



An Adaptive Whale Optimization Algorithm Guided Smart City Big Data Feature Identification for Fair Resource Utilization

Kapil Sharma, Sandeep Tayal

Abstract: World improvement is the development of every single province of the world. Smart city implies changed hardware to adjusted individuals. Smart cities have the most indispensable part in altering distinctive regions of human life, touching segments like transportation, wellbeing, vitality, and instruction. Productively to make measurements to improve distinctive smart city benefits huge information frameworks are put away, prepared, and mined in smart cities. For the change and course of action of huge information applications for smart cities, different difficulties are faces. In this paper, we propose a wrapper display based ideal element recognizable proof calculation for ideal use of assets given highlight subset age. Nine component determination techniques used for compelling element extraction. At last, which includes best add to the ideal usage of assets got by means of a novel element recognizable proof calculation made by the application out of a Whale Optimization Algorithm with Adaptive Multi-Population (WOA-AMP) system as inquiry process in a wrapper display driven by the notable relapse demonstrate regression model Random Forest with Support Vector Machine (RF-SVM). Our proposed calculation gives the exact method to choose the most agreeable feature blend, which prompts ideal asset usage.

Keywords: Feature selection, Whale Optimization Algorithm (WOA), Adaptive Multi-Population (AMP), Random forest, Smart city.

I. INTRODUCTION

As another sort of reasonable improvement, the thought "Smart City" knows a tremendous expansion among the current years [1]. The little and substantial regions are proposing another city display, called "the smart city," which speaks to a gathering of normal innovation estimate, alluring and secure, agreeable, interconnected and doable [2]. The innovative Internet of Things (IoT) solutions is empowering Smart City practices the overall world. It empowers to control contraptions and manage, remotely screen, and to make novel encounters and important information from huge surges of steady data [3]. The basic characteristics of a brilliant city join an irregular condition of information development compromise and an expansive usage of information resources [4].

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Kapil Sharma*, Faculty of Engineering and Technology, M. D. University, Rohtak (Haryana), India. Email: kapil@ieee.org

Sandeep Tayal, Assistant Professor, Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Delhi, India. Email: tayal.mait@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

The fundamental parts of urban change for a splendid city should fuse shrewd industry, keen development, keen organization, wise life and splendid organizations [5].

Each part of an economy of nations is changed by using smart city big data. Such change engages urban communities to finish the learning norms and necessities of the employments of the savvy city by understanding the essential keen condition characteristics [6]. The smart city abuses preferred standpoint of rising advancements, for instance, wireless sensor network (WSN), to limit asset utilization and cost [7]. The rising advancements among one with the enormous probable to improve smart city administrations are huge information investigation [8]. At the exhibit, a great deal of information is being made from different information sources like PCs, advanced mobile phones, cameras, sensors, interpersonal interaction destinations, worldwide situating frameworks, diversions and business exchanges [9]. The dataset made in the current world of digitizing creates, powerful information stockpiling and handling offices have presented challenges on the logical stages and traditional information mining [10] [11]. The use of huge information in a smart city has numerous points of interest and difficulties, together with the openness of substantial mathematical and storerooms to practice floods of information delivered inside a brilliant city condition [11][12].

Data Computation and Processing Phase (DCPP) takes after nine element feature techniques for successful component extraction. They are DIA affiliation factor, Chi measurement, Information Gain, Document Frequency, OCFS, DPFS, Comprehensive Measurement Feature Selection, Mutual Information, and enhanced Gini record [13],[14],[15]. The huge test here is we need to distinguish which technique is appropriate for particularly recognizing the highlights.[38] To defeat this test, we have required a viable approach in light of recognized highlights for ideal usage of assets.[37]

The upcoming sections of the research paper contains, section 1 contains the introduction of the smart city feature selection, section 2 has the recent related techniques as a literature survey, section 3 contains the proposed Whale optimization with Adaptive Multi-Population (WOA-AMP), section 4 represents the results of the anticipated system and section 5 provides the conclusion of the proposed method.



II. RELATED WORKS

Yintong Wang et al. [16] proposed a proficient Semi-administered Representatives Feature Selection calculation given data hypothesis (SRFS).

SRFS has a place with channel approaches and has performed especially well in evacuating the unimportant highlights and bunching the excess highlights. At first, they built up importance pick up the system through which the pertinence of highlights can estimate in the unlabelled information. Promote they presented the parcel of the guided non-cyclic diagram to group the repetitive highlights. At long last, they expanded the current Markov blanket calculations to abuse the extra data entropy contained in the unlabelled information. Also, they found, pertinence picks up measure is an essential advance to investigate the data of unlabelled information, and lessening the time unpredictability.

Ahamad Tajudin Khader and Laith Mohammad Abualigah et al. [17] exhibited a mixture of PS (particle swarm) improvement calculation through administrators of genetic for the component feature issue. The adequacy of the got highlights subsets assessed by using the k-means bunching. The element choice issue is defined as an improvement issue to locate an ideal subset of useful content highlights. Besides, it dispenses with uninformative highlights. The outcomes demonstrated that the proposed technique diminishes the number of uninformative highlights as well as essentially improves the execution of the content clustering calculation; the accomplished outcomes are similarly superior to the aggressive strategies.

Kuan-Cheng Lin et al. [18] proposed a Feature choice [39] strategy in light of an enhanced particle swarm advancement calculation for enormous information arrangement. The proposed calculation is known as 'Improved CSO' (ICSO) and adjusted starting the amicable particle swarm advancement. They led investigations to decide if the two proposed techniques enhance the execution of conventional CSOs when connected exclusively. The execution of the general ICSO in include choice was assessed utilizing an SVM. They utilized different datasets to encourage correlations of ICSO and CSO and likewise led a trial to research the impacts of fluctuating SVM restrictions on CSO and ICSO yields.

Sina Tabakhi et al. [19] displayed an unsupervised component choice strategy given subterranean ant colony optimization, called UFSACO. The technique looks to locate the ideal element subset through a few emphases without utilizing any learning calculations. Additionally, the element importance will be processed in light of the comparability between highlights, which prompts the minimization of the repetition. Along these lines, it can be named a channel based multivariate strategy. The proposed technique has a low computational many-sided quality; along these lines, it can be connected for high dimensional datasets. The test comes about on a few habitually utilized datasets demonstrate the productivity and adequacy of the UFSACO technique and additionally enhancements over past related strategies.

Simon Fong et al. [20] proposed a novel lightweight component determination calculation, especially to mine gushing information on the fly, and it is utilized the APSO

(accelerated particle swarm optimization). The assessment comes about demonstrated that the incremental technique acquired a higher pick up in exactness every second caused in pre-handling. Specifically, APSO is intended to utilize for information mining of information streams on the fly. A combinatorial blast is tended to utilized swarm seek approach connected incrementally. This approach likewise fits better with genuine applications where their information land in streams.[36] What's more, an incremental information mining approach is probably going to take care of the demand for enormous information issue in benefit registering.

