ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2022.0362

AN EFFECTIVE DATA MINING TECHNIQUE TO IDENTIFY AND CLASSIFY RESPIRATORY DISEASES IN CHILDREN AND ADULTS

K. Nithyanandan and S. Prakasam

Department of Computer Science and Applications, Sri Chandrasekharendra Saraswathi Viswa Maha Vidyalaya, India

Abstract

The authors of this study have built an outstanding data mining model for the classification of respiratory issues in children and adults, which they have applied in their research. Deep learning ensembles are built by utilising support vector regression (SVR), long short-term memory neural networks (LSTMs), and a metaheuristic optimization (MHO) strategy that incorporates nonlinear learning in the DL ensemble. A collection of LSTMs with variable hidden layers and neurons is used to detect and exploit the underlying relationships in order to overcome the limitations of a single deep learning approach limited generalisation skills and robustness when faced with diverse input. The LSTM classification is then combined with a nonlinear-learning SVR and MHO to optimise the top-layer parameters. Nonlinear-learning meta-layer and LSTM classification. Finally, the final classification of the ensemble is provided by the fine-tuning meta-layer. Using data from six benchmark studies as well as energy consumption data sets, the proposed EDL is put to the test in two classification scenarios: ten-ahead and one-ahead classification.

Keywords:

Data Mining, Ensemble Deep Learning, Support Vector Regression, Metaheuristic Optimization Algorithm

1. INTRODUCTION

Since years, the biological characteristics, as well as the amount of time they spend outdoors, it is proposed that children are vulnerable to ambient particulate matter (PM) than adults (e.g., playing outside) [1]. Over 300 million children throughout the world are exposed to outdoor air pollution levels that are harmful to their health on a regular basis [2]. The level of toxins in China air has increased considerably over the past few decades, posing a particular threat to children, who are particularly vulnerable [3] [4].

Environmental pollutants in urban areas, such as particulate matter (PM), have been identified as one of the most significant airborne pollutants because they are composed of inhalable particles that can infiltrate the respiratory system and produce substantial health consequences [5]. These health concerns may be more prevalent in people who are predisposed to them (for example, children, people with chronic respiratory or cardiovascular disorders, and people with a weakened immune system). People who live in close proximity to air pollution sources are at an increased risk of developing serious health problems as a result of being exposed to high amounts of particulate matter (PM) [6]. The World Health Organization (WHO) [7] and other research [8] have found a connection between exposure to PM and respiratory morbidity (exacerbation of asthma symptoms, worsening of respiratory symptoms, and a rise in hospitalizations) [9,10] Adults with respiratory diseases and lung cancer are also more likely than the general population to die prematurely. It has been established that fine PM (PM0.1–PM2.5) is associated with cardiovascular illnesses [11].

Children are exposed to environmental risks such as air pollution at a higher rate than adults [12]. Inhalation of particulate matter (PM) has been shown to damage the development of children lungs, resulting in both permanent functional defects and a chronically slower rate of lung growth [13]-[15]. Chronic childhood disease is the most common chronic childhood disease, and it is the leading source of morbidity in children with chronic disorders. In recent years, an increasing number of children have developed asthma as a result of increased urbanisation, exposure to household allergens, and rising levels of air pollution. Identifying and reducing asthma triggers is an important stage in the course of the disease. Because asthmatic children are restricted in their daily physical activity, they are more likely to miss school than other children. There is an increase in asthma morbidity due to a variety of factors, including air pollution, which is a contributing factor. According to a recent study conducted in two Romanian cities, the prevalence of childhood asthma is increasing at an average annual rate of 8–11%.

Finding safe exposure levels or a threshold at which adverse health consequences do not occur has proven to be difficult. The effects of diverse chemical species of PM on human health are currently being investigated in a multidisciplinary way. As a result of the fact that children under the age of six are the most vulnerable, epidemiology studies suggest that greater research on respirable dust and its chemical speciation is required.

Long-term PM concentration reductions have been linked to improved health outcomes and lower national health expenditures. (For example, a reduction in the number of hospitalizations and accompanying therapy). It is also critical to identify the sources and quantities of primary particles, as well as trends in precursor gas emissions, in order to develop the most effective control strategy for reducing risks over time.

