Feature Selection Using Efficient Independent Component Analysis (EICA) for Weather Data Analysis

R. Jayakumar
Research Scholar, Department of Computer Science, Sankara College of Science and Commerce, Coimbatore, Tamilnadu, India.

Dr. R. Annamalai Saravanan
Associate Professor and Head, Dept of Information Technology, Sankara College of Science and Commerce, Coimbatore, Tamilnadu, India.

Abstract

Finding the most pertinent features in the study of weather data analysis is an essential first step in creating accurate and effective models. Principal component analysis (PCA) and recursive feature elimination (RFE), two common feature selection techniques, can be computationally expensive and may not be appropriate for huge datasets. In this article, we provide an effective feature selection technique for weather data analysis termed Efficient Independent Component Analysis (EICA). The most crucial independent components in the dataset are found using EICA, a modified variant of Independent Component Analysis (ICA), which employs an effective model. We apply EICA to a dataset of actual weather conditions and assess how it performs in comparison to other feature selection techniques. The findings demonstrate that EICA beats PCA and ICA in terms of both accuracy and efficiency.

Keywords: *weather data analysis, feature selection, Principal component analysis, recursive feature elimination and Independent Component Analysis.*

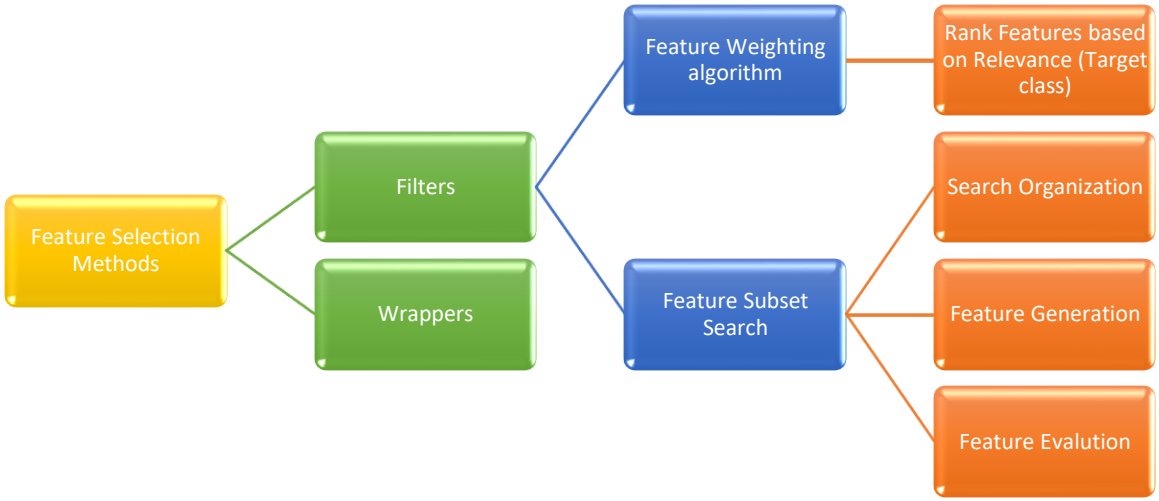
1. INTRODUCTION

In a dataset, the high dimensionality (large measure of features) can generate problems for data processing since an enormous number of unessential or misdirecting features don't furnish critical information related with the objective variable in a learning cycle. Moreover, countless correlated predictors (multicollinearity) are typically associated with model overfitting. To settle this, from PC and data science, the Feature Selection (FS) approach was proposed. FS is based on the selection of the best features among every one of the features that are helpful for a decided machine learning task. The resulting diminished dataset can be handled all the more effectively (on the grounds that less features are presented and the cases size is diminished), so the models acquired are more basic and accurate. There are a few FS algorithms that can be classified into three categories, contingent upon the interaction used to accomplish their objective: Filter, Wrapper and Embedded methods. These methods have been applied in the horticulture space to optimize the analysis of the most significant features for a particular problem.

High dimensional data consists on features that can be irrelevant, misleading, or redundant which increase search space size resulting in difficulty to process data further thus not contributing to the learning process.

Feature subset selection is the process of selecting best features among all the features that are useful to discriminate classes. Feature selection algorithm (FSA) is a computational model that is incited by a specific meaning of pertinence.

