RESEARCH ARTICLE OPEN ACCESS

# Optimised Approach for Traffic Prediction by Hybridizing KNN with Euclidean Distance with WSN

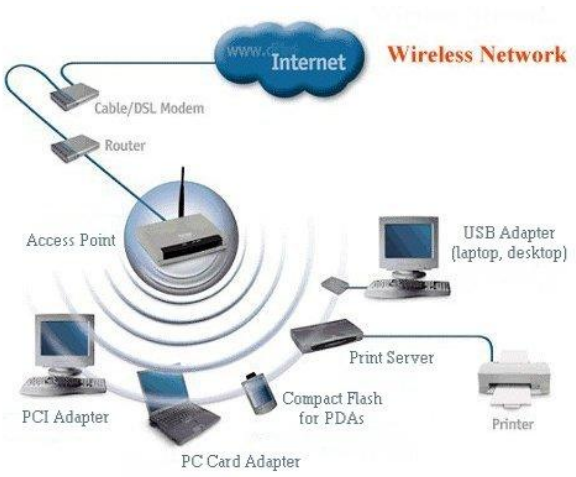
Kawaljit kaur [1], Meenakshi Sharma [2]
Student [1], HOD [2]
Department of Computer Science Engineering
GIMET Amritsar
Punjab - India

**ABSTRACT**
Traffic prediction within WSN is critical since it leads to decay in time consumption as packet is being transmitted from source to destination. Traffic prediction mechanisms are researched over and work towards accurate prediction of traffic where critical parameter for prediction is accuracy. This paper proposed a enhanced KNN mechanism for predicting traffic on route and convey to neighbouring nodes for following different routes. In order to accomplish this modified KNN approach with Euclidean distance is proposed. The data set used for prediction of derived from UCI machine learning website. Proposed approach produce better result as compared to existing approach without KNN and Euclidean distance. Applications along with examples of KNN with Euclidean distance is presented to prove worth of the study.
***Keywords :—*** Traffic prediction, WSN, KNN , Euclidean distance

## I. INTRODUCTION

Wireless network is defined as a type of computer network that uses wireless data connections for connecting nodes in the network. It is a 'shared network' network technology that is based on the transmission of radio frequencies between a wireless networking card and the base station. In recent years, the growth of economy has led to the advancement of wireless technologies. Wireless technology is the easiest and fastest way of connecting with one another whether it is a personal life or professional.[1] Due to this reason, it has gained popularity all over the world in past few years. These wireless technologies include the cellular networks, wireless local area networks, bluetooth, etc. The emergence of these technologies has largely facilitated the world. In the present time there exist two main categories under mobile wireless networks. First is the infrastructured networks that has fixed and wired gateways. The network consists of bridges which are known as base stations. [2]The basic structure of wireless networks is shown in the Figure1. The mobile unit in the network is responsible for connecting and communicating with the base stations within a specific communication range. A situation when a mobile goes out of range of one base station and enters another base station's Traffic prediction within WSN is required for reducing propagation delay as packets are being transmitted from source towards destination. Traffic predictions process using KNN and various other models are described in this section. Before describing models data collection process comes into existence.range, there occurs a "handoff" from the old base station to the new one and the mobile continues the communication throughout the network. Wireless local area network (WLAN) is the application of this type of network.[3] Another category under mobile wireless network is the infrastructure-less network which is also known as an ad hoc network. This type of network has no fixed network structure like the infrastructure network. It has no fixed routers and the nodes are independent of moving anywhere in the network. [4] The nodes in this network act as routers and these routers discover and maintain the routes to other nodes within the network. Other applications are the business applications, educational applications, etc.

Fig. 1:Wireless Network[5]