III. FEATURE SELECTION BASED WHALE OPTIMIZATION USING WRAPPER BASED METHOD

The way toward disposing of the insignificant features and excess features from the database is known as features choice, which is used to improve the learning algorithm. The two criteria, named as assessment and search criteria feature collection approaches arranged. By utilizing the wrapper and channel approaches the chose feature subsets assessed. Algorithm of Learning is utilized as a part of the wrapper strategy in the determination procedure [21].

Big data smart cities [34] reasonable usage of assets is the fundamental goal of our proposed strategy. For this activity, numerous improvement strategies are utilized. In our proposed article, whale enhancement technique is used to get the reasonable used ideal list of capabilities.[35] By utilizing this ideal technique list of capabilities to be chosen and decrease the computational period; additionally, the forecast institution of metrics has the preferred qualities over the current strategies.

3.1 Wrapper Method

The feature selection in the basis of novel Whale Optimization Algorithm is working based on wrapper-based strategy. The usage of the classifier as in the form of the controller of the feature selection system is the foremost feature of the wrapper strategies. The subsequent three main schemes are used in the feature selection in wrapper [22]:

- Classification Scheme
- Criteria of Feature Evaluation
- Search Scheme

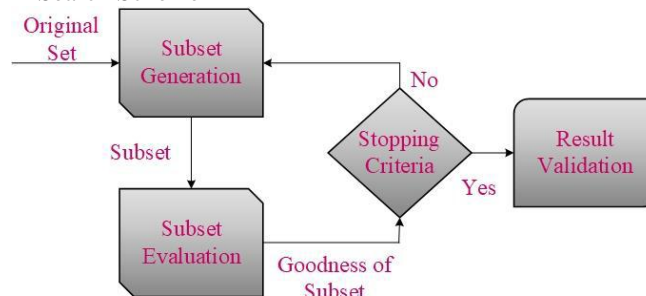


Fig 3.1 Process of Feature Subset Selection

A Subset Generation is a heuristic pursuit process where look for space contains states, each one of which demonstrates a cheerful subset for evaluation. Two things must be settled for subset age, Search starting stage, and Search framework.

Hunt starting stage can be forward, backward, bi-directional, and unpredictable. An interesting framework must pick the contender subsets [6]. An Evaluation display is used to survey each as of late made candidate. In perspective of the dependence on learning calculations that will be associated with pick feature set an evaluation demonstrate is arranged into two social affairs, one is needy criteria second one is self-ruling criteria. Wrapper demonstrate uses subordinate criteria, and for feature choice, it needs a learning computation. Fairness of feature or its subset is assessed with the help of critical features of the arrangement data without associating any learning algorithm.

In include subset decision, a stopping model speaks to when the feature decision process will stop. A bit of the routinely used halting criteria are according to the accompanying:

- Exhaustive search finishes.
- If a progressive expansion or evacuation of any element does not influence comes about component determination process could be ceased.
- If the agreeably decent subset chosen.

Toward the end, comes about are approved by utilizing grouping blunder rate of classifiers as an execution marker. Examinations are led to compare the characterization error rate on the full arrangement of the classifier learned on highlights, and that prepared on the chose include subset [8].

Wrapper based component subset assessment strategies prompt learning calculations amid assessment venture to calculate the decency of a chose include subset given the calculation's precision so are computationally costly when contrasted with channels. As far as prescient or arrangement, exactness wrapper strategies viewed as better than the channel. The utilized element order criteria or target work in the wrapper highlight determination is regularly mirroring the grouping execution and also the number of highlights. A non-specific portrayal of the wellness work speaking to for both order execution and the number of chose includes as portrayed in condition (10) In wrapper-based techniques, the single assessment of guaranteed arrangement is exorbitant as it generally applies preparing and testing of the given more tasteful. Along these lines, the proficient choice of the look technique is fundamental.

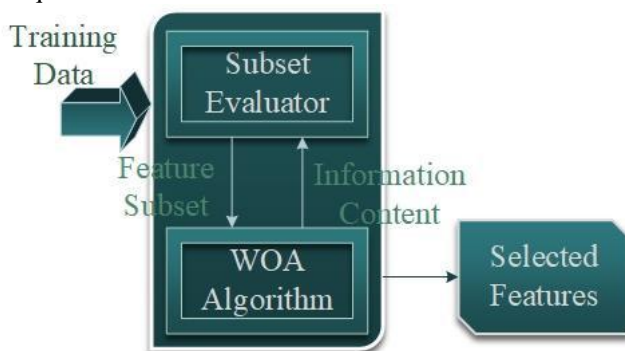


Fig 3.2: Wrapper Approach for Feature Selection

The wrapper is a moderate component determination strategy. As per the reliance standard embraced by the wrapper display, it requires a foreordained acceptance calculation and utilizations its execution measure to be connected on the chosen subset to figure out which highlights are chosen. It by and large gives better execution as it chooses includes more

qualified to the fated acceptance calculation. However, it is computationally more costly and moderate, which may not be appropriate for other acceptance calculations once in a while. The highlights chose by the classifier are utilized to anticipate the class marks of covered up occurrences; exactness is typically the evaluator of the produced subsets, however, cost high to figure precision for each element subset. Consequently, in the wrapper technique, it is imperative to distinguish the choice strategy most appropriate for the given dataset and characterization strategy. Besides, the subset of highlights chose is the negligible containing all the pertinent highlights for the objective class. Fig 3.2 shows the wrapper-based feature selection approach. The training set produced to subset evaluator. The dataset is too big, so the set is given to the optimization algorithm to evaluate the optimal dataset. The reduced dataset value is processed under the correlation process to get the optimal feature set. The most relevancy data considered as the optimal set and the other values are given back to the subset evaluator.

3.2 WOA

Here the WOA is said to be a fresh optimization algorithm. The WOP algorithm used to mimic the Humpback whales natural behavior. Hunting behavior is the main character of these whales in the path of survival. This technique is to address the optimization problems, and the hunting strategy has been introducing and to chase the prey in the search space; the capability to employ the best agent or random is the whale algorithms uniqueness. By using spirals, humpback whales bubble-net attaching mechanism is simulated [23].

