MAUSAM, **77**, 1(January 2026), 203-218

MAUSAM

DOI: https://doi.org/10.54302/q6zxvb08
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No.551.509.316:004.431.2(540)

Development of machine learning models for the nearcasting of the duration of low-visibility events in the Indo-gangetic plains regions of India

ANAND SHANKAR[1,2]*, BIKASH CHANDRA SAHANA[1], SUNNY CHUG[3]

[1]*Department of Electronics & Communication Engineering National Institute of Technology, Patna, India*

[2]*India Meteorological Department, Ministry of Earth Sciences, Govt. of India, Patna, India*

[3]*India Meteorological Department, Ministry of Earth Sciences, Govt. of India, Kolkata, India*

(*Received 16 May 2024, Accepted 13 August 2024*)

***Corresponding author's email: anand.shankar@imd.gov.in**

सार – किसी कैलेंडर दिवस में कम-दृश्यता घटनाओं की अवधि का पूर्वानुमान लगाना एक कठिन प्रक्रिया है, क्योंकि इन घटनाओं के आरंभ और समाप्ति की प्रक्रियाएँ जटिल तथा अव्यवस्थित (chaotic) होती हैं। फिर भी, यह जानकारी हवाई अड्डा सेवाओं के संचालन (विमानों की समय-सारणी, हवाई अड्डों का अनुकूल संचालन) तथा विभिन्न गतिविधियों (यात्रा, पर्यटन, कृषि आदि) की योजना के लिए अत्यंत उपयोगी है। यह अध्ययन 1500 UTC (समन्वित सार्वभौमिक समय) की प्रारंभिक परिस्थितियों के आधार पर किसी दिन में कम-दृश्यता घटनाओं–कोहरा (सतही दृश्यता<1000 मीटर) और घना कोहरा (सतही दृश्यता<200 मीटर)–की अवधि का सटीक निकट-पूर्वानुमान करने हेतु आधारभूत मशीन लर्निंग (ML) मॉडलों [लाइट ग्रेडिएंट बूस्टिंग मशीन (LightGBM), रैंडम फॉरेस्ट (RF) और सपोर्ट वेक्टर रिग्रेशन (SVR)] के सर्वोत्तम संयोजन से एक गतिशील भारित एन्सेम्बल मॉडल विकसित करने का प्रयास करता है।सतही मौसमीय मानदंडों–वायु तापमान, ओसांक तापमान, सापेक्ष आर्द्रता, पवन (प्रत्येक 3 घंटे), वर्षा (दैनिक) और धूप अवधि (दैनिक)–तथा ऊपरी वायुमंडलीय मानदंडों–1000, 925 और 850 hPa स्तरों पर पवन, तापमान और सापेक्ष आर्द्रता (प्रत्येक 3 घंटे)–को कम-दृश्यता घटनाओं की अवधि के सटीक निकट-पूर्वानुमान हेतु सर्वोत्तम व्याख्यात्मक कारकों के चयन के लिए सम्मिलित किया गया। अंतिम मॉडल व्याख्यात्मक चर चुनने हेतु पीयरसन सहसंबंध गुणांक और स्पीयरमैन रैंक सहसंबंध गुणांक का उपयोग किया गया। डाटासेट्स का पर्यवेक्षित ML एल्गोरिद्मों द्वारा प्रशिक्षण, परीक्षण, मॉडलिंग और क्रॉस-वैलिडेशन के विभिन्न चरणों में विस्तृत विश्लेषण किया गया। सभी सर्वोत्तम संयोजन मॉडलों की सटीकता का मूल्यांकन और तुलना MAE (मीन एब्सोल्यूट एरर), RMSE और R^2 (निर्धारण गुणांक) के माध्यम से की गई। R^2 के आधार पर यह पाया गया कि प्रस्तावित गतिशील भारित एन्सेम्बल मॉडल ने सर्वश्रेष्ठ पूर्वानुमान सटीकता प्रदर्शित की–एक दिन के अग्रिम समय (lead time) के लिए कोहरे और घने कोहरे की अवधि हेतु क्रमशः 0.89 और 0.88। यह LightGBM (0.79), RF (0.78) और SVR (0.76) की तुलना में बेहतर है। अतः यह अध्ययन कोहरा-प्रवण इंडो-गंगेटिक मैदान (IGP) में हवाई अड्डा समय-सारणी के स्वचालन और संचालन के अनुकूलन में मशीन लर्निंग की क्षमता को रेखांकित करता है।

ABSTRACT. The prediction of the duration of the occurrence of low-visibility events in a calendar day is a difficult process because of the complex and chaotic mechanisms of the onset and dissipation of the low-visibility events. However, it is most useful for the operation of airport services (scheduling of aircraft, optimal operations of the airports) and the planning of any activities (travel, tourism, agriculture, *etc.*). This research tries to build the best dynamic weighted ensemble of the best combination of base machine learning (ML) models (Light Gradient Boosting Machine (Light GBM), Random Forest (RF), and Support Vector Regression (SVR)) to accurately nearcast the duration of low visibility events (fog (surface visibility <1000 m) and dense fog (surface visibility <200 m) for a calendar day based on the initial conditions of 1500 UTC (Universal Time Co-Ordinate). Conditions such as surface meteorological parameters (air temperature, dew point temperature, relative humidity, wind (every 3 hours), rainfall (daily), and sunshine (daily)) and upper air meteorological parameters (wind, temperature, and relative humidity of 1000, 925, and 850 hPa (every 3

hours)) were taken into account to find the best set of explanatory factors for the accurate nearcasting of the duration of the low visibility events. The Pearson correlation coefficient and Spearman's rank correlation coefficient were used to choose the final set of model explanatory variables. The datasets were thoroughly examined using supervised ML algorithms at the various stages of training, testing, modelling, and cross validation. All the best combination models' accuracy was evaluated and compared using performance measures, namely MAE (mean absolute error), RMSE (root mean square error), and R^2 (R squared error). Based on the coefficient of determination (R^2), it can be observed that the suggested dynamic weighted ensemble model exhibits the best level of prediction accuracy, specifically 0.89 and 0.88 for the duration of fog and dense fog for a given lead time of a day. This surpass the accuracy of LightGBM (0.79), RF (0.78), and SVR (0.76) for the prediction of the duration of fog. Therefore, this study highlights the potential of ma-chine learning in facilitating the advancement of automation in airport scheduling and optimizing the operations of airports, specifically in the fog-prone Indo-Gangetic Plains (IGP).

Key words – Duration of low-visibility events, Weighted ensemble, Machine learning, Nearcasting, Airport operations.

1. Introduction

Fog is a meteorological phenomenon characterized by a boundary layer containing a significant accumulation of water droplets or ice crystals; as a consequence, visibility is diminished to a distance of less than 1 kilometer (World Meteorological Organization 2019). Fog has been the subject of an abundance of research, which has utilized a wide range of methodologies and perspectives (Gultepe *et al*., 2007; Long *et al*., 2021; Lakra and Avishek, 2022; Bari *et al*., 2023; Shankar and Sahana, 2023a). The effects of fog on humans and the local economy have been the subject of an abundance of research (Pérez-Díaz *et al*., 2017; Peng *et al*., 2018; Gu *et al*., 2019). Extensive fog considerably hinders the movement of sea, land, rail, and air transportation, leading to considerable economic consequences (Belaroussi and Gruyer, 2014; Gultepe *et al*., 2017; Wu *et al*., 2018; Kulkarni *et al*., 2019; Chandu *et al*., 2022; Shankar and Giri, 2024). (Tyagi *et al*., 2017, 2020) unveiled worrisome patterns of increased fog prevalence and land, and air pollution in the Indo-Gangetic Plain (IGP) from November to February. Concerns have been expressed regarding the socioeconomic ramifications of these environmental changes in light of these findings (Gautam *et al*., 2007). (Hosea, 2019; Mitsokapas *et al*., 2021) posit that the occurrence of dense fog at airports causes aircraft to be diverted, delayed, or cancelled, thereby causing passenger inconvenience and financial detriment to airlines (Kulkarni *et al*., 2019). Severe visibility conditions, specifically those falling below 1000 meters, impede the operations of major airports situated in a specific geographic region (Hosea 2019). Diminished visibility can have a substantial detrimental effect on air navigation. In 2017, a dense fog episode in India resulted in the tragic loss of 11,000 lives due to road accidents, and 21 flights were disrupted at Patna airport in December 2017, causing substantial economic losses (Shankar and Sahana, 2023b). The expenses related to fog events and the duration of fog (the period between onset and dissipation) are currently as expensive as the occurrence of thunderstorms (Gultepe *et al*., 2007). Slower operations at airports during the duration of the fog cost several thousand dollars every day (Dietz *et al*., 2019). So, for airports to run as smoothly and efficiently as possible, they need better nearcasting (with a one- to two-day advance) for the duration of fog events, mostly so that airlines and operators can plan their schedules of flight. The advancement of observation and monitoring platforms and networks improves data quality as well as the historical database (Izett *et al*., 2019). AI/ML (artificial intelligence/machine learning) analytical capabilities may improve next-generation fog-episode predictions based on historical datasets. This strategy can improve decision support systems for low-visibility events and improve the decision-making process. The nearcasting of the duration of fog is a huge challenge, as it is associated with complex atmospheric processes. However, the forecasters' understanding of the local conditions, ability to extract the desired input, and understanding of the algorithms improve the data-driven nearcasting of the duration of low-visibility events (fog and dense fog). The proposed nearcasting model, which is a dynamic weighted ensemble of the best combination of the base ML models (SVR, RF, and light GBM), predicts the durations of the low-visibility events (fog and dense fog) of a calendar day at the initial conditions of 1500 UTC. The surface meteorological datasets of Patna Airport for the synoptic hours (03 hourly) of the parameters Temperature, Dew Point Temperature, Relative Humidity, and u and v components of winds and daily rainfall and sunshine and corresponding upper air data derived from the Indian Monsoon Data Assimilation and Analysis reanalysis (IMDAA) dataset (Indirarani *et al*., 2021) and the target duration of low visibility events (fog (visibility <1000 m) and dense fog (visibility <200 m)) derived from the instrumental visibility dataset. The IMDAA datasets and observed datasets of the Patna airport are representative of the fog-prone IGP regions. Therefore, the proposed techniques may be used at any location in the IGP regions

without any further changes in their algorithms or data ingestion. The novelty of the research article is outlined as

