ADBA
Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.37-48
https://doi.org/10.69882/adba.ai.2026016

Predicting Upper Respiratory Tract Infections: The Role of Weather Data and Explainable AI

Türker Berk Dönmez[*,1], **Mustafa Kutlu**[α,2] and **Chris Freeman**[β,3]

[*]Sakarya University of Applied Sciences, Faculty of Technology, Department of Biomedical Engineering, Sakarya 54050, Turkey, [α]Department of Mechatronics Engineering, Sakarya University of Applied Sciences, Sakarya, Türkiye, [β]University of Southampton, School of Electronics and Computer Science (ECS), Southampton SO17 1BJ, United Kingdom.

ABSTRACT Upper Respiratory Tract Infections (URTIs) are a global health issue, affecting myriad individuals and encompassing infections of the nose, sinuses, pharynx, or larynx. The diverse symptoms and varying severity of URTIs, coupled with their potential to be influenced by meteorological factors, underscore the importance of understanding the interplay between weather conditions and URTI incidence. This research, conducted in the Pamukova District from December 15, 2020, to December 31, 2022, delves into this relationship by integrating weather data from Meteoblue and patient data from the Pamukova Family Medicine Center. The comprehensive data cleaning, harmonization, and preprocessing considered the 3 or 5 preceding days in alignment with the URTI incubation period. Utilizing the Catboost machine learning model on two separate datasets, the study revealed enhanced performance with a 5-day data frame. The model yielded 67 true positives, 24 true negatives, 8 false positives, and 22 false negatives, resulting in an F1-score of 0.6154, an accuracy of 75.21%, precision and recall values of 0.75 and 0.5217, respectively, and an AUC value of approximately 0.7768. These results emphasize the critical role of an extended temporal frame in understanding the connection between environmental factors and URTI incidence, offering substantial insights for the development of targeted public health interventions in the Pamukova District.

KEYWORDS
Explainable AI
Upper respiratory tract infections
Weather data

INTRODUCTION

When the upper respiratory tract is thought of, the anatomical formations located in the head and neck and above the thorax, especially the nose, paranasal sinuses, nasopharynx, oropharynx and larynx, are thought of (Tu *et al.* 2013). Viral infections of the upper respiratory tract, also known as the common cold (acute chorea, cold, common cold), are the most common respiratory diseases in humans. They are occurred on average 2-4 times a year in adults and every year in children. URTIs, which can be seen 6-10 times a year on average, are usually self-limiting and rarely lead to complications (Derbyshire and Calder 2021).

URTIs may be brought on by a variety of diseases, including bacterial, viral, and fungal infections (Kolawole and Idris 2020). The rhinovirus and coronavirus are two of the viruses that are most commonly discovered in connection with URTIs. They may be spread by direct contact with respiratory droplets from ill persons or by touching contaminated surfaces and then coming in contact with the face. The frequency of URTIs might also be impacted by the weather and other environmental variables (Stolz *et al.* 2019). URTIs are placing a heavy strain on both the healthcare system and the general public. These infections can raise the chance of acquiring later-life illnesses like wheeze, asthma, as well as recurring doctor visits, parental stress, and other health issues (Chen *et al.* 2021).

In Turkey, URTI is the most common condition among children and adults, according to the Turkey Health Survey 2022. The results of the survey show that in the previous six months, URTI accounted for 31.3% of illnesses seen in children aged 0–6 years and 27.1% of illnesses seen in children aged 7–14 years. Furthermore, with a prevalence of 9.6% in the previous 12 months, URTI ranks fourth among the illnesses experienced by people over the age of 15 years. According to this findings, URTI is a significant public health issue in Turkey, and preventive and therapeutic efforts are required to lessen the burden it places on the populace .

Many academics have been interested in the connection between weather and URTIs which are common conditions that can be brought on by a variety of bacterial or viral agents and affect the nose, sinuses, throat, larynx, or trachea (Zeru *et al.* 2020). The productivity, quality of life, and cost of healthcare of those who are affected by URTIs can all be significantly impacted. In order to create effective preventative and treatment plans, it is crucial to understand the factors that affect the frequency and severity of URTIs. As they can modify the host's immune response, the pathogen's survival and transmission, and the exposure to other respiratory irritants and allergens, weather conditions are one of the environmental factors that may alter the risk and outcome of URTIs (D'Amato *et al.* 2018). The epidemiology and pathophysiol-

Manuscript received: 5 November 2025,
Revised: 26 December 2025,
Accepted: 18 January 2026.

[1]turkerberkdonmez@yahoo.com
[2]mkutlu@subu.edu.tr
[3] cf@ecs.soton.ac.uk (**Corresponding author**).

ogy of URTIs can be affected by a variety of meteorological factors, including temperature, humidity, precipitation, wind speed, and air quality (Zou *et al.* 2021).

The Eccles and Wilkinson (2015) research found a link between exposure to cold air and a greater frequency of URTI. Cold air may influence the nasal mucosa, the body's first line of defense against viral viruses, which might explain this. Cold air may slow down and make less efficient mucociliary clearance, the process of removing mucus and trapped particles from the respiratory system. Cold air may also impair the capacity of immune cells in the nasal cavity, such as macrophages, to ingest and destroy viruses. The link between cold and URTI has also been shown to be stronger in northern than in southern areas in UK research, indicating that the influence of cold air on URTI may vary by geographic location. However, rather than happening immediately after the temperature shift, the effect of cold air on URTI occurs two to three weeks later. The cold air also has less of an impact on URTI than it does on lower respiratory infections like pneumonia or bronchitis. In conclusion, exposure to cold air may impair the mucosal immune system, raising the danger of upper respiratory tract viral infections .

In order to learn more about the connections between various meteorological factors and URTIs and lower respiratory tract infections (LRTIs), Falagas et al. retrospectively analyzed meteorological and clinical data from the Attica region of Greece. The incidence of URTIs was observed to positively correlate with cold weather conditions, peaking when the weekly average temperature fell below 10 degrees Celsius, according to the researchers. In addition, they discovered that LRTIs were more commonly associated with chilly temperatures than URTIs were. The association between cold weather and URTIs, according to researchers, is caused by a number of factors, including direct effects of cold on the viability and infectivity of the viruses that cause URTIs, indirect effects of cold on immune and respiratory system function, and direct effects of cold on people's behavior (Falagas *et al.* 2008).

A comprehensive analysis of the impact of meteorological and air pollution factors on respiratory diseases in Linyi, China, was conducted. According to the study, a 0.31 increase in the concentration of NO2 is associated with a rise in pneumonia cases. Similarly, increased levels of PM2.5 and PM_{10}—specifically, by 0.23 and 0.24, respectively—were linked to higher pneumonia incidence. Low temperature and humidity levels, particularly a decrease in daily average temperature and humidity, were associated with a reduction in chronic lower respiratory diseases and pneumonia cases. Conversely, these same factors increased the incidence of acute upper respiratory infections by 0.04 and 0.05. High wind speeds also correlated positively with respiratory diseases. The SVR model used in the study showed a significant prediction potential, with an R^2 value of 0.308 for pneumonia, highlighting the intricate relationship between environmental factors and respiratory health (Yang *et al.* 2023).

