# A Comparative Study of Winter Fog Prediction in Haryana using Machine Learning Models

Deepak Prakash Londhe[1], Dr. Pravin Kumar[2], Dr. Suryanshu Choudhary[3]

[1]Research Scholar, Department of Physics, P. K. University, Shivpuri, M.P.
[2]Assistant Professor, Department of Physics, P. K. University, Shivpuri, M.P.
[3]Associate Professor,Department of Physics, P. K. University, Shivpuri, M.P.

## ABSTRACT

**Fog forecasts in Haryana, India, from 1973 to 2022, with a concentration on the winter months (November–February), are the subject of this research. Observations of visibility culled from METAR recordings are used into the investigation. With a focus on 3-hour fog prediction, we investigated several time series and machine learning models for regression and classification tasks. In comparison to more conventional approaches, the research shows that advanced sequence models, especially LSTM and GRU versions, significantly improve fog forecasting accuracy. This research sheds light on how to best use machine learning algorithms to forecast local weather.**

**Keywords: Winter, Fog, Machine Learning, Regression, Meteorological**

## INTRODUCTION

Transportation, farming, and public safety are all severely impacted by the thick and low-lying winter fog. It is very important to be able to anticipate when it will happen with a high degree of precision and advance notice in areas like Haryana, India, where fog episodes are often and severe. Because of their superior pattern-finding capabilities in massive datasets, machine learning (ML) models become invaluable in this context. While physically and empirically based weather forecast models have their uses, they aren't always up to the task of capturing the complex and non-linear interactions between meteorological components that cause fog to occur. Machine learning provides an attractive alternative by analyzing massive information and spotting small patterns.

Many variables, such as air pressure, humidity, wind speed, and temperature, contribute to the production of fog. It may be difficult to anticipate how these components would interact within the framework of winter fog, especially in areas like Haryana. In the winter, especially between November and February, the northern Indian state of Haryana is beset by heavy fog. During this time, visibility is low, which is dangerous for anyone traveling by car, train, or plane. Typical weather patterns in this area of the nation cause thick fog to roll in around this time of year. Forecasting the occurrence, severity, and length of fog is a formidable task in and of itself.

Statistical approaches and physical models based on weather historical data and meteorological equations have traditionally been used for fog prediction. The complicated, non-linear interactions between the several components that lead to fog formation are sometimes difficult for these strategies to account for, notwithstanding their limited effectiveness. Here is where the influence of machine learning models has started to grow. Algorithms that can learn from data and generate predictions are at the heart of machine learning, a branch of artificial intelligence. When it comes to fog prediction, ML models can sift through mountains of meteorological data from the past to find connections and patterns that conventional approaches may miss.

When it comes to winter fog prediction, ML models' capacity to manage big datasets with various factors is a major plus. For instance, in order to forecast the probability of fog on a certain day, a model may examine data on humidity, temperature, wind speed, and atmospheric pressure in addition to past occurrences of fog. Using this data, the model may be trained to gradually become more accurate. Particularly in time series forecasting applications, complex ML models like Gated Recurrent Unit (GRU) networks and Long Short-Term Memory (LSTM) have shown great potential in recent years. Because fog is fundamentally a time-dependent phenomena, these models excel at forecasting it because they are built to capture temporal dependencies in data. To address the shortcomings of classic RNNs in capturing long-term dependencies in sequential data, two variants of recurrent neural networks (RNNs) were developed: long short-term memory (LSTM) and generalized recurrent unit (GRU) models. Since the order and timing of occurrences are so important in weather prediction, they are ideal for this kind of work. These models may be taught the patterns that cause fog to occur using meteorological data that has been collected in the past, which is useful for fog prediction. Following training, they might be used to anticipate future fog outbreaks, perhaps offering precise forecasts up to several hours in advance. Problems do arise when trying to use ML models for fog prediction. The accessibility and accuracy of the data is one of the main obstacles. Predicting fog accurately calls on long-term, high-resolution data on a wide range of meteorological variables. The accuracy of the ML models is sometimes compromised by missing or noisy data. In addition, high-quality training data is crucial for ML model performance. The accuracy of the model's predictions is dependent on how well the training data captures the actual circumstances of fog formation.

