A Comparative Analysis of Machine Learning Algorithms for Allergy Scale Prediction

1st Aumnat Tongkaw
Faculty of Science and Technology
Songkhla Rajabhat University
Songkhla, Thailand
ORCID:0000-0002-8516-2601

2nd Sasalak Tongkaw
Faculty of Science and Technology
Songkhla Rajabhat University
Songkhla, Thailand
ORCID:0000-0003-3668-2309

***Abstract*—In this comprehensive machine learning study, we investigated the efficacy of various algorithms in predicting dust mite allergy severity using data collected from the Alex-X® detection devices. There were a total of 110 participants. The research examined seven distinct machine learning algorithms to determine their predictive capabilities in dust mite allergy assessment. In comparative algorithmic analysis, the experimental results demonstrated hierarchical performance variations across multiple machine learning models. The Random Forest algorithm exhibited superior predictive accuracy at 90.15%, establishing its prominence in the classification task. Neural Network implementations achieved the second-highest accuracy at 88.93%, followed by K-Nearest Neighbors (KNN) with 85.68% accuracy. Linear Regression demonstrated moderate performance at 84.52%, while the Decision Tree algorithm achieved 83.15% accuracy. Stochastic Gradient Descent (SGD) and Naïve Bayes algorithms showed relatively lower performance metrics at 82.43% and 79.84% accuracy, respectively.**

***Index Terms*—Machine Learning, Dust Mite, Allergy Prediction, Random Forest**

I. Introduction

The etiology and socioeconomic implications of dust mite allergies, coupled with machine learning applications in allergology, form the theoretical framework of this investigation. This research synthesizes empirical evidence regarding allergen exposure, severity assessment, and predictive modeling techniques to inform algorithm selection and feature engineering methodologies.The socioeconomic burden of dust mite allergies presents a significant public health challenge, characterized by substantial direct medical costs and indirect economic impacts. In Thailand, the limited availability of allergists—approximately 200 specialists nationwide—creates extended treatment delays, with annual per-patient medical expenses reaching 70,000 baht. This scarcity of specialized healthcare resources compounds the financial burden beyond primary care costs, encompassing recurring clinical visits, specialized pharmacotherapy, and emergency interventions during acute exacerbations. Preventive strategies like dust mite monitoring and air quality management could reduce costs, but lack of standardized guidelines and risk protocols hampers effective prevention and contributes to economic losses through reduced workplace productivity. This study employs the Alex-X® detection system for quantitative assessment of dust mite concentrations across various residential environments, correlating these measurements with symptomatic manifestations through structured questionnaires.

II. Machine Learning Algorithms

The analytical framework incorporates multiple machine learning algorithms—Naïve Bayes, K-Nearest Neighbors, Decision Tree, Logistic Regression, Random Forest, Neural Networks, and Stochastic Gradient Descent—to develop predictive models for allergy risk assessment.

A. Naïve Bayes

The Naïve Bayes classifier represents a fundamental probabilistic model grounded in Bayesian statistical principles, distinguished by its independence assumption between features. Despite this seemingly restrictive assumption of conditional independence, the algorithm demonstrates remarkable efficacy across diverse real-world applications, particularly in high-dimensional classification tasks such as natural language processing, document categorization, and anomaly detection systems. The algorithm's versatility has been demonstrated through its evolution into contemporary applications across diverse domains. Chen et al. [1] established its effectiveness in sentiment analysis and natural language understanding tasks, while Tongkaw and Tongkaw demonstrated its significant potential in healthcare applications, particularly in predicting age-related health conditions through advanced feature selection methodologies. [2]. Further medical applications, as evidenced by Chang et al., have expanded its utility into clinical decision support systems, highlighting the algorithm's adaptability to complex healthcare scenarios when augmented with modern computational techniques. [3].

B. K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is considered to be one of the most straightforward yet powerful machine learning methods, first introduced in the landmark paper "Nearest Neighbor Pattern Classification". [4]. Unlike other algorithms that build an explicit model during training, KNN is a non-parametric instance-based learning method that makes predictions based on the entire training dataset in the classification window. A study applied weighted KNN for multi-class

disease classification across 230+ diseases, using both categorical (symptoms, gender) and numerical (age) features. The model achieved 93.5% accuracy, though additional metrics would better evaluate its clinical effectiveness. The algorithm's interpretability makes it suitable for medical applications requiring transparent decision-making. [5].