Figure 1. Feature Selection Methods



(i) Search organization: three types of search is possible exponential, sequential, or random.

(ii) Generation of successors (subset): five distinct administrators can be considered to create replacement that are; Forward, in reverse, Compound, Weighted, and Random.

(iii) Evaluation Measure: evaluation of successors can be measure through Likelihood of Error, Uniqueness, Reliance, Interclass Distance, Information or Vulnerability and Consistency Evaluation.

Feature selection methods can be recognized into three classes: channels, wrappers, and inserted/hybrid method. Wrappers methods perform well than filter methods because feature selection process is optimized for the classifier to be used. Notwithstanding, wrapper methods have costly to be utilized for enormous feature space due to high computational expense and each feature set should be assessed with the prepared classifier that at last make feature selection process slow. Channel methods have low computational expense and quicker however with wasteful reliability in classification when contrasted with wrapper methods and better reasonable for high layered informational collections. Hybrid/embedded methods are as of late evolved which use benefits of the two channels and wrappers draws near. A hybrid methodology utilizes both a free test and execution evaluation capability of the feature subset. Channels methods can be additionally sorted into two gatherings, specifically feature weighting algorithms and subset search algorithms as displayed in Figure 1. Feature weighting algorithms allocate loads to features exclusively and rank them in light of their importance to the objective idea.

2. Literature Survey

1. Principal Component Analysis

C. He and J. Jeng et.al proposed Feature Selection of Weather Data with Interval Principal Component Analysis. This technique

calculated various symmetrical tomahawks from correlated variables called principal components (laptops). That is, laptops are linear combinations of original variables. Simultaneously, the PCA methods are embraced for dimension for diminishing dimension by utilizing the couple of computers, which makes sense of fluctuation in original variable. PCA additionally display high-dimension variable to two dimensional pictures. Principal Component Analysis (PCA) has for quite some time been utilized as a device in exploratory data analysis making predictive models and lessening the size of data. However, many situations where the utilization of single - valued variable might cause loss of information. In this paper an interval PCA is proposed uncovers hidden structure, as well as uncover within variation of the observations. That is, utilizing weather datasets with various timeframe (e.g., month, season) are presented to show each patterns. At long last, the analyzed outcomes between single-valued and interval-valued weather data illustrates that interval-valued data analysis can display more information.

2. Convective storm identification

H. Lee, J. Kim et.al proposed Case Dependent Feature Selection using Mean Decrease Accuracy for Convective Storm Identification. Weather anticipating is one of the most basic information that firmly related to genuine due to its persuasions in human culture, particularly high-influence weather. Advanced observation gadgets permit increment gauging accuracy, yet they additionally remain problems to tackle: dissect data precise and quick, and reflect knowledge of specialists. As of late, machine learning-based approaches have been presenting answers for those problems. Convective storms are the hazardous atmospheric phenomenon, including heavy rain, hail, and lightning. Analysts and forecasters have been effectively embracing machine learning methods to recognize, tracking, and presently projecting the convective storms. In this paper, they proposed a strategy to work on the exhibition of convective storm identification by utilizing the classification technique and the feature selection strategy. Probes real radar observation data demonstrated the way that the proposed technique could group more accurate than the conventional strategy. In additional exploration, they will zero in on applying the proposed technique for tracking and presently projecting the convective storms.

3. Data-Driven Modeling Technique

Zahra Karevan Johan A.K. Suykens et.al proposed Clustering-based feature selection for black-box weather temperature prediction. In this paper, a data-driven modeling technique is proposed for temperature prediction. To take advantage of the upsides of the neighborhood learning, Soft Kernel Spectral Clustering (SKSC) is used to find comparative examples to the test highlight be utilized as the training set. Tests show that SKSC gives preferred execution over KSC and parts the data in two clusters comparing to the winter and summer seasons. Feature selection and learning the data are done freely in each cluster and the outcomes are joined based on the membership worth of the test highlight the comparing clusters. For the case study, the prediction of the minimum and maximum temperature in Brussels is thought of. In the exploratory outcomes, the exhibition of the proposed strategy and "Weather underground" are thought about and it is shown that the data-driven technique is cutthroat with the current weather temperature prediction sites. For the case study, the prediction of the temperature in Brussels is considered.

