An IoT-aware Air Quality Prediction System utilising Hybrid Optimization and Fuzzy Temporal Rules Enabled Auto Encoded Bi-LSTM

Ayshwarya Lakshmi S[1*], Sannasi Ganapathy[2]

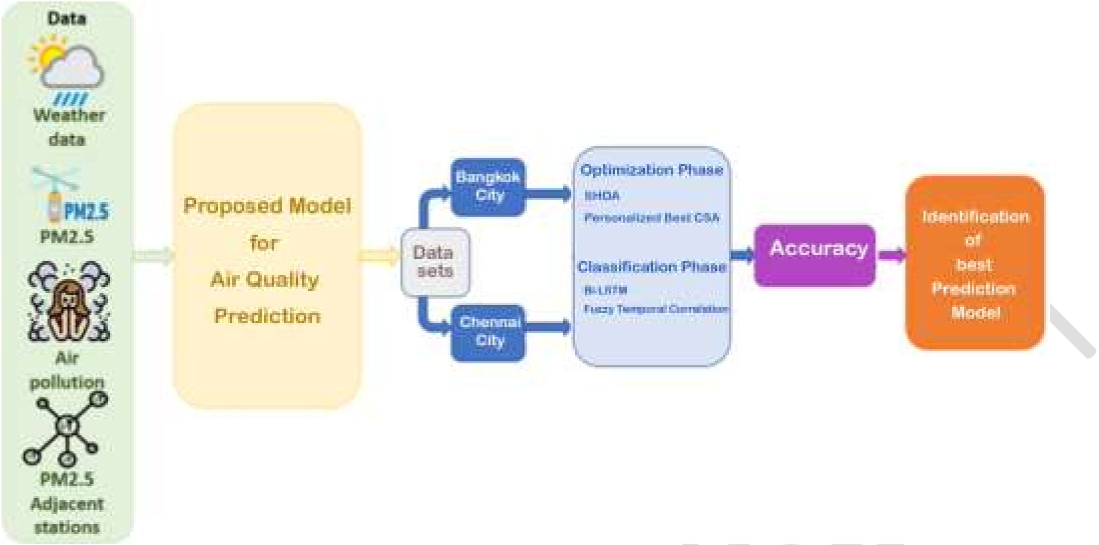
[1] *Department of Computer Science and Engineering, University College of Engineering Panruti, INDIA.*

[2] *Centre for Cyber-Physical Systems & School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, INDIA.*

[*] Corresponding Author:

E-mail*:* *ayshresearch@gmail.com*, tel: *+91 9444206138* fax: -

Graphical Abstract



Abstract

Recently, the environmental pollution is becoming a challenging issue due to the lacking of awareness about the importance of the environment and the growth of transportation facility and the various factories. The Internet of Things (IoT) technology is useful today for collecting and updating the data dynamically from anywhere anytime in less expense. Air quality of different cities is predicted by considering and maintaining the above listed technique which is useful for identifying and predicting the air quality and measure the air pollution level as well. This paper proposes a new IoT aware air quality prediction system that incorporates the newly proposed classifier with fuzzy temporal correlation and the auto-encoded Bidirectional Long Short-Term Memory (Bi-LSTM) to predict the diseases such as heart, cancer and diabetes. Moreover, we propose a new optimization algorithm called Spotted Hyena and Personalized best Cuckoo Search Optimization Algorithm (SHPCSA) for selecting and optimizing the most contributed features that are helpful for enhancing the prediction accuracy with short span of time. The air quality data is collected from various cities like Chennai, Bong Kong, etc. to conduct experiments. The proposed model is proved better than standard classifier with respect to precision, recall, F1-score and accuracy by conducting experiments.

Keywords **–** Air quality, Pollution**,** Machine Learning, Deep Learning, Optimization, SHOA.

1. INTRODUCTION

Quality Air is necessary in this fast world to live a healthy life. The absence of air quality leads to diseases especially related to respiratory diseases like asthma, cancer, etc. Now-a-days the quality of the air is decreasing in fast rate due to the increases in number of vehicles plying on the roads, waters and also in air. In this challenging environment, it is necessary to predict the air quality every day and also needs to provide necessary alert to the public. For this purpose, many air quality prediction systems have been developed by various researchers. Even though, the people are affecting various breath related diseases due to the lack of efficiency and effectiveness in the process of prediction. This work proposes a new air quality prediction system to fulfil the current requirements and it is capable of reducing the air pollution by taking precaution action.

Data pre-processing is the most important task for enhancing the performance of the air quality prediction system. Generally, it contains the feature selection, extraction and optimization processes that are playing vital role and widely used by many researchers and academicians in many applications. The major advantage of the feature selection techniques is the manner in which processing of data analysis is done including understanding the data, data visualization and prediction accuracy. The two types of feature selection approaches are there namely, filter and wrapper methods. First, the filter method is done according to the common behaviours and the wrapper approach is done by applying the ML algorithms for evaluating the various subset of features. Later on, the feature selection algorithms are applied in many prediction systems to enhance the prediction accuracy of the classifier. In addition, the various optimization techniques are incorporated on feature selection process for identifying the important features that are helpful for enhancing the accuracy of the classifier on various data mining algorithms, machine learning and deep learning algorithms as well. Finally, it selects the best feature set that is capable of producing better classification accuracy.

Data mining and Machine Learning (ML) algorithms are normally applied to predict the air quality in various cities of the different countries. The air quality prediction works based on the data of different cities to extract the knowledge. The ML algorithm is very useful in the process of predicting the air quality of the various cities and alerts the people about the pollution level as per admin perspective. Moreover, this work is moved to the next level of training process that can be done by applying deep learning algorithm to predict the quality of the air in the specific city or environment. Many research works have been done in air quality prediction by applying Machine Learning (ML) technique. The ML has demonstrated that their capabilities in handling the air quality datasets efficiently and also dealt with many features while designing a novel air quality prediction model. The supervised ML methods are focusing to find the relevant terms in the form of the relevant features and terms.

With the help of the feature selection process, the ML algorithms are achieving better result on various datasets such as medical datasets, network traffic datasets, e-learning datasets, amazon datasets and weather datasets. Despite this, ML is not able produce the expected accurate prediction result for the various kinds of datasets, even though, the availability of huge data. For overcoming these all, advancement in ML algorithms is introduced as Deep Learning algorithms including Long-Short Term Memory (LSTM), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) for enhancing the performance in training process. In addition, the optimization techniques are also used to improve the prediction accuracy by reducing the number of features that are most important and also contribute more to make effective decisions on the various datasets by deep learning algorithms. For this purpose, we propose a new air quality prediction system in this paper with the combination of an existing feature optimization technique and the effective deep learning technique.

The main contributions and the novelty of this paper are as follows:

1. To propose a new air quality prediction system to predict the quality of the air and the pollution level effectively.
2. To propose a new featured optimization technique namely Spotted Hyena and Personalized best Cuckoo Search Optimization Algorithm (SHPCSA) for effective performance and optimization process for enhancing the prediction accuracy.

3. To apply K-Means clustering algorithm for grouping the patient records with respect to the feature levels of the various datasets.
4. To apply new proposed Fuzzy Temporal Correlation Aware Classifier with Auto-encoded Bi-LSTM for predicting the air quality and their levels are termed as Very high, High, Medium, Low, High and Very High.
5. To obtain more prediction accuracy than other fuzzy temporal and deep learning algorithm model in short span of time and it is compared to other DL based prediction models.
6. To prove that the newly developed air quality prediction system as better by compared with other classifiers with respect to the accuracy, f-measure, recall and precision.