Lim *et al.* (2023) conducted a comprehensive study on forecasting URTIs using high-dimensional time series data and forecast combinations. Their research indicated that a 1-week lag in lower temperature is associated with a significant increase in URTI attendances. Similarly, past relative humidity and absolute humidity levels showed notable effects on URTI forecasts. For example, a 1% increase in relative humidity decreased URTI attendances by approximately 3-4%, while an increase in absolute humidity at longer forecast windows (4-8 weeks ahead) was associated with a decrease in URTI attendances. The study also highlighted the superior predictive performance of forecast combinations, with mean absolute percentage errors ranging from 10% to 25% across different horizons. These findings emphasize the intricate relationship between climatic factors and URTI incidence, providing valuable insights for public health resource planning and outbreak preparedness .

Jhuo *et al.* (2019) conducted a comprehensive study on predicting trends in influenza and associated pneumonia in Taiwan using machine learning. Their research utilized meteorological parameters, such as temperature and relative humidity, and air pollution parameters, including PM 2.5 and CO, alongside the number of acute upper respiratory infection (AURI) outpatients as inputs. They used data from December 2009 to December 2017 and made predictions for January 2010 to January 2018. Patients were categorized into low, moderate, and high volume levels. The multilayer perceptron (MLP) model developed in their study achieved an accuracy of 81.16% for the elderly population and 77.54% for the overall population. The study found that larger data sets from bigger areas improved accuracy, whereas lower accuracy was observed for children aged 0-4 years due to fewer samples and less exposure to environmental factors. These findings underscore the intricate relationship between environmental factors and the incidence of influenza and pneumonia .

A thorough investigation of the effects of climatic factors on URTIs was undertaken by Makinen et al. According to their research, a $1^{\circ}C$ drop in temperature is associated with a 4-5% rise in URTI incidence. Similar to this, low humidity levels—more precisely, those below 40%—were linked to an increase in instances. High wind speeds were similarly linked to an increased risk of URTIs despite having received less research. Additionally, it was proposed that a minor increase in infections may be associated with situations of declining barometric pressure. These observations highlight the complex interaction between weather and the frequency of URTIs (Mäkinen *et al.* 2009).

Kern *et al.* (2016) explored the relationship between weather data and the incidence of ophthalmological conditions using model-agnostic methods. Through Spearman's correlation analysis, they examined clinical data from the University Eye Hospital Munich from January 2014 to July 2015. They linked patient visits to weather variables like sunshine duration, temperature, and wind speed, finding a weekly increase of one sunshine hour correlated with an additional patient visit per week ($\rho = 0.44$, $P < 0.01$). Temperature increase of 1°C correlated with 2.6 more patients per week ($\rho = 0.29$, $P < 0.01$). Specifically, higher temperatures and longer durations of sunshine were positively correlated with increased visits for conditions like conjunctivitis and foreign body injuries. The model-agnostic approach allowed them to uncover significant correlations without being constrained by underlying data structure assumptions .

Santhanam *et al.* (2024) extended the application of model-agnostic methods by incorporating machine learning models to predict daily acute ischemic stroke (AIS) admissions based on weather data. Employing techniques such as Support Vector Machines (SVR), Random Forests (RF), and Extreme Gradient Boosting (XGB), they effectively managed the complex, nonlinear relationships between environmental factors and health outcomes. Their study identified maximum air pressure as a critical predictive variable, with extreme temperature conditions and stormy conditions also playing significant roles. The XGB model's robust predictive capability was evidenced by a low mean absolute error (MAE) of 1.21 cases/day on the test set, supporting better healthcare resource allocation and preparedness .

Mansour *et al.* (2023) employed the Lorenz equation and numerical techniques like the Runge-Kutta method to develop a novel chaotic system for forecasting respiratory disease outbreaks, using a model-agnostic approach to integrate weather variables such as maximum temperature, air pressure, and humidity with patient data from the Pamukova Region. By utilizing a NARX network for input-output data processing, they established a high correlation coefficient of 90.16% between predicted and actual patient numbers. Their findings suggest a robust framework for employing chaotic systems in real-time health warning systems, potentially enhancing preemptive responses to environmental health risks. This model-agnostic methodology underlines the adaptability of chaotic systems in predicting complex health-related events .

In this study, the necessity is underscored by the global health impact of URTIs and the limited understanding of how weather conditions influence their incidence. By focusing on the Pamukova District, this research provides localized insights using model-agnostic SHAP (SHapley Additive exPlanations) values to reveal how specific weather conditions affect URTI patients. Integrating weather data with patient records and utilizing advanced machine learning models, the study highlights the importance of considering an extended temporal frame for accurate predictions. These findings are crucial for developing effective public health interventions and offer significant insights that can be applied to similar contexts globally, demonstrating how studies in smaller provinces can contribute to the bigger picture of improving public health outcomes.There is still a need for a more thorough understanding that takes into account regional variations, seasonal patterns, and potential interactions with other health determinants. The existing literature has given valuable insights into how specific meteorological factors affect the prevalence of URTIs (Mansour *et al.* 2023). In order to fully understand the epidemiology of URTIs, which continue to place a heavy strain on healthcare systems and communities, a comprehensive approach is required. This study contributes to the creation of more effective public health treatments and policies meant to lessen the effects of URTIs by investigating these aspects holistically.

MATERIALS AND METHODS

Data Collection and Preprocessing

Data on URTIs from the Pamukova District were combined with corresponding meteorological data. The datasets were then cleaned, harmonized, and assessed for outliers before being analyzed.

Weather Data Meteoblue was used to gather weather information, which included metrics for maximum, minimum, and mean temperatures, sunlight duration, shortwave radiation levels, precipitation, snowfall, humidity levels, cloud cover, air pressure, and wind speeds. The dataset is exclusive to the Pamukova area and runs from December 15, 2020, through December 31, 2022.

Patient Data Patients' information was gathered for this study with Sakarya University's ethical permission (E-71522473-050.01.04-15185-157). The research, which covers the period from January 1, 2021, to December 31, 2022, is concentrated on the Pamukova District in Sakarya Province. The information was given by the Pamukova Family Medicine Center, which is closed on weekends and major holidays. The clinic treated 52,792 patients in total during the course of 484 workdays, 4,454 of whom had upper respiratory tract infections (ICD codes J09–J18). As a result, during the course of the trial, URTIs accounted for around 9.2% of all patient visits.

Preprocessing Data and Model Training The weather data underwent various preprocessing. The averages, standard deviations, and value gaps of the data from the previous 3 or 5 days, including the current day, were calculated. This approach was chosen because the average incubation period for any URTI agent varies between approximately 1-5 days. The average number of patients was calculated as approximately 9.2 patients per day. The number of patients distributed over the days was classified binarily as above average and below average and was determined as the main target.

