TPR, PPV and ROC based Performance Measurement and Optimization of Human Face Recognition of IoT Enabled Physical Location Monitoring

Ajitkumar S. Shitole, Manoj H. Devare

Abstract: ***This paper describes the construction of Internet of Things (IoT) enabled system which not only captures the sensors data in textual and numeric form but also performs live human face recognition to monitor physical location effectively. The dataset used in order to apply supervised machine learning algorithms is the combination of automatically captured live sensor data along with name of the human face recognized or unknown and additional manually introduced class label. Performance measurement of face recognition is done with the help of Decision Tree (DT), K-Nearest Neighbors (KNN), Naïve Bayes (NB) and Logistic Regression (LR). The results show that DT gives the best performance with respect to classifier's accuracy; True Positive Rate, Positive Predictive Value and area under curve of Receiver Operating Characteristics (ROC) for face recognition prediction whether the recognized face is true or false.***

Index Terms: ***Machine Learning, Physical Location Monitoring, Confusion Matrix, ROC, Decision Tree, Naive Bayes, Logistic Regression, K-Nearest Neighbors.***

I. INTRODUCTION

Internet of Things (IoT) is the one of the emerging and rapidly developing technology in the field of Information Technology and Communication Engineering. Lots of devices can be connected to each other with the help of IoT to communicate and exchange their information and data. In today's life, it is necessary to monitor the physical location with the help of IoT where numbers of different sensors are connected to single board computer. Analysis of physical location is required in order to identify any abnormal conditions in the environments like home locations, sensitive laboratories, hospitals, educational institute, industries etc. Abnormal conditions can be sudden increase or decrease in temperature and humidity, increase in intensity of light, increase in gas sensor values, unknown person's detection in the premises which in turn can cause severe damage to the location and surroundings. So it is essential task to capture sensor data continuously on regular intervals and perform statistical as well as systematic analysis of the same to create decision support system which is required to avoid further loss in the environment. IoT enabled system with multimedia data such as digital images of human faces are useful for face detection and recognition. Face recognition is useful in various scenarios such as intrusion detection, identifying the several actions such as switch ON/OFF various devices, identifying user's routine in the environment to know when user is at home and interacting with the devices and so on. Development of IoT enabled system with face recognition makes significant change in safety and security of premises. More robust and powerful system can be achieved with the help of IoT and face recognition. The objective of this paper is to present prescient scientific models for IoT enabled with face recognition system for monitoring physical location. Location considered here is the living room of a home and data is captured for one month continuously. The system employs four supervised machine learning predictive models with DT, KNN, NB and LR for analysis of human face recognition to find accuracies of applied classifiers, precision, recall and ROC curve and compare them.

II. RELATED WORK

Sankar Mukherjee et al. addressed an issue of meeting sensor connect with the Mobile Adhoc Network (MANET) organizes on the grounds that hubs have distinctive power levels, heterogeneous conventions and have odds of co-channel obstructions another design of IoT systems, where sensor systems and MANET are joined together for proficient correspondence with the Internet Gateways [1].

Neelesh Mishra et al. presented an overview of different congestion control calculations utilized at transport layer. IoT requires a vehicle layer convention which offers blockage control, adaptability and dependability as indicated by necessity of gadgets [2]. Dragos Mocrii et al. presented a survey of real advancements of IoT-based smart homes and current difficulties of brilliant home advances and their scattering, and indicate some interesting arrangements and future patterns [3]. Adel Alkhalil et al. recommended the usage of information provenance as an imperative instrument that can improve the security and protection of IoT frameworks and reviewed the most difficult issues in IoT information provenance. Seven issues have been talked about including provenance security, monstrous measure of information, ordering, different customers, change, question, and interoperability [4]. Nallapaneni Manoj Kumar et al. expounded the conceivable security and protection issues considering the segment cooperation in IoT and concentrates how the Distributed Ledger based Block Chain (DL-BC) innovation add to it [5].

Revised Manuscript Received on July 20, 2019.

Ajitkumar S. Shitole, Research Scholar, Amity University Mumbai, India, Asso. Prof, I²IT, Hinjawadi, India.
Dr. Manoj H. Devare, HoI, AIIT, Amity University Mumbai, India.

Mustafa Alper Akkaş et al. displayed a Wireless Sensor Network model comprising of MicaZ hubs which are utilized to quantify nurseries' temperature, light, weight and stickiness. With this framework farmers can control their nursery from their cell phones or PCs which have web association [6].

Partha Pratim Ray reviewed mainstream IoT cloud stages in light of explaining a few administration areas such as application advancement, gadget the executives, framework the board, heterogeneity the executives, information the board, devices for investigation, arrangement, checking, perception, and research. An examination is displayed for in general spread of IoT mists as per their appropriateness [7].

Nawaz Mohamudally et al. featured the difficulties significant to center components engaged with the advancement of an Anomaly Detection Engine (ADE). It was discovered that an exact and dependable ADE depends on three fundamental determination factors to be specific, the nature of the information focuses, the time arrangement change, and where investigation are executed [8].

Mohammad Saeid Mahdavinejad et al. evaluates the different machine learning strategies that bargain with the difficulties exhibited by IoT information by considering shrewd urban areas as the fundamental use case. The key commitment of this investigation is the introduction of a scientific classification of machine learning calculations clarifying how unique strategies are connected to the information so as to remove larger amount data [9].

Bill Karakostas proposed an engineering that utilizes a Bayesian occasion expectation display that utilizes chronicled occasion information produced by the IoT cloud to ascertain the likelihood of future occasions. Framework anticipated outbound flight defer occasions, in view of inbound flight delays, in light of authentic information gathered from avionics measurements databases [10].

Huseyin Yildirim et al. concentrated to break down the variables that impact representatives' aim to utilize wearable gadgets at the work environment. In this examination, an audit of the writing with respect to acknowledgment of innovations and affecting elements, for example, hazard and trust is utilized to build up an applied model [11].

Ajitkumar Shitole et al. clarified about proposed showing of relevant adjustment approach and the executives customization that abuses distinctive identification methodology to give a proactive control advantage at home is conceivable and attainable. Principle based administration customization technique that utilizes a standard happenstance strategy dependent on semantic separation to settle on choices about the unique situation and a set hypothesis strategy dependent on set hypothesis to screen benefit customization [12].

Manoj Devare explained about the huge amount of statistics values captured from the sensor want to be analyzed because one cannot forget the ideal values. The hassle may additionally arise at some stage in the handshaking of the sensors with the libraries established inside the SBCs. The sensor information gathered within the SBC, and pushing it either at the internet-server or Cloud is likewise having a few synchronization issues. The handshaking and synchronization troubles can be detected and appropriately analyzed the usage of the statistical gear and techniques which applied to the accrued sample data within the preliminary checking out [13].

Ajitkumar Shitole and Manoj Devare depicted about the observing of a physical area isn't only a basic action yet suggests vital restorative measures after efficient investigation, to stay away from the further misfortune in the materials and also dangers in nature. The sensor information caught as time arrangement is helpful for examination of the anomalous conditions in the environment. The content based and numerical qualities from the sensor are valuable for the examination utilizing the factual instruments and procedures [14].

Alexandra Moraru et al. presented vertical framework mix of a sensor hub and a toolbox of machine learning calculations for anticipating the quantity of people situated in a shut space. The dataset utilized as a contribution for the learning calculations is made out of consequently gathered sensor information and extra physically presented information. The framework broke down the dataset and assessed the execution of two kinds of machine learning calculations on this dataset. The investigations demonstrated that enlarging sensor information with appropriate data can enhance forecast results and furthermore the arrangement calculation performed better [15].

Joseph Siryani et al. depicted machine learning Decision-Support System (DSS) which enhances the IoT Smart Meter Operations. The model is observationally assessed utilizing informational indexes from a business organize. The framework shows the effectiveness of methodology with a total Bayesian Network forecast model and contrast and three machine learning expectation demonstrate classifiers: Naïve Bayes, Random Forest and Decision Tree. Results show that approach creates factually critical estimations and that the DSS will enhance the cost effectiveness of Electric Smart Meter (ESM) arrange tasks and support [16].

Purnendu Shekhar Pandey et al. acquainted significance with advice the individual about his undesirable way of life and even alert him/her before any intense condition happens. To distinguish the pressure heretofore framework have utilized heart beat rate as one of the parameters. IoT alongside ML is utilized to alert the circumstance when the individual is in genuine hazard [17].

Go Takami et al. outlined the ML techniques and described a sensor identification experiment and the results of a deterioration determination experiment that suggests the possibility of understanding the sensor deterioration process. System believed that there is a great possibility that analysis of sensor data using the ML techniques can be used for the preventive maintenance such as sensor deterioration estimation [18].

Rui Madeira et al. clarified ML Approach for Indirect Human Presence Detection Using IoT Devices. The gave data was anonymized at the source. The initial step was to extricate satisfactory highlights for this issue. A naming advance is presented utilizing a blend of heuristics to affirm the probability of anybody being home at a given time, in light of all data accessible, including, yet not constrained to, coordinate nearness indicators.

The arrangement lays chiefly on the utilization of regulated learning calculations to prepare models that recognize the nearness with no data dependent on direct nearness finders [19]. Che-Min Chung et al. examined about new methodology that utilized ML strategies to determine the gigantic information issue in the quickly business of the IoT. This arrangement is represented considerable authority in IoT information and connected to a genuine case of a keen working with more than 100 associated sensors and its execution is contrasted with industry benchmarks [20].

III. EXPERIMENTATION

IoT enabled device is created to reveal the physical area in real time the usage of sensors connected the usage of the jumper wires. Various sensors together with digital temperature and humidity sensor, light intensity, physical presence, and gasoline detection sensors are connected to Raspberry Pi3 to optimize the physical location tracking. The device is evolved to fetch actual-Time facts from the sensors. The web camera is likewise linked to Raspberry Pi3 to capture snap shots of human face for recognition. In order to monitor physical location, the sensor data is captured regularly in real time fashion and stored onto local server. Whenever human face is detected and recognized as either known or unknown person, same data is also stored onto Go Daddy Cloud Service for further use and analysis. Subset of original dataset is pushed onto cloud to create labeled dataset and then apply supervised machine learning algorithms to measure the performance of human face recognition. Fig. 1 shows the IoT Enabled System with Face Recognition for live face recognition and sensor data capturing. Fig. 2, Fig. 3, and Fig. 4 show Face Recognition of Known Persons. Face recognition and sensor readings are captured simultaneously using multithreading programming in python. The system is used to monitor the home location continuously for one month. To create labeled dataset, cloud data was downloaded daily to add the class label manually. Whenever the person was recognized incorrectly, false entry was registered manually in register to create labeled dataset to apply supervised machine learning algorithms.

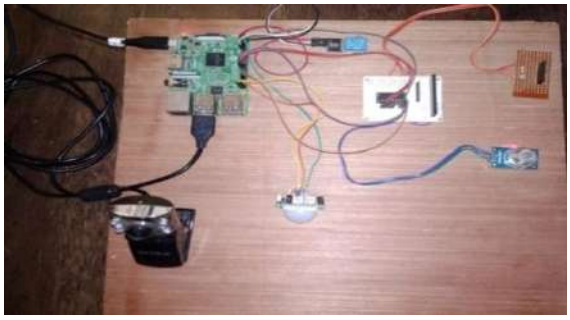


Fig. 1 IoT Enabled System with Face Recognition

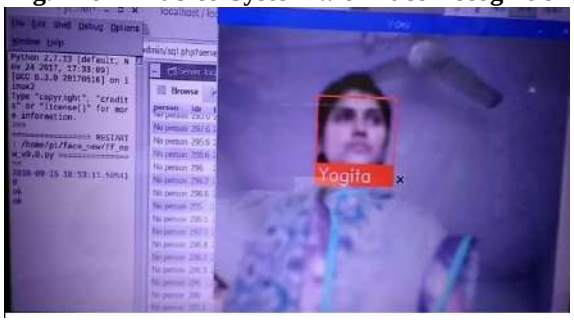


Fig. 2 Face Recognition of Known Person: Yogita

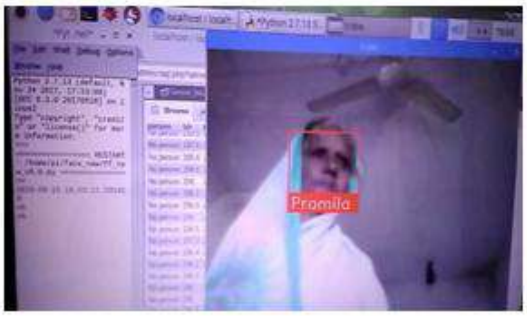


Fig. 3 Face Recognition of Known Person: Pramila

Fig. 4 Face Recognition of Known Person: Swaroop

In order to capture sensor values and to recognize human face in real time, multithreading programming in python is applied as Raspberry Pi 3 supports a quad core processor. Main thread along with two additional threads is created to achieve simultaneous processing which in turn to get maximum throughput. Main thread is used to capture the live image frame by frame and to perform processing on that captured image for face recognition. First thread out of two additional threads is used to read temperature and humidity sensor values. Second thread is used to read LDR, Gas, and PIR sensor values. To perform face recognition activity effectively in IoT enabled environment, face recognition library which recognizes and manipulates faces from Python is installed onto system. Local database of known faces is created to compare with live captured images frame by frame. Face recognition library consists of various in built methods to perform tasks such as to load image file, to get face locations, to get face encodings, to compare faces etc. Every known image is loaded into temporary variable for encoding of facial characteristics that can be contrasted with some other picture of a face. Two arrays are initialized to represent known face encoding and known face names. Live image is captured and processed to get areas and frameworks of every individual's eyes, nose, mouth, and jaw. Face location is applied to get face encodings. Captured image's face encodings are compared with known face encodings and if match is found known face name is displayed on screen otherwise unknown string is displayed. System consists of heterogeneous data as it combines numeric, string, and image data. Although image data is combined with sensor data, captured images are not stored either on local database or cloud. Whenever face is recognized, the names of known persons along with other sensor values are stored onto local server as well as cloud.

Irrespective of face detection and recognition, all entries with sampling rate of 2 to 4 seconds are maintained onto local database. Cloud database is a subset of local database as it contains entries when face is recognized either as known or unknown. Local database contains dataset in csv file with 5, 86,506 entries of size 40.2 MB where as cloud database contains dataset in csv file with 3025 entries of size 213 KB.

IV. MACHINE LEARNING MODELS

Classification is utilized to discover in which gather every datum example is connected inside a given dataset. It is utilized for characterizing information into various classes as indicated by some obliges. A few remarkable sorts of arrangement calculations including Decision Tree, K-Nearest Neighbors, Naive Bayes, and Logistic Regression are utilized for it. Arrangement is a two stage process. During initial step the model is made by applying arrangement calculation on preparing informational collection at that point, in second step the extricated model is tried against a predefined test informational collection to gauge the model prepared execution and precision. So grouping is the procedure to allot class mark from informational index whose class name is unclear.