(*i*) This study investigates the optimal combinations of the best base ML models that are well-suited for the conditions of Indo-Gangetic Plains (IGP) regions. Additionally, it proposes a dynamic weighted ensemble approach that exhibits superior generalization ability and default sample recognition ability while also preserving the robustness and interpretability of the model's results.

(*ii*) Ensemble modelling reduces bias in prediction results by creating sub-datasets with varying sample imbalance ratios and training base models with targeted prediction abilities for specific sample classes. Integration leads to improvements in both base-model complementarity and overall performance.

(*iii*) It is planned to make a dynamic weighted ensemble method that is more flexible and has a changing balancing effect. The ensemble weight is changed for each sample that needs to be predicted based on the different prediction results of each base model and how well they recognized specific class samples in the validation stage. This makes setting the weights even more flexible.

(*iv*) Therefore, this research suggests an alternate forecasting technique for the most essential nearcasting of the duration of low visibility in a calendar day.

The subsequent sections of this work are structured in the following manner: Section 2 discusses the previous related works and the following: Section 3 of the paper presents the observed dataset and the architecture of the proposed ML dynamic weighted ensemble model. Section 4 contains a comprehensive evaluation and analysis of the assessment outcomes, along with a detailed discourse on these findings. The findings and implications of our study are presented in Section 5.

1.1. *Related work*

Two kinds of current state-of-the-art visibility prediction algorithms are summarized in Table 1. The first category includes creating NWP models that employ fluid mechanics and thermodynamic equations to forecast weather and its progression. To forecast the onset and dissipation of fog, *i.e*., the duration of fog, meteorologists have created numerous NWP models. For instance, (Bergot *et al*., 2005) created a one-dimensional numerical model to anticipate reduced visibility near Charles de Gaulle Airport. To anticipate reduced visibility in coastal zones, (Müller *et al*., 2010; Dhangar *et al*., 2021) and (Parde *et al*., 2022) used three-dimensional numerical models. Moreover, mesoscale models like WRF models

TABLE 1

Various state-of-the-art methods for the prediction of low visibility.

Categories	Author	Year	Method	Study Area	Predicted Time Interval
NWP Based	(Bergot *et al*., 2005)	2005	One dimensional COBEL model	Airport	30 min to 6 h
	(Müller *et al*., 2010)	2010	Three dimension model	Complex topographic terrain	03 hours
	(Melo *et al*., 2023)	2023	One Dimension PAFOG model	North East Brazil	06 hours
	(Parde *et al*., 2022)	2022	WRF model	IGP regions	03 hours
Meteoro-logical Feature Based	(Koziara, M., Robert, J., Thompson 1983)	1982	Multiple linear regression model	Marine Area	24 hours
	(Miao *et al*., 2020)	2020	Deep Learning	Traffic Freeway	01 to 04 hours
	(Shankar and Sahana 2023b)	2023	Ensemble modelling	Airport	01 to 06 hours
	(Zhai *et al*., 2023)	2023	Ensemble learning	Freeway	15 Min

replicate fog generation, dissipation, and development (Román-Cascón *et al*., 2012, 2016; Steeneveld *et al*., 2015; Ryerson and Hacker 2018; Pithani *et al*., 2019; Pahlavan *et al*., 2021). While these models may mimic fog generation and progression, they mostly forecast visibility in mesoscale places like airports and coastal areas. Certain geographical places, such as airports and coastal regions, exhibit a wide range of microclimatic conditions that vary across both temporal and spatial dimensions. It is imperative to incorporate this information into models, which therefore leads to an escalation in the expenses associated with state-of-the-art monitoring devices. Numerical weather prediction (NWP) models need the use of precise data and robust computational resources, rendering them challenging for accurately forecasting short-term visibility under location-specific circumstances.

A different way to predict visibility is to build prediction models by looking at the relationship between fog formation and certain weather conditions, such as wind speed, humidity, air temperature, barometric pressure, rainfall, *etc*. Fog formation occurs when the temperature approaches the dew point and there are enough condensation nodules in the air (Pulugurtha *et al*., 2019). These weather elements are fed into visibility prediction statistical models (Román-Cascón *et al*., 2016; Cornejo-Bueno *et al*., 2021). Most statistical models are simpler than NWP models. While most fog forecasting research focuses on the sea (Gultepe *et al*., 2017; Han *et*

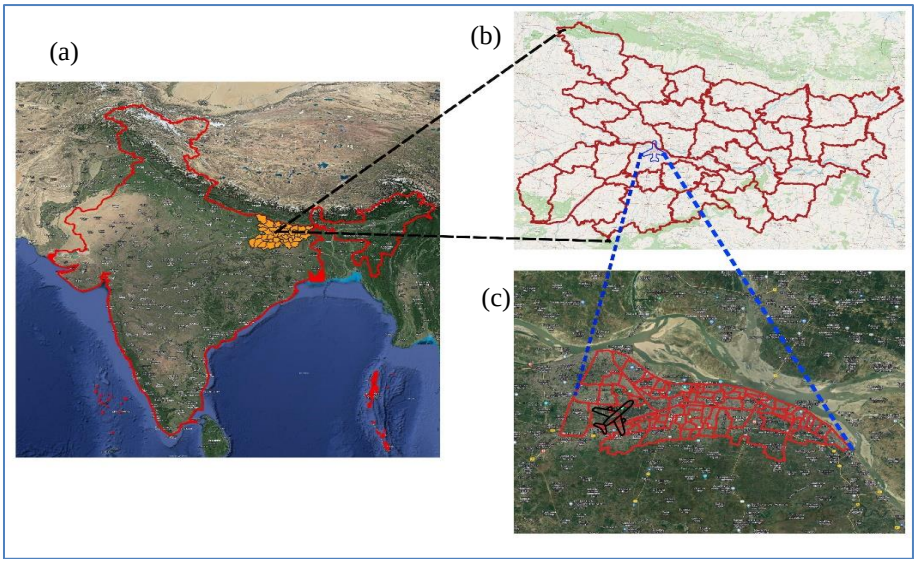


Fig. 1. The geographical location of Jay Prakash Narayan International Airport Patna (a) India (b)State of Bihar, India (c) Capital Cities of Patna and its Airport

al., 2021) or airports (Teixeira and Miranda 2001; Dutta and Chaudhuri 2015; Cornejo-Bueno *et al*., 2017; Shankar and Sahana, 2023b), statistical models can anticipate fog events hours in advance, enabling proactive management measures. Statistical approaches have limited prediction capability; hence, machine learning methods are being used to create models that can handle non-linear connections. Many machine-learning approaches have been studied, such as probabilistic neural networks, multilayer perceptron's, Bayesian decision networks, ordinal classification, and support vector machines (Guijo-Rubio *et al*., 2018; Ortega *et al*., 2020; Bartok *et al*., 2022). These strategies train base learners using historical meteorological data and develop a correlation between visibility and meteorological parameters. Limited information in training data may restrict the selection of the best learner, resulting in lower prediction capabilities. Ensemble learning will be used to forecast visibility to compensate for the errors of the previous approaches. Multiple researchers tackle the same issue in this machine learning paradigm (Zhu *et al*., 2018; Shahhosseini *et al*., 2022; Shankar and Sahana 2023b). Ensemble learning typically outperforms base learners in generalization ability. Ensemble learning techniques may improve prediction accuracy and computational time and prevent overfitting in theoretical and practical investigations (Huang *et al*., 2018; Shahhosseini *et al*., 2022; Shankar and Sahana 2023b). Ensemble learning offers efficient computation and the capacity to handle the complicated non-linear connection between visibility and meteorological factors. Most of these studies try to predict fog (visibility<1000) or no fog (visibility>1000 m), dissipation of fog as a classification issue, and visibility as a regression problem. None of these studies tries to predict the duration of fog in calendar days, which includes the onset and dissipation of low-visibility events in advance. Also, most of these studies work well in a matter of a few hours (Nowcasting). Previous studies have had lots of issues with estimating weakness, as mentioned (Vorndran *et al*., 2022), and none of them attempted to forecast the duration of the fog. The accurate prediction of it directly serves the needs of the end users. This study tries to predict and suggests dynamic weighted ensemble models predict the duration of fog and dense fog with practical implementation in the optimal operations of the airports in a lead time of a few days (nearcasting) with satisfactory operations. Pearson correlation coefficients and Spearman's rank correlation coefficients identify the best meteorological variables linked to the duration of low-visibility events. Next, the dynamic weighted and simple ML algorithms (RF, Light GBM, and SVR) create short-term prediction models (nearcast model). The prediction performance and computational costs of the proposed dynamic weighted ensemble models are compared to those of baseline benchmarked models. Also, the proposed approach is tested for the different sets of data from the representative station of the IGP regions (Patna Airport).