This paper is organized as below: The related existing works that are available in the literature are discussed in detail with respect to the contributions, achievement and drawbacks in section 2. Section 3 shows overall workflow of the proposed air quality prediction model with necessary details by mentioning the algorithms used in this work. Section 4 provides the relevant background information about the proposed model and the algorithms developed in this paper in detail with necessary formulae. Section 5 shows the performance of the proposed system by showing the results in the form of graphs and tables. Section 6 concludes the proposed model by mentioning the achievements of the proposed air quality prediction system quantitatively and it also provides the possible future works.

2. RELATED WORKS

Many weather forecasting and prediction systems are available in the literature that are developed by using neural networks, feature selection, fuzzy rules and temporal constraints and learning techniques by different researchers (Yuval et al 2012, Qi et al 2018, Liu et al. 2021, Atakan et al. 2008, Yi et al. 2022). Among them, Atakan et al. (2008) developed a new online air pollution forecasting system to predict the air pollution in online by considering the pollution indicators for a specific duration in a region. They conducted a comparison with the current systems. According to synoptic forecasts, Yuval et al. (2012) developed a novel approach to evaluate future air pollution levels. They have proven to be superior to other forecasting strategies in terms of predicting future air pollution levels by using the air quality index datasets as input for their algorithm. For the purpose of diagnosing disorders, Sethukkarasi et al. (2014) created a novel cognitive map with fuzzification. Their FTCM used neural networks and time restrictions to analyse datasets and make judgements. In order to effectively perform prediction processes on medical datasets, Ganapathy et al. (2014) introduced a new neural classifier dubbed Fuzzy Temporal Min-Max Aware Neural Network with the common particle swarm optimizer.

In order to achieve high accuracy on the input dataset, Yoichi et al. (2016) created a novel rule extraction approach that combines sampling selection methods. This method achieved about 84% accuracy. Hassan et al. (2016) developed a brand-new hybrid technique that combines SVM and neural networks to classify inter- and intra-class transitions and forecast the beta range. Additionally, their approach makes use of a sliding window aware attribute selection method to extract the appropriate attributes that can improve the performance of the classifier during training

In a detailed review of the many ML and data mining algorithms previously used to forecast diabetes, Ioannis et al. (2017) beat other classifiers in terms of accuracy. A brand-new fuzzy rule-based reinforcement learning-based evolutionary system for diagnosing diabetes was proposed by Fatemeh et al. in 2017. They have incorporated the rule base and the optimization technique for enhancing the performance of their system. They have developed a new rule based which has learnt about the initial rules, numeric values and removed the redundant rules. They have applied suitable rules with interpretability for performing pruning process and uses genetic algorithm and reinforcement learning to achieve better accuracy. Moreover, they have proved that their system is performed well than other models by achieving highest prediction accuracy on diabetes datasets. Amir et al (2018) considered the ten thousand records that are collected from 36 different hospitals in the years between 2008 and 2016. They have performed the prediction process using the well-known classifiers namely SVM, Neural classifier, regression and Decision Trees for identifying the disease affected patient records.

Namrata et al (2019) developed a new hybrid approach to extract the rules using SVM classifier. Moreover, they incorporated a new attribute selection technique for selecting the useful attributes that are grouped significantly from benchmark dataset. In addition, they have applied XG Boost for converting the SVM black box model with the consideration of live data as input. In the end, they demonstrated that their model outperformed other models in terms of accuracy by taking into account both conventional datasets and the live dataset for conducting tests. Ma et al. (2019) present a novel methodology framework integrating a deep learning network, specifically a bidirectional long short-term memory (BLSTM) network with the inverse distance weighting (IDW) technique, for the spatiotemporal forecasts of air pollutants at various time granularities. The BLSTM efficiently captures the long-term temporal mechanism of air pollution. While accounting for the geographic correlation of air pollution, the IDW layer, on the other hand, has the capacity to interpolate the spatial distribution. A case study is conducted to ensure the effectiveness of the given methodology. PM2.5 concentration in Guangdong, China is predicted. The LSTM network's hourly, daily, and weekly prediction performances for various time intervals are shown. Analysis is done on both the prediction errors and the spatial distribution of the expected PM2.5 concentrations. The experimental results demonstrate that the recommended strategy can produce superior prediction performance for the PM2.5 concentration when compared to existing models.

Piotr et al (2019) developed a new cluster aware ensemble approach to predict the air pollution by evolving the Spiking Neural Networks and it is trained on a specific group of time duration. In their method, they have generated a new training dataset by applying clustering method and conducted experiments with air quality index dataset called PM10 dataset. Ping et al (2019) conducted a detailed analysis by incorporating the hybrid air quality and also provides the alert system which comprises the features like feature estimation, prediction and evaluation. Also, a novel hybrid forecasting technique that combines improved data processing, a neural classifier, and a new heuristic technique. They have considered the various forecasting criterion to develop the warning system and achieved better performance in terms of scalability, efficiency and effectiveness.

To keep track of the adjacent industrial zones and the air quality monitoring station, Yue et al. (2020) created a novel Aggregated LSTM approach. In order to improve Taiwan's ability to forecast events between 2012 and 2017, they combined three LSTM models. They have conducted numerous tests and established themselves as superior to the current systems in terms of mistake rates and forecast accuracy. In order to anticipate the amount of air pollution, Masoomeh et al. (2020) created a novel adaptive neuro-fuzzy classifier that analyses the air quality index with the aid of an adaptive neuro-fuzzy inference system. They have carried out numerous trials and demonstrated that they are more accurate predictors than the current models.

Adil et al. (2021) offered a thorough analysis of AI-aware approaches to forecast air pollution using logistic regression, ANN, DNN, and SVM. They conducted a comparison of the suggested model's prediction accuracy with that of the current systems. Mokhtari et al (2021) developed a new multi-point DL aware LSTM with convolutional to predict the air quality dynamically. The proposed model is constructed by applying LSTM, CNN along with spatial and temporal features for evaluating the air quality dynamically and done a comparative analysis with the available air quality prediction models.

An innovative integrated composition and reconstruction technique for forecasting air quality was created by Zhenni et al. in 2022. For carrying out the numerous experiments, they took into account the unique geographical data, and they also showed to be a more accurate model than the current forecasting methods. Sheen et al. (2022) carried out a thorough evaluation while taking the ANN into account to address the uncertainty issues. They have considered 128 research papers that are published in the duration between 2000 and 2022 for this review and it reveals uncertainty. Moreover, the various ensemble techniques are also deployed on neural classifiers with fuzzy logic for predicting the air quality. Krishna et al (2022) developed a new multi-output LSTM with auto-encoder for conducting the multi-step forecasting process. Moreover, their model applied a multiple imputation method for handling the missing values of the dataset and also deployed both the spatial and climate features for improving the forecasting process. Eventually, they have demonstrated that their model has a higher prediction accuracy than the current systems.