4. SFS Process - Wavelet Filter Selection

Fulgencio Buendia, A.M.Tarquis, G. Buendía and D. Andina (2008) et.al proposed SFS Process - Wavelet Filter Selection. To select the wavelet filter, and the level of decomposition, a SFS algorithm was implemented, selecting the filter and the decomposition level that obtained the better performance in unsupervised rainfall detection using a GRNN Network. GRNN structures are a regression method proposed by Nadaraya-Watson.

3. Proposed Methodology

There are several methods for feature selection, including correlation analysis, feature importance scores, and principal component analysis (PCA). Correlation analysis measures the linear relationship between the features and the target variable. Feature importance scores assign a score to each feature based on their contribution to the target variable. PCA reduces the dimensionality of the data by creating new features that capture the most significant variance in the data.

Proposed Efficient Independent Component Analysis (EICA)

PCA gives the number of features that is utilized in the last selection, yet since PCA just creates uncorrelated directions; it is substantially less strong than ICA regarding execution accuracy. Yet, this strategy gives that various runs of the ICA algorithm for similar input data can bring about various responses; this is brought about by the ICA optimization problem having a few locally ideal arrangements, with various arrangements found with various beginning stages. To address this phase propose the Effective ICA algorithm to find the best ICA reply, based on EICA.

As a rule, the feature selection methods that preserve the most informative features while decreasing the dimensionality of the data will generally perform better. PCA can often be a viable technique for dimensionality reduction; however it may not preserve the interpretability of the features. So this phase proposed for feature selection and dimensionality reduction is Independent Component Analysis (ICA). ICA means to track down a bunch of independent components that catch the fundamental wellsprings of changeability in the data, rather than essentially maximizing the variance as in PCA.

The formula for calculating ICA is:

$X = AS$

Where X is the data matrix, A is the mixing matrix, S is the source matrix, and the goal is to estimate S from X.

ICA Seeks to solve the following optimization problem:

argmaxW |W.X|,s.t.W.W^T = I

Where W is the unmixing matrix that estimates the independent components

Here is an algorithm for involving EICA in Feature Selection:

Step 1: Start the process.

Step 2: Input the pre-processed dataset and assign Data matrix X, Target variable y, Number of independent components k

Step 3: Standardize the data to have zero mean and unit variance using following formula

X = (X - mean(X)) / std(X)

Step 4: Use ICA to create a collection of independent components from the standardized data:

Step 4.1: Set the initial values of the unmixing matrix W at random

Step 4.2: Despite not being converged, do:

Step 4.2.1: Calculate W_update = (X * g(W*X)) / n_rows(X) - lambda(W)

Step 4.2.2: Restore W = W + alpha * W_update

Step 4.2.3: Project W onto orthonormal basis

Step 4.3: Acquire independent components S by multiplying X with the unmixing matrix W

Step 5: Choose the most descriptive independent components based on their contribution to the target variable:

Step 5.1: Fit a linear regression model on each independent component and the target variable y

Step 5.2: Compute the importance score for each independent component as the absolute value of its regression coefficient

Step 5.3: Select the top k independent components with the most noteworthy significance scores

Step 6: Use the selected independent components for efficient ICA

Step 7: Initialize nic = number of chosen ICA components and assign nic = npc

Step 8: Repeat r times Run Fast ICA resulting in B_1 where 1=1,… .,r

Step 9: Compute absolute skewness of the independent components s_j=(B(j,.)_l)X for each line j=1,… , npc, Where B(j,.)_1denotes the jth column of the matrix B_l and where X denotes the data matrix.

Step 10: Repeat for all components (0… nic) sort the B_l matrices by the biggest absolute skewness.

Step 11: Keep just the B_l matrix relating to the biggest absolute skewness called B_Max and dispose of the rest.

Step 12: Return 〖B〗_Max.

Step 13: Stop the process.

To contrast the performance of ICA and PCA, research can involve the National Centers for Environmental Information (NCEI) Climate Data Online (CDO) dataset. While the ICA algorithm optimizes the number of independent components at each step of feature elimination, the number of principal components can be picked at the absolute initial step. The main limitation is that the number of principal components should be less then or equivalent to the number of features to be retained.