### COLLECTION OF PARAMETERS THROUGH DATASETS

The parameters collection is integral part of traffic prediction. Collection of parameters is organised in the form of tabular structure. [6], [7]as more and more data is collected Big Data is formed, it is organised to form dataset. Parameter collection

process involves sensors placed on different parts of the body. As the persons moves or perform distinct activities, sensor produces information which is recorded in memory. Overall organization of internet of things in parameters collection is organised as follows

| PARAMETER COLLECTION "SETTINGS" ALONG WITH SENSOR PLACEMENT | | |
|---|---|---|
| Attribute | Description | Utilization Example |
| Lanes | Devices attached inside or outside Lanes | Devices used to maintain well being of humans. Applications include disease management, increased productivity etc. |
| Home based environment | Homes and Building where people live in | Sensors used in security systems |
| Business Store | Places where customers engage in transactions | Stores, Banks, maul etc involving large number of people. |
| Offices | Place where intellectuals interact with each other for business | Management of energy and security enhancement services in buildings |
| Organization like factories, industries etc. | Mostly used in production | Places where repetitive work is done like in hospitals, inventory systems. |
| Sites where actual work is done | User specific customer environment | Oil Mining and construction environment |
| Cars and other moving vehicles | System which work inside moving vehicles | Vehicles including cars, jeeps etc used to monitor consumption of fuel. |
| Urban Environment | Cities | Smart Cities |
| Miscellaneous | Between Urban and rural area | Including rail tracks , roads etc. used to detect blockage if any |

Table 1: Parameter Collection settings Source

Collection of parameters collected through the above listed source form dataset. For detection of disease related to Activities, dataset from UCI website can be drawn. The parameter collection process is listed in following diagram

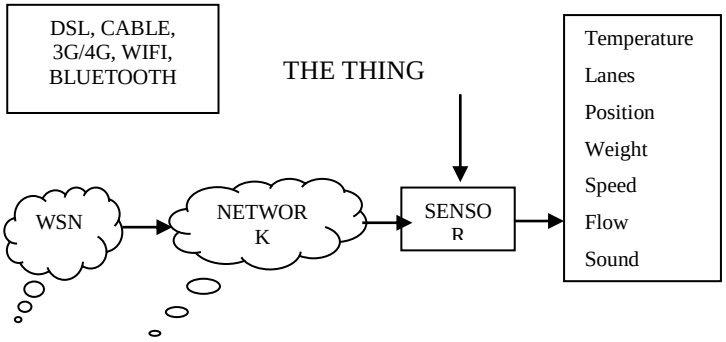



Fig. 2: Parameter Collection process through Dataset

## II. LITERATURE SURVEY CONCERNING TECHNIQUES USED FOR PREDICTION OF TRAFFIC

Predicting traffic situations and monitoring it is critical for safeguard of humans. Techniques used for prediction purposes are discussed in this section.

T. Zhou et.al. proposed that Road traffic conditions are normally influenced by occasions, for example, outrageous climate or game recreations. With the progress of Web, occasions and climate conditions can be promptly recovered continuously. In this paper, we propose an activity condition expectation framework consolidating both on the web and disconnected in-development. RFID-based framework has been conveyed for mon-itoring street activity. By joining information from both street movement checking framework and online data, we star represent a various leveled Bayesian system to foresee street activity condition. Utilizing chronicled information, we set up a progressive Bayesian system to describe the connections among occasions and street activity conditions. To assess the model, we utilize the activity information gathered in Western Massachusetts and in addition online data about occasions and climate. Our proposed expectation accomplishes an exactness of 93% generally speaking.[8]

As street systems are ending up progressively used, it is progressively critical to have the capacity to precisely anticipate travel times for scattering data among street clients and to help movement administration choices and arranging. In the course of recent years, another method for making such forecasts using the space-time autoregressive incorporated moving normal (STARIMA) show has been presented. The outcomes have been exceptionally encouraging so far with great exactness revealed for expectation times of a few several minutes. Be that as it may, up until now, the writing just concerns consistent state freeflow or Manhattan lattice based situations. In this paper, we sum up on the past work and explore how this approach performs in urban rush hour gridlock situations. We explore the models expectation precision both on an informational index of estimated travel times in the metropolitan Sydney locale and additionally a

steadfast portrayal of a substantial area of Sydney's urban scene. We break down the execution of STARIMA under six level of administration (LOS) in the two settings and locate that despite the fact that the model performs well in the relentless state case, the fundamental approach isn't appropriate for displaying the urban movement setting. We along these lines propose to broaden the STARIMA display with input control circles all together for the way to deal with be appropriate for exceedingly differing conditions.[9]