A novel technique for estimating air quality by Hongmin et al. (2022) combines the probabilistic method with fuzzy logic, outlier detection, and correction. By including the best weights, their solution addresses the non-deterministic and stochastic problem in the process of predicting air quality. Their approach uses probabilistic forecasting to make interval predictions that follow the ideal distribution. Lastly, they have conducted numerous trials and demonstrated that they are more accurate at predicting air quality than the current techniques. Gu et al. (2022) created a new temporal aware support vector regression that takes into account end-to-end records while taking the crowd and pertinent factors into consideration. At the end, they have developed a multitask TSVR to predict the pollution level in air and also considered the various datasets including PM 2.5, PM10 and O3 for evaluating the proposed model and done a comparative analysis with other air quality prediction models. Lo et al (2022) developed a new PM 2.5 prediction system that uses the dataset to resolve the missing value problem by incorporating the auto encoder and linear interpolation. They have used Spearman's correlation coefficient for extracting the more relevant features for PM 2.5 and also incorporated the LSTM and K-Means clustering method incorporated LSTM for predicting the PM 2.5 value for every regions. The K-Means clustering approach using LSTM is superior to the current ML and DL methods, they have demonstrated through comparison study.

Hashmy et al (2023) developed a modular air quality prediction system which is capable of resolving the issues of unreliability by applying the advanced ML techniques that leverages the potential of IoT architecture. Moreover, their system used to calibrate the data with additional features. IN addition, they have done a validation process and demonstrate the air quality monitoring process in a specific region. In order to estimate the air quality, Kefei et al. (2023) created a new air quality prediction system that combines the CNN with spatial and temporal information. The existing needs in terms of efficiency and effectiveness cannot be satisfied by the air quality forecast technologies that are now available. In order to do this, this study suggests a new air quality prediction system that can forecast the level of air pollution and provide alerts when it crosses a certain threshold.

The presently available air quality prediction systems are failed to fulfil the current requirements in terms of predicting the air quality efficiently and effective manner. The major challenge is to predict the air pollution level or air quality of the specific regional space and time duration. There is no system is considered the time as an important factor to make decisions over the air quality index values that are available as a standard dataset for the specific city. Generally, time is a key factor to predict the air quality for the specific regions. Moreover, the various Machine Learning and Deep Learning algorithms are incorporated in the air quality prediction model in the past. However, the available models are not obtained the required prediction accuracy with less time. For this purpose, this paper proposes a new air quality prediction system to predict the air pollution level and also provides alert when cross the level on time.

3. SYSTEM ARCHITECTURE

The user interface module, the air quality database, the rule base, the decision manager, the rule manager, the data pre-processing module, and the prediction module are among the ten components of the proposed model that are depicted and described in detail in Figure 1.

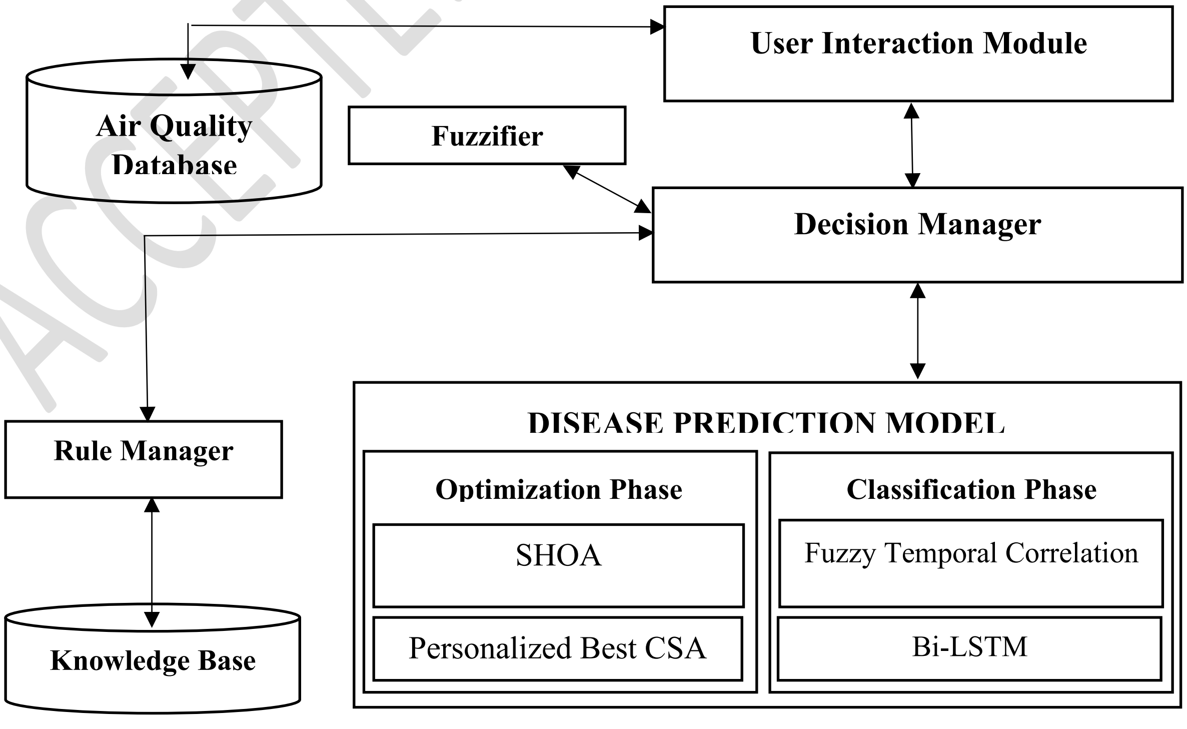


Figure 1. System Architecture

The relevant information is gathered by the user interaction module from the air quality database. The user asked the decision manager to make a choice. With a user interface module, the decision manager pulls the pertinent data from the diabetes dataset. The prediction module is used by the decision maker to process the data. The prediction module comprises two distinct phases, including the classification phase and the optimization phase. The feature optimization is done in the optimization phase utilising a recently proposed hybrid optimizer that combines SHOA and Personalized Best Cuckoo Search Algorithm (SHPCSA). In the classification phase, the optimized feature sets of records are considered as input for the classification module and it performs the classification process by using the deep learning algorithm. Before that, the relevant air quality datasets can be summarized by applying the existing clustering algorithm called K-means clustering for grouping the datasets effectively. Moreover, it classifies by applying a fuzzy temporal correlation and the Auto-encoder aware Bi-LSTM. The decision manager uses the fuzzifier to execute the fuzzification as well. When making judgements on air quality record sets, the decision manager employs the crucial information and guidelines included in the knowledge base.

4. RECOMMENDED MODEL

Using the recently developed Spotted Hyena- Personalized Best Cuckoo Search Optimization Algorithm (SHPCSA) and the Personalized Best Cuckoo Search Algorithm (PBestCSA) (Gandomi et al. 2013,Yang et al. 2009, Wang et al. 2016 and Yang et al. 2010), we propose a new air quality prediction system in this paper. A fresh fuzzy temporal correlation-based ensemble classifier is also recommended for effective classification. In this study, a bidirectional long short-term memory (Bi-LSTM) based auto encoder is coupled with a fuzzy temporal correlation-based classifier for effective categorization. After describing the feature optimization process in great depth, the classification approach is covered in this section.

The novelty of the proposed air quality prediction model are the introduction of new feature optimization algorithm called SHPCSA and the introduction of fuzzy temporal rules incorporated correlation based classifier along with the auto-encoder based Bi-LSTM for performing effective prediction process on time.

4.1 Feature Optimization

The newly proposed Spotted Hyena-Personalized Best Cuckoo Search Optimization Algorithm (SHPCSA), which is employed for efficient feature selection, is described in depth in this paragraph. It begins by providing background information on the Spotted Hyena Optimization Algorithm. Finally, it provides a thorough explanation of the recently introduced SHPCSA along with the procedures that are used in this work to achieve effective feature selection and optimization.

