RespRisk: A Novel Real-Time Asthma-Related Hospitalization Risk Assessment System Using Air Pollutant Concentrations

Arnav Dhar, Erik Siavichay & Adrian Williams

Received May 28, 2024
Accepted August 20, 2024
Electronic access September 15, 2024

Asthma is a chronic lung disease affecting 1 in 12 (27 million +) United States (US) citizens. Statewide air pollutant concentrations affect asthma-based hospitalization rates, and increased air pollutant concentrations can cause unprecedented surges in asthma-related hospitalizations, sometimes overwhelming hospitals. Hospitals can use artificial intelligence (AI) to provide them with an estimated number of asthma-related hospitalizations, which allows for effective preparation and resource allocation. To contribute to this field, we propose RespRisk, a device optimized specifically for efficient hospitalization rate prediction. RespRisk is based on Raspberry Pi 3B+ and uses local pollutant concentrations as input for an Extreme Gradient Boost Regression (XGBoost) Model, which outputs predicted hospitalization rates per 10,000 individuals. RespRisk is trained using a concatenation of California Health and Human Services Asthma Hospitalization data and Environmental Protection Agency (EPA) Daily Air Quality Data for all months from 2000-2020 in New York State. A Conditional Tabular Generative Adversarial Network (CTGAN) is fitted to this data to generate additional synthetic training data for the RespRisk model, and multiple model modalities are created and evaluated via Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), and Mean Absolute Error (MAE). RespRisk uses an API to access local Ozone (O_3) concentrations and air quality sensors to access Nitrogen Dioxide (NO_2), Particulate Matter 2.5 (PM2.5), and Carbon Monoxide (CO) concentrations. RespRisk achieves a test MAE of just 0.10 via only 169 original training data points. Future work for the model includes training the model on larger datasets accounting for diverse populations and locations.

Keywords: asthma, asthma-related hospitalization, asthma exacerbation, air pollution, hospitalization prediction, respiratory hospitalization

Introduction

Asthma is a chronic respiratory condition characterized by inflammation and narrowing of the airways. Its etiology is complex, involving genetic predisposition, environmental exposures, and immunological factors. Often, asthma is triggered by an allergic reaction to foreign particles. Around 300 million people worldwide suffer from asthma, and asthma prevalence increased from 226.9 million in 1990 to 262.41 million in 2019[1]. Asthma is particularly dangerous due to its ability to cause "asthma attacks," or sudden asthma exacerbations characterized by coughing, wheezing, shortness of breath, and, in some cases, asphyxiation. During an asthma attack, a person's airways become inflamed, constrict, and produce extra mucus, making it difficult for the asthmatic to breathe. During severe and extended asthma attacks, the narrowing of airways can prevent adequate oxygen levels from reaching the lungs and brain. This eventually leads to reduced breathing and possibly death by mixed acidosis and hypercapnia[2]. Asthma cannot be "cured" in a traditional sense, but its symptoms can be managed using medication and inhalers, which treat and prevent symptoms and are well tolerated by the majority of individuals.

Under-resourced nations sometimes experience hospital overcrowding due to sudden surges in asthma attacks, leading to long wait times for treatment. Delayed treatment for an asthma attack can lead to worsening of symptoms or even death[2]. Figure 1 details how countries without access to consistently quality healthcare – particularly developing nations – experience increased asthma mortality rates, sometimes due to lack of preparedness and poor management of asthma surges. A device that could warn hospitals in underprivileged areas about predicted surges in asthma-related hospitalizations can allow preparedness and more effective handling of mass treatment. Alerting hospitals about surges in asthma-related hospitalizations can allow better preparedness and overall resource allocation, allowing more efficient and effective treatment of people with asthma exacerbations.

Early hospital preparedness requires continuously calculating an estimated number of asthma-related hospitalizations, which can be predicted based on air pollution concentration. Air pollution is often a major factor in causing asthma attacks. Carbon monoxide, nitrogen dioxide, ozone, and fine particulate matter

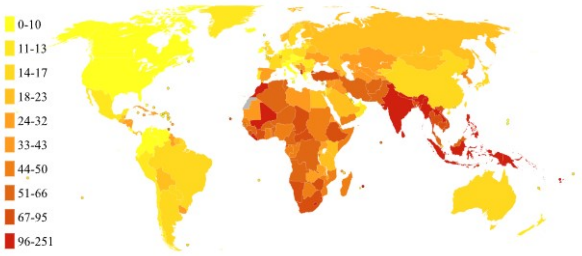

Fig. 1 Map Depicting Worldwide Asthma Related Deaths Per Million in 2012 (Taken from World Health Organization)

(PM2.5) stand out as key factors in predicting asthma attacks[3]. These pollutants originate mostly from human activities: CO and NO_2 are byproducts of combustion and vehicle exhausts; O_3 forms due to reactions with volatile organic compounds (VOCs) and nitrogen oxides (NOx) in sunlight; and fine particulate matter includes all tiny particles from vehicles, construction, smoke, and natural sources[4]. These four pollutants intensify respiratory conditions and can harm asthmatics. Concentrations of these pollutants fluctuate throughout the year due to fluctuations in temperature, natural events, human energy consumption, traffic, and more.