[10] With the accessibility of movement sensors information, various methods have been proposed to make blockage prediction by using those datasets. One key test in anticipating activity clog is the amount to depend on the verifiable information v.s. the continuous information. To better use both the authentic and ongoing information, in this paper we propose a novel online system that could take in the present circumstance from the continuous information and anticipate the future utilizing the best indicator in this circumstance from an arrangement of indicators that are prepared utilizing recorded information. Specifically, the proposed structure utilizes an arrangement of base indicators (e.g. a Support Vector Machine or a Bayes classifier) and learns progressively the best one to use in various settings (e.g. time, area, climate condition). As constant activity information arrives, the setting space is adaptively apportioned so as to proficiently gauge the adequacy of every indicator in various settings. We get and demonstrate both here and now and long haul execution ensures (limits) for our online calculation. Our examinations with certifiable information, all things considered, conditions demonstrate that the proposed approach essentially outflanks existing arrangements.

M. T. Asif et.al. discussedThe capacity to precisely anticipate activity speed in a huge and heterogeneous street arrange has numerous helpful applications, for example, course direction and clog shirking. On a basic level, information drivenmethods, for example, SupportVector Regression (SVR) can anticipate activity with high precision, since movement tends to show normal examples after some time. Notwithstanding, practically speaking, the forecast execution can fluctuate essentially over the system and amid various eras. Knowledge into those spatial and fleeting patterns can enhance the execution of Intelligent Transportation Systems (ITS). Customary forecast mistake measures such asMean Absolute Percentage Error (MAPE) give data about individual connections in the system, however don't catch worldwide patterns. We propose unsupervised learning techniques, for example, k-implies grouping, Principal Component Analysis (PCA), and Self Organizing Maps (SOM) to mine spatial and worldly execution patterns at both system level and for singular connections. We perform expectation for an expansive, interconnected street organize, for various forecast skylines, with SVR based calculation. We demonstrate the viability of the proposed execution examination strategies by applying them to the forecast information of SVR.[11]

X. Pang et.al. proposed a Short-term traffic flow is one of the core technologies to realize traffic flow guidance. In this article, in view of the characteristics that the traffic flow changes repeatedly, a short-term traffic flow forecasting method based on a three-layer K-nearest neighbor non-parametric regression algorithm is proposed. Specifically, two screening layers based on shape similarity were introduced in K- nearest neighbor non-parametric regression method, and the forecasting results were output using the weighted averaging on the reciprocal values of the shape similarity distances and the most-similar-point distance adjustment method. According to the experimental results, the proposed algorithm has improved the predictive ability of the traditional K-nearest neighbor non- parametric regression method, and greatly enhanced the accuracy and real-time performance of short-term traffic flow forecasting.[12]

G. Zhu et.al. Travel time parameters got from street activity sensors information assume an imperative part in rush hour gridlock administration rehearse. In this paper, a movement time investigation and expectation show was built up for urban street activity sensors information in light of the change point examination calculation and ARIMA demonstrate. Right off the bat, time arrangement of movement time parameters were bunched by utilizing change point mining calculation after activity sensors information preprocessing. At that point, a movement time expectation display was set up in light of ARIMA show. At last, the model was confirmed with high exactness through reproduction by utilizing different arrangements of information and investigation of its practicability was finished. Catchphrases—Travel. [13]

M. A. Jabbar et.al. & I. K. A. Enriko et.al. proposes a KNN technique for detecting heart disease and performing prediction accurately by simplifying parameters. The nearest neighbourhood algorithm is used to identify elements having similar attributes values. These attribute values are grouped together using grouping functions. Grouping function generates certain value which is compared against the threshold value to determine problems[14]. Problems are reflected in the form of deviation. The process is described by considering two points ‘A’ and ‘B’. Let distance(A,B) is the distance between points A and B then

a. distance(A,B)=0 and distance(A,B) >=0 iff A=B
b. distance(A,B)=distance(B,A)
c. distance(A,C)<=distance(A,C)+distance(C,B)

Property 3 is also known as transitive dependency. Distance if close to zero then prediction is accurate otherwise error is recorded. Error calculating metric is applied to determine accuracy of the approach. Accuracy is given as

Accuracy=1-Error_rate

KNN is used in many distinct environments such as classification, interpolation, problem solving, teaching and learning etc. Major limitation of KNN is that its performance depends upon value of k. Accuracy is low and further work is required to be done to improve accuracy.