The demonstrating of this calculation incorporates three administrators to mimic the scan for prey (investigation stage), the surrounding prey, and the air pocket net searching (abuse stage) conduct of humpback whales. The numerical definition is displayed and clarified as takes after:

• **Prey Encircling:** By using the best hunt operator whale calculation begins here. The present activities are the best and also whether it is the area of the prey or near this expects it. Whatever remains of the operators subsequently refresh their areas in the direction of the finest pursuit specialist.

$$\vec{P} = \left| \vec{B} \cdot \vec{x}^*(t) - \vec{x}(t) \right| \quad (1)$$

$$\vec{x}(t+1) = \vec{x}^*(t) - \vec{A} \cdot \vec{P} \quad (2)$$

Here, a recent repetition of the method represented as t ; coefficient vectors are denoted as \vec{A} and \vec{B} . The best solution in the location vector represented as \vec{x}^* , and the location vector is \vec{x} . \vec{x}^* must be updated iteratively for the actuality of better solution. \vec{A} and \vec{B} Vectors are deliberate in the following expression:

$$\vec{A} = 2 \cdot \vec{a} \cdot \vec{r} - \vec{a} \quad (3)$$

$$\vec{B} = 2 \cdot \vec{r} \quad (4)$$

where, over the number of iterations from 2 to 0, \vec{a} is linearly decreased and random vector in [0, 1] denoted as \vec{r} . The above demonstrating gives some operator to refreshing that one area in the district of the present finest arrangement also emulates enclosing the victim. For n measurements it also encourages the look space, also the specialists encourage the hypercubes development using the about the finest arrangement accomplished.

Exploitation Phase: It is also known as bubble-net attacking also two tactics are used to work the task:

Mechanism of Shrinking encircling: here, based on the eqn

3, \vec{a} value is reduced and subsequently the changeability range of \vec{A} is also reduced by \vec{a} . it indicates that \vec{a} is arbitrarily positioned in $[-\vec{a}, \vec{a}]$. Here over the optimization time from 2 to 0 the value of \vec{a} is reduced. Where, The unpredictability of \vec{A} in $[-1, 1]$, determine the search agents new location next to the best current location and past agent location.

▪ **Spiral updating position:** The distance amongst the locations of whale and its prey is estimated. By using humpback whales of helix shape, the spiral equation is formed among prey locations also whale to enhance the movement. It exposed in the following equation:

$$\vec{x}(t+1) = \vec{P} \cdot e^{bl} \cdot \cos(2\Pi l) + \vec{x}^*(t) \quad (5)$$

$$\vec{P} = \left| \vec{x}^*(t) - \vec{x}(t) \right| \quad (6)$$

The t^{th} whale distance is given in eqn 6, and the logarithmic spiral shape constant is denoted as b , and in the region of $[-1, 1]$ random number is denoted as l . It has a whale movement in the direction of its prey is simultaneously spread over the shrinking circling, and it placed in a spiral-shaped path. Consequently, the whale's next position updated by the 50% postulation of the opportunity to control between the binary modes.

Algorithm: 1
Input: $D(f_1, f_2, \dots, f_{|M|})$ // a dataset D with $|M|$ features
 S_n // an initial subset of features
 δ // a stopping criterion
Output: S_n^* // a (sub)-optimal set of features
Begin
 Initialize $S_n^* = S_n$;
 $J(S_n^*) \leftarrow \text{Evaluate}(S_n^*, D, L)$; // Evaluate goodness of fit of S_n^* on D w.r.t. learning task L .
Step 1: Start the Whale Population $X_i(t=1, 2, 3, \dots, n)$.
Step 2: Calculate the fitness of each whale.
Step 3: Set the best whale is X^* .
Step 4: While ($t < \text{maximum number of iterations}$) do
 for (each search whale) do
 Update a, A, C, l and p .
 if ($p < 0.5$) then
 if ($|A| < 1$) then
 Updating whale position is by Eq. (1).
 else
 if ($|A| \geq 1$) then
 Choose the random whale X_{rand} .
 Updating whale position is by Eq. (9).
 end
 end
 else
 if ($p \geq 0.5$) then
 The whale position is modifying by the Eq. (5).
 end
 end
 end
 Verify if any search agent goes beyond the search space and amend it.
 Calculate each search agent's fitness.
 if there is a better solution update X^* .
 $t = t + 1$
 end
 while δ not reached

Fig 3.3: Algorithm of WAO with Feature Subset Selection

$$\vec{x}(t+1) = \begin{cases} \vec{x}^*(t) - \vec{A} \cdot \vec{P} & \text{if } p < 0.5 \\ \vec{P} \cdot e^{bl} \cdot \cos(2\Pi l) + \vec{x}^*(t) & \text{if } p \geq 0.5 \end{cases} \quad (7)$$

where, p is a random number in $[0, 1]$.

Exploration Phase: A global optimization is attaining in

WOA in this phase. To accommodate the search agent \vec{A} is set casually from $[-1, 1]$ and which is used to retreat as of the reference whale. It means the value of \vec{A} ranges from either less than the -1 or greater than 1. Search agent location simplified through choosing an agent in randomly and that permits WOA to execute the global search.

The mathematical expression of the displaying of this assessment appliance given below:

$$\vec{P} = \left| \vec{C} \cdot \vec{x}_{rand} - \vec{x} \right| \quad (8)$$

$$\vec{x}(t+1) = \vec{x}_{rand} - \vec{A} \cdot \vec{P} \quad (9)$$

Where random location for the arbitrary whale denoted as

\vec{x}_{rand} and that preferred since the present residents. The

WOA pseudo-code presented in Algorithm 1.

3.3 Feature selection of WOA

Whale optimization algorithm approach utilized for consist of optimal in a wrapper-form strategy. The significant standard for the wrapper-based technique is used to put on the order of methodology as a manual for determination strategy includes some enhanced part which chose the feature set. This article applies the WOA adaptively for the ideal element subset to discover that expanding order execution. With an underlying best pursuit operator, the whales ceaselessly modify their areas to any point in the space begins the suitable operator in the WOA.

As in conditions (1) and (2) their work to refresh the situations toward the best pursuit specialist from that forward point.

With a similar measurement in the dataset, the individual arrangement portrayed as a constant vector. The arrangement vector esteems limited to $[0, 1]$ and constant. While the estimations of arrangement fitness calculation are binary signified. A connected fitness work is normally coordinating several chose features and the grouping execution. This can be spoken to the following condition.

$$f_{\theta} = \beta \cdot H + (1 - \beta) \frac{\sum_i \theta_i}{N} \quad (10)$$

Based on the above equation, f_{θ} is represented as a function of fitness specified a vector θ , N sized with $1/0$ components, which is elaborating the unselected or selected feature subset. In the database, the whole amount of features described as N . From the given selected feature subset, H is represented as a classification error. β is a constant. Finally, the constant handling the trade-off amongst classification errors to feature subset quantity. The organization enactment is the main objective, so $\beta=1$ use in this work.

