Journal of Statistical Research *https://doi.org/10.3329/jsr.v58i1.75419*
2024, Vol. 58, No. 1, pp. 131-149 *ISSN 0256 - 422 X*

PREDICTING PEDIATRIC ASTHMA SEVERE OUTCOMES USING MACHINE LEARNING METHODS FOR EHR DATA WITH REPEATED CLINIC VISITS

JIAQIAN LIU, NA BO, XUEPING ZHOU

Department of Biostatistics, University of Pittsburgh, USA

Email: {jil289, nab177, xuz37}@pitt.edu

ERICK FORNO

Department of Pediatrics, Indiana University, USA

Email: eforno@iu.edu

YING DING*

Department of Biostatistics, University of Pittsburgh, USA

Email: yingding@pitt.edu

SUMMARY

Asthma is the most common multifactorial chronic disease among children. Identifying children at high risk of severe asthma outcomes, such as emergency department (ED) visits and hospitalizations due to asthma exacerbation, is essential in asthma care and clinical management. Existing studies have employed different machine learning methods to predict pediatric asthma occurrence or progression using electronic health records (EHR) data. However, these studies often neglected the correlated nature of EHR data (e.g., repeated clinic visits of the same patients). To address this issue, this research applied and evaluated two types of machine learning-based methods for longitudinal or clustered data, including random forests with mixed effects and generalized neural networks with mixed effects. We applied these methods to the real-world large asthma EHR data obtained from the Children's Hospital of Pittsburgh in a four-year period expanded from pre to post-COVID-19 pandemic, focusing on predicting the chance of having ED visits due to asthma exacerbation and the length of stay (LOS) when hospitalized. Moreover, we characterized the importance of predictors using the kernel SHAP metric and identified vulnerable patient groups that are more likely to experience asthma exacerbation or have a longer LOS. Our findings provide valuable guidance to improve pediatric asthma care by prioritizing the protection of these vulnerable patients, especially when a disruptive health crisis occurs.

Keywords and phrases: COVID-19 pandemic, EHR data, mixed effects model, neural networks, pediatric asthma, random forests

* Corresponding author

1 Introduction

Asthma, a chronic and multifactorial respiratory disease, remains present worldwide, with 300 million people affected (Lizzo and Cortes, 2019). In the United States, recent data from the Centers for Disease Control and Prevention (CDC) reveals that 4,675,475 children under 18 years old were reported to have current asthma. Uncontrolled asthma leads to recurrent Emergency Department (ED) visits and hospitalizations, where 270,330 ED visits and 27,055 hospitalizations were reported in a recent 12-month period. Moreover, managing asthma symptoms during the COVID-19 pandemic is challenging due to the disruptions in normal hospital visits and patient care. It is important to study the factors influencing pediatric asthma severe outcomes (e.g., ED visits and hospitalizations) during and after the COVID-19 disruption.

Recently, the emergence of machine learning (ML) and deep learning (DL) methods has inspired research in predicting pediatric asthma outcomes through data-driven approaches using electronic health records (EHR) data. For example, AlSaad et al. (2022) employed recurrent neural networks to predict the frequency of ED (categorized as no ED visits, 1-2 ED visits, and $>$2 ED visits) in the next 12 months among pediatric patients from the Cerner Health Facts EHR database. Sills et al. (2021) used the H2O AutoML platform to build ensembled ML models to predict the need for hospitalization among pediatric patients presenting at ED for asthma exacerbation using data from five hospital emergency departments in Colorado. Wang et al. (2019) compared deep neural networks (DNN) and penalized Lasso LR in predicting pediatric ED visits within 3 months using the Medicaid claims data from Parkland Community Health Plan in 2012 to 2014, revealing similar performance between the two methods. Patel et al. (2018) compared decision trees (DT), Lasso LR, RF, and gradient boosting in predicting the need for hospitalizations using EHR data among pediatric patients who visited two urban emergency departments affiliated with a single children's hospital for asthma exacerbation in 2012 to 2015. The DT exhibited the lowest AUCs, while other methods demonstrated comparable performance. Das et al. (2017) utilized EHR data in 2013 to predict frequent pediatric ED users (defined as $\geq$2 ED visits for asthma exacerbation) in 2014 at Weill Cornell Medical Center. Notably, all these papers ignored the potential correlations inherent in the EHR data. For instance, the same patient might have multiple ED encounters during a specific follow-up period, while the previous work either treated them independently or transformed the longitudinal binary ED occurrence into the frequency (i.e., counts within a time period) in the analysis.

Multiple ML and DL methods have recently been extended to incorporate the longitudinal structure using a generalized linear mixed models (GLMM) framework. Tree-based methods have gained increasing attention for their advantages of interpretation through the tree structure. For Gaussian response data, linear mixed effects regressions through a single tree (e.g., MERT) or random forests (RF) (e.g., MERF) have been developed (Hajjem et al., 2011; Ahlem Hajjem and Larocque, 2014), while Sela and Simonoff (2012) proposed random effects expectation maximization trees (RE-EM) to incorporate autocorrelation of errors within subjects. For non-Gaussian response data, several groups have considered replacing the fixed effects with a single tree within the GLMM framework (Hajjem et al., 2017; Fokkema et al., 2018; Fontana et al., 2021), while another group considered incorporating RF to estimate fixed effects (Pellagatti et al., 2021). In addition, Bayesian methods become popular due to their fast computation and ability to make inferences through posterior inter-

vals. For example, Binary mixed effects model forest (BiMM Forest) (Speiser et al., 2019) focuses on binary outcomes by employing a Bayesian GLMM to mitigate the convergence issues commonly associated with traditional GLMMs.

Neural network (NN) methods have also been developed in building prediction models for longitudinal or clustered data. For example, DNN with feed-forward structures were considered to approximate fixed and random components for Gaussian responses (Tandon et al., 2006; Simchoni and Rosset, 2021; Maity and Pal, 2013). Recently, multiple works have been done to extend the feed-forward DNN to model non-Gaussian responses (Wörtwein et al., 2023; Mandel et al., 2021) . For example, Mandel et al. (2021) incorporated feed-forward neural networks into the mixed model framework (GNMM), using stochastic gradient descent (SGD) to simultaneously update the random effects, network weights, and biases by maximizing the quasi-likelihood function.