B. Veytsman et.al. suggested Euclidean distance is one of the simplest mechanisms for classification and prediction. Distance is the prime criteria used to evaluate the deviation in this case. Distance can be defined in several ways.[16]

All the components of vectors are taken equally and no correlation is evaluated in this case. The result of Euclidean distance equation can be normalized.Where averaging is taken over all the vectors in the dataset. The scaled distance is adjusted value so that obtained result lie between the specified range. The metric is used to evaluate errors. [17]–[19]Mean root square error is one such mechanism for observing accuracy. Accuracy and error rate is inversely proportional to each other.

This equation is used to evaluate Root Mean square error. Lower the value of RMS more accurate a prediction. Advantage of this approach is, convergence rate is better but disadvantage is that it can work over limited values. Non negative values are allowed and hence result always lies between 0 and 1.[15]

D. V Jose et.al. suggested that Auto regressive moving average model is used for accurate forecasting in case of disease detection. Changes in time series using mathematical model is used in ARIMA. This model is based on adjustment of observed values. The goal is to obtain the differences of observed value and value obtained from the model close to zero. This model can predict accurately difference between the stationery and non stationery series. [20]-[22]

ARIMA model has multiple phases associated with it. This model can be merged with Euclidean distance and KNN for better performance in traffic prediction.

## III. PROPOSED SYSTEM FOR TRAFFIC PREDICTION

Proposed system considers traffic prediction by the application of KNN and Euclidean distance. The KNN approach forms clusters having equivalent distance within the group based on time and traffic count. The training is performed depending upon the value of K. Higher the value of K larger will be the size of the cluster. For accurate prediction of traffic, value of K must be within max and min range.

The proposed algorithm using the hybridization of KNN with Euclidean distance caused the accurate traffic prediction and has the following pseudo code.

**Algorithm ARIMA (KNN+EUCLIDEAN)**

* INPUT: Dataset with attribute values including Lanes and traffic metrics
* Output: Prediction Accuracy
a) Perform Pre-Processing
Convert attribute values to nominal form for analysis.
b) Select classifier through inputting the value of K
c) Obtain accuracy and record it for comparison
d) Apply auto regression mechanism for calculating most probable values from each attributes
e) Use value of K suggesting neighbour distance and Euclidean distance to determine values from dataset to be tested and form clusters
f) Replaced missing values calculated from KNN+EUCLIDEAN with most probable values for traffic prediction.
g) Calculate accuracy.

In the above algorithm input the attributes like lane no, latitude, longitude, no. of vehicles etc. firstly perform the pre-processing step on the data to convert the attribute values to nominal values. After that KNN classification is performed to generate clusters from that input data and extract the accuracy from them to record the results for comparison purposes. In the next step apply auto regression mechanism for calculating most probable values from each attributes. Later on neighbouring distance and Euclidean distance is used to determine values from dataset to be tested and formulate clusters from them. Now replace all the missing values calculated from KNN+EUCLIDEAN with most probable values for traffic prediction. Calculate the accuracy of the value to get optimal result.

Obtained result is in the form of classification accuracy. Performance analysis and result section is given in the next section.

## IV. PERFORMANCE ANALYSIS AND RESULTS

Results are obtained in terms of accuracy and future prediction. In case missing value are present the Existing work cannot tackle this issue. In order to overcome this problem, missing values are tackled and classification accuracy is enhanced using KNN and Euclidean distance.

| Number of Tuples Analysed | Existing without KNN+Euclidean | With KNN and Euclidean |
|---|---|---|
| 300 | 85.5 | 93.3 |
| 290 | 82.365 | 92.12 |
| 250 | 75.235 | 85.212 |
| 150 | 65.555 | 70.646 |
| 100 | 48.999 | 53.457 |