A. Decision Tree

Decision tree fabricates regression or classification models as a tree structure. The last outcome is a tree with choice nodes and leaf nodes. A choice node has at least two branches, each speaking to values for the quality tried. Decision trees utilized in information mining are of two principle types: Classification tree investigation is the point at which the anticipated result is the class to which the information has a place. Regression tree examination is the point at which the anticipated result can be viewed as a genuine number. The Gini coefficient is a factual proportion of conveyance. The coefficient ranges from 0 (or 0%) to 1 (or 100%), with 0 speaking to consummate fairness and 1 speaking to consummate disparity. The impurity measure used in building decision tree in Classification and Regression Trees (CART) is Gini Index. The decision tree built by CART algorithm is always a binary decision tree.
Gini index for a given node *t*:

$$GINI(t) = 1 - \sum_j [p(j|t)]^2 \qquad (1)$$

$p(j|t)$ is the relative frequency of class *j* at node *t.*
When a node t is split into k partitions (child nodes), the quality of split is computed as,

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \qquad (2)$$

Where n_i = number of records at child node i
n = number of records at parent node t
Trait that boosts the decrease in impurity or having minimum gini index is chosen as dividing attribute.

B. Naïve Bayes

The Naive Bayes Classifier system depends on Bayesian hypothesis and is especially utilized when the dimensionality of the information sources is high. Bayes hypothesis gives a method for computing the posterior probability P(A|B), from P(A), P(B), and P(B|A). Naive Bayes classifier thinks about that the impact of the estimation of an indicator (B) on a given class (A) is autonomous of the estimations of different indicators.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad (3)$$

Where, P (A|B) is the posterior probability of class (target) given predictor of a class, P (A) is called the prior probability of a class, P (B|A) is the likelihood which is the probability of predictor of given class, and P (B) is the prior probability of predictor of a class.

C. K Nearest Neighbors (KNN)

KNN recognizes the order of unclear information point based on its nearest neighbor whose class is as of now known. It makes usage of the more than one nearest neighbor to decide the class in which the given information point has a place with and subsequently it is called as KNN.
The Euclidean distance between the points x and u is

$$d(x,u) = \sqrt{\sum_{i=1}^{n}(x_i - u_i)^2} \qquad (4)$$

D. Logistic Regression (LR)

Logistic Regression is utilized to portray information and to clarify the connection between one dependent binary variable and at least one nominal, ordinal, interim or proportion level autonomous factors. LR is a factual strategy for breaking down a dataset in which there is at least one autonomous factor that decide a result. The result is estimated with a dichotomous variable.
The "logit" function is given below

$$ln\left[\frac{p}{(1-p)}\right] = \alpha + \beta X + e \qquad (5)$$

Where, p is the probability that the event Y occurs, p(Y=1)
p/(1-p) is the "odds ratio"
ln[p/(1-p)] is the log odds ratio, or "logit"
The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
The estimated probability is:

$$p = \frac{1}{[1 + \exp(-\alpha - \beta X)]} \qquad (6)$$

Where if $\alpha + \beta X = 0$, then p = .50
as $\alpha + \beta X$ gets really big, p approaches 1 , as $\alpha + \beta X$ gets really small, p approaches 0.

E. Open Source Distribution for ML Predictive Models

For ML predictive models, Anaconda Jupyter is used. Anaconda is open supply circulation of the Python and R programming languages for information technology and device studying related applications. The Jupyter notebook is open-source web software that lets in you to create and percentage files that include stay code, equations, visualizations and all. Four supervised machine learning algorithms: DT, NB, KNN and LR are applied.

First dummy variables are created for categorical attributes using pandas in python. Input and output variables are created using pandas. Input variable consists of values of all features except class label. Output variable with class label is created. Data set is divided into training and testing dataset. To create the models machine learning algorithms are applied on training data set. Classifier's performance is measured with the help of testing dataset. Graph of all confusion matrices and Receiver Operating Characteristics (ROC) are plotted for interpretations of accuracies.

V. EXPERIMENTAL RESULTS

To monitor physical location i.e. home, sensor values are collected for one complete month and to detect outliers, if any, box whisker plots are created for temperature, humidity, LDR, and gas sensor values.

A. Box Whisker Plots

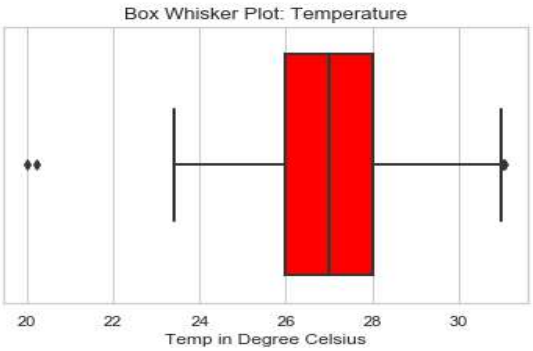


Fig. 5 Box Whisker Plot for Temperature

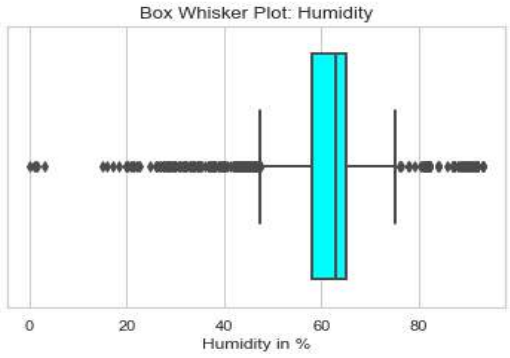


Fig. 6 Box Whisker Plot for Humidity

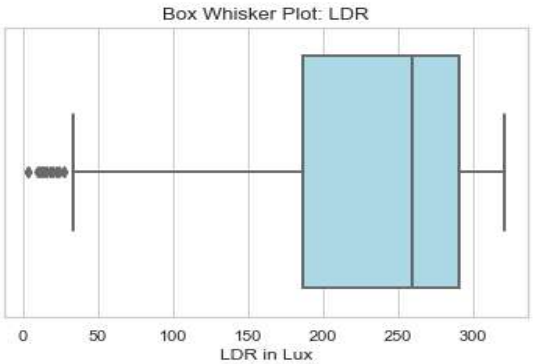


Fig. 7 Box Whisker Plot for LDR

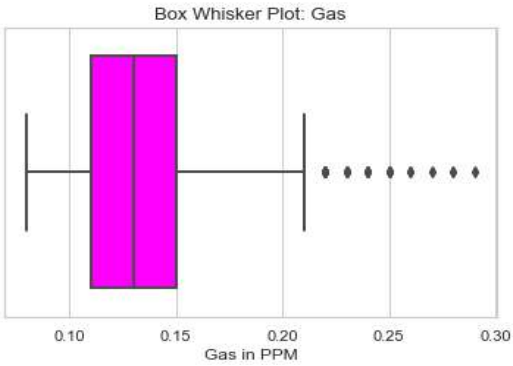


Fig. 8 Box Whisker Plot for Gas

Fig. 5, Fig. 6, Fig. 7, and Fig. 8 show box-whisker plots of temperature, humidity, LDR, and gas sensors respectively. Box whisker plot divides entire data into four regions and every region consists of 25% of total data. The first region is from minimum value to first quartile (Q1), second region is from first quartile to median (Q2), the third region is from median to third quartile (Q3) and fourth region is from third quartile to maximum value. Difference between Q3 and Q1 is called as Inter Quartile Range (IQR).

$$IQR = Q_3 - Q_1 \qquad (7)$$

The data points having values less than 1.5 times IQR and values greater than 1.5 times IQR are called as outliers. Outliers indicate unusual happenings in the environment where the system is located.

$$Outliers < 1.5 * IQR \qquad (8)$$

$$Outliers > 1.5 * IQR \qquad (9)$$

Fig. 5 and Fig. 8 show that very few outliers exist for temperature and gas sensor values. Fig. 6 shows too many outliers exist for humidity sensor and Fig. 7 shows outliers for LDR more than temperature and gas sensor values but less than humidity sensor values.
Various experiments are also carried out to assess classification accuracy, classification report, ROC curves, evaluation and the analytical model selection based on ML classifiers.

B. Classification Accuracy

A confusion matrix is an abstract of forecast results on a classification problem. It is a two dimensional matrix of order 2*2 for binary classification problem. Row is reserved to indicate actual values of negative and positive samples. Column is reserved to indicate predicted values of negative and positive samples. Matrix is divided into four cells such as True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) respectively. Entries along the diagonal from left most upper corner to right most bottommost corner represent true entries representing either TN or TP otherwise remaining entries are false. FPs are called as type-I error and FNs are called as type-II error.

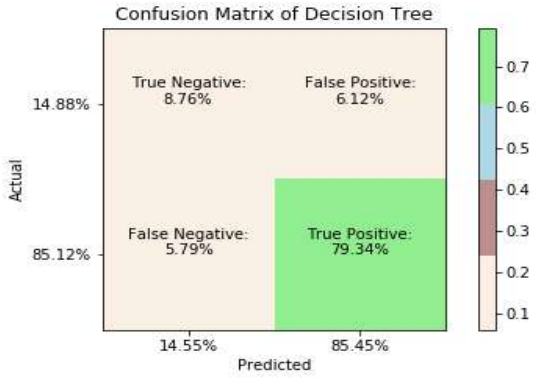


Fig. 9 Confusion Matrix of DT

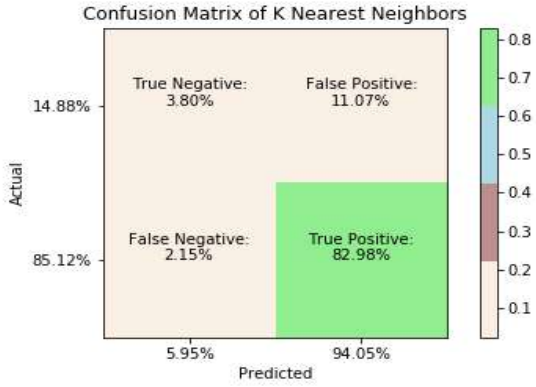


Fig. 10 Confusion Matrix of KNN

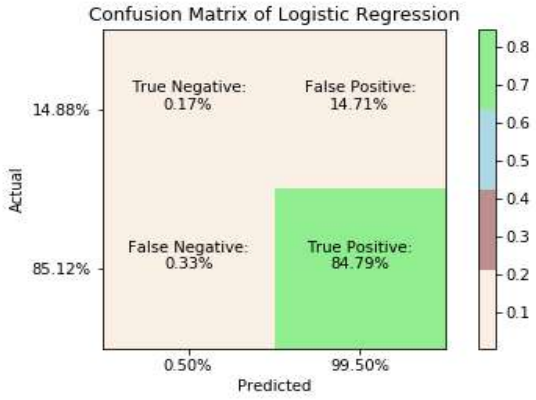


Fig. 11 Confusion Matrix of LR

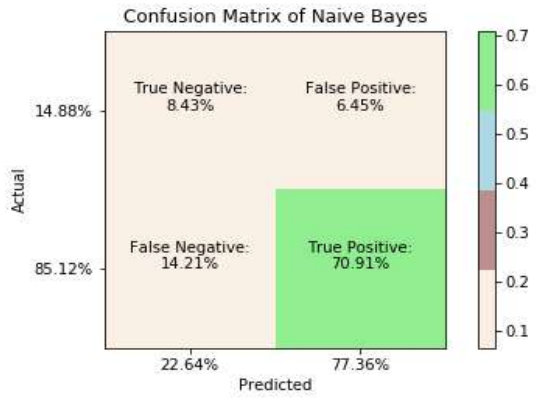


Fig. 12 Confusion Matrix of NB

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \qquad (10)$$

Accuracy of a classifier is the ratio of summation of TP and TN to total number of samples or instances. Accuracy is also called as recognition rate which specifies the proportion of total samples that are correctly identified. Misclassification rate is the difference between 1 and recognition rate.

Fig. 9, Fig. 10, Fig. 11 and Fig. 12 show confusion matrices of ML classifiers: DT, KNN, LR and NB respectively. Out of total instances, there are 85.12 % instances are true instances and 14.88 % instances are false instances. Fig. 9 shows that 8.76 % instances are correctly predicted as false instances and 79.34 % instances are correctly predicted as true instances for DT. Fig. 10 shows that 3.80 % instances are correctly predicted as false instances and 82.98 % instances are correctly predicted as true instances for KNN. Fig. 11 shows that 0.17 % instances are correctly predicted as false instances and 84.79 % instances are correctly predicted as true instances for LR. Fig. 12 shows that 8.43 % instances are correctly predicted as false instances and 70.91% instances are correctly predicted as true instances for NB.

C. Classification Report

The classification report summarizes, gives the precision, recall, f1-score, and support for the model. Precision is a classifier's ability not to label a positive instance which is in fact negative. It is the percentage of predicted positive instances that are correctly predicted as true positives. It is the ratio of true positive values to the summation of true and false positive values. Precision is also called as Positive Predictive Value (PPV).

$$Precision = PPV = \frac{TP}{TP + FP} \qquad (11)$$

Recall is a classifier's ability to find all positive events. It is the ratio of true positive values to the summation of true positive and false negative values. Sensitivity also called the True Positive Rate (TPR), the recall, measures the proportion of actual positives that are correctly classified as true positives. In binary classification, recall of the positive category is also recognized as sensitivity and recall of the negative category is recognized specificity.

$$Sensitivity = TPR = \frac{TP}{TP + FN} \qquad (12)$$

$$Specificity = TNR = \frac{TN}{TN + FP} \qquad (13)$$

Specificity also called the True Negative Rate (TNR) measures the proportion of actual negatives that are correctly classified as true negatives. Another appraise is F1-score which is the harmonic mean of precision and recall such that greatest score is 1.0 and bad score is 0.0.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (14)$$

In general, F1-scores are lower than accuracy measures as they include precision and recall into their calculations but they can be used to evaluate classifier models, not universal accuracy. Support is the number of real class occurrences in the data set specified. The support does not alter between models, but diagnoses the evaluation procedure as an alternative.

Table 1: Classification Report of DT

	Precision	Recall	F1-Score	Support
FALSE	0.60	0.59	0.60	90
TRUE	0.93	0.93	0.93	515
Avg / Total	0.88	0.88	0.88	605

Table 2: Classification Report of KNN

	Precision	Recall	F1-Score	Support
FALSE	0.64	0.26	0.37	90
TRUE	0.88	0.97	0.93	515
Avg / Total	0.85	0.87	0.84	605

Table 3: Classification Report of LR

	Precision	Recall	F1-Score	Support
FALSE	0.33	0.01	0.02	90
TRUE	0.85	1.00	0.92	515
Avg / Total	0.77	0.85	0.79	605

Table 4: Classification Report of NB

	Precision	Recall	F1-Score	Support
FALSE	0.37	0.57	0.45	90
TRUE	0.92	0.83	0.87	515
Avg / Total	0.84	0.79	0.81	605

Table 1, Table 2, Table 3 and Table 4 show the classification report of DT, KNN, LR and NB respectively. Table 1 shows that TPR and PPV of a DT is 93 % each. Table 2 shows that TPR of a KNN is 97 % and PPV is 88%. Table 3 shows that TPR of a LR is 100 % and PPV is 85%. Table 4 shows that TPR of a NB is 83 % and PPV is 92%. System's major class of interest is TRUE class and task is to minimize False Positives as well as False Negatives as much as possible to get good performance of a model. There is a trade-off between PPV and TPR. Among four predictive models, DT gives good results for TPR as well as PPV.