A. Spotted Hyena Optimization

Generally, the spotted hyenas are the huge carnivorous canines which are living in different kinds of dry environments and open environments. The different sizes like medium size and large size herbivores namely zebras, impala and wildebeests are preyed by spotted hyenas preys. Spotted hyenas are the social animals with intelligence and it can be identified by relatives applying the various senses. Moreover, they are ranked in race according to the relationships between them. In nature, the spotted hyenas are capable of achieving good success rate in the process of group hunting. It contains five different steps such as encircling prey, hunting, exploitation and exploration processes. Here, the encircling process selects the best candidate that is nearer to the target prey. Moreover, update the search agent locations with definition. In the hunting process, the spotted hyenas hunt in packs, rely on friends of trusted network and also capable of spot prey. For defining the behaviour of the spotted hyena, consider the best agent that is optimal location of the targeted prey. The location of the targeted prey is updated based on the best solution. The spotted hyena group attack the preys and change the search agent's location based on the best agent solutions. This SHO permits the search agents for updating the locations and the relevant attacks of the prey. In the exploration process, the SH's find key based on the SH's group. Moreover, they are located in different places for finding the prey of attacks. It permits the global search that makes the possible exploration processes. In addition, the random weights are assigned to every prey and also takes precedence for demonstrating the randomness that avoids the local optimization and the global search process. Especially, the local optimization is avoided in final iteration process and terminates them when satisfied the given condition.

B. Personalized Cuckoo Search

The conventional cuckoo search method makes advantage of the top global solution put into place for the coming generation. It performs the same functions as a PSO that manages the memory that is easily accessible on social media. Although this is the case, it does not maintain the RAM for recording each population detail independently. The cuckoo search method is used to replace the current answer when the new one outperforms the previous ones and there is no individual memory. The cuckoo search creates new solutions from the earlier solution's data with random methods. After employing Levy flights and solution switching processes to locate the new solution, the cuckoo search uses random ways to create new solutions from the data of the prior ones. This kind of solution lacks diversity due to a lack of information on the search space. Along with substituting for the switching parameter, the variable Pa completes the current cuckoo search, which designates a certain number of unsuccessful solutions that should be abandoned.

The convergence rate is dynamically adjusted by this parameter while simultaneously accounting for the limitations of the conventional cuckoo search method. For maintaining the lack of environmental data, the Enhanced Personal Best Cuckoo Search (EpBestCS) method, which deals with each person's best memory, is recommended. In the cuckoo search method, this kind of data is used to generate new solutions, which are known as solution creation stages. EpBestCS - employs personal best data as a result of the recently introduced system. The variations are demonstrated in the equations (1) and (2).

$$StepSize_i = 0.01.\left[\frac{Levy(\lambda).Rand_i}{Rand_i}\right]^{\frac{1}{x}}\left((x_i - \chi_{gbest}) + (pBest_i - x_i)\right) \tag{1}$$

$$\chi_{new} = x_i + (pBest_i - x_i) + Rand_1 \oplus H(P_a - Rand_2) \oplus (p_i - x_k) \tag{2}$$

Where, $pBest_i$ indicates that the personal best memory of a population ith individual. According to the weight adjustment procedure of PSO, the suggested EpBestCSA dynamically modifies the switching parameter or abandon rate [20]. The switching value in this study is high at the beginning of the iterations, which minimises the abandonment rate. As a result, maxPa and minPa are added as extra parameters in the recently proposed EpBestCS.

$$P_a = maxp_a - \frac{maxp_a - minp_a}{iter_{max}} \times k \qquad (3)$$

Where, $maxp_a$ indicates the maximum abandon rate and $minp_a$ represents the minimum abandon rate. The number k denotes the current iteration, whereas the variable 〖iter〗_max represents the maximum number of iterations. The cuckoo search incorporates the best data to support the exploitation process and constantly modifies the switching parameter to carry out the exploration process. By using the suggested EpBestCSA, the two features help maintain the proper balance between the exploration and exploitation processes, which is crucial for any optimization technique that has been proven to be effective.

C. Proposed Spotted *Hyena and Personalized Cuckoo Search Optimizer*

This subsection explains the Spotted Hyena based Personalized Cuckoo Search Algorithm (SHPCSA) in detail with necessary steps. This new optimization algorithm is helpful for selecting the most contributed features that are capable of enhancing the performance of the classifier.

Algorithm: Spotted Hyena and Personalized Cuckoo Search Optimizer (SHPCSO)

Input: Air Quality Dataset

Output: Optimised feature set

Phase 1:

Step 1: Perform the encircling prey using the optimal features.

Step 2:Find the most useful features by perform hunting process.

Step 3: Exploits the optimal features

Step 4: The exploration task is to be performed over the optimal features

Phase 2:

```
Cuckoo-Search (Pa, a, λ)
{
 Nests = n;
 While (i<=n)
  {
     Geteach and every Cuckoo (say i)
Calculate the fitness value for all features (Fi)
Identify a suitable group (nest) from N number of features that are selected randomly.
     IF Fi >Fj THEN
         Fi=Fj
         Pa is declared as unrestrained
         NA = Pa
Rank (groups)
         Best group (n)
   }
  Return (group)
}
```

The Improved Personalised Cuckoo Search Algorithm (EpBestCSA), which combines the SHOA and the proposed SHPCSO, is intended to improve the attributes used to categorise the contents. It typically contains two phases, such Phase 1 and Phase 2. The initial level feature selection is carried out in Phase 1, which is referred to as the Improved SHOA. The EpBestCSA stages are covered in phase 2, which is used to carry out the second level optimization in the suggested optimizer. The optimal featured record sets are grouped by using the K-Means clustering algorithm for further process.

In this proposed feature optimization algorithm, the feature selection and optimization are done effectively by using the existing optimization techniques. In phase 1, first it performs the encircling prey by applying the most important limited number of features. Then, it identifies the optimal features by initiating the effective hunting process. The optimal features are to be filtered by performing the exploration. Finally, the phase 1 will produce the optimal features that are suitable and useful for predicting the air quality in a specific region and time duration. In phase 2, the Cuckoo Search algorithm is applied for getting the further more optimal features from the set of features. Here, it considers all the features as cuckoo and calculates the fitness values. Then, it finds the suitable nest according to the fitness values of the features randomly. Now, the features are to be compared with each other features and rank them according to the groups and return as final optimal features that are useful for making effective decisions over the instances.

4.2 Classification

This section explains the newly proposed classifier called fuzzy temporal correlation-based ensemble classifier along with the necessary background information. First, it explains the fuzzy correlation with temporal constraints. Then, it explains the fuzzy temporal cognitive map and auto-encoder Bi-LSTM consecutively.

A. Fuzzy Correlation with Temporal Constraints

In the proposed prediction system, the fuzzy correlation values and temporal constraints are also considered to make final decision on disease dataset according to the work [8]. The co- efficient value of fuzzy correlation is taken consider in the process of classification to achieve better prediction accuracy. The methodology is explained as follows:

First, the $Xn = \{x1, x2, ... , xn\}$ and $Yn = \{y1, y2, \quad , yn\}$ are the independent observation then the correlation coefficient relation is derived from equation (4).