In this paper, we consider two outcomes: ED visits due to asthma exacerbation (a binary outcome) and length of stay (LOS) for hospitalization (a continuous outcome). "ED exacerbation" is defined as Yes if the patient visited ED for asthma exacerbation (evidenced by getting the albuterol treatment during their asthma-related ED visit), while other types of visits are coded as No for "ED exacerbation". The LOS is calculated as the difference between the discharge date/time and the hospitalization date/time for patients admitted to the hospital from ED visits with exacerbation. All the patient visit data were recorded longitudinally. We consider both tree-based methods (i.e., BiMM Forest, MERF) and NN-based methods (i.e., GNMM) to build prediction models for ED exacerbation and LOS, respectively. The RF and GNMM methods, without taking into account the correlations among repeated visits, are also implemented to serve as benchmarks. In Section 2, we introduce the pediatric asthma EHR data. Section 3 introduces the three prediction methods used in this paper, including the BiMM Forest, MERF, and GNMM. Section 4 applies different prediction models on the pediatric asthma EHR data and compares the performance in terms of prediction accuracy. Subgroups with high risks for ED admission or hospitalization are also identified and characterized. Section 5 concludes our analysis and discusses the findings.

2 Data Description

In this study, our EHR data were obtained from the Children's Hospital of Pittsburgh (IRB:STUDY 22040043). The dataset curated more than 31,000 encounters (i.e., visits) for asthma-related symptoms from 13,264 children (aged between 2 to 21 years old) between January 2019 and January 2023. Patient demographics variables include age, sex (female and male), race (black, white, others), state (live in Pennsylvania (PA) or not), written asthma action plan (WAAP) (yes or no), influenza vaccination (yes or no), the existence of chronic diseases (existence of any bronchomalacia, bronchopulmonary dysplasia, immunodeficiency, pulmonary hypertension or tracheomalacia), existence of acute respiratory diseases (existence of acute bronchospasms or subcutaneous emphysema) and pandemic period. WAAP is an educational intervention recommended by the National Heart, Lung, and Blood Institute (NHLBI), which is tailored to the patient's individual medical history and symptoms for asthma self-care. At each visit, WAAP is recorded as Yes if the patient (or caregiver) is provided with WAAP. Otherwise, it is recorded as No. For each encounter, we categorize it into

three pandemic periods, defined as the pre-pandemic period (January 1, 2019, to March 14, 2020), the pandemic period (March 15, 2020, to April 30, 2021), and the post-pandemic period (after May 1, 2021). The cut-off dates are based on the COVID-19 lockdown and major back-to-school dates.

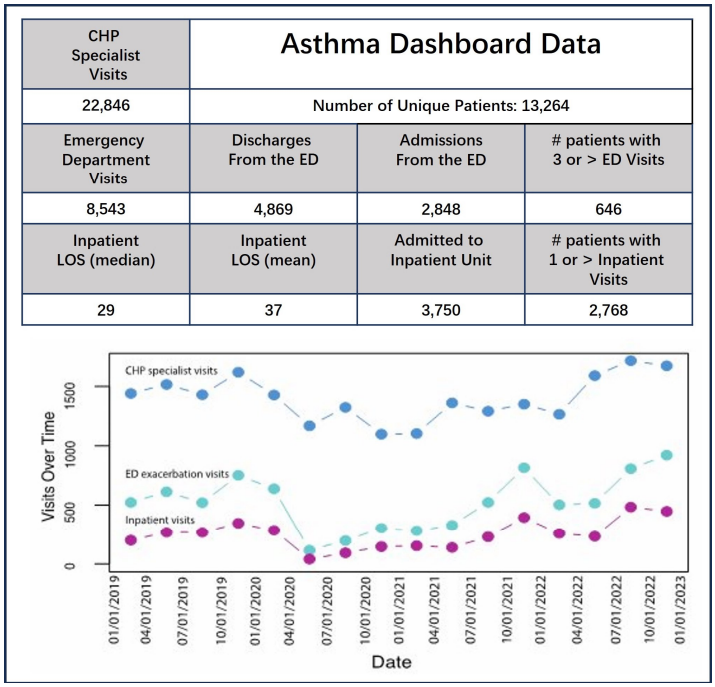

Figure 1: Summary of pediatric asthma dashboard from January 2019 to January 2023 in Children's Hospital of Pittsburgh

In this dataset, 8,500+ encounters (from 5,348 patients) are ED visits due to asthma exacerbation in which 3,810 (71.2%) patients had one ED visit and 1,538 (28.8%) patients had two or more ED visits (Figure 1). Table 1 presents the patient characteristics grouped by ED exacerbation status. We use 'n' to denote the number of patients and 'N' to denote the number of encounters. Four variables, sex, race, PA, and existence of chronic diseases, do not change over time, while all the other variables may change over time. The majority of patients live in PA (97.3%); 41.1% of patients are female; more than two-thirds of the patients are white (68.4%); and the majority of patients are between 5 and 12 years old (53.5%). There are more black patients among those who had ED exacerbation than among those who did not have ED exacerbation (40.6% vs. 12.2%), while there are more white patients among those who did not have ED exacerbation than those who had ED exacerbation (80.1% vs. 51.1%) ($p < 0.001$). Patients with ED exacerbation are younger than those without ED asthma exacerbation ($p < 0.001$). The distributions of chronic diseases, influenza vaccination, pandemic period, WAAP, acute respiratory diseases, and hospitalization are all significantly different between the groups with and without ED exacerbation (all $p < 0.001$). For example, patients with ED exacerbation are less vaccinated and have less chance of having WAAP (percentage of having

influenza vaccination: 50.5% vs 68.8%, percentage of having WAAP: 21.1% vs 44.6%). Moreover, in pre-and post-pandemic, more encounters are ED exacerbation visits than non-ED exacerbation visits (35.3% vs. 31.6% in pre-pandemic, 52.5% vs. 45.3% in post-pandemic); while during the pandemic, fewer encounters are ED exacerbation compared to non-ED exacerbation (12.3% vs. 23.1%).

Table 2 compares the LOS values across different covariate groups. Most covariates are significantly associated with LOS values. For example, younger patients (5-12 years old) have shorter LOS than older patients (13-21 years old) with the youngest group (2-4 years old) having the shortest LOS ($p < 0.001$). Notably, the LOS is significantly different across three pandemic periods ($p < 0.001$). Patients who were hospitalized during the pandemic have the shortest LOS while patients who were hospitalized after the pandemic have the longest LOS (33.0 hours during-pandemic vs 38.5 hours post-pandemic). In addition, patients who are male, live in PA, do not have chronic diseases have significantly shorter LOS compared to their corresponding complementary group (all $p < 0.001$). Note that WAAP is not included for analyzing LOS since WAAP was given after the hospitalization, which does not impact the current LOS. These exploratory analyses reveal strong associations between the covariates we explored and ED exacerbation or LOS. We decided to include all these covariates into our prediction models.

