ORIGINAL ARTICLE

Development of an Atopic Dermatitis Incidence Rate Prediction Model for South Korea Using Air Pollutants Big Data: Comparisons Between Regression and Artificial Neural Network

Byeonggeuk Lim[1] · Poong-Mo Park[2,3] · Da-Mee Eun[4] · Dong-Woo Kim[2] · Cheonwoong Kang[2] · Ki-Joon Jeon[2,3] · SeJoon Park[1] · Jong-Sang Youn[4]

Abstract

We have developed models to predict the incidence of atopic dermatitis using regression analysis and artificial neural networks (ANN). Initially, the prediction models were created using various inputs, including air pollutants (SO_2, CO, O_3, NO_2, and PM_{10}), meteorological factors (temperature, humidity, wind speed, and precipitation), population rates, and clinical data from South Korea, referred to as the average model. Subsequently, we developed models that use sex and age as variables instead of population rates, named the sex and age model. Both sets of models were designed to forecast incidence rates on a nationwide scale (NW), as well as for 16 administrative districts (AD) in South Korea, which includes seven metropolitan areas and nine provinces. We found that SO_2 significantly affected the incidence rate, and the inclusion of regional variables in the AD models helped account for regional variations in incidence rates. The average models generally provided accurate predictions of incidence rates, with SO_2 chosen as the key independent variable in the regression models for the five air pollutants studied. The R^2 values for the average models using regression are 0.70 for the NW model and 0.89 for the AD model. Among the ANN-based models, the R^2 values are 0.84 for the NW model and 0.90 for the AD model, this indicated a slightly higher predictive accuracy. For the sex and age models, we differentiated between children under 10 years of age and those older. In these models, ANN demonstrated greater accuracy than regression, with R^2 values of 0.95, 0.92, 0.96, and 0.92 for the sex and age NW model under 10 years old, sex and age AD model under 10 years old, sex and age NW model over 10 years old, and sex and age AD model over 10 years old, respectively.

Keywords Atopic dermatitis · Air pollutants · Environmental big data · ANN · Regression · Prediction model

Byeonggeuk Lim and Poong-Mo Park have contributed equally.

✉ SeJoon Park
  sonmupsj@kangwon.ac.kr

✉ Jong-Sang Youn
  jsyoun@catholic.ac.kr

[1] Division of Energy Resources Engineering and Industrial Engineering, Kangwon National University, Chuncheon 24341, South Korea

[2] Department of Environmental Engineering, Inha University, Incheon 22212, South Korea

[3] Particle Pollution Research and Management Center, Incheon 22212, South Korea

[4] Department of Energy and Environmental Engineering, The Catholic University of Korea, Bucheon 14662, South Korea

Introduction

Atopic dermatitis, a chronic inflammatory skin condition that significantly disrupts sleep, induces emotional and physical stress, and hampers social interaction. This widespread disease has seen a two to threefold increase in prevalence over the last 30 years [1, 2]. Commonly affecting infants and young children, atopic dermatitis typically manifests within the first 6 months of life, severely diminishing the quality of life of both patients and their families, primarily through sleep disturbances. Spergel and Paller [3] highlight the substantial financial impact on families and society. Despite its prevalence across all age groups and the resulting socioeconomic strain, comprehensive research encompassing all ages currently remains insufficient [4]. Often emerging in childhood, atopic dermatitis is regarded as an initial indicator of the 'atopic march,' a progression towards

respiratory allergies such as asthma and allergic rhinitis with aging. This progression typically follows a specific order: atopic dermatitis, asthma, and allergic rhinitis, and is linked to an approximately 20% increased long-term risk of atrial fibrillation [5–7]. Environmental factors heavily influence diseases, such as atopic dermatitis, asthma, and allergic rhinitis. In Korea, the distribution of atopic dermatitis cases shows regional variations, with higher incidences in areas with numerous industrial complexes and rapid development [8, 9].

Atopic dermatitis is influenced by a combination of genetic susceptibilities, environmental elements, and also immunological irregularities. The interplay between an individual's genetic makeup and environmental conditions significantly contributes to disease occurrence. Kim et al. [4] suggests that the escalating occurrence of atopic conditions is more due to heightened exposure to diverse environmental factors than to alterations in human genetic makeup. Symptoms are exacerbated by factors such as skin infections, climatic conditions, and air pollution. The effective management of these environmental factors is essential for the treatment of atopic dermatitis [7]. Notably, the correlation between air pollutants and atopic dermatitis has garnered increasing attention, as highlighted in research by Guo et al. [10].

Air pollutants have a profound impact on human health and increase mortality rates [11]. The Ministry of Environment has identified key air pollutants, including sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), fine particulate matter (PM_{10}), and ultrafine particulate matter ($PM_{2.5}$). These pollutants, along with meteorological factors, play a crucial role in the incidence of atopic dermatitis, as evidenced by Vocks et al. [12], Kim [13], and Silverberg et al. [14]. The detrimental effects of air pollutants on skin include inflammation and accelerated aging [15, 16]. Furthermore, research has indicated a significant association between temperature, humidity, air pollution, and atopic dermatitis, highlighting the interactions between meteorological factors and air pollutants [10]. Given these findings, it is evident that research on atopic dermatitis should incorporate both environmental and climatic considerations.

