OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

The Role of Information Theory in Asthma Prediction

Elaine Peng
elaine.zijin.peng@gmail.com

ABSTRACT

Asthma is a chronic condition where early prediction is often hindered by complex interactions between genetic and environmental factors. Traditional feature selection methods usually rely on assumptions that might not fully capture the complexity of these interactions. This study investigates whether Information Gain (IG), a non-parametric entropy-based metric, can improve asthma prediction accuracy compared to traditional statistical methods. It was hypothesized that IG will outperform correlation, ANOVA, and logistic regression by more effectively capturing complex relationships seen in asthma development. Using a shared framework for prediction, four different feature selection protocols were compared against each other: correlation matrices, ANOVA F-tests, multivariable logistic regression, and Information Gain. The models using ANOVA and logistic regression mostly matched the baseline performance (F1: 0.83/0.84, AUC: 0.90). On the other hand, the Information Gain model achieved better results with an F1 score of 0.87 and an AUC of 0.93. It can possibly be concluded that entropy-based feature selection effectively identifies informative predictors in complex biological datasets to offer a potentially more effective feature selection approach for asthma disease prediction within this experimental framework.

INTRODUCTION

Asthma is a chronic lung disease that causes breathing difficulties associated with inflammation in the lungs and narrowing of the airways. In 2019, this disease affected around 262 million people from around the world. It also continues to be the most common chronic disease in children [1]. The condition can pose serious dangers which include potentially fatal asthma attacks that require immediate emergency care. When controlled poorly, asthma might create permanent lung damage, reduced lung function, and can negatively impact sleep and the ability to participate in physical activities. Resultingly, early identification of asthma is very important, since intervention during early immune development helps to prevent progression of the disease.

Effectively predicting a person's likelihood of contracting asthma at a young age based on other observable characteristics such as allergy histories and family histories can help to proactively manage the condition, allowing potential asthma patients to receive early treatment and intervention, preventing uncontrolled development of the condition. Previous attempts to do so have shown that this is a difficult task because of the complex interaction between genetic factors, unique triggers among individuals, and environmental influences. For instance, as summarized by Castro-Rodriguez et al., a study conducted on

95 children in Australia showed that airway responsiveness at the age of 1 month is an effective predictor of airway function at the age of 6 years, while on the other hand, a study conducted on 129 children in France showed that early bronchial hyperresponsiveness in wheezing infants is not an effective predictor of asthma later in life [2,3,4]. In addition, two studies showed that wheezing during the first three years was not an accurate predictor of later asthma, while early atopy is [2,5]. This complicated combination of factors that may or may not relate to the development of asthma suggests that the ability to select the most important and informative features that directly predict the onset of asthma is extremely important to the success of prediction models. Due to the lack of accurate, non-invasive measurements of airway inflation, early prediction and diagnosis of asthma often relies on subjective clinical features rather than standardized metrics [2].

For this reason, currently, accurate predictive modeling depends heavily on selecting the most informative features from clinical datasets. Feature selection is particularly important in biomedical contexts to reduce overfitting, improve interpretability and generalizability. Prior research has relied primarily on traditional statistical approaches. Blakey et al. utilized multivariate logistic regression to identify significant predictors [6], while Tsang et al. employed a feature-target correlation matrix to rank variables based on linear association strength [7]. Although effective in certain contexts, these methods rely on assumptions such as linear relationships between predictors and outcomes, proportional log-odds, and homoscedasticity. Such assumptions may not reflect the complexity of factors underlying asthma development.