3 Methods

This section describes the tree-based and neural-network-based methods we implemented for the pediatric asthma EHR data. We also introduce the evaluation metrics to compare the model performance.

3.1 Notation

The notations for the methods we described below are defined as follows. y_{ij} is the outcome for cluster i at jth visit ($j = 1, \ldots, n_i$). In our application, each patient is considered as a cluster (Section 4). X_{ij} is the fixed effect covariate vector (of length p); Z_{ij} is the random effect covariate vector (of length k) and b_{ij} is the associated unknown random effect coefficient vector.

3.2 Random Forests with Mixed Effects

BiMM Forest. For binary outcomes, the BiMM Forest model is defined by:

$$\text{logit}(y_{ij}) = \beta_0 + \beta_1 RF(X_{ij}) + Z_{ij}^T b_{ij},$$

where $\text{logit}(y_{ij}) = \log(\frac{y_{ij}}{1-y_{ij}})$ is the logistic link function; $RF(X_{ij})$ is the predicted probability for the ith cluster at the jth visit obtained from the standard Random Forest model. β_0 represents the coefficient for the intercept, while β_1 is the coefficient for the $RF(X_{ij})$.

The algorithm for BiMM Forest consists of three steps:

Table 1: Patient characteristics by ED exacerbation status

	ED Exacerbation (No) (n=7,916) (N=23,371)	ED Exacerbation (Yes) (n=5,348) (N=8,543)	Total (n=13,264) (N=31,914)	p-value*
Sex (n, %)				
Female	3,280 (41.4)	2,171 (40.6)	5,451 (41.1)	0.628
Male	4,636 (58.6)	3,177 (59.4)	7,813 (58.9)	
Race (n, %)				
White	6,340 (80.1)	2,734 (51.1)	9,074 (68.4)	<0.001
Black	968 (12.2)	2,173 (40.6)	3,141 (23.7)	
Others	608 (7.7)	441 (8.2)	1,049 (7.9)	
PA (n, %)				
Yes	7,639 (96.5)	5,261 (98.4)	12,900 (97.3)	<0.001
No	277 (3.5)	87 (1.6)	364 (2.7)	
Existence of chronic diseases (n, %)				
Yes	340 (4.3)	187 (3.5)	527 (4.0)	<0.001
No	7,576 (95.7)	5,161 (96.5)	12,737 (96.0)	
Age (N, %)				<0.001
2 YR - 4 YR	4,414 (18.9)	3,338 (39.1)	7,752 (24.3)	
5 YR - 12 YR	13,165 (56.3)	3,908 (45.7)	17,073 (53.5)	
13 YR - 21 YR	5,792 (24.8)	1,297 (15.2)	7,089 (22.2)	
Influenza vaccination (N, %)				
Yes	16,082 (68.8)	4,310 (50.5)	20,392 (63.9)	<0.001
No	7,289 (31.2)	4,233 (49.5)	11,522 (36.1)	
Pandemic period (N, %)				
Pre-pandemic	7,385 (31.6)	3,012 (35.3)	10,397 (32.6)	<0.001
During-pandemic	5,410 (23.1)	1,048 (12.3)	6,458 (20.2)	
Post-pandemic	10,576 (45.3)	4,483 (52.5)	15,059 (47.2)	
WAAP (N, %)				
Yes	10,422 (44.6)	1,804 (21.1)	12,226 (38.3)	<0.001
No	12,949 (55.4)	6,739 (78.9)	19,688 (61.7)	
Existence of acute respiratory diseases (N, %)				
Yes	17 (0.1)	115 (1.3)	132 (0.4)	<0.001
No	23,354 (99.9)	8,428 (98.7)	31,782 (99.6)	

*p-values were computed using the Chi-square test.

n denotes the number of unique patients; N denotes the number of visits.

Table 2: Comparison of LOS among hospitalized patients for each variable

	Number of visits (N=3,750)	**LOS (hours) Mean (SD)**	**p-value***
Sex (N, %)			<.001
Female	1,534 (40.9)	39.0 (25.9)	
Male	2,216 (59.1)	36.2 (23.7)	
PA (N, %)			0.007
Yes	3,678 (98.1)	37.1 (24.5)	
No	72 (1.9)	46.8 (29.1)	
Race (N, %)			0.071
White	1,975 (52.7)	37.8 (25.7)	
Black	1,484 (39.6)	37.3 (24.1)	
Others	291 (7.7)	34.2 (19.6)	
Age (N, %)			<.001
2 YR - 4 YR	1,722 (46.0)	34.8 (23.6)	
5 YR - 12 YR	1,596 (42.6)	38.8 (24.7)	
13 YR - 21 YR	432 (11.4)	41.8 (27.5)	
Existence of chronic diseases (N, %)			<.001
Yes	246 (6.6)	50.7 (33.3)	
No	3,504 (93.4)	36.4 (23.7)	
Influenza vaccination (N, %)			0.007
Yes	2,106 (56.2)	38.2 (25.5)	
No	1,644 (43.8)	36.1 (23.6)	
Pandemic period (N, %)			<.001
Pre-pandemic	1,235 (32.9)	36.8 (24.5)	
During-pandemic	439 (11.7)	33.0 (21.8)	
Post-pandemic	2,076 (55.4)	38.5 (25.3)	
Existence of acute respiratory diseases (N, %)			0.365
Yes	93 (2.5)	40.0 (28.6)	
No	3,657 (97.5)	37.2 (24.6)	

*p-values were computed using the Chi-square test.

- Step 1: Treat y_{ij} as independent observations and train the standard random forests using covariates X_{ij} and obtain the predicted probabilities $RF(X_{ij})$.
- Step 2: Fit a Bayesian GLMM with y_{ij} as outcome and $RF(X_{ij})$ from Step 1 as the fixed effects covariate. Obtain a new predicted probability from Bayesian GLMM, denoted as q_{ij}.
- Step 3: Compute the updated binary response value y^*_{ij} by adding the q_{ij} to the original y_{ij} and applying a split function $h()$: $y^*_{ij} = h(y_{ij} + q_{ij})$.
- Step 4: Repeat step 1 to 3 using y^*_{ij} as the new outcome. The iteration ends until the change of the posterior log-likelihood of Bayesian GLMM is below a tolerance value.