The interpretability of machine learning models is another obstacle. Although these models are capable of producing very precise results, they are often "black boxes," making it impossible to trace their reasoning behind their forecasts. In the field of weather prediction, where knowledge about the causes of certain weather phenomena is crucial, this might be a major limitation. Fortunately, this problem is starting to get some attention from the explainable AI (XAI) community, which is developing methods and tools to make ML models easier to understand and work with.

## REVIEW OF LITERATURE

Castillo-Botón, C. et al., (2022) Fog often accompanies atmospheric low-visibility occurrences. Every year, accidents and traffic issues are caused by extreme low-visibility occurrences, which have a profound impact on air and ground transportation, airports, and motor-road infrastructure. Numerous fog generation and low-visibility prediction issues have been effectively addressed using Machine Learning (ML) methods. Both the kind of ML technique utilized and the quality of the predictions obtained are affected by whether the related issue is stated as a regression or a classification challenge. Predicting low-visibility occurrences as a combination of regression and classification issues is the focus of this paper's comprehensive examination. Using a standard comparison framework, we analyze and rank the efficacy of several ML methods across all issue types.

Dewi, Ristiana et al., (2020)Airport operations are impacted by several meteorological phenomena, including fog. A decrease in vision may have an effect on aircraft operations such as taxiing, take-off, and landing. Consequently, in order to ensure the safety of flights, fog forecast is essential. The atmosphere's intricate and chaotic process is the greatest obstacle to accurate weather prediction. This study employs AI in an effort to foretell when fog may form at Wamena Airport. Forecasting model construction based on an hourly synoptic data collection covering the period from 2015 to 2018 (May).

Dry and wet ball temperatures, dew point, relative humidity, cloud cover, wind direction and speed, visibility, and current meteorological conditions from the last six hours are some of the elements used to forecast the likelihood of fog or lack thereof. For this grid search, we used five different algorithms to fine-tune their parameters: Deep Learning, Generalized Linear Model, Distributed Random Forest, GLM, and XRT. The Stacked Ensemble (SE) ensemble model, which achieves an accuracy of more than 90%, is the most effective at predicting fog one to three hours later.

Durán-Rosal, Antonio et al., (2018) This research presents a solution to the real-world issue of fog event categorization using meteorological input variables: new artificial neural networks trained evolutionary-style with three distinct basic functions: sigmoidal, product, and radial. In order to train an artificial neural network to identify fog occurrences at Valladolid Airport (Spain), a Multiobjective Evolutionary Algorithm is specifically studied. The goal is to generate a binary classification model. The generated evolutionary neural models are guided by two performance metrics: classic accuracy and minimal sensitivity.

Bartoková, Ivana et al., (2015) The coastal desert area of Dubai is nowcasting fog occurrences using a technology based on decision trees. The examined region depended on automated weather sensors' high-frequency readings to assess useful trends. According to a study published in Boundary-Layer Meteorol in 2012 by Bartok et al., induced decision trees outperform the combined WRF and PAFOG fog models in the first six hours of forecasting. Incorporating the findings of the linked numerical fog forecasting models into the decision tree's training database improved the outcomes even more. Results from Gilbert's skill score (0.69), detection probability (0.88), and false alarm ratio (0.19) were all significantly improved after this therapy. Using the newly-developed machine-learning method for the first six forecast hours yields the best fog prediction in the Dubai area, according to these data. On the other hand, coupled numerical models are the way to go for forecast durations longer than six hours.

## RESEARCH METHODOLOGY

### Study Area and Datasets

Our research focuses on the state of Haryana. Metar recordings are used to get the visibility observations. The reliability and trustworthiness of these METAR observations are enhanced since they are measured in-situ. The observations are used during the winter months (November to February) from 1973 to 2022. Perceived visibility is the metric for fog. On the other hand, reduced vision is not limited to fog; dust storms, strong rains, high levels of air pollution, etc., may also cause this. Relevant climatic variables, such as relative humidity, etc., become crucial variables in such situations alongside visibility.