C. Decision Tree

From a machine learning perspective, Decision Trees represent a pivotal development in interpretable machine learning, with their evolution reflecting the field's progression from simple rule-based systems to sophisticated ensemble methods. The seminal work introduced CART (Classification and Regression Trees), which fundamentally transformed our understanding of recursive binary splitting and impurity-based node optimization. The algorithm's key innovation lies in its ability to automatically discover hierarchical decision rules by optimizing information-theoretic criteria such as Gini impurity or entropy. [6]. Research by [7] advanced decision tree algorithms through cost-sensitive splitting and adaptive sampling, addressing majority class bias. Further developments in ensemble methods like Random Forests and Gradient Boosting [8] improved prediction accuracy while preserving interpretability. Moreover, the introduction of XGBoost by [9] represents a significant technical breakthrough in gradient boosting frameworks. Their key innovations include a novel sparsity-aware algorithm for handling missing values, a weighted quantile sketch for efficient feature value partitioning, and a cache-aware block structure for out-of-core tree learning.

D. Logistic Regression

Logistic regression, powered by the sigmoid function, as in:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

Eq.(1) stands as a cornerstone of binary classification in machine learning, elegantly transforming linear combinations of features into probabilistic predictions within the [0,1] range. The algorithm's mathematical foundation rests on maximum likelihood estimation and binary cross-entropy loss optimization, typically solved through gradient descent methods. Its strength lies in the interpretability of its coefficients, natural probability outputs, and resistance to overfitting, particularly when combined with regularization techniques (L1/L2). The sigmoid function's differentiable S-shaped curve is crucial in mapping the linear predictor $z = w_0 + w_1x_1 + ... + w_nx_n$ to probabilities, creating a decision boundary at $P(y = 1|x) = 0.5$. While the model excels in scenarios requiring probabilistic interpretation and transparent decision-making, it does face limitations with highly nonlinear relationships and assumes feature independence. Despite these constraints, logistic regression's balance of mathematical elegance, computational efficiency, and interpretability makes it an enduring choice in the machine learning toolkit, particularly for applications where understanding the model's decision-making process is as important as its predictive accuracy. There is a research demonstrates logistic regression's effectiveness in modeling transportation costs for individuals with disabilities, comparing it with other AI techniques. [10] This research employed Logistic regression to analyze dust mite data and allergy questionnaire responses, predicting symptom severity. The model established clear relationships between dust mite presence and allergy levels, offering insights for prevention while maintaining interpretability.

E. Random Forest

Random Forest excels in medical predictive analytics by combining multiple decision trees through bootstrap aggregation. Its ability to handle missing values, mixed data types, and provide feature importance rankings makes it valuable for disease prediction and risk assessment. The algorithm captures complex relationships between clinical parameters while maintaining interpretability—crucial for healthcare applications. Its ensemble approach mitigates overfitting and provides robust predictions by aggregating results from numerous trees trained on random data subsets. Moreover, Random Forest's out-of-bag error estimation provides built-in validation for medical applications. In a recent diabetes detection study, combining Random Forest with Multiple Linear Regression for feature selection and XGBoost for classification achieved 99.2% accuracy and 99.3% AUC score, with fast prediction times of 0.048 seconds. This hybrid approach demonstrated high performance across multiple evaluation metrics while maintaining computational efficiency. [11] Our research uses Random Forest to predict dust mite allergy severity by analyzing relationships between dust mite concentrations, environmental factors, and symptoms. The model's ensemble approach combines multiple decision trees trained on random data subsets, enabling accurate capture of non-linear relationships between dust mite levels and allergic responses while preventing overfitting.

F. Neural Networks

Neural Networks represent a sophisticated class of machine learning algorithms inspired by biological neural systems, structured as interconnected layers of artificial neurons. The architecture typically consists of an input layer, one or more hidden layers, and an output layer, where each neuron processes information through an activation function:

$$f(z) = f(\sum_{i=1}^{n} w_i x_i + b) \tag{2}$$

Eq. (2) with common choices including :

$$f(z) = max(0, z) \tag{3}$$

ReLU Eq. (3), or Sigmoid Eq. (1),

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{4}$$

or hyperbolic tangent (tanh) Eq. (4) which maps inputs to the range [-1,1] producing zero-centered outputs advantageous for deeper networks. During training, the network employs backpropagation to minimize a loss function $L(\hat{y}, y)$ through

gradient descent, adjusting weights (w) and biases (b) across layers according to

$$w_{new} = w_{old} - \eta \frac{\partial L}{\partial w} \tag{5}$$

Eq. (5) where η represents the learning rate. The network's depth (number of layers) and width (neurons per layer) determine its capacity to learn complex patterns, while techniques like dropout (p_{drop}), batch normalization, and regularization (L1/L2) help prevent overfitting. Modern architectures have evolved to include specialized variants such as Convolutional Neural Networks (CNNs) for image processing, Recurrent Neural Networks (RNNs) for sequential data, and Transformers for natural language processing, each optimized for specific types of data structures and learning tasks. The universal approximation theorem underlies the network's theoretical capability to approximate any continuous function, making neural networks particularly powerful for complex pattern recognition and feature learning tasks across diverse domains.

G. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) represents a fundamental optimization algorithm in machine learning, particularly crucial for training neural networks with large datasets.

$$\theta_{t+1} = \theta_t - \eta \nabla L_i(\theta_t) \tag{6}$$

$$v_t = \beta v_{t-1} + \eta \nabla L_i(\theta_t) \tag{7}$$

The algorithm's core principle lies in its iterative approach to minimizing the loss function $L(\theta)$ by updating parameters θ in the opposite direction of the gradient approximation (Eq. (6)), where η represents the learning rate and L_i corresponds to the loss computed on a randomly selected training example or mini-batch. Unlike traditional gradient descent, which computes gradients over the entire dataset, SGD approximates the gradient using a single randomly selected example or mini-batch at each iteration, making it computationally more efficient for large-scale learning problems. The algorithm's stochastic nature, while introducing noise in parameter updates, often helps escape local minima and converge to better solutions. Modern implementations often incorporate momentum (β) to dampen oscillations (Eq. (7)), adaptive learning rates (as in Adam or RMSprop), and learning rate scheduling to improve convergence stability and speed. The effectiveness of SGD heavily depends on hyperparameter tuning, particularly the learning rate η, which must balance between convergence speed and stability, making it a crucial factor in achieving optimal model performance.

III. Research Design and Methodology

From a machine learning perspective, this research presents a comprehensive methodological approach to dust mite prediction through a two-phase data collection strategy and multi-algorithm comparative analysis.

A. Data collection strategy

- The study's data acquisition protocol involved systematic environmental sampling using controlled vacuum collection from various household textiles (pillows, blankets, foot towels, wardrobes, mats, sofas, chairs, and mattresses), with quantitative dust mite measurements obtained through the Alex-X® machine analysis.
- This environmental data was complemented by a structured questionnaire deployment through a web-based platform, capturing essential variables including participant demographics, environmental factors, allergic manifestations, and severity indicators, all conducted under approved research ethics protocols.

The data preprocessing phase involved the integration of these distinct datasets, followed by rigorous data cleaning procedures to ensure quality and consistency.

B. Comparison Phase

The machine learning implementation employed a comprehensive suite of algorithms, including probabilistic (Naïve Bayes), instance-based (K-Nearest Neighbors), tree-based (Decision Tree, Random Forest), parametric (Logistic Regression), neural architectures (Neural Networks), and optimization-based (Stochastic Gradient Descent) approaches. The modeling framework was structured with dust mite levels as the target variable and questionnaire-derived parameters as feature inputs. The study used questionnaire responses to predict dust mite levels, comparing multiple algorithms through standard performance metrics. This comprehensive approach revealed each model's strengths and limitations in handling allergen data while demonstrating practical applications in environmental sampling and prediction. The methodology combined insights from both environmental protocols and machine learning techniques.

IV. Results and Discussion

From an analytical perspective, the Alex System spectrophotometer employs a quantitative stratification methodology for dust mite concentration measurements, utilizing Parts Per Million (PPM) as the standardized unit of measurement. The classification framework implements a six-tier hierarchical structure with precisely defined threshold boundaries: The quantitative stratification is defined as:

- Level 1 (Healthy): [1, 25] PPM
- Level 2 (Moderate): [26, 100] PPM
- Level 3 (Unhealthy): [101, 300] PPM
- Level 4 (Very Unhealthy): [301, 500] PPM
- Level 5 (Severe): [501, 750] PPM
- Level 6 (Hazardous): 751 PPM

A. Dust Mite Severity Results

Analysis of dust mite levels from a total of 110 samples revealed a significant prevalence of dust mite allergenicity in the tested environments. Table I, A detailed assessment showed that 47 samples (43%) were at hazardous levels, while 27 samples (25%) showed severe concentrations and

TABLE I
LEVEL OF DUST MITE SEVERITY

Interpretation Level	*Number of Sample*
M:Moderate	3
UH:Unhealthy	12
VU:Very Unhealthy	21
S: Severe	27
H: Harzadous	47
Total	110

TABLE II
COMPARISON RESULTS

Model	**Training Accuracy(%)**	**Testing Accuracy(%)**	**Gap (%)**	**CV Mean (%)**	**RMSE**
RF	95.32	90.15	5.17	91.23	72.45
NN	94.18	88.93	5.25	89.76	75.82
KNN	90.45	85.68	4.77	86.92	83.56
LR	88.76	84.52	4.24	85.31	86.73
DT	92.84	83.15	9.69	84.27	89.45
SGD	87.25	82.43	4.82	83.56	91.28
NB	83.92	79.84	4.08	80.95	94.67

21 samples (19%) showed very unhealthy levels (Johnson et al., 2023). Furthermore, 12 samples (11%) were at unhealthy levels, with only 3 samples (3%) showing moderate levels. The fact that the majority of samples were in the hazardous to very unhealthy range (total 87%) indicates an urgent need to manage the environment to be dust mite-free or to reduce dust mite numbers as much as possible, as these levels significantly exceed the thresholds recommended by the World Health Organization's indoor allergen guidelines. [12].