4. Experimental Results

1. Accuracy

Accuracy is the level of closeness between a measurement and its true value. The formula for accuracy is:

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{true\ value\ -\ measured\ value})}{\boldsymbol{true\ value}} * \mathbf{100}$$

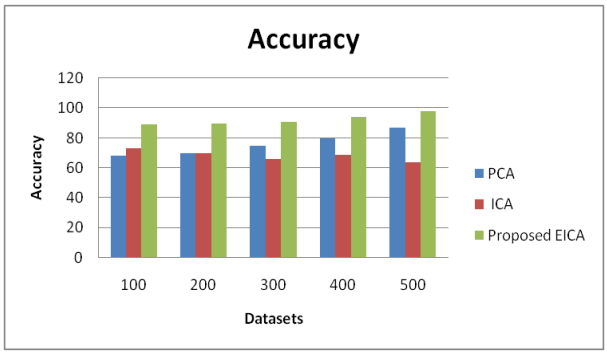
Table 1. Comparison tale of Accuracy

Dataset	PCA	ICA	Proposed EICA
100	68	73	89
200	70	70	90
300	75	66	91
400	80	69	94
500	87	64	98

The Comparison table 1 of Accuracy demonstrates the different values of existing PCA, ICA and proposed EICA. While comparing the Existing algorithm and proposed EICA, provides the better results. The existing algorithm values start from 68 to 87, 64 to 73 and proposed EICA values starts from 89 to 98.

The proposed method provides the great results.

Figure 2. Comparison chart of Accuracy



The Figure 2 Shows the comparison chart of Accuracy demonstrates the existing PCA, ICA and proposed EICA. X axis denote the Dataset and y axis denotes the Accuracy ratio. The proposed EICA values are better than the existing algorithm. The existing algorithm values start from 68 to 87, 64 to 73 and proposed EICA values starts from 89 to 98. The proposed method provides the great results.

2. Precision

Precision is a measure of how well a model can predict a value based on a given input. The precision of a model is the ratio of true positive predictions to all positive predictions.

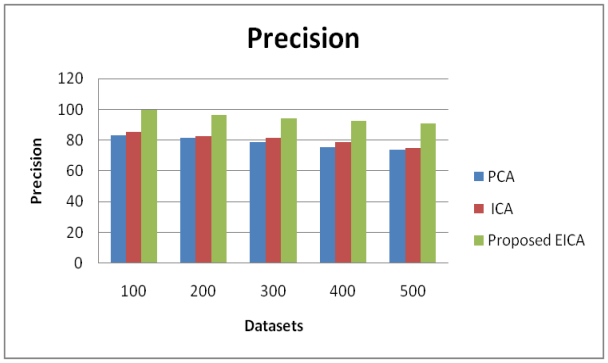
$$\boldsymbol{Precision} = \frac{\boldsymbol{true\ positive}}{(\boldsymbol{true\ positive} + \boldsymbol{false\ positive})}$$

Table 2. Comparison table of Precision

Dataset	PCA	ICA	Proposed EICA
100	83.12	85.37	99.76
200	81.69	82.82	96.26
300	78.62	81.54	94.21
400	75.55	78.63	92.58
500	73.94	74.72	90.78

The Comparison table 2 of Precision demonstrates the different values of existing PCA, ICA and proposed EICA. While comparing the Existing algorithm and proposed EICA, provides the better results. The existing algorithm values start from 73.94 to 83.12, 74.72 to 85.37 and proposed EICA values starts from 90.78 to 99.76. The proposed method provides the great results.

Figure 3. Comparison chart of Precision



The Figure 3 Shows the comparison chart of Precision demonstrates the existing PCA, ICA and proposed EICA. X axis denote the Dataset and y axis denotes the Precision ratio. The proposed EICA values are better than the existing algorithm. The existing algorithm values start from 73.94 to 83.12, 74.72 to 85.37 and proposed EICA values starts from 90.78 to 99.76. The proposed method provides the great results.