Changes in air pollution concentrations have been mapped to corresponding fluctuations in respiratory hospitalization, and the positive correlation between air pollution and asthma-related hospitalizations has been documented extensively[5]. The six criteria air pollutants – PM2.5, PM10, Sulfur Dioxide (SO_2), lead (Pb), NO_2, CO, and O_3 – play an important role in asthma-related hospitalizations, particularly PM2.5, CO, NO_2 and O_3. A recent meta-analysis by the United States EPA suggests a strong correlation between these four pollutants and asthma exacerbation rates. A large body of literature suggests that increased concentrations of air pollution, specifically the criteria pollutants, volatile organic compounds, allergens, etc. are positively correlated with increases in asthma-related hospitalizations[6]. However, other studies demonstrate highly mixed results by pollutant, or even no correlation[7]. This lack of consistency may be due to variations in predicted values (e.g., hospitalizations vs. Emergency Department [ED] visits), locations, methods of air pollutant exposure measurement (e.g., ambient monitor data vs. personal exposure data), and model evaluation methods[8].

Current methods for quantifying the relationship between air pollution and asthma-related hospitalizations involve evaluating the accuracy of a predictive model. For example, Shen et al.[9] utilized daily pollutant concentrations and asthma hospitalization rates from January 1 to December 31 in 2014 to train and evaluate a generalized additive model. They demonstrated proficiency in predicting hospitalization rates after controlling for seasons and holidays. Kheirbek et al.[?] investigated the health impacts of PM2.5 and ozone in New York City, highlighting the role of data resolution and socioeconomic disparities in asthma-related emergencies. Delamater et al.[10] investigated the effectiveness of Bayesian regression and temporal random effects in correlating air pollution with asthma hospitalizations and evaluated their model based on a variety of 'goodness of fit' metrics such as the D statistic. Similarly, Hwang et al.[6] proposed a recurrent neural network, a long short-term memory model, and a gated recurrent unit model to predict emergency room visit rates using air pollutants, weather conditions, pollen, and influenza[6]. The authors utilized a training dataset consisting of 18 environmental factors and corresponding daily Emergency Room visit rates in Korea from 2015-2019.

A common theme among these studies is the use of large and complex models, resulting in subsequently large models that take correspondingly more computational power to form predictions. While this is not an issue in developed countries like the United States with consistent, quality healthcare, developing countries, such as African ones, often experience understaffing in emergency services. In such contexts, small, reliable devices prove to be more valuable due to their ability to function independently of other complex software, minimizing points of failure in the system. With this in mind, we propose a device designed to be transportable and computationally efficient, fit with a model designed for edge computing using a relatively small dataset. To combat the common problem of lacking public healthcare data, we analyze the effects of synthetically generating additional training data with a CTGAN model while still testing and evaluating the model using the original dataset. RespRisk, created as a proof of concept, involves an AI model fit into the Raspberry Pi minicomputer. Because APIs can sometimes lag in reporting recorded data to clients, sensors are attached to the Raspberry Pi, to feed PM2.5, NO_2, and CO data into the RespRisk model, allowing continuous and almost instantaneous output of a predicted number of asthma-related hospitalizations per 10,000. Because of the weak correlation between Ozone concentrations and hospitalization rates, an API is utilized to access local concentrations to keep the device budget-friendly for widespread usage.

RespRisk is designed to tackle the unique challenges faced by under-resourced regions by employing lightweight computational models that can run on minimal and basic hardware, navigating the challenge of minimal resources and infrastructure that these countries face. Moreover, RespRisk's cost-effective implementation makes it accessible even in regions with limited financial resources, allowing doctors to be notified of hospitalization surges with reasonable accuracy at a fraction of the computational and hardware-related cost of other similar models.

Results

Model Variations

The RespRisk model accuracy progressed with each of the three trials. Table 2 depicts the results of employing multivariate regression, neural network regression, and XGBoost regression on the same testing dataset with default hyperparameters, along with the effects of synthetic data. XGBoost Regression outperformed the other alternatives by all evaluation metrics, achieving nearly one-half the MAE and MAPE scores of Neural Network and Multivariate Regression. The inclusion of synthetic data had minimal impact on model performance, and, in some cases, the model performed better without it. Therefore, we opted to train the final RespRisk model using only the original dataset.

Hyperparameter Tuning

After using KerasTuner to optimize the hyperparameters of the RespRisk depicted in Table 1, we trained and evaluated the model five times using five unique train-validation-test splits, all following a 70% - 15% - 15% ratio. Each trained model was discarded before starting a new trial, and models were evaluated on Root Mean Square Error, Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error, and Coefficient of Determination (R2). These are accuracy metrics measuring the average discrepancy between true and predicted values for each prediction the model makes. Each of these evaluations is described as a "trial" in Table 2, and the average model evaluation is presented.

Model Evaluation

Figures 2 and 3 below depict the relation between the true and predicted Asthma-Related Hospitalizations. The dotted red line represents the line y = x.