$r\ (t1, t2) = n\ i{=}1\ (Xi{-}\overline{\overline{X}})(Yi{-}\overline{\overline{Y}})$ (4)

$n\ I{=}1\ (Xi{-}\overline{\overline{X}})2\ \sum n\ (Yi{-}\overline{\overline{Y}})2$

Apply the fuzzification process with the variables Xn and Yn by applying the triangular membership function and also defined the non-membership defined in equations (5) and (6).

$\mu(x) < t1, t2 >= x{-}p\ if\ p \leq x \leq q$

$q{-}p\ r{-}x\ if\ q \le x \le (t1, t2)$ (5)

$(r{-}q\ |0 \qquad if\ x < p\ and\ x >\ (t1, t2)$

And $v(x) < t1, t2 >= \{q{-}xq{-}p{*}\ x{-}q\ r{*}{-}q\ if\ p\ {*}{\le}\ x \le q$

$if\ q \le x \le r{*}$ (6)

$if\ x < p\ {*}\ and\ x > r{*}$

where $p\ {*}{<}\ p < q < r < r\ {*}$ now, to perform fuzzification for the value r and it calculates the alpha-beta cut values of y as:

$(t1, t2) = [\{(t1, t2), (t1, t2) \ge \alpha\}, \{y(t1, t2), \mu y(t1, t2) \ge \alpha\}]$ (7)

$(t1, t2) = [\{(t1, t2), (t1, t2) \ge \beta\}, \{y(t1, t2), \mu y(t1, t2) \ge \beta\}]$ (8)

And from equations (3) and (4), this work has

$r\alpha(t1, t2) = [min\{\ r\ (t1, t2) \in [-1,1], r(t1, t2) \ge \alpha\ \}, max\{\ r(t1, t2), r(t1, t2) \ge \alpha\}]$ (9)

$r\beta(t1, t2) = [min\{\ r(t1, t2) \in [-1,1], r(t1, t2) \ge \beta\ \}, max\{\ r(t1, t2), r(t1, t2) \ge \beta\}]$ (10)

From equations (5) to (7), it is found that to calculate the correlation value (r') is easy as below:

$r'(t1, t2) = (r\alpha(t1, t2) + r\beta(t1, t2) - \gamma)\ def\mu + (1 - r\alpha(t1, t2) - r\beta(t1, t2) + \gamma) defv$ (11)

Where, $\gamma = (r\alpha, r\beta)$ indicates the contradictory methodology.

A meta-classifier, the newly suggested ensemble classifier combines two or more machine learning methods to forecast the disease using majority voting. Two voting processes, such as hard voting and soft voting, are combined in a soft voting categorization approach. In the hard voting technique, the final decision is made on predictive analysis by considered the majority voting that aggregate and chosen the class which appears repeatedly from the models. In the soft voting methodology, the base models must incorporate the Predict_proba technique. The voting classification method presents the better results than the base models which are combines the various models. In the proposed model, the Fuzzy Temporal Cognitive Map [6] and the auto encoding Bi-LSTM have been ensembled. The predict_proba feature is applied for every targeted feature and it mixed the training data and the points. These data points are passed to the classifiers FTCM and Bi-LSTM.

Every model calculates the voting aggregation value and also finds the yields the final disease prediction process. The steps of the proposed fuzzy temporal aware soft voting classifier are below:

```
Procedure Replace_data (dd)
Return dd [“pollution”, “air-quality index”].replace (‘0’, median()) Procedure DS (dd)      //      DS represents the data split
TRD, TSD = split (Pollution Feature, label)
Return (TRD, TSD) Voting =” soft”
M1= FTCM (TRD, TRL, TSD)        // TRL – Training Labels M2= SVM (TRD, TRL, TSD)
Procedure EM (TRD, TRL, TSD)
SVC = concatenate (M1, M2)// SVC – Soft Voting Classifier SVC.fit (TRD, TRL)
Predictions = SVC.predict(TSD)
```

The necessary rules were developed for predicting the air quality according to the air quality index values in prediction process.

B. Classification using Auto-encoding Bi-LSTM

This section explains the workflow of the Bi-LSTM in detail by demonstrating the necessary formulae that are used in this work for performing air quality prediction. This section explains the standard deep learning algorithms namely LSTM and Bi-LSTM that are used for predicting the air quality according to the work (Xinghan and Minoru 2021).

The three different gates that are useful for figuring out the states of the cells are discussed along with the LSTM method. Here, one of the layers in the neural network that may be represented by an equation is the forget gate, or equ. (12).

$$f_t = \sigma\ (W_f[h_{i-1}, X_t]\ +\ b_f) \qquad (12)$$

where σ stands for the weights, b_f stands for the biases, h_(i-1) stands for the previous output, X_t stands for the current input, and stands for the sigmoid function. Here, the input gate is viewed as an additional neural network capable of generating a new gate (f_t) with a formula identical to that in equation (2).

$$i_t = \sigma\ (W_i[h_{i-1}, X_t]\ +\ b_i) \qquad (13)$$

where the weights and biases are holding different values.

The candidate value is represented by using the variable C_t along with the values of h_{t-1} and X_t in the equation (13).

$$\tilde{C}_t =\ tanh\ (W_C[h_{i-1}, X_t] + b_C) \qquad (14)$$

Next, change the cell state C_t and also multiply the historical state C_{t-1} along with forget gate. Moreover, multiply the input gate along with candidate state $\tilde{C}_t$ and also add them together. Finally, it provides the structure which is given in equation (15).

$$C_t\ =\ f_t\ *\ C_{t-1} + i_t\ *\ \tilde{C}_t \qquad (15)$$

Still, the cell state is updated. Finally, the output of the LSTM according to the state of the cell that is the combination of cell state and output gate.

$$o_t = \sigma(W_o[h_{i-1}, X_t] + b_o)$$

$$h_t = o_t\ *\ tanh\ (C_t) \qquad (16)$$

The necessary features are trained by applying the backpropagation through time method that is shortly called BPTT which has backpropagation variant. This work has been implemented by using tensor flow and also applies an optimization method names Adam optimizer to learn the features of the dataset.

5. RESULTS AND DISCUSSION

This section explains the dataset, evaluation parameters, experimental results and comparative results. First, it explains the datasets used in this work in detail.

A. Datasets

The PM2.5 data is gathered from the standard database called Berkeley Earth (R. Muller and E. Muller). This database is developed by considering the data that are taken from various weather stations in different cities in all over the world. Moreover, the gathered data from the various provinces of Bangkok city, Thailand with 40000 observations from the year 2016. Their air quality data is divided into training dataset and testing dataset. Here, 90 percent of the data is considered for performing training process and the 10% of the data is used for performing testing process. The evaluation metrics of cross entropy loss is also considered while performing training process in various number of experiments with different epochs. Moreover, the air quality is predicted by using the IoT kit and create a database for the cities Chennai, Bangalore and Delhi in India. The accuracy, precision, recall, and other common metrics are used to assess the prediction system. In addition, the error rate and false alarm rate are also considered for evaluating the system.

B. Experimental Setup

The effectiveness and efficiency of the suggested disease prediction model, which makes use of Raspberry PI software, IoT devices, and the relevant source code, have been demonstrated through experiments.. Also, in order to transfer data to the disease prediction model, the IoT devices, along with the necessary hardware devices, a microcontroller, and a LoRa communicator, are all incorporated into the disease prediction model. Here, the Omron HeartGuid-bp8000m is used for measuring the blood pressure and the heart monitor board is also used to measure the heart rate. Here, the data read from the USB port connected to the Arduino Uno is stored in a text file. The carbon monoxide concentration from this text file is concatenated with the particulate matter concentrations, temperature and humidity read simultaneously by applying the file read options in Python programming language. This data is stored with time in the CSV file which is placed in the root folder of the lightly server installed. The lightly server supports HTML, CSS and JavaScript. D3.js is used to create data visualization of the data stored in the CSV file.