3.4 Regression Model Random Forest with Support

3.4.1 Vector Machine (RF-SVM):

Random Forest:

This procedure comprises of an outfit of various relapse trees. Random forest regression is a tree-based strategy that includes stratifying or portioning the predictor space into various straightforward locales. To make a forecast for a given perception, the mean of the reaction estimations of the preparation perceptions in a similar area is regularly connected [24].

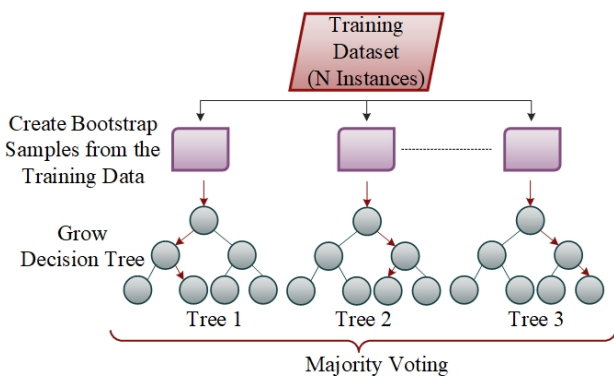


Fig 3.4: Random Forest Pictorial Representation

It is the piece of programmed learning technique. "Bagging" and irregular subspaces ideas consolidated by utilizing this calculation. The choice tree forest calculation prepares on numerous choice trees driven on somewhat unique subsets of information.

The arbitrary backwoods is a piece of the set strategies that takings the choice tree as a separable indicator, and depend on the techniques for Random Subspace pardoning boosting, Randomizing Outputs, and Bagging [25]. The calculation of random forest is truly outstanding among

characterization calculations - ready to order a lot of information with precision. It is an outfit learning technique for characterization and relapse that builds various tree choice at preparing period and conveys the class that is the method of the programs yield through singular trees.

Algorithm: 2

```

For  $b = 1$  to  $B$  Make
    • Draw  $Z^*$  of size  $N$ .
    • Construct a random forest tree  $R_f$  to the initial condition, through the repeatedly reconstructing the associated steps for every end child of the tree, till the point that the base hub evaluates  $n_{\min}$  is makes to a minimum.
        • From the  $p$  variables, select the  $m$  variables randomly.
        • Among the  $m$  pick the finest variable/split-point.
        • Divide the node
    • Output the group of trees  $\{T_{b_1}^B\}$ .
    
```

Fig 3.5: Algorithm of Random Forest

The prediction making at the novel point x is denoted by:

Regression:

$$\hat{F}_{rf}^B(X) = \frac{1}{B} \sum_{b=1}^B T_b(X) \quad (11)$$

Classification:

The b^{th} random forest tree prediction is assumed by $\hat{C}_b(x)$.

$$\hat{C}_{rf}^B(X) = \text{majority vote} \left\{ \hat{C}_b(X) \right\}_1^B \quad (12)$$

In random forest arrangement technique, numerous classifiers are produced from littler subgroups of the information and late their distinct outcomes are collected in light of a voting mechanism to create the coveted yield of the informational info index. This outfit learning system has, as of late turned out to be exceptionally prevalent. Previously from the RF, Bagging, Boosting and were the main two group education strategies utilized. RF has been broadly connected in different zones together with sedate present-day disclosure, organize interruption identification, arrive cover examination, FICO assessment investigation, remote detecting and quality microarrays information examination and so on [26] [27].

Two approaches are used to assess the blunder proportion. They are the training and testing part. To fabricate the backwoods, the preparation part is used, and to figure the blunder rate, the test part-use. An alternative path is to utilize the Out of Bag (OOB) blunder assess. Since irregular backwoods calculation figures the OOB mistake amid the preparation stage, we don't have to part the preparation information.

3.4.2 SVM:

It is an assessable grouping approach and depends on the boost of edge among the occasions also the hyper-plane partition. The technique observed as the best content arrangement of performance.

It is said to be a non-probabilistic paired direct classifier, which can straightly isolate the classes by a substantial edge, it winds up a standout amongst the most intense classifier equipped for taking care of limitless dimensional component vectors.

SVM is a coordinated machine learning computation which can be used for both backslide tasks and gathering. Regardless, it is, for the most part, utilized as a few portrayal issues. In this check, we plot every datum thing in n-dimensional space point with the estimation of each segment being the estimation of a specific mastermind.

By at that point, we execute assembling through discovery the hyper-plane and that distinctive the double classes to an incredible degree well.

It has demonstrated effective in characterizing supposition archives, including style. The guideline for help vector machine calculation is to tackle arrangement and relapse issues. It has connected to numerous fields. SVMs were produced in the 1990s given the hypothetical contemplations of Vladimir Vapnik on the advancement of a measurable hypothesis learning is said to be Vapnik-Chervonenkis hypothesis. SVMs were immediately received aimed at their capacity to work through vast information, the modest hyper constraints number, their hypothetical certifications, also their great outcomes by and by.

A linear classifier is measured to make the SVM is less demanding for twofold arrangement issue with highlights x and y names. $y \in \{-1, 1\}$ is used to designate the constraints w , b and class labels:

$$f(x) = w^T x + b \quad (13)$$

where,

The normal to the line denoted as w . Bias displayed by b .

It is spoken to via an isolating the hyperplane of $f(x)$ in the sense of geometrically separates the information space into two assorted districts subsequently bringing about the grouping of the info information space into two classifications [28]:

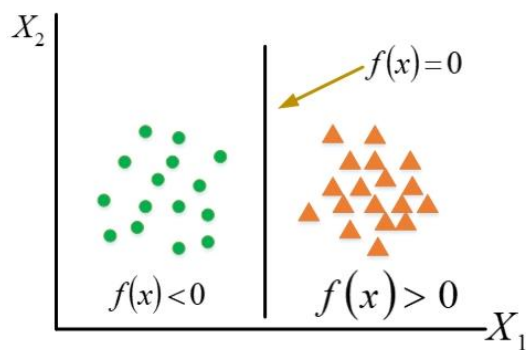


Fig 3.6: Separating the Hyperplane $f(x)$

The capacity $f(x)$ means that the isolation of hyperplane in two locales also encourages in a grouping of the informational index. Hyperplane creates two locales symmetrically and compares to two classifications of information underneath two class names. An information point "a" has a place with both of the area relying upon the estimation of $f(a)$. On the off chance that $f(a) > 0$ it has a place with one district and if $f(a) < 0$ it has a place with another region.