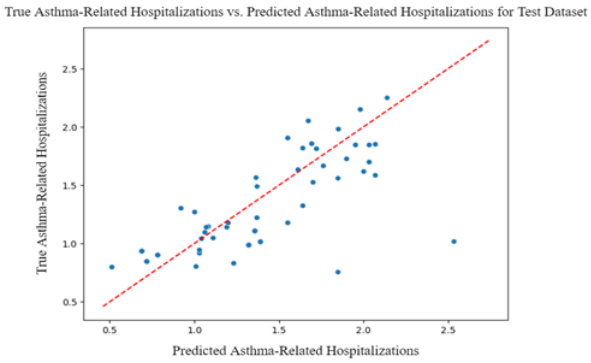

Fig. 2 Model Predictions of Asthma-Related Hospitalizations Per 10,000 Individuals in New York State for Testing Dataset vs. True Values

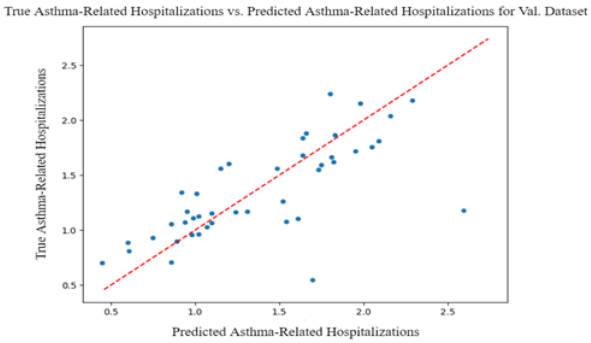

Fig. 3 Model Predictions of Asthma-Related Hospitalizations Per 10,000 Individuals in New York State for Training Dataset vs. True Values

Discussion

Model Variations

Table 2 shows that synthetic data generation did not significantly enhance the accuracy of the three regression modalities. This is evidenced by the small difference in error between predictions made using normal training data alone and those made using a combination of synthetic and normal training data. This may be due to the complexity and uniqueness of real-world environmental and health data, which synthetic data may not fully capture with just 169 reference data points. Additionally, the training data may have been sufficiently robust to begin with, making the synthetic data's contribution negligible. Lastly, overfitting to the patterns present in true data may have detracted from performance by causing the synthetic data to look too similar to the real data. Because of the negligible performance increase caused by synthetic data augmentation, we did not include the synthetically generated data in the training dataset for the final model.

Our first approach to designing the RespRisk model was a simple multivariate regression model under the assumption that air pollutant concentrations directly increase asthma hospitalization rates3. However, this model produced relatively high error metrics (RME, MAE, MSE, and MAPE scores, as shown in Table 2), indicating poor performance. This is likely due to the complex correlation of the target and predictor variables in the dataset.

Our second approach used neural network regression, since a neural network can discover complicated trends in the data which a linear regression model could not account for. However, despite slight improvement, the neural network regression model still achieves a below-proficient accuracy (Table 2). This is presumably because neural networks require a substantial amount of training data to operate at optimal performance, resulting in the model's inability to fully learn the variable correlations with the low amount of training data provided.

Table 1 Effects of Modeling Modalities and Synthetic Data without Hyperparameter Tuning

Metric	Linear Regression		Neural Network Regression		XGBoost Regression	
	No synthetic data	With synthetic data	No synthetic data	With synthetic data	No synthetic data	With synthetic data
Test RMSE	0.59	0.53	0.46	0.43	0.39	0.37
Test R^2	0.76	0.72	0.79	0.76	0.86	0.84
Test MAE	0.34	0.41	0.32	0.38	0.16	0.17
Test MSE	0.34	0.27	0.19	0.23	0.29	0.30
Test MAPE	0.55	0.53	0.39	0.37	0.25	0.27

Table 2 Final Model Evaluation after Running for 10 Trials

Trial	RMSE	MSE	MAE	MAPE	R^2
Trial 1	0.30	0.10	0.14	0.19	0.88
Trial 2	0.40	0.28	0.08	0.16	0.84
Trial 3	0.31	0.15	0.09	0.16	0.86
Trial 4	0.24	0.11	0.05	0.18	0.82
Trial 5	0.23	0.16	0.10	0.19	0.87
CI(95%)	$\pm$0.066	$\pm$0.063	$\pm$0.029	$\pm$0.013	$\pm$0.045
Average	0.33	0.23	0.10	0.19	0.85

Our third and best approach used the XGBoost regression model (Table 2). We chose this model due to its gradient-boosting capabilities and its ability to work with smaller datasets, making it suitable for our limited dataset of 241 rows. We believe gradient boosting allowed the model to capture complex relationships similar to a neural network regression model without requiring large quantities of training data like neural networks. Additionally, the Grid Search methodology implied more increasingly accurate predictions, given a higher number of decision trees in the model. However, we have chosen to limit the model to 170 decision trees to lower model complexity and improve runtime efficiency. This ensures that the model can reliably run on the Raspberry Pi's small Graphics Processing Unit (GPU) and Central Processing Unit (CPU) without additional software. Because this model had the lowest error rates out of the six model variations tested in Table 2, we chose to use XGBoost regression as our final model.

Dataset

The weak correlation between our target and predictor variables, as seen in Figure 4, creates a glass-ceiling effect for model accuracy. This may be caused by two main reasons. First, averaging $\sim$30 air pollution quantities to create monthly representative values led to significant data loss, increasing the influence of outliers on the representative number. Second, observation period of the study – 2000 to 2020 – has significant implications. Sensors and hospital legislation may have been updated during this period, causing variations in data measurement techniques. Additionally, New York State's population has increased from 8 million in 2000 to nearly 20 million in 2020, causing increased pollution, industrial activity, and development.