To address these limitations, this study evaluates three predictive models based on Tsang's architecture, replacing the original correlation-based feature selection approach with ANOVA testing, multivariate logistic regression, and Information Gain (IG). With previously common feature selection methods for asthma prediction including univariate logistic regression and stepwise regression, Information Gain was often overlooked as a potential feature selection approach due to its requirement for discretization and its failure to capture dependencies between features, being a univariate method [8,9]. However, Information Gain also holds distinct advantages over the previously used parametric methods. Unlike correlation and regression-based methods, Information Gain quantifies the reduction in entropic uncertainty when a feature is observed, allowing it to capture nonlinear and distribution-free relationships between predictors and outcomes [10]. While overlooking dependencies between features, the ability of IG to capture nonlinear and distribution-free relationships can potentially outweigh its shortcomings. In addition, Mutual Information serves as an effective variant to address the existence of continuous variables. Because asthma development likely arises from complex, interacting mechanisms rather than simple linear associations, we hypothesize that IG will outperform correlation, ANOVA, and logistic regression in identifying predictive features. By comparing these approaches, this study aims to determine whether entropy-based feature selection methods provide measurable improvements in pediatric asthma risk prediction.

MATERIALS AND METHODS

The dataset used in this study was the cross-sectional Asthma Disease Dataset derived from Kaggle [11]. A total of 2392 patient records were analyzed, and the primary outcome was a binary diagnosis of asthma

(Yes/No), with 2268 patients being diagnosed with "No" and 124 diagnosed with "Yes". Categorical variables were encoded using one-hot encoding, and there were no missing values to be treated in the dataset. Initial features included demographic details, lifestyle factors, environmental and allergy factors, medical history, clinical measurements, and asthma-related symptoms. Data unrelated to asthma diagnosis, particularly patient ID and confidential information, were removed from the dataset.

This study used the predictive framework established by Tsang et al. [7]. To ensure consistency for the purpose of comparison, several parameters were held constant for all models. The algorithm used for all models was Random Forest and the data was partitioned into an 80/20 train-test split. The dataset showed significant class imbalance (5.2% positive class). This inspired the use of SMOTE. SMOTE was applied only to the training data in order to address this imbalance, using k=5 nearest neighbors, which deviates from the Tsang model in which SMOTE was applied to the entire dataset. The Random Forest classifier was employed with 100 trees, maximum depth unrestricted, Gini impurity criterion, random_state=42, along with 5-fold cross validation, which deviates from the single train-test split in the Tsang model. Python 3.x was used to perform the analysis along with Scikit-Learn. Four different feature selection methods were applied to the training data (this also deviates from the Tsang model, in which feature selection was applied to the entire dataset) in order to select the top ten features for the asthma prediction models. First, for the Correlation Matrix (Baseline) model, features were selected based on the highest absolute Pearson correlation coefficients relative to the target variable. For the ANOVA F-Test model, features were ranked according to their F-statistic, assuming homoscedasticity and a linear relationship between group means. For the Multivariable Logistic Regression model, features were selected based on the largest magnitude of coefficients (β values) after fitting a logistic model using L2 regularization and lbfgs solver. Features were standardized before applying Logistic regression. Finally, for the Information Gain (IG) model, features were selected based on the Mutual Information (MI) score,

$$I(X;Y) = H(X) + H(Y) - H(X,Y),$$

where $H(X)$ denotes entropy of feature X, and $H(X,Y)$ denotes joint entropy of the two features. Entropy was calculated using base-2 logarithms. MI was calculated using sklearn's mutual_info_classif, which estimates mutual information for both discrete and continuous predictors using a k-nearest neighbor–based entropy estimator.

The effectiveness of each feature selection method was evaluated using four metrics: precision, the ratio of true positive observations to the total predicted positives; recall (sensitivity), the ratio of true positive observations to all observations in the actual class; F1 score, the harmonic mean of precision and recall, providing a balance between the two; and area under the ROC Curve (AUC-ROC), measuring the model's ability to distinguish between classes across all possible thresholds. All performance metrics were computed on the test set.