Accept that the information comprises of n information vectors where every datum vector is spoken to by

$x_i \in R^n$, where $i=1, 2 \dots n$. Class is given, and a chance to mark that should exist allocated to the information directions to execute managed order stay signified using y_i , which is $+1$ aimed at information vectors of one classification and -1 aimed at the further classification of information vectors. By using hyperplane, the informational collection can be geometrically isolated. Subsequently, the hyperplane is spoken to by a line it can likewise be scientifically spoken to the following eqn[29]:

$$w^T x_i + b \geq +1 \quad (14)$$

$$w^T x_i + b \leq -1 \quad (15)$$

The hyperplane mathematical formation indicated below:

$$f(x) = \text{sgn}(w^T x + b)$$

(16)

Where, sign function is indicated as $\text{sgn}()$, which is scientifically characterized through the equivalence, as shown below [31]:

$$\text{sgn}(X) = \begin{cases} 1 & \text{if } X > 0 \\ 0 & \text{if } X = 0 \\ -1 & \text{if } X < 0 \end{cases} \quad (17)$$

The following condition gives the separation D . It is from the hyperplane to the point is spoken to scientifically:

$$D = \frac{|w^T x + b|}{w} \quad (18)$$

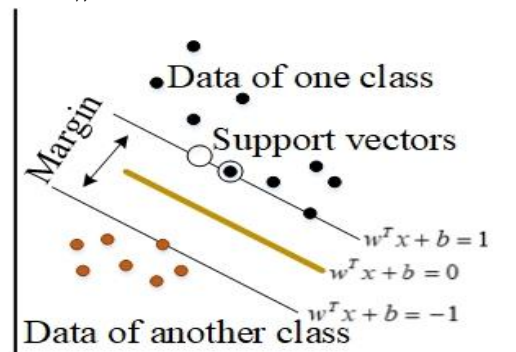


Fig 3.7: Support Vector Machines Hyperplane

The margin is displayed by:

$$\frac{W}{\|W\|} \cdot (X_+ - X_-) = \frac{W^T (X_+ - X_-)}{\|W\|} \quad (19)$$

$$= \frac{W^T \left(\left(\frac{+1-b}{W^T} \right) - \left(\frac{-1-b}{W^T} \right) \right)}{\|W\|} = \frac{2}{\|W\|} \quad (20)$$

Many hyperplanes use in this optimization, and these hyperplanes have split the data into two regions. But the hyperplane is selected via the SVM, in the two regions the hyperplane is the extreme distance from the nearby data points. But few hyperplanes are there, and that should fulfill this condition. Finally, the accurate classification results are provided the SVM by ensuring that criterion [32].

The means took after while utilizing SVM in arranging information are specified in the beneath algorithm:

Algorithm: 3

I/P: I: Given data

O/P: V: Maintenance vectors set

Start

Separate the given data set into two sets of data items

To support the vector set, V then add them

The divided n data items are then loop

If a data element is not allocated to any of the class labels means then add that to set V.

If inadequate data substances are created then break

end the loop

Train and test using the SVM classifier

End

Fig 3.8: SVM Algorithm

IV. EXPERIMENTS RESULTS

A. Dataset:

City Pulse of EU FP7 Assignment [33] includes a few SC applications in light of IoT. Inside the extent of the pollution, task, road traffic, climate, social, parking and library information was gathered starting the urban areas of Brasov and Aarhus in Denmark and also Romania individually in the vicinity of 2015 and 2013. In this investigation, CityPulse EU FP7 pollution dataset Task is utilized to understand the novel framework. The dataset has eight features such as particulate issue, ozone, sulfur dioxide, carbon monoxide, nitrogen dioxide, timestamp, longitude, and latitude utilized for exploring. Seventeen thousand five hundred sixty-eight examples are presented in the dataset and that are assembled at five-minute intervals. Each test esteem set as standard of EPA's AQI. In this examination, nitrogen dioxide and ozone toxins are chosen as expectation of air quality.

B. Thresholds:

In this part, by considering the AQI (Air Quality Index) critical level (100), its levels are separated into three. Based on the alarm color, the threshold values are defined, and it is described in the below table.










Table 4.1: Threshold value for AQI

Index value	Level	Alarm Colour
0 - 50	Healthy	Green
51 – 100	Medium Healthy	Yellow
101 - 500	Unhealthy	Red

According to these threshold values, the test datasets of nitrogen dioxide and ozone are independently labeled as Healthy, Medium Healthy, and Unhealthy. And it is enumerated in table 4.2.

Table 4.2: The Alarm based Decision Table

Ozone	Nitrogen Dioxide	Alarm based Colour

Healthy	Healthy	
Healthy	Medium Healthy	
Healthy	Unhealthy	
Medium Healthy	Healthy	
Medium Healthy	Medium Healthy	
Medium Healthy	Unhealthy	
Unhealthy	Healthy	
Unhealthy	Medium Healthy	
Unhealthy	Unhealthy	

As shown in table 4.3, evaluating the predicted values and observed values of pollutant attentions together, the final alarm colors are labeled.

C. Performance Metrics:

The following well-known methods such as Precision, Confusion Matrix (CM), False Negative Rate (FNR), Accuracy metrics, F1-Score, Specificity, False Positive Rate (FPR) and G-mean are used to estimate the alarm color matching performance of the novel prediction strategy. Precision and Recall then showed as;

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (21)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (22)$$

$$\text{Specificity} = \frac{TN}{(FP + TN)} \quad (23)$$

$$\text{FPR} = \frac{FP}{FP + TN} \quad (24)$$

$$\text{FNR} = \frac{FN}{FN + TP} \quad (25)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (26)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (27)$$

$$G - \text{mean} = \sqrt{\text{Sensitivity} \times \text{precision}} \quad (28)$$

The precision and recall harmonic mean is said to be a combine measure between recall and precision,

An Adaptive Whale Optimization Algorithm Guided Smart City Big Data Feature Identification for Fair Resource Utilization

the balanced F-score or traditional F-measure:

$$F - Score = 2 \times \frac{Precision \times recall}{Precision + recall} \quad (29)$$

Where, True Negative is indicated by *TN*, True Positive is indicated by *TP*, False Negative is represented by *FN* and False Positive is denoted as *FP*.

D. Results Analysis:

Here the outcomes of double air pollutant concentrations and assessment of air quality prediction model are introduced. To appraise that prediction performance from several perspectives, we used Precision, False Negative Rate, G-Mean, False Positive Rate, F1-Score, Recall, Accuracy metrics, and Specificity.

In terms of error criteria, the proposed strategy gives a significant development. Precision, Confusion Matrix (CM),

F1-Score, and Recall values are obtained according to the results from the arrangement and which is prepared by using threshold values. Table 4.1 gives the threshold values, and table 4.2 gives the decision values.

Next, Table 4.3 represents the performance metrics values of the Precision, False Negative Rate, G-Mean, False Positive Rate, Recall, F1-Score, specificity and Accuracy metrics based on the existing methods HMM-SVM, ANN, and KNN.