Model Evaluation

Figures 5 and 6, depicting model predictions and true values, show their consistently accurate predictions on unseen testing data. Because this line represents the y = x function, points falling upon this line represent instances where the model's prediction is the same as the true value. Because most points fall within a reasonable distance of this line, and due to the proficient coefficient of determination, there is some evidence that the model can generalize and predict hospitalization rates, but this claim cannot be verified without further testing. We believe the gradient boosting framework, which aptly captures complex data relationships without requiring an extensive volume of training data, allows this to happen. Additionally, the decision to limit the model to 170 decision trees allows the model to learn to form predictions more effectively without simply memorizing the trends of the training dataset. This also allows the model to operate effectively on the constrained hardware resources of a Raspberry Pi.

The model's evaluation metrics in Table 3 show its proficiency. The model's MAE, being the best metric, suggests that it was quite effective in minimizing average errors in predictions. Specifically, MAE treats all errors linearly, without excessively penalizing larger errors. Therefore, MAE, being the model's best metric, implies that the model consistently produces predictions with low error but occasionally completely misjudges a data point. This trend can be seen in the outlier points in Figure 2 and Figure 3, and explains the relatively higher RMSE and MSE scores, as they would punish such outliers more harshly than MAE.

On the other hand, the relatively higher MAPE indicates that the model was less adept at handling outliers. MAPE calculates

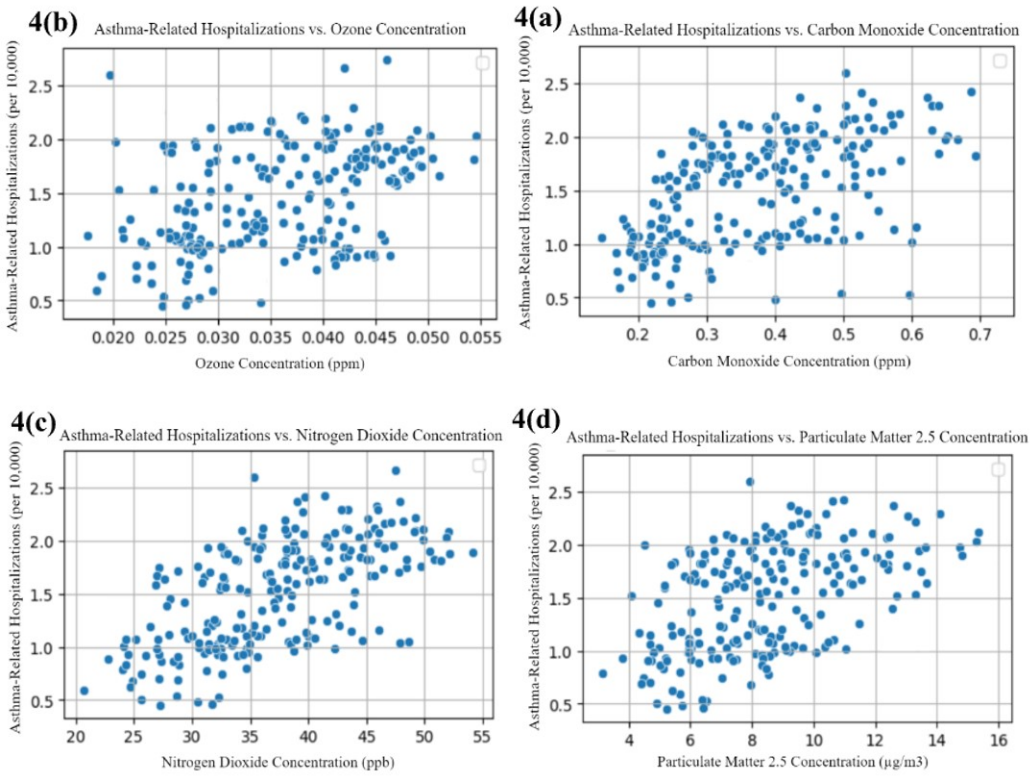

Fig. 4 (a, b, c, d). Scatter Plots of Hospitalization Rates against Predictor Variables

errors as a percentage of actual values, making it sensitive to relative differences. This implies that the model sometimes struggles with data points which should, in reality, have low values, as it tends to occasionally severely overestimate the number of hospitalizations that will happen in such cases and produce an outlier. The equation to calculate MAPE and RMSE penalizes larger errors more heavily than smaller ones. The higher RMSE and MAPE, in comparison to the MAE, indicates that while the average error is low, the model occasionally significantly overestimates the number of hospitalizations, creating an outlier point that increases the RMSE and MAPE more than the MAE. This is consistent with the presence of outlier points in Figure 2 and Figure 3, where the model's predictions were notably off. This could be attributed to the model's limited training data, which may not have adequately captured the variability in the data, increasing error. Furthermore, an R2 value of 0.87 indicates a large proportion of the variability in hospitalizations is accounted for by the prediction variables, demonstrating that the model is a good fit for the data.

In Table 4, we compare RespRisk's MSE with that of various models proposed by Delamater et al., who examined the accuracy of seven Bayesian Regression-based models in predicting monthly asthma hospitalizations per 10,000 using different features. Similarly, Table 5 compares the R^2 values of RespRisk to those of various deep learning methods for asthma hospitalization prediction in South Korea proposed by Hwang et al. These tables collectively demonstrate that RespRisk's error rates and model capabilities are on par with current state-of-the-art methods for predicting asthma hospitalizations.