C. Performance Evaluation Parameters

The suggested approach measures success based on f1 score, recall, accuracy, and precision. Here, using the formulae provided in equations, the True Positive (TP), True Negative (TN), False Negatives (FN), and False Positives (FP) are taken into consideration while computing the prediction accuracy:

$$Accuracy = TP+TN \,/\, TN+TP+FP+FN \qquad (17)$$

$$Precision = TP/TP+FP \qquad (18)$$

$$Recall = \quad TP/TP+FN \qquad (19)$$

$$F1-Score = 2 \times Precision \times Recall/\ Precision + Recall \qquad (20)$$

These metrics are used in this work to evaluate the performance of the proposed air quality prediction model by computing precision, recall, f1-score, and accuracy from the values of true positives, true negatives, false positives, and false negatives.

D. Fuzzy Temporal Rules

Using the input air quality datasets, this work implements a set of fuzzy temporal rules for decision-making. All of these regulations were developed with the assistance of medical professionals and subject matter experts. This section offers a variety of guidelines for predicting air quality. The rules developed to accurately anticipate the quantity of air pollution are first listed.

E. Experimental Results

By taking into account all the aspects of the datasets, the air quality prediction procedure can be performed by taking assessment factors like precision, recall, and f1-score into consideration. Five different experiments consider the different set of records as input for performing the prediction process. Here, the randomly selected full featured records or instances are to be considered as input for the prediction model. In this figure 2, the Y-axis stated that the prediction accuracy. X-axis indicates the experiment numbers.

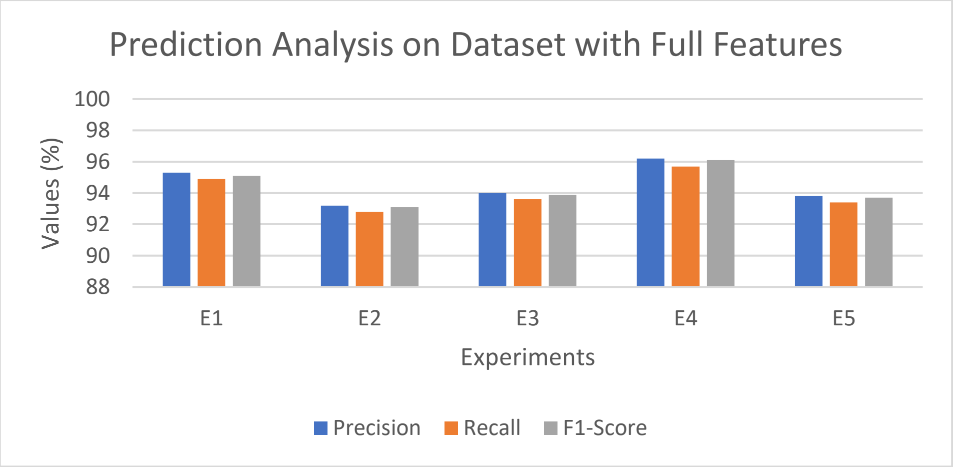


Figure 2. Prediction Analysis on Dataset with full features

Figure 2 demonstrates that the suggested air quality prediction model performs better across the board in terms of assessment criteria including precision, recall, and f1-score. In this case, the precision values are closer to the f1-score of all the studies. The accuracy of predictions is more influenced by these three factors. Due to the use of an efficient deep learning method with fuzzy temporal correlation, full featured datasets are taken into consideration and performed well.

Figure 3 indicates that the prediction analysis on datasets with fewer features by considering the precision, recall and f1-score. Five distinct experiments, including E1, E2, E3, E4, and E5, have been carried out to demonstrate the viability of the suggested air quality prediction model. For the five tests, several sets of records were taken into consideration. These five different experiments consider the different set of records as input for performing the prediction process. Here, the randomly selected optimal featured records or instances are to be considered as input for the prediction model. In this figure 3, the Y-axis stated that the prediction accuracy. X-axis indicates the experiment numbers.

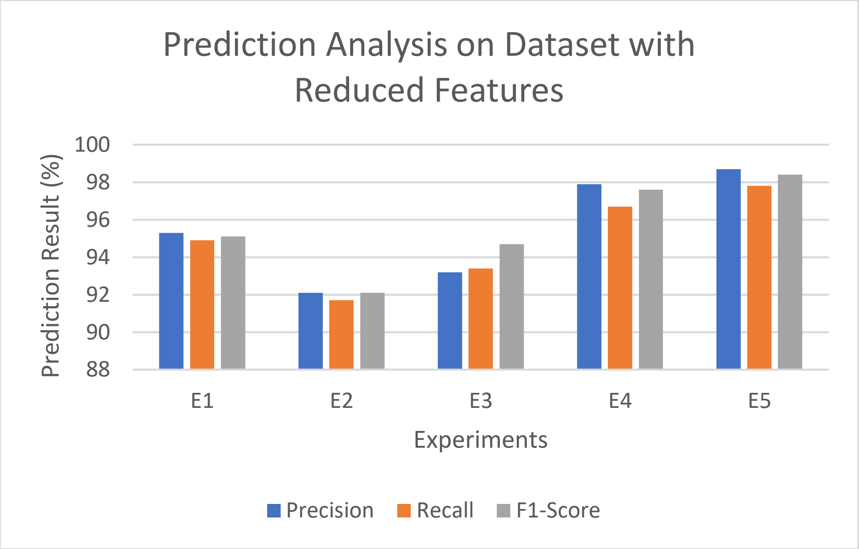


Figure 3. Prediction result analysis on datasets with selected features

The usefulness of the suggested disease prediction model on distinct sets of records taken into account in five separate tests is demonstrated in Figure 3. The suggested prediction model outperforms the full featured dataset in every assessment metric, including precision, recall, and f1-

score. The improvement was achieved by combining the proposed feature optimization technique with fuzzy temporal correlation-based Auto-encoding Bi-LSTM.

Table 1 compares the prediction accuracy between the complete featured dataset and the chosen featured dataset using assessment metrics including precision, recall, and f1-score. Two different types of information, including PM2.5 and live data gathered from several cities utilising IoT kit, were analysed in this study. Even with the same number of records being evaluated for experimental analysis, the performance varies between these two distinct types of datasets. Here, all three air quality datasets were taken into account, and the combined result was only taken into account for the original dataset and the highlighted dataset.

Table 1. Prediction Accuracy Analysis on UCI and Live Records

Datasets	**Overall Air Quality Prediction Accuracy (%)**	
	Original Dataset	**Selected featured Dataset**
Standard PM2.5 Datasets	99.41	99.92
Live patient records	98.32	98.23

The proposed air quality prediction model's accuracy was shown in Table 1 for both the original dataset and the live records. Furthermore, it is demonstrated that the prediction model performs better on real data than on benchmark datasets, both for the original full features dataset and for the selected features dataset. The utilisation of fuzzy temporal correlation, auto-encoding Bi-LSTM, and feature optimization accounts for the improvement.

The time taken for making decision on various datasets by the proposed air quality prediction model is shown in table 2. Here, the heart disease dataset is considered for performing time analysis on various datasets in the processes of training and testing. The same number of records only considered for time analysis.