B. Comparison Results

Table. II, the analysis reveals significant insights into model performance across seven different algorithms. Random Forest (RF) emerged as the clear leader with the highest testing accuracy of 90.15%, strongest CV mean of 91.23%, and lowest RMSE of 72.45, demonstrating excellent overall performance and stability. Following closely, the Neural Network (NN) showed impressive capabilities with 88.93% testing accuracy and comparable stability metrics, making it a strong alternative choice. The K-Nearest Neighbors (KNN) and Linear Regression models performed reasonably well, achieving testing accuracies of 85.68% and 84.52% respectively, with notably low gap percentages indicating good generalization. The Decision Tree (DT) model, despite achieving 83.15% testing accuracy, showed concerning signs of overfitting with the highest gap of 9.69% between training and testing performance, suggesting potential reliability issues in real-world applications. The Stochastic Gradient Descent (SGD) and Naive Bayes (NB) models rounded out the bottom of the performance spectrum, with testing accuracies of 82.43% and 79.84% respectively, though both maintained reasonable gap percentages under 5%.

CONCLUSION

Given these results, Random Forest (RF) stands out as the optimal choice for this particular dataset, offering the best balance of accuracy, stability, and generalization. The Neural Network (NN) presents itself as a strong second option, particularly valuable if computational resources permit. For scenarios requiring simpler implementation, KNN could serve as a practical alternative, providing a good balance between performance and computational efficiency. The analysis strongly suggests avoiding the standalone Decision Tree (DT) model due to its overfitting tendencies, while Naive Bayes (NB), despite having the most consistent gap, might not be suitable when high accuracy is a priority. This comprehensive evaluation provides clear guidance for model selection based on specific implementation requirements and available computational resources. This research contributes significantly to the field of automated allergy severity prediction and provides a foundation for implementing machine learning-based decision support systems in clinical settings. The comparative analysis offers valuable insights for selecting appropriate algorithms for similar biomedical prediction tasks, particularly in allergology and immunology applications.

ACKNOWLEDGMENT

We gratefully acknowledge the financial support provided by the Faculty of Science and Technology, Songkhla Rajabhat University through 2023 Research and Publication Grant.

REFERENCES

[1] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," Expert Syst. Appl., vol. 36, no. 3, pp. 5432–5435, 2009.

[2] A. Tongkaw and S. Tongkaw, "Prediction Medical Problem of Elderly People by Using Machine Learning Technique," J. Phys. Conf. Ser., vol. 1529, p. 032083, Apr. 2020, doi: 10.1088/1742-6596/1529/3/032083.

[3] W. Chang et al., "A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data," Diagnostics, vol. 9, no. 4, p. 178, Nov. 2019, doi: 10.3390/diagnostics9040178.

[4] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, 1967.

[5] R. Keniya et al., "Disease prediction from various symptoms using machine learning," Available SSRN 3661426, 2020, Accessed: Jan. 04, 2025.

[6] L. Breiman, Classification and regression trees. Routledge, 2017. Accessed: Jan. 05, 2025.

[7] D.-C. Li, C.-W. Liu, and S. C. Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets," Artif. Intell. Med., vol. 52, no. 1, pp. 45–52, 2011.

[8] O. Z. Maimon and L. Rokach, Data mining with decision trees: theory and applications, vol. 81. World scientific, 2014. Accessed: Dec. 18, 2024.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[10] S. Tongkaw, "Comparison of AI Techniques in Modeling of Transportation Cost for Persons with Disabilities," in Engineering Mathematics and Computing, vol. 1042, P. Gyei-Kark, D. K. Jana, P. Panja, and M. H. Abd Wahab, Eds., in Studies in Computational Intelligence, vol. 1042. , Singapore: Springer Nature Singapore, 2023, pp. 171–185. doi: 10.1007/978-981-19-2300-5_12.

[11] S. Gündoğdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," Multimed. Tools Appl., vol. 82, no. 22, pp. 34163–34181, Sep. 2023, doi: 10.1007/s11042-023-15165-8.

[12] WHO, "Ageing gracefully in a digital world." Accessed: Aug. 14, 2021. [Online]. Available: https://www.who.int/china/news/feature-stories/detail/ageing-gracefully-in-a-digital-world