Three types of split functions, denoted as h_1, h_2, and h_3, are employed in the algorithms. These functions are designed to optimize different criteria, where h_1 aims to maximize sensitivity, h_2 targets specificity maximization, and h_3 is intended to balance both sensitivity and specificity. Formulas of the three split functions can be found in section 3 of Speiser et al. (2019). Users can choose the split function based on their preference. Source code available at `github.com/bcjaeger/bimm`.

MERF . MERF is similar to the expectation–maximization algorithm used in the linear mixed-effects model. The model is defined by:

$$y_{ij} = RF(X_{ij}) + Z_{ij}^T b_{ij} + \epsilon_{ij},$$
$$b_i \sim N(0, D_i), \epsilon_i \sim N(0, R_i),$$

where $RF(X_{ij})$ is the predicted value for the ith cluster at jth visit ($j = 1, \ldots, n_i$) obtained from the standard Random Forest model; $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})$ is the $n_i \times 1$ vector of errors; D_i is the covariance matrix of $b_i = (b_{i1}, \cdots, b_{in_i})$, and R_i is the covariance matrix of ϵ_i. It assumes that b_i and ϵ_i are independent of each other and only the between-cluster variation is considered in the correlation, indicating that R_i is diagonal (i.e., $R_i = \sigma^2 I_{n_i}$).

The MERF algorithm iterates through the following three steps:

- Step 1: Treat y_{ij} as independent observations and train the standard random forests using covariates X_{ij} and obtain the predicted values $RF(X_{ij})$.
- Step 2: Calculate the estimated random effects $\hat{b}_{ij}$ by applying a linear mixed effects model with y_{ij} as outcome and $RF(X_{ij})$ from step1 as the fixed effects component.
- Step 3: Compute y^*_{ij} by subtracting the random component $Z_{ij}^T \hat{b}_{ij}$ estimated in step 2: $y^*_{ij} = y_{ij} - Z_{ij}^T \hat{b}_{ij}$.
- Step 4: Repeat step 1 to 3 using y^*_{ij} as the new outcome. The iteration ends until the change of the generalized log-likelihood is below a tolerance value.

The algorithm reaches convergence when the generalized log-likelihood between two iterations falls below a pre-specified threshold or the maximum number of iterations is reached. Source code available at `github.com/manifoldai/merf`. The flowchart of the algorithms for BiMM Forest and MERF are illustrated in Figure 2.

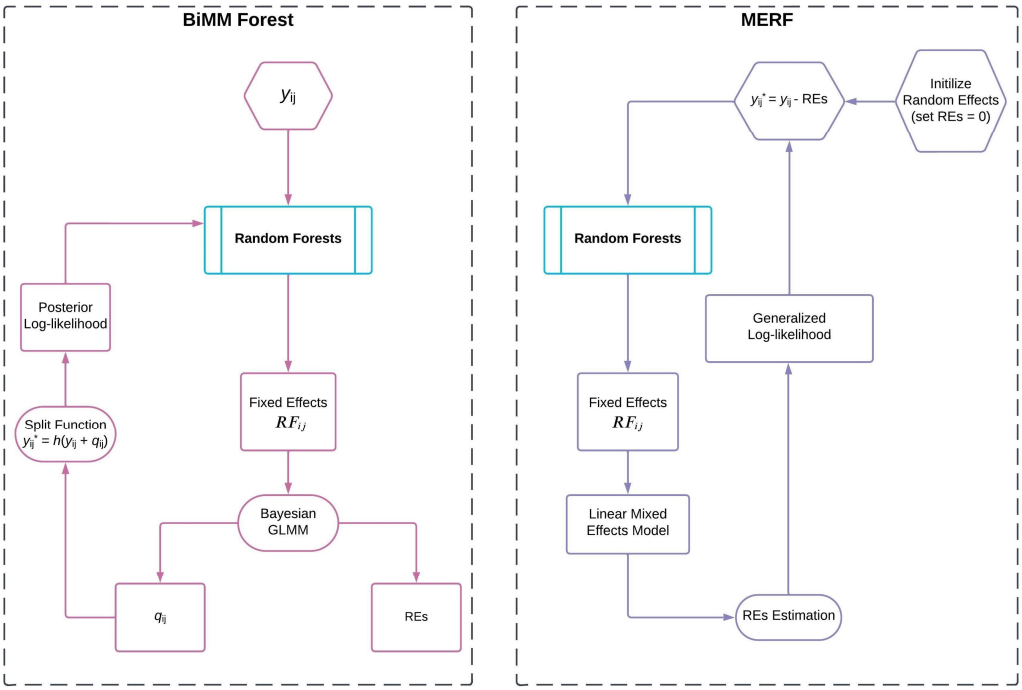

Figure 2: The flowchart of the algorithms used in BiMM Forest and MERF.

3.3 Generalized Neural Network Mixed Model (GNMM)

GNMM is a multi-layer neural network with covariates X_{ij} being the input layer and μ_{ij} being the output layer depending on weights $w^{(l)}$, bias $\delta^{(l)}$ across all layers $l = 1, \ldots, L+1$. The linear predictor of the random effects $Z_{ij}^T b_i$ is included in the final layer. An example of GNMM with $L = 2$ hidden layers is shown in Figure 3. Specifically, when x_{ij} enters into the NN with L hidden layers consisting of K_l nodes without random effects, we have

$$a_{ij}^{(1)} = g_1\{w^{(1)}x_{ij} + \delta^{(1)}\}, \tag{3.1}$$

$$a_{ij}^{(l)} = g_l\{w^{(l)}a_{ij}^{(l-1)} + \delta^{(l)}\},\ l = 2, ..., L, \tag{3.2}$$

$$\mu_{ij}^b = g_{L+1}\{w^{(L+1)}a_{ij}^{(L)} + \delta^{(L+1)} + Z_{ij}^T b_i\},\ b_i \sim N(0, D_i(\theta)). \tag{3.3}$$

Here, $a_{ij}^{(l)}$ is the output of the layer l, $g_l(\cdot)$ is the activation functions for layer l, $\delta^{(l)}$ is the $K_l \times 1$ bias vector. $w^{(1)}$ is a $K_1 \times p$ weight matrix for the first layer in (3.1) while $w^{(l)}$ is a $K_l \times K_{l-1}$ weight matrix for layer $l = 2, \ldots, L$ in (3.2). The final output μ_{ij}^b is produced by equations (3.1) to (3.3), and the random effects coefficient vector b_i is assumed to follow a multivariate normal distribution with mean 0 and covariance D_i. The covariance D_i may depend on an unknown vector

of variance components θ. $g_{L+1}(\cdot)$ is the final activation function, which is the same as the link function in GLMM. Specifically, the logit function is used to predict binary outcomes, while the identity function is employed for continuous outcomes.