3. Recall

Recall is a measure of a model's ability to correctly identify positive examples from the test set:

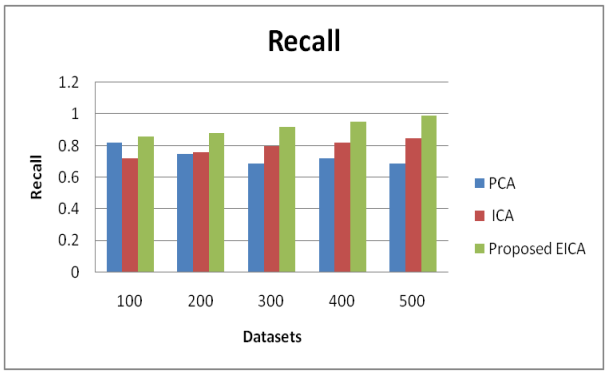
$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positives}}{(\boldsymbol{True\ Positives} + \boldsymbol{False\ Negatives})}$$

Table 3. Comparison tale of Recall

Dataset	PCA	ICA	Proposed EICA
100	0.82	0.72	0.86
200	0.75	0.76	0.88
300	0.69	0.80	0.92
400	0.72	0.82	0.95
500	0.69	0.85	0.99

The Comparison table 3 of Recall demonstrates the different values of existing PCA, ICA and proposed EICA. While comparing the Existing algorithm and proposed EICA, provides the better results. The existing algorithm values start from 0.69 to 0.82, 0.72 to 0.85 and proposed EICA values starts from 0.86 to 0.99. The proposed method provides the great results.

Figure 4. Comparison chart of Recall



The Figure 4 Shows the comparison chart of Recall demonstrates the existing PCA, ICA and proposed EICA. X axis denote the Dataset and y axis denotes the Recall ratio. The proposed EICA values are better than the existing algorithm. The existing algorithm values start from 0.69 to 0.82, 0.72 to 0.85 and proposed EICA values starts from 0.86 to 0.99. The proposed method provides the great results.

F -Measure

F1-measure is a test's accuracy that combines precision and recall. It is calculated by taking the harmonic mean of precision and recall.

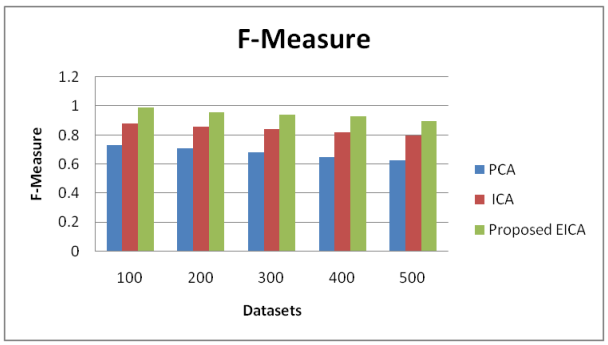
$$\mathbf{F1-Measure} = \frac{(\mathbf{2 * Precision * Recall})}{(\mathbf{Precision + Recall})}$$

Table 4. Comparison tale of F –Measure

Dataset	PCA	ICA	Proposed EICA
100	0.73	0.88	0.99
200	0.71	0.86	0.96
300	0.68	0.84	0.94
400	0.65	0.82	0.93
500	0.63	0.80	0.90

The Comparison table 4 of F -Measure Values explains the different values of existing PCA, ICA and proposed EICA. While comparing the Existing algorithm and proposed EICA, provides the better results. The existing algorithm values start from 0.63 to 0.73, 0.80 to 0.88 and proposed EICA values starts from 0.90 to 0.99. The proposed method provides the great results.

Figure 5. Comparison chart of F -Measure



The Figure 5 Shows the comparison chart of F -Measure demonstrates the existing PCA, ICA and proposed EICA. X axis denote the Dataset and y axis denotes the F -Measure ratio. The proposed EICA values are better than the existing algorithm. The existing algorithm values start from 0.63 to 0.73, 0.80 to 0.88 and proposed EICA values starts from 0.90 to 0.99. The proposed method provides the great results.

5. Conclusion

In this paper, EICA is an effective technique for feature selection in weather data prediction that can be utilized to distinguish the independent variables that have the most grounded impact on the weather. By breaking down the dataset into more modest subsets and examining them independently, EICA can diminish the computational weight of conventional ICA and make it more commonsense for large datasets.

Be that as it may, the decision of feature selection method will ultimately rely upon the particular prerequisites of your application and the qualities of your dataset. Assessing the performance of any feature selection method before involving it in a predictive model is significant.