2. Data and methodology

2.1. *Study area*

The specific sites within the Indo-Gangetic Plain (IGP) regions, which are sandwiched between the southern Plateau and the Northern Himalaya (as shown in Fig. 1.), are the focus of estimating the duration of low-

TABLE 2

The details of the input datasets and targets used in the prediction of the duration of fog and dense fog specific to the conditions of IGP regions (representative station: Patna Airport)

Type	Input/Target Variables	Unit	Indicator	Source
Surface Meteorological Data (Instrumental)	Air Temperature (Synoptic Hours *i.e* 03 hourly)	°C	18DB,21DB,00DB,03DB,06DB,09,DB,12DB,15DB	Automatic Weather Observing Station (AWoS), IMD
	Dew Point Temperature (Synoptic Hours *i.e* 03 hourly)	°C	18DP,21DP,00DP,03DP,06DP,09,DP,12DP,15DP	
	Relative Humidity (Synoptic Hours *i.e* 03 hourly)	%	18RH,21RH,00RH,03RH,06RH,09RH,12RH,15RH	
	U wind (Synoptic Hours *i.e* 03 hourly)	knots	18UWIND,21UWIND,00 UWIND,03UWIND,06 UWIND,09UWIND,12UWIND,15UWIND	
	V wind (Synoptic Hours *i.e* 03 hourly)	knots	18VWIND,21VWIND,00 VWIND,03VWIND,06 VWIND,09VWIND,12VWIND,15VWIND	
	Rainfall (Daily)	In mm	Rainfall	Class I Observatory at Patna (IMD)
	Sunshine (Daily)	Hours	Sunshine	
Upper Air Sounding Data (IMDAA)	Temperature (03 hourly) of 1000,925 and 850 hPa	°C	18T1000,21T1000,00T1000,03T1000,06T1000,09,T1000,12T1000,15T1000,18T925,21T925,00T925,03T925,06T925,09,T925,12T925,15T925,18T850,21T850,00T850,03T850,06T850,09,T850,12T850, reanalysis 15T850.	Derived point Data from IMDAA Gridded dataset
	Relative humidity(03 hourly) of 1000,925 and 850 hPa	%	18RH1000,21RH1000,00RH1000,03RH1000,06RH1000,09,RH1000,12RH1000,15RH1000,18RH925,21RH925,00RH925,03RH925,06RH925,09,RH925,12RH925,15RH925,18RH850,21RH850,00RH850,03RH850,06RH850,09,RH850,12RH850,15RH850.	
	U Wind (03 hourly) of 1000,925 and 850 hPa	Knots	18VW1000,21VW1000,00VW1000,03VW1000,06VW1000,09,VW1000,12VW1000,15VW1000,18VW925,21VW925,00VW925,03VW925,06VW925,09,VW925,12VW925,15VW925,18VW850,21VW850,00VW850,03VW850,06VW850,09,VW85 ,12VW850,15VW850.	
	V Wind(03 hourly) of 1000,925 and 850 hPa	knots	18UW1000,21UW1000,00UW1000,03UW1000,06UW1000,09,UW1000,12UW1000,15UW1000,18UW925,21UW925,00UW925,03UW925,06UW925,09,UW925,12UW925,15UW925,18UW850,21UW850,00UW850,03UW850,06UW850,09,UW850,12UW850,15UW850.	
Target (duration of fog and dense fog)	The period between Onset and Dissipation of Fog	hours	Duration of fog (Surface visibility <1000 m)	Derived Parameters (Transmissometers & Scatterometer) Installed at the Patna.
	The period between Onset and Dissipation of Dense Fog	hours	Duration of dense fog (Surface visibility <200 m)	

visibility events (fog and dense fog). The sites of Jay Prakash Narayan International (JPNI) Airport, which lies in the IGP region, have been taken into consideration for the evaluation of the proposed models. There are two primary justifications for this particular choice: The initial aspect pertains to the ongoing surveillance of data, and the presence of Class 1 observatories located at the JPNI Airport in Patna and the Automatic Weather Observing Station (AWoS) facilitates ongoing surveillance of meteorological data sets, encompassing visibility measurements that offer ample training data for our models. Furthermore, limited visibility leads to notable social and economic repercussions in the IGP regions. Low-visibility incidents have had a significant negative impact on the operational effectiveness of aviation services in recent years, causing delays, rescheduling, diversion, and cancellations of flights. Hence, the provision of precise forecasts about the duration of low-

visibility events in advance of one or two days can effectively contribute to the mitigation and improvement of the economic repercussions experienced by the aviation sector.

2.2. *Datasets*

During this research, both ground-based observation data from the Automatic Weather Observing System (AWoS) and instrumental visibility (Transmissometers and scatterometer) and the Upper Air dataset (reanalysis dataset of IMDAA) of the study area (for the period January 2017 to February 2023) of the Jay Prakash Narayan International Airport (JPNI), Patna (25.5947° N, 85.0908° E) have been taken for the analysis and prediction of the duration of low visibility events. The designated temporal period for fog prediction is during the nocturnal hours at 1500 UTC. The models will commence running and generate simulations for the following day. These simulations will include predictions regarding the duration of low-visibility events (fog and dense fog hours) for the next few days. The details of the datasets are presented in Table 2.

The surface air temperature, dew point temperature, relative humidity, u and v components of wind in the synoptic hours (from 1800 UTC of previous days to 1500 UTC), daily rainfall, sunshine, and upper air sounding data (u and v wind, temperature and relative humidity) of the synoptic hours of 1000, 925 and 850 hPa are used to train and test the proposed models with the historical dataset and nearcast the duration of fog and dense fog for the next few days at 1500 UTC. As per WMO rules, the accuracy of the instruments is checked regularly for the target variables, which are the duration of the fog and dense fog as measured by the visibility instruments (Transmissometers or scatterometer), as well as the surface meteorological parameter taken from the Automatic Weather Observing System (AWoS) at Patna Airport. The IMD's certification standards are 0.1 °C for air temperature and 1% for relative humidity, 0.2 m/s for wind speed, and 0.5 mm for precipitation. The different sets of the dataset for the periods November, December, January, and February from 2017 to 2023 are used as training (80%) and testing (20%) sets to evaluate and compare the proposed dynamic weighted ensemble models and their base models. In the investigation, instances where visibility was documented to be below 1000 meters were regarded as the length of fog. Similarly, if visibility measures below 200 meters, it is categorized as dense fog during the hourly observation. The monthly distribution of the mean and standard deviation of the duration of fog and dense fog for the studied period is shown in Fig. 2(a). Fig. 2(b) shows the temporal distribution of the duration of fog and dense fog for one of the fog seasons of 2022-23 (November-December).

2.3. *Methodology*

Pearson's correlation coefficients and Spearman's rank correlation coefficient are employed to identify meteorological elements correlated with the duration of low-visibility events. The best combinations of three representative algorithms are used to nearcast the duration of low-visibility events. These are RF, light GBM, and SVR. Additionally, the best combination of dynamic-weighted ensemble and simple ensemble models performs well against the benchmarked model. The methodology outlined in this study is utilized to estimate the duration of low-visibility events (fog and dense fog), *i.e*., the duration between their onset and dissipation. This prediction task is explicitly approached as a non-linear regression problem. The detailed methodology is presented in Fig. 3.