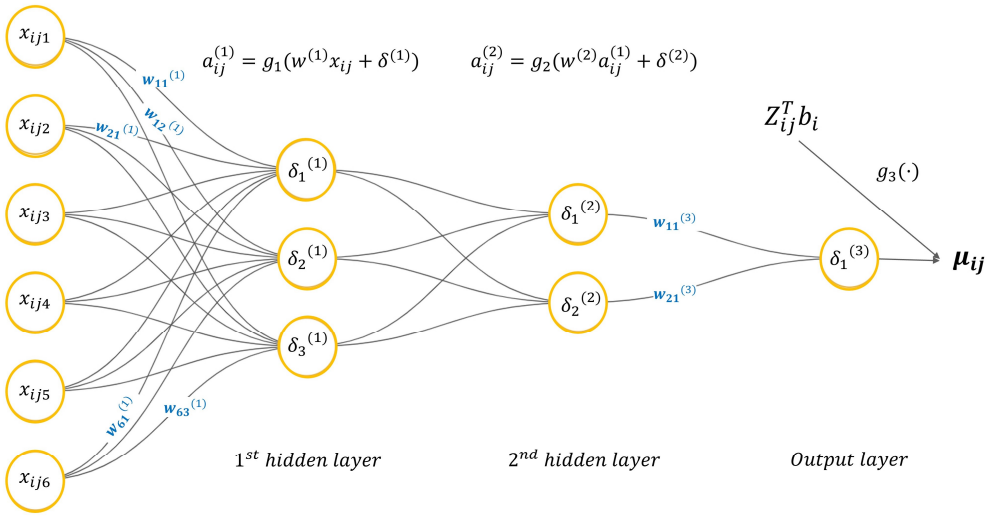

Figure 3: An example of GNMM with $p = 6$ input nodes x_{ij} fed into $L = 2$ hidden layers with $K_1 = 3$ and $K_2 = 2$ nodes, respectively. The random effects $Z_{ij}^T b_i$ are included in the final layer.

Instead of minimizing the cost function in traditional NN, GNMM maximizes the quasi-likelihood by solving quasi-score equations through back-propagation. The quasi-likelihood is written as

$$\exp\{ql(w, \delta, \theta)\} \propto$$
$$|D|^{-\frac{1}{2}} \int \exp\left\{ \frac{1}{\phi} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \int_{y_{ij}}^{\mu_{ij}^b} \frac{y_{ij} - u}{\alpha_{ij} v(u)} du - \frac{1}{2} b^T D^{-1} b - \lambda(w^T w + \delta^T \delta) \right\} db,$$

where α_{ij} is a known constant, $v(\cdot)$ is a known variance function, and ϕ is a fixed dispersion parameter. w is a single column vector generated by the combined vectorization of $w^{(1)}, ..., w^{(L+1)}$, δ is the concatenation of $\delta^{(1)}, ..., \delta^{(L+1)}$ and λ is the penalty term to avoid over-fitting. Since the integration with respect to b does not have an analytical form, Laplace approximation is used for that integration. The final objective function that GNMM maximizes is the approximated quasi-

likelihood. Details can be found in section 2 of Mandel et al. (2021). Source code available at the supporting information of Mandel et al. (2021).

Since our binary outcome has a large sample size with $N > 30,000$, we implement batching in the GNMM algorithm by creating batches containing subsets of samples to reduce the computational burden. This adjustment enables GNMM to be trained on batches in each epoch. The batches are generated by splitting the data at the patient level, ensuring that encounters from the same patients are not separated into different batches.

3.4 Evaluation Metrics

To examine the model performance, we use the area under the receiver operating characteristic (ROC) curve (AUC) to evaluate the model's discriminatory power in predicting asthma ED exacerbation. The AUC values, ranging from 0 to 1, indicate the model's predictive performance, where an AUC greater than 0.5 suggests the predictive capability is better than a random guess.

Root mean squared error (RMSE) is used as the evaluation metric to assess the model performance in predicting the continuous outcome LOS, where RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_{ij})^2}{\sum_{i=1}^{n} n_i}}$$

where $\hat{y}_{ij}$ is the predicted value of jth encounter of the ith patient, n_i is the number of encounters of the ith patient and n is the total number of patients.

Variable importance (VIP) is calculated using the SHapley Additive exPlanation (SHAP) metric introduced by Lundberg and Lee (2017). This method quantifies the contribution of each covariate to the prediction of a specific sample. Specifically, the SHAP values for covariate x_k, $k = 1, 2, \ldots, p$, represent the average difference in predicted values with and without the inclusion of x_k across all coalitions. We use the Kernel SHAP, an extended version of SHAP (Lundberg and Lee, 2017), to compute VIP as it improves computational efficiency.

4 Application to Pediatric Asthma EHR Data

In the analysis of predicting asthma exacerbation, we applied three methods: standard RF, BiMM Forest, and GNMM with the logit link. Similarly, in the analysis of predicting LOS, we also applied three methods: standard RF, MERF, and GNMM with the identity link. For the GNMM method, we tried with and without the inclusion of random effects during training. Specifically, BiMM Forest was fitted with h_1 using $k = 0.1$ as the split function ($h_1(y_{ij} + q_{ij}) = \mathbf{1}(y_{ij} + q_{ij} > k, 0 < k < 1)$). GNMM, with or without random effects, was fitted with 10 nodes in 1 hidden layer. For the ED exacerbation prediction, due to the large sample size, we implemented the stochastic mini-batch approach (with 25 batches) to save the computational load.