Recent advancements have enabled researchers to leverage artificial intelligence (AI) to analyze the influence of air pollutants and meteorological conditions on atopic dermatitis and develop prediction models for the disease. Historically, studies have predominantly employed statistical methods, with limited exploration using AI. There is also a notable gap in atopic dermatitis prediction models, especially when compared to other environmental disease models. Most previous studies have focused on specific age groups, creating models that categorize atopic dermatitis symptoms into to 3–5 levels among populations ranging from 60 to 177 individuals [17–19]. For instance, Patella et al. [17] applied artificial neural networks (ANN) to test the severity of atopic dermatitis in response to air pollutants and weather changes, achieving a high prediction accuracy of 75.46% in a study involving 60 patients over 5 years old (average age 23.5 years). Research on predicting the incidence of atopic dermatitis using AI techniques included a survey sample of 1439 pregnant women [20]. Previous studies on the link between air pollutants and atopic dermatitis have used sample sizes ranging from 21 to 787,646 individuals, with one notable study involving 246,844 participants [21, 22]. Park et al. [8] conducted a domestic study examining the correlation between atopic dermatitis cases and air pollutants across eight regions encompassing a population of 23,288,000. These studies typically segmented atopic dermatitis symptoms into stages for model development or patient number estimation. However, these previous studies were often limited by its focus on specific age groups and the inconsistent sample sizes, which affect the generalizability of findings. Despite recent advancements, the use of AI in atopic dermatitis prediction is still underexplored, and existing models often oversimplify the disease by categorizing symptoms into a few levels. Additionally, high prediction accuracies reported in some studies may not be consistently reproducible across different populations or conditions, necessitating further validation.

This study represents a significant expansion of the survey population size, encompassing 51,778,544 individuals across all age groups in Korea. Its primary aim was to investigate the incidence of atopic dermatitis, aiming for a more precise identification of regional variations and risk factors associated with the disease. This study involved a detailed analysis of the air pollutants and meteorological conditions that contribute to the incidence of atopic dermatitis. Two types of atopic dermatitis incidence prediction models were developed: an average incidence prediction model and one segmented by sex and age groups. The methodology encompasses both statistical and artificial intelligence techniques. Statistical approaches include regression analysis, whereas from the AI perspective, ANN are employed. To capture regional variations effectively, this study developed prediction models for atopic dermatitis incidence both nationwide (NW) and across various administrative districts (AD). One of the key objectives was to compare the efficacy of the regression and ANN models in predicting atopic dermatitis. Through this approach, the study seeks to discover and present improved strategies for forecasting the occurrence of atopic dermatitis, deepen insights into the environmental determinants of the condition, and support more precise and efficacious public health interventions.

Materials and Methods

This study has focused on South Korea and a prediction model for the incidence rate of atopic dermatitis was developed by considering both NW and AD levels. South Korea is divided into 17 administrative districts, including eight metropolitan cities and nine states, as shown in Fig. 1. However, it's important to note that Sejong, one of the AD, is not included in the development of the AD model. This exclusion was due to the lack of available data on air pollutants and meteorological factors specific to Sejong. The absence of such data makes it challenging to accurately model the incidence of atopic dermatitis in the district. By focusing on the remaining districts, this study aimed to provide a comprehensive and accurate prediction model that reflects regional variations in the incidence rates of atopic dermatitis across South Korea. This approach will help understand the distribution of diseases and potential environmental risk factors at both the national and local levels.

Atopic Dermatitis Incidence Rate

This study utilized outpatient data for atopic dermatitis from the National Health Insurance Service of South Korea, accessible through their website (https://nhiss.nhis.or.kr/, accessed on January 12, 2023). The dataset spans

Fig. 1 Nationwide monthly atopic dermatitis incidence rate

from January 1, 2013, to December 31, 2017. To minimize errors stemming from the lag between air pollutant exposure and subsequent hospital outpatient visits, this study used monthly outpatient data. The incidence rate of atopic dermatitis was calculated using the following formula: the number of atopic dermatitis outpatients divided by the population, as indicated in Eq. (1). This calculation was performed for both the NW context and each AD, with further segmentation by sex and age group. The age groups were categorized in 5-year increments, allowing for a detailed analysis of the incidence rates across different demographics. By adopting this methodology, this study has aimed to provide a nuanced understanding of atopic dermatitis incidence rates across various regions of South Korea, considering age and sex differences. This approach is crucial for identifying specific patterns and trends in disease occurrence, which can inform targeted healthcare strategies and policy interventions.

$$\text{Incidencerate} = \frac{\text{The number of outpatients}}{\text{Population}}. \tag{1}$$