Deep learning ensembles are built by combining support vector regression (SVR), and a metaheuristic optimization strategy and long short-term memory neural networks (LSTMs) that incorporates nonlinear learning in the DL ensemble. A collection of LSTMs with variable hidden layers and neurons is used to detect and exploit the underlying relationships in order to overcome the limitations of a single deep learning approach limited generalisation skills and robustness when faced with diverse input.

2. PROPOSED METHOD

The results of the back propagation training procedure can differ depending on the number of epochs used in the training process. Consequently, it is possible to combine the outputs of LSTM models that have been trained with different epochs. After examining the relationships between the obtained outputs and the planned output values, it is possible to assign a weight value to each output in order to determine the overall anticipated output

value after calculating the weight value for each output. The final classification was achieved by utilising an ensemble of deep learning algorithms in the base layer composed of LSTMs trained with variable numbers of epochs and an SVR in the meta layer with input derived from the LSTM yields, as well as a combination of the two. According to the proposed model, there are three steps: the first is data preparation, followed by basic deep learning, and the third and final phase is meta-kernel learning.

2.1 DATA PREPROCESSING

When performing classification tasks, noise has a major impact on the data points. In statistical terms, noise is defined as a data point whose value differs significantly from the values of the other data points in the series. The failure to pay sufficient attention to noise reduction can be blamed for poor prediction outcomes.

Noise may be removed from classification data using a variety of techniques, one of which is the use of a second-order value as an indicator of the classification data. Data points are deemed noisy if their absolute values are four times larger in absolute value than the absolute median points immediately preceding and following them, respectively. This means that it is deemed noise if its value falls into one or more categories defined by the following parameters:

$$Y_i \geq 4 \times max\left\{\left|m_a\right|,\ \left|m_b\right|\right\} \tag{1}$$

where, $m_a = median\left(Y_i - 3,\ Y_i - 2,\ Y_i - 1\right)$ and $m_b = median\left(Y_i + 3,\ Y_i + 2,\ Y_i + 1\right)$.

2.2 LAG SELECTION

After this phase is completed, the training set will be generated in the following manner:

(a) Create an initial ensemble member consisting of N LSTM networks as a starting point. The greatest lag of the series is denoted by the parameter l_{max}, and in this case, N is a user-defined and larger than the l_{max} parameter.

(b) The lag parameter l_i is assigned to the ensemble LSTM i^{th} at random. l_{max} is the maximum number of times a random number between 1 and l_{max} can be generated on a regular basis.

(c) The architecture of each LSTM is composed of three layers: an input layer, a hidden layer, and an output layer. There are a total of three nodes in the input and hidden levels of the LSTM network, but only one node in the output layer of the network.

(d) In order to produce N training sets, one for each network in the ensemble, the lags assigned to each of the N networks should be used to create N training sets.

(e) Utilize the LSTM technique to train the whole base-layer network using the training data that was generated.

(f) It is recommended that each LSTM in the base layer is assessed using n data points from the testing set.

2.2.1 Base-Layer Deep Models:

LSTMs are a distinct type of RNN that uses memory blocks to replace the standard neurons in the hidden layers of the neural network. There are three gate components in memory blocks, known as input gates, output gates, and forget gates, which allow LSTMs to update and control the flow of information in the memory block. LSTM outputs can be followed in the implementation of cell state updates and cell state calculations below.

Writing, reading, and erasing from the cell memory state is used to deal with information drift in the LSTM. Three gates (input, forget, and output) are used to control each LSTM cell operation:

$$i_t = \sigma(W_{ix} x_t + W_{ip} p_{t-1} + W_{ie} e_{t-1} + b_i)$$

$$o_t = \sigma(W_{ox} x_t + W_{op} p_{t-1} + W_{oe} e_{t-1} + b_o)$$

$$f_t = \sigma(W_{fx} x_t + W_{fp} p_{t-1} + W_{fe} e_{t-1} + b_f)$$

$$\tilde{e}_t = g(W_{ex} x_t + W_{ep} p_{t-1} + b_f)$$

$$e_t = i_t \odot \tilde{e}_t + f_t + \odot e_{t-1};\ p_t = o_t \odot h(e_t);\ y_t = \varphi(W_{yp}.p_t + b_y) \tag{2}$$

where x_t-input, y_t-output, i_t-input gate output, o_t-output gate, f_t-forget gate, $\tilde{e}_t$-temporary cell state in a block, e_t-finishing cell state in a block, p_t-memory block output, σ-gate activation function, g-input activation function, h-output activation function, $\odot$-element-wise multiplication and φ-output activation function of LSTMs.