RESULTS

To evaluate the impact of different feature selection techniques on asthma prediction, three alternative methods were integrated into the established Tsang model architecture. Model performance was evaluated

using precision, recall, F1 score, and Area Under the ROC Curve (AUC). The results are summarized in Table 1, shown below.

	correlation matrix (baseline)	ANOVA f-test	multivariable logistic regression	information gain (IG)
precision	0.8075	0.8058	0.8053	0.8447
recall	0.8609	0.8609	0.8675	0.9007
F1 score	0.8333	0.8324	0.8353	0.8718
AUC	0.9033	0.9042	0.9035	0.9284

Table 1. Comparative Performance Metrics of Asthma Prediction Models by Feature Selection Method. Table displaying the performance of the asthma prediction models across four different feature selection protocols. Assessment is based on precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

The two models using ANOVA F-statistics and logistic regression coefficients for feature selection both produced metrics largely similar to the original correlation-based model, with an F1 score of around 0.83/0.84 and an AUC of around 0.90 on the held-out test set. Despite the differences in feature ranking approach, these methods did not produce any measurable improvement in performance under this particular experimental configuration.

The model using Information Gain (IG) achieved the highest observed performance, with an F1 score of 0.87 and an AUC of 0.93. This measurable improvement indicates that entropy-based feature selection may be better for capturing nonlinear dependencies between predictors and asthma diagnosis within this dataset.

DISCUSSION

The main goal of this study was to evaluate whether non-parametric feature selection methods, specifically Information Gain (IG), could improve the predictive accuracy of asthma onset models compared to traditional statistical methods. The results support the hypothesis that IG demonstrates improved performance for identifying predictive features in complex biological datasets. While the ANOVA-test and multivariable logistic regression models matched the baseline performance of Tsang's correlation-based model (F1: 0.83, AUC: 0.90), the IG-based model demonstrated a measurable improvement (F1: 0.87, AUC: 0.93).

The similarity between the ANOVA, logistic regression, and correlation matrix methods suggests a shared

limitation: their feature selection criteria all rely on specific assumptions regarding the underlying distribution of data. As noted in the introduction, these methods assume linear relationships or specific properties like homoscedasticity. However, the nature of asthma development involves a complicated interplay of genetics and environment that rarely abide by these conditions.

The failure of ANOVA and logistic regression to improve upon the baseline suggests that these methods may be capturing similar high-level, obvious predictors while overlooking more nuanced, non-linear interactions. In contrast, Information Gain, derived from the concept of entropy, measures the reduction in uncertainty regardless of the "shape" of the relationship between the feature and the target.

The visible increase in F1 score and AUC in the IG model indicates that entropy-based selection likely identified informative features that the other three methods dismissed as statistically insignificant or non-correlated. In the context of asthma prediction, this could mean capturing environmental triggers or family history markers that only become predictive when viewed in the context of other variables. These findings suggest that entropy-based feature selection may help identify clinically relevant predictors that warrant further investigation, though this should be interpreted with caution due to the limited scope and exploratory nature of this study.

This particular study is limited by the use of a single dataset, which may not fully capture data that is universally representative. Additionally, while this study demonstrates the utility of Information Gain, it is limited by the specific architecture of the Tsang model. Future research could investigate hybrid models combining Information Gain with recursive feature elimination (RFE) and utilize more diverse datasets, as testing these selection methods on more diverse longitudinal datasets can further confirm or deny whether the increased performance of IG holds across different demographic populations. Future research should also investigate feature interaction, specifically whether or not IG is able to handle multi-way interactions between genetic and environmental factors in disease development, and whether or not this ability is the reason for its success.

Overall, this study potentially indicates that traditional statistical feature selection methods may not be enough to take into account the complex, multi-faceted nature of asthma disease development. By transitioning to an Information Gain approach, a more effective prediction model was achieved. This suggests that assumption-light feature selection methods may be advantageous when modeling complex disease processes, so much so that the benefits of these methods can potentially outweigh their shortcomings in terms of ability to capture dependencies between features. This study also demonstrates how information-theoretic feature selection methods can be applied in student-level biomedical machine learning research, providing an accessible framework for exploring nonlinear predictive modeling.