Population data crucial for this study were sourced from the Ministry of the Interior and Safety of South Korea, available on their website (https://jumin.mois.go.kr/) and accessed on January 6, 2023. These data are essential for calculating the incidence rates of atopic dermatitis by considering both the total population and number of outpatients. Figure 2a shows the nationwide monthly incidence of atopic dermatitis. An interesting observation from these data was the fluctuation in incidence rates based on seasonal changes. Notably, the incidence rate of atopic dermatitis tended to be higher during the summer months than during the winter. This seasonal variation in the incidence rate is a critical factor to consider, as it suggests potential links between environmental conditions such as temperature and humidity and the prevalence of atopic dermatitis. By integrating these insights with the previously mentioned outpatient data and environmental factors, this study aimed to provide a comprehensive analysis of the patterns and triggers of atopic dermatitis in South Korea. Understanding these seasonal trends is crucial for developing effective prevention and treatment strategies as well as for guiding public health policies and awareness campaigns. Figure 2b shows the incidence rate of atopic dermatitis, broken down by sex and age. There was a marked change in the incidence rates across the different age groups. The variation in incidence rates between sexes was relatively minor when compared with the more significant differences observed across various age levels. Figure 2c illustrates the incidence rates of atopic dermatitis across the 17 administrative districts. Each district has a distinct incidence rate. Notably, there was no substantial disparity in incidence rates between metropolitan cities and states. The population ratio of each administrative district
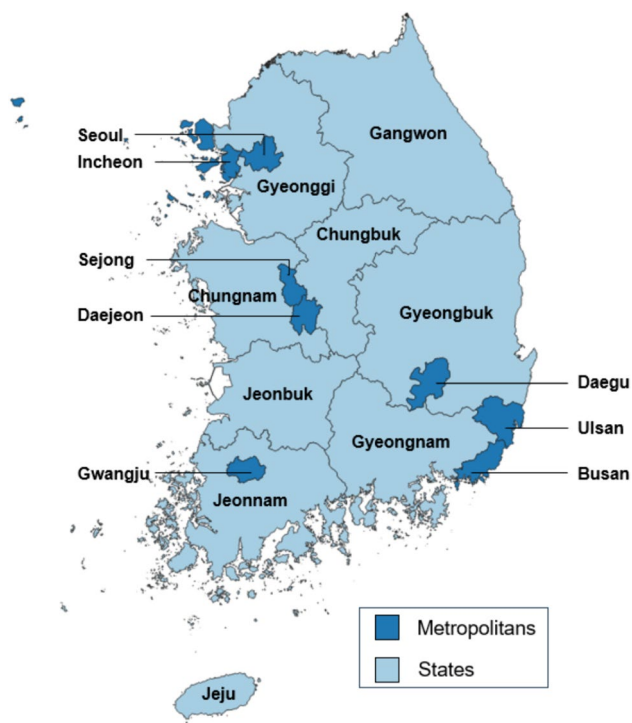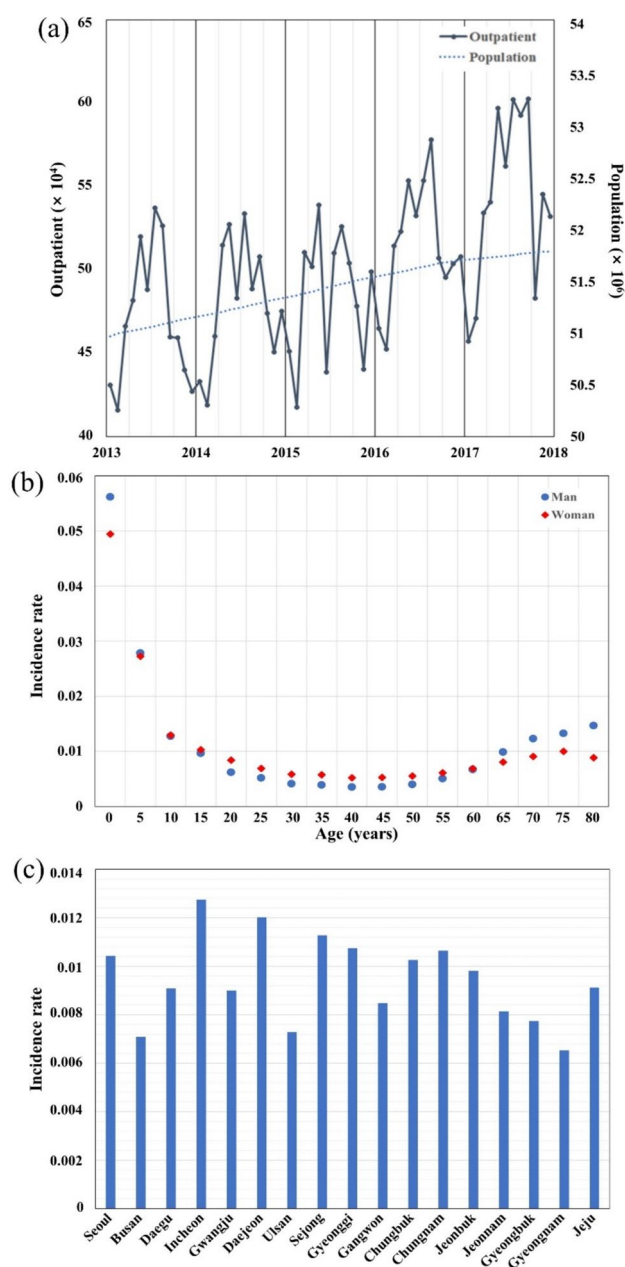
Fig. 2 **a** Nationwide monthly atopic dermatitis outpatients and population, **b** atopic dermatitis incidence rate by sex and age level, and **c** atopic dermatitis incidence rate by administrative district

may have influenced these rates. Other factors not covered in this study may also play a role. These aspects are further explored in the Results and Discussion section.

Air Pollutants and Meteorological Data

In this research, we utilized monthly data on air pollutants (SO_2, CO, O_3, NO_2, and PM_{10}) and meteorological variables (temperature, humidity, wind speed, and precipitation). Air pollutant data were sourced from Air Korea

(accessible at https://www.airkorea.or.kr/) as of November 19, 2022, whereas meteorological data were obtained from the Open Met Data Portal (https://data.kma.go.kr/, accessed on November 19, 2022). We calculated the monthly averages for each type of data across the nation and its administrative districts. However, Sejong was not included in the regional analysis because of the availability of air pollutant data starting only from January 2016 and meteorological data from May 2019, while the atopic dermatitis data were available from January 1, 2013, to December 31, 2017. More detailed information on the monthly averages of air pollutant concentrations and meteorological factors is provided in the Supplementary Information (S1). Air pollutant and meteorological data were standardized to develop prediction models using regression and ANN, as described by Eqs. (2) and (3), respectively. Equation (2) employs z-score normalization (adjusted by adding 4, as per [23], facilitating the comparison of the influence of independent variables in the regression models. For ANN models, input variables are typically normalized to fall between 0 and 1, often using minimum–max normalization to meet this requirement.

The normalization using z-score is obtained by Eq. (2)

$$\text{Normalized data} = \frac{\text{Original data} - \mu}{\sigma} + 4, \qquad (2)$$

where μ is mean and σ is standard deviation and the min–max normalization is calculated by Eq. (3)

$$\text{Normalized data} = \frac{\text{Original data} - \min}{\max - \min}, \qquad (3)$$

where min is the minimum value of all data and max is the maximum value of all data.