D. ROC Curve

The ROC curve is formed by plotting the True Positive Rate (TPR) along Y axis and the False Positive Rate (FPR) along X axis at various threshold settings. The ROC curve is thus the sensitivity as a function of FPR. The model with TPR=1 and FPR=0 is called as perfect model. Area Under Curve (AUC) is applied in classification examination in order to find out which of the used models predicts the best results. Fig. 13, Fig. 14, Fig. 15 and Fig. 16 show ROC Curves of ML classifiers: DT, KNN, LR and NB respectively where dotted line indicates the random guessing with AUC=0.5. Curve below the dotted line indicates bad performance of a model and curve above the random guessing shows good performance of a model. Fig. 13 shows that AUC of a DT is 0.76. Fig. 14 shows that AUC of a KNN is 0.62. Fig. 15 shows that AUC of a LR is 0.50. Fig. 16 shows that AUC of a NB is 0.70. Among four predictive models, Decision Tree gives the maximum AUC.

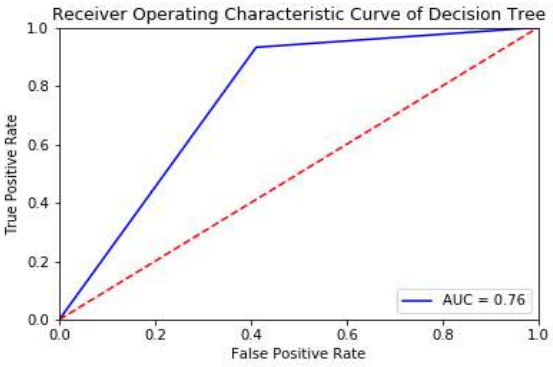


Fig. 13 ROC Curve of DT

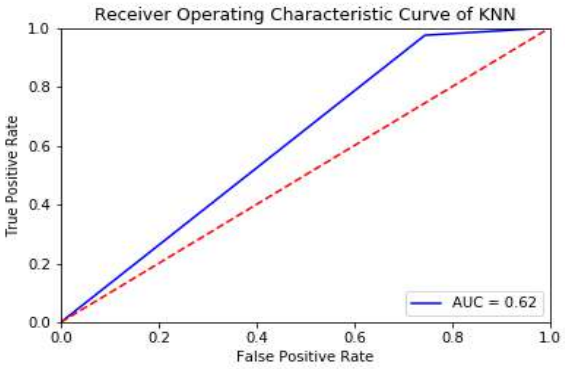


Fig. 14 ROC Curve of KNN

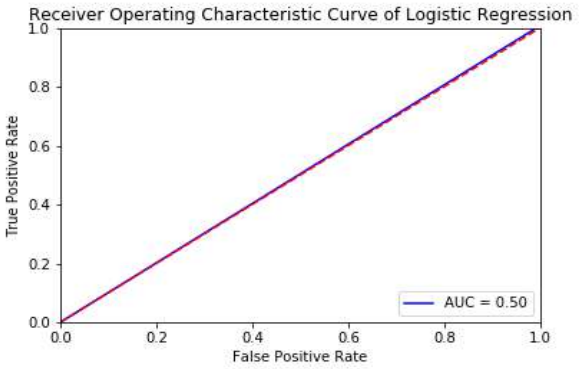


Fig. 15 ROC Curve of LR

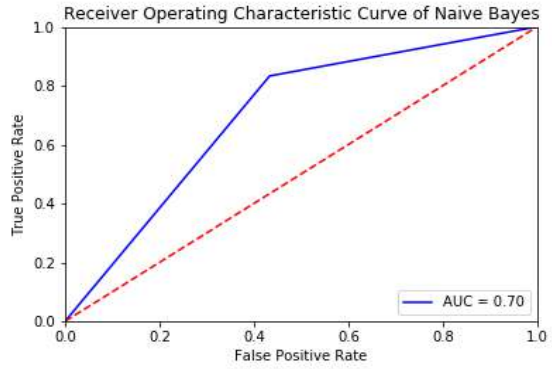


Fig. 16 ROC Curve of NB

E. Comparison and Selection of the Best ML Model

Fig. 17 shows that accuracy performance comparison of four ML models. In holdout method, the dataset is distributed into separate training and testing dataset - the former is used for creation of model and later is used to estimate its performance. As dataset applied belongs to imbalanced binary class, stratified k-fold cross validations is also used to compare and select the model effectively. In stratified k fold cross validation, k indicates number of folds which is equal to 10 and the type proportions are maintained in every fold to make sure that each fold is consultant of the category proportions in the training dataset. Among these four models, DT to be the best model for prediction of human face recognition with the highest accuracy of 87.77 % using hold out method and 89.81 % using stratified k-fold cross validation. Precision and recall of DT model is also very good as compared to other ML models. It is also observed that AUC of DT is the highest with 0.76 units. The model having curve nearest to the uppermost left area indicates the best performance. So DT is proved to be the best model with the highest classification accuracy, very good recall, precision, and highest AUC.

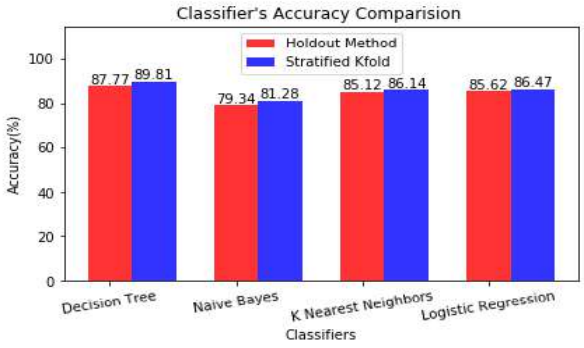


Fig. 17 Accuracy Comparison Chart of Four Classifiers

VI. CONCLUSION

Live sensor captured data along with multimedia data is useful for analysis of abnormal conditions in the environment of various physical locations. Sensor data analysis with the support of digital images of human face for human presence detection and recognition is useful for confirmation of abnormal conditions in the surroundings. As only sensitive information is pushed on to the cloud whenever the human presence is recognized either as known or unknown face, optimization of IoT enabled physical location monitoring is achieved because only the subset of original dataset is stored onto the cloud. The projected scheme is new and competent because it offers proper accuracies of classifiers and effectiveness of the approach. The Decision Trees, amongst the quite a number of analytical models, is a remarkable method for the evaluation of multimedia sensor records, with the maximum correctness of 87.77 % using hold out method and 89.81 % using stratified 10-fold cross validation, very good TPR and PPV of 0.93 each and ROC with AUC of 0.76, followed by NB, KNN and LR respectively. The prediction of person who is either known or unknown using sensor data analysis in the physical location, sending notifications and alert messages to mobiles and email accounts will be extended work of this system to enhance the robustness.

REFERENCES

1. Sankar Mukherjee , G.P. Biswas, "Networking for IoT and applications using existing communication technology", Egyptian Informatics Journal 19 (2018) 107–127.
2. Neelesh Mioshra, Lal Pratap Varma, Prabhat Kumar Srivastava, Ajay Gupta, "An Analysis of IoT Congestion Control Policies", Procedia Computer Science 132 (2018) 444–450
3. Dragos Mocrii, Yuxiang Chen, Petr Musilek, "IoT-based smart homes: A review of system architecture, software, communications, privacy and security", Internet of Things 1–2 (2018) 81–98
4. Adel Alkhalil, Rabie A. Ramadan, "IoT Data Provenance Implementation Challenges", Procedia Computer Science 109C (2017) 1134–1139.
5. Nallapaneni Manoj Kumar,Pradeep Kumar Mallick, " Blockchain technology for security issues and challenges in IoT"International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Procedia Computer Science 132 (2018) 1815–1823.
6. Mustafa Alper Akkaş, Radosveta Sokullu, "An IoT-based greenhouse monitoring system with Micaz motes", International Workshop on IoT, M2M and Healthcare (IMH 2017), Procedia Computer Science 113 (2017) 603–608
7. Partha Pratim Ray, "A survey of IoT cloud platforms", Future Computing and Informatics Journal 1 (2016) 35-46.
8. Nawaz Mohamudally, Mahejabeen Peermamode-Mohaboob, " Building An Anomaly Detection Engine (ADE) For IoT Smart", The 15th International Conference on Mobile Systems and Pervasive Computing, Procedia Computer Science 134 (2018) 10–17.
9. Mohammad Saeid Mahdavinejad , Mohammadreza Rezvan, Mohammadamin Barekatain , Peyman Adibi, PayamBarnaghi, Amit P. Sheth, "Machine learning for internet of things data analysis: a survey", Digital Communications and Networks 4 (2018) 161–175.
10. Bill Karakostas, "Event prediction in an IoT environment using naïve Bayesian models", The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016), Procedia Computer Science 83 (2016) 11 – 17.
11. Huseyin Yildirim, Amr M.T. Ali-Eldin, "A model for predicting user intention to use wearable IoT devices at the workplace", Journal of King Saud University-Computer and Information Sciences (2018)
12. Ajitkumar S. Shitole, Dr. Kamatchi R. Iyer, "SMART HOME CONTEXT-AWARE AUTOMATION BY CUSTOMIZATION STRATEGY", International Journal of Advanced in Management, Technology and Engineering Sciences(pp. 277-283), April 2018, ISSN NO: 2249-7455.
13. Devare M. (2018). Analysis and Design of IoT Based Physical Location Monitoring System. In Lucio Grandinetti, Seyedeh Leili Mirtaheri, Reza Shahbazian, Thomas Sterling, Vladimir Voevodin (Eds.), Advances in Parallel Computing,Volume 33: Big Data and HPC: Ecosystem and Convergence (pp. 120 - 136). IOS Press. doi: 10.3233/978-1-61499-882-2-120.
14. Ajitkumar S. Shitole, Manoj Devare, "Machine Learning Supported Statistical Analysis of IoT Enabled Physical Location Monitoring Data", International Conference On Computational Vision and Bio Inspired Computing, Nov 2018.

15. Alexandra Moraru, Marko Pesko, Maria Porcius, Carolina Fortuna and DunjaMladenic,"Using Machine Learning on Sensor Data", Journal of Computing and Information Technology - CIT 18, 2010, 4, 341–347 doi:10.2498/cit.1001913, 341.
16. Joseph Siryani, Bereket Tanju, and Timothy Eveleigh,"A Machine Learning Decision-Support System Improves the Internet of Things', Smart Meter Operations", IEEE 2017.
17. Purnendu Shekhar Pandey, "Machine Learning and IoT for Prediction and Detection of Stress", 978-1-5386-3893-4/17/$31.00 ©2017 IEEE.
18. Go Takami, Moe Tokuoka, Hirotsugu Goto, Yuuichi Nozaka, "Machine Learning Applied to Sensor Data Analysis ",Yokogawa Technical Report English Edition Vol. 59 No. 1(2016), pp. 27-30.
19. Rui Madeira, Luís Nunes, "A Machine Learning Approach for Indirect Human Presence Detection Using IOT Devices", The Eleventh International Conference on Digital Information Management 2016, pp. 145-150.
20. Che-Min Chung, Cai-Cing Chen, Wei-Ping Shih, "Automated Machine Learning for Internet of Things", 2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW) pp. 295-296.

AUTHORS PROFILE

Ajitkumar S. Shitole is pursuing Ph.D. in CSE from Amity University Mumbai. He has published more than 21 research papers in various International Journals/ Conferences. His area of interest is Data Mining, Machine Learning, and Algorithms. Currently he is working as an Associate Professor in Computer Engineering Department at I^2IT Hinjawadi, Pune.

Dr. Manoj Devare currently holding position of Associate Professor at Amity Institute of Information Technology, Amity University Mumbai. He has been served as Post Doctorate Fellow at Centre of Excellence on HPC, University of Calabria, Italy. His research deals with Multi-Scale High Performance Computing in Grids, Virtualization, and Clouds. He is also involved in the review process of the International Journals papers, edited chapters from the reputed journals like Elsevier's Future Generation Computer Systems (FGCS), Springer's review system, International Journal of Computing, IEEE Transactions on Parallel and Distributed Systems (TPDS), and SCIT International Journal.

ISBN: 978-93-5351-521-8

Analyzing Impact of Proper and Improper Feature Scaling on Performance of IoT Enabled Physical Location Monitoring Data

Ajitkumar Shitole[#], Dr. Manoj Devare[$]
[#]*Research Scholar (CSE), Amity Uni. Mumbai, Asst. Prof., I²IT Hinjawadi, Pune*
ajitkumars@isquareit.edu.in
[$]*Associate Professor, AIIT, Amity University Mumbai.*
mhdevare@mum.amity.edu

ABSTRACT

Internet of Things (IoT) and Machine Learning (ML) are two promising technologies in today's globe. To monitor any physical location on regular time interval, IoT is used to capture real time sensor data aligned with timestamp and combined with multimedia data. Abnormal conditions in the environment can be recognized and ML is used to perform statistical as well as efficient analysis to produce quality decision support system. Data preprocessing before applying supervised ML algorithms is required to acquire quality data. Handling missing values, noise, outliers, converting categorical features into one or more new features, transformation, normalization, feature scaling and feature selection are the some data preprocessing techniques which are required to enhance the performance of the model. Four feature scaling techniques such as Min Max Scalar, Standard Scalar, Normalization, and Robust Scalar are used to understand the impact of proper and improper scaling on the performance of the binary as well as multi class classifier. Experimental results show that performance of the model is approximately remains same before and after applying feature scaling, if dataset is originally almost free from missing values, noise and outliers. Study also reveals that if improper scaling is used, it somehow diminishes performance of the model. Even though proper feature scaling improves the performance of the classifier, its computational time including data preprocessing and learning model is exceptionally high for very large datasets.

Keywords: Physical Location Monitoring, Data Preprocessing Techniques, Feature Scaling, Min Max Scalar, Standard Scalar, Normalization, Robust Scalar.

INTRODUCTION

IoT is the one of the developing and quickly creating innovation within the field of Data Innovation and Communication Designing. Parts of gadgets can be connected to each other with the support of IoT to communicate and deal their data and information. In today's life, it is essential to screen the physical area with the assistance of IoT where numbers of diverse sensors are linked to single board computer. Assessment of physical area is required in arrange to distinguish any anomalous conditions within the surroundings. Real time sensor and multimedia data is gathered regularly with the help of IoT system.

Data preprocessing may perhaps be information mining procedure that includes altering rough information into reasonable format. Real-world information is regularly poor, contradictory, and/or missing in certain behaviors or patterns, lacking feature values, noisy containing random error or outliers and is likely to hold several blunders. Information preprocessing could be an established policy of settling such issues. Information preprocessing plans rough information for further processing. In this paper, different feature scaling strategies are analyzed to observe the performance of the supervised machine learning algorithms using holdout method.