The study utilized a range of meteorological attributes to predict Upper Respiratory Tract Infections (URTIs). These attributes include the mean temperature (*meantemp*), mean humidity (*meanhumidity*), mean pressure (*meanpressure*), mean wind speed (*meanwind*), and mean sunshine duration (*sunshine*) measured over the last 3 or 5 days. Additionally, shortwave radiation (*radiation*), total precipitation (*precipitation*), and snowfall amount (*snowfall*) were considered. Cloud cover (*cloudcover*) was also included, along with calculated variables such as the standard deviation of min-max values (*minmaxSDtemp*, *minmaxSDhumidity*, *minmaxSDpressure*, *minmaxSDwind*), the standard deviation of average values (*meanSDtemp*, *meanSDhumidity*, *meanSDpressure*, *meanSDwind*), and the value range of the highest and lowest values (*VRtemp*, *VRhumidity*, *VRpressure*, *VRwind*) over the specified period. The *minmaxSD* values were calculated by measuring the standard deviation of the minimum and maximum values on a daily basis within the specified period.

Twenty-one different variables were created for these two separate datasets and are shown in detail in Table 1. These two datasets were evaluated with various machine learning models as shown in Table 2, and among them, the Catboost model showed the highest success.

75% of the data for each of the two sets is used for training and 25% is utilized for testing, resulting in a 75/25 split of the data. Additionally, the "Discussions" section includes findings that shed light on the model's stability.

CatBoost has been chosen as the main model for more investigation. CatBoost stands out for its practical efficiency and simplicity of use, qualities that are especially well-aligned with the study goals, even if other models show equivalent performance. CatBoost is thought to be the best solution for the challenges associated with illness forecasting using meteorological data because it combines great computing speed with powerful predictive skills.

Categorical Boosting (CatBoost) CatBoost, which stands for *Category Boosting*, is a gradient boosting library developed by Yandex. It has gained popularity in the machine learning community for its performance and its built-in support for categorical features, thus eliminating the need for extensive preprocessing like one-hot encoding or label encoding.

The mathematical foundation of CatBoost is rooted in the gradient boosting framework. The primary objective of gradient boosting is to optimize a cumulative objective function, which is a sum of a loss function and a regularization term (Prokhorenkova *et al.* 2018). CatBoost introduces several enhancements to the traditional gradient boosting technique:

$$\mathcal{L}(\mathbf{y}, \mathbf{F}) = \sum_{i=1}^{N} l(y_i, F(x_i)) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

Where $\mathbf{y}$ is the vector of true labels, $\mathbf{F}$ is the ensemble model, l is a differentiable convex loss function, f_k are the individual trees, and Ω is a regularization term.

Table 1 Attribution of the dataset for 3 or 5-day metrics

No.	Abbreviation	Name of the attribute	Units
1	meantemp	Last 3 or 5-day mean temperature	℃
2	meanhumidity	Last 3 or 5-day mean humidity	%
3	meanpressure	Last 3 or 5-day mean pressure	hPa
4	meanwind	Last 3 or 5-day mean wind speed	km/h
5	sunshine	Last 3 or 5-day mean sunshine duration	min
6	radiation	Last 3 or 5-day mean shortwave radiation	W/m^2
7	precipitation	Last 3 or 5-day mean total precipitation	mm
8	snowfall	Last 3 or 5-day mean snowfall amount	cm
9	cloudcover	Last 3 or 5-day mean total cloud cover	%
10	minmaxSDtemp	Standard deviation of min-max temperature values in the last 3 or 5 days	Calculated
11	meanSDtemp	Standard deviation of the average temperature over the last 3 or 5 days	Calculated
12	VRtemp	Range of the highest and lowest temperature values in the last 3 or 5 days	Calculated
13	minmaxSDhumidity	Standard deviation of min-max humidity values in the last 3 or 5 days	Calculated
14	meanSDhumidity	Standard deviation of the average humidity over the last 3 or 5 days	Calculated
15	VRhumidity	Value range of the highest and lowest humidity values in the last 3 or 5 days	Calculated
16	minmaxSDpressure	Standard deviation of min-max pressure values in the last 3 or 5 days	Calculated
17	meanSDpressure	Standard deviation of the average pressure over the last 3 or 5 days	Calculated
18	VRpressure	Value range of the highest and lowest pressure values in the last 3 or 5 days	Calculated
19	minmaxSDwind	Standard deviation of min-max wind speed values in the last 3 or 5 days	Calculated
20	meanSDwind	Standard deviation of the average wind speed over the last 3 or 5 days	Calculated
21	VRwind	Value range of the highest and lowest wind speed values in the last 3 or 5 days	Calculated

A standout feature of CatBoost is its treatment of categorical features. The algorithm leverages a technique called ordered boosting, which involves random permutations to prevent overfitting. Another notable method is mean encoding, where categories are replaced with the average target value for that category, with certain regularization techniques applied to avoid overfitting.

For model interpretation, CatBoost offers built-in support for SHAP values, making it easier to explain the predictions and understand feature importances. This integration of SHAP values is particularly beneficial as it offers a consistent methodology for model interpretation without needing external tools (Chelgani *et al.* 2023).

In practice, CatBoost has proven to be competitive with other gradient boosting implementations, often outperforming them, especially when dealing with datasets with a high number of categorical features. Its efficiency, coupled with its user (Bentéjac *et al.* 2021). Table 3 in this research displays the hyperparameters for Catboost that were chosen.

Table 2 Model Metrics and Confusion Matrices

Model	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
CatBoost	0.7521	0.7500	0.5217	0.6154	67 8 22 24
XGBoost	0.7273	0.6667	0.5652	0.6118	62 13 20 26
Extra Trees	0.7273	0.7097	0.4783	0.5714	66 9 24 22
Random Forest	0.7107	0.6571	0.5000	0.5679	63 12 23 23
LightGBM	0.6942	0.6047	0.5652	0.5843	58 17 20 26
Explainable Boosting Machine	0.6860	0.6111	0.4783	0.5366	61 14 24 22
Logistic Regression	0.6777	0.6061	0.4348	0.5063	62 13 26 20
Adaboost	0.6446	0.5385	0.4565	0.4941	57 18 25 21
Decision Tree	0.6281	0.5116	0.4783	0.4944	54 21 24 22
KNN	0.6281	0.5111	0.5000	0.5055	53 22 23 23
Support Vector Machine	0.6198	0.5000	0.3478	0.4103	59 16 30 16
Naive Bayes	0.6116	0.4915	0.6304	0.5524	45 30 17 29

Table 3 Hyperparameter values for the CatBoost model

Hyperparameter	Value
iterations	1000
depth	6
l2_leaf_reg	3.0
model_size_reg	0.5
border_count	254

Model Interpretation with SHAP Shapley Additive Explanations (SHAP) provides a robust methodology for understanding and interpreting the output of any machine learning model. Drawing its foundation from cooperative game theory, SHAP was proposed in 2017 with an ambition to unify the various methods of model interpretation. By allocating an "importance value" to each feature, SHAP gives an indication of how much each feature contributes to a given prediction. This methodology serves as a consistent and locally accurate lens through which we can understand model behavior (Kavzoglu and Teke 2022).