Regression and Artificial Neural Network

Regression models are commonly utilized to analyze time-series data, whereas artificial neural networks (ANN) are among the most robust techniques in artificial intelligence for developing predictive models. A regression model was used to statistically delineate the relationship between the response variable and its independent predictors. In contrast, although ANNs often yield more accurate predictions than the regression models, they do not explicitly clarify the relationship between the response and the independent variables. This study conducted a comparative analysis of predictive models using regression and an ANN. Figure 3 illustrates the structure of the ANN model, which incorporates two hidden layers. This design choice was based on the findings of Thomas et al. [24], who found that adding more than three hidden layers did not significantly enhance ANN performance. The number of nodes in each hidden layer and the threshold were carefully optimized to develop

Fig. 3 Schematic of the ANN model
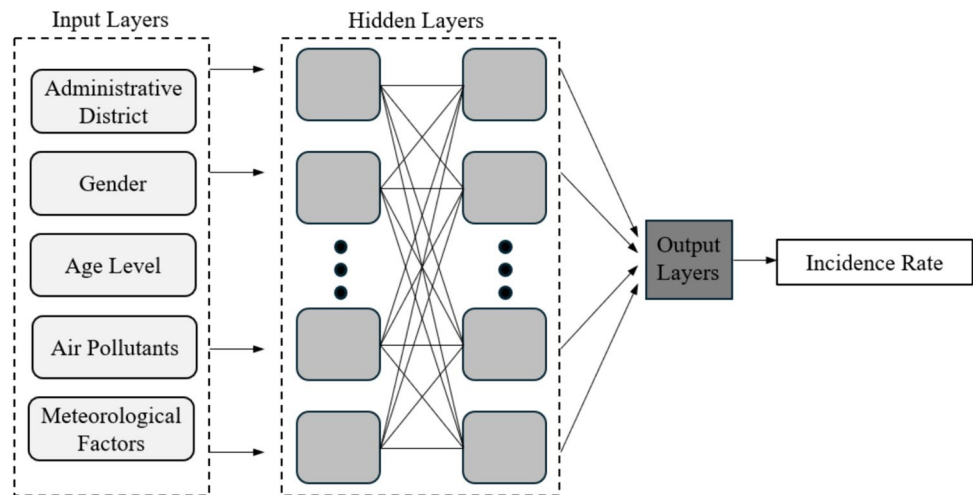
the ANN-based predictive models. In ANN, the threshold plays a crucial role in determining the extent of learning. Lowering the threshold typically enhances the precision of the ANN predictions. However, this often leads to overfitting, a scenario in which the model excessively conforms to the specific data at hand and increases the computation time required for model training.

Results and Discussion

In this study, two types of prediction models for atopic dermatitis incidence rates have been developed: "average models" and "sex and age models." As detailed in Table 1, 'y' represents the response variable, while the remaining factors are independent variables. The "average NW model" forecasts the overall incidence rate across all sexes and ages, taking into account air pollutant concentrations, meteorological conditions, and population proportions. For this model, the administrative districts mentioned in Table 1 were incorporated as an additional independent variable. Similarly, the "sex and age NW model" estimates incidence rates for each sex and age group, based on air pollutant concentrations, meteorological factors, and the demographic variables of sex and age. Here too, the administrative district from Table 1 is included as an extra independent variable in the "sex and age AD model." Each predictive model was developed using regression and artificial neural networks (ANN). Furthermore, all categorical variables listed in Table 1 are binary (0 or 1).

In this study, the data were split into in-sample (80%) and out-of-sample (20%) to develop regression models. The in-samples were employed to construct the model, whereas the out-of-samples were reserved for model validation and assessment. For the artificial neural network (ANN) models, the in-samples were further divided into training (60%) and

Table 1 Variables for atopic dermatitis incidence rate prediction models

Variable type	Data	Categories	Variables
Continuous variables	Incidence rate	Atopic dermatitis incidence rate	y
	Air pollutants concentrations	SO_2, CO, O_3, NO_2, and PM_{10}	$x1, x2, x3, x4, x5$
	Meteorological factors	Temperature, humidity, wind speed, and precipitation	$z1, z2, z3, z4$
	Population rate	Man and age levels (0–19, 20–39, 40–59, 60–79, more than 80)	M MW0, MW20, MW40, MW60, MW80
Categorical variables	Administrative district	Seoul, Busan, Daegu, Incheon, Gwangju, Daejeon, Ulsan, Gyunggi, Gangwon, Choongbuk, Choongnam, Jeonbuk, Jeonnam, Kyungbuk, Kyungnam, and Jeju	Area
	Sex	Man and woman	Sex
	Age level	Age levels (0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, more than 80)	Age

validation (20%) data. The out-of-sample data used in the regression models served as the test data (20%). Training data were utilized for the initial development of the model, and validation data were used to select the best model. Finally, the test data were used for model validation and evaluation, paralleling the use of the in-sample regression models.

Average Model

The population rate variables used to develop the average model are listed in Table 2. M was selected to account for the differences between men and women. Correlations between the population and atopic dermatitis incidence rates are shown in Table 2. MW20 and MW60 were selected based on their correlation with the incidence rate. The correlations for MW0, MW40, and MW80 are nearly zero. The correlations between MW20 and MW60 are positive and negative, respectively.

Average Model by Regression

To determine the impact of different air pollutants on the incidence of atopic dermatitis, separate regression models were created for each pollutant. Each model included a specific air pollutant (x), population rate, an interaction term involving the population rate, and meteorological factors, as outlined in (4). In models focusing on administrative districts (AD), an additional area variable was incorporated, as indicated in Eq. (5).

$$y \sim (MW20 + MW60)M + x + z1 + z2 + z3 + z4 \quad (4)$$

$$y \sim (MW20 + MW60)M + x + z1 + z2 + z3 + z4 + Area. \quad (5)$$

Table 3 present the summaries of the regression results at the nationwide level and for administrative districts, respectively. For all air pollutants except O_3, the coefficients were positive, and the R^2 values of these models were relatively similar. Notably, SO_2 showed the highest standardized coefficient (coefficient divided by standard deviation) in both the nationwide (NW) model and the administrative district (AD) model. This indicates that variations in SO_2 levels have

the most significant impact on the incidence rate of atopic dermatitis. Consequently, $\times 1$ (SO_2) was chosen as the representative independent variable among air pollutants for the regression models.

Table 4 details the coefficients and R^2 values for the NW model, the AD model excluding the area variable, and the AD model including the area variable. The R^2 values for the NW models are 0.79, 0.74, and 0.70, respectively, while the AD models incorporating the area variable each have an R^2 value of 0.89. AD models without the area variable struggle to predict the incidence rate accurately, with R^2 values

Table 2 Correlations between the population rate and atopic dermatitis incidence rate

Population rate	Correlation
M	− 0.088
MW0	0.036
MW20	0.336
MW40	− 0.060
MW60	− 0.239
MW80	− 0.089

Table 3 Summary of nationwide (NW) and administrative district (AD) regression results

x	Coefficients	SD	Coefficients/SD	R^2	Adjusted R^2
NW					
x1 (SO_2)	0.278	0.001	285.997	0.787	0.743
x2 (CO_2)	0.643	0.103	6.229	0.808	0.769
x3 (O_3)	-0.299	0.009	-34.720	0.782	0.737
x4 (NO_2)	0.465	0.005	101.299	0.789	0.746
x5 (PM_{10})	0.123	11.296	0.011	0.775	0.729
AD					
x1 (SO_2)	0.140	0.002	80.911	0.892	0.889
x2 (CO_2)	0.134	0.134	0.999	0.891	0.888
x3 (O_3)	0.028	0.010	2.920	0.887	0.883
x4 (NO_2)	0.060	0.008	7.970	0.887	0.884
x5 (PM_{10})	0.100	11.995	0.008	0.891	0.888

SD standard deviation

Table 4 Coefficients and R^2 values of the average models by regression

	NW model	AD model without area variable	AD model with area variable
Intercept	12.496	14.080	− 11.262
MW20	0.456	− 1.043	1.5059
MW60	− 1.007	− 1.515	1.312
M	− 2.666	− 4.612	2.820
MW20:M	0.062	0.603	− 0.412
MW60:M	− 0.120	0.589	− 0.129
x1	0.256	− 0.191	0.131
z1	1.056	0.104	0.370
z2	− 0.340	0.174	− 0.092
z3	0.163	− 0.003	0.094
z4	0.141	− 0.149	0.034
R^2			
In-sample	0.794	0.422	0.891
Adjusted	0.738	0.414	0.888
Out-of-sample	0.698	0.292	0.887

of 0.42, 0.41, and 0.29, indicating the influence of regional factors beyond air pollutants, meteorological conditions, and population rates in South Korea. Furthermore, Table 4 indicates that the presence of SO_2 (×1) in both the NW model and the AD model with the area variable is associated with an increase in the atopic dermatitis incidence rate. The data also suggest that increases in temperature, wind speed, and precipitation correlate with a rise in the incidence rate, whereas an increase in humidity correlates with a decrease. For coefficients related to the area variable in the AD model, refer to supplementary information (S2).

Average Model by ANN

Factorial designs were employed for model development, as illustrated in Table 5. The performances of the ANN models were evaluated based on the number of nodes in the hidden layer and the threshold value. The research suggests that ANNs perform well when the number of nodes equals the number of independent variables or the logarithm of base 2 of the number of independent variables [25, 26]. For the NW and AD models, the numbers of independent variables were 15 (comprising six population rates, five air pollutants, and four meteorological factors) and 31 (including six population

rates, five air pollutants, four meteorological factors, and 16 area variables), respectively. The number of nodes is set by halving these values. The threshold value is crucial, because it dictates the error rate at which learning ceases in the ANN. Lowering the threshold extends the learning duration and increases the risk of overfitting, whereas raising the threshold can compromise the model accuracy. Threshold ranges were set based on the balance between learning time and accuracy, as determined through preliminary experiments. The ANN models were constructed using a full-factorial design experiment with the parameters listed in Table 5. Details of the parameter combinations for the average models by ANN can be found in Supplementary Information (S3).

Figure 4 shows a comparison between the observed and predicted values for the regression and ANN models across the NW model, AD model excluding the area variable, and AD model incorporating the area variable. The out-of-sample R^2 values for parts (a) and (b) of Fig. 4 stand at 0.70 and 0.84, respectively. A closer look at Fig. 4, specifically (c), (d), (e), and (f), reveals that the inclusion of the area variable in the model enhanced its accuracy. The corresponding out-of-sample R^2 values for these sections were 0.29, 0.89, 0.76, and 0.90. For average model assessments, both the NW

Table 5 Parameters of ANN for the average models

NW model		AD model without area variable		AD model with area variable	
Hidden node	Threshold	Hidden node	Threshold	Hidden node	Threshold
2	0.0000005	2	0.000005	2	0.000005
4	0.000001	4	0.00001	4	0.00001
8	0.000005	8	0.00005	8	0.00005
15	0.00001	15	0.0001	16	0.0001
				31	

Fig. 4 Comparison of the actual and prediction incidence rates for the average models: **a** regression for NW, **b** ANN for NW, **c** regression for AD without area variable, **d** ANN for AD without area variable, **e** regression for AD with area variable, and **f** ANN for AD with area variable
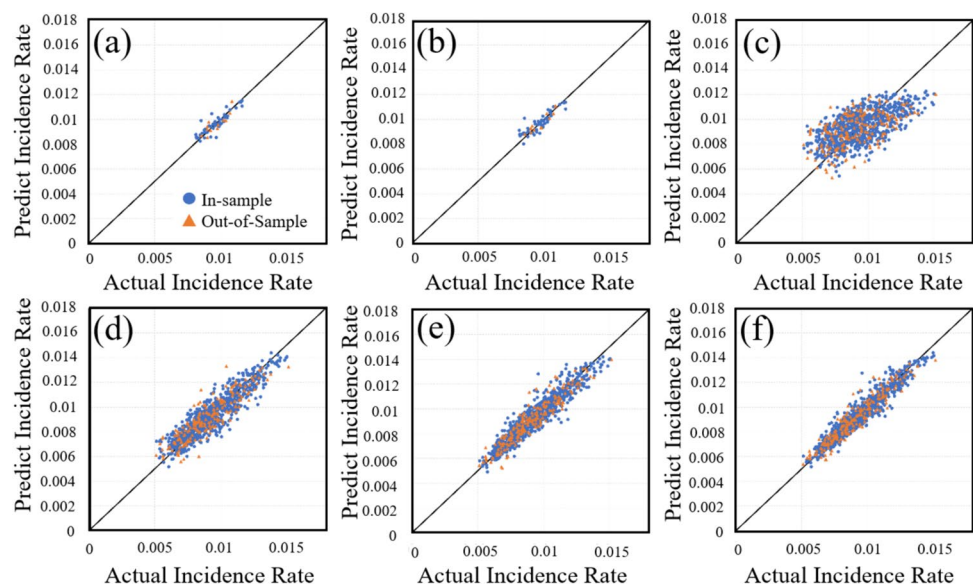
and AD models that included the area variable demonstrated that the regression and ANN models effectively predicted the atopic dermatitis incidence rate.

Sex and Age Model

Sex and age models differentiate themselves by utilizing sex and age as independent variables, rather than relying on population rates, as is common in average models. This approach enables the prediction of atopic dermatitis incidence with specificity for sex and age groups.

Sex and Age Model by Regression

The regression models for both NW and AD were tailored for each sex and age group, as shown in Table 6. These models were adapted from the averaging model mentioned in the previous section by specifically substituting the incidence rate with data segmented by sex and age. Notably, the coefficient of the effect of $\times 1$ (SO_2) was not significant in individuals over 10 years of age. Table 6 presents the coefficients of $\times 1$ and R^2 values for the regression model for the under 10 age group. Consequently, sex and age models have been structured to differentiate between individuals under 10 years old and those over 10 years old. This discussion focuses on the models pertaining to the under-10 age group, with models for the over-10 age group detailed in Supplementary Information (S4).

In the sex and age NW models, the independent variables included $\times 1$ (SO_2), meteorological factors, sex, age, and the interaction between sex and age, as shown in Eq. (6).

$$y \sim x1 + z1 + z2 + z3 + z4 + \text{Gender} \cdot \text{Age}. \qquad (6)$$

For the sex and age AD models, area variables were incorporated as independent variables, as indicated in Eq. (7). This approach allows for a more nuanced analysis

that considers not only environmental and meteorological factors but also demographic variables and their interactions, providing a comprehensive framework for understanding the incidence rates of atopic dermatitis.

$$y \sim x1 + z1 + z2 + z3 + z4 + \text{Gender} \cdot \text{Age} + \text{Area}. \qquad (7)$$

Table 7 presents the coefficients and R^2 values for the sex and age regression models for children under 10 years of age. Consistent with the average model, an increase in the incidence rate was observed with an increase in $\times 1$ (SO_2), $z1$, $z3$, and $z4$. The R^2 values for the NW and AD models, which included area variables, exceeded 0.8, indicating a strong fit. However, these regression models are not recommended for predicting the incidence of atopic dermatitis in the context of AD. The reasons for this recommendation are further discussed in conjunction with the sex and age models using an ANN. Additionally, the AD model, excluding the area variable, showed an R^2 values of 0.57 across the board, indicating that it falls short of accurately predicting incidence rates, akin to the average model.

Sex and Age Model by ANN

The development of the ANN model followed the same methodology as the average model, with a key distinction: instead of using the population rate, sex and age levels were employed as independent variables. Consequently, the total number of independent variables was 13 and 29 for the NW and AD models, respectively. Table 8 lists the ANN parameters of the ANN for the experimental design. For the detailed outcomes of the parameter combinations for the sex and age models by ANN, refer to Supplementary Information (S3).

Figure 5 compares the observed and predicted values for the regression and ANN models across the NW model, AD model excluding the area variable, and the AD model

**Table 6** Summary of regression results for each sex and age under 10 years old

Sex	Age level	Coefficients of $\times 1$	R^2
Nationwide			
Man	0–4	0.260	0.436
Woman	0–4	0.227	0.401
Man	5–9	0.041	0.556
Woman	5–9	0.044	0.529
Administrative districts (with area variable)			
Man	0–4	0.226	0.766
Woman	0–4	0.201	0.760
Man	5–9	0.083	0.729
Woman	5–9	0.103	0.702

Table 7 Coefficients and R^2 values of the sex and age models under 10 years old by regression

	NW model	AD model without area variable	AD model with area variable
Intercept	3.286	5.118	3.637
x1	0.131	− 0.017	0.110
z1	0.304	0.053	0.199
z2	− 0.113	− 0.016	− 0.090
z3	0.077	0.002	0.111
z4	0.061	− 0.069	0.037
R^2			
In-sample	0.916	0.571	0.837
Adjusted	0.913	0.570	0.836
Out-of-sample	0.901	0.569	0.830

Table 8 Parameters of ANN for the sex and age models under 10 years old

NW model		AD model without area variable		AD model with area variable	
Hidden node	Threshold	Hidden node	Threshold	Hidden node	Threshold
2	0.00001	2	0.0001	2	0.0001
4	0.00005	4	0.0005	4	0.0005
7	0.0001	7	0.001	8	0.001
13	0.0005	13	0.005	15	0.005
				29	