We applied the 5-fold cross-validation (CV) procedure for each method, which involves dividing the data into five folds, with training on four folds and testing on the remaining one fold. We repeated this process five times with different splits to ensure robust validation. We employed the

fixed effects component of each method (i.e., extracted the random forest model from BiMM and MERF, and the neural network without random effects from GNMM) to make predictions on the testing data. We expect that incorporating mixed effects will assist in adjusting the fixed effects models. Note that the data split in CV was performed based on individuals (not on encounters) to avoid the separation of encounters from the same patients into different folds. After the comparison of prediction performance across all models, the preferred model for each outcome was selected for the downstream analysis, which includes calculating the VIP through kernel SHAP metric and identifying the vulnerable subgroups using the top predictors.

4.1 Prediction Performance

Figure 4(A) presents the AUC values for the ED exacerbation prediction across four methods with five data splits. It shows that the incorporation of random effects in the model training contributes to an enhanced AUC. The average AUCs of GNMM, with or without random effects, are both 0.782, which is slightly higher than the average AUC of BiMM Forest (0.778). However, BiMM Forest shows a more stable prediction performance (i.e., more robust against different data splits) and is notably more efficient in terms of computational time. As such, BiMM Forest is chosen as the preferred method for predicting ED exacerbation. Figure 4(B) presents the RMSE of each method for predicting LOS. RF and MERF show slightly better performance than GNMM, with an average RMSE of 0.542. GNMM produces an average RMSE of 0.548, regardless of whether including the random effects or not.

4.2 VIP and Top Predictors

Figure 5(A) presents the VIP of each covariate calculated by Kernel SHAP based on the BiMM Forest model. Race, Age, and WAAP are the top three predictors, contributing significantly to the model's predictive performance. Figure 5(B-D) plots the distribution of predicted ED exacerbation probabilities grouped by each top predictor. Compared to white patients or those of other races, black patients have a higher probability of having ED visits due to asthma exacerbation, evidenced by a significantly taller bar at probability around one. Similarly, patients with younger ages (2-4 years old) also show a higher risk of ED visits in comparison to patients aged 5-12 or 13-21 years old. Patients who received WAAP show a reduced risk of having ED visits compared to those who did not receive WAAP.

To examine the interaction effects by top predictors, we implemented the partial dependence plots (PDP) to visualize the predicted probabilities of ED exacerbation due to asthma exacerbation by individual covariates. Figure 6(A) presents the PDP of predicted probabilities by age and race. Black patients exhibit significantly higher risks of ED exacerbation compared to white and other race patients across all age groups. For white and other race patients, those aged 5 years or younger show significantly higher risks of ED exacerbation compared to older patients (aged 5-21). In contrast, both younger (5 years or below) and older (16 years or above) Black patients show higher risks of ED exacerbation compared to those aged 5-16 years. Similarly, Figure 6(B) demonstrates that WAAP aids in reducing the risk of ED exacerbation, especially among younger patient groups,

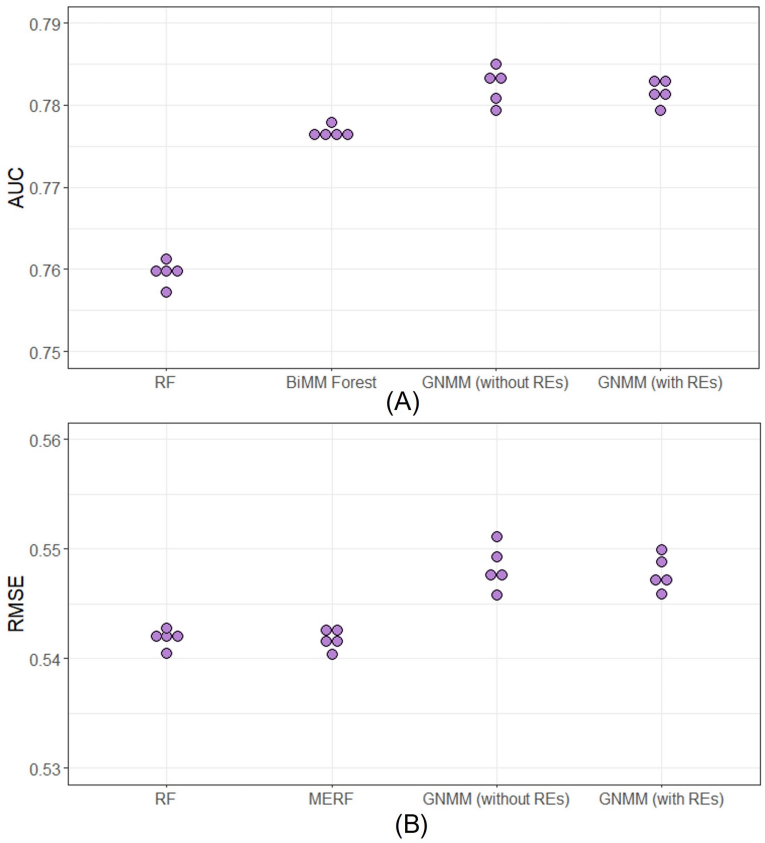

Figure 4: (A) Dot plot of AUC for ED with exacerbation prediction. (B) Dot plot of RMSE for LOS prediction.

as evidenced by the lower predicted probabilities of ED exacerbation in patients aged 5 years or younger with WAAP. Additionally, the overall predicted probabilities of ED exacerbation are lower in those with WAAP compared to those without it.

Figure 7(A) presents the VIP of each covariate calculated by Kernel SHAP based on the MERF model. Among all predictors, age, pandemic period, and sex exhibit the highest absolute median SHAP values. We identified two groups when fitting the Gaussian mixture model on the predicted