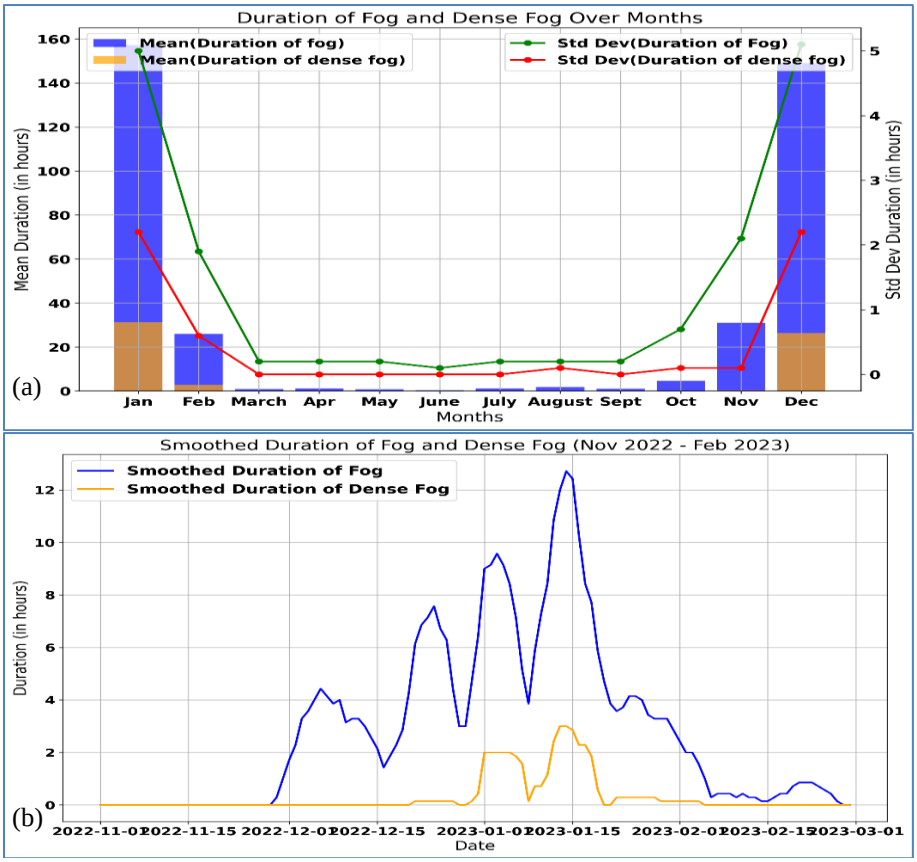
All the benchmarked models and their proposed dynamic weighted ensembles have been developed using the Python programming language on the Anaconda Platform. The evaluation of the performance of these models is conducted by utilizing the performance metrics outlined in sub-section 3.4.

2.3.1. *Random forest*

Since classification and regression trees (CART) use machine learning and are created on the same data, they are sometimes correlated and statistically dependent, making them more varied and uncorrelated. (Breiman, 2001) recommended growing each bagged tree split using random characteristics and observation samples. This is called a random forest (RF) based on bagging methodology. The detailed procedures of random forest models are presented in Fig. 4. RF requires setting the number of trees B (forest size), the number of predictors m out of p variables (features) for each randomly chosen split, and the n_{min} minimum number of observations per node (leaf size). Three main steps comprise the random forest algorithm (Hastie, Tibshirani, 2009).

(*i*) Create B-size training datasets of size N; these datasets can be replaced and overlapped randomly.

(*ii*) Using the following procedures, create a random forest tree T_b for each sample dataset until the minimal node size n_{min} is reached. (*a*) Randomly choose m predictors from p variables. (*b*) Choose the best-split point predictors from m. (*c*) Set certain decision rules to split this node into two daughter nodes.



Figs. 2(a & b). (a) The Mean and Standard Deviation of the Duration of Fog and Dense Fog across the Years (b) The Duration of Fog and Dense Fog for the Season (November to February)

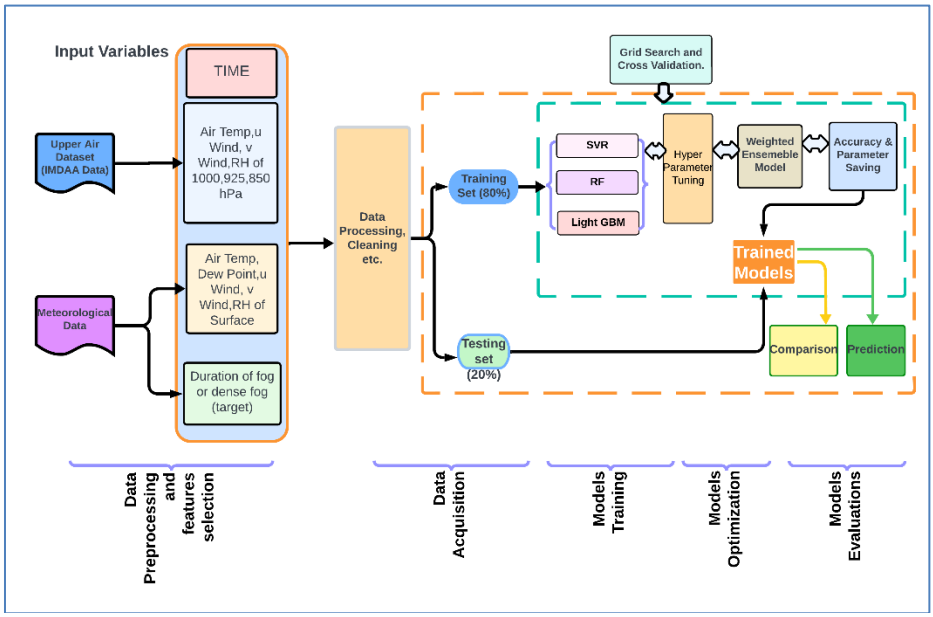


Fig. 3. The detailed procedures (step by step) of the applied methodology (data source, feature selection, data acquisition, model training, model optimisation, and model evaluation) in the predictions of the duration of fog and dense fog

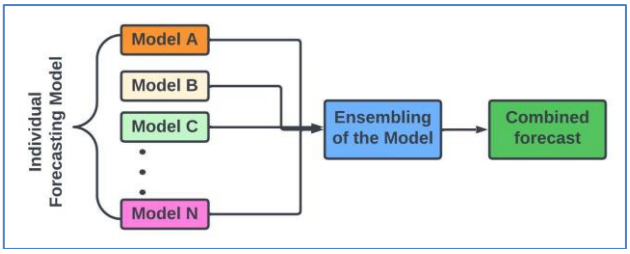


Fig. 4. Schematic of RF Models

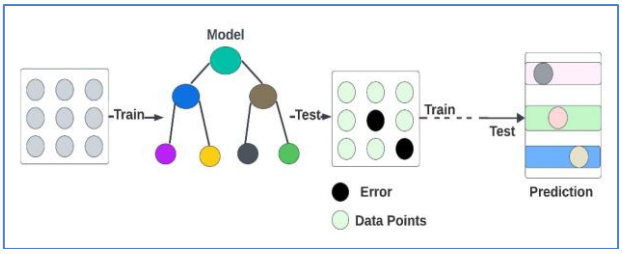


Fig. 5 Schematic of Light Gradient Boosting Machine (Light GBM) Regression

(*iii*) Finally, get the tree ensemble $\{T_b\}_1^B$ where B is the random forest's tree count.

The individual tree's outputs may be averaged to predict a response variable at a point x.

$$\widehat{f_{RF}} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (1)$$

Ensemble learning methods reshape input data to create regression trees that best match feature-output relationships. This decorrelation of the trees makes random forest outputs less variable and more reliable. Its efficacy stems from its capability to aggregate the predictions of numerous randomized decision trees through the process of averaging. Analyzing nonlinear, collinear, and interactive data, the algorithm avoids the overfitting issue with commendable performance. The best hyper parameters of the RF algorithms achieved in the local conditions of IGP regions: max_depth: none, min_samples_leaf=4, min_samples_split: 10, and n_estimators: 150.

2.3.2. *Support Vector Regression (SVR)*

Support vector machines (SVM) can handle continuous and categorical data for regression and classification. Kernel-based SVM reduces over-fitting by minimizing structural risk with a regularization parameter (cortes *et al*., 1995). A linear Support Vector Regression model may be constructed using the following equation, often known as regression-based SVM.

$$f(x) = \sum_i^n \varphi(x_i)w + b \quad (2)$$

The function f(x) is used to denote the output of a model. The variable x_i represents an input variable, while the symbol φ denotes a non-linear mapping. The weight vector w and the regression function bias b are also involved in the model. The loss function is known as the ε-insensitive loss, which was specifically designed for ε-insensitive Support Vector Regression (ε-SVR) (Smola & Scholkopf, 2004). The basic recommendation of Support Vector Regression (SVR) is to minimize the squared norm of the weight vector, $|w|^2$ & the cumulative number of permissible errors. To evaluate the deviation of training samples outside the ε-insensitive zone, non-negative slack variables ξ_i & ξ_i^-(where i = 1, 2,., n) are introduced. Therefore, the ε-SVR algorithm aims to minimize the following function:

$$C\sum_{n=1}^{N}(\xi_n + \xi_i^-) + \frac{1}{2}||w||^2 \quad (3)$$

The following equations apply to the minimization function.

$$\xi_n \geq 0 \; and \; t_n \leq f(x_n) + \epsilon + \xi_n \quad (4)$$

$$\xi_n^- \geq 0 \; and \; t_n \geq f(x_n) - \epsilon - \xi_n^- \quad (5)$$

The best hyper parameters of the SVR algorithms were achieved at C:0.1, Epsilon:0.1, and kernel: linear in the studied areas.