Table 1 Accuracy of classification of existing and proposed approach

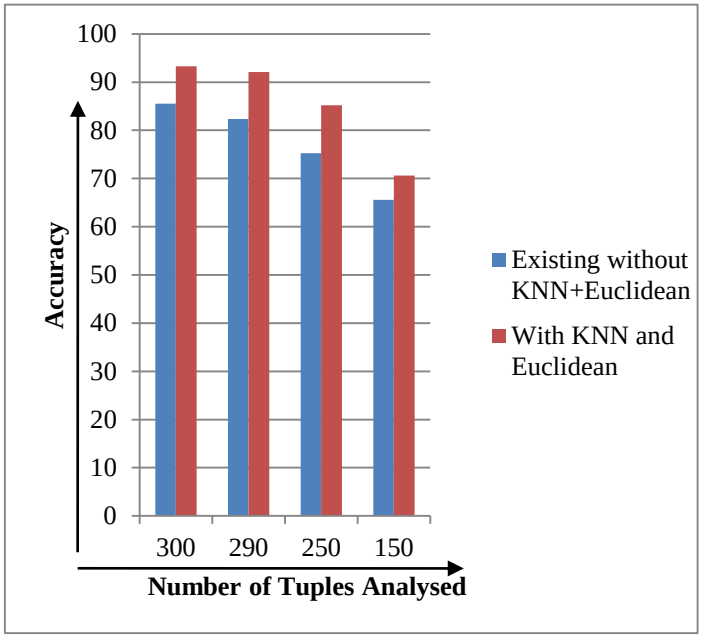



Fig. 3. Accuracy of classification of existing and proposed approach

| Number of Tuples Analysed | Existing without KNN and Euclidean | With KNN and Euclidean |
|---|---|---|
| 300 | 2013 | 435 |
| 290 | 2011 | 432 |
| 250 | 1990 | 399 |
| 150 | 1880 | 390 |
| 100 | 1473 | 380 |

Table 3. Accuracy of classification of existing and proposed appssroach

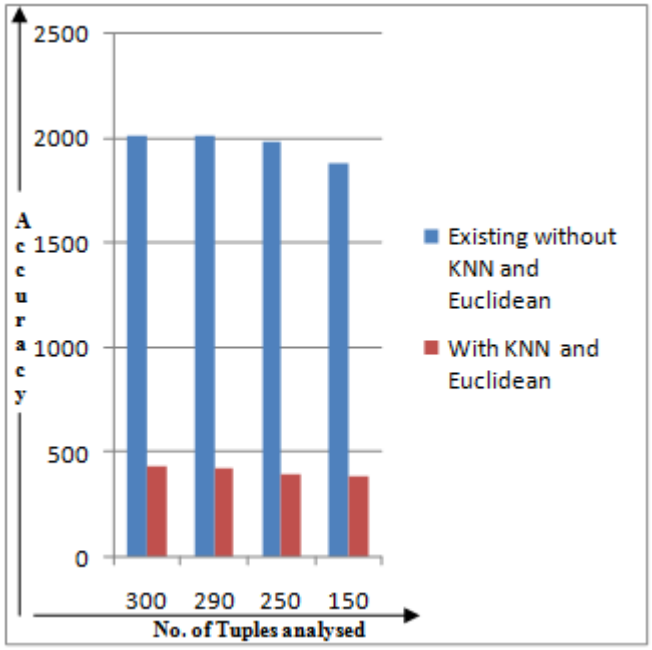



Fig. 4 latency of Support vector machine classifier with modified KNN

The prediction generated through the proposed approach is going to help the person in predicted heart diseases and take preventing measures.

## V. CONCLUSION AND FUTURE SCOPE

In this paper we proposed a framework for online traffic prediction. Proposed framework utilizes the real-time data to select the most effective predictor in different contexts, thereby self-adapting to the dynamically changing traffic conditions as well. The missing values handling is a critical part which is accomplished with the modified KNN model. The existing approach is unsupervised classifier usually used for numerical value. The missing values are generally string in nature which cannot be classified with existing approach. Hence accuracy degrades. In order to overcome this problem modified KNN with Euclidean distance is used. The result indicates betterment in terms of accuracy, prediction, and latency. As a future work, we plan to also adapt the individual base predictors using real-time data in addition to selecting the most effective one to use.

In future ARIMA with KNN and Manhattan distance can be combined or hybridized to improve prediction accuracy.

## REFERENCES

[1] Q. Xu and J. Zhao, "Multi-Head Track-Sector Clustering Routing Algorithm In WSN," no. Icitmi, pp. 707–713, 2015.

[2] Z. Zhou, C. Du, L. Shu, G. Hancke, J. Niu, and H. Ning, "An Energy-Balanced Heuristic for Mobile Sink Scheduling in Hybrid WSNs," *IEEE Trans. Ind. Informatics*, vol. 12, no. 1, pp. 28–40, 2016.