Table 4.3: Evaluation of Performance Metrics

Methods	Precision	Recall	Specificity	FPR	FNR	F1-measure	G-mean	Accuracy
Proposed	0.94	0.95	0.92	0.57	0.15	0.941	0.943	0.978
HMM-SVM	0.92	0.9	0.9	0.5	0.23	0.92	0.91	0.96
ANN	0.84	0.81	0.86	0.45	0.65	0.55	0.82	0.85
KNN	0.74	0.79	0.79	0.49	0.61	0.81	0.76	0.81

Table 4.4: Pollution Level Ranking of Cities

CITIES	POLLUTION LEVEL RANKING
1	9
2	3
3	10
4	2
5	1
6	4
7	8
8	7
9	5
10	6

Table 4.4 gives the pollution level ranking of cities. Here, the ranking of the cities is evaluated by computing the pollution rates between the pollutants ozone and nitrogen dioxide. According to the air quality index, the health concern level calculated, and the alarm is predicted the color. The cities ranked by using this alarm. Based on the unhealthy cities are ranked. Let us consider ten cities, among that we find out the hazard level of each city. According to the hazardous level of the city, the lowest level is said to be rank 1, further levels are mentioned as rank 2, rank 3...etc. The graphical representing the ranking of cities is shown in fig 4.9. The following figures are the performance analysis of the special matrices.

The performance analysis of graphical representation is shown in Fig 4.1. It represented the proposed method has good specificity performance compared to other existing methods HMM-SVM, ANN, KNN.

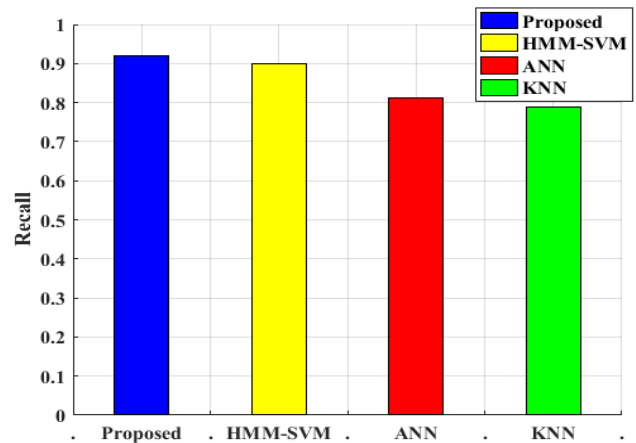


Fig 4.1: Performance of Specificity

Fig 4.2 and Fig 4.3 gives the graphical representation of the recall and precision performance. It has high values rather than the HMM-SVM, ANN and KNN

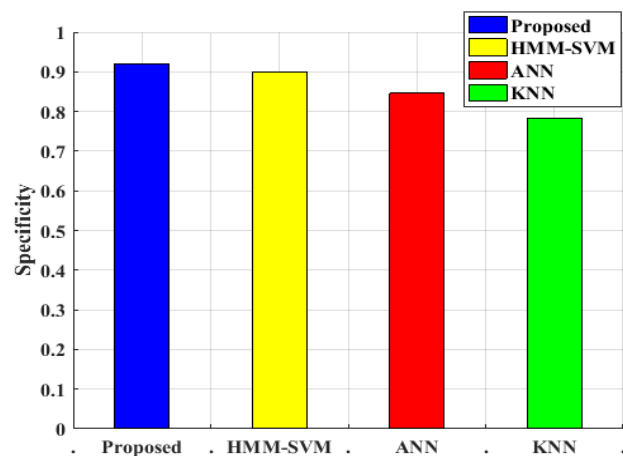


Fig 4.2: Recall Performance

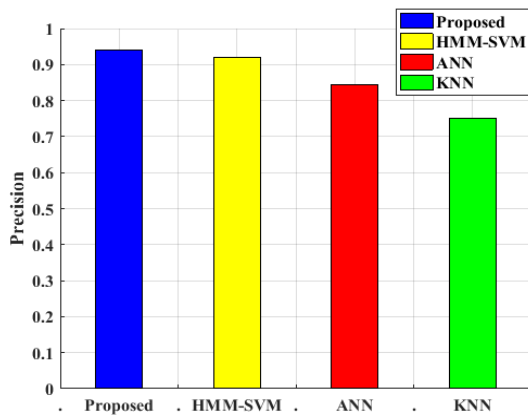


Fig 4.3: Precision Performance

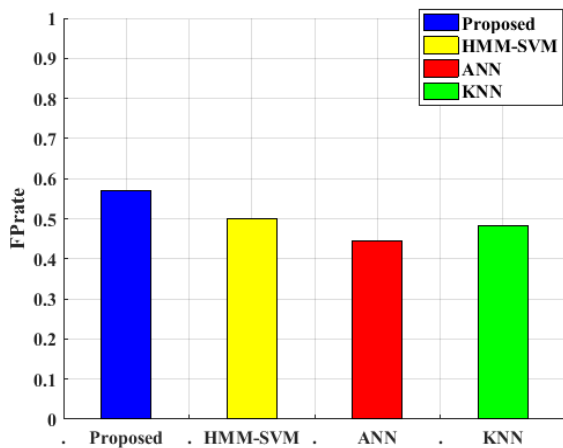


Fig 4.4: FPR Performance

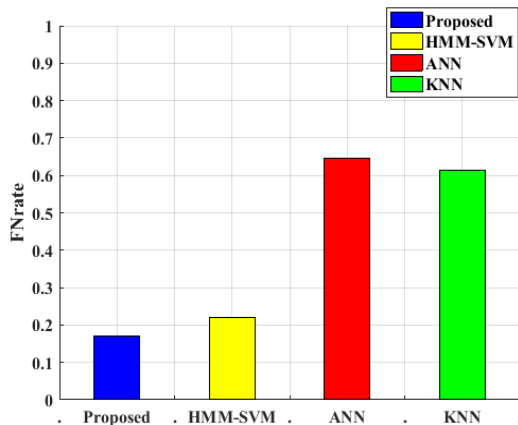


Fig 4.5: FNR Performance

Fig 4.4 and fig 4.5 represents the false positive and negative rate performance. When compared to the existing approaches such as HMM-SVM, ANN, and KNN, our proposed method has good performance rate.