**Evaluation Metrics**
While assessing the regression models, we make use of the following measures:

- MAE (Mean Absolute Error): The mean absolute deviation (MAE) between the actual and anticipated values is computed.
- RMSE (Root Mean Squared Error): Subtracting the square of the difference between the actual and projected numbers yields this value.

**The categorization problem's metrics are as follows:**

- Accuracy: If the actual and predicted classes are same, then the prediction is correct. We take two metrics into account: the overall accuracy and the average accuracy across all classes.
- (Macro) F1-score: A class's average F1-score is known as its macro F-score.

**Classification and Regression Models**
Multiple regression and classification models were tested. Ensemble techniques, sequence models, time series approaches, and conventional machine learning models are all part of this category. The following machine learning algorithms are included: LR, DT, MLP, NN, SVM, NB, and KN.

The LSTM and GRU models, both of which are recurrent neural network (RNN) implementations, are examples of sequence models. We also tried both stack and bidirected versions of these models. Using MAE as a monitoring measure, the hyper-parameters of machine learning, ensemble, and sequence models were fine-tuned on the validation set using a grid search strategy.

The Akaike Information Criterion (AIC), a metric for predicting error, was used as the optimization metric for ARIMA models. For two hidden states, HMM performs best on the validation set.

## RESULTS AND DISCUSSION

**Multiclass Classification Results**
Table 1 shows that after using a hybrid class balancing method that incorporates NCR and SMOTE, Gradient Boosting is the highest performing model for the 5-class classification issue with a 3-hour lead time for Haryana. The acronym GBC-NCR-SMT stands for this model.

**Table 1: Multi-class results for 3 hr lead time at Haryana**

| Model Name | Accuracy | Macro-F1 | Macro-F1 of Fog |
|---|---|---|---|
| Persistence | 0.73 | 0.46 | 0.29 |
| Gradient BoostingClassifier - Neighborhood Cleaning Rule-SMOTE | 0.81 | 0.50 | 0.40 |
| Gradient Boosting Classifier | 0.81 | 0.46 | 0.35 |
| Gradient Boosting Classifier -SMOTE | 0.78 | 0.47 | 0.35 |
| Gradient Boosting Classifier -OVR | 0.78 | 0.49 | 0.37 |
| Logistic Regression | 0.72 | 0.40 | 0.28 |
| Logistic Regression -OVR | 0.72 | 0.40 | 0.28 |
| Random Forest -OVR | 0.70 | 0.42 | 0.34 |
| Random Forest | 0.60 | 0.45 | 0.31 |
| DecisionTree -OVR | 0.61 | 0.40 | 0.29 |
| DecisionTree | 0.52 | 0.35 | 0.26 |

The table also shows the average F1-score of the 4 fog classes.

Table 2 displays the confusion matrix that corresponds to it. As anticipated, the majority of the mistakes occur within nearby fog classes.

**Table 2: Multi Class Confusion Matrix for 3-hr lead time at Haryana**

<table>
<tr><th colspan="8">Predicted</th></tr>
<tr><td rowspan="7">Actual</td><td></td><th>Very Dense Fog</th><th>Dense Fog</th><th>Moderate Fog</th><th>Shallow Fog</th><th>No Fog</th><th>Total</th></tr>
<tr><td>Very Dense Fog</td><td>257</td><td>99</td><td>108</td><td>59</td><td>18</td><td>541</td></tr>
<tr><td>Dense Fog</td><td>109</td><td>129</td><td>156</td><td>100</td><td>64</td><td>558</td></tr>
<tr><td>Moderate Fog</td><td>34</td><td>117</td><td>353</td><td>198</td><td>197</td><td>899</td></tr>
<tr><td>Shallow Fog</td><td>22</td><td>65</td><td>627</td><td>1268</td><td>1703</td><td>3685</td></tr>
<tr><td>No Fog</td><td>4</td><td>11</td><td>109</td><td>732</td><td>15523</td><td>16379</td></tr>
<tr><td>Total</td><td>426</td><td>421</td><td>1353</td><td>2357</td><td>17505</td><td>22062</td></tr>
</table>

**Regression Results**

The experimental findings for several regression models using Haryana data with a 3-hour lead time are summarized in Table 3.