2.3.3. *Light GBM Regression*

A novel machine learning approach is used for more accurate residual value modelling and prediction in data processing. LightGBM excels in data categorization and regression with reduced processing time. A new methodology combines Exclusive Feature Bundling (EFB) with Gradient-based One-Side Sampling (GOSS) for data sampling and classification (Guolin Ke, et al., 2017). Combining these attributes enables efficient and accurate data scanning, sampling, grouping, and classification in less time than traditional methods. When considering memory consumption, processing time, and arithmetic performance, LightGBM excels in training speed, efficiency, memory utilization, accuracy, parallelism, and large-scale data processing. This study evaluates the effectiveness of LightGBM regression algorithms in predicting the duration of fog and dense fog. Also examined gradient boosting's effectiveness in reducing variation and improving model accuracy. The best hyper parameters attained are: colsample_bytree:1.0, learning_rate:0.01, n_estimators:150, num_leaves:31, and subsample:0.8.The LightGBM regression process is presented in Fig. 5.

TABLE 3

Procedures of the proposed weighted ensemble model

	Algorithms
1	Consider a set of independent variables $\hat{y}_i$ as an input (where i=1, 2, 3…n) The prediction is the duration of the daily fog and dense fog.
2	Regress independent variables using SVR, Light GBM, & RF estimate.
3	Calculate a weighted average of equation (7) and regress values of all three models to ensure $0 \leq w_i \geq 1$.
4	Calculate expected values using a weighted ensemble model.
5	Compare individual forecast errors using a weighted ensemble mode

TABLE 4

Mathematical representation of performance metrics

Measure	Formula	Description of Variables
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	n is the sample size; y_i, $\hat{y}_i$ are the actual and predicted values of the ith case.
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert(y_i - \bar{y}_i)\rvert$	$\bar{y}_i$ is the arithmetic mean of Y, and SSE is the sum of squares of the residuals, which is equal to $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. SST is the total sum of squares equal to
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{SSE}{SST}$	

2.3.4. *Proposed ensemble models*

A weighted ensemble, derived from a model averaging ensemble, weighs each member's impact on the final forecast based on the model's efficiency. Each model has moderate positive weights with a sum of one, representing the proportion of trust or projected performance. Ensemble forecasts are created by averaging individual regression forecasts (Pawlikowski and Chorowska, 2020). The mode of member forecasts is used to compute class-label projections. Use the 'arguments of the maxima' of accumulated probabilities for each class label to construct class probability predictions.

2.3.5. *Model description*

Consider $\hat{y}_i$ as the set of independent variables where (i=1, 2, 3……n). The series data set is defined as $Y=[y_1+ y2_{+....+}\ y_n]^T$ with $\hat{y}_i = [y_{1i} + y_{2i} + \cdots + y_{ni}]^T$ as the prediction obtained from the i^{th} method.

The basic ensemble is formed by weighting each forecast model equally. The weight assigned to each model is labelled as w_i.

$$\hat{y}^{(c)} = [y_1 + y_2 + \cdots + y_n]^T \quad (6)$$

$$\hat{y}^{(c)} = w_0 + w_1\ \hat{y}_k{}^{(1)} + w_2\ \hat{y}_k{}^{(2)} + w_3\ \hat{y}_k{}^{(3)}/3 \quad (7)$$

A speculative weighting ensemble algorithm with steps 1–3 is presented in Table 3.

2.4. *Evaluation of the Model's Performances*

After allocating 80% of the datasets for training, the remaining portion as test data was used to evaluate the model's performance. Several assessment measures may be used to evaluate performance. This study used three capacity measuring measures (MAE, RMSE, and R^2) to evaluate continuous qualities. Measurements were taken throughout the testing datasets to ensure the accuracy and validity of the proposed prediction models. The coefficient of determination, R^2, represents prediction accuracy. More accurate predictions are produced with R^2 closer to 1. For continuous dependent variables, MAE and RMSE are prominent performance indicators. The objective of this study was to provide highly precise predictions while acknowledging the potential margin of error associated with these estimations. These metrics better assess accuracy since they reveal the prediction's mistake. The R^2 is the best way to evaluate error in this research; thus, the models were trained on the datasets and tested using it. Table 4 shows the adopted measures mathematically.

To get a fair and consistent evaluation, performance indicators were calculated for all base models, the proposed ensemble model, and the weighted ensemble model. The study measured the divergence between observed and predicted values. As performance metrics decline, model accuracy increases. The coefficient of determination (R2) indicates prediction accuracy. Additionally, MAE and RMSE metrics may evaluate the algorithms' prediction abilities.

3. Results and discussion

In this section, we compare the proposed dynamic weighted ensemble and ensemble models' performance to that of the state-of-the-art best combination of benchmark models SVR, RF, and LightGBM for multivariate time series prediction of the duration of fog and dense fog for random sets of datasets (training (80%) and testing (20%)) with forecast horizons of one day and two days. The

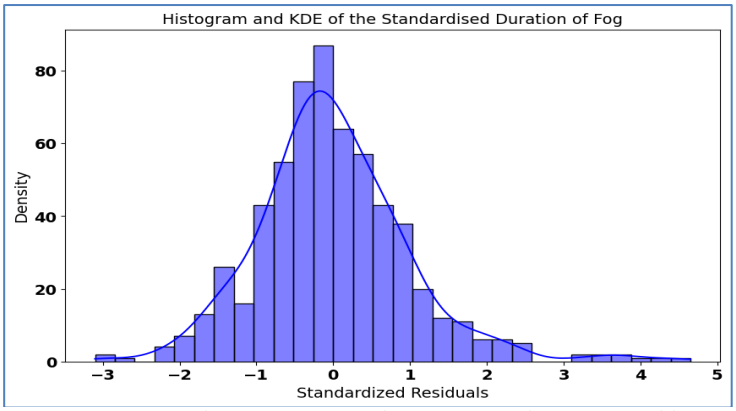


Fig. 6. Normality of the residual value of the predictors of the duration of fog

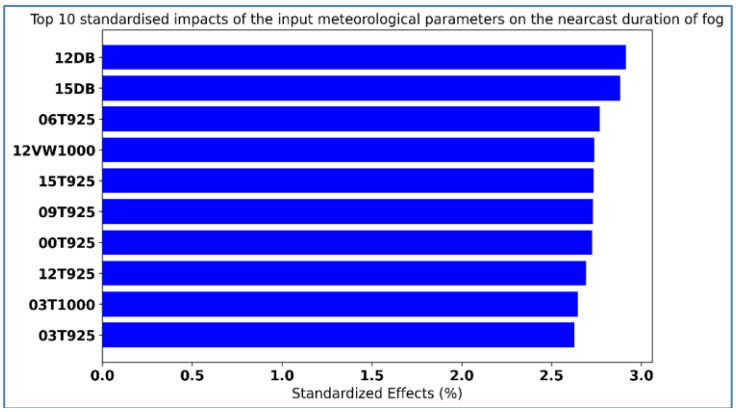
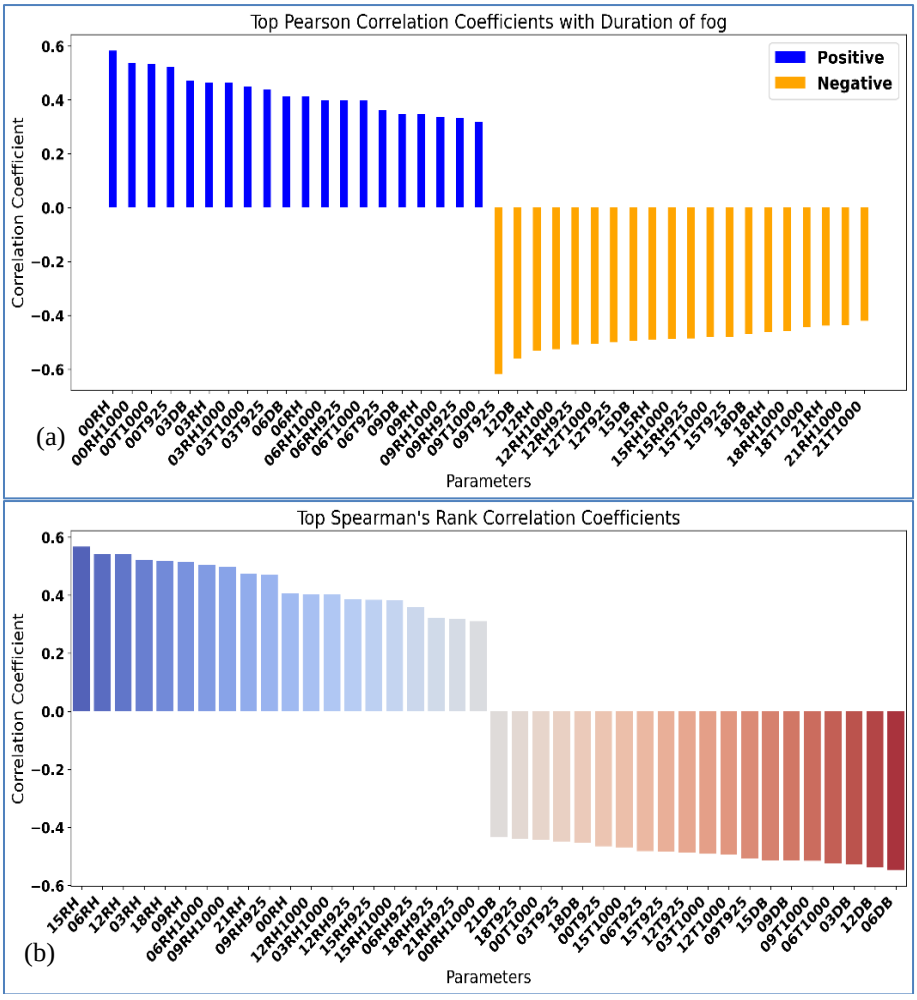


