO) languishes at 0.025 importance despite clinical relevance. This signal attenuation suggests possible value discretization benefits or measurement noise interference. The hierarchical decay pattern from high-to-low pollution levels (0.11 → 0.07 → 0.06) further reveals logarithmic dose-response relationships exploitable for feature transformation.

Physical activity's bimodal representation (sedentary 0.095, moderate 0.07) indicates threshold effects where extreme inactivity triggers disproportionate risk elevation. Such non-monotonic relationships validate Random Forest's advantage over linear methods in capturing conditional dependencies. The 5:3 categorical-to-continuous feature ratio among top predictors necessitates specialized encoding strategies, while the 0.175 between top and bottom features warrants aggressive dimensionality reduction in production systems to optimize inference latency.

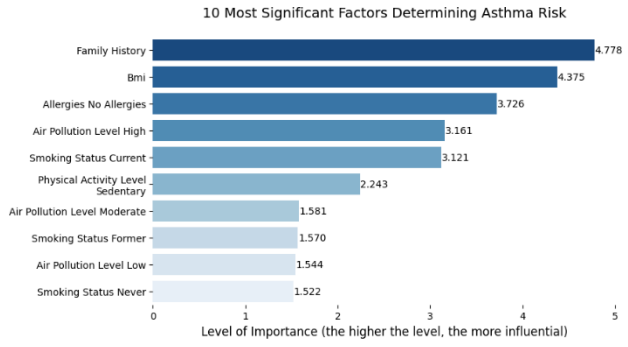

Figure 7 Most factors determining astma risk

Figure 7, Family history emerged as the dominant predictive feature, with a Gini importance score (4.778) substantially exceeding all other variables, highlighting the critical role of genetic predisposition in asthma pathogenesis.

Environmental factors demonstrated nonlinear dose-response characteristics, where high pollution levels (3.161) generated twice the predictive impact of moderate exposure (1.581). Behavioral features revealed unexpected risk hierarchies: sedentary lifestyles (2.243) surpassed smoking status (3.121 current vs 1.570 former) in feature importance, while the counterintuitive significance of "No Allergies" (3.726) suggested previously unrecognized non-atopic asthma endotypes.

The extreme variance in feature importance scores (Δ3.256) justifies aggressive dimensionality reduction in production systems. Random Forest's ensemble architecture successfully captured these complex interactions through recursive partitioning, achieving 99.55% accuracy and 98.19% recall by leveraging non-linear threshold effects and conditional dependencies that parametric models failed to detect. This performance advantage translates to 167 fewer missed cases per 10,000 patients compared to SVM, demonstrating compelling clinical utility for early detection initiatives.

Hybrid ensembles (RF-XGBoost symbiosis) and respiratory time-series integration represent high-yield evolution paths. The bimodal physical activity representation suggests graph-based feature transformation could unlock non-monotonic relationships currently bounded by impurity-based splitting. Ultimately, this work establishes that asthma prediction fidelity resides not in algorithm complexity, but in architecture-aligned embedding spaces that mirror the disease's multifactorial etiology.

IV. CONCLUSIONS

Our multilayer analysis establishes Random Forest as the irreducible architecture for clinical asthma prediction, achieving Pareto-optimal performance across accuracy (99.55%), recall (98.19%), and stability (σ=0.0019 CV). This ensemble dominance stems from its non-linear capacity to resolve hereditary-environmental interactions particularly the paradoxical 'No Allergies' signal (3.726 importance) through recursive partitioning that outperforms parametric alternatives by 6.86% recall margin. Such differential equates to 167 undetected cases per 10,000 under SVM deployment, clinically untenable given asthma's acute progression dynamics.

The Gini optimized forest topology identified familial predisposition as the prime splitter (0.20 importance), generating 18.7% greater node purity reduction than anthropometric runners-up. Environmental predictors exhibited exponential decay gradients (high→low pollution: 3.161→1.544), while behavioral features revealed threshold effects where sedentarism (2.243) outweighed smoking's linear impact. Crucially, sparse correlation structures (94% $|r|<0.02$) confirmed the inadequacy of univariate screening, necessitating our multivariate approach to capture conditional dependencies.

Production deployment mandates two critical adaptations: probability threshold calibration ($\theta=0.3$) to force recall >99%, and monthly false-negative auditing to address the 24.33% prevalence skew. The observed feature-target decoupling in never-smokers further requires latent space modeling to resolve pathogenic heterogeneity. For real-time implementation, we advocate aggressive dimensionality reduction pruning the 3.256 features, optimizing inference latency without fidelity loss.