RELATED WORK

Ajitkumar Shitole and Manoj Devare (2018) portrayed around the watching of a physical region is not only a fundamental activity but also recommends imperative corrective measures after productive assessment, to remain absent from the encourage mishap within the materials additionally perils in nature. The sensor data caught as time course of action is helpful for examination of the atypical conditions within the environment. S. B. Kotsiantis et al. (2006) explained that there's much unimportant and redundant information display or noisy and untrustworthy information, at that point knowledge discovery construction of model is more troublesome and therefore data preprocessing becomes crucial task before building final model. Ventseslav Shopov and Vanya Markova (2013) presented that number of missing data points and number of outliers could have major impact on classification and predict performance of the model but data preprocessing shows the impact on performance of the models. Nazri Mohd Nawi et al. (2013) proved that efficiency of Artificial Neural Network can be improved with the help of proper data preprocessing techniques like Min Max, Z-score, and decimal scaling normalization. Shaik Shahul et al. (2016) applied different data preprocessing techniques on various datasets and showed that prediction of software cost judgment changed by using data preprocessing steps. Sven F. Crone et al. (2006) investigated the impact of various preprocessing methods of feature scaling, sampling, handling of nominal as well as continuous features on the model performance of decision trees, neural networks and support vector machines.

XIAOLONG XU et al. (2018) proposed new missing value imputation algorithm based on the verification series for auxiliary judgment of lost values. MIRIAM SEOANE SANTOS et al. (2019) explained about diverse approaches to synthetic missing data generation found in the literature and discussed their practical details, elaborating on their qualities and shortcomings. HUI LU et al. (2018) explained about outlier detection method which is based on Cross-correlation Analysis consists of three parts such as data preprocessing, outlier analysis, and outlier rank which in turn shows strong detection capability for high-dimensional time series datasets. RUBEN TOLOSANA et al. (2015) presented data preprocessing phase where data acquired from diverse devices is pre-processed in order to reach a high similarity between signatures coming from different devices. The second phase is a selection of the best features in order to further decrease the effect of device interoperability, selecting features which are robust in these conditions. PHILIP JORIS et al. (2018) attempted to handle heteroscedastic data more adequately and proposed the concepts of intensity-specific distributions and intensity-specific variances.

RESEARCH METHOD

Different sensors like temperature, humidity, Passive Infra Red (PIR), Light Dependent Resistor (LDR), Gas, and web camera to recognize the face of a person are connected to Raspberry Pi 3 to monitor physical location continuously in real time fashion to detect unusual conditions in the environment if any and same data is stored onto cloud as well as local server for further systematic analysis to extract meaningful information which supports decision making system. Face recognition library is installed in the python environment to recognize face of a human being in real time and aligned with sensor data in the form of 'Person' feature. Multithreading programming concept in python is used to create multiple threads for capturing live sensor data and face recognition efficiently. System was continuously observed in the home location for one month to determine whether the recognized face is true or false and same class label is manually added in the dataset to perform binary classification for evaluation of performance of the binary model. Data is also pushed onto cloud whenever face is recognized as true or false to optimize the system. Local dataset size is very large as compared to cloud dataset as every entry is stored over there irrespective of face recognition. Different data preprocessing and feature scaling techniques are applied on cloud dataset to observe the performance of binary and multi class classifiers with proper and improper scaling. General outline of the data preprocessing is shown in Figure1.

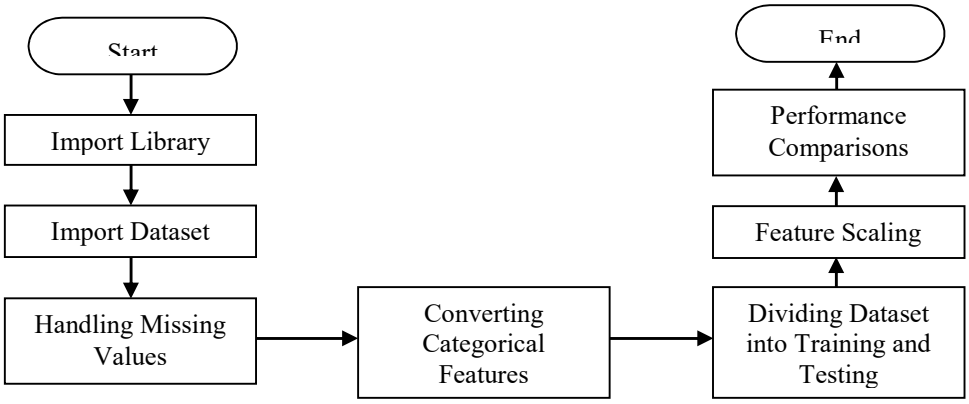


Figure1. Data Preprocessing Steps

As shown in the Figure1, first respective libraries and dataset in comma separated values (csv) file is imported using python. After data preprocessing like handling of missing values and converting categorical features into additional features, dataset is partitioned into training and testing sets to apply feature scaling techniques.

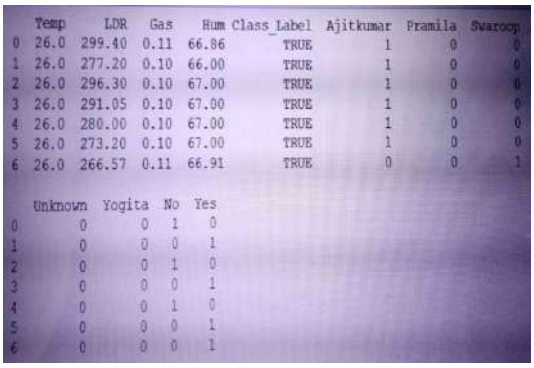
RESULT

Understanding Dataset and Data Preprocessing

Table 1 shows that portion of IoT enabled physical location monitoring dataset. Dataset is labeled and consists of eight features such as Person, Temp (^{0}C), LDR (Lux), Gas (PPM), PIR, Hum (%), Timestamp, and Class Label. Dataset is used for binary classification with Class Label as output feature and remaining features as input features. After removal of Class Label feature, same dataset is also used for multi class classification with Person feature as target feature to observe the performance of the classifier using proper and improper feature scaling techniques.

Table1. Portion of IoT enabled physical location monitoring dataset

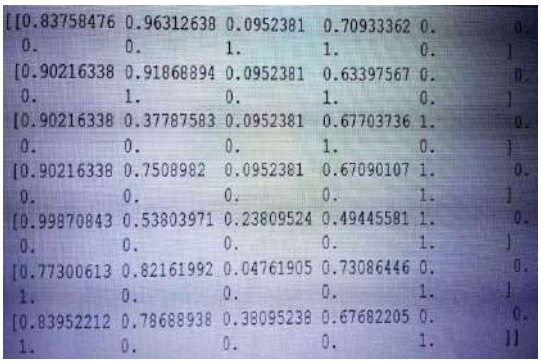
Person	Temp (^{O}C)	LDR (Lux)	Gas (PPM)	PIR	Hum (%)	Timestamp	Class Label
Ajitkumar	26	299.4	0.11	No	66.86	2018-09-13 10:59:19.319301	TRUE
Ajitkumar	26	277.2	0.1	Yes	66	2018-09-13 10:59:25.436754	TRUE
Ajitkumar	26	296.3	0.1	No	67	2018-09-13 10:59:31.036016	TRUE
Ajitkumar	26	291.05	0.1	Yes	67	2018-09-13 10:59:41.162160	TRUE
Ajitkumar	26	280	0.1	No	67	2018-09-13 10:59:46.225767	TRUE
Ajitkumar	26	273.2	0.1	Yes	67	2018-09-13 10:59:51.290994	TRUE
Swaroop	26	266.57	0.11	Yes	66.91	2018-09-13 11:03:06.093508	TRUE
Swaroop	26	261.15	0.12	No	66.75	2018-09-13 11:03:24.441326	TRUE
Swaroop	26	260	0.13	Yes	67	2018-09-13 11:03:29.496944	TRUE
Unknown	26	260.2	0.14	No	67	2018-09-13 11:03:47.653924	FALSE
Yogita	26	261.1	0.14	Yes	66	2018-09-13 11:03:53.520794	TRUE
Ajitkumar	26	299.4	0.11	No	66.86	2018-09-13 10:59:19.319301	TRUE



	Temp	LDR	Gas	Hum	Class_Label	Ajitkumar	Pramila	Swaroop
0	26.0	299.40	0.11	66.86	TRUE	1	0	0
1	26.0	277.20	0.10	66.00	TRUE	1	0	0
2	26.0	296.30	0.10	67.00	TRUE	1	0	0
3	26.0	291.05	0.10	67.00	TRUE	1	0	0
4	26.0	280.00	0.10	67.00	TRUE	1	0	0
5	26.0	273.20	0.10	67.00	TRUE	1	0	0
6	26.0	266.57	0.11	66.91	TRUE	0	0	1

	Unknown	Yogita	No	Yes
0	0	0	1	0
1	0	0	0	1
2	0	0	1	0
3	0	0	0	1
4	0	0	1	0
5	0	0	0	1
6	0	0	0	1

Figure2. Portion of dataset after one hot encoding with Person and PIR features



```
[[0.83758476 0.96312638 0.0952381  0.70933362 0.         0.
  0.         0.         1.         1.         0.         1.
 [0.90216338 0.91868894 0.0952381  0.63397567 0.         0.
  0.         1.         0.         1.         0.         1.
 [0.90216338 0.37787583 0.0952381  0.67703736 1.         0.
  0.         0.         0.         1.         0.         1.
 [0.90216338 0.7508982  0.0952381  0.67090107 1.         0.
  0.         0.         0.         0.         1.         1.
 [0.99870843 0.53803971 0.23809524 0.49445581 1.         0.
  0.         0.         0.         0.         1.         1.
 [0.77300613 0.82161992 0.04761905 0.73086446 0.         0.
  1.         0.         0.         0.         1.         1.
 [0.83952212 0.78688938 0.38095238 0.67682205 0.         0.
  1.         0.         0.         0.         1.         ]]
```

Figure3. Portion of dataset after feature scaling: Min Max Scalar

As dataset is free from the missing values, there is no need to handle missing values. Apart from Timestamp and Class Label features, it consists of two categorical features such as Person and PIR which are converted into one or more new features using dummy variables with the help of pandas in scikit-learn library of anaconda jupyter environment. Figure2 shows portion of dataset after one hot encoding that is dummy variables with Person and PIR features. Person feature is now converted into five more additional features 'Ajitkumar', 'Pramila', 'Swaroop', 'Unknown', 'Yogita' having values 0 and 1 based on the name of the person recognized. Similarly PIR feature is also converted into two more additional features 'No' and 'Yes'.

Feature Scaling

Temp, LDR, Gas, and Hum features are numeric and having different scale of measurements. It is necessary to bring all numeric features into same standardize scale before building the models to enhance the performance of the same. Decision Tree, Random Forest, and Gradient Boosting are some of the supervised ML algorithms which don't need to perform feature scaling to bring them onto same scale. Feature scaling is necessary for K Nearest Neighbor (KNN), Logistic Regression (LR), Artificial Neural Network (ANN) and Support Vector Machine (SVM) to enhance the performance metric. To understand the impact of proper and improper feature scaling, binary classification is performed on KNN and LR where as multi class classification is performed only on KNN. Four different scaling techniques are applied to reveal the performance of the models. Dataset is now separated into input and output features. Figure3 shows portion of dataset after Min Max feature scaling on input features. As shown in the Figure3, all input feature values are converted into the scale of 0 to 1.

Min Max Scalar

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \ (new\_\max(x) - new\_\min(x)) + new\_\min(x) \qquad (1)$$

Where x is original value, x' is normalized value, $\min(x)$ is the minimum value and $\max(x)$ is the maximum value of that feature, $new\_\max(x)$ is the new maximum value and $new\_\min(x)$ is new minimum value.

Min Max scalar is the simplest feature scaling and uses minimum and maximum values to rescale the values in the range usually [0, 1] or [-1, 1]. In the event that the dispersion is not Gaussian or the standard deviation is little, the min-max scalar works better.

Standard Scalar

$$x' = \frac{x - \mu}{\sigma} \qquad (2)$$

Where x is original value, x' is normalized value, μ is mean, and σ is standard deviation.
Standard scalar uses mean and the standard deviation of the feature vector and rescale the values in such a way that data distribution is centered on zero mean and unit variance so that features take the shape of normal distribution. It preserves the valuable information of outliers and makes algorithms fewer sensitive to them.

Normalization

$$x' = \frac{x}{\|x\|} \qquad (3)$$

Where x is original value, x' is normalized value, and $\|x\|$ is Euclidean length of the feature. It projects samples on the circle or sphere having radius equal to one and usually used when direction of the data matters.

Robust Scalar

$$x' = \frac{x - Q_1(x)}{Q_3(x) - Q_1(x)} \qquad (4)$$

Where x is original value, x' is normalized value, $Q_1(x)$ is first quartile, $Q_3(x)$ is third quartile.
Robust scalar uses quartiles instead of mean and standard deviation to rescale the points. It ignores the very extreme data points called as outliers and makes difficulty to further scaling techniques.

Effect of Proper and Improper Feature Scaling on Performance of Binary Classifier

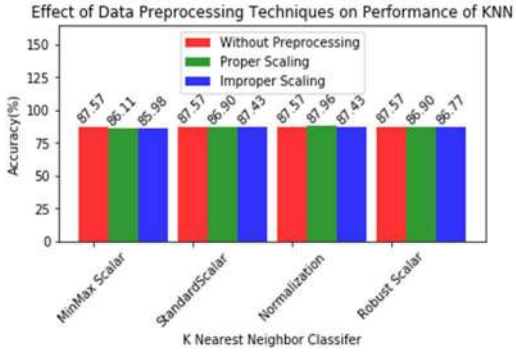


Figure4. Effect of feature scaling techniques on performance of KNN binary classifier

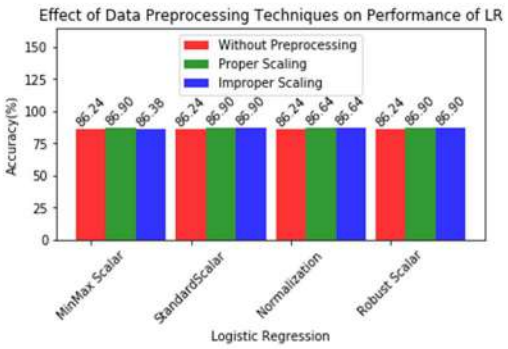


Figure5. Effect of feature scaling techniques on performance of LR binary classifier

Dataset is partitioned into training and testing sets using hold out method and then feature scaling techniques are applied. Feature scaling techniques supports fit and transform methods to rescale the original data into new representations. Proper scaling allows applying fit method on training dataset followed by transform method on training as well as testing datasets. In improper scaling, fit method is applied on training and testing datasets separately followed by transform method which most of the times degrade the performance of the models. Figure4 and Figure5 show the effect of feature scaling techniques on performance of KNN and LR binary classifiers respectively. Performance of the binary classifiers is measured in terms of accuracy. Accuracy is the proportion of total number of samples that are correctly identified by the model. Both figure show that there is no significant difference in performance of classifiers before and after feature scaling. Figure4 shows that accuracy of KNN classifier before normalization is 87.57% and after proper normalization there is small enhancement in accuracy which is 87.96%, but it decreases somewhat after improper normalization. Figure5 shows that accuracy of LR classifier before applying any feature scaling techniques is 86.24% and after proper scaling it has been slightly increased to 86.90% using all four feature scaling techniques. Improper scaling of Min Max scalar brings down performance of LR classifier slightly.