[3] K. Kaushal and V. Sahni, "Early Detection of DDoS Attack in WSN," *Int. J. Comput. Appl.*, vol. 134, no. 13, pp. 14–18, 2016.

[4] A. Kaur and H. Kaur, "A REVIEW ON A HYBRID APPROACH USING MOBILE SINK AND FUZZY LOGIC FOR REGION BASED CLUSTERING IN WSN," vol. 16, no. 2, pp. 7586–7590, 2017.

[5] Q. Nadeem, M. B. Rasheed, N. Javaid, Z. A. Khan, Y. Maqsood, and A. Din, "Multi-Hop Routing Protocol for WSNs."

[6] D. Li, H. W. Park, E. Batbaatar, Y. Piao, and K. H. Ryu, "Design of Health Care System for Disease Detection and Prediction on Hadoop Using DM Techniques," pp. 124–129.

[7] G. Vaishali and V. Kalaivani, "Big Data Analysis for Heart Disease Detection System Using Map Reduce Technique," vol. V.

[8] T. Zhou, L. Gao, and D. Ni, "Road Traffic Prediction by Incorporating Online Information," *IEEE Access*, 2014.

[9] "Predicting Travel Times in Dense and Highly Varying Road Traffic Networks using STARIMA

Models .," *IEEE*, no. February, 2012.

[10] J. Xu, D. Deng, U. Demiryurek, C. Shahabi, and M. Van Der Schaar, "Context-Aware Online Spatiotemporal Traffic Prediction."

[11] M. T. Asif, J. Dauwels, C. Y. Goh, A. Oran, E. Fathi, M. Xu, M. M. Dhanya, N. Mitrovic, and P. Jaillet, "Spatial and Temporal Patterns in Large-Scale Traffic Speed Prediction."

[12] X. Pang, C. Wang, and G. Huang, "A Short-Term Traffic Flow Forecasting Method Based on a Three-Layer K-Nearest Neighbor Non-Parametric Regression Algorithm," no. July, pp. 200–206, 2016.

[13] G. Zhu, K. Song, and P. Zhang, "A Travel Time Prediction Method for Urban Road Traffic Sensors Data," *2015 Int. Conf. Identification, Information, Knowl. Internet Things*, pp. 29–32, 2015.

[14] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," *Procedia Technol.*, vol. 10, pp. 85–94, 2013.

[15] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient ' s Health Parameters," vol. 8, no. 12, 1843.

[16] B. Veytsman, L. Wang, T. Cui, S. Bruskin, and A. Baranova, "Distance-based classifiers as potential diagnostic and prediction tools for human diseases," *BMC Genomics*, vol. 15 Suppl 1, no. Suppl 12, p. S10, 2014.

[17] M. M. El-Hattab, "Applying post classification change detection technique to monitor an Egyptian coastal zone (Abu Qir Bay)," *Egypt. J. Remote Sens. Sp. Sci.*, vol. 19, no. 1, pp. 23–36, 2016.

[18] C. Chen, M. Won, R. Stoleru, and G. G. X. Member, "Energy-Efficient Fault-Tolerant Data Storage & Processing in Mobile Cloud," vol. 3, no. 1, pp. 1–14, 2014.

[19] D. Bui, S. Hussain, E. Huh, and S. Lee, "Adaptive Replication Management in HDFS based on Supervised Learning," vol. 4347, no. c, pp. 1–14, 2016.

[20] D. V Jose and G. Sadashivappa, "a N Ovel E Nergy E Fficient R Outing a Lgorithm for W Ireless S Ensor N Etworks," *Int. J. Wirel. Mob. Networks*, vol. 6, no. 6, pp. 15–25, 2014.

[21] Y. Pan, M. Zhang, Z. Chen, M. Zhou, and Z. Zhang, "An ARIMA based model for forecasting the patient number of epidemic disease," *2016 13th Int. Conf. Serv. Syst. Serv. Manag. ICSSSM 2016*, pp. 31–34, 2016.

[22] I. A. Permanasari, A.E. Hidayah, I.Bustoni, "SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence," *Inf. Technol. Electr. Eng. (ICITEE),2013 Int. Conf.* ., no. 2, pp. 2–6, 2013.