2.3 META-LAYER KERNEL MODEL

Ensemble deep learning for temporal data prediction using LSTMs and SVRs is being developed in order to improve classification accuracy. Stacking LSTM predictions are sent into a nonlinear-learning to achieve the classification in nonlinear-learning ensemble learning. SVR is offered as a nonlinear layer because of its capacity to solve complex regression issues, widespread application, and notable effectiveness in classification.

Complexity Factor C and Kernel Parameter σ are critical to EL performance, but so are other SVR meta-layer factors. In the next step, real-coded metaheuristic search is used to successfully handle this tuning issue. SVR fundamental concepts are introduced before the optimization problem formulation for classification is given. Eq.(3) can be used to describe the aforementioned operation as follows:

$$f(x_i) = W^T \Phi(x) + b;\ R[f] = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{N} L(x_i, y_i, f(x_i)) \tag{3}$$

2.4 METAHEURISTIC OPTIMIZATION

In the case of the SVR algorithm, for example, if the loss function ε-insensitive (ε), regularisation constant (C) and kernel parameter (λ) are big, the goal is to minimise the empirical risk. This is because hyper-parameters affect the predictive model classification accuracy. A decrease in the number of support vectors (SVs) will be caused by an increased. The proposed approach uses ensemble tuning to adjust EDL parameters more precisely, as previously stated.

2.5 ENSEMBLE PREDICTION

Meta-kernel is a term used to characterise the output of base-layer deep models together. Stacking employs a similar notion to

K-folds cross-validation to handle two important challenges in this context: building the second level of data Out-of-sample prediction is the first step. To capture unique locations where each model works best, the second goal must be achieved. By inferring the generalizer biases on the given learning set, the stacking procedure examines. Cross-validation was utilised to build the 'good' mix of variables.

$$f_{stack}(x) = \sum_{i=1}^{m} \alpha_i f_i(x) \quad (4)$$

where *a* - weight vector

To generate the output that will be used for final outcomes, the meta models can be trained on dataset *D*. stacking provides for the selection of several sub-learners, as well as a variety of alternative models to learn how to most effectively integrate their predictions. Meta-models combine the best sub-models into a single prediction; therefore, this process is referred to as blending.

The proposed EDL approach has two primary advantages over typical data-based ensembles. In the first case, unlabeled data are used to uncover additional evidence that can be analysed further. For example, stacking generalisation can be employed in the second phase to avoid overfitting in a single LSTM model, to some extent. The EDL model is therefore better suited for categorization data prediction due to these two advantages.

3. RESULTS AND DISCUSSIONS

There are four primary categories in the RespiRare® database structure for rare respiratory diseases: infantile lung disease (ILD), respiratory malformations, and other uncommon respiratory insufficiencies. Patients' information is entered into a secure database through a web interface using a secure Internet protocol. On a DMZ network, the interface is accessible via the https secure protocol. With a MySQL database, PHP scripts for accessing the database, and an Apache 2 webserver, it is a LAMP system. An encrypted IP to IP tunnel connects the database server housing MySQL to the web server, which is protected by a firewall. A central identification number is used to restrict access to the patient.

A total of eight centres linked with university hospitals make up the Reference Center, which is located throughout France. Each of them has a local network of connected centres that spans a French region. Patient data from each participating centre is solely accessible to that centre. This agreement must be made in writing before any data may be sent between the different research centres. General organisational norms and rules controlling data access are among the topics covered in a charter. Participants are invited to sign the charter of good use at the outset of their participation. Users are assigned a unique user name and password in order to access the database. Tracking systems can be used to go back to an earlier version of data if a user notices a change in the file content.