**Table 3: Regression results for 3-hr lead time at Haryana**

| Model Name | RMSE (km) | Fog RMSE | MAE (km) |
|---|---|---|---|
| Persistence | 0.69 | 0.58 | 0.51 |
| Univariate Vanilla Long Short-TermMemory | 0.22 | 0.14 | 0.12 |
| Stacked Long Short-TermMemory | 0.22 | 0.19 | 0.09 |
| Bidirectional Long Short TermMemory | 0.19 | 0.16 | 0.15 |
| Stacked Gated Recurrent Unit | 0.19 | 0.17 | 0.11 |
| Univariate Vanilla Gated Recurrent Unit | 0.25 | 0.17 | 0.15 |
| Bidirectional Gated Recurrent Unit | 0.25 | 0.16 | 0.11 |
| LightGBM Custom Scoring | 0.50 | 0.31 | 0.37 |
| PloyLogistic Regression | 0.52 | 0.36 | 0.29 |
| LightGBM | 0.50 | 0.37 | 0.31 |
| Multi-layer Perceptron | 0.51 | 0.45 | 0.36 |
| Logistic Regression | 0.49 | 0.45 | 0.35 |
| GBR | 0.46 | 0.44 | 0.36 |
| Random Forest | 0.58 | 0.47 | 0.38 |
| DecisionTree | 0.59 | 0.51 | 0.39 |
| Hidden Markov Model | 0.52 | 1.05 | 0.98 |
| Auto-regressive Integrated Moving Average | 1.45 | 2.15 | 1.22 |

In the table, you can see the regression results for several models' performance in forecasting fog in Haryana with a 3-hour advance time. Top performances among the models studied were Stacked GRU and Bidirectional LSTM. Bidirectional LSTM had the lowest RMSE of 0.19 km and an MAE of 0.15 km. With a little lower MAE of 0.11 km and an RMSE of 0.19 km, the Stacked GRU was right behind. Stacked LSTM and Univariate Vanilla LSTM both did well in fog data modeling; their RMSE values were 0.22 km and 0.22 km, respectively.

On the other hand, more conventional models such as ARIMA and HMM had a very hard time. ARIMA had the worst performance in this regard, with an RMSE of 1.45 km and an MAE of 1.22 km. With RMSE values ranging from 0.46 to 0.52 km and MAE values ranging from 0.29 to 0.37 km, LightGBM, MLP, and GBR exhibited reasonable performance among the machine learning models. The LSTM and GRU versions outperformed these models, nevertheless. Almost all data-driven models beat the baseline Persistence model, which had an RMSE of 0.69 km and an MAE of 0.51 km; this shows the benefit of using more advanced methodologies.

## CONCLUSION

Research on the use of machine learning models for winter fog prediction in Haryana shows how promising sophisticated computational methods are for weather forecasting. This study demonstrated that sequence models such as LSTM and GRU are more accurate and reliable than conventional forecasting approaches by using data from METAR records and a specialized weather monitoring station at IIT Kanpur. Fog formation involves complicated, time-dependent patterns, which these models successfully handle for a 3-hour lead time forecast. Fog is still a big deal in the area, especially in the winter. By adding these machine learning models to operational weather forecasting systems, we can hopefully lessen the social and economic toll that fog takes by giving people early and accurate warnings.

## REFERENCES

[1]. C. Castillo-Botón, D. Casillas-Perez, C. Casanova, S. Ghimire, E. Cerro Prada, P. A. Gutiérrez, R. Deo, and S. Salcedo-Sanz, "Machine learning regression and classification methods for fog events prediction," Atmospheric Research, vol. 272, no. 37, p. 106157, 2022, doi: 10.1016/j.atmosres.2022.106157.

[2]. Y. Zhang, W. Yangjun, Y. Zhu, L. Yang, L. Ge, and C. Luo, "Visibility Prediction Based on Machine Learning Algorithms," Atmosphere, vol. 13, no. 7, p. 1125, 2022, doi: 10.3390/atmos13071125.