Table 2. Time Analysis

Medical Datasets	Time Taken (Sec) for the proposed Air Quality prediction model		Time Taken (Sec) for the work proposed by Voravarun and Ferdin (2022)	
	Training	Testing	Training	Testing
Standard Dataset	0.39	0.21	0.42	0.23
Streaming Dataset	0.38	0.18	0.43	0.20

To carry out efficient training and testing procedures, Table 2 shows the temporal analysis between the proposed air quality prediction model and the current air quality monitoring system. The use of an efficient feature optimization technique, which aids in limiting the selection of features to those that have contributed the most, is the cause for the improved performance. It takes less time than the current disease prediction system to undertake training and testing operations since just around one-fifth of the records are chosen and taken into account.

The prediction accuracy analysis between the domain expert and the proposed air quality prediction model is shown in figure 4 that considers the various cities air quality datasets.

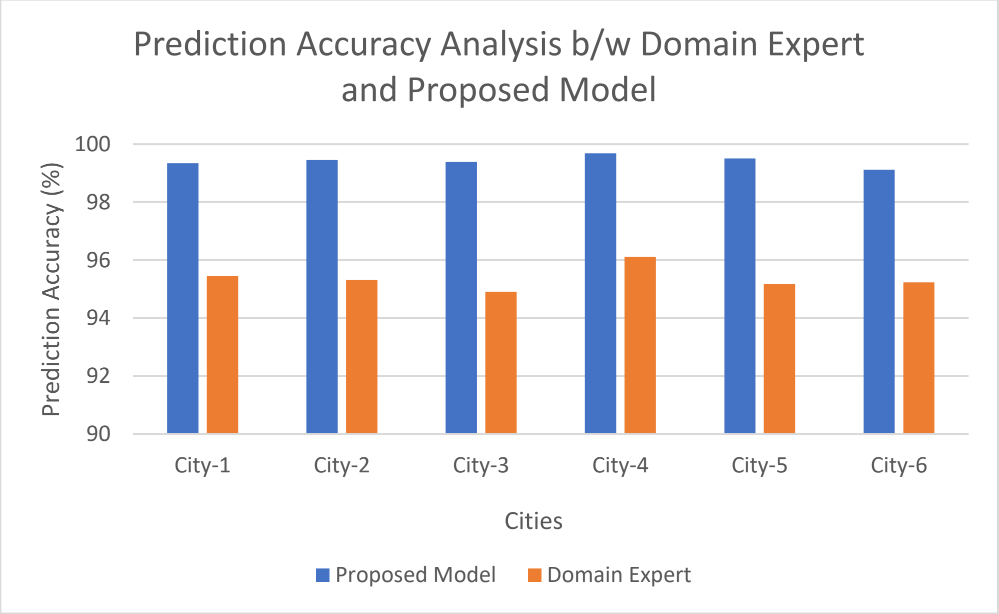


Figure 4. Prediction Accuracy analysis

Figure 4 demonstrated that the proposed air quality prediction model outperformed the judgements of domain experts on the various pollution levels that were recorded in the benchmark datasets' records. By taking into account the 1000 records in each category, there is a disparity of about 2-3%

between the domain expert result and the suggested air quality prediction model. The use of efficient feature optimization and the fuzzy temporal correlation-based Bi-LSTM is what has caused the performance to increase.

Figure 5 compares the performance of the proposed air quality prediction model with the existing prediction systems proposed by Voravarun and Ferdin (2022), Jujie et al (2022), Marjan et al (2022), and Gokul et al (2022) in terms of prediction accuracy, which is determined by using the evaluation metrics such as precision, recall, and f1-score (2023). Five experiments—E1, E2, E3, E4, and E5—have been conducted in this study to evaluate the suggested air quality prediction model while accounting for different sets of records. The combined datasets of the several cities taken into consideration in this work are used to calculate the overall prediction accuracy.

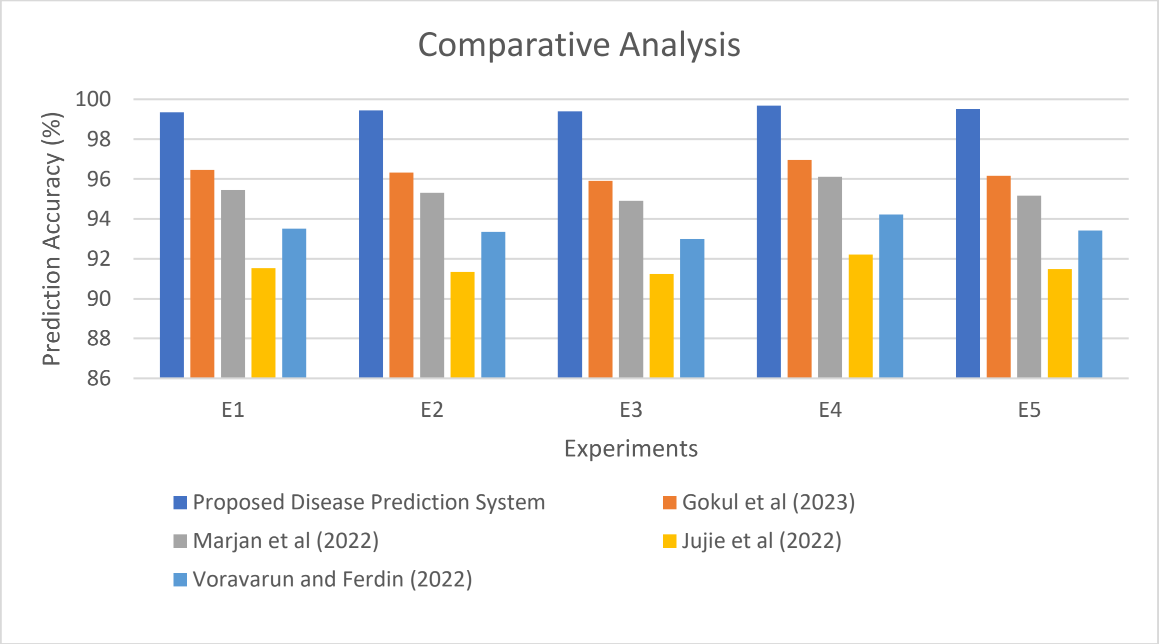


Figure 5. Comparative Analysis

When the suggested air quality prediction model was compared to the current air quality prediction models created by Voravarun and Ferdin (2022), Jujie et al (2022), Marjan et al (2022), and Gokul et al., Figure 5 demonstrated the effectiveness of the proposed model in terms of prediction accuracy (2023). The use of an efficient feature optimization technique that combines the Spotted Hyena Optimization Algorithm and the Personalized Best Cuckoo Search Optimization Algorithm, as well as a newly proposed classifier that uses new fuzzy temporal correlation and the existing

Auto-encoding based Bidirectional Long Short-Term Memory, are what's responsible for the performance improvement (Bi-LSTM).