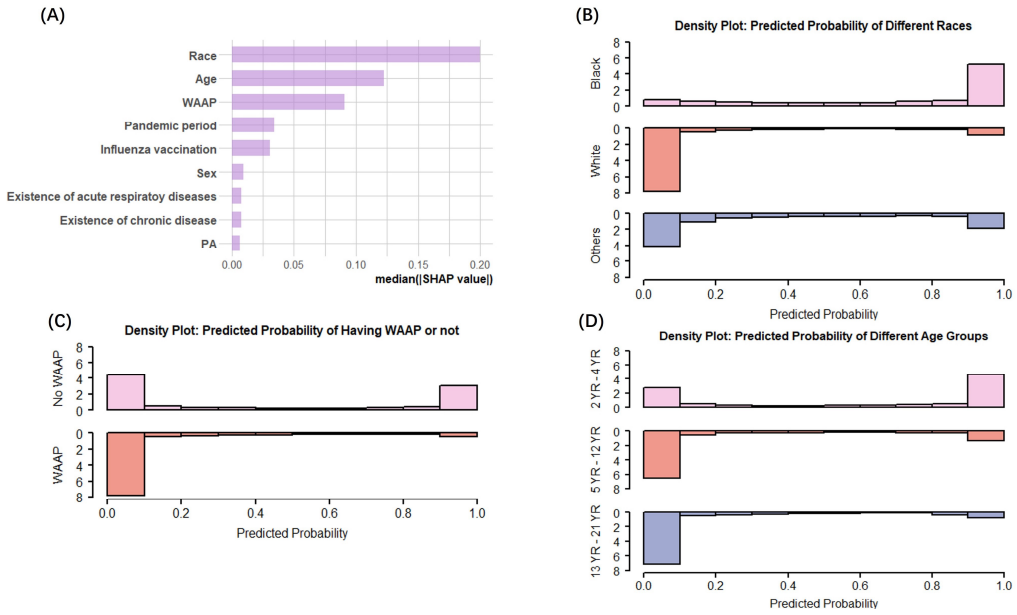

Figure 5: Results for ED exacerbation from BiMM Forest: (A) VIP measured by Kernel SHAP. The order of the covariate display is determined by the median of absolute Kernel SHAP values. (B-D) Predicted probability densities by each top three VIP predictor.

LOS values, indicated by the short-LOS group and the long-LOS group (Figure 7(B)). We compared the patient characteristics between the two groups in terms of the top three predictors. Among those hospitalized patients, children with longer LOS are older compared to children with shorter LOS (Figure 7(C), $p < 0.001$). The majority of patients with long LOS are hospitalized post-pandemic (Figure 7(D), 62.1% high-risk are in post-pandemic, $p < 0.001$). This could be attributed to the delayed hospital visits due to the pandemic, leading to worse symptoms and longer hospital stays during the post-pandemic period. Additionally, patients who have longer LOS are female patients compared to the low-risk group (Figure 7(E), 55.8% female in high-risk o vs 39.4% female in low-risk, $p < 0.001$).

Figure 8(A) presents the PDP by pandemic period and age. Although the overall trends indicate longer predicted LOS for older patients across all three pandemic periods, the increase in LOS with age is more pronounced in the pre- and post-pandemic periods compared to during the pandemic. Patients hospitalized during the pandemic were discharged quicker compared to those hospitalized before or after the pandemic. Figure 8(B) illustrates the PDP by sex and age, showing that female patients are predicted to have longer hospital stays than male patients across most age groups, although

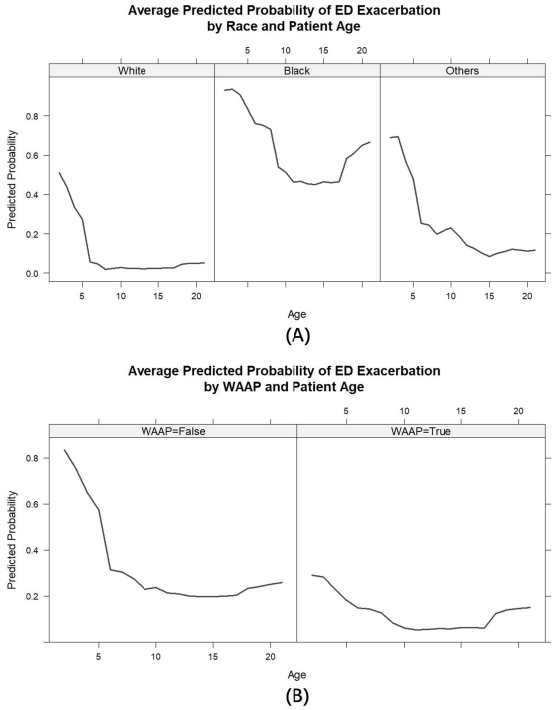

Figure 6: Partial dependence between top three predictors: (A) Average predicted probability from BiMM by race and age. (B) Average predicted probability from BiMM by WAAP and age.

the increase in LOS with age is more pronounced in males.

5 Conclusion and Discussion

In this paper, we introduced two types of machine learning-based methods for longitudinal or clustered data: (1) BiMM Forest and MERF, which are the random forest-based models with random effects for binary data and continuous data, respectively; (2) GNMM, which is the generalized neural network model with mixed effects. We applied all these models to the pediatric asthma EHR data that involved repeated clinic visits. Specifically, we analyzed two severe asthma outcomes: ED visits due to asthma exacerbation and length of stay after admission to the hospital. Among all methods, random forest-based methods, i.e., BiMM Forest and MERF, outperform the other methods with satisfactory and robust prediction performance and efficient computation. We also used the kernel

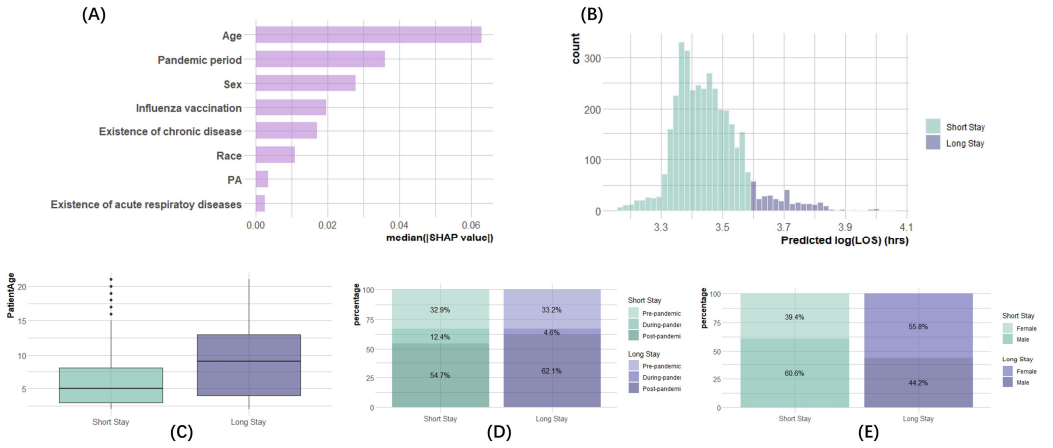

Figure 7: Results for LOS from MERF: (A) VIP measured by Kernel SHAP. The order of the covariate display is determined by the median of absolute Kernel SHAP values. (B) Two clusters were identified using the Gaussian mixture model. (C-E) Patient characteristics include age, pandemic period, and sex, separated by two risk groups.