[3]. S. Cornejo-Bueno, D. Casillas-Pérez, L. Cornejo Bueno, M. Chidean, A. Caamaño, E. Cerro Prada, C. Casanova, and S. Salcedo-Sanz, "Statistical Analysis and Machine Learning Prediction of Fog-Caused Low-Visibility Events at A-8 Motor-Road in Spain," Atmosphere, vol. 12, no. 6, p. 679, 2021, doi: 10.3390/atmos12060679.

[4]. R. Dewi, P. Prajitno, and H. Harsa, "Fog prediction using artificial intelligence: A case study in Wamena Airport," Journal of Physics: Conference Series, vol. 1528, no. 1, p. 012021, 2020, doi: 10.1088/1742-6596/1528/1/012021.

[5]. A. Durán-Rosal, J. C. Fernández, C. Casanova, J. Sanz-Justo, S. Salcedo-Sanz, and C. Martínez, "Efficient Fog Prediction with Multi-objective Evolutionary Neural Networks," Applied Soft Computing, vol. 70, no. 4, 2018, doi: 10.1016/j.asoc.2018.05.035.

[6]. A. Roman-Cascón, G. J. Steeneveld, C. Yagüe, M. Sastre, J. A. Arrillaga, and G. Maqueda, "Forecasting radiation fog at climatologically contrasting sites: Evaluation of statistical methods and WRF," Quarterly Journal of the Royal Meteorological Society, vol. 142, no. 695, pp. 1048-1063, 2016.

[7]. I. Bartoková, A. Bott, J. Bartok, and M. Gera, "Fog Prediction for Road Traffic Safety in a Coastal Desert Region: Improvement of Nowcasting Skills by the Machine-Learning Approach," Boundary-Layer Meteorology, vol. 157, no. 3, 2015, doi: 10.1007/s10546-015-0069-x.

[8]. G. J. Steeneveld, R. J. Ronda, and A. A. M. Holtslag, "The challenge of forecasting the onset and development of radiation fog using mesoscale atmospheric models," Boundary-Layer Meteorology, vol. 154, no. 2, pp. 265-289, 2015.

[9]. D. Dutta and S. Chaudhuri, "Nowcasting visibility during wintertime fog over the airport of a metropolis of India: Decision tree algorithm and artificial neural network approach," Natural Hazards, vol. 75, no. 2, pp. 1349-1368, 2015.

[10]. T. Boneh, G. T. Weymouth, P. Newham, R. Potts, J. Bally, A. E. Nicholson, and K. B. Korb, "Fog forecasting for Melbourne airport using a Bayesian decision network," Weather and Forecasting, vol. 30, no. 5, pp. 1218-1233, 2015.

[11]. B. Zhou and J. Du, "Fog prediction from a multimodel mesoscale ensemble prediction system," Weather and Forecasting, vol. 25, no. 1, pp. 303-322, 2010.

[12]. I. Gultepe, G. Pearson, J. A. Milbrandt, B. Hansen, S. Platnick, P. Taylor, M. Gordon, J. P. Oakley, and S. G. Cober, "The fog remote sensing and modeling field project," Bulletin of the American Meteorological Society, vol. 90, no. 3, pp. 341-360, 2009.

[13]. C.-H. Bang, J. Lee, and S.-Y. Hong, "Predictability Experiments of Fog and Visibility in Local Airports over Korea using the WRF Model," Journal of Korean Society for Atmospheric Environment, vol. 24, no. 2, pp. 92-101, 2008.

[14]. D. Fabbian, R. De Dear, and S. Lellyett, "Application of artificial neural network forecasts to predict fog at Canberra international airport," Weather and Forecasting, vol. 22, no. 2, pp. 372-381, 2007.

[15]. G. M. Forthun, M. B. Johnson, W. G. Schmitz, J. Blume, and R. J. Caldwell, "Trends in fog frequency and duration in the southeast United States," Physical Geography, vol. 27, no. 3, pp. 206-222, 2006.