6. CONCLUSION AND FUTURE WORKS

In order to accurately estimate the degree of air pollution, a new system for predicting air quality has been developed and put into use in this study. The proposed system employs the recently developed feature optimization algorithm known as Spotted Hyena-Cuckoo Search Optimization Algorithm (SHCSOA), which combines the Spotted Hyena Optimization Algorithm (SHOA) and the Personalized Best Cuckoo Search Algorithm, to find and select the most contributed features (PBestCSA). A fresh fuzzy temporal correlation-based ensemble classifier is also recommended for effective classification. The auto-encoders used to create this bidirectional long short-term memory have fuzzy temporal correlation (Bi-LSTM). The experimental results showed that the proposed air quality prediction system is superior to the existing systems in terms of precision, recall, f1-measure, and forecast accuracy. The creation of a lightweight optimizer and classifier for using illness datasets as decision-making inputs will enable future work in this approach.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

DATA AVAILABILITY STATEMENT

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an on-going study.

REFERENCES

Adil Masood. and Kafeel Ahmad, (2021), A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance, *Journal of Cleaner Production*, **322**, 129072, 1-10.

Amir Talaei-Khoei. and James M. Wilson. (2018), Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables, *International Journal of Medical Informatics*, **119**, pp. 22-38.

Atakan Kurt. et al. (2008), An online air pollution forecasting system using neural networks, *Environment International*, **34(5)**, 592-598.

Fatemeh Mansourypoor. and Shahrokh Asadi. (2017), Development of a Reinforcement Learning-based Evolutionary Fuzzy Rule-Based System for diabetes diagnosis, *Computers in Biology and Medicine*, **91**, 337-352.

Ganapathy S. *et al.* (2014), An intelligent temporal pattern classification system using fuzzy temporal rules and particle swarm optimization, *Sadhana*, **39(2)**, 283-302.

Gandomi A. H., Yang X. S. and Alavi A. H. (2013), Cuckoo search algorithm: A metaheuristic approach to solve structural optimization problems, *Eng. Comput.*, **29**, 17–35.

Gokul P. R. *et al.* (2023), Spatio-temporal air quality analysis and PM2.5 prediction over Hyderabad City, India using artificial intelligence techniques, *Ecological Informatics*, 1-17.

Gu K. *et al.* (2022), Air Pollution Prediction in Mass Rallies with a New Temporally-Weighted Sample-Based Multitask Learner, *IEEE Transactions on Instrumentation and Measurement*, **71,** 1-15.

Hashmy Y. *et al.* (2023), Modular Air Quality Calibration and Forecasting Method for Low-Cost Sensor Nodes, *IEEE Sensors Journal*, **23(4)**, 4193-4203.

Hassan Kazemian. *et al.* (2016), NN approach and its comparison with NN-SVM to beta-barrel prediction, *Expert Systems with Applications*, **61**, 203-214.

Heming Jia. (2019), Spotted Hyena Optimization Algorithm with Simulated Annealing for Feature Selection, *IEEE Access*, **7**, 71943-71962.

Hongmin Li. *et al.* (2022), Air quality deterministic and probabilistic forecasting system based on hesitant fuzzy sets and nonlinear robust outlier correction, *Knowledge-Based Systems*, **237**, 107789, 1-12.

Ioannis Kavakiotis. *et al.* (2017), Machine Learning and Data Mining Methods in Diabetes Research, *Computational and Structural Biotechnology Journal*, **15**, 104-116.

Jujie Wang. *et al.* (2022), A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization, *Chaos, Solitons & Fractals*, **158**, 112098, 1-12.

Kefei Zhang. *et al.* (2023), Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning, *Environment International*, **171**, 107691, 1-17.

Krishna Rani Samal K., Korra Sathya Babu. and Santos Kumar Das. (2022), Multi-output Spatio-temporal air pollution forecasting using neural network approach, *Applied Soft Computing*, **126**, 109316, 1-12.

Liu B. et al. (2021), A Spatiotemporal Recurrent Neural Network for Prediction of Atmospheric PM2.5: A Case Study of Beijing, *IEEE Transactions on Computational Social Systems*, **8(3)**, 578-588.

Lo C. Y. *et al.* (2022), Recurrent Learning on PM2.5 Prediction Based on Clustered Airbox Dataset, *IEEE Transactions on Knowledge and Data Engineering*, **34(10)**, 4994-5008.

Ma J. *et al.* (2019), Spatiotemporal Prediction of PM2.5 Concentrations at Different Time Granularities Using IDW-BLSTM, *IEEE Access*, **7**, 107897-107907.

Marjan Asgari, Wanhong Yang. and Mahdi Farnaghi. (2022), Spatiotemporal data partitioning for distributed random forest algorithm: Air quality prediction using imbalanced big spatiotemporal data on spark distributed framework, *Environmental Technology & Innovation*, **27**, 102776, 1-13.

Masoomeh Zeinalnezhad. *et al.* (2020), Air pollution prediction using semi-experimental regression model and Adaptive Neuro-Fuzzy Inference System, *Journal of Cleaner Production*, **261**, 121218, 1-12.

Mokhtari W. *et al.* (2021), Uncertainty-Aware Deep Learning Architectures for Highly Dynamic Air Quality Prediction, *IEEE Access*, **9**, 14765-14778.

Namrata Singh. and Pradeep Singh. (2020), Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus, *Biocybernetics and Biomedical Engineering*, **40(1)**, 1-22.

Ping Jiang. *et al.* (2019), An innovative hybrid air pollution early-warning system based on pollutants forecasting and Extenics evaluation, *Knowledge-Based Systems*, **164**, 174-192.

Piotr S. *et al.* (2019), Air pollution prediction with clustering-based ensemble of evolving spiking neural networks and a case study for London area, *Environmental Modelling & Software*, **118**, 262-280.

Qi Z. *et al.* (2018), Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality, *IEEE Transactions on Knowledge and Data Engineering*, **30(12)**, 2285-2297.

Sethukkarasi R. *et al.* (2014), An intelligent neuro fuzzy temporal knowledge representation model for mining temporal patterns, *Journal of Intelligent & Fuzzy Systems*, **26(3)**, 1167-1178.

Sheen Mclean Cabaneros. and Ben Hughes. (2022), Methods used for handling and quantifying model uncertainty of artificial neural network models for air pollution forecasting, *Environmental Modelling & Software,* **158**, 105529, 1-13.

Voravarun Pattana-Anake. and Ferdin Joe John Joseph. (2022), Hyper Parameter Optimization of Stack LSTM Based Regression for PM 2.5 Data in Bangkok, *7th International Conference on Business and Industrial Research (ICBIR)*, 13-17.

Wang G. G. *et al.* (2016), Chaotic cuckoo search, *Soft Computing*, **20**, 3349–3362.

Yang X. S. and Deb S. (2009), Cuckoo search via Lévy flights, *In Proceedings of the 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), Coimbatore, India*, 210–214.

Yang X. S. and Deb S. (2010), Engineering optimisation by cuckoo search, *Int. J. Math. Model. Numer. Optim.*, **1**, 330–343.

Yi X. *et al.* (2022), Predicting Fine-Grained Air Quality Based on Deep Neural Networks, *IEEE Transactions on Big Data*, 8(5), 1326-1339.

Yoichi Hayashi. and Shonosuke Yukita (2016), Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset, *Informatics in Medicine Unlocked*, **2**, 92-104.

Yue-Shan Chang. *et al.* (2020), An LSTM-based aggregated model for air pollution forecasting, *Atmospheric Pollution Research*, **11(8)**, 1451-1463.

Yuval. *et al.* (2012), Exploring the applicability of future air quality predictions based on synoptic system forecasts, *Environmental Pollution*, **166**, 65-74.

Zhenni Ding. *et al.* (2022), A forecasting system for deterministic and uncertain prediction of air pollution data, *Expert Systems with Applications*, **208**, 118123, 1-12.