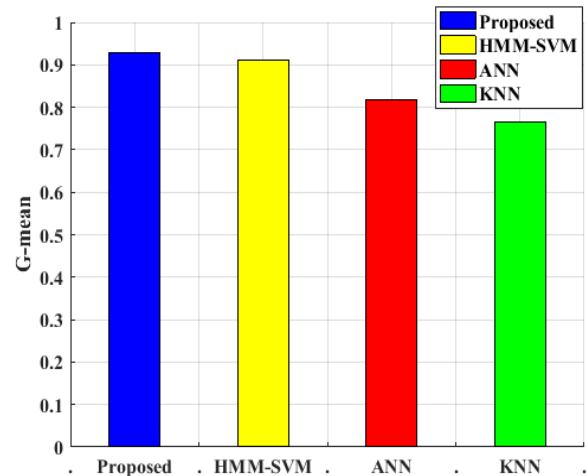


Fig 4.6: G-mean Performance

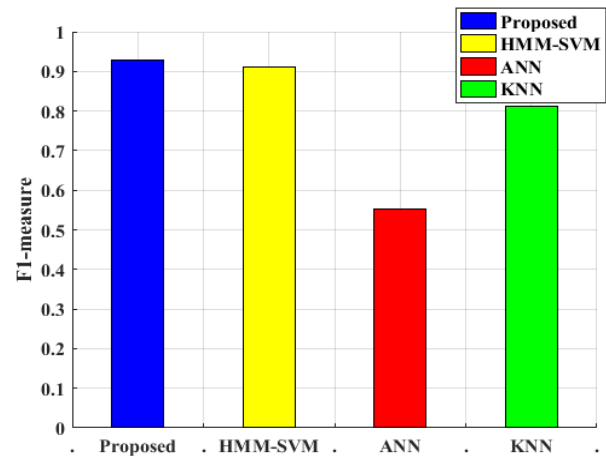


Fig 4.7: F1-measure Performance

Fig 4.6 and 4.7 shows the graphical representation of the performance of G-mean and F1-measure. By seeing this representation, our proposed method has a high-performance value.

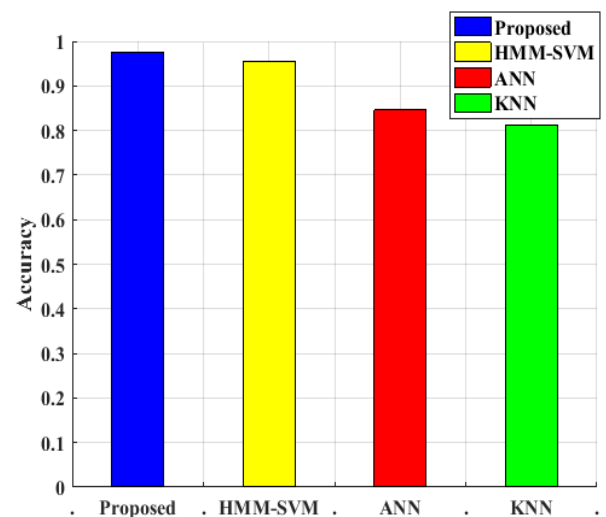


Fig 4.8: Performance of Accuracy

Fig 4.8 represents the performance accuracy graphical representation. Accuracy of the proposed method is high compared to existing methods such as HMM-SVM, ANN, and KNN.

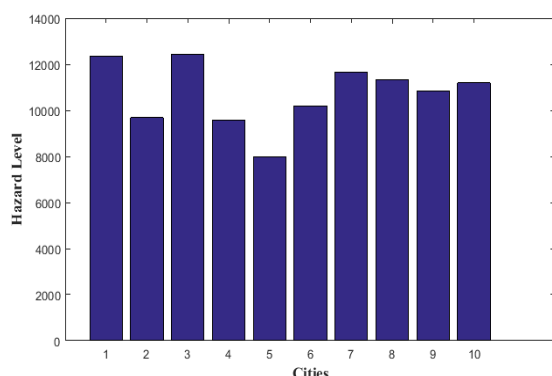


Fig 4.9: Ranking Level of Different Cities

Fig 4.9 gives the graphical representation of different cities. In which, the city having the lowest hazardous level marked as rank 1. City 5 having rank one position that is hazardous level is low.

V. CONCLUSION

Smart cities have the most indispensable part in altering distinctive regions of human life, touching segments like transportation, wellbeing, vitality, and instruction. Big data smart cities fair resource utilization is the main goal of this method. In this method, Whale optimization is used to get the optimal resource set, and the whale optimization method is performed based on the wrapper method. From the City pulse dataset, air pollution datasets are utilized for the performance analysis. Finally, the regression model Random Forest with Support Vector Machine (RF-SVM) classifier is developed to get the performance analysis. The novel approach is implemented in the MATLAB platform the enactment is evaluated based on the identified subset of features leads to optimal resource utilization and obtained results compared to earlier wrapper model-based algorithm. Our proposed method has a good performance compared to the other existing methods, and the cities are ranked based on the hazardous level of pollution.

REFERENCES

1. Sta, Hatem Ben, "Quality and the efficiency of data in "Smart-Cities," Future Generation Computer Systems, Vol. 74, pp. 409-416, 2017.
2. Lazaro, George Cristian, and Mariacristina Rosica, "Definition methodology for the smart cities model," Energy, Vol. 47, No. 1, pp. 326-332, 2012.
3. Rodríguez-Mazahua, Lisbeth, Cristian-Aarón Rodríguez-Enríquez, José Luis Sánchez-Cervantes, Jair Cervantes, Jorge Luis García-Alcaraz, and Giner Alor-Hernández, "A general perspective of Big Data: applications, tools, challenges and trends," The Journal of Supercomputing, Vol. 72, No. 8, pp. 3073-3113, 2016.
4. Zanella, Andrea, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi, "Internet of things for smart cities," IEEE Internet of Things Journal, Vol. 1, No. 1, pp. 22-32, 2014.
5. Bhosale, Amar S. "Critical and Comparative Study of Indian Tv Industry With Special Reference To Channel Trp." (2017).