Effect of Proper and Improper Feature Scaling on Performance of Multi Class Classifier

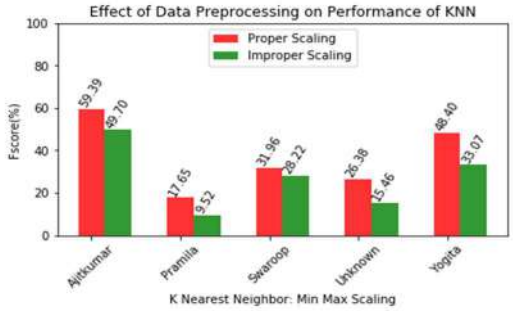


Figure6. Effect of proper and improper Min Max Scaling on f-score of KNN multi class classifier

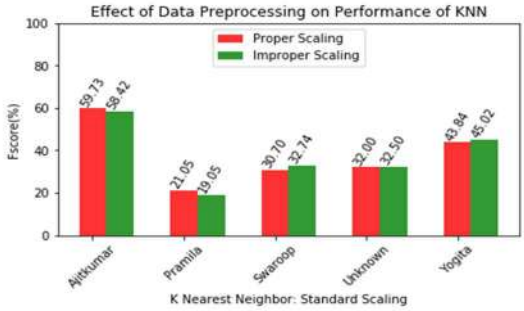


Figure7. Effect of proper and improper Standard Scaling on f-score of KNN multi class classifier

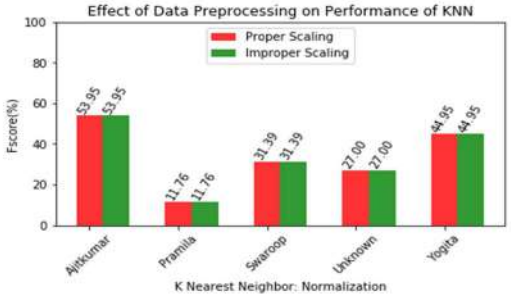


Figure8. Effect of proper and improper Normalization on f-score of KNN multi class classifier

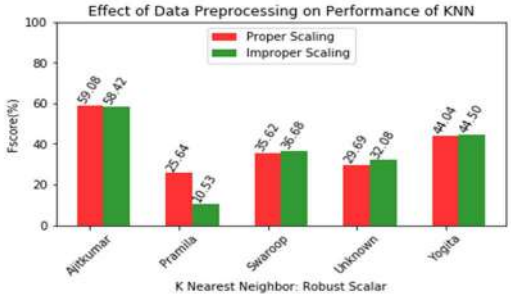


Figure9. Effect of proper and improper Robust Scaling on f-score of KNN multi class classifier

One vs. all multi class classifier is used and f-score of the KNN classifier is compared to analyze the impact of the proper and improper scaling. Multi class classification consists of five classes such as 'Ajitkumar', 'Pramila', 'Swaroop', 'Unknown', and 'Yogita' for which harmonic mean of precision and recall are calculated to get f-score. Figure6, Figure7, Figure8, and Figure9 show the effect of proper and improper scaling on performance of KNN multi class classifier using Min Max, Standard, Normalization, and Robust Scaling respectively. Figure6 depicts that performance of the KNN model is reasonably high using proper

min max feature scaling than its improper scaling across all class labels. Figure7 reveals that for few classes there is little growth in performance using proper standard feature scaling. Figure8 tells that there is no change in performance using proper and improper normalization. Figure9 explains that there are variations in performance using proper and improper robust feature scaling.

CONCLUSION

To enhance the performance of the supervised machine learning models, data preprocessing and feature scaling techniques are important aspects to be applied before construction of the models. Robustness of the model depends on quality of the data. If the data is missing, noisy, contains outliers or consists of contradictory data, it may degrade the performance and to resolve these issues data preprocessing and proper feature scaling techniques are required. Categorical features like nominal and ordinal must be converted into additional features having values 0 and 1 using either one hot encoding or dummy variables. Feature scaling is utilized for bringing all features into same scale to create robust model. SVM, KNN, LR, and ANN are sensitive to feature scaling. Proper and improper scaling of Min Max Scalar, Standard Scalar, Normalization, and Robust Scalar give variations in performance of the model. Impact of proper and improper feature scaling is tested and analyzed for binary as well as multi class classifier and results show that proper scaling gives better enhancement of performance metric over improper scaling.

REFERENCES

Ajitkumar S. Shitole, Manoj Devare (2018). "Machine Learning Supported Statistical Analysis of IoT Enabled Physical Location Monitoring Data", International Conference On Computational Vision and Bio Inspired Computing.

Hui Lu, Yaxian Liu, Zongming Fei, And Chongchong Guan (2018). "An Outlier Detection Algorithm Based on Cross-Correlation Analysis for Time Series Dataset", *10.1109/ACCESS.2018.2870151,* Volume 6, 53593-53610.

Miriam Seoane Santos, Ricardo Cardoso Pereira, Adriana Fonseca Costa, Jastin Pompeu Soares, João Santos, And Pedro Henriques Abreu (2019). "Generating Synthetic Missing Data: A Review by Missing Mechanism", Volume 7, 11651-11667

Nazri Mohd Nawi, Walid Hasen Atomi, M. Z. Rehman (2013). "The Effect of Data Pre-Processing on Optimized Training of Artificial Neural Networks", The 4th International Conference of Electrical Engineering and Informatics (ICEEI 2013), Procedia Technology 11 (2013) 32–39.

Philip Joris, Wim Develter, Wim Van De Voorde, Paul Suetens, Frederik Maes, Dirk Vandermeulen, And Peter Claes (2018). "Preprocessing of Heteroscedastic Medical Images", *10.1109/ACCESS.2018.2833286,* Volume 6, 26047-26058.

Ruben Tolosana, Ruben Vera-Rodriguez, Javier Ortega-Garcia, And Julian Fierrez (2015). "Preprocessing and Feature Selection for Improved Sensor Interoperability in Online Biometric Signature Verification", *10.1109/ACCESS.2015.2431493,* Volume 3, 478-489.

S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas (2006). "Data Preprocessing for Supervised Leaning", International Journal of Computer Science Volume 1 Number 1 2006 ISSN 1306-4428.

Shaik Shahul, Sristi Suneel, M.A. Rahaman, Swathi J. N (2016). "A Study of Data Pre-Processing Techniques for Machine Learning Algorithm to Predict Software Effort Estimation", Imperial Journal of Interdisciplinary Research(IJIR), Vol-2, Issue-6, 2016.

Sven F. Crone, Stefan Lessmann, Robert Stahlbock (2006). "The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing", European Journal of Operational Research 173 (2006) 781–800.

Ventseslav Shopov and Vanya Markova (2013). "Impact Of Data Preprocessing On Machine Learning Performance", Proceedings of the International conference on Information Technologies (InfoTech-2013) 19-20 September 2013, Bulgaria.

Xiaolong Xu, Weizhi Chong, Shancang Li, Abdullahi Arabo, And Jianyu Xiao (2018). "Miaec: Missing Data Imputation Based on the Evidence Chain", *10.1109/ACCESS.2018.2803755,* Volume 6, pp.12983-12992.

Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018)
IEEE Xplore Compliant Part Number: CFP18K74-ART; ISBN:978-1-5386-2842-3

Prediction of Antimicrobial Resistance for disease-causing agents using Machine Learning

Using Machine Learning algorithms to predict individuals susceptibility of developing AMR for drugs

Srajan Kulshrestha
Department of Computer Engineering
International Institute of Information Technology
Pune, India
srajan.kulshreshtha@gmail.com

Dinesh Nayar
Department of Computer Engineering
International Institute of Information Technology
Pune, India
dineshnyr@gmail.com

Sanjana Panda
Department of Computer Engineering
International Institute of Information Technology
Pune, India
sanjana.96panda@gmail.com

Dr. Vaishali Dohe
Associate Prof. Department of Microbiology
B. J. Medical College
Pune, India

Ashwini Jarali
Asst. Prof. Department of Computer Engineering
International Institute of Information Technology
Pune, India

***Abstract*—Antimicrobial resistance (AMR) occurs when disease-causing microorganisms are resistant towards prescribed drugs, nullifying its effect. As a consequence, there is a delay in recovery which worsens the patient's health. Antimicrobial resistance is identified as a global threat by the medical fraternity and various government bodies.**

Objective of the proposed system is to integrate technology with the field of bio-medical, in context with AMR. We applied various machine learning algorithms on datasets, to identify patterns and use them to predict resistance towards various drugs. This model would help in closing the gap between Doctors and Labs.

In this model, we used ML and data mining techniques to predict AMR for individual patients based on trends identified from datasets. For building the model we use results of Patients undergoing antibiotic susceptibility test as datasets.

***Keywords*—Machine Learning, Classification, Decision Tree, Association Rule, Apriori Algorithm, Data Mining, Pathogens, Drugs, Combination Therapy, Antimicrobial Resistance, CLSI Guidelines, Staphylococcus aureus (sau), Pseudomonas aeruginosa elastase (pae).**

I. INTRODUCTION

Pathogens are disease-causing microorganisms. An antimicrobial is that agent that prevents the multiplication effect of pathogens. Antimicrobial resistance (AMR) is a Darwinian selection process that pathogens adapt to survive. Pathogens show resistant towards the drugs they are exposed to. Hence, these drugs have no effect on the patient, delaying the healing process. Sometimes, this may be fatal to the patient's life. In a recent report, it is said that an estimated population equivalent to 10 million people will perish by 2050 because of AMR. [1]

One of the major triggers for AMR is the use, misuse, or overuse of antimicrobial drugs. Antimicrobial resistance has become a serious global threat. World Health Organization (WHO) and governments of various countries have also identified this threat.

Knowledge engineering methods can prove to become a powerful tool in finding unexpected patterns and hidden knowledge, and establishing new rules from large datasets.

There are antibiotic susceptibility tests that determine AMR for patients, but these tests take time to deliver results.

Doctors constantly deal with impatient Patients, who do not wait for elementary diagnostic results, they want instant results. Due to which there is a gap between doctors and clinics. Our project aims to bridge this gap, by creating a tool that predicts AMR in Patient. In this model, we will use ML algorithms to classify drugs for individual patients, into two groups Resistant and Sensitive. A Resistant classified drug means the patient has grown resistive towards the drug. Similarly, a Sensitive classified drug implies that the patient is responding towards the drug. We would be applying classification algorithms on the datasets, to predict resistance towards certain drugs.

II. LITERATURE SURVEY

Antibiotics have been used since the 1940s. Since then, deaths from several infections and illness have been significantly reduced. Various reasons contribute to the growing Antibiotic resistance. Over-prescription of antibiotics is a problem that has to be tackled [2]. 2010 study results declared India as the world's biggest consumer of antibiotics for human health at 10.7 units per person [3]. Patients often do not complete their entire antibiotic course which allows the strongest bacteria to survive [4]. A survey conducted in 2015 by WHO which involved multiple countries suggested that there was a ubiquitous public misconception about antibiotic usage and resistance. The results indicated that 42% do not know that they should stop taking antibiotics only when they complete the dosage as administered [5]. Once a patient has grown resistant towards a drug, there is a transfer of resistant determinants between microorganisms. This brings a change in the genome sequence of the patient. Interaction between humans, animals and agricultural host create a platform where resistant genes can be transferred thus habilitating the spread of resistance [6]. Antimicrobials are heavily used in animal food production industry for disease prevention, treatment, and growth promotion. But the large-scale use of antimicrobials in agriculture (livestock and fish farming) results in human exposure to antimicrobial-resistant bacteria via direct and indirect pathways. Poor infection prevention, control practices and unsanitary conditions in healthcare facilities are also responsible to further spread and increase of antimicrobial resistance. Also, as new, the rate at which bacteria are getting resistant to existing medicines is a lot faster than newer antibiotics are being developed [7]. Incorrect and excessive use of antibiotics, as well as poor infection control, has boosted antibiotic resistance. With proper steps, society can reduce and limit the spread of resistance. The World Health Organization has created various guidelines to help organizations tackle AMR [8]. To tackle the challenge of resistance and infections, antibiotic stewardship and hospital infection control have been deployed worldwide [9]. Persistent reconnaissance of local antimicrobial susceptibility patterns is a must for fighting rising antimicrobial resistance. WHONET is a compelling computerized microbiology research facility information administration and examination program that can give direction for empiric treatment of contaminations, alarm clinicians of patterns of antimicrobial resistance, direct drug-policy choices and preventive measures. The program encourages sharing of information among different hospitals by keeping a common format which can be collaborated for global or national antibiotic resistance surveillance [10]. In a paper published by the University of Athens proposed a framework in which data produced by various hospitals were integrated into a data warehouse and data mining approaches like Apriori algorithm was used to detect hidden and previously unknown patterns on large datasets [11]. A paper published in PLOS used decisions trees to find special relationships among variables and were used to establish new rules from datasets [12].

Association rule learning is knowledge engineering method which uses rules to find interesting relations between different variables in large datasets. Apriori is an algorithm for frequent itemset mining and association rule learning over large databases [13]. Support says how popular an item set is in the datasets. Confidence says how likely item A is purchased when item B is purchased. Lift says how likely item A is purchased when item B is purchased [14].

III. PROPOSED SYSTEM

We aim to make a model that takes patient details and predicts whether the person is sensitive or resistant to first line of drugs of treatment. In an ideal system, the model needs to know vast patient history for training and accurate prediction of future instances but in India there is no system for keeping a track of patients, therefore we use results of antibiotic susceptibility test as datasets. To increase the efficiency, association rules were discovered among various drugs for individual micro-organisms. And these patterns were used to predict results for further drugs.

This section provides the description of components of proposed model, as illustrated in "Fig. 1," (A) Data Cleaning and Transformation, (B) Association Rules Generation, (C) Features Selection, (D) Model, (E) User Interface.

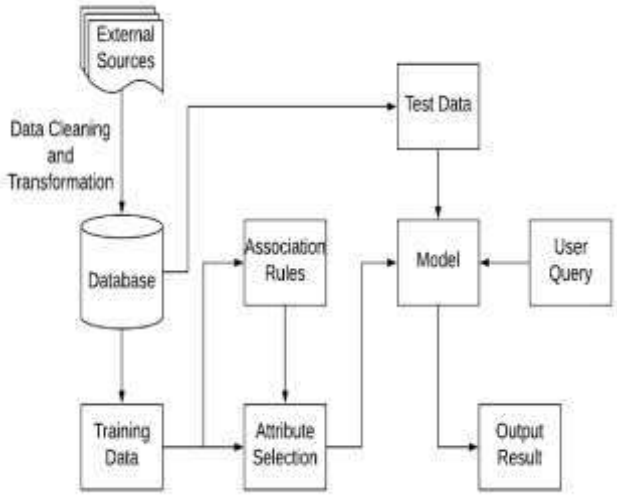


Fig. 1 Architecture Diagram

A. *Data Cleaning and Transformation*

Data from Antibiotic Susceptibility Test from various Patients were used. Data available was highly noisy data, hence cleaning was required. Mainly the data indicates sensitivity or resistivity of drugs. Various mathematical formulas were used with reference to CLSI guidelines to fill those empty data cells. Thereafter a transformed and cleaned data was created on which all further processing was carried out.