Table 3 Model Comparison between RespRisk Model and Delamater et al.'s Models

Model	Features	MSE
RespRisk	NO_2 + PM2.5 + O_3 + CO	0.23
Delamater et al.	CO	0.38
	NO_2	0.24
	NO_2 + RH*	0.31
	PM2.5 + RH*	0.24
	NO_2 + PM10 + O_3 + CO + RH	0.21
	NO_2 + Tmax*	0.31
	PM2.5	0.39

RH: relative humidity
Tmax: maximum monthly temperature

We combatted high RMSE and MSE scores by adding L1 and L2 Regularization to the XGBoost Regression and the neural network regression models. Running the tests depicted in Table 2 again on these updated models, we noticed insignificant

Table 4 Model Comparison between RespRisk Model and Hwang et al.'s Models

	Features	**Model**	**R^2**
RespRisk	NO_2 + PM2.5 + O_3 + CO	XGBoost	0.85
Hwang et al.	CO + NO_2+ O_3+ PM10 + PM2.5 + SO_2 + temperature + humidity + precipitation + solar radiation + wind speed	GLM	0.792
		GAM	0.943
		RM	0.857
		GBM	0.957
		RNN	0.894
		LSTM	0.886
		GRU	0.905

improvements in the accuracy of the model. We also tested decreasing the z-score limits for outlier removal, discussed in Section B of the Methods section, from 3 and -3 to 2.5 and -2.5. This classifies more data points as "outliers" and removes them from the dataset. After doing so, we again conducted the tests depicted in Table 2 on the new training data, but noticed insignificant effects on reducing the error of the model. Therefore, we decided to reset the z-score limit for outlier removal back to 3 and -3, and did not include L1 or L2 regularization in the final model.

Similarly, a moderate RMSE value suggests that the model's predictions are generally close to the actual values, but still have some variability, particularly with more significant errors. As mentioned earlier, RMSE, as the square root of MSE, gives a direct interpretation of the error magnitude, meaning larger errors significantly affect it. The MSE, also moderate, emphasizes the squaring of errors, which disproportionately penalizes larger errors more than smaller ones. This could mean that while the model is fairly accurate on average, it may struggle more with data points that are outliers or have larger variances from the norm. These metrics indicate the model's overall reliability in prediction but also highlight areas where its performance might falter, particularly in handling outliers or extremely variable data.

Table 6 compares the specifications – CPU, Random Access Memory (RAM), and GPU – of the Raspberry Pi 3B+ with those of four other popular single-board computers across various price points. The table demonstrates that although the specifications of the Raspberry Pi 3B+ are slightly less powerful than those of comparable options, it still performs adequately for running RespRisk. This choice allows for a reduced overall cost of the device.

As illustrated in Figure 5, RespRisk can generate predictions in approximately 0.18 seconds after receiving input. The Raspberry Pi is capable of running the XGBoost model at specified intervals based on real-time weather conditions, thereby providing doctors with regular updates on asthma-related hospitalization risks. During testing, we successfully configured the model to operate at intervals of 30 seconds, 1 minute, and 5 minutes. In dangerous, time-sensitive situations, RespRisk can rapidly formulate predictions without significantly compromising response time.

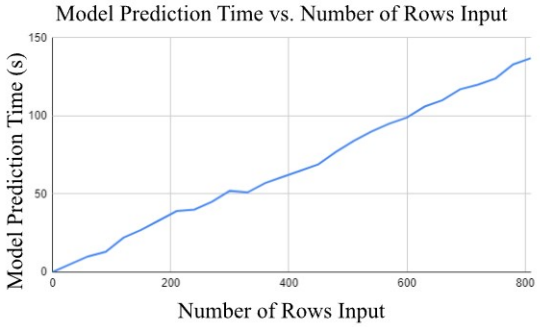

Fig. 5 Total Time Taken to Generate Predictions as a Function of the Number of Rows Input

Limitations and Future Work

Our primary limitation was the dataset's scale. The mere 169 training data points limit our model's capability to capture the full complexity of environmental and health dynamics. Our reliance on public data, while necessary due to strict privacy regulations on most healthcare data, further constrained our analysis to broader trends, potentially overlooking localized health impacts. The collection of more such health data in the present can allow future AI models to learn at an even deeper level.

Other limitations include the geographical limitation to New York. These findings may not be generalizable to other regions

Table 5 Performance Comparison between Four Popular Single-Board Computers

Model	Price	CPU	RAM	GPU
Qualcomm DragonBoard 410C	$132.81	Cortex-A53 4x 1.2 GHz	1 GB	Qualcomm Adreno 306 400 MHz
Rock64 (4GB)	$79.99	Rockchip RK3328 4x 1.5 GHz	4 GB	ARM Mali-450MP2 450 MHz
Raspberry Pi 3 B+	$35.71	Broadcom BCM2837 4x 1.2 GHz	1 GB	Broadcom VideoCore IV 250 MHz
ODROID-XU4	$49.00	Samsung Exynos 5 Octa 5422 4x 2 GHz	2 GB	ARM Mali-T628 MP6 600 MHz