In September of that year, the RespiRare® database went live. At the close of 2008, the database was created. After completing their initial training, centres began to accept patients one at a time. To begin with, all patients under the age of 18 with ILD currently being monitored in the centres were included in the database after its certification by the database committee. Prospective entry of new patients was also made into the database. Because some patients had been tracked for an extended period of time when the database was created, it has information going back as far as 17 years. Centers' capacity to submit new patients for inclusion in the database is still sporadic, as this is an ongoing effort. We can't yet talk about prevalence or incidence because of a variety of biases, such as people who have been recovered or who have died.

A total of 217 patients have been proposed for the database as of right now. There were 12 patients whose ILD diagnoses were deemed invalid by the database committee due to a lack of evidence or data. 205 ILD patients have been added to the database as a result.

Table.1. Interstitial Lung Disease (ILD) In Children

Diagnoses	Patients (%)	Median Diagnosis Age (years)	Range
Surfactant disorders	17.6	0.3	(0–12.3)
Haemosiderosis	11.2	4.8	(0.7-14)
Granulomatous Disease	10.2	9.6	(5.6-14.4)
Alveolar Proteinosis	9.7	0.6	(0–13.6)
Exposure Diseases	4.4	10.6	(2.2-14.4)
Connective Disorders and Vascularitis	4.4	12.4	(0.1-15.7)
Metabolic Disorder	2.4	1.7	(0–4)
Langerhan cell histiocytosis	2.4	1.1	(0.5-7.3)
Infectious disease	2.0	0.2	(0–6.6)
Eosinophilic lung diseases	2.0	10.0	(5.1-16.9)
Lymphatic disorder	2.0	0	(0–14.5)
Others	4.4	1.3	(0-12.8)
Undiagnosed	27.3	0.7	(0-16.4)

Suboptimal respiratory function later in life is a result of early lung function problems (Table.2-Table.5) with severe socioeconomic repercussions. Many lung disorders become more severe when youngsters are exposed to airborne contaminants for an extended period of time. The children respiratory health consequences were consistent regardless of geography, economic status, or PM concentration levels in the countries studied in chosen articles.

Table.2. Comparison of Accuracy of Different Algorithms

Samples	SVR	LSTM	MHO	EDL
10	69.057	69.361	79.073	92.540
20	70.321	70.096	83.601	92.638
30	70.076	70.066	83.601	92.706
40	69.949	69.263	83.141	92.745
50	70.027	69.174	82.945	92.785

Even at exposure levels much below the existing national regulations, PM2.5, PM0.1, and soot may pose a risk to public health. With other air pollutants, PM has a considerable synergistic effect on children normal pulmonary function, particularly in indoor microenvironments.

Another drawback of the current strategy is the lack of knowledge on the relationship between PM chemistry, particle morphology, and the impacts on respiratory health. This suggests that a combination of chemical and physical features of aerosols, such as particle mass, particle number, or surface area, may lead to unfavourable health outcomes.

Table.3. Precision of Accuracy of Different Algorithms

Samples	SVR	LSTM	MHO	EDL
10	69.900	70.399	82.337	92.530
20	70.390	70.399	83.405	92.638
30	71.125	71.213	84.523	92.696
40	71.125	70.497	84.170	92.745
50	70.096	69.410	82.670	92.785

Table.4. Recall of Accuracy of Different Algorithms

Samples	SVR	LSTM	MHO	EDL
10	61.775	78.760	58.315	90.570
20	60.148	73.016	65.019	90.677
30	59.707	74.134	66.283	90.736
40	64.303	73.016	68.381	90.785
50	64.686	72.154	66.842	90.824

Table.5. F1-Score of Accuracy of Different Algorithms

Samples	SVR	LSTM	MHO	EDL
10	65.636	74.388	73.261	89.590
20	64.940	71.683	73.056	89.697
30	65.009	72.644	74.290	89.756
40	67.567	71.742	75.457	89.805
50	67.302	70.762	73.908	89.844

4. CONCLUSIONS

Classification utilising SVR, LSTMs and MSO is proposed in the EDL model. A set of LSTMs with varying hidden layers and neurons is used to identify and exploit the hidden relation in order to solve the limitation of limited generalisation capabilities and robustness of a single deep learning approach when faced with diverse input. Then, in order to optimise the top-layer parameters, LSTM classification is paired with a nonlinear-learning meta-layer composed of SVR and the MHO. The fine-tuning meta-layer provides the final classification of the ensemble. Data from six benchmark studies and energy consumption data sets are used to test the proposed EDL in a ten-ahead and one-ahead classification scenario.