B. Generation of Association Rules

Patterns were generated amongst different drugs for individual micro-organisms using Apriori Algorithm. Hidden patterns among drugs were discovered. By this thousand of association rules were generated; filtering of patterns was done using support, confidence and lift as qualifying metrics.

C. Features Selection

Antecedents of identified association rules were used as attributes. In the system antecedents were the features and consequent was our target class. The qualifying metrics depends on the organism. Hence, for each organism a different Decision Tree was to be built.

D. Model

System works on data tailoring, association generation, and decision tree. Role of former two is explained now comes the decision tree.

The system was trained with data to build a decision tree classifier; it worked on entropy as split criteria. This trained model is now used to predict the target class for an unseen instance.

E. User Interface

Our system was highly client interactive in nature, end users being medical practitioner.

An interface is provided to feed details of any patient to the system such as his age, gender, site of infection, any prior known resistant drugs. These details are used as input to our model which in turn classify first line of drug for his/her treatment as resistant or sensitive as output on our robust dashboard.

Fig 2. Doctor Dashboard UI

IV. Mathematical Model

Decision tree algorithm works by recursive partitioning of data set into subsets. Each node of the tree is given a particular set of record T that is split by a specific test on feature.

Attributes are categorical in nature. Antibiotics can be split according to its nature. An antibiotic can belong to various subcategories i.e. Resistant, Sensitive, Intermediate or Not Determined. To split into entropy is used as criteria.

Let us consider following set of tuples:

S = {D, X, Y, F}

Where,

D = datasets

X = {basic patient information, site of infection, patient medical record}

Y = {pool of resistant drugs or sensitive drugs}

F = {data cleaning, apriori algorithm, decision classifier}

'D' is the training datasets with only essential features. 'X' is the input filled by the medical practitioner. 'Y' is the end result for any patient about resistivity or sensitivity of him to first line of drugs. 'F' is the various function used in the implementation of the system.

V. Results

Table I.

Organism	Features	Target Drug	Accuracy (%)
Sau	Penicillin, Clindamycin, Cefoxitin	Erythromycin	91.67
Sau	Clindamycin, Cefoxitin	Erythromycin	91
Sau	Clindamycin, Cefoxitin	Penicillin	85
Sau	Erythromycin, Cefoxitin	Penicillin	95.83
Pae	Aztreonam, Ceftazidime	Cefepime	96
Pae	Amikacin, Ceftazidime	Cefepime	92
Pae	Amikacin, Imipenem	Ceftazidime	76

As you can observe from the table, the average accuracy is around 85-90% for few microorganisms, the accuracy can be increased with more personalized Patient details.

VI. Conclusion

We believe that prediction of AMR can be a vital step to fighting AMR. It can also act as a tool to prevent AMR.

As stated earlier for building an ideal model to predict the resistivity or sensitivity of any drug in an individual requires vast data with minute details is needed. Such datasets require organizations to invest more in technology and build systems for the same.

Feedback loop can be used to increase discover new patterns and keep the system updated. As the patterns for AMR keeps changing a feedback loop will keep the model efficient.

VII. Acknowledgment

Our work wouldn't have been possible without the datasets is sourced from Microbiology Department, B. J. Medical College, Pune. We are deeply grateful of Doctor Renu Bharadwaj (HOD Microbiology Dept.) to believe in us and

give us suggestions and constant guidance with various aspects of AMR.

VIII. References

[1] Marlieke E. A. de Kraker, Andrew J. Stewardson, and Stephan Harbarth, "Will 10 Million People Die a Year due to Antimicrobial Resistance by 2050?" *ncbi.nlm.nih.gov*, Nov 13, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5127510/. [Accessed: Aug. 2, 2017]

[2] Andrew Duong and Michelle Jaelin, "6 Factors That Have Caused Antibiotic Resistance," *infection-control.tips*, Nov 18, 2015. [Online]. Available: https://infectioncontrol.tips/2015/11/18/6-factors-that-have-caused-antibiotic-resistance/. [Accessed: Aug. 10, 2017].

[3] Ramanan Laxminarayan and Ranjit Roy Chaudhury, "Antibiotic Resistance in India: Drivers and Opportunities for Action," *journals.plos.org*, March 2, 2016. [Online]. Available: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001974. [Accessed:20 Aug, 2017]

[4] "CDC: 1 in 3 antibiotic prescriptions unnecessary," May 3, 2016. [Online]. Available: https://www.cdc.gov/media/releases/2016/p0503-unnecessary-prescriptions.html. [Accessed: Aug 30, 2017]

[5] WHO, "Combating Antimicrobial Resistance in India," April 30, 2014. [Online]. Available: http://www.searo.who.int/india/topics/antimicrobial_resistance/Combating_Antimicrobial_Resistance_in_India/en/. [Accessed: Sept. 10, 2017]

[6] WHO, "The world health report 2007 - A safer future: global public health security in the 21st century," *who.int*, 2007. [Online]. Available: http://www.who.int/whr/2007/en/. [Accessed: Sept. 20, 2017]

[7] Charles H. Brower, Siddhartha Mandal, Shivdeep Hayer, Mandeep Sran, Asima Zehra, Sunny J. Patel, Ravneet Kaur, Leena Chatterjee, Savita Mishra, B.R. Das, Parminder Singh, Randhir Singh, J.P.S. Gill, and Ramanan Laxminarayan, "The Prevalence of Extended-Spectrum Beta-Lactamase-Producing Multidrug-Resistant Escherichia coli in Poultry Chickens and Variation According to Farming Practices in Punjab, India," *ehp.niehs.nih.gov*, July 2017. [Online].Available:https://ehp.niehs.nih.gov/ehp292/. [Accessed: Sept. 30, 2017]

[8] WHO, "Antibiotic resistance," *who.int*, Feb 5, 2018. [Online]. Available: http://www.who.int/en/news-room/fact-sheets/detail/antibiotic-resistance. [Accessed: Oct. 10, 2017]

[9] Sujith J. Chandy, Joy Sarojini Michael, Balaji Veeraraghavan, O.C. Abraham, Sagar S. Bachhav, and Nilima A. Kshirsagar, "ICMR programme on Antibiotic Stewardship, Prevention of Infection & Control (ASPIC)," *ncbi.nlm.nih.gov*, Feb 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4001333/. [Accessed: Oct 20, 2017].

[10] A Agarwal, K Kapila, and S Kumar, "WHONET Software for the Surveillance of Antimicrobial Susceptibility" *ncbi.nlm.nih.gov*, Jul 21, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4921382/. [Accessed: Oct 30, 2017].

[11] Eugenia G. Giannopoulou, Vasileios P. Kemerlis, and Michalis Polemis, "A Large Scale Data Mining Approach to Antibiotic Resistance Surveillance" Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07), June 15, 2007.

[12] Joana Rosado Coelho, João André Carriço, Daniel Knight, Jose-Luis Martínez, Ian Morrissey, Marco Rinaldo Oggioni and Ana Teresa Freitas, "The Use of Machine Learning Methodologies to Analyse Antibiotic and Biocide Susceptibility in Staphylococcus aureus" *journals.plos.org*, Feb. 19, 2013. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0055582#s2. [Accessed: Nov 10, 2017]

[13] Rakesh Agrawal and Ramakrishnan Srikant, "Fast algorithms for mining association rules." Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

[14] Annalyn Ng, "Association Rules and the Apriori Algorithm: A Tutorial," *kdnuggets.com*, April 2016. [Online]. Available: https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html. [Accessed: Nov 20, 2017

Principles of IoT, Robotics and Automation Systems

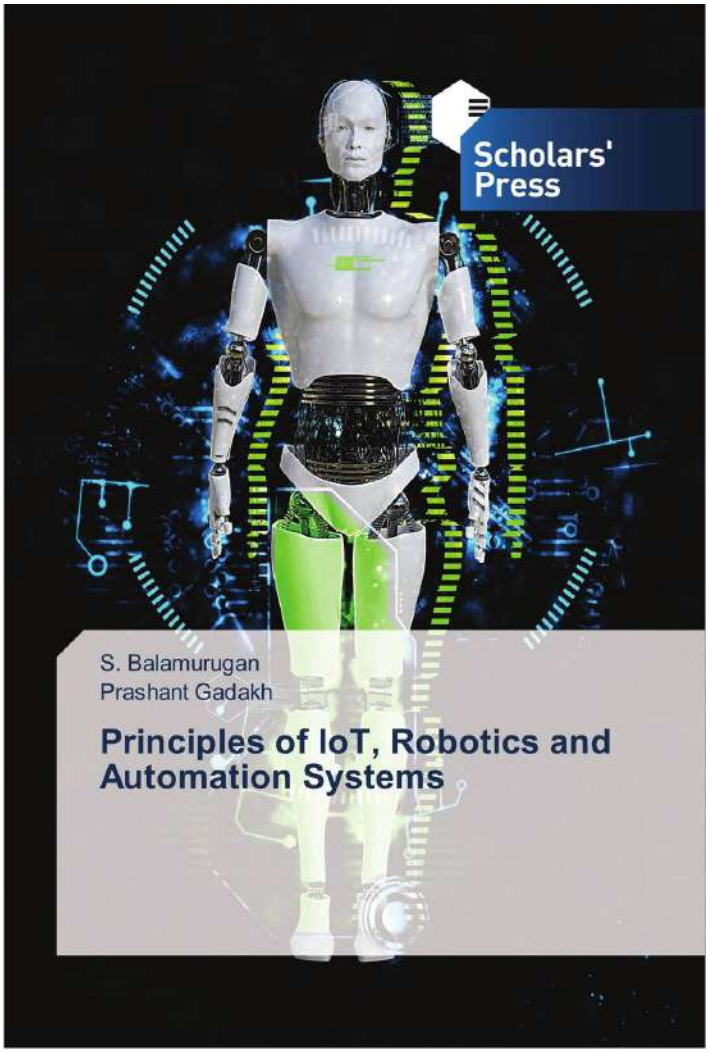
This book would serve as an ideal guide for B.E., B.Tech., B.S., B.Sc, B.C.A., undergraduate students of Computer Science and Engineering, Information Technology, Electronics and Communication Engineering who wish to take up projects on IoT, Robotics and Automation Systems. Students pursuing postgraduate course in Science and Engineering, M.E., M.Tech., M.S., M.Sc., M.C.A. students will find this book useful for their projects. Research Scholars working in the area of IoT, Robotics and Automation Systems, will find this book as a handy reference guide for their M.Phil., Ph.D. D.Sc., and other post-doctoral research works. Software Engineers and Hardware Analysts, involved in IT and ITES sector specifically on IoT, Robotics and Automation Systems, would find this book as a useful resource. As a word of conclusion, we believe that the reader will find this book as a really helpful guide and a valuable source of information about IoT, Robotics and Automation Systems.

Dr.S.Balamurugan is the Director - Research and Development, Mindnotix Technologies, India. He has to his credit 175 papers, 22 Books and 34 International Awards for Excellence in Research. Prof.PrashantGadakh is working as Assistant Professor at International Institute of Information Technology,Hinjawadi,Pune, India.He has to his credit 30 papers.

978-620-2-31580-7

IoT, Robotics and Automation Systems

Balamurugan, Gadakh



Principles of IoT, Robotics and Automation Systems

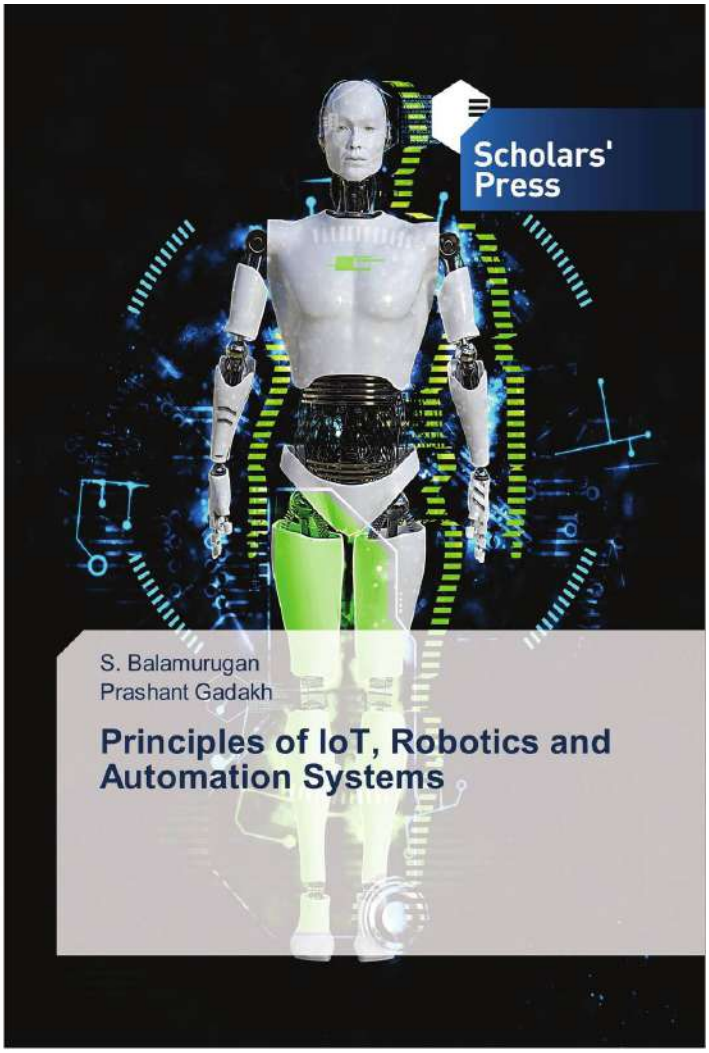
This book would serve as an ideal guide for B.E., B.Tech., B.S., B.Sc, B.C.A., undergraduate students of Computer Science and Engineering, Information Technology, Electronics and Communication Engineering who wish to take up projects on IoT, Robotics and Automation Systems. Students pursuing postgraduate course in Science and Engineering, M.E., M.Tech., M.S., M.Sc., M.C.A. students will find this book useful for their projects. Research Scholars working in the area of IoT, Robotics and Automation Systems, will find this book as a handy reference guide for their M.Phil., Ph.D. D.Sc., and other post-doctoral research works. Software Engineers and Hardware Analysts, involved in IT and ITES sector specifically on IoT, Robotics and Automation Systems, would find this book as a useful resource. As a word of conclusion, we believe that the reader will find this book as a really helpful guide and a valuable source of information about IoT, Robotics and Automation Systems.

Dr.S.Balamurugan is the Director - Research and Development, Mindnotix Technologies, India. He has to his credit 175 papers, 22 Books and 34 International Awards for Excellence in Research. Prof.PrashantGadakh is working as Assistant Professor at International Institute of Information Technology,Hinjawadi,Pune, India.He has to his credit 30 papers.