with different environmental and healthcare contexts. Additionally, fitting the model to operate on a Raspberry Pi requires compromises in model complexity and, subsequently, accuracy. Lastly, the ongoing but relatively new integration of AI in healthcare, especially for predictive modeling of hospitalization rates, limited the number of established methodologies we could use as a reference for the proficiency of our model. This reveals the compelling need to expand the scope and depth of data collection in the domain of healthcare as a whole. Using a larger and more diverse dataset, which encompasses different geographical locations and more granular health data, could significantly enhance the model's accuracy and applicability. Furthermore, the addition of an ozone sensor allows the device to run with no internet connection so it can help people in remote locations. Additionally, various factors other than air pollution, such as average income, education level, weather conditions, and patient demographics, have been proven to have a significant correlation with higher rates of health crises[11]. The absence of these predictors in the model undermines its ability to comprehensively predict hospitalizations by taking into account all possible indicators of increased hospitalization rates. These factors were intentionally excluded in RespRisk to focus the study on investigating the effects of air pollution on asthma-related hospitalizations and providing an accessible tool, as developing nations may lack the resources to measure and include all these additional features in the model. Testing the model in different regions and hospitals would provide further insight into its ability to generalize to different datasets, and the usage of hourly or daily data rather than monthly data could help account for seasonal pollution concentration fluctuations. However, the 2-3-day lag between air pollution surges and asthma-related hospitalizations may be a flaw in this approach. Lastly, in addition to using weather information, browsing the internet for keywords like "asthma" or "inhaler" could be added to enhance the accuracy of the prediction system as well.

This project's significance lies in its ability to ensure hospital preparedness. In scenarios such as having a wildfire or high winds nearby, this model could prepare hospitals to allocate more resources and doctors for asthmatics and, in the future, other sensitive groups. With this model, hospitals would be able to prepare sooner and act effectively when hospitalizations do happen, saving as many lives as possible and reducing the burden on the healthcare system as a whole. While it is established that asthma incidence increases with severe air pollution, RespRisk presents a more advantageous alternative to merely monitoring weather conditions for healthcare professionals. Certain pollutants exhibit a stronger correlation with increased asthma-related hospitalizations, as illustrated in Figure 4. This complexity adds another layer for physicians to consider when manually predicting the expected number of asthma-related hospitalizations based on weather forecasts. Furthermore, given the constant demands and stress associated with hospital work, utilizing RespRisk to obtain a numeric prediction is significantly more efficient than consulting multiple weather forecasts for various air pollution indicators. Lastly, if RespRisk forecasts 20,000 asthma-related hospitalizations for a particular county, an individual hospital can estimate that it will need to prepare for approximately 400 cases based on its historical share of county-wide admissions. This targeted prediction enables hospitals to better anticipate and manage patient influx, thereby improving response efficiency and healthcare outcomes. This work can prevent the extended suffering and even death of an asthmatic individual due simply to limited hospital availability. Additionally, given sufficient notice, hospitals can save up to 5% of the costs associated with an unexpected surge in hospitalization rates via early preparation and resource allocation[12].

Methods

To create the model, we first investigated the correlation between annual asthma hospitalizations and criteria pollutant concentrations in New York State. We chose New York State due to its extensive public healthcare data availability, which is crucial for reliable analysis, and its historically high rates of asthma attacks, which makes it an ideal case study for developing predictive models. We compared annual records of asthma hospitalizations with ambient levels of NO_2, CO, PM2.5, and O_3 recorded by the EPA. After concatenating EPA air quality data for 2000-2020 with asthma hospitalizations per 10,000 people, we used three regression modalities to predict hospitalization rates.

Dataset Collection

The studied region of New York State is a densely populated state in the northeastern United States. Annual rates of asthma-related hospitalizations from 2000-2020 were collected by the New York State Department of Health (NYSDH). The data contain values for annual hospitalizations for New York State

sorted by age, ethnicity, and other strata. We define asthma-related hospitalizations as the total number of asthma-related hospitalizations in New York State per month for all ages and ethnicities per 10,000 people. We sourced hospitalization data from all licensed New York State hospitals, and air pollution data through EPA-managed air quality monitors in New York State from 2000-2020. Because asthma-related hospitalization data in New York State is recorded monthly, we averaged daily data for each air pollutant from all monitors across New York State to create a single representative monthly pollutant concentration value per pollutant from 2000-2020. We then averaged all daily pollution concentration values in a given month to form a single representative pollutant concentration value for each month from 2000-2020. This pollutant dataset was then concatenated to the asthma-related hospitalization dataset, forming a monthly dataset of hospitalization rates per 10,000 people and corresponding monthly air pollutant concentrations for New York State from 2000-2020.

Dataset Preprocessing

To remove outliers, data points with a z-score exceeding +3 or dropping below -3, relative to their respective feature distributions, were identified and excluded. We also utilized the Interquartile Range (IQR) method for the same purpose, which involved removing data points beyond 1.5 times the IQR above the third quartile or below the first quartile. This preprocessing step was to ensure that we trained with quality data by removing outliers, which decreased the total number of rows in the dataset from 252 to 241 rows. We also averaged 30 daily air pollution quantities to create monthly representative values, paired with corresponding monthly asthma-related hospitalization values.

Model Variations

We employed various Machine Learning (ML) models and techniques to analyze a small dataset and identify the most effective model for the task at hand. This paper details the three most significant methodologies used and their consequent results.