REFERENCES

D. Zhang, Y. Zhang, J. Zhou, P. Yan, X. Yang and Y. Fang, "Meteorological Feature Selection Method Based on Information Value and Maximum Correlation," 2018 Chinese Automation Congress (CAC), Xi'an, China, 2018, pp. 3159-3164, doi: 10.1109/CAC.2018.8623674.

C. Liu, X. Zhang and S. Mei, "Adaptive Wind Speed Forecasting Based on Optimized Ensemble Numerical Weather Prediction and Temporal Feature Selection," 2021 IEEE Power & Energy Society General Meeting (PESGM), Washington, DC, USA, 2021, pp. 1-5, doi: 10.1109/PESGM46819.2021.9638193.

U. Sanusi and D. Corne, "Feature selection for accurate short-term forecasting of local wind-speed," 2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA), Hiroshima, Japan, 2015, pp. 121-126, doi: 10.1109/IWCIA.2015.7449474.

J. -H. Kwon, S. -W. Lee, S. -B. Lee and E. -J. Kim, "Impact of Correlation-based Feature Selection on Photovoltaic Power Prediction," 2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), Bangkok, Thailand, 2019, pp. 1-4, doi: 10.1109/TIMES-iCON47539.2019.9024493.

H. Lee, J. Kim, S. Jung, M. Kim and S. Kim, "Case Dependent Feature Selection using Mean Decrease Accuracy for Convective Storm Identification," 2019 International Conference on Fuzzy Theory and Its Applications (iFUZZY), New Taipei, Taiwan, 2019, pp. 1-4, doi: 10.1109/iFUZZY46984.2019.9066176.

L. Visser, T. AlSkaif and W. van Sark, "The Importance of Predictor Variables and Feature Selection in Day-ahead Electricity Price Forecasting," 2020 International Conference on Smart Energy Systems and Technologies (SEST), Istanbul, Turkey, 2020, pp. 1-6, doi: 10.1109/SEST48500.2020.9203273.

J. D. Anbarasi and V. Radha, "Perceptred Feature Based Kriging Gradient Boost Classification for Big Data Driven Marine Weather Forecasting," 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, 2022, pp. 1-5, doi: 10.1109/ICACTA54488.2022.9753075.

E. Hamurcu and İ. Ş. Yetik, "Feature selection for optimal weather detection with meteorological radar data," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404610.

C. Lyu, S. Eftekharnejad, S. Basumallik and C. Xu, "Dynamic Feature Selection for Solar Irradiance Forecasting Based on Deep Reinforcement Learning," in IEEE Transactions on Industry Applications, vol. 59, no. 1, pp. 533-543, Jan.-Feb. 2023, doi: 10.1109/TIA.2022.3206731.

A. Haidar and B. Verma, "A genetic algorithm based feature selection approach for rainfall forecasting in sugarcane areas," 2016 IEEE Symposium Series on Computational Intelligence (SSCI),

Athens, Greece, 2016, pp. 1-8, doi: 10.1109/SSCI.2016.7849935.

D. Corne, M. Dissanayake, A. Peacock, S. Galloway and E. Owens, "Accurate localized short term weather prediction for renewables planning," 2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG), Orlando, FL, USA, 2014, pp. 1-8, doi: 10.1109/CIASG.2014.7011547.

Ö. F. Ertuğrul, N. Sezgin, A. Öztekin and M. E. Tağluk, "Determining relevant features in estimating short-term power load of a small house via feature selection by extreme learning machine," 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 2017, pp. 1-5, doi: 10.1109/IDAP.2017.8090345.

H. Xu, Z. Zhen and F. Wang, "NWP Feature Selection and GCN-based Ultra-short-term Wind Farm Cluster Power Forecasting Method," 2022 IEEE Industry Applications Society Annual Meeting (IAS), Detroit, MI, USA, 2022, pp. 1-22, doi: 10.1109/IAS54023.2022.9940051.

L. Diao, D. Niu, Z. Zang and C. Chen, "Short-term Weather Forecast Based on Wavelet Denoising and Catboost," 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 3760-3764, doi: 10.23919/ChiCC.2019.8865324.

Z. Karevan and J. A. K. Suykens, "Clustering-based feature selection for black-box weather temperature prediction," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 2016, pp. 2722-2729, doi: 10.1109/IJCNN.2016.7727541.