The mathematical foundation of SHAP is rooted in the Shapley value from cooperative game theory. To calculate the SHAP value for a particular feature, denoted as f_i, we consider a set of all features, F, and all the potential feature subsets, S, that can be created after removing the i-th feature. The equation is:

$$f_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|! \times (|F|-|S|-1)!}{|F|!}\left(f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right) \quad (2)$$

In this equation, $f_{S\cup\{i\}}$ and f_S represent the predictions of models trained with feature sets $S \cup \{i\}$ and S respectively. The terms

$x_{S\cup\{i\}}$ and x_S denote the values of the input features in the sets $S \cup \{i\}$ and S.

While SHAP provides a comprehensive methodology, the direct computation of Shapley values can be intensive, especially when dealing with a large number of features. To address this, SHAP offers approximations such as Shapley sampling and Shapley quantitative influence.

SHAP values can be interpreted from both global and local perspectives. On a global scale, features with consistently high absolute SHAP values across many samples generally have a greater influence on model predictions. Conversely, at the local level, for a given prediction, SHAP values provide information on the variables (Kannangara *et al.* 2022).

The versatility of SHAP is one of its standout attributes. It can be seamlessly applied to a multitude of models, ranging from decision trees and ensemble methods to neural networks. However, it's worth noting that the computational demand of SHAP can sometimes be a bottleneck, especially when the model has a vast number of features or when dealing with large datasets.

RESULTS

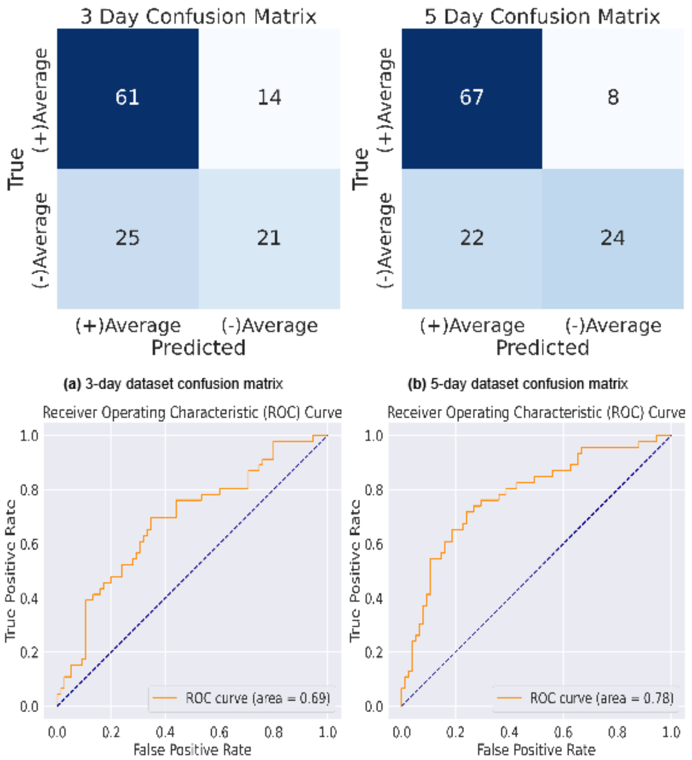
Results of Classification

In our quest to measure the potential of environmental variables in predicting the occurrence of Upper Respiratory Tract Infections (URTIs), we applied the Catboost machine learning model on two individually built datasets, incorporating both meteorological and patient data. For the dataset structured around metrics from the prior 3 days, the model defined 61 true positives and 21 true negatives, while misidentifying 14 and 25 examples as false positives and negatives, respectively. This resulted in an F1-score of around 0.519, an accuracy rate of 67.77%, and precision and recall values of 0.6 and 0.4565, respectively. Moreover, the model's AUC value, a vital statistic evaluating its discriminative power, was at around 0.6861. (Figure 1a and 1c)

Contrastingly, when the model was trained on data covering the prior 5 days, the results were considerably improved. The confusion matrix indicated 67 true positives, 24 true negatives, 8 false positives, and 22 false negatives. The measures indicated improvement across the board: an F1-score of 0.6154, an accuracy of 75.21%, and precision and recall values of 0.75 and 0.5217, respectively. The AUC value likewise experienced a boost, reaching roughly 0.7768. These findings underline the Catboost model's heightened competence with a longer 5-day data frame as compared to a 3-day one. It emphasizes the benefit of adopting an extended temporal frame while examining the link between environmental elements and URTI occurrences in the Pamukova District. Gleaning from this, specific public health interventions may be designed, aiming to limit the effect of URTIs on the population. (Figure 1b and 1d)

Explaining Model with SHAP

Understanding the choices of complicated models, such as Catboost, is vital for exposing the role of numerous environmental factors in forecasting Upper Respiratory Tract Infections. SHapley Additive exPlanations values give a complete measure of feature importance, providing a better comprehension of the model's decision-making process.



(a) 3-day dataset confusion matrix

(b) 5-day dataset confusion matrix

(c) 3-day dataset ROC Curve

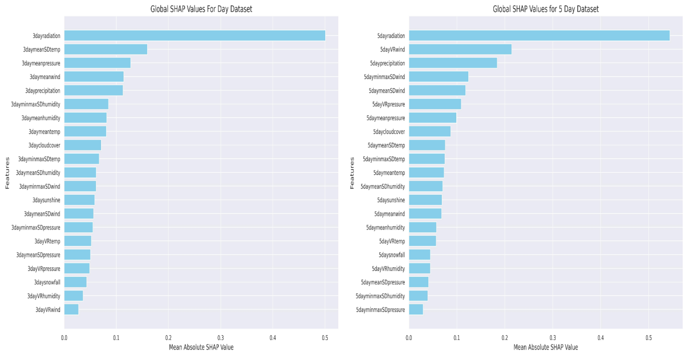
(d) 5-day dataset ROC Curve

Figure 1 Confusion Matrices and ROC Curves for 2 dataset

For the model developed using meteorological data from the prior three days, shortwave radiation emerged as the most significant feature with a SHAP value of 0.50, underlining its substantial influence on predicting Upper Respiratory Tract Infections. Other significant determinants included the variability in temperature over these three days with a SHAP value of 0.16, the average atmospheric pressure (SHAP value: 0.13), and the average wind speed (SHAP value: 0.11), further emphasizing the role of these environmental conditions in the model's predictions. Features including the length of sunlight (SHAP value: 0.06), the fluctuation in wind speed (SHAP value: 0.06), and snowfall during the same time (SHAP value: 0.04) also affected the model's outputs, albeit to a lesser amount as shown in Figure 2a.

When assessing the model trained with data covering the preceding five days, shortwave radiation consistently appeared as the dominating feature with a SHAP score of 0.54. This reinforces the persistent relevance of shortwave radiation levels in predicting Upper Respiratory Tract Infection prevalence over an extended duration. Other parameters, such as the variation in wind speed (SHAP value: 0.21), precipitation levels (SHAP value: 0.18), and the difference between the lowest and highest wind speeds over these five days (SHAP value: 0.12), also played a considerable part in the model's classifications. However, features like snowfall (SHAP value: 0.05), the variance in humidity (SHAP value: 0.04), and the difference between the minimum and maximum atmospheric pressures over this period (SHAP value: 0.03) exhibited a relatively reduced influence in the five-day dataset compared to the three-day one as shown in Figure 2b.