CONCLUSION

This study investigates the effectiveness of non-parametric feature selection methods for asthma prediction. It particularly emphasizes Information Gain (IG) as an alternative to traditional statistical approaches. By incorporating four different feature selection techniques into a common predictive framework, the results showed that entropy-based selection saw a visible improvement in model

performance, producing an F1 score of 0.87 and an AUC of 0.93. On the other hand, ANOVA-based selection, logistic regression coefficient ranking, and correlation-matrix methods resulted in essentially identical performance metrics, suggesting that parametric statistical feature selection might be less effective for capturing the complex, nonlinear relationships within asthma development.

The findings of this study suggest that Information Gain might provide a more effective framework for identifying predictive features in biological datasets characterized by complicated interactions between genetic, environmental, and unique lifestyle factors. In contrast to methods that rely on distributional or linearity assumptions, entropy-based feature selection evaluates predictors based on uncertainty reduction, allowing potentially informative variables to be retained even when their relationships with the target outcome do not strictly abide by the assumptions of parametric methods.

In spite of these promising indications, the generalizability of this study is limited due to the use of a single dataset. This restricts the applicability of the findings. Future research could explore validation across multiple datasets and hybrid feature selection frameworks that combine Information Gain with other dimensionality reduction techniques.

In conclusion, this study demonstrates the potential value behind assumption-light, non-parametric feature selection methods such as Information-theoretic feature selection in asthma prediction modeling and attempts to address the historical lack of standardized metrics in asthma diagnosis, allowing for non-invasive as well as accurate measurements of asthma development.

REFERENCES

[1] World Health Organization. (2024). Asthma. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/asthma

[2] The Asthma Predictive Index: A very useful tool for predicting asthma in young children Journal of Allergy and Clinical Immunology, 2010; 126, 212-216

[3] Palmer, L. J., Rye, P. J., Gibson, N. A., Burton, P. R., Landau, L. I., & LeSOUËF, P. N. (2001). Airway responsiveness in early infancy predicts asthma, lung function, and respiratory symptoms by school age. American Journal of Respiratory and Critical Care Medicine, 163(1), 37-42.

[4] Delacourt, C., Benoist, M. R., Le Bourgeois, M., Waernessyckle, S., Rufin, P., Brouard, J. J., ... & Scheinmann, P. (2007). Relationship between bronchial hyperresponsiveness and impaired lung function after infantile asthma. PLoS One, 2(11), e1180.

[5] Clough, J. B., Keeping, K. A., Edwards, L. C., Freeman, W. M., Warner, J. A., & Warner, J. O. (1999). Can we predict which wheezy infants will continue to wheeze?. American journal of respiratory and critical care medicine, 160(5), 1473-1480.

[6] Blakey, J. D., Price, D. B., Pizzichini, E., Popov, T. A., Dimitrov, B. D., Postma, D. S., Josephs, L. K., Kaplan, A., Papi, A., Kerkhof, M., Hillyer, E. V., Chisholm, A., & Thomas, M. (2017). 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*(4), 1015-1024.e8. https://doi.org/10.1016/j.jaip.2016.11.007

[7] yeemeitsang. (2024, July 9). *Asthma Prediction*. Kaggle.com; Kaggle. https://www.kaggle.com/code/yeemeitsang/asthma-prediction/notebook

[8] Kothalawala, D. M., Murray, C. S., Simpson, A., Custovic, A., Tapper, W. J., Arshad, S. H., Holloway, J. W., & Rezwan, F. I. (2021). Development of childhood asthma prediction models using machine learning approaches. Clinical and Translational Allergy, 11(9). https://doi.org/10.1002/clt2.12076

[9] Nathan, R. A., Sorkness, C. A., Kosinski, M., Schatz, M., Li, J. T., Marcus, P., Murray, J. J., & Pendergraft, T. B. (2004). Development of the asthma control test: a survey for assessing asthma control. The Journal of Allergy and Clinical Immunology, 113(1), 59–65. https://doi.org/10.1016/j.jaci.2003.09.008

[10] Shapiro, L. (n.d.). *Information Gain Which test is more informative? Split over whether applicant is employed*. https://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf

[11] Rabie El Kharoua. (2024). 🌬️ *Asthma Disease Dataset* 🌬️. Kaggle.com. https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset