PollenNet - a deep learning approach to predicting airborne pollen concentrations

Rebeka Čorić[1,*], **Domagoj Matijević**[1] and **Darija Marković**[1]
[1] *Department of Mathematics, J. J. Strossmayer University of Osijek, Trg Ljudevita Gaja 6, Osijek , Croatia*
E-mail: ⟨{*rcoric, domagoj, darija*}*@mathos.hr*⟩

Abstract. The accurate short-term forecasting of daily airborne pollen concentrations is of great importance in public health. Various machine learning and statistical techniques have been employed to predict these concentrations. In this paper, an RNN-based method called PollenNet is introduced, which is capable of predicting the average daily pollen concentrations for three types of pollen: ragweed (*Ambrosia*), birch (*Betula*), and grass (*Poaceae*). Moreover, two strategies incorporating measurement errors during the training phase are introduced, making the method more robust. The data for experiments were obtained from the RealForAll project, where pollen concentrations were gathered using a Hirst-type 7-day volumetric spore trap. Additionally, five types of meteorological data were utilized as input variables. The results of our study demonstrate that the proposed method outperforms standard models typically used for predicting pollen concentrations, specifically the pollen calendar method, pollen predictions based on patterns, and the naive approach.

Keywords: LSTM, pollen, predictions, RNN

Received: March 2, 2023; accepted: May 23, 2023; available online: July 10, 2023

DOI: 10.17535/crorr.2023.0001

1. Introduction

One of the more important elements in the public health system is the short-term forecasting of daily airborne pollen concentrations. The quality of life for those with allergic rhinitis is significantly improved by forecasting these amounts and using real-time monitoring. Accurate forecasts a few days in advance enable people to schedule their activities to prevent exposure to excessive pollen concentrations. The focus of this study is on predicting concentrations for three types of pollen: ragweed (*Ambrosia*), birch (*Betula*), and grass (*Poaceae*). Birch and grass are responsible for the majority of allergy reactions in Europe, according to Bousquet et al. [3], whereas ragweed is the second-most prominent allergen with the vastly enhanced clinical relevance [19].

To precisely predict the pollen patterns, it is beneficial to account for various aspects such as the life cycle of the plant species, meteorological data, and ecological data, aside from the concentration of airborne pollen. As these factors differ between different locations, the forecasting models are typically formulated for a particular plant category and region. Various elements such as temperature, rainfall, sunshine hours, cloud cover, relative humidity, wind speed, and wind direction can potentially impact the pollen trends [20]. According to previous studies, temperature affects flowering season, wind speed tends to increase pollen emission, while humidity and precipitation reduce it.

As mentioned in [20], the most extensively employed pollen observation system involves the manual examination of a tape collected from a volumetric spore sampler that records data over a period of 7 days. Despite the disadvantages of the system, the existing set of input data is the optimal choice for daily operational pollen forecasting.

*Corresponding author.

1.1. Previous work

A variety of methods is used for predicting airborne pollen concentrations [13, 20]. One of the most used approach in practice is a straightforward approach known as a pollen calendar. To predict pollen concentration on a given day, it uses means/medians of concentrations on the same day in previous years that are available in the dataset.

Many research papers use classical machine learning and statistical methods to determine pollen concentrations. Csépe et al. [6] used a neural network with one hidden layer to determine the category to which ragweed pollen concentration for a given day will belong. They used pollen concentrations from the previous days and meteorological data based on temperature, precipitation, wind speed, and humidity to determine the correct category.

Zewdie et al. [22] used a deep neural network with two hidden layers, random forest, extreme gradient boosting, and Bayesian ridge algorithm to determine daily pollen concentration for ragweed pollen. They used daily pollen concentrations and extensive meteorological and land surface contextual data for their models.

Liu et al. [11] used random forest, neural network, and LASSO regression to determine daily ragweed pollen concentration. They used daily pollen concentration from previous years and 85 environmental parameters from NASA MERRA meteorological analysis as input to their models.

Lara et al. [10] used the decomposition of time series of pollen data for short- and long-term patterns of variations and an additive model for combining components of decomposed time series to obtain thresholds for plane tree pollen concentrations. They made predictions based on previous daily pollen concentrations and variables based on temperature, rainfall, and humidity.

Am Seo et al. [2] used a deep neural network (DNN)-based estimation model containing five hidden layers to compute the oak pollen concentration, associated risk levels, and the duration of the yearly pollen season. They employed seven distinct weather variables for this purpose. To prevent the DNN model from over-fitting and underestimating the data, they integrated a bootstrap aggregating-type ensemble model.

Lo et al. [12] developed and evaluated different random forest models to predict the daily pollen concentrations for four types of pollen 1-14 days in advance. The forecasted pollen concentration relied on a combination of meteorological and vegetational variables, and pollen observations.

Goudarzi et al. [8] employed an artificial neural network for forecasting total pollen concentration and investigated the interdependence between pollen concentrations and environmental parameters. The ANN included an input layer containing 13 parameters, a hidden layer with five neurons, and an output layer. Their findings indicated a negative correlation between the pollen concentration and the relative humidity, precipitation, and air pressure, while displaying a positive correlation with temperature.

1.2. Contribution

Since they were first proposed in the papers [17] and [21], recurrent neural networks (RNN) have played an interesting and significant role in neural network research. They have been successfully applied to solve a variety of problems involving time sequences of events and ordered data [14], including machine translation [4], speech recognition [16], human action recognition [7], robot control [9], etc. To the best of our knowledge, the application of sequence-to-sequence RNNs to the problem of predicting pollen concentrations has not been done, although it was used to predict air quality [15]. In our model, both encoder and decoder RNN's use long short-term memory (LSTM). On top of the encoder, we implemented an attention mechanism to allow the network to learn the influence of past days on the forecast of current day's pollen

concentration. Moreover, we implemented two different simulation-based approaches to make our method robust to unavoidable measurement errors in real-world scenarios.

To maintain a simplistic and easy-to-use model, while preserving its expressive capacity, only those variables that are easily accessible and have been previously identified as influential have been incorporated.

Unfortunately, implementations of classical machine learning methods listed in section 1.1 are not publicly available. Moreover, the data they used is also not publicly available. Thus, comparing our approach against classical machine learning methods was not possible in practice. Hence, we experimentally evaluated and demonstrated that our approach outperforms the pollen calendar method, which is typically used in modern pollen prediction systems. In addition, we show that our method outperforms pollen predictions based on patterns [5] and the naive approach that predicts concentration for a given day by copying the pollen concentration seen on the previous day.

1.3. Paper organization

The paper is organized as follows: Section 2 describes the methodology used for predicting pollen concentrations and additional strategies for training, which incorporate measurement error into the training process to make the model more robust. In section 3, an experimental setup was given, and the results are described. Finally, section 4 gives final remarks and concludes the paper.

2. Method

Here we formally introduce and define sequence-to-sequence RNN with attention. We refer an interested reader to the standard textbook for more details [1] (Section 10.2.2. Attention Mechanisms for Machine Translations), which focuses on several ways attention can be incorporated into neural machine translation. The schematic representation of RNN is given in Figure 1. Let $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)})^T \in \mathbb{R}^n$ denote the vector of input features and without the loss of generality suppose $m^{(t)} = (x_1^{(t)}, \ldots, x_{n-1}^{(t)})$ represent meteorology data and $x_n^{(t)}$ pollen concentration at time t. We will assume that the length of the input to the encoder is hyperparameter k, i.e. the k days in the past relevant for the prediction, and the length of the decoder is l, i.e. the prediction is being made for l days in the future.

Furthermore, let $(h_{enc}^{(t)})_{t=0}^{k}$ and $(h_{dec}^{(t)})_{t=0}^{l}$ denote hidden states from the encoder and decoder, respectively. Our decoder RNN is forward LSTM on top of which a particular attention mechanism is implemented enhancing the decoder hidden state. Note that each LSTM unit has additional hidden vectors $(c_{enc}^{(t)})_{t=0}^{k}$ and $(c_{dec}^{(t)})_{t=0}^{l}$ (*cell states*) associated with it. The transition from the encoder to the decoder is defined as follows:

$$\begin{aligned} h_{dec}^{(0)} &= h_{enc}^{(k)} \in \mathbb{R}^h \\ c_{dec}^{(0)} &= c_{enc}^{(k)} \in \mathbb{R}^h, \end{aligned} \tag{1}$$

where h is a hyperparameter denoting the size of internal hidden LSTM vectors. After the decoder is initialized, we provide it with meteorology data (weather forecasting data) $m^{(t)} \in \mathbb{R}^{n-1}$. More generally, in time t decoder is defined as

$$h_{dec}^{(t)}, c_{dec}^{(t)} = \text{Decoder}(m^{(t)}, h_{dec}^{(t-1)}, c_{dec}^{(t-1)}). \tag{2}$$

Given $h_{dec}^{(t)}$ we further compute attention weights over $h_{enc}^{(1)}, \ldots, h_{enc}^{(k)}$ in order to give our model additional chance to understand which days from the past are influencing the pollen concen-

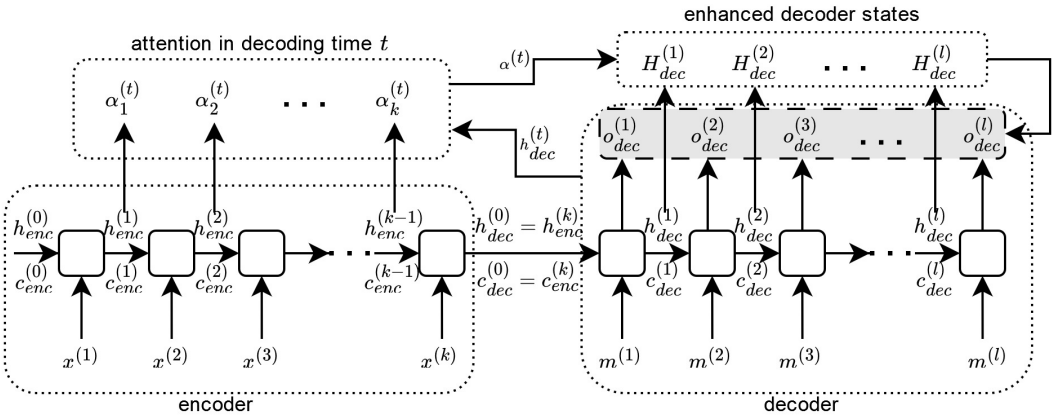


Figure 1: The architecture follows the sequence-to-sequence (encoder/decoder) approach, with an additional attention mechanism implemented on top. At each decoding time $t = 1, \ldots, l$, the attention mechanism produces a softmax distribution α^t that leverages the importance of days in the past and helps compute the enhanced state $H_{dec}^{(t)}$. Given the enhanced state $H^{(t)}$, the pollen prediction $o_{dec}^{(t)}$ is computed.

tration at time t. The attention weights are computed as follows:

$$\begin{aligned} e_i^{(t)} &= (h_{enc}^{(i)})^T W_1 h_{dec}^{(t)}, i = 1, \ldots, k, \text{ and } e^{(t)} \in \mathbb{R}^h, W_1 \in \mathbb{R}^{h \times h} \\ \alpha^{(t)} &= \text{softmax}(e^{(t)}). \end{aligned} \tag{3}$$

Letting the network learn the attention weights will enable our model to understand better what days in history are relevant for prediction at time t. Given the attention weights, we can compute additional contextual information $a^{(t)}$ of the encoder hidden states that is most relevant from the decoder hidden state at a given time t. In other words, we define

$$a^{(t)} = \sum_{i=1}^{k} \alpha_i^{(t)} h_{enc}^{(i)}. \tag{4}$$

Finally, we create a new enhanced decoder hidden state $H_{dec}^{(t)}$ that combines the information in $a^{(t)}$ and $h_{dec}^{(t)}$ as follows:

$$H_{dec}^{(t)} = \tanh(W_2 \begin{bmatrix} a^{(t)} \\ h_{dec}^{(t)} \end{bmatrix}), \text{ and } W_2 \in \mathbb{R}^{h \times 2h}, H_{dec}^{(t)} \in \mathbb{R}^h. \tag{5}$$

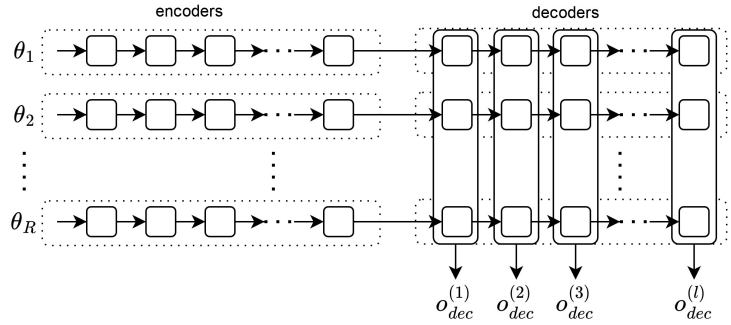
Our attention layer introduces new weight matrices $W_1 \in \mathbb{R}^{h \times h}$ and $W_2 \in \mathbb{R}^{h \times 2h}$, i.e. we introduce an additional set of $O(h^2)$ parameters into our model. In order to gain control over this new richer model, we introduce dropout as an additional regularization step and compute the final scalar value in time t as pollen concentration prediction as follows:

$$o_{dec}^{(t)} = W \cdot \text{dropout}(\tanh(H_{dec}^{(t)})), \text{ and } W \in \mathbb{R}^{h \times 1}. \tag{6}$$

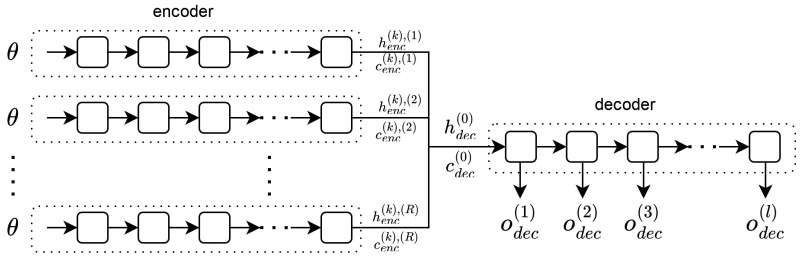
We trained our networks with the standard mean square error (MSE) loss function

$$\sum_{t=1}^{l} (o_{dec}^{(t)} - x_n^{(t)})^2, \tag{7}$$

where $x_n^{(t)}$ stands for an actual pollen concentration value in decoding time t.



(a) RM strategy



(b) SM strategy

Figure 2: We demonstrate for both approaches the logic during the training. Note that in the RM approch we train R different models in parallel. SM strategy evaluates the very same encoder on R different inputs, combine the given encoder hidden states into one, and move on to the decoding phase.

2.1. Measurement errors

To accommodate the known measurement error uncertainties we use two different simulation-based approaches: an approach in which we will train separate models for each dataset that simulates measurement error in pollen concentration; an alternative and more involving approach in which different datasets simulating pollen measurement error will be exposed to just a single RNN model.

The basic idea in both approaches is to simulate multiple data sets from a given one using the known measurement error uncertainties (e.g. errors in the measurement system, measured by careful calibration of the system). The common assumption is that the errors measured in time t are i.i.d. Gaussian, i.e. $\epsilon_t \sim N(0, \sigma_t^2)$, where σ_t^2 is known or can be accurately estimated. Note that for each time t variance σ_t^2 may differ, hence we refer to it as a heteroscedastic measurement error.

Let $x^{(i)} \in \mathbb{R}^n$, for $i = 1, \ldots, m$ denote our given training dataset. As before, we assume that first $n-1$ components of $x^{(i)}$, i.e. $x_1^{(i)}, \ldots, x_{n-1}^{(i)}$, represent meteorology data and $x_n^{(i)}$ represent pollen concentration. By sampling R times each $x_n^{(i),(r)} \sim N(x_n^{(i)}, \sigma_i^2)$, $r = 1, \ldots, R$, for all $i = 1 \ldots, m$, we generate R perturbed datasets $S_1, \ldots, S_R$. Note that the meteorology data in perturbed datasets doesn't change.

By generating R simulated datasets and providing all of them (in some sense) to our model we will not necessarily improve the model accuracy. Our intention is to incorporate more reliable and more trustworthy results by giving our model the chance to absorb the knowledge of uncertainty introduced by the measurement error.

In order to utilize R different datasets we adopt two different strategies which we will refer

to as RM and SM strategies (RM will stand in short for computing R different **M**odels, while SM stands for considering R different datasets with a **S**ingle **M**odel). In the following we will explain both strategies in detail.

RM strategy. The easiest way of incorporating the measurement error uncertainty is to train our model R times, i.e. on each of the R simulated data sets. Very similar approach has been used and showed promising in a recent result of [18] in the context of incorporating measurement error in astronomical object classification.

Given $S_1, \ldots, S_R$ let $\theta_1, \ldots, \theta_R$ denote the optimal set of model parameters, respectively. During the inference, all R models are evaluated and the average pollen prediction value is reported. More precisely, the prediction scalar in time t is computed according to the following:

$$o_{dec}^{(t)} = \frac{1}{R} \sum_{i=r}^{R} o_{dec}^{(t),(r)}, \tag{8}$$

where $o_{dec}^{(t),(r)}$ denotes the output given by equation (6), for r'th simulated data set (see Figure 2a).

SM strategy. The more involving strategy to deal with the uncertainty that comes from the measurement error is to use a single model with weights θ and to give model a chance to 'see' all simulated data sets during the training. Since in our model an encoder RNN is the part of the network that is responsible for learning, we will tend to provide all R simulated data sets to the encoder. Let $h_t^{enc,(r)}$ denote the encoder hidden state in time t for data set r, where $t = 1, \ldots, k$ and $r = 1, \ldots, R$. Similar to (1), the transition from the encoder to the decoder is now defined as follows:

$$\begin{aligned} h_{dec}^{(0)} &= \frac{1}{R} \sum_{r=1}^{R} h_{enc}^{(k),(r)} \in \mathbb{R}^h \\ c_{dec}^{(0)} &= \frac{1}{R} \sum_{r=1}^{R} c_{enc}^{(k),(r)} \in \mathbb{R}^h. \end{aligned} \tag{9}$$

After the transition from the encoder to the decoder has been made, the decoding decoder RNN proceeds exactly as before. Note that we indeed trained R different encoder RNNs but they all share the same weights (see Figure 2b).

3. Experiments

All experiments were done using pollen concentrations and meteorological data for the city of Novi Sad from the year 2000 to the year 2021. The data came from RealForAll project ("Real-time measurements and forecasting for successful prevention and management of seasonal allergies in Croatia-Serbia cross-border region" - RealForAll (2017HR-RS151) - https://www.realforall.com/) which was active from 15.07.2017. to 14.01.2020. and dealt with forecasting pollen concentrations from day to day.

All our implementations are publicly available under the MIT Licence and can be accessed at https://github.com/dmatijev/polen_forcasting. The data presented in this study are available on request from the corresponding author.

The data consists of average daily pollen concentrations and meteorological data. The daily pollen concentrations are measured for three types of pollen: ragweed (*Ambrosia*), birch (*Betula*), and grass (*Poaceae*). Pollen concentrations were collected using the 7-day volumetric

Pollen	Model	dropout	lrn. rate	batch size	hidden dim.	seq. length
ragweed	PollenNet	0.1	0.001	32	256	7
birch	PollenNet	0.1	0.005	128	256	7
grass	PollenNet	0.25	0.005	128	64	7
ragweed	RM	0.1	0.001	32	256	7
birch	RM	0.1	0.005	128	256	7
grass	RM	0.25	0.005	128	64	7
ragweed	SM	0.25	0.005	1	64	5
birch	SM	0.1	0.005	1	64	7
grass	SM	0.1	0.001	1	64	7

Table 1: *Hyperparameters values selected after cross-validation*

spore trap of the Hirst design and are measured as the number of particles per cubic meter of air (P/m^3). The meteorological data mainly consists of temperature, wind, and rainfall data. For the experiments, the following meteorological data was used: minimal daily temperature (MNT), maximal daily temperature (MKT), precipitation (PAD), relative humidity (VLZ), and maximum daily wind speed (MBV).

For every of the three pollen types, all three models (PollenNet, RM, and SM strategy) described in section 2 were used, and the obtained results are compared to the pollen calendar, the naive model, and the prediction based on patterns ([5]).

The pollen calendar predicts pollen concentration for a given day by taking the mean value of pollen concentrations on that day in all previously available years.

The naive model predicts pollen concentration for a given day simply by copying the pollen concentration from the previous day, i.e., the model states that pollen concentration for today will be the same as yesterday's observed pollen concentration. If one wants to predict pollen concentration two days in advance using this naive model, the pollen concentration for today and tomorrow will be the same as yesterday's pollen concentration. This naive model serves as a baseline to see if the proposed new models improve pollen predictions.

The prediction based on patterns uses clustering of patterns, and computes a prediction of the target value as the mean of values from the same cluster, minimizing the total squared deviation between predicted and actual values of the target variable. Every pattern consists of consecutive days where pollen concentrations grow to the local maxima and fall down until the end of the pattern. Several features are calculated for every pattern, and patterns are clustered into similar groups based on them. In order to predict pollen concentration for a new day, one should make a pattern consisting of previous days leading up to that day, determine to which cluster the new pattern belongs, and then predict the pollen concentration value based on the mean of values in the same cluster. This approach can predict pollen concentration only one day in advance.

For all approaches, the dataset was reduced to only those months of the year where the corresponding pollen is present. So for ragweed pollen, only days in July, August, September, and October were used; for birch pollen, only days in March, April, and May were used, while for grass pollen, days in April, May, June, July, August, September and October were used. Outside of given months, predictions of pollen concentrations are equal to zero.

3.1. Results

For all our models and different data inputs (ragweed, birch or grass), we have determined the optimal values for the following hyperparameters: the batch size, the input sequence length (i.e., the number of days before the one for which the prediction is calculated), the learning rate, the hidden dimension of hidden states and the dropout rate using cross-validation.

Experiment	Training years	Validation years	Test years
exp1	2000 - 2011	2012, 2013	2014, 2015
exp2	2000 - 2012	2013, 2014	2015, 2016
exp3	2000 - 2013	2014, 2015	2016, 2017
exp4	2000 - 2014	2015, 2016	2017, 2018
exp5	2000 - 2015	2016, 2017	2018, 2019
exp6	2000 - 2016	2017, 2018	2019, 2020
exp7	2000 - 2017	2018, 2019	2020, 2021

Table 2: *List of years used for training, validating and testing the models*

3.1.1. Cross-validation

Cross-validation is a commonly used technique for determining optimal hyperparameters in machine learning models. Cross-validation can prevent overfitting and help evaluate the model performance more robustly than a simple train-test approach. In cross-validation, the available data is split into k equally sized subsets (known as "folds"). The model is then trained on $k-1$ of the folds and evaluated on the remaining fold. This process is repeated k times, with each of the k folds being used as the evaluation set once. The results from each of the k evaluations are averaged to produce a final performance metric for the model. Specifically, our approach chooses two consecutive years for testing and all the remaining years for training. Then the next two consecutive years are used for testing, and the remaining year for training, etc.

To perform cross-validation for hyperparameter tuning, we set the hyperparameter search space as follows: the dropout rate was selected from $\{0.1, 0.25\}$, the learning rate from $\{0.001, 0.005\}$, the batch size from $\{1, 32, 64, 128\}$, the hidden dimension from $\{64, 128, 256\}$ and the input sequence length from $\{3, 5, 7\}$.

We trained our models on the training data for each combination of hyperparameters (the grid search approach) using k-fold cross-validation. The MSE loss was calculated for each fold, and the average performance over all k folds was used as the final evaluation for each hyperparameter combination. The hyperparameters with the best performance were chosen for the different data inputs (see Table 1).

3.1.2. Models evaluation

After determining the needed parameters, models were trained on a given set of years, validated on two consecutive years, which come after the training years, and tested on two consecutive years, which come after the validation years. The complete set of years for conducted experiments can be seen in Table 2, where column *Experiment* denotes the ordinal number of the experiment, columns *Training years*, *Validation years* and *Test years* show which years were used for training, validation, and testing, respectively in a given experiment.

Tables 3, 4 and 5 give a comparison of MSE loss on given test years for predictions of pollen concentrations one day (rows *day 1*) and two days (rows *day 2*) in advance by using PollenNet (column *PollenNet*), RM strategy (column *RM*), SM strategy (column *SM*), the pollen calendar (column *pollenCal*), naive model (column *naive*) and predictions based on patterns (column *patterns*) for ragweed, birch and grass pollen, respectively. Additionally, average MSE loss values and standard deviations of given values are calculated for all test years to compare how different models behave on all test sets.

All tables denote the best-obtained values for given test years and a given day in bold. RNN-based models almost always outperform the naive model and consistently outperform predictions obtained by the pollen calendar and using patterns. Usually, RM and SM outperform the PollenNet model, which justifies their usage. Incorporating measurement error into

learning helps the network to obtain better predictions on unseen examples.

test years		PollenNet	RM	SM	pollenCal	naive	patterns
2014-2015	day 1	106.314	**95.249**	98.784	100.001	115.941	106.185
	day 2	120.153	**99.715**	102.820	100.001	133.056	-
2015-2016	day 1	100.807	**100.652**	102.666	111.911	122.055	115.452
	day 2	108.617	**107.512**	109.855	111.911	132.505	-
2016-2017	day 1	101.173	91.016	**86.328**	93.521	102.238	103.855
	day 2	111.628	102.970	93.722	**93.521**	113.899	-
2017-2018	day 1	96.686	98.359	103.111	119.500	**95.853**	121.691
	day 2	**106.915**	108.648	111.725	119.500	123.665	-
2018-2019	day 1	102.442	102.431	**99.631**	152.352	103.634	153.635
	day 2	106.448	110.341	**101.504**	152.352	132.972	-
2019-2020	day 1	87.802	**80.247**	84.586	128.206	97.041	141.626
	day 2	92.040	**86.790**	89.940	128.206	116.936	-
2020-2021	day 1	73.778	**71.071**	75.133	91.161	90.273	107.852
	day 2	83.544	**80.818**	84.569	91.161	105.044	-
average	day 1	95.572	**91.289**	92.891	113.807	103.862	121.471
	day 2	104.192	99.542	**99.162**	113.807	122.582	-
stdev	day 1	11.240	11.606	**10.859**	21.772	11.364	19.177
	day 2	12.357	11.460	**10.154**	21.772	11.046	-

Table 3: *Comparison of MSE loss for ragweed pollen predictions for different models*

test years		PollenNet	RM	SM	pollenCal	naive	patterns
2014-2015	day 1	252.237	277.768	**240.754**	358.024	311.701	360.312
	day 2	282.447	298.123	**278.078**	358.024	421.365	-
2015-2016	day 1	145.674	**143.772**	146.431	235.990	172.992	234.667
	day 2	174.732	**172.630**	210.986	235.990	255.135	-
2016-2017	day 1	135.099	**132.285**	150.880	203.065	158.929	201.156
	day 2	**151.369**	156.970	172.168	203.065	225.499	-
2017-2018	day 1	107.902	100.303	**99.380**	132.278	121.699	128.166
	day 2	**109.783**	119.650	115.024	132.278	155.646	-
2018-2019	day 1	**108.938**	109.199	116.159	135.729	111.567	115.825
	day 2	132.168	128.211	**115.195**	135.729	153.656	-
2019-2020	day 1	160.434	**154.132**	157.206	191.875	188.293	187.979
	day 2	177.151	**172.947**	183.739	191.875	237.441	-
2020-2021	day 1	173.777	**159.672**	166.626	186.587	194.097	204.086
	day 2	184.510	**182.441**	190.867	186.587	247.253	-
average	day 1	154.866	**153.876**	153.920	206.221	179.897	204.599
	day 2	**173.166**	175.853	180.865	206.221	242.285	-
stdev	day 1	49.430	58.903	**45.052**	76.342	66.065	80.790
	day 2	**55.171**	58.875	56.553	76.342	89.374	-

Table 4: *Comparison of MSE loss for birch pollen predictions for different models*

Based on the average MSE loss for pollen concentrations of ragweed (Table 3), for one day in advance, the RM gives the smallest loss, i.e., it gives the best pollen predictions on average for unseen instances, while for two days in advance, the SM gives the smallest loss, although the RM is also very close in value. Also, based on standard deviations of MSE losses for pollen

test years		PollenNet	RM	SM	pollenCal	naive	patterns
2014-2015	day 1	11.742	**11.690**	12.172	14.793	13.358	12.684
	day 2	**12.663**	12.738	12.762	14.792	17.293	-
2015-2016	day 1	15.472	**15.351**	15.428	18.025	19.123	15.962
	day 2	**16.041**	16.193	16.164	18.025	22.542	-
2016-2017	day 1	12.647	**12.539**	12.645	13.746	16.729	14.417
	day 2	12.787	12.645	**12.598**	13.746	18.003	-
2017-2018	day 1	8.248	8.256	**8.108**	12.585	8.943	9.373
	day 2	8.923	8.939	**8.671**	12.585	10.194	-
2018-2019	day 1	8.293	7.679	7.705	13.238	**7.402**	7.395
	day 2	9.467	8.639	**8.382**	13.238	9.148	-
2019-2020	day 1	5.052	4.954	**4.858**	9.837	6.230	6.618
	day 2	6.031	5.409	**5.332**	9.837	7.280	-
2020-2021	day 1	9.217	**9.193**	9.384	13.614	9.605	11.573
	day 2	10.791	**10.676**	10.859	13.614	12.321	-
average	day 1	10.096	**9.952**	10.043	13.691	11.627	11.146
	day 2	10.958	10.748	**10.681**	13.691	13.826	-
stdev	day 1	3.440	3.475	3.578	**2.459**	4.890	3.516
	day 2	3.234	3.499	3.560	**2.459**	5.562	-

Table 5: *Comparison of MSE loss for grass pollen predictions for different models*

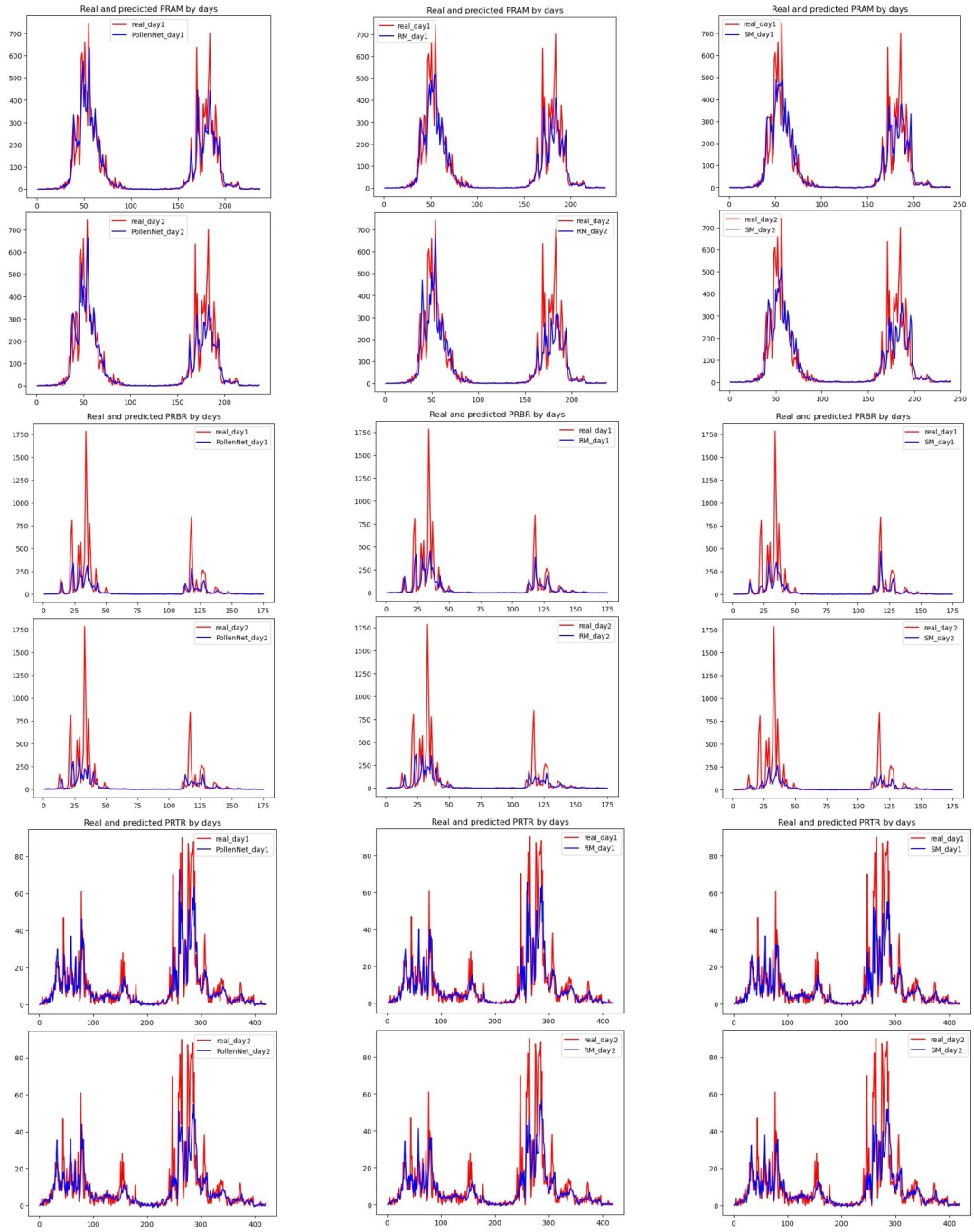
predictions, the SM has the least dispersed values, although the standard deviations for other models are not much more different than those of SM.

In Table 4, for one day in advance, RM has the smallest average MSE loss for birch pollen concentration predictions, similar to ragweed pollen, but the average loss for SM is very similar to that of RM. For two days in advance, the PollenNet has the smallest mean MSE loss and the smallest standard deviation, i.e., the best pollen predictions, which are the least dispersed of all observed models. On the other hand, the dispersion is the smallest on SM for one day in advance, which can indicate (considering the small average MSE loss) that for one day in advance, SM could be used to predict birch pollen concentrations.

For grass pollen concentrations (Table 5), all three RNN-based models perform similarly and have almost the same standard deviations of obtained MSE losses, which means that any of the three proposed methods can be used to predict grass pollen concentrations.

Figure 3 compares daily pollen concentrations and daily pollen predictions obtained by PollenNet, RM, and SM for one day and two days in advance in 2020 and 2021 for ragweed, birch, and grass pollen, respectively. For all types of pollen and all types of proposed models, the predicted concentrations follow the trend of movement of daily pollen concentrations. It can also be noticed that the birch pollen concentrations achieve the highest values, which explains the largest MSE loss values in Table 4. It can also be seen that for that type of pollen, proposed models never achieve those peak values. They always underestimate pollen concentrations in those cases. However, it can be noted that the predicted values for those days are larger than those for the days before and after the peak in real value, which means that the models can predict when the peak will occur. The possible reason for this behavior is that the extreme values are scarce in the dataset, and the model adapts to most values in the dataset, which are significantly smaller than those peak values.

For ragweed and grass pollen predictions, the predicted peaks are much closer to the actual peaks because the range of real pollen concentrations is much smaller than for the birch. This can also be noted in corresponding tables where MSE loss values are much smaller than for birch pollen.



(a) PollenNet (b) RM (c) SM

Figure 3: The comparison of real daily pollen concentrations and daily pollen predictions obtained by PollenNet, RM, and SM for one day and two days in advance in 2020 and 2021 for ragweed, birch, and grass pollen. The values on the x-axis represent the ordinal number of datapoint in the dataset, while the values on the y-axis represent average daily pollen concentrations.

4. Conclusion

In the present study, the PollenNet model, which utilizes a sequence-to-sequence RNN with an attention layer, was employed to establish the nonlinear relationship between pollen concentrations and meteorological data. To ensure a simple and user-friendly model without compromising its explanatory power, only those variables that are easily accessible and have been previously identified as influential were included. To make our approach more robust to unavoidable measurement errors, we have adopted two simulation-based approaches that have significantly improved the learning phase of our method.

In order to validate the model's efficacy, it was compared against established methods commonly used for predicting airborne pollen concentration. In the majority of experiments conducted, the new model demonstrated superior performance compared to the traditional approaches.

However, the prediction of peak pollen concentrations requires improvement, as the results indicate that the model cannot properly simulate severe and extreme grades. Moreover, as expected, the forecasting accuracy diminishes over time. Possible strategies to improve the model in both cases could be adding or removing different variables. Further research is required to identify the optimal and most appropriate model.

This research can serve as a foundation for utilizing RNNs in predicting airborne pollen concentrations, given that modern RNNs are ideally suited for such tasks.

References

[1] Aggarwal, C. C. (2018). *Neural networks and deep learning: A textbook* (1st). Springer Publishing Company, Incorporated.