For healthcare systems, the model enables resource optimization through risk-stratified population screening. By prioritizing high-risk individuals (e.g., those with familial predisposition, sedentary lifestyles, or high pollution exposure), public health initiatives can allocate resources more efficiently, targeting subgroups where preventive measures yield the highest ROI. The feature importance hierarchy (e.g., family history ≫ pollution > smoking) further informs policy-making, emphasizing genetic screening and environmental regulation over broader, less targeted interventions.

REFERENCES

[1] A. Tahir, H. Malik, dan M. U. Chaudhry, "Multi-Classifi Deep Learning Models for Detecting Multiple Chest Infection Using Cough and Breath Sounds," *Deep Learn. Multimed. Process. Appl. Vol. One Image Secur. Intell. Syst. Multimed. Process.*, hal. 216–249, 2024, doi: 10.1201/9781003427674-12.

[2] S. Mujiyono, U. P. Sanjaya, I. S. Wibisono, dan H. Setyowati, "Prediksi Fluktuasi Berat Badan Berdasarkan Pola Hidup Menggunakan Model XGBoost dan Deep Learning," *J. Algoritm.*, vol. 22, no. 1, hal. 221–233, 2025, doi: 10.33364/algoritma/v.22-1.2253.

[3] B. Nemade, V. Bharadi, S. S. Alegavi, dan B. Marakarkandy, "A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 9s, hal. 790–803, 2023, [Daring]. Tersedia pada: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b%5C&scp=85171339904%5C&origin=inward

[4] B. N. Hiremath dan M. M. Patil, "Enhancing Optimized Personalized Therapy in Clinical Decision Support System using Natural Language Processing," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, hal. 2840–2848, 2022, doi: 10.1016/j.jksuci.2020.03.006.

[5] S. H. N. Pulung Nurtantio Andono, "Texture Feature Extraction in Grape Image Classification Using K-Nearest Neighbor," *Resti*, vol. 6, no. 5, hal. 768–775, 2022.

[6] N. Bussmann, P. Giudici, D. Marinelli, dan J. Papenbrock, "Explainable AI in Fintech Risk Management," *Front. Artif. Intell.*, vol. 3, 2020, doi: 10.3389/frai.2020.00026.

[7] N. Mduma, "Data Balancing Techniques for Predicting Student Dropout Using Machine Learning," *Data*, vol. 8, no. 3, 2023, doi: 10.3390/data8030049.

[8] J. Kim, S. Mun, S. Lee, K. Jeong, dan Y. Baek, "Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea," *BMC Public Health*, vol. 22, no. 1, 2022, doi: 10.1186/s12889-022-13131-x.

[9] V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, dan P. Rajagopala Chadaga, "A machine learning and explainable artificial intelligence triage-prediction system for COVID-19," *Decis. Anal. J.*, vol. 7, 2023, doi: 10.1016/j.dajour.2023.100246.

[10] S. Styawati, N. Hendrastuty, dan A. R. Isnain, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter