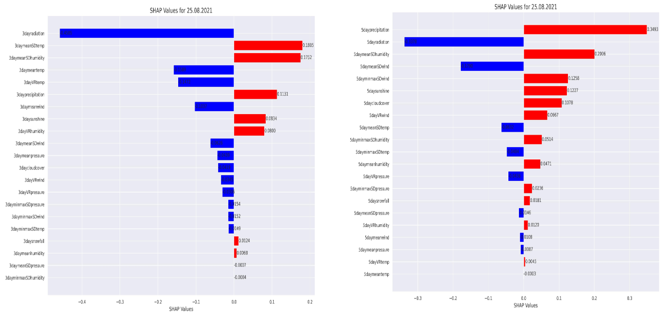
Drawing from these findings, it's obvious that although certain climatic factors, like shortwave radiation, continuously increase the incidence of Upper Respiratory Tract Infections, others change in their relevance dependent on the observational period. This sophisticated knowledge reveals the delicate association between environmental conditions and Upper Respiratory Tract Infection incidences, underlining the necessity for region-specific, data-driven interventions, especially in locations like the Pamukova District.



(a) Global SHAP Values for 3-Day Dataset

(b) Global SHAP Values for 5-Day Dataset

Figure 2 Global SHAP Values for 3-Day and 5-Day Datasets



(a) Local SHAP values for 25.08.2021 from 3-day dataset

(b) Local SHAP values for 25.08.2021 from 5-day dataset

Figure 3 Local SHAP values analysis for 25.08.2021

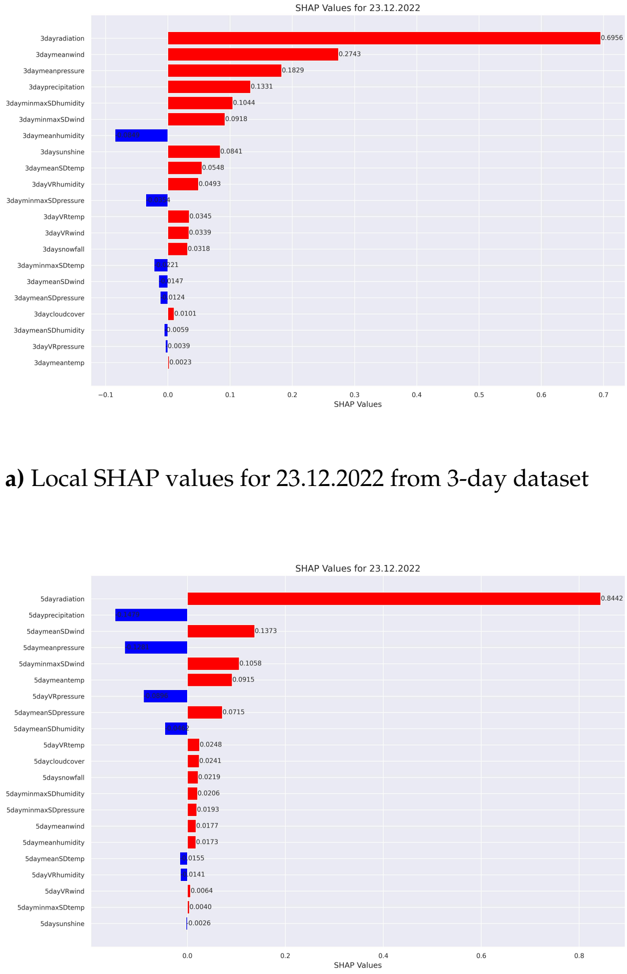
The predictions of machine learning models may be understood by SHAP values, since both global and local explanations can be offered by them. How each characteristic impacts the model's output on average is explained by global explanations, while how each feature affects the model's output for a given instance is indicated by local explanations. The 2 days on which common and accurate forecasts were provided by both models are investigated in this case: 25.08.2021 and 23.12.2022. These days are noteworthy since diverse seasons and weather conditions are represented by them, and it is desirable to understand how these differences were recorded by the models. The table provides meteorological readings and SHAP values of the data for these two days. The relative value of each attribute for the two models and the two days can be compared by evaluating the table, and which features contributed favorably or adversely to the predictions may be discovered. It can also be seen how the meteorological data impact the SHAP values, and how the nonlinear and interactive effects of the characteristics on the model's output are represented by them.

For 25.08.2021, placed in the warmer season, the model generated a True Negative forecast, suggesting that URTI diagnoses were below average—a prediction that was successfully determined by the model. Shortwave radiation on this day was significantly high, and its accompanying negative SHAP values of -0.46 (3-day model) and -0.34 (5-day model) imply that excessive radiation levels lowered the chance of URTIs. This makes shortwave radiation one of the most significant factors for this day's prognosis. Additionally, the lack of precipitation, as represented in its zero actual value, bore positive SHAP values of 0.11 (3-day model) and 0.35 (5-day model), designating it as another important contributor.

Wind speed and its variations also appear relevant. The average wind speed, with its SHAP value of 0.11 for the 3-day model, shows that wind could have had a role in raising URTI prevalence. Similarly, the variation in humidity across the three days, albeit tiny in real value, nonetheless exhibited a positive SHAP value of 0.05 (3-day model), showing its subtle effect on the model's predictions.

On 23.12.2022, a day indicative of the colder season, the model returned a True Positive prediction, accurately projecting an above-average URTI diagnosis. The reduced shortwave radiation value for this day, followed with its noticeable positive SHAP value of 0.84 (5-day model), clearly stands out as the most impactful variable, highlighting its crucial significance in the model's predictions. Atmospheric pressure, however constant across both models in actual values, displayed differing SHAP values, indicating its subtle effect.

Furthermore, the variation in wind speed for the 5-day model, with its SHAP value of 0.21, also emerges as an important driver, illustrating the model's sensitivity to varying wind conditions throughout this winter day.



a) Local SHAP values for 23.12.2022 from 3-day dataset

b) Local SHAP values for 23.12.2022 from 5-day dataset

Figure 4 Local SHAP values analysis for 23.12.2022

Piecing together these findings, it's evident that the model precisely evaluates many climatic parameters when forecasting URTIs. Shortwave radiation, precipitation, and wind-related factors constantly appear as the most relevant drivers, altering the model's predictions throughout various seasons. This research underlines the complex, multiple character of environmental determinants on health outcomes and underscores the need of examining a spectrum of environmental variables, particularly when creating tailored treatments for varied weather conditions and seasons.

SHAP Dependences for Variables

Using SHAP dependency plots, it was showed how the model output is impacted by critical meteorological factors. This variation is not formed entirely by an individual component but is also impacted by interactions with other weather-related variables. In each figure, the SHAP value and the individual variable value are depicted on the axes, and the feature with the most noticeable interaction impact is indicated by the color of the dots. Through these representations, crucial meteorological components were discovered, and their complicated interaction was recognized.