Fig. 7. The standardized impacts of the input features on the nearcasting of the duration of fog

implementation code was written in Python 3.10. The proposed machine learning algorithms are developed on a laptop with a Windows 11 operating system and an Intel(R) Core (TM) i5-1035G1 processor running at 1.00 GHz and 8 GB of memory. The details about the datasets are mentioned in sub-section 3.2. Also, it is crucial to verify the normality of the duration of fog and dense fog for revising model assumptions and making necessary adjustments to improve prediction accuracy. Proper feature selection and correlation are essential for accelerating prediction and avoiding overfitting by minimizing the number of attributes. Additionally, the proposed models' performance evaluation includes comparing the accuracy of prediction outcomes and the competency of the proposed weighted ensemble and the base model algorithms. To examine the correlation between the duration of fog and dense fog with the predictors, model assumptions were examined and corrected before the final proposed dynamic weighted ensemble model. Fig. 6 preserves the normality of the duration of fog residual values. The response variable, the duration of fog and the dense fog was not skewed and did not require a normality analysis and possible adjustment based on its distribution. Results-based adjustments were applied to fulfil near-normal distribution assumptions (mean = 0). Before picking the optimal model, assumptions were examined to ensure accurate findings and corrective actions were taken

3.1. *Correlation of the Attributes (Input Features)*

Various methodologies can be employed to ascertain the significance of features. The backward elimination method, which is a wrapper approach, utilizes the performance of features as a criterion for evaluating and selecting influential factors. After that, the model's performance is checked using an iterative process. This process keeps getting rid of the features that don't work well until the model's overall accuracy is within a reasonable range. The performance parameter employed in this study to assess the effectiveness of features is the p-value. Features with a p-value < 0.05 are eliminated. Fig. 7. illustrates the capacity of separate predictors to identify the most crucial elements in the prediction process. The crucial factors for estimating the duration of fog and dense fog include the air temperature at 12 and 15 UTC, as well



Figs. 8(a&b). (a) Top significant Pearson correlation coefficients of the model's attributes; (b) Top Spearman's rank correlation coefficients of the model's attributes

as the temperature at 925 hPa at 06 UTC, which is associated with the inversion layers. Additionally, the wind at 1000 hPa at 12 UTC is also a significant factor.

As per feature engineering, the most influential attributes are evaluated for correlation analysis (Hauke and Kossowski, 2011). The association values around 1 imply a strong and direct association between predictor and input features, as recognized by the Pearson correlation coefficient. However, correlation levels around -1 suggest a significant inverse association. The Pearson correlation coefficient of the top features for both direct and inverse correlation is shown in Fig. 8. (a). Relative humidity of 00 UTC on the surface and 1000 hPa and air temperature of 1000 hPa are the most directly influencing factors, as well as a temperature of 925 hPa and an air temperature of 12 UTC, which are negatively correlated with the duration of fog and dense fog. As per the analysis of Spearman's rank correlation coefficient, which defines the monotonicity of the input features and predictors, the relative humidity of 15 and 06 UTC is the most directly correlated, and the air temperature of 6 UTC and 12 UTC is negatively correlated with the duration of fog and dense fog. The top-highest spearmen rank correlation coefficients are presented in Fig. 8(b).

Therefore, surface observation along with the upper air datasets of 1000, 925, and 850 hPa is important for the prediction of the duration of fog and dense fog. But particularly the meteorological parameters of the surface to 925 are most important for the proposed nearcasting model

3.2. *Evaluation of Models' Performance*

Following the random allocation of 80% of the datasets to train the model, the remaining portion was used to evaluate the model's performance. The

TABLE 5

Performance metrics of the prediction of the duration of fog and dense fog for a lead time of (a) one day and (b) two days.

(a)

Proposed Models	MAE		RMSE		Coefficient of Determination(R^2)	
	Duration of fog	Duration of dense fog	Duration of fog	Duration of dense fog	Duration of fog	Duration of dense fog
LightGBM	1.81	0.42	2.84	0.93	0.79	0.78
SVR	2.48	0.36	3.40	0.89	0.76	0.79
RF	1.93	0.35	3.01	0.88	0.78	0.80
Ensemble model	1.79	0.29	2.83	0.82	0.83	0.84
Weighted Ensemble Model	1.54	0.27	2.66	0.81	0.88	0.89

(b)

Proposed Models	MAE		RMSE		Coefficient of Determination(R^2)	
	Duration of fog	Duration of dense fog	Duration of fog	Duration of dense fog	Duration of fog	Duration of dense fog
LightGBM	1.54	0.34	2.66	0.86	0.73	0.71
SVM	1.88	0.31	2.85	0.84	0.70	0.74
RF	1.61	0.27	2.69	0.81	0.72	0.76
Ensemble model	1.49	0.27	2.57	0.71	0.75	0.80
Weighted Ensemble Model	1.48	0.25	2.51	0.69	0.79	0.82

performance evaluation process can be implemented by utilizing several assessment metrics discussed in subsection 3.4 (Table 4). These measures were shown to be highly valuable in assessing the performance of the appraisal process. The metrics were computed on the entire set of testing datasets to assess the sufficiency and validity of the proposed dynamic weighted ensembles, simple ensemble models, and their best combination of base benchmarked models. The objective of this study was to generate highly precise predictions while acknowledging the probable margin of error in these estimations. These measurements serve as more reliable indicators of accuracy in this context, as they offer valuable information regarding the potential margin of error in the predictions. Performance indicators were calculated for all models to provide a consistent and accurate evaluation of the prediction of the duration of low visibility events (fog and dense fog). The study measured the divergence between observed and predicted values. As performance metrics decline, model accuracy increases. The Coefficient of Determination (R^2) indicates prediction accuracy. The three algorithms' prediction abilities can be compared using MAE and RMSE performance metrics. The results of the performance evaluation for the discussed models for the nearcast duration of fog and dense fog for the lead times of a day and two days are presented in Tables 5(a&b), respectively.

Grid search is used to find the best model and optimal hyper parameter settings. Section 3.3 provides more detail about the best hyper parameter combinations and the proposed algorithms. Lower values for mean absolute error (MAE) and root mean squared error (RMSE) are indicative of superior model performance. Consequently, among the benchmarked machine learning (ML) models, the performance order for predicting fog is Light GBM > RF > SVR, and for dense fog, RF > SVR > Light GBM, with lead times of one and two days. Additionally, the implementation time for the models follows the order LightGBM > RF > SVR. In summary, the ensemble models demonstrate superior performance compared to their base models (presented in Table 5). Notably, the dynamic weighted ensemble outperforms both the basic ensemble and the benchmarked base models. Consequently, the dynamic weighted ensemble

model emerges as the top-performing model for predicting both fog and dense fog for the next few days.

The coefficient of determination (R^2) gauges the extent to which the variance in the dependent variable can be foreseen from the independent variables. A higher R^2 value, approaching 1, indicates a better fit for the model. In the context of the benchmarked models, performance ranks as follows for nearcasting of fog: Light GBM > RF > SVR, and for dense fog: RF > SVR > Light GBM. Notably, the accuracy of predicting dense fog surpasses that of fog in the same time domain. The dynamic weighted ensemble model stands out with the highest R^2 values (0.88 and 0.89 for lead times of fog and dense fog, respectively), signifying its superior ability to elucidate a larger proportion of the data's variance compared to other models. The dynamic weighted ensemble model works the best, as it has the lowest errors (MAE and RMSE) and the highest coefficient of determination (R^2) for both fog and dense fog predictions.

The proposed multivariate dynamic weighted ensemble models required only 80-90 seconds to train and less than 30 seconds to get the results of the validation datasets. Also, the proposed models achieve very high accuracy (presented in Table 5). From this, we can draw some concluding technical observations based on the findings. Therefore, we may summarise some technical findings based on the preceding results:

(*i*) It is important to use data-driven methods to quickly find the most important features and build multivariate models for comparison, since there are many things about the duration of low-visibility events (like fog or dense fog) that can change how well and accurately they are predicted.

(*ii*) Because of the limitations of most models, it is important to present a dynamic weighted ensemble model for nearcasting the duration of low-visibility events (fog or dense fog).

The final findings demonstrate that the proposed model increases prediction accuracy while decreasing training time. The model's input and output can also be modified to meet changing needs.