Due to our small dataset consisting of 241 total rows, we analyzed the effects of synthetic data generation via a custom-trained CTGAN model for all three methodologies to provide more examples for the model to train on. To determine the optimal data generation model, we created multiple model versions, each distinguished by its specific training duration, ranging from 200 to 2000 epochs. We evaluated the performance of each model through its Generator and Discriminator loss graph (Figure 6).

Additionally, the Synthetic Data Vault library was utilized to generate synthetic data to augment the original training data. We trained a CTGAN generator on the authentic training data for 2,000 epochs, noting that the Discriminator loss and Generator loss stabilized and became generally parallel around the 600th epoch. We therefore utilized 600 epochs to train the generator model, and then used the trained generator to produce synthetic data.

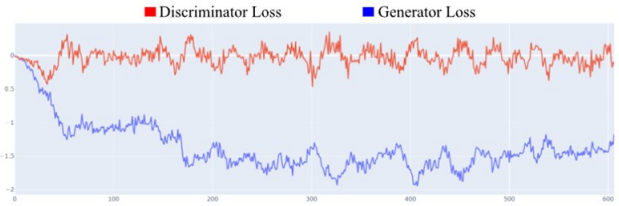

Fig. 6 Results of Concatenating Different Quantities of Synthetic Data to the Training Dataset

We also used the Synthetic Data Vault library to assess the quality of the synthetic data produced, and the highest quality synthetic data was used to train the RespRisk model. Quality tests included utilizing the Kolmogorov–Smirnov test to quantify discrepancies in numerical distributions, the Total Variation Distance to compare categorical distributions, and the Correlation Similarity between different columns. The synthetic data's fidelity to the true data increased linearly as a function of the CTGAN model's number of training epochs. Our synthetic data achieved a 94.66% statistical similarity to the authentic training data based on these statistical tests, proving that it is high-quality and effectively replicates the trends of the authentic training data. We concatenated 300 rows of the synthetic data points to our training dataset, after using a 70%-15%-15% split for training, validation, and testing data, respectively. The original split resulted in 169 training data points, 36 validation data points, and 36 testing data points. Concatenating synthetic data raised the total number of training data points, while the testing and validation datasets remained composed of solely authentic data.

We examined three different methodologies – linear regression, neural network regression, and XGBoost regression – along with the use of synthetic data and their effects on model performance. Our first approach utilized a simple multivariate regression model, which formed a linear function to describe the relation between air pollutant concentrations and asthma-related hospitalizations. Our second approach employed a neural network regression model, which, instead of a best-fit line, constructed a complex, nonlinear function for the same purpose. Our third approach utilized the XGBoost regression model, a powerful state-of-the-art method that employs gradient-boosting techniques for predictive analysis [13]. This model is a novel take on the traditional decision tree framework, where each new tree attempts to correct the errors of the previous one. XGBoost stands out for its ability to handle various types of datasets and work with a variety of data types, proving more accurate at hospitalization rate prediction than all other models tested.

Final Model and Hyperparameter Tuning

Due to its high accuracy, we selected the XGBoost Regression model for RespRisk and tuned the model's hyperparameters, in which all possible combinations of provided values for each hyperparameter were tested to find the ideal set of parameters. We present all the hyperparameters tuned, along with a brief explanation of each in Table 6. With these hyperparameters, we evaluate the RespRisk model and fit it into the Raspberry Pi.

The XGBoost Regression model uses the mean squared error loss function, as shown in Equation 1. This loss function measures the average squared difference between predicted and actual values, meaning that a lower mean squared error is better.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

Device Construction

To create RespRisk, we attached the MICS-6814 air quality sensor to provide NO_2 and CO concentration readings, and the SDS011 sensor to provide PM2.5 concentration readings, to a Raspberry Pi. A Raspberry Pi is an affordable, single-board minicomputer that can utilize AI models to make predictions using data from various sensors. We connected the VK-172 Global Positioning System (GPS) module for location data, allowing the use of Google's Air Quality Application Programming Interface (API) to access local Ozone concentrations. The VK-172, MICS-6814, and SDS011, sourced from Viking, SGX Sensortech, and Nova, respectively, are reliable and commonplace sensors that have undergone extensive evaluation and have been proven reliable [14]. This GPS and air quality data, measured by the VK-172, SDS011, and MICS-6814, is transmitted by Bluetooth to Adafruit IO's data storage system, where it can be downloaded in real-time and inputted to the RespRisk model (Figures 7 and 8).

Device Cost

The components used to create RespRisk were chosen for their ability to provide accurate measurements for a lower cost, making RespRisk more accessible. The Raspberry Pi, case, and breadboard cost $34, forming the core of the device. The Pimoroni Enviro+ Sensor, used for measuring SO_2 and NO_2, also costs roughly ∼$35. Additional components include an I2C Liquid Crystal Display (LCD) Module for $10, a Viking VK-172 GPS module for $7, the Nova SDS011 Sensor costing $11, and wiring, bringing the final cost to $99.60.