SHAP metrics to measure the variable importance and identified vulnerable subgroups at a high risk of ED visits due to asthma exacerbation or longer LOS.

Our analysis suggests that patients who are younger, black, and not receiving the WAAP intervention are more likely to experience asthma exacerbation. Moreover, among those patients who are hospitalized after ED visits, older and female children are more likely to have a longer LOS. The COVID-19 pandemic also impacts pediatric asthma care, with longer LOS observed post-pandemic. These findings provide valuable insights for improving the surveillance strategy for pediatric asthma care by prioritizing the protection of vulnerable patient subgroups, especially when a future disruptive health crisis occurs.

Acknowledgements

This work has been supported by the CTSI Public Health Trans-Disciplinary Collaboration Pilot grant by the University of Pittsburgh, USA.

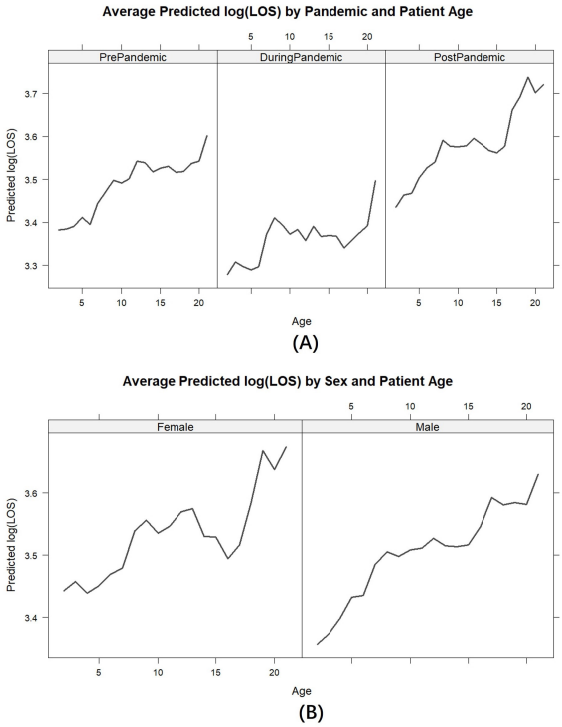

Figure 8: Partial dependence between top three predictors: (A) Average predicted log(LOS) by pandemic and age. (B) Average predicted log(LOS) by sex and age.

References

Ahlem Hajjem, F. B. and Larocque, D. (2014), "Mixed-effects random forest for clustered data," *Journal of Statistical Computation and Simulation*, 84, 1313–1328.

AlSaad, R., Malluhi, Q., Janahi, I., and Boughorbel, S. (2022), "Predicting emergency department utilization among children with asthma using deep learning models," *Healthcare Analytics*, 2, 100050.

Das, L. T., Abramson, E. L., Stone, A. E., Kondrich, J. E., Kern, L. M., and Grinspan, Z. M. (2017),

"Predicting frequent emergency department visits among children with asthma using EHR data," *Pediatric Pulmonology*, 52, 880–890.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., and Kelderman, H. (2018), "Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees," *Behavior research methods*, 50, 2016–2034.

Fontana, L., Masci, C., Ieva, F., and Paganoni, A. M. (2021), "Performing Learning Analytics via Generalised Mixed-Effects Trees," *Data*, 6, 74.

Hajjem, A., Bellavance, F., and Larocque, D. (2011), "Mixed effects regression trees for clustered data," *Statistics & Probability Letters*, 81, 451–459.

Hajjem, A., Larocque, D., and Bellavance, F. (2017), "Generalized mixed effects regression trees," *Statistics & Probability Letters*, 126, 114–118.

Lizzo, J. M. and Cortes, S. (2019), "Pediatric asthma," StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. Available from `https://www.ncbi.nlm.nih.gov/books/NBK551631/`.

Lundberg, S. M. and Lee, S.-I. (2017), "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, 30.

Maity, T. K. and Pal, A. K. (2013), "Subject specific treatment to neural networks for repeated measures analysis," *Proc Int MultiConf Eng Comput Sci*, 1, 60–5.

Mandel, F., Ghosh, R. P., and Barnett, I. (2021), "Neural Networks for Clustered and Longitudinal Data Using Mixed Effects Models," *Biometrics*, 79.

Patel, S. J., Chamberlain, D. B., and Chamberlain, J. M. (2018), "A Machine Learning Approach to Predicting Need for Hospitalization for Pediatric Asthma Exacerbation at the Time of Emergency Department Triage," *Academic Emergency Medicine*, 25, 1463–1470.

Pellagatti, M., Masci, C., Ieva, F., and Paganoni, A. M. (2021), "Generalized mixed-effects random forest: A flexible approach to predict university student dropout," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14, 241–257.

Sela, R. J. and Simonoff, J. S. (2012), "RE-EM trees: a data mining approach for longitudinal and clustered data," *Machine learning*, 86, 169–207.

Sills, M. R., Ozkaynak, M., and Jang, H. (2021), "Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning," *International Journal of Medical Informatics*, 151, 104468.

Simchoni, G. and Rosset, S. (2021), "Using Random Effects to Account for High-Cardinality Categorical Features and Repeated Measures in Deep Neural Networks," in *Advances in Neural Information Processing Systems*, eds. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., Curran Associates, Inc., vol. 34, pp. 25111–25122.

Speiser, J. L., Wolf, B. J., Chung, D., Karvellas, C. J., Koch, D. G., and Durkalski, V. L. (2019), "BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes," *Chemometrics and Intelligent Laboratory Systems*, 185, 122–134.

Tandon, R., Adak, S., and Kaye, J. A. (2006), "Neural networks for longitudinal studies in Alzheimer's disease," *Artificial intelligence in medicine*, 36, 245–255.

Wang, X., Wang, Z., Pengetnze, Y. M., Lachman, B. S., and Chowdhry, V. (2019), "Deep Learning Models to Predict Pediatric Asthma Emergency Department Visits," .

Wörtwein, T., Allen, N. B., Sheeber, L. B., Auerbach, R. P., Cohn, J. F., and Morency, L.-P. (2023), "Neural Mixed Effects for Nonlinear Personalized Predictions," in *Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 445–454.

Received: February 1, 2024

Accepted: May 12, 2024