4. Conclusions

This research endeavors to forecast the most challenging periods of fog and dense fog in terms of calendar days. The practical implications extend beyond aviation services, impacting areas such as tourism, agriculture, transportation, maritime, and rail services. The research uses machine learning (ML) models to find the best combinations of algorithms and how to tune their hyper parameters based on the knowledge of forecasters and local conditions of the fog-prone Indo-Gangetic Plain (IGP) regions about the things that cause low visibility events, such as the onset and dissipation of the most noticeable radiation, advection fog, and its combination. The proposed dynamic weighted ensemble models utilize the potential of the three different best combinations of the ML models: random forest (RF) for bagging, light GBM for boosting, and support vector regression (SVR) for the robust and most generalizable output. The real database for the research is sourced from historical observations of Surface Meteorological Instruments (AWoS) at synoptic hours (03 hours), daily observations of rainfall and sunshine from Class I Observatories, duration of low visibility events data derived from Transmissometers and scatterometer visibility readings, and upper air data from the IMDAA reanalysis dataset. The selection of pertinent explanatory variables for the models employs statistical measures such as the Pearson correlation coefficient and Spearman's rank correlation coefficient. Comparing the prediction accuracy of the benchmarked models for the prediction of fog in a calendar day, Light GBM has superior performances compared to RF and SVR, and for the duration of dense fog, RF has superior performances compared to SVR and Light GBM in terms of RMSE, MSE, and R^2. The proposed dynamic weighted ensemble model outperforms a simple ensemble, and its benchmarked models demonstrate the effectiveness of the models across different locations. The models exhibit significant variability in their performance, with some excelling while others struggle to predict the duration of low-visibility events. Ensemble models prove valuable in striking a balance, delivering the best and most balanced results across different datasets, notably surpassing the performance of individual models. While the study achieves promising results using training and testing of the real-time observational datasets and the IMDAA Reanalysis datasets, in practical applications, the trained models perform well on the real-time observation data of the upper air data. Additionally, for the improved performance of the proposed models and the timing of the low-visibility events in the implementation of the proposed models, the earlier proposed study (Shankar and Sahana, 2023b) sorted out the most important issues of the location-specific forecast of low-visibility events.

Data availability

The three hourly(synoptic) surface meteorological datasets (observed from AWoS) of Patna Airport, Class-I Observatory Data of Patna, taken from the National Data Centre, Climate Research Station of the India

Meteorological Department, where weather data of the India Meteorological Department is available through the portal https://dsp.imdpune.gov.in/, Also, the Upper Air IMDAA dataset is accessible from its portal, https://rds. ncmrwf.gov.in/. It is noted that these portals can be accessed publicly. Also, data can be shared after the request.

Acknowledgements

The authors acknowledge the National Centre of Medium Range Weather Forecasting for IMDAA reanalysis datasets. For providing the real-time datasets of the visibility observation and corresponding instrumental visibility, the India Meteorological Department Patna is also acknowledged with thanks. AS is thankful to officials of the India Meteorological Department, Patna, especially Ashish Kumar, for the useful discussions.

Authors' Contributions

Anand Shankar: Conceptualization, methodology, software, validation, data curation, writing-original draft preparation.
Bikash Chandra Sahana: writing-revised draft, supervision. (*email- sahana@nitp.ac.in*).
Sunny Chug: writing-revised draft. (*email-sunny.chug@ imd.gov.in*).

All authors have read and agreed to the published version of the manuscript.

Disclaimer: The contents and views presented in this research article/paper are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Alex, J. Smola and Bernhard Scho¨ Lkopf, 2004, "A tutorial on support vector regression", *Stat Comput*, **14**, 199-222.

Bari, D., Bergot T. and Tardif R, 2023, "Fog Decision Support Systems: A Review of the Current Perspectives", *Atmosphere*, Basel, **14**, 1-13. doi: https://doi.org/10.3390/atmos14081314.

Bartok, J., Šišan P, Ivica L, et al, 2022, "Machine Learning-Based Fog Nowcasting for Aviation with the Aid of Camera Observations", *Atmosphere*, Basel, **13**, https://doi.org/10.3390/atmos 13101684.

Belaroussi, R. and Gruyer D, 2014, "Impact of reduced visibility from fog on traffic sign detection", *IEEE Intell Veh Symp Proc*, 1302–1306, https://doi.org/10.1109/IVS.2014.6856535.

Bergot, T., Carrer D, Noilhan J and Bougeault P, 2005, "Improved site-specific numerical prediction of fog and low clouds: A feasibility study", *Weather Forecast*, **20**, 627–646,https://doi.org/10.1175/WAF873.1.

Breiman, L, 2001, "Random Forests", *Mach Learn*, **45**, 5–32.

Chandu, K., Dharma Raju A, Kumar SVJ, et al, 2022, "Operational Constraints on Flight Navigation due to Fog and Consequent Economic Implications at the Rajiv Gandhi International Airport, Hyderabad, Telangana, India", *Asian J Water,Environ Pollut*, **19**, 25–32, https://doi.org/10.3233/AJW220052.

Cornejo-Bueno, L., Casanova-Mateo C, Sanz-Justo J, et al, 2017, "Efficient Prediction of Low-Visibility Events at Airports Using Machine-Learning Regression", *Boundary-Layer Meteorol*, **165**, 349–370, https://doi.org/10.1007/s10546-017-0276-8.

Cornejo-Bueno, S., Casillas-Pérez D, Cornejo-Bueno L, et al, 2021, "Statistical analysis and machine learning prediction of fog-caused low-visibility events at a-8 motor-road in spain" ' *Atmosphere*, Basel, **12**, 1–22, https://doi.org/10.3390 /atmos12060679.

Corte, Corinna and Vapnik V, 1995, "Support-Vector Networks", *Mach Learning* , **20**, 272–297.

Dhangar, NG, Lal DM, Ghude SD, et al, 2021, "On the Conditions for Onset and Development of Fog Over New Delhi: An Observational Study from the WiFEX", *Pure Appl Geophys*, **178,** 3727–3746, https://doi.org/10.1007/s00024-021-02800-4.

Dietz, SJ., Kneringer P, Mayr GJ and Zeileis A, 2019, "Low-visibility forecasts for different flight planning horizons using tree-based boosting models", *Adv Stat Climatol Meteorol Oceanogr*, **5**, 101–114, https://doi.org/10.5194/ascmo-5-101-2019.

Dutta, D. and Chaudhuri S, 2015, "Nowcasting visibility during wintertime fog over the airport of a metropolis of India: decision tree algorithm and artificial neural network approach", *Nat Hazards*, **75**, 1349–1368, https://doi.org/10.1007/s11069-014-1388-9.

Gautam, R., Hsu NC, Kafatos M and Tsay SC, 2007, "Influences of winter haze on fog/low cloud over the Indo-Gangetic plains",J Geophys Res Atmos, **112**, 1–11, https://doi.org/ 10.1029/2005JD007036.

Gu, Y., Kusaka H, Doan VQ and Tan J, 2019, "Impacts of urban expansion on fog types in Shanghai, China: Numerical experiments by WRF model", *Atmos Res*, **220**, 57–74, https://doi.org /10.1016/j.atmosres.2018.12.026.

Guijo-Rubio, D., Gutiérrez PA, Casanova-Mateo C, et al, 2018, "Prediction of low-visibility events due to fog using ordinal classification", *Atmos Res*, **14**, 64–73. https://doi.org/ 10.1016/j.atmosres.2018.07.017.

Gultepe, I., Milbrandt JA and Zhou B, 2017, "Marine Fog: A Review on Microphysics and Visibility Prediction" , *Springer, Cham*, 345–394, https://doi.org/10.1007/978-3-319-45229-6_7.

Gultepe, I., Tardif R, Michaelides SC, et al, 2007, "Fog research: A review of past achievements and future perspectives", Pure Appl. Geophys, **164**, 1121–1159, https://doi.org/10.1007 /s00024-007-0211-x.

Guolin, Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma and Qiwei Ye T-YL, 2017, "Lightgbm: A highly efficient gradient boosting decision tree", *Adv Neural Inf Process Syst*, **30**, 3146–3154.

Han, JH., Kim KJ, Joo HS, et al, 2021, "Sea fog dissipation prediction in incheon port and haeundae beach using machine learning and deep learning", *Sensors*, **21**, https://doi.org/10.3390/s21155232.

Hauke, J. and Kossowski T, 2011, "Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data", *Quaest Geogr*, **30**,87–93, https://doi.org/10.2478/v10117-

011-0021-1.

Hosea, MK., 2019, "Effect of Climate Change on Airline Flights Operations At Nnamdi Azikiwe International Airport Abuja, Nigeria", *Sci World J*, **14**.

Huang, D., Wang CD, Lai JH, 2018, "Locally Weighted Ensemble Clustering", IEEE Trans Cybern, **48**, 1460–1473, https://doi.org /10.1109/TCYB.2017.2702343.

Indirarani, S., Arulalan T, George JP, et al, 2021, "IMDAA: High-resolution satellite-era reanalysis for the indian monsoon region", *J Clim*, **34**, 5109–5133, https://doi.org/10.1175/JCLI-D-20-0412.1.