978-620-2-31580-7

IoT, Robotics and Automation Systems

Balamurugan, Gadakh



International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8 Issue-8, June, 2019

Supervised Machine Learning Supported Time Series Prediction and Analysis of IoT Enabled Physical Location Monitoring

Ajitkumar S. Shitole, Manoj H. Devare

Abstract: ***Internet of Things (IoT) is one of the evolving technologies in the recent days to exchange the information from one device to another using any type of network, at anytime, and at anywhere. With the introduction of IoT and Machine Learning (ML) to monitor physical location in real time fashion is necessary to identify abnormal conditions in the surroundings. The proposed system depicts that different sensors in addition to camera are used to monitor and identify abnormal environment conditions of the same and send alert message to the user to take corrective action to avoid any future loss in the environment. Real time sensor data which is aligned with multimedia data is stored onto local system and ThingsSpeak server as well as it is pushed onto Go Daddy cloud whenever camera detects person to perform systematic and statistical analysis using different supervised machine learning algorithms. This paper presents time series prediction of different sensor values such as temperature, humidity, gas, light dependent resistor, and person prediction using timestamp (day and time) to understand the physical location well in advance to take appropriate decision. Experimental results show that decision tree is the best predictive model to predict person when timestamp is given in the form of date and time. Study also reveals that Decision Tree Regression (DTR) and Random Forest Regression (RFR) give good results with approximately same minimum Root Mean Squared Error (RMSE) to predict different sensor values.***

Index Terms: ***Physical Location Monitoring, Time Series Prediction, RMSE, Supervised Machine Learning***

I. INTRODUCTION

The Analysis and structure of IoT is the way towards providing data and giving a forecast utilizing the sensor. IoT chips away at smart items that interface with the sensor and accumulate data and speak with neighboring individuals utilizing versatile, remote and sensor advances. The valuable data from sensor information and process on this data utilizing machine learning are separated. Physical area for gathering the data from sensor and work on this data is required to extract the knowledge.

Proposed framework utilizes the sensor to ask for feeling of the earth. The significance of installed is the association of two distinct things and the coordinated framework in which the product is incorporated into the equipment. The incorporated framework that has the benefit of low power utilization enhances framework execution and does it effortlessly. IoT alongside Machine Learning (ML) is utilized to caution the circumstance when the individual is in genuine hazard. ML is utilized to do systematic analysis of the dataset. It utilizes Raspberry-Pi as the fundamental base of our task for preparing information. The Graphical User Interface (GUI) is created for work areas or workstations and applications for mobiles to display different sensor values with the status of the physical location to indicate whether the location is in normal condition or not. Framework likewise gives the predication of the data. It utilizes the machine learning calculation for giving precision of the framework and arranges the data originate from sensor and gives the predication of this data. The construction utilizes four various machine learning predictive models with Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Random Forest (RF) for person prediction using time series analysis. The proposed system also performs time series prediction of different sensor values using DTR.

Revised Manuscript Received on June 05, 2019
Ajitkumar S. Shitole, Research Scholar, Amity University Mumbai, India
Manoj H. Devare, HoI, AIIT, Amity University Mumbai, India

II. RELATED WORK

Aras Can Onal et al widened IoT skeleton that consolidates the data recuperation, getting ready, and knowledge layers is given a use container on atmosphere data gathering examination. The learning model made uses batching unsubstantiated learning procedure in the learning time of the skeleton with an explicit ultimate objective to best use the related immense data for this issue. The US climate data got from 8000 assorted atmosphere stations around North America is received through log records. Wind Speed 3 Clusters, Sensor Fault and submitted to learning stage for the learning system. In this explicit examination, air temperature, wind-speed, relative wetness, detectable quality, and weight data are used as a piece of the data examination. Traditional k-infers gathering count is associated and the results are presented. As interesting miracles, framework watched that the data packing matches the geological game plan of the stations. In a manner of speaking, a segment of the fundamental land locale inside the North American terrain (and the territory USA) shape obvious atmosphere gatherings and easily isolated from one another. Likewise, possible sensor inadequacies and quirks are produced with using gathering technique. This use case empowered to show an instance of how such an IoT Big Data framework can be used for such utilization [1].

Peng Sun et al inspects the endorsement of accelerating sensors in a support assistant examination using both Naive Bayesian Classifier (NBC) and Tree Augmented Naïve Bayesian Classifier (TAN) figuring. Through a bracket helper preliminary the counts are affirmed. The examination comes about confirm that the future techniques in this paper in perspective of NBC and TAN are effective. In addition, the results similarly instigate that

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Principles of Network Security and Information Security

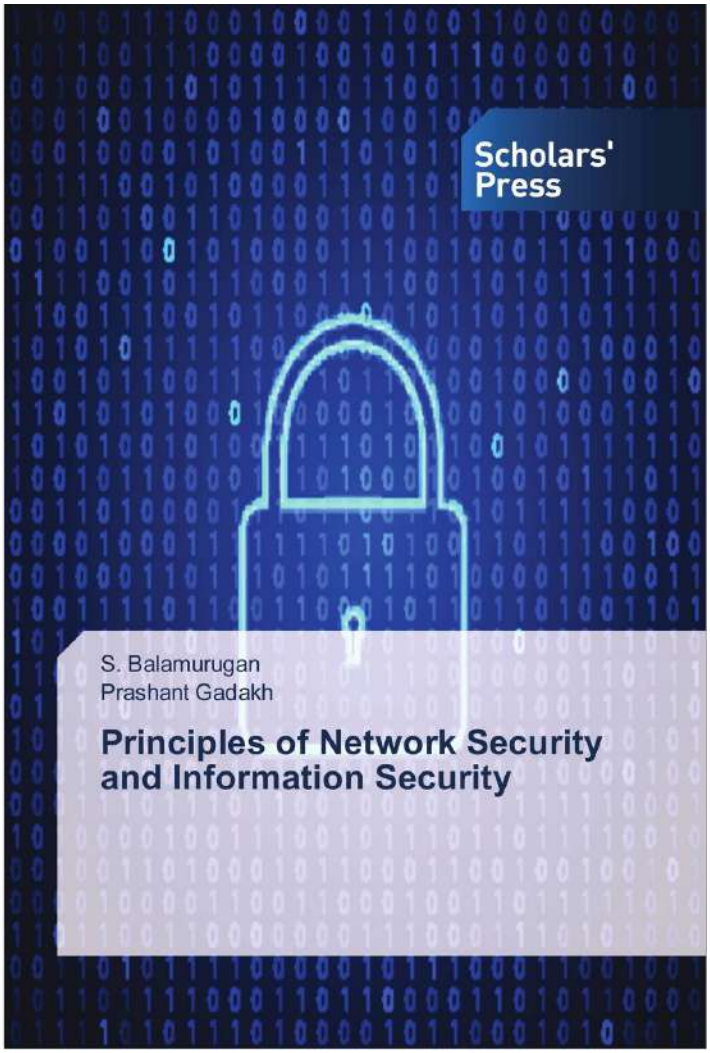
This book would serve as an ideal guide for B.E. B.Tech., B.S., B.Sc, B.C.A., undergraduate students of Computer Science and Engineering, Information Technology, Electronics and Communication Engineering who wish to take up projects on Network Security. Students pursuing postgraduate course in Science and Engineering, M.E., M.Tech., M.S., M.Sc., M.C.A. students will find this book useful for their projects. Research Scholars working in the area of Network Security, will find this book as a handy reference guide for their M.Phil., Ph.D. D.Sc., and other post-doctoral research works. Software Engineers and Hardware Analysts, involved in IT and ITES sector specifically on Network Security, would find this book as a useful resource. As a word of conclusion, we believe that the reader will find this book as a really helpful guide and a valuable source of information about Network Security.

Dr.S.Balamurugan is the Director - Research and Development, Mindnotix Technologies, India. He has to his credit 175 papers, 22 Books and 34 International Awards for Excellence in Research. Prof.PrashantGadakh is working as Assistant Professor at International Institute of Information Technology,Hinjawadi,Pune, India.He has to his credit 30 papers.

978-3-639-66544-4

Network and Information Security

Balamurugan, Gadakh



IEEE Bombay Section

Certificate of Presentation

This certificate is proudly presented to

Mr/Ms. Deptii Chaudhari

for presenting paper entitled

A review of Learning Strategies applied to the Arcade Learning Environment

In 5th International Conference for Convergence in Technology 2019

Pune,India from 29th - 31st March,2019

International Conference for Convergence in Technology
PUNE
2019

Pallavi

Dr. Pallavi
Conference Chair

This research study is focused on investigating the problem of recognition of human identity from its walking pattern in presence of view angle and clothing condition as covariate. The problem of gait recognition is challenging due to fact that gait has spatio-temporal phenomena with very high dimensional tensorial data distribution, large amount of redundancy, complex pattern distribution and very large variability of appearance within the same class of subject. The extraction of discriminative features in the presence of covariates for robust human gait recognition is a challenging task. The study presents processing aspects of gait recognition system through systematic framework of pre-processing, gait representation, feature extraction, dimensionality reduction and classification for human identity recognition. It has contributed to understanding, interpretation and development of effective gait representation and multilinear subspace learning algorithm. The key contribution, understanding and interpretation are summarized here and direction for future work is provided.

Dr. Risil Rameshbhai Chhatrala - Degree of Doctor of Philosophy in the Faculty of Engineering. Research Center JSPM's, Rajarshi Shahu College of Engineering, Pune, Department of Electronics and Telecommunication Engineering.

978-613-9-93634-2

Human Identity Recognition

Risil Rameshbhai Chhatrala
Dattatray Jadhav
Shailaja Patil

Gait Based Human Identity Recognition

LAP LAMBERT Academic Publishing

Chhatrala, Jadhav, Patil

Spatial and Temporal Characteristics of Ionospheric Total Electron Content over Indian Equatorial and Low-Latitude GNSS Stations

*G Sivavaraprasad, **Yuichi Otsuka, ***Nitin Kumar Tripathi, ***V Rajesh Chowdhary, and *D Venkata Ratnam, Senior Member IEEE, *Mohammed Afzal Khan
*Department of ECE, KLEF, KL University, Vaddeswaram, Guntur Dt, 522502, Andhra Pradesh, India
** Institute for Space-Earth Environmental Research, Nagoya University, Nagoya, Japan
***Asian Institute of Technology, School of Engineering and Technology, RS&GIS Field of Study, Thailand
Email: dvratnam@kluniversity.in

***Abstract*— The study and understanding of the intermittent characteristics of equatorial and low latitude ionosphere is crucial for modelling and forecasting the ionosphere and space weather conditions. The performance of space-based navigation systems such as Global Positioning System (GPS) is affected by the sporadic temporal and spatial variations of ionospheric Total Electron Content (TEC). The variability of ionospheric electron density over Indian low latitude sector is difficult to model due to Equatorial Ionization Anomaly (EIA). In this paper, Multi-fractal aspects of the GPS measured TEC is investigated during both high and low solar activity periods of 24th solar cycle. The vertical TEC (VTEC) data sets are obtained from two Indian low latitude stations namely, Bangalore (Geographical Latitude: 13.02^0 N, Geographical Longitude: 77.57^o E), and Lucknow (Geographical Latitude: 26.83^0 N, Geographical Longitude: 80.92^o E) for two year long period 2013 and 2015.The experimental results shows that the respective geographic sites have important scaling differences as well as similarities when their Multi-fractal signatures for VTEC are compared. These differences and similarities are interpreted in terms of the EIA conditions, where this phenomenon is an important source of intermittence due to the presence of the VTEC peaks at ±30^0 geomagnetic latitudes. During the high solar activity period, the intermittence characteristics of VTEC over EIA region (Lucknow) are relatively more complex than equatorial (Bengaluru) station, whereas during low solar activity period the scenario is reciprocal.**

***Index Terms*— Global Navigation Satellite System (GNSS), Global Positioning System (GPS), Vertical Total Electron Content (VTEC), Equatorial Ionization Anomaly (EIA), Multi-fractal, Detrending Fluctuation Analysis (DFA), Solar Activity.**

I. INTRODUCTION

Satellite dependent navigation systems experience range errors due to non-linear characteristics of ionosphere. The ionospheric dynamic medium is one of main effecting error source of Global Positioning System (GPS) performance for precise range and position determining capability [1]. Phase advance and group delay are commonly termed as ionospheric phenomenon, moreover amplitude fading and phase scintillation phenomenon cause the loss of carrier lock and interrupt GPS receiver operations [2]. Total Electron Content (TEC) fluctuations are large with respective to the solar radiations, geo magnetic activities, magnetic storms [6]. The activity of Vertical Total Electron Content (VTEC) is dependent on geo magnetic storms and solar cycles or solar activity [2], [3]. According to the studies carried out by [4], the intermittent characteristics like ionospheric TEC are responsible for improper Global Navigation Satellite System (GNSS) dual-frequency receiver's measurements. To reduce GPS data transmission errors, analysis of ionospheric phenomenon is necessary [4].

Spatial temporal characteristics of TEC are highly variant at equatorial and low latitude regions than at mid latitude regions [5]. A small change in TEC data is enough to change the dimensional view of ionosphere layer [6]. This makes fractal dimensions of TEC data crucial in understanding ionospheric space weather for robust design and operations of GNSS receiver. [6]. Study of small changes in TEC is termed as fractal analysis. Holder exponent and local Hurst exponents are used to define the fractal structure to be whether multi or mono fractal signal [7]. The TEC data has the characteristics of Multi-fractality which shows extreme magnitudes and long memory [8].

To study these highly variant VTEC (nonlinear characteristics of a signal) many statistical methods like Fourier transform, wavelet analysis [9], wavelet Transform modulus maxima method (WTMM) [4], Multi-fractal detrended fluctuation analysis (MFDFA) [1], [10], are implemented. For studying the intermittent characteristics of VTEC Fourier transform method fails because of its restriction to handle multiple high range frequencies [11]. Wavelet transform method localises the random fluctuations of non-linear properties of non-stationary data at different scales and positions [7], [9], [12]. Similarly MFDFA method works on the principle of finding the q-order moments in the signal [7], [9]. However, these methods give accurate results when the fractal property of the signal is known [7], [9]. WTWM method uses the calculation of the average for the total signal (VTEC) at a time calculating but in MFDFA

The above work has been carried out under the joint research collaboration of K L University, Vaddeswaram, Guntur, A.P, India and Institute for Space-Earth Environmental Research, Nagoya University, Nagoya, Japan. The authors would like to express their thanks to the Department of Science and Technology, New Delhi, India SR/FST/ESI-130/2013(C) FIST program.

method, the average for specific time intervals of the signal is considered to get the more information about the multi-fractral nature in the temporal and spatial variations in the ionospheric TEC. Thus, MFDFA becomes more subtle than WTWM method. Moreover, when the fractal property of the VTEC is unknown the complexity in WTMM increases more than MFDA [7], [9]. The computation of Root-Mean-Square (RMS) for given VTEC time series provides the average variations in it. In the case of mono-fractal detrended fluctuation analysis (DFA), the power law relation between the multiple segment sample sizes and overall RMS is known as Hurst exponent and in the case of MFDFA, *q*-order Hurst exponent can distinguish the RMS for segments with small variations and large fluctuations. For the negative values of q-order Hurst exponent, the WTMM method becomes more programmable and does not give the accurate results when compared to the MFDFA method which will be of easy functioning [12]. This makes MFDA more suitable and reliable method to determine the Multi-fractality nature of the VTEC whether it is a short time series or long time series [7], [9].