Fig. 5 Comparison of the actual and prediction incidence rates for the sex and age models under 10 years old: **a** regression for NW, **b** ANN for NW, **c** regression for AD without area variable, **d** ANN for AD without area variable, **e** regression for AD with area variable, and **f** ANN for AD with area variable
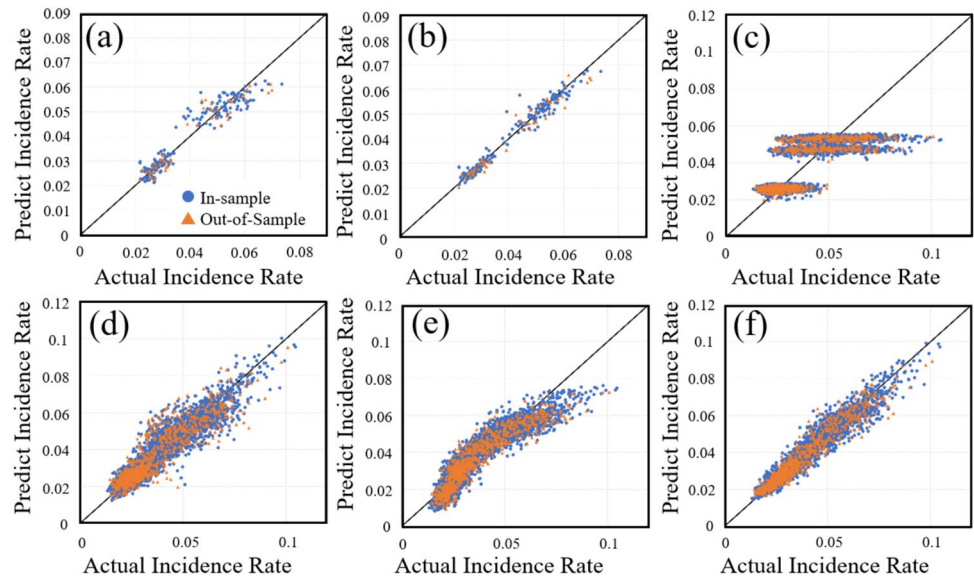
including the area variable. In Fig. 5a and b, the sex and age models, both via regression and ANN, accurately predict the incidence rate of atopic dermatitis, achieving out-of-sample R^2 values of 0.90 and 0.95, respectively. Figure 5c–f highlight the significance of including the area variables in the models. The out-of-sample R^2 values for these parts were 0.57, 0.79, 0.83, and 0.92, respectively, indicating an enhanced accuracy with the inclusion of the area variable. The notable disparity in the R^2 values between the regression and ANN models suggests a preference for the ANN model for more reliable predictions. Figure 5e illustrates the distortion in the predictions made by the regression model within the AD context, signifying that the sex and age AD model via regression warrants further investigation in future studies.

Table 9 displays the R^2 values for all the developed prediction models, indicating their predictive accuracy. The average models effectively predicted the atopic dermatitis incidence rate in both NW and AD contexts. Sex and age models provide detailed predictions for atopic dermatitis incidence rates across different sex and age groups. The models developed using ANN techniques exhibited higher accuracy than those based on regression analysis. However, regression models offer valuable insights into the

relationships between independent variables and the incidence of atopic dermatitis. The inclusion of the area variable enhanced the precision of the AD models, underscoring its importance in improving model performance.

Conclusion

In this study, we have explored the impact of air pollutants on the incidence of atopic dermatitis and formulated various predictive models. We developed average models for estimating the incidence rates across both nationwide and administrative district levels and sex and age models for forecasting the incidence rates by sex within specific age groups, applicable to both nationwide and administrative districts. Each model was constructed using the regression and ANN methods. In the regression models, a significant influence of SO_2 on the incidence rate. For the averaged models, both the regression and ANN approaches demonstrated strong predictive capabilities. However, in the sex and age models, the ANN models outperformed the regression models in terms of accuracy, particularly for AD predictions. While regression models have the advantage of clarifying the relationship between independent variables

Table 9 R^2 of the regression and ANN models by sex and age

	Regression			ANN		
	In-sample	Adjusted	Out-of-sample	Train	Valid	Test
Average						
NW model	0.794	0.738	0.698	0.752	0.863	0.836
AD model without area variable	0.422	0.414	0.292	0.839	0.812	0.794
AD model with area variable	0.891	0.888	0.887	0.919	0.906	0.903
Sex and age under 10 years old						
NW model	0.916	0.913	0.901	0.974	0.952	0.947
AD model without area variable	0.571	0.570	0.569	0.891	0.822	0.791
AD model with area variable	0.837	0.836	0.830	0.937	0.927	0.923
Sex and age over 10 years old						
NW model	0.909	0.907	0.885	0.985	0.975	0.965
AD model without area variable	0.677	0.677	0.680	0.778	0.763	0.766
AD model with area variable	0.804	0.803	0.800	0.930	0.931	0.916

and the atopic dermatitis incidence rate, the sex and age AD model via regression highlights the need for further investigation in future studies. Future studies should focus on enhancing model accuracy by incorporating more advanced ANN techniques and a wider time range of variables. Additionally, expanding research to include diverse population groups and different environmental conditions will improve the models' generalizability. Further exploration of sex and age-specific models using advanced ANN methods is essential to gain deeper insights into the factors influencing atopic dermatitis incidence.