Izett, JG., van de Wiel BJH, Baas P, et al, 2019, "Dutch fog: On the observed spatio-temporal variability of fog in the Netherlands". *Q J R Meteorol Soc*, **145**, 2817–2834, https://doi.org /10.1002/qj.3597.

Koziara, M., Robert, J. and Thompson W, 1983, "Estimating Marine Fog Probability Using a Model Output Statistics Scheme", Mon Weather Rev, **111**, 2333–2340, https://doi.org/ https://doi.org/10.1175/1520-493.

Kulkarni, R., Jenamani RK, Pithani P, et al, 2019, "Loss to aviation economy due to winter fog in New Delhi during the winter of 2011-2016", *Atmosphere*, Basel, **10**, 1–10, https://doi.org/ 10.3390/ATMOS10040198.

Lakra, K. and Avishek K, 2022, "A review on factors influencing fog formation, classification, forecasting, detection and impacts" , Rendiconti Lincei. Scienze Fisiche e Naturali, **33**, 2, 319-353.

Long, Q., Wu B, Mi X, et al, 2021, "Review on parameterization schemes of visibility in fog and brief discussion of applications performance", *Atmosphere*, Basel, **12**, https://doi.org/ 10.3390/atmos12121666.

Melo, DL de, Oliveira CG de, Fedorova N and Levit V, 2023, "Fog Analysis and Forecast by PAFOG Model in Brazilian Northeast", *Rev Bras Meteorol*, **38**, https://doi.org/10.1590 /0102-77863810215.

Miao, K chao, Han T ting, Yao Y qing, et al, 2020, "Application of LSTM for short term fog forecasting based on meteorological elements", Neurocomputing, **408**, 285–291, https://doi.org/ 10.1016/j.neucom.2019.12.129.

Mitsokapas, E, Schäfer B, Harris RJ and Beck C, 2021, "Statistical characterization of airplane delays", *Sci Rep*, **11**, 1–11, https://doi.org/10.1038/s41598-021-87279-8.

Mülle,r MD, Masbou M and Bott A, 2010, "Three-dimensional fog forecasting in complex terrain", *Q J R Meteorol Soc*, **136**, 2189–2202, https://doi.org/10.1002/qj.705.

Ortega, LC, Otero LD, Otero CE and Fabregas A, 2020, "Visibility forecasting with deep learning" *SYSCON 2020 - 14th Annu IEEE Int Syst Conf Proc*, https://doi.org/10.1109/Sys Con47679.2020.9275833.

Pahlavan, R., Moradi M, Tajbakhsh S, et al, 2021, Fog probabilistic forecasting using an ensemble prediction system at six airports in Iran for 10 fog events", *Meteorol Appl*, **28**, 1–16. https://doi.org/10.1002/met.2033.

Parde, AN, Ghude SD, Dhangar NG, et al, 2022, "Operational Probabilistic Fog Prediction Based on Ensemble Forecast System: A Decision Support System for Fog", Atmosphere, Basel, **13**, 1–17, https://doi.org/10.3390/atmos13101608.

Pawlikowski, M, Chorowska A, 2020, "Weighted ensemble of statistical models", *Int J Forecast*, **36**, 93–97. https://doi.org/10.1016/j. ijforecast.2019.03.019.

Peng, Y., Abdel-Aty M, Lee J, Zou Y, 2018, "Analysis of the Impact of Fog-Related Reduced Visibility on Traffic Parameters", J Transp Eng Part A Syst, **144**, 04017077, https://doi.org/ 10.1061/jtepbs.0000094.

Pérez-Díaz JL, Ivanov O, Peshev Z, et al, 2017, "Fogs: Physical basis, characteristic properties, and impacts on the environment and human health", *Water (Switzerland)*, **9**, 1–21, https://doi.org /10.3390/w9100807.

Pithani, P., Ghude SD, Chennu VN, et al, 2019, "WRF Model Prediction of a Dense Fog Event Occurred During the Winter Fog Experiment (WIFEX)", Pure Appl Geophys, **176**, 1827–1846, https://doi.org/10.1007/s00024-018-2053-0.

Pulugurtha, SS., Mane AS, Duddu VR and Godfrey CM, 2019, "Investigating the influence of contributing factors and predicting visibility at road link-level" Heliyon, **5**, 02105, https://doi.org/10.1016 /j.heliyon.2019.e02105.

Román-Cascón C, Steeneveld GJ, Yagüe C, et al, 2016, "Forecasting radiation fog at climatologically contrasting sites: Evaluation of statistical methods and WRF', *Q J R Meteorol Soc*, **142**, 1048–1063, https://doi.org/10.1002/qj.2708.

Román-Cascón, C, Yagüe C, Sastre M, et al, 2012, "Observations and WRF simulations of fog events at the Spanish Northern Plateau", *Adv Sci Res*, **8**, 11–18, https://doi.org/10.5194/asr-8-11-2012.

Ryerson, WR. and Hacker JP, 2018, "A nonparametric ensemble postprocessing approach for short-range visibility predictions in data-sparse areas", *Weather Forecast*, **33**, 835–855, https://doi.org/10.1175/WAF-D-17-0066.1.

Shahhosseini, M., Hu G and Pham H, 2022, "Optimizing ensemble weights and hyperparameters of machine learning models for regression problems", *Mach Learn with Appl*, **7**, 100-251, https://doi.org/10.1016/j.mlwa.2022.100251.

Shankar, A. and Giri RK, 2024, "The Impacts of Low Visibility on the Aviation Services of Patna Airport During the Period from 2016 to 2023", *Journal of Airline Operations and Aviation Management*, **3**, 1, https://doi.org/10.56801/jaoam.v3i1.5.

Shankar, A. and Sahana BC, 2023a, "Efficient prediction of runway visual range by using a hybrid CNN-LSTM network architecture for aviation services", *Theor Appl Climatol*, **155**, 3, 2215–2232, https://doi.org/10.1007/s00704-023-04751-3.

Shankar, A. and Sahana BC, 2023b, "Early warning of low visibility using the ensembling of machine learning approaches for aviation services at Jay Prakash Narayan International (JPNI) Airport Patna", *SN Appl Sci*, https://doi.org/10.1007/s42452-023-05350-7.

Steeneveld, GJ, Ronda RJ and Holtslag AAM, 2015, "The Challenge of Forecasting the Onset and Development of Radiation Fog Using Mesoscale Atmospheric Models", *Boundary-Layer Meteorol*, **154**, 265–289, https://doi.org/10.1007/s10546-014-9973-8.

Hastie, T., Tibshirani R and Friedman J, 2009, "The elements of statistical learning", *Springer New York*, NY, **2** Edition, https://doi.org/ 10.1007/978-0-387-84858-7.

Teixeira, J. and Miranda PMA, 2001, "Fog prediction at Lisbon airport using a one-dimensional boundary layer model", *Meteorol Appl*, **8**, 497–505, https://doi.org/10.1017/S135048270100411X.

Tyagi, A., Kharb L and Chahal D , 2020, "Scrutinizing Patterns of Air Pollution in India", *Proc - IEEE 2020 2nd Int Conf Adv Comput Commun Control Networking, ICACCCN 2020*, 915–920, https://doi.org/10.1109/ICACCCN51052.2020.9362990.

Tyagi, S., Tiwari S, Mishra A, et al, 2017, "Characteristics of absorbing aerosols during winter foggy period over the National Capital Region of Delhi: Impact of planetary boundary layer dynamics and solar radiation flux", *Atmos Res*, **188**, 1–10, https://doi.org/10.1016/j.atmosres.2017.01.001.

Vorndran, M., Schütz A, Bendix J and Thies B, 2022, "Current Training and Validation Weaknesses in Classification-Based Radiation Fog Nowcast Using Machine Learning Algorithms", *Artif Intell Earth Syst*, **1**, 1–17, https://doi.org/10.1175/AIES-D-21-0006.1.

World Meteorological Organization, 2019, "Manual on Codes International Codes", *2019 editi. WMO-No. 306 © World Meteorological Organization, 2019*.

Wu, Y., Abdel-Aty M and Lee J, 2018, "Crash risk analysis during fog conditions using real-time traffic data", *Accid Anal Prev*, **114**, 4–11, https://doi.org/10.1016/j.aap.2017.05.004.

Zhai, B., Wang Y and Wu B, 2023, "An ensemble learning method for low visibility prediction on freeway using meteorological data", *IET Intell Transp Syst*, https://doi.org/10.1049/itr2.12404.

Zhu, X., Ni Z, Cheng M, et al, 2018, "Selective ensemble based on extreme learning machine and improved discrete artificial fish swarm algorithm for haze forecast", *Appl Intell*, **48**, 1757–1775, https://doi.org/10.1007/s10489-017-1027-8.

Abbreviations

This manuscript employs the following abbreviations

JPNI Airport	Jay Prakash Narayan International Airport Patna
IGP	Indo Gangetic Plains
RF	Random Forest
Light GBM	Light Gradient Boosting Machine
SVR	Support Vector Regression
UTC	Universal Time Co-ordinate
ML	Machine Learning
AI/ML	Artificial Intelligence/Machine Learning
IMDAA	Indian Monsoon Data Assimilation and Analysis Reanalysis