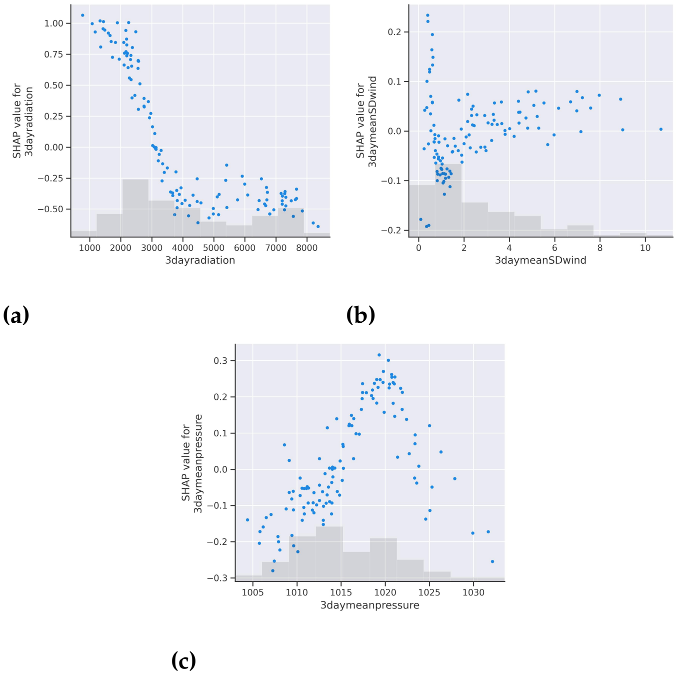
Both the 3-day and 5-day datasets were employed to produce the SHAP dependency plots, revealing insights into short-term weather patterns and their possible cumulative consequences. By merging information from both periods, a thorough knowledge of the climatic conditions' immediate and prolonged consequences was produced. However, it's crucial to remember that although an instructive summary is offered by these plots, they have not been submitted to in-depth statistical analysis. Their major objective is to illustrate the link between certain weather conditions and the predictions provided by our model.

A constant trend was found across both periods in the Shortwave Radiation dependency plots shown in Figures 6a and 6e for the 3-day and 5-day datasets respectively. Looking at both graphs, it is seen that values below 3000 W/m2 trigger an increase in the number of patients, while values above trigger a decrease in the number of patients. This discovery was confirmed by the remarkable SHAP values reported for shortwave radiation in both data sets.

In Figure 6f, the impact of average wind speed fluctuations on forecasts is shown by the Standard Deviation of Average Wind Speed in the Last 3 Days. Although there is no obvious sign of an increase in the standard deviation, it seems that the low wind speed change within 3 days triggered the number of patients. Its important role is emphasized by the corresponding SHAP values.

The significant impact of wind dynamics on model predictions is highlighted by the Value Range of Highest and Lowest Wind Speed Values in the Last 5 Days, as shown in Figure 6g. It has been observed that all "wind speed range" within 5 days being below 20 km/h triggers an increase in the number of diseases, while being above 30 km/h triggers a decrease in the number of patients. This discovery was supported by the prominent SHAP values in both datasets.

The importance of atmospheric pressure in shaping model outputs is illustrated in the Last 3-Day Average Pressure dependency plot, which can be seen in Figure 6h. Looking at the graph, it . as seen that the 3-day average pressure value being below 1015 hPa pushed the number of patients to a decreasing trend, and being between 1015 and 1020 hPa triggered a relative increase in the number of patients. The relevance was further emphasized by the SHAP value in the 3-day data set.

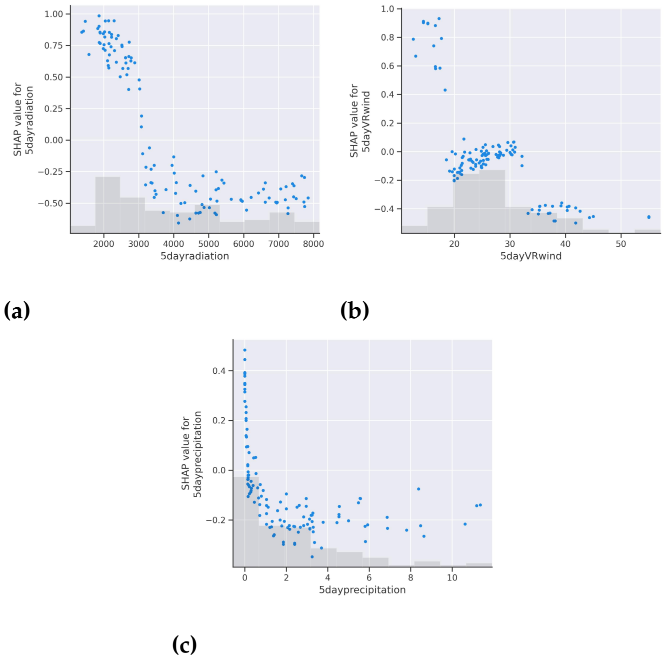


(a) (b)

(c)

Figure 5 Dependence plots for variables **a)** shortwave radiation (3-day), **b)** mean SD wind (3-day), **c)** mean pressure (3-day)

Finally, the impact of precipitation on our model's predictions over the 5-day period is clarified in the Precipitation dependency plot in Figure 6d. Looking at the graph, it is seen that the 5-day average rainfall amount being different from zero tends to reduce the number of patients relatively.



(a) (b)

(c)

Figure 6 Dependence plots for variables **a)** shortwave radiation (5-day), **b)** VR wind (5-day), **c)** precipitation (5-day)

DISCUSSION

This research aims to bridge the gap in understanding the complicated association between climatic factors and the prevalence of

Table 4 Actual Values and SHAP Values table for 25.08.2021 and 23.12.2022

	25.08.2021 (True Negative)				23.12.2022 (True Positive)			
Feature	3-Day A.V.	3-Day SHAP	5-Day A.V.	5-Day SHAP	3-Day A.V.	3-Day SHAP	5-Day A.V.	5-Day SHAP
meantemp	23.22	-0.16	22.48	-0.00	6.88	0.00	6.19	0.09
meanhumidity	77.03	0.01	76.58	0.05	63.53	-0.08	73.78	0.02
meanpressure	1012.54	-0.04	1013.31	-0.01	1019.02	0.18	1024.24	-0.13
meanwind	9.70	-0.10	10.96	-0.01	6.84	0.27	11.35	0.02
sunshine	650.60	0.08	660.33	0.12	554.56	0.08	361.82	-0.00
radiation	6635.25	-0.46	6618.57	-0.34	2561.42	0.70	1987.19	0.84
precipitation	0.00	0.11	0.00	0.35	0.00	0.13	2.52	-0.15
snowfall	0.00	0.01	0.00	0.02	0.00	0.03	0.22	0.02
cloudcover	44.85	-0.04	41.04	0.11	10.56	0.01	44.31	0.02
minmaxSDtemp	5.25	-0.01	5.60	-0.05	5.99	-0.02	5.15	0.00
meanSDtemp	0.72	0.18	1.11	-0.06	1.95	0.05	1.86	-0.02
VRtemp	11.83	-0.15	14.99	0.00	14.97	0.03	14.97	0.02
minmaxSDhumidity	1.03	-0.00	27.89	0.05	7.18	0.10	22.23	0.02
meanSDhumidity	1.16	0.18	1.76	0.20	9.20	-0.01	14.82	-0.05
VRhumidity	2.61	0.08	66.00	0.01	18.39	0.05	58.00	-0.01
minmaxSDpressure	1.61	-0.02	2.21	0.02	4.32	-0.04	8.03	0.02
meanSDpressure	1.62	-0.00	1.57	-0.01	4.11	-0.01	7.16	0.07
VRpressure	3.89	-0.03	7.10	-0.04	10.20	-0.00	24.20	-0.09
minmaxSDwind	2.14	-0.02	13.49	0.13	1.11	0.09	9.02	0.11
meanSDwind	1.91	-0.06	2.14	-0.18	1.56	-0.01	5.97	0.14
VRwind	4.52	-0.03	30.65	0.07	3.02	0.03	25.77	0.01