6. Chen, Min, Shiwen Mao, and Yunhao Liu, "Big data: A survey," Mobile Networks and Applications, Vol. 19, No. 2, pp. 171-209, 2014.
7. Gubbi, Jayavardhana, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future generation computer systems, Vol. 29, No. 7, pp. 1645-1660, 2013.
8. Hashem, Ibrahim Abaker Targio, Victor Chang, Nor Badrul Anuar, Kayode Adewole, Ibrar Yaqoob, Abdullah Gani, Ejaz Ahmed, and Haruna Chiroma, "The role of big data in the smart city," International Journal of Information Management, Vol. 36, No. 5, pp. 748-758, 2016.
9. Al Nuaimi, Eiman, Hind Al Neyadi, Nader Mohamed, and Jameela Al-Jaroodi, "Applications of big data to smart cities," Journal of Internet Services and Applications, Vol. 6, No. 1, pp. 25, 2015.
10. Kambatla, Karthik, Giorgos Kollias, Vipin Kumar, and Ananth Grama, "Trends in big data analytics," Journal of Parallel and Distributed Computing, Vol. 74, No. 7, pp. 2561-2573, 2014.
11. Vilajosana, Ignasi, Jordi Llosa, Borja Martinez, Marc Domingo-Prieto, Albert Angles, and Xavier Vilajosana, "Bootstrapping smart cities through a self-sustainable model based on big data flows," IEEE Communications Magazine, Vol. 51, No. 6, pp. 128-134, 2013.
12. Yang, Jieming, Yuanning Liu, Xiaodong Zhu, Zhen Liu, and Xiaoxu Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," Information Processing & Management, Vol. 48, No. 4, pp. 741-754, 2012.
13. Miorandi, Daniele, Sabrina Sicari, Francesco De Pellegrini, and Imrich Chlamtac, "Internet of things: Vision, applications and research challenges," Ad Hoc Networks, Vol. 10, No. 7, pp. 1497-1516, 2012.
14. Liu M, Lu X, Song J, "A New Feature Selection Method for Text Categorization of Customer Reviews," Communications in Statistics-Simulation and Computation, Vol.45, No.4,pp.1397-409,2016.
15. Yang J, Liu Y, Zhu X, Liu Z, Zhang X, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," Information Processing and Management, Vol.48, No.4,pp.741-754,2012.
16. Wang Y, Wang J, Liao H, Chen H, "An efficient semi-supervised representatives feature selection algorithm based on information theory," Pattern Recognition, Vol.61,pp.511-523,2017.
17. Abualigah LM, Khader AT, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering," The Journal of Supercomputing, pp.1-23, 2017.
18. Lin KC, Zhang KY, Huang YH, Hung JC, Yen N. Feature selection based on an improved cat swarm optimization algorithm for big data classification. The Journal of Supercomputing. 2016 Aug 1; 72(8):3210-21.
19. Tabakhi S, Moradi P, Akhlaghian F. An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence. 2014 Jun 30; 32:112-23.
20. Fong S, Wong R, Vasilakos AV. Accelerated PSO swarm search feature selection for data stream mining big data. IEEE transactions on services computing. 2016 Jan 1; 9(1):33-45.
21. Mafarja, Majdi, and Seyedali Mirjalili. "Whale optimization approaches for wrapper feature selection." Applied Soft Computing 62 (2018): 441-453.
22. Zambia, Hossam M., Eid Emary, Aboul Ella Hassanien, and Bazil Parv. "A wrapper approach for feature selection based on swarm optimization algorithm inspired by the behavior of social-spiders." In Soft Computing and Pattern Recognition (SoCPaR), 2015 7th International Conference of, pp. 25-30. IEEE, 2015.
23. M. Seyedali and A. Lewis, "The Whale Optimization Algorithm," Adv. Eng. Soft. vol. 95, pp. 51-67, 2016.
24. Al Amrani, Yassine, Mohamed Lazaar, and Kamal Eddine El Kadiri. "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis." Procedia Computer Science 127 (2018): 511-520.
25. Gender, Robin. "Forêts aléatoires: aspects théoriques, sélection de variables et applications." Ph.D. diss., Université Paris Sud-Paris XI, 2010.

26. V.F. Rodríguez-Galiano, F.Abarca-Hernández, B. Ghimire, M. Chica-Olmo, P.M.Akinson, C. Jeganathan, "Incorporating Spatial Variability Measures in Land-cover Classification using Random Forests," *Procedia Environmental Sciences*, vol. 3, pp. 44-49, 2011.
27. Reda M. Elbasiony, Elsayed A.Sallam, Tarek E. Eltobely, Mahmoud M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means," *An in Shams Engineering Journal*, Available online 7 Mar. 2013.
28. Al-Amrani, Yassine, Mohamed Lazaar, and Kamal Eddine Elkadiri. "Sentiment Analysis using supervised classification algorithms." In *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications*, p. 61. ACM, 2017.
29. Chun-Xia Zhang, Jiang-She Zhang, Gai-Ying Zhang, "An efficient modified boosting method for solving classification problems," *Journal of Computational and Applied Mathematics*, vol. 214, issue 2, 1 May 2008, pp. 381-392.
30. Xinjun Peng, Yifei Wang, Dong Xu, "Structural twin parametric-margin support vector machine for binary classification, *Knowledge-Based Systems*," vol. 49, Sept. 2013, pp. 63-72.
31. J. T. Lalis, "A New Multiclass Classification Method for Objects with Geometric Attributes Using Simple Linear Regression," *IAENG International Journal of Computer Science*, vol. 43, no. 2, pp.198–203, 2016.
32. Hsun-Jung Cho, Ming-Tseng, "A support vector machine approach to CMOS-based radar signal processing for vehicle classification and speed estimation," *Mathematical and Computer Modelling*, vol. 58, issues 1–2, Jul. 2013, pp. 438- 448.
33. T. C. Consortium, "CityPulse Annual Report," The CityPulse Consortium2016.
34. S. Tayal, N. Nagwal, and K. Sharma, "Role of big data in make in India," in *Advances in Intelligent Systems and Computing*, 2018, vol. 564, pp. 431–437.
35. K. Sharma and S. Tayal, "Indian Smart City Ranking Model," *Int. J. Recent Technol. Eng.* ISSN2277-3878, vol. 8, no. 2, pp. 4820–4832, 2019.
36. S. Tayal, S. K. Goel, and K. Sharma, "A comparative study of various text mining techniques," in *2015 International Conference on Computing for Sustainable Global Development, INDIACom 2015*, 2015, vol. 11, no. 5, pp. 76–81.
37. R. Garg, R. K. Sharma, and K. Sharma, "Ranking and selection of commercial off-the-shelf using fuzzy distance-based approach," *Decis. Sci. Lett.*, vol. 5, no. 2, pp. 201–210, 2016.
38. K. Sharma, R. Garg, C. K. Nagpal, and R. K. Garg, "Selection of optimal software reliability growth models using a distance-based approach," *IEEE Trans. Reliab.*, vol. 59, no. 2, pp. 266–276, 2010.
39. S. Tayal and K. Sharma, "The Recommender Systems Model for Smart Cities," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2S7, pp. 451–456, 2019.

AUTHORS PROFILE



Kapil Sharma has completed his Doctors Degree in Computer Science and Engineering under the Faculty of Engineering and Technology at the M. D. University, Rohtak (Haryana), India in 2011. he has worked as a Faculty of the Department of Information Technology at Guru Premsukh Memorial College of Engineering, Budhpur, Delhi, India. Presently, he is working as Professor and Head of Information Technology Department at Delhi Technological University (DTU), Delhi,.



Sandeep Tayal is Pursuing a Doctorate Degree in Information Technology under the Department of Information Technology at Delhi Technological University (DTU), Delhi, India. Presently, he is working as an Assistant Professor in the Department of Computer Science and Engineering at Maharaja Agrasen Institute of Technology, Rohini Sector- 22, Delhi, India