[2] Am Seo, Y., Kim, K. R., Cho, C., Oh, J.-W., and Kim, T. H. (2020). Deep neural network-based concentration model for oak pollen allergy warning in south korea. Allergy, Asthma & Immunology Research, 12(1), 149–163. doi: 10.4168/aair.2020.12.1.149

[3] Bousquet, P.-J., Chinn, S., Janson, C., Kogevinas, M., Burney, P., and Jarvis, D. (2007). Geographical variation in the prevalence of positive skin tests to environmental aeroallergens in the european community respiratory health survey I. Allergy, 62(3), 301–309. doi: 10.1111/j.1398-9995.2006.01293.x

[4] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734. doi: 10.3115/v1/D14-1179

[5] Čorić, R., umić, M., and Jelić, S. (2019). A clustering model for time-series forecasting. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1105–1109. doi: 10.23919/MIPRO.2019.8756806

[6] Csépe, Z., Leelőssy, Á., Mányoki, G., Kajtor-Apatini, D., Udvardy, O., Péter, B., Páldy, A., Gelybó, G., Szigeti, T., Pándics, T., Kofol-Seliger, A., Simčič, A., Leru, P. M., Eftimie, A.-M., Šikoparija, B., Radišić, P., Stjepanović, B., Hrga, I., Večenaj, A., Vucić, A., Peroš-Pucar, D., Škorić, T., Ščevková, J., Bastl, M., Berger, U., and Magyar, D. (2020). The application of a neural network-based ragweed pollen forecast by the Ragweed Pollen Alarm System in the Pannonian biogeographical region. Aerobiologia, 36(2), 131–140. doi: 10.1007/s10453-019-09615-w

[7] Gaur, D., and Kumar Dubey, S. (2022). Development of activity recognition model using lstm-rnn deep learning algorithm. Journal of Information and Organizational Sciences, 46(2), 277–291.

[8] Goudarzi, G., Tahmasebi Birgani, Y., Assarehzadegan, M.-A., Neisi, A., Dastoorpoor, M., Sorooshian, A., and Yazdani, M. (2022). Prediction of airborne pollen concentrations by artificial neural network and their relationship with meteorological parameters and air pollutants. Journal of Environmental Health Science and Engineering, 20. doi: 10.1007/s40201-021-00773-z

[9] Khan, A. H., Li, S., and Luo, X. (2019). Obstacle avoidance and tracking control of redundant robotic manipulator: An rnn-based metaheuristic approach. IEEE transactions on industrial informatics, 16(7), 4670–4680. doi: 10.1109/tii.2019.2941916

[10] Lara, B., Rojo, J., Fernández-González, F., and Pérez-Badia, R. (2019). Prediction of airborne pollen concentrations for the plane tree as a tool for evaluating allergy risk in urban green areas. Landscape and Urban Planning, 189, 285–295. doi: 10.1016/j.landurbplan.2019.05.002

[11] Liu, X., Wu, D., Zewdie, G. K., Wijerante, L., Timms, C. I., Riley, A. T., Levetin, E., and Lary, D. J. (2017). Using machine learning to estimate atmospheric ambrosia pollen concentrations in tulsa, ok. Environmental Health Insights, 11. doi: 10.1177/1178630217699399

[12] Lo, F., Bitz, C. M., and Hess, J. J. (2021). Development of a random forest model for forecasting allergenic pollen in north america. Science of The Total Environment, 773, 145590. doi: https://doi.org/10.1016/j.scitotenv.2021.145590

[13] Maya-Manzano, J. M., Smith, M., Markey, E., Hourihane Clancy, J., Sodeau, J., and O'Connor, D. J. (2021). Recent developments in monitoring and modelling airborne pollen, a review. Grana, 60(1), 1–19. doi: 10.1080/00173134.2020.1769176

[14] Medsker, L., and Jain, L. C. (1999). *Recurrent neural networks: Design and applications*. CRC press.

[15] Méndez, M., Merayo, M. G., and Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. Artificial Intelligence Review, 1–36. doi: 10.1007/s10462-023-10424-4

[16] Miao, Y., Gowayyed, M., and Metze, F. (2015). Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 167–174. doi: 10.1109/asru.2015.7404790

[17] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533–536. doi: 10.7551/mitpress/1888.003.0013

[18] Shy, S., Tak, H., Feigelson, E. D., Timlin, J. D., and Babu, G. J. (2022). Incorporating measurement error in astronomical object classification. The Astronomical Journal, 164(1), 6. doi: 10.3847/1538-3881/ac6e64

[19] Smith, M., Cecchi, L., Skjøth, C., Karrer, G., and Šikoparija, B. (2013). Common ragweed: A threat to environmental health in europe. Environment International, 61, 115–126. doi: 10.1016/j.envint.2013.08.005

[20] Suanno, C., Aloisi, I., Fernandez Gonzalez, D., and Duca, S. (2021). Pollen forecasting and its relevance in pollen allergen avoidance. Environmental research, 200, 111150. doi: 10.1016/j.envres.2021.111150

[21] Werbos, P. (1990). Backpropagation through time: What it does and how to do it. Proceedings of the IEEE, 78(10), 1550–1560. doi: 10.1109/5.58337

[22] Zewdie, G. K., Lary, D. J., Levetin, E., and Garuma, G. F. (2019). Applying Deep Neural Networks and Ensemble Machine Learning Methods to Forecast Airborne Ambrosia Pollen. International Journal of Environmental Research and Public Health, 16(11). doi: 10.3390/ijerph16111992