REFERENCES

[1] D. Tomar and S. Agarwal, "A Survey on Data Mining Approaches for Healthcare", *International Journal of Bio-Science and Bio-Technology*, Vol. 5, No. 5, pp. 241-266, 2013.

[2] M. Saberi-Karimian and M. Ghayour-Mobarhan, "Potential Value and Impact of Data Mining and Machine Learning in Clinical Diagnostics", *Critical Reviews in Clinical Laboratory Sciences*, Vol. 58, No. 4, pp. 275-296, 2021.

[3] S. Vijiyarani and S. Sudha, "Disease Prediction in Data Mining Technique-A Survey", *International Journal of Computer Applications and Information Technology*, Vol. 2, No. 1, pp. 17-21, 2013.

[4] M.H.B.M. Adnan and F. Damanhoori, "A Survey on Utilization of Data Mining for Childhood Obesity Prediction", *Proceedings of Asia-Pacific Symposium on Information and Telecommunication Technologies*, pp. 1-6, 2010.

[5] D. Piedra, A. Ferrer and J. Gea, "Text Mining and Medicine: Usefulness in Respiratory Diseases", *Archivos De Bronconeumología (English Edition)*, Vol. 50, No. 3, pp. 113-119, 2014.

[6] H. Baek, M. Cho and S. Yoo, "Analysis of Length of Hospital Stay using Electronic Health Records: A Statistical and Data Mining Approach", *PloS One*, Vol. 13, No. 4, pp. 1-14, 2018.

[7] A.S. Monto and B.M. Ullman, "Acute Respiratory Illness in an American Community: the Tecumseh Study", *Jama*, Vol. 227, No. 2, pp. 164-169, 1974.

[8] P. Ahmad, S. Qamar and Q.A. Rizvi, "Techniques of Data Mining in Healthcare: A Review", *International Journal of Computer Applications*, Vol. 120, No. 15, pp. 1-16, 2015.

[9] M. Mozaffarinya, A.R. Shahriyari and G. Vahedi, "A Data-Mining Algorithm to Assess Key Factors in Asthma Diagnosis", *Revue Française d'Allergologie*, Vol. 59, No. 7, pp. 487-492, 2019.

[10] C.E. Wheelock, V.M. Goss and P.J. Skipp, "Application of Omics Technologies to Biomarker Discovery in Inflammatory Lung Diseases", *European Respiratory Journal*, Vol. 42, No. 3, pp. 802-825, 2013.

[11] Y. Feng, Y. Wang and H. Mao, "Artificial Intelligence and Machine Learning in Chronic Airway Diseases: Focus on Asthma and Chronic Obstructive Pulmonary Disease", *International Journal of Medical Sciences*, Vol. 18, No. 13, pp. 2871-2879, 2021.

[12] M. Anthimopoulos and S. Mougiakakou, "Lung Pattern Classification for Interstitial Lung Diseases using a Deep Convolutional Neural Network", *IEEE Transactions on Medical Imaging*, Vol. 35, No. 5, pp. 1207-1216, 2016.

[13] A. Srivastava, S. Jain and K. Kotecha, "Deep Learning based Respiratory Sound Analysis for Detection of Chronic Obstructive Pulmonary Disease", *PeerJ Computer Science*, Vol. 7, pp. 1-13, 2021.

[14] K.S. Alqudaihi, N. Aslam and M.S. Alshahrani, "Cough Sound Detection and Diagnosis using Artificial Intelligence Techniques: Challenges and Opportunities", *IEEE Access*, Vol. 9, pp. 102327-102344, 2021.

[15] E. Oostveen, D. MacLeod and F. Marchal, "The Forced Oscillation Technique in Clinical Practice: Methodology, Recommendations and Future Developments", *European Respiratory Journal*, Vol. 22, No. 6, pp. 1026-1041, 2003.