URTIs in the Pamukova District, utilizing modern machine learning methods. Several critical discoveries arose from this analysis, having substantial implications for both the scientific community and public health measures.

The results of this study underscore the significant improvements achieved by extending the temporal frame of environmental data from 3 to 5 days when predicting URTIs using the Catboost model. The extended data frame led to a substantial increase in predictive performance, with the F1-score improving from 0.519 to 0.6154, accuracy from 67.77% to 75.21%, precision from 0.6 to 0.75, recall from 0.4565 to 0.5217, and the AUC value from 0.6861 to 0.7768. These results highlight the advantage of considering a broader temporal context, which captures more comprehensive environmental patterns that influence URTI occurrences.

Additionally, the integration of SHAP values significantly enhances the interpretability of the Catboost model, providing clear insights into feature importance. Key findings revealed that shortwave radiation was the most influential predictor, with its SHAP value increasing from 0.50 in the 3-day model to 0.54 in the 5-day model. Other important factors included variations in wind speed, precipitation levels, and atmospheric pressure. The improved interpretability and predictive accuracy underscore the potential of this method for developing effective, data-driven public health interventions. This approach not only improves the robustness of predictions but also enables targeted strategies tailored to specific environmental conditions, ultimately contributing to better health outcomes in regions like the Pamukova District.

The continuous relevance of shortwave radiation across both 3-day and 5-day datasets highlights its severe influence on URTI incidences. This coincides with some recent study, which has highlighted the possible immunomodulatory effects of sun radiation, possibly altering virus transmission and susceptibility. The negative link identified between high shortwave radiation levels and URTIs shows that greater sunshine exposure, and maybe its related vitamin D synthesis, could give some protection against URTIs. This underscores the necessity of addressing regional and seasonal changes when creating public health interventions.

The association between wind dynamics and URTIs is complicated. While wind may scatter respiratory droplets, possibly lowering transmission, it can also worsen respiratory symptoms and increase exposure to allergens. Our results highlighting the impact of wind speed changes, particularly over extended periods, may open the way for more nuanced study addressing the interactions between wind patterns, allergen dispersion, and URTI occurrences. Specifically, the 5-day model showed that variations in wind speed, with a SHAP value of 0.21, were significant predictors of URTIs. Additionally, the dependency plot analysis revealed that wind speed ranges below 20 km/h increased the number of cases, whereas ranges above 30 km/h decreased the number of cases, further emphasizing the importance of wind dynamics in predicting URTI occurrences.

Atmospheric pressure, another crucial element in our model, has been little examined in connection to respiratory diseases. Our results reveal prospective pathways for investigation into how atmospheric pressure could alter air quality, respiratory function, and therefore, susceptibility to infections. The observed association between pressure levels and URTIs could also indicate indirect consequences, such as behavioral changes in reaction to climatic circumstances. For instance, the 3-day average atmospheric pressure below 1015 hPa was associated with a decrease in URTI cases, whereas pressure between 1015 and 1020 hPa triggered a relative increase, as shown in the SHAP dependency plot. The SHAP values for mean atmospheric pressure were -0.04 for the 3-day model and 0.18 for the 5-day model, indicating its significant but complex role in influencing URTI rates.

Precipitation appeared as a key variable across the 5-day period. This is in accordance with prior research, which has generally correlated damp circumstances with higher virus survival and transmission, particularly in cold settings. Our findings indicate that the presence of precipitation in the 5-day dataset, with a SHAP value of 0.18, was a significant factor in predicting URTIs. The dependency plot demonstrated that non-zero precipitation values tended to reduce the number of URTI cases. This lays the ground for further extensive investigations that might shed light on how various precipitation types affect URTIs, emphasizing the importance of understanding the specific climatic conditions that promote or hinder the transmission of respiratory infections.

The strength of this work comes in its holistic approach, incorporating a variety of climatic factors and applying powerful machine learning methods to decode their cumulative influence on URTIs. CatBoost, with its capacity to handle categorical characteristics without preprocessing, emerged as a useful tool, offering insights with great accuracy. However, some limits must be noted. While the research caught a broad variety of environmental factors, additional possible confounders such as indoor air quality, personal habits, and vaccination rates were not evaluated. The reliance on data from a specific area further restricts the generalizability of the results. Furthermore, although SHAP values give a comprehensive comprehension of feature relevance, they do not always suggest causation.

Future research should seek to replicate and build upon these results in other geographical locations, integrating additional possible confounders and examining causal processes. Longitudinal research covering longer time periods might give insights into the long-term impact of climatic factors on URTIs. There's also a need for in-depth investigation of the molecular and physiological pathways via which these environmental influences impact respiratory health.

In conclusion, our analysis underlines the complicated interaction of climatic circumstances in producing URTI patterns. As the global community grapples with respiratory illnesses, information from such research are vital. They not only increase our knowledge but also give actionable information for public health authorities, allowing the creation of tailored treatments that account the particular environmental and climatic context of a place.

CONCLUSION AND FUTURE WORK

This research provides a complete examination of the interaction between several climatic conditions and the prevalence of URTIs in the Pamukova District of Sakarya Province, Turkey. Through the application of the CatBoost machine learning model, we discovered that certain environmental characteristics, notably shortwave radiation, precipitation, and wind-related variables, continuously emerge as key drivers in forecasting URTI occurrences.

Notably, shortwave radiation's consistent effect throughout varied seasons highlights its importance as a critical element. This underlines the delicate balance between the environment and human health, indicating that although certain elements have a consistent influence, others fluctuate dependent on the time range studied. Additionally, the model's precise detection of detailed weather patterns, particularly when employing a 5-day observing period, shows that protracted environmental circumstances could play a more essential role in determining URTIs than previously believed.

The use of SHAP values greatly improved our comprehension, allowing for both global and local interpretations of the model's predictions. These numbers not only strengthened the findings reached from the model's raw outputs but also offered insight on the complicated connections between numerous weather factors.