Declarations

References

1. H.C. Williams, Epidemiology of atopic dermatitis. Clin. Exp. Dermatol.. Exp. Dermatol. **25**, 522 (2000)
2. B.J. Kim, J.-A. Jung, J.S. Lee, Association between social economic status and atopic dermatitis in Korean adult: an analysis of the fifth Korea national health and nutrition examination survey (2010–2012). Allergy Asthma Respir. Dis. **3**, 128 (2015)
3. J.M. Spergel, A.S. Paller, Atopic dermatitis and the atopic march. J. Allergy Clin. Immunol.Clin. Immunol. **112**, S118 (2003)
4. K. Kim, Influences of environmental chemicals on atopic dermatitis. Toxicol Res **31**, 89 (2015)
5. T. Zheng, J. Yu, M.H. Oh, Z. Zhu, The atopic march: progression from atopic dermatitis to allergic rhinitis and asthma. Allergy Asthma Immunol. Res. **3**, 67 (2011)
6. S.A.J. Schmidt, M. Olsen, M. Schmidt, C. Vestergaard, S.M. Langan, M.S. Deleuran, J.L. Riis, Atopic dermatitis and risk of atrial fibrillation or flutter: a 35-year follow-up study. J. Am. Acad. Dermatol.Dermatol. **83**, 1616 (2020)
7. J. Wang, C. Janson, A. Malinovschi, M. Holm, K.A. Franklin, L. Modig, A. Johannessen, V. Schlunssen, T. Gislason, N.O. Jogi, D. Norback, Asthma, allergic rhinitis and atopic dermatitis in association with home environment—the Rhine study. Sci. Total. Environ. **853**, 158609 (2022)
8. T.H. Park, S. Park, M.K. Cho, S. Kim, Associations of particulate matter with atopic dermatitis and chronic inflammatory skin diseases in South Korea. Clin. Exp. Dermatol.. Exp. Dermatol. **47**, 325 (2022)
9. S. Jung, J.-H. Kim, S.-W. Jeong, J.-W. Lee, B.-S. Son, Association between exposure to air pollutants and allergic diseases among residents near the Gwangyang industrial complex in korea. Toxicol. Environ. Health Sci. **15**, 425 (2023)
10. Q. Guo, X. Xiong, F. Liang, L. Tian, W. Liu, Z. Wang, X. Pan, The interactive effects between air pollution and meteorological factors on the hospital outpatient visits for atopic dermatitis in Beijing, china: a time-series analysis. J. Eur. Acad. Dermatol. Venereol. Dermatol. Venereol. **33**, 2362 (2019)

11. H. Kim, G. Byun, Y. Choi, S. Kim, S.Y. Kim, J.T. Lee, Effects of long-term exposure to air pollution on all-cause mortality and cause-specific mortality in seven major cities of south Korea: Korean national health and nutritional examination surveys with mortality follow-up. Environ. Res. **192**, 110290 (2021)

12. E. Vocks, R. Busch, C. Frohlich, S. Borelli, H. Mayer, J. Ring, Influence of weather and climate on subjective symptom intensity in atopic eczema. Int. J. Biometeorol.Biometeorol. **45**, 27 (2001)

13 K.H. Kim, Climate change and atopic dermatitis. J. Korean Med. Assoc. Taehan Uisa Hyophoe Chi **54**, 169 (2011)

14. J.I. Silverberg, J. Hanifin, E.L. Simpson, Climatic factors are associated with childhood eczema prevalence in the United States. J. Investig. Dermatol.Investig. Dermatol. **133**, 1752 (2013)

15 S.Y. Park, E.J. Byun, J.D. Lee, S. Kim, H.S. Kim, Air pollution, autophagy, and skin aging: Impact of particulate matter (pm(10)) on human dermal fibroblasts. Int. J. Mol. Sci. **19**, 2727 (2018)

16. S.-P. Jin, Z. Li, E.K. Choi, S. Lee, Y.K. Kim, E.Y. Seo, J.H. Chung, S. Cho, Urban particulate matter in air pollution penetrates into the barrier-disrupted skin and produces ros-dependent cutaneous inflammatory response in vivo. J. Dermatol. Sci.Dermatol. Sci. **91**, 175 (2018)

17. V. Patella, G. Florio, M. Palmieri, J. Bousquet, A. Tonacci, A. Giuliano, S. Gangemi, Atopic dermatitis severity during exposure to air pollutants and weather changes with an artificial neural network (ann) analysis. Pediatr. Allergy Immunol.. Allergy Immunol. **31**, 938 (2020)

18. G. Hurault, V. Delorieux, Y.M. Kim, K. Ahn, H.C. Williams, R.J. Tanaka, Impact of environmental factors in predicting daily severity scores of atopic dermatitis. Clin. Transl. Allergy **11**, e12019 (2021)

19. Y.-Y. Hsiao, Y.-H. Chen, W.-T. Hung, K.-T. Tang, The relationship between outdoor air pollutants and atopic dermatitis of adults: a systematic review and meta-analysis. Asian Pac. J. Allergy Immunol.Immunol. **40**, 295 (2022)

20. Y. Huang, H.-J. Wen, Y.-L.L. Guo, T.-Y. Wei, W.-C. Wang, S.-F. Tsai, V.S. Tseng, S.-L.J. Wang, Prenatal exposure to air pollutants and childhood atopic dermatitis and allergic rhinitis adopting machine learning approaches: 14-year follow-up birth cohort study. Sci. Total. Environ. **777**, 145982 (2021)

21. C. Wang, C.C. Wei, L. Wan, C.L. Lin, J.D. Tsai, Association of exposure to hydrocarbon air pollution with the incidence of atopic dermatitis in children. Ital. J. Pediatr.Pediatr. **47**, 202 (2021)

22. R.P. Fadadu, K. Abuabara, J.R. Balmes, J.M. Hanifin, M.L. Wei, Air pollution and atopic dermatitis, from molecular mechanisms to population-level evidence: a review. Int. J. Environ. Res. Public Health **20**, 2526 (2023)

23 J.S. Youn, J.W. Seo, S. Park, K.J. Jeon, Prediction model for dry eye syndrome incidence rate using air pollutants and meteorological factors in south Korea: analysis of sub-region deviations. Int. J. Environ. Res. Public Health **17**, 4969 (2020)

24. A.J. Thomas, M. Petridis, S.D. Walters, S.M. Gheytassi, R.E. Morgan, Two hidden layers are usually better than one. Commun. Comput. Inf. Sci. **744**, 279 (2017)

25. Wanas, N., Auda, G., Kamel, M. S., Karray, F., On the optimal number of hidden nodes in a neural network, Conference proceedings, *IEEE Canadian conference on electrical and computer engineering* (Cat. No. 98TH8341) (IEEE, 1998). p. 918–921.

26. G. Shmueli, P.C. Bruce, P. Gedeck, N.R. Patel, *Data Mining for Business Analytics: Concepts, Techniques and Applications in Python* (Wiley, New York, 2019)