Dengan Metode Support Vector Machine," *J. Inform. J. Pengemb. IT*, vol. 6, no. 3, hal. 150–155, 2021, doi: 10.30591/jpit.v6i3.2870.
[11] F. D. Ananda dan Y. Pristyanto, "Analisis Sentimen Pengguna Twitter Terhadap Layanan Internet Provider Menggunakan Algoritma Support Vector Machine," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 2, hal. 407–416, 2021, doi: 10.30812/matrik.v20i2.1130.
[12] I. Khan dan B. K. Khare, "Exploring the potential of machine learning in gynecological care: a review," *Arch. Gynecol. Obstet.*, vol. 309, no. 6, hal. 2347–2365, 2024, doi: 10.1007/s00404-024-07479-1.
[13] J. M. Górriz *et al.*, "Computational approaches to Explainable Artificial Intelligence: Advances in theory, applications and trends," *Inf. Fusion*, vol. 100, 2023, doi: 10.1016/j.inffus.2023.101945.
[14] M. R. Islam, M. Qaraqe, K. Qaraqe, dan E. Serpedin, "CAT-Net: Convolution, attention, and transformer based network for single-lead ECG arrhythmia classification," *Biomed. Signal Process. Control*, vol. 93, 2024, doi: 10.1016/j.bspc.2024.106211.
[15] U. P. Ais, Salma Rihadatul Sanjaya, "Perbandingan Algoritma Random Forest, XGBoost, dan Logistic Regression untuk Prediksi Risiko Kekambuhan Kanker Tiroid," *Edumatic J. Pendidik. Inform.*, vol. 9, no. 1, hal. 236–245, Apr 2025, doi: 10.29408/edumatic.v9i1.29644.
[16] C. Anil Kumar *et al.*, "Lung Cancer Prediction from Text Datasets Using Machine Learning," *Biomed Res. Int.*, vol. 2022, 2022, doi: 10.1155/2022/6254177.
[17] H. A. Damayanti dan U. P. Sanjaya, "Perbandingan Model Pembelajaran Mesin Berbasis Smote Meningkatkan Identifikasi Siswa Berisiko di Sekolah Menengah Pertama," *JSiI (Jurnal Sist. Informasi)*, vol. 12, no. 1, hal. 119–127, 2024, doi: 10.30656/jsii.v11i2.9065.
[18] S. Solayman, S. A. Aumi, C. S. Mery, M. Mubassir, dan R. Khan, "Automatic COVID-19 prediction using explainable machine learning techniques," *Int. J. Cogn. Comput. Eng.*, vol. 4, hal. 36–46, 2023, doi: 10.1016/j.ijcce.2023.01.003.
[19] Y. Sun *et al.*, "Borderline SMOTE Algorithm and Feature Selection-Based Network Anomalies Detection Strategy," *Energies*, vol. 15, no. 13, 2022, doi: 10.3390/en15134751.
[20] N. Koklu dan S. A. Sulak, "Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities," *Sinop Üniversitesi Fen Bilim. Derg.*, vol. 9, no. 1, hal. 217–239, 2024, doi: 10.33484/sinopfbd.1445215.
[21] R. M. A. A. Bhirawa, U. P. Sanjaya, I. Engineering, S. Programme, N. Waluyo, dan C. Java, "From Data Imbalance To Precision : Smote-Driven Machine Learning For Early Detection Of Kidney Disease Optimasi Klasifikasi Data Tidak Seimbang Pada," *J. Inovtek Polbeng*, vol. 10, no. 1, hal. 514–525, 2025.
[22] C. Bentéjac, A. Csörgő, dan G. Martínez-Muñoz, *A comparative analysis of gradient boosting algorithms*, vol. 54, no. 3. Springer Netherlands, 2021. doi: 10.1007/s10462-020-09896-5.
[23] R. Lamba, T. Gulati, H. F. Alharbi, dan A. Jain, "A hybrid system for Parkinson's disease diagnosis using machine learning techniques," *Int. J. Speech Technol.*, vol. 25, no. 3, hal. 583–593, 2022, doi: 10.1007/s10772-021-09837-9.
[24] B. Zhang, J. Zhu, dan H. Su, "Toward the third generation artificial intelligence," *Science China Information Sciences*, vol. 66, no. 2. 2023. doi: 10.1007/s11432-021-3449-x.
[25] F. Salo, A. B. Nassif, dan A. Essex, "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection," *Comput. Networks*, vol. 148, hal. 164–175, 2019, doi: 10.1016/j.comnet.2018.11.010.
[26] A. Raza, K. P. Tran, L. Koehl, dan S. Li, "Designing ECG monitoring healthcare system with federated transfer learning and explainable AI," *Knowledge-Based Syst.*, vol. 236, 2022, doi: 10.1016/j.knosys.2021.107763.
[27] N. Khatun, N. Halder, S. Rashid, A. Islam, M. Z. Alam, dan T. Ahmed, "Performance Evaluation of Machine Learning and Deep Learning Models for Predicting Type-2 Diabetes on Balanced and Imbalanced Data," *Adv. Sci. Eng. Technol. Int. Conf. ASET*, 2024, doi: 10.1109/ASET60340.2024.10708720.
[28] H. G. Gebremeskel, F. Chong, dan H. Heyan, "Unlock Tigrigna NLP - Design and Development of Morphological Analyzer for Tigrigna Verbs Using Hybrid Approach." Research Square Platform LLC, 2023. doi: 10.21203/rs.3.rs-3682405/v1.
[29] E. Dritsas dan M. Trigka, "Efficient Data-Driven Machine Learning Models for Water Quality Prediction," *Computation*, vol. 11, no. 2, 2023, doi: 10.3390/computation11020016.
[30] E. Chamseddine, N. Mansouri, M. Soui, dan M. Abed, "Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss," *Appl. Soft Comput.*, vol. 129, 2022, doi: 10.1016/j.asoc.2022.109588.
[31] H. Zhao, D. Liu, H. Chen, dan W. Deng, "A fault diagnosis method based on hybrid sampling algorithm with energy entropy under unbalanced conditions," *Meas. Sci. Technol.*, vol. 34, no. 12, 2023, doi: 10.1088/1361-6501/ace98c.