Conclusion

Despite data constraints, the RespRisk model proves its proficiency in effectively predicting local asthma-related hospitalizations based on air pollutant levels. The model evaluation – an RMSE of 0.33, MSE of 0.23, MAE of 0.10, and MAPE of 0.19 – has verified this hypothesis, particularly highlighting NO_2's strong correlation with asthma hospitalizations, which was an interesting finding that aligned with prior studies. The relatively low MAE and high MSE is explained by the model's tendency to sometimes severely overestimate hospitalization rates, creating outliers punishable by the MSE function. The practical usage of this work is in its potential to help hospitals with resource allocation and preparedness. This mitigates the risk of hospital overcrowding and also asthma attacks. The exclusion of socioeconomic factors from the model and the improvable quality of the dataset represent limitations for this work. Thus, future work on this project will focus on integrating socioeconomic variables into the model and reducing random overestimation by adding training data and utilizing a better system for data collection to predict rates on a daily or hourly basis based on air pollution. Overall, this project provides a groundwork for the field of using AI to predict health crises, demonstrating the connection between environmental factors and public health.

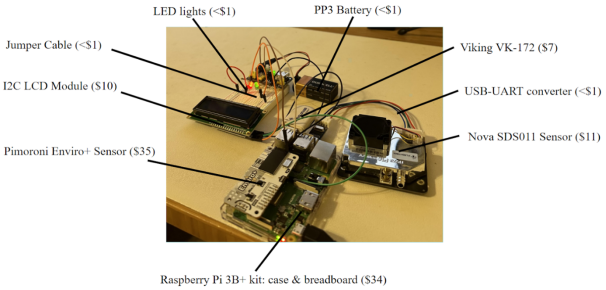

Fig. 7 RespRisk model hardware

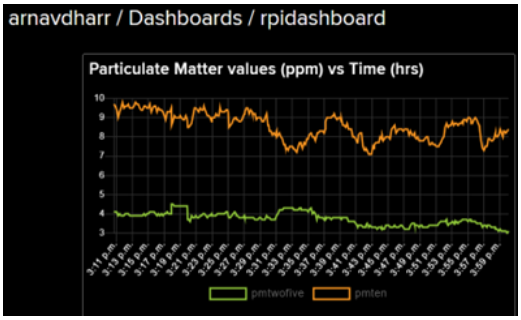

Fig. 8 Particulate Matter Data Recorded by SDS011 Sensor in Adafruit IO System

References

1 Z. Wang, Y. Li, Y. Fu, J. Lin, X. Lei, J. Zheng and M. Jiang, *Respiratory Research*, 2023, **24**, 1–13.

2 J.-B. Wasserfallen, M.-D. Schaller, F. Feihl and C. H. Perret, *The American Review of Respiratory Disease*, 1990, **142**, 108–11.

Table 6 XGBoost Hyperparameters and Specific Values Chosen for RespRisk Model

Hyperparameter	Explanation	Chosen Value
Feature Subsampling	Determines the fraction of columns used for each tree, affecting diversity in the model	0.6
Minimum Loss Reduction	Sets the minimum loss reduction for further splits, influencing tree complexity	0.2
Learning Rate	Controls the weighting of new trees, affecting model convergence and overfitting	0.15
Tree Depth	Determines the maximum depth of each tree, impacting model complexity and overfitting	5
Minimum Sum of Instance Weight (Hessian)	Sets the minimum sum of instance weight needed in a child node, guiding decision-making in tree creation	3
Number of Trees	Specifies the number of trees in the ensemble, affecting the model's capacity and potential overfitting	170
L1 Regularization	Alpha, contributing to higher model sparsity	0.1
Row Subsampling	Fraction of samples used for fitting individual trees, promoting diversity in the model	0.5

3 A. Tiotiu, P. Novakova, D. Nedeva, H. J. Chong-Neto, S. Novakova, P. Steiropoulos and K. Kowal, *International Journal of Environmental Research and Public Health*, 2020, **17**, year.

4 S. A. Meo, A. A. Abukhalaf, O. M. Alessa, A. S. Alarifi, W. Sami and D. C. Klonoff, *International Journal of Environmental Research and Public Health*, 2021, **18**, year.

5 R. J. Fiter, L. J. Murphy, M. N. Gong and K. L. Cleven, *Expert Review of Respiratory Medicine*, 2023, **17**, 1237–47.

6 H. Hwang, J.-H. Jang, E. Lee, H.-S. Park and J. Y. Lee, *Respiratory Research*, 2023, **24**, year.

7 K. Moore, R. Neugebauer, F. Lurmann, J. Hall, V. Brajer, S. Alcorn and I. Tager, *Environmental Health Perspectives*, 2008, **116**, 1063–70.

8 L. J. Akinbami, C. D. Lynch, J. D. Parker and T. J. Woodruff, *Environmental Research*, 2010, **110**, 294–301.

9 W. Shen, Y. Ming, T. Zhu and L. Luo, *Annals of Translational Medicine*, 2023, **11**, year.

10 P. L. Delamater, A. O. Finley and S. Banerjee, *The Science of the Total Environment*, 2012, **425**, 110–118.

11 A. L. Kozyrskyj, G. E. Kendall, P. Jacoby, P. D. Sly and S. R. Zubrick, *American Journal of Public Health*, 2010, **100**, 540–46.

12 K. Gribben, H. Sayles, S. Roy, R. J. Shope, J. S. Ringel and S. Medcalf, *Health Security*, 2020, **18**, 409–17.

13 T. Chen and C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–94.

14 M. Badura, P. Batog, A. Drzeniecka-Osiadacz and P. Modzel, *Hindawi*, 2018, **2018**, 1–16.