The results of this study uncover various exciting paths for both additional research and practical applications. The usage of SHAP values in our study has shown to be a rewarding venture, enabling a granular comprehension of the multiple meteorological aspects that contribute to the occurrence of URTIs. While our work has found a plethora of discoveries, it's obvious that the full potential of SHAP values and other machine learning interpretability tools needs to be utilized, especially in the context of URTIs and weather conditions. Some of the planned futureworks are:

- Enhancing Patient-Centric Reporting Using SHAP Values: Presently, SHAP values enable us to discover which climatic parameters in our dataset play a crucial role in forecasting URTI occurrences. An extension of this method would be to describe the percentage contribution of each variable, supplying a more explicit and accurate grasp of the risk factors. This revised technique might provide healthcare practitioners with a personalized strategy to predict URTI spikes, basing their preventative actions around the most relevant weather circumstances.
- Integration of Additional Data Sources: Beyond meteorological considerations, incorporating statistics relating to air pollution, pollen counts, and other environmental variables might increase the forecasting powers of the model. This would offer a more thorough view of the environmental triggers of URTIs.
- Expansion of Geographical Scope: Delving into different geographical locations, both within Turkey and worldwide, could show whether the linkages detected in the Pamukova District are generally applicable or feature regional quirks.
- Temporal Analysis using SHAP: By applying SHAP values in a temporal context, it would be feasible to discover patterns linked to the seasonality of URTIs and how various meteorological conditions interact through time to impact URTI prevalence.
- Real-time URTI Predictive Systems: Capitalizing on the insights obtained, there's a chance to create real-time prediction systems that can estimate URTI prevalence based on present and impending climatic conditions, thereby helping healthcare institutions to plan appropriately.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

Participants' information was gathered from January 2020 to December 2022 at the Pamukova Family Health Centre, where T.B.D. is affiliated as MD, after receiving ethical approval from the Sakarya University of Applied Sciences Ethical Committee (E-26428519-044-77759). The Pamukova Family Health Centre provided the dataset for URTI. T.B.D., the study's corresponding author, will provide the data that back up its conclusions upon request. Please be aware, though, that the data cannot be made public because it contains details that would jeopardize the research participants' right to privacy even if it was anonymized.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used artificial intelligence tools in order to improve the readability and language quality of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

LITERATURE CITED

Bentéjac, C., A. Csörgő, and G. Martínez-Muñoz, 2021 A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review **54**: 1937–1967.

Chelgani, S. C., H. Nasiri, A. Tohry, and H. Heidari, 2023 Modeling industrial hydrocyclone operational variables by shap-catboost-a "conscious lab" approach. Powder Technology **420**: 118416.

Chen, X., L. Huang, Q. Li, M. Wu, L. Lin, *et al.*, 2021 Exposure to environmental tobacco smoke during pregnancy and infancy increased the risk of upper respiratory tract infections in infants: a birth cohort study in wuhan, china. Indoor air **31**: 673–681.

Derbyshire, E. J. and P. C. Calder, 2021 Respiratory tract infections and antibiotic resistance: a protective role for vitamin d? Frontiers in Nutrition **8**: 652469.

D'Amato, M., A. Molino, G. Calabrese, L. Cecchi, I. Annesi-Maesano, *et al.*, 2018 The impact of cold on the respiratory tract and its consequences to respiratory health. Clinical and translational allergy **8**: 1–8.

Eccles, R. and J. Wilkinson, 2015 Exposure to cold and acute upper respiratory tract infection. Rhinology **53**: 99–106.

Falagas, M. E., G. Theocharis, A. Spanos, L. A. Vlara, E. A. Issaris, *et al.*, 2008 Effect of meteorological variables on the incidence of respiratory tract infections. Respiratory medicine **102**: 733–737.

Jhuo, S.-L., M.-T. Hsieh, T.-C. Weng, M.-J. Chen, C.-M. Yang, *et al.*, 2019 Trend prediction of influenza and the associated pneumonia in taiwan using machine learning. In *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1–2, IEEE.

Kannangara, K. P. M., W. Zhou, Z. Ding, and Z. Hong, 2022 Investigation of feature contribution to shield tunneling-induced settlement using shapley additive explanations method. Journal of Rock Mechanics and Geotechnical Engineering **14**: 1052–1063.

Kavzoglu, T. and A. Teke, 2022 Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (xgboost) and natural gradient boosting (ngboost). Arabian Journal for Science and Engineering **47**: 7367–7385.

Kern, C., K. Kortüm, M. Müller, F. Raabe, W. J. Mayer, *et al.*, 2016 Correlation between weather and incidence of selected ophthalmological diagnoses: a database analysis. Clinical ophthalmology pp. 1587–1592.

Kolawole, O. M. and O. O. Idris, 2020 Erythromycin resistance in bacterial isolates from patients with respiratory tract infections in ikere-ekiti, nigeria. Annals of Science and Technology **5**: 49–57.

Lim, J. T., K. B. Tan, J. Abisheganaden, and B. L. Dickens, 2023 Forecasting upper respiratory tract infection burden using high-dimensional time series data and forecast combinations. PLOS Computational Biology **19**: e1010892.

Mäkinen, T. M., R. Juvonen, J. Jokelainen, T. H. Harju, A. Peitso, *et al.*, 2009 Cold temperature and low humidity are associated with increased occurrence of respiratory tract infections. Respiratory medicine **103**: 456–462.

Mansour, M., T. B. Donmez, M. KUTLU, and C. Freeman, 2023 Respiratory diseases prediction from a novel chaotic system. Chaos Theory and Applications **5**: 20–26.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, 2018 Catboost: unbiased boosting with categorical features. Advances in neural information processing systems **31**.

Santhanam, N., H. E. Kim, D. Ruegamer, A. Bender, S. Muthers, *et al.*, 2024 Machine learning-based forecasting of daily acute ischemic stroke admissions using weather data. medRxiv pp. 2024–07.

Stolz, D., E. Papakonstantinou, L. Grize, D. Schilter, W. Strobel, *et al.*, 2019 Time-course of upper respiratory tract viral infection and copd exacerbation. European Respiratory Journal **54**.

Tu, J., K. Inthavong, G. Ahmadi, J. Tu, K. Inthavong, *et al.*, 2013 The human respiratory system. Computational fluid and particle dynamics in the human respiratory system pp. 19–44.

Yang, J., X. Xu, X. Ma, Z. Wang, Q. You, *et al.*, 2023 Application of machine learning to predict hospital visits for respiratory diseases using meteorological and air pollution factors in linyi, china. Environmental Science and Pollution Research **30**: 88431–88443.

Zeru, T., H. Berihu, G. Buruh, and H. Gebrehiwot, 2020 Magnitude and factors associated with upper respiratory tract infection among under-five children in public health institutions of aksum town, tigray, northern ethiopia: an institutional based cross-sectional study. Pan African Medical Journal **36**.

Zou, Z., C. Cheng, and S. Shen, 2021 The complex nonlinear coupling causal patterns between pm2. 5 and meteorological factors in tibetan plateau: A case study in xining. IEEE Access **9**: 150373–150382.

How to cite this article: Dönmez, T. B.,Kutlu, M. and Freeman, C. Predicting Upper Respiratory Tract Infections: The Role of Weather Data and Explainable AI. *Artificial Intelligence in Applied Sciences*, 2(1),37-48, 2026.