In this paper the seasonal, annual temporal and spatial variations of ionosphere TEC has been investigated during high and low solar activity period of the 24th solar cycle (2013 & 2015). Seasonal and annual variations of ionosphere TEC is measured over two low latitude stations. The main objective of this paper is to understand the seasonal dependence of the Multi-fractal behaviour compared to the annual variations.

II. Data Processing

Fig. 1 depicts two low latitude stations in India, Bengaluru station (Southern Indian region nearer to geomagnetic equator), Lucknow station (Northern Indian region nearer to Equatorial Ionization Anomaly (EIA)). The GPS satellite data recorded at these IGS stations are processed to determine the VTEC [1].Using the calculated TEC data [5], studied the effect of TEC on satellite communication [1].The GPS-TEC data obtained at the stations are processed to find out the slant TEC, baseline parameters and then stored in format of RINEX_GPS_TECH [3], [13].

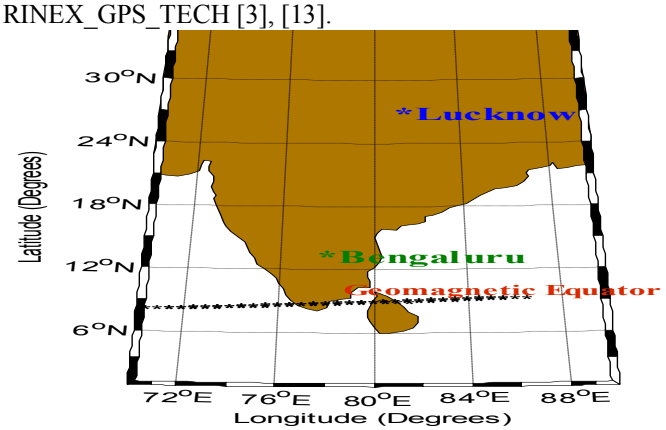


Fig. 1. India map plotting the Lucknow and Bengaluru Stations

The RINEX observation files data of 2013, year of solar maxima [3] and 2015, year of solar minima are processed further for investigating the seasonal and annual variations of Multi-fractality, at low latitude Stations.

The GPS-TEC data is obtained from Scripps Orbit and Permanent Array Centre (SOPAC) for two low latitude stations Bengaluru (Geographical latitude: 13.02^0 N, Geographical longitude: 77.57^0 E) and Lucknow (Geographical latitude: 26.83^0 N, Geographical longitude: 80.92^0 E) are collected.

III. Implementation of MFDFA Algorithm for Fractal Analysis of GPS-TEC Data

Mathematical equations are applied on the recorded TEC data for over a year, 2013, and 2015. Steps for performing Multi-fractality analysis are:

Step 1: Generate VTEC data using the formulae below.

$$x(n)=\sum_{i=1}^{n}(y(i)-\bar{y});n=1,2,3,.....N \tag{1}$$

Where, y= VTEC data recorded for a year period.

Step 2: The series x (n) is divided into N_l equal non overlapping segments of length l each. But it is not always possible for the N_l to be an integer hence for accuracy the grouping process is performed from both forward and backward directions leading to form $2N_l$ number of segments in total [10, 14].

Step 3: The average fluctuation of the TEC is calculated by mean of the both forward and backward group's data.

$$P(l,m)=\sqrt{\frac{1}{l}\sum_{i=r+1}^{lm}\left[x(i)-x_i(i)\right]^2} \tag{2}$$

where P (l, m) is the average fluctuation in both forward and backward directions.

r = (m-1)* l .

m = the window number.

For forward operation m=1, 2, 3....., and

For backward operation m = -1, -2, ... , 3, 2, 1.

Step 4: Using the variance of both backward and forward operations, qth order fluctuation function is given by

$$P_q(l)=\left\{\frac{1}{2N_l}\sum_{m=1}^{2Nl}\left[P^2(l,m)\right]^{\frac{q}{2}}\right\}^{\frac{1}{q}} \tag{3}$$

where, q (index variable) can be of any value other than 0. Negative q values are used to capture the small fluctuations where as positive q value is used to capture the large fluctuations of VTEC data. Eq. 3 is repetitively calculated for different window lengths (k) and q values. P (l, m) follows the power law relation

$$P_q(l)\alpha l^{H(q)} \tag{4}$$

where H_q is the slope for the graph between log (P_q (l)) vs log (k). H_q is termed as local Hurst exponent and varies with respective to q. Generally q-order is 2 for mono-fractality and above 2 for Multi-fractality [10]. Similarly H_q in the range of

0.5 refers to mono-fractality and Hq in the range of 0.7-0.8 refers to Multi-fractality [10].

For q=0 the Hurst exponent can be given as

$$P_0(l) = \exp[\frac{1}{4N_l}\sum_{m=1}^{2N_l}\ln\left\{P^2(l,m)\right\}] \approx l^{h(0)} \quad (5)$$

Now for the measurement of Multi-fractality of the VTEC data Multi-fractal spectrum (Dq) is calculated through MFDFA method in the following step 5.

Step 5 (Calculation of Dq):

Measurement of very small variations present in the TEC data is termed as Detrending analysis. This can be done through the analysis of Multi-fractality spectrum.

$$t_q = q\left(H_q\right) - 1 \quad (6)$$

Holder exponent (or singularity exponent) is calculated by

$$hq = \frac{dt_q}{dq} \quad (7)$$

$$Dq = \left(q \times hq\right) - t_q \quad (8)$$

Finally, a plot is drawn between singularity exponent (hq) and Singularity Dimension (Dq) which is termed as Multi-fractality spectrum.

IV. Results and Discusion

Fig. 2 depicts the TEC variations in the year 2013 and 2015 for the stations Bengaluru and Lucknow. Amplitude of TEC (Eq. 1) is plotted against the number of samples i.e., for a 365 day period. It's clear that in the year 2015 equinox seasons experiences large fluctuation in TEC data this phenomenon is explained by the analysis carried out by the [15], [16] which states that higher variations in TEC data is observed in equinox's because of vertical E×B drifts along with EUV ionization in both high and lower solar activity periods. Moreover, during the 2013 and 2015 year, large fluctuation in TEC data are observed during March month, supporting the geomagnetic disturbed conditions during, 17 March and 29 June 2013 due to geo-magnetic storms [6] which plays major role in variation in electron density in ionospheric F layer. Moreover, it has been observed that the random VTEC signatures over both the GPS stations due to the largest geomagnetic storm in 24th solar cycle occurred on 17 March 2015.The dependency of VTEC variations on the solar activity and geomagnetic storms has learnt from the [8] which here used for better understanding of the variation in TEC through multi-fractral analysis.

In Fig. 3 Annual Multi-fractality nature of the VTEC data is examined by the plot between singularity exponent (hq) and singularity spectra (Dq) (Eq. 6 & 8). q-order in MFDFA varies according to the VTEC data. The width of the Multi-fractality spectra defines the degree of Multi-fractality. This proves the statements made by [6] i.e., earth atmosphere got effected by the solar storms and geo magnetic fields and it's evident through the width that is measured in the Multi-fractality spectrum for the year 2013 in both stations. Bengaluru and Lucknow stations in 2013 have higher widths in Multi-fractal spectra when compared to the same stations in the year 2015 (solar minima). i.e., 2015 (year of solar minimum) gets least effected with the solar radiations proving through the width obtained from the Multi-fractal spectrum at both the stations.

Bengaluru stations also experiences highest peak in TEC recorded assuming that there have been large and sudden changes in the electron density of the ionosphere. Using Singularity exponent (hq), and holder exponent the signal can be predicted whether it is mono-fractal or multi fractal. Hurst exponent at a range of 0.7-0.8 (Eq. 6) can be termed as the

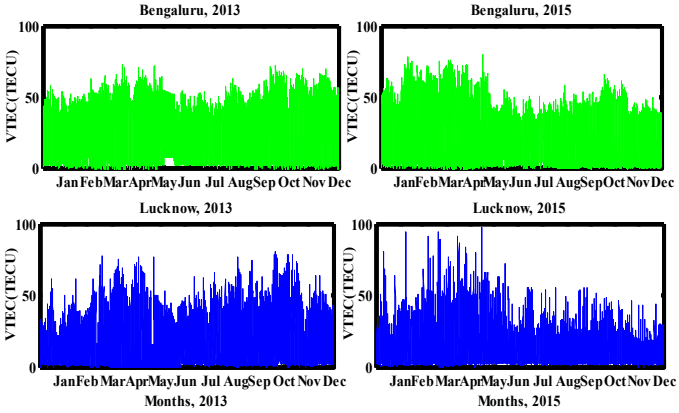


Fig. 2. Annual plots TEC data at Bengaluru and Lucknow Stations for the years 2013 and 2015.

Multi-fractal and when it is at rate of 0.5 it is called as mono-fractal [10]. This is evident from the Fig. 3 where Hurst exponent (H_q) & Singularity Exponent (hq) values are observed monotonically decreasing order with increase in scaling factor (Eq. 4 & 6).

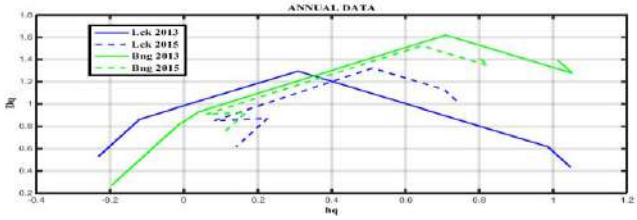


Fig. 3. Annual Multi-fractality spectrum at Lucknow and Bengaluru stations for the years 2013 and 2015.

Fig. 4 depicts the seasonal variations of TEC at Bengaluru and Lucknow station in 2013 and 2015 years. The amplitude of VTEC is shown against the Time in UT i.e., for 24 hours. The presence of winter anomaly is observed for the year 2013 [17]. Here the absence of winter anomaly is observed for the year 2013 solar maximum that is the correlation between equinoxes and summer solstice is minimum but when comes to December solstice it shows different behavior from others. Winter anomaly occurs due to O/N_2 ratio changes (i.e., winter season has low solar activity but due to enormous increase in O/N_2 ratio in the ionosphere layer TEC gets increased compared to the June equinox) in the ionosphere layer [3]. Winter anomaly effect is absent at Bengaluru station and Lucknow, EIA station in the year of solar maximum, 2013 and its effect is negligible over both the stations during year of solar minimum,2015 [3] which is clearly depicted in Figure 4. Similar conclusions are drawn for winter anomaly got missing for solar maximum year, 2013 [15].

Multi-fractal spectrum for seasonal variations is carried out for the TEC variations during 2013 and 2015. It is observed that Bengaluru station (equatorial region) experiences more ionization rate when compared to the Lucknow station due to EIA and ExB drifts at geomagnetic equator. Seasonal plots gives use the detrending analysis of the TEC data. March, September equinoxes and December solstice are left skewed and right truncated [10]. Insensitive to the local fluctuations having lower amplitudes is termed as right truncation.

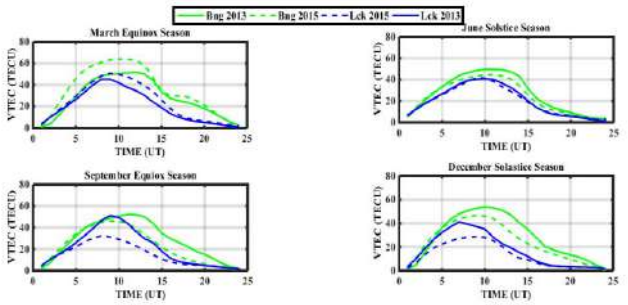


Fig. 4. Seasonal plots TEC data at Bengaluru and Lucknow Stations for the years 2013 and 2015.

In contrast June solstice has unique features when compared to other seasons it is right skewed and also experience left truncation. From the graphs plotted seasonally and annually singularity exponent (hq) values are decreasing as there is increase in the scaling size.

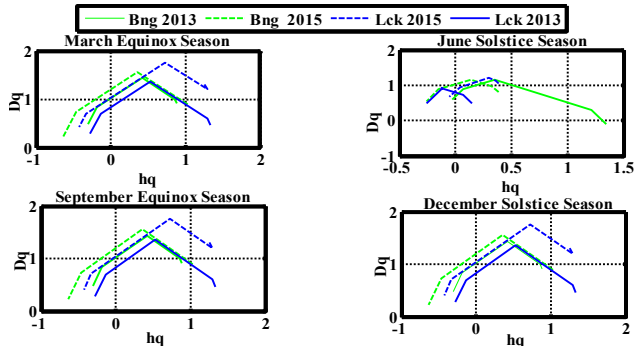


Fig. 5. Seasonal Multi-fractality spectrum at Lucknow and Bengaluru stations for the years 2013 and 2015.

The seasonal Multi-fractality characteristics are show in Fig. 5. It is observed that the temporal and spatial variations ionosphereic TEC over equatorial and north Indian station near EIA have been illustrated using MFDFA technique. The non linear multi-scale features and intermittences are high over both the stations during all the season for solar maximum period, 2013(Fig. 5). However, the degree of Multi-fractality is found more during 2015 year over Lucknow station. Thus, the MFDFA is a useful fractal analysis tool for the investigation of intermittencies in GPS observation.

V. Conclusion

This paper mainly concentrates on the use of efficient method, MFDFA which overcomes the difficulty in previous methods like WTWM, Fourier transform. Multi-fractral spectrums have been analysed to study the effects of solar activity (solar maximum, solar minimum, geomagnetic storms) in varying the TEC data at the stations Bangalore and Lucknow for the years 2013 and 2015. The effect of winter anomaly is also observed in the seasonal TEC variation at both the stations. It can be concluded that TEC data is largely sensible to the space weather environmental conditions and should be analysed fractally in order to eliminate the errors that occurs in GPS and radio communication systems.

References

[1] Thomas, M., J. Norton, A. Jones, A. Hopper, N. Ward, P. Cannon, N. Ackroyd, P. Cruddace, and M. Unwin. "Global navigation space systems: reliance and vulnerabilities." *The Royal Academy of Engineering, London* 2011.

[2] J. A. Klobuchar, "Ionospheric effects on GPS," *Global Positioning System: Theory and applications.,* vol. 1, pp. 485-515, 1996.

[3] S. Hamzah and M. J. Homam, "The correlation between total electron content variations and solar activity," 2015.

[4] M. Bolzan, R. Rosa, and Y. Sahai, "Multi-fractal analysis of low-latitude geomagnetic fluctuations," in *Annales geophysicae: atmospheres, hydrospheres and space sciences*, 2009, p. 569.

[5] P. R. Rao, S. G. Krishna, K. Niranjan, and D. Prasad, "Temporal and spatial variations in TEC using simultaneous measurements from the Indian GPS network of receivers during the low solar activity period of 2004? 2005," in *Annales Geophysicae*, 2006, pp. 3279-3292.

[6] B. Zhao, M. Wang, T. Yu, W. Wan, J. Lei, L. Liu, *et al.*, "Is an unusual large enhancement of ionospheric electron density linked with the 2008 great Wenchuan earthquake?," *Journal of Geophysical Research: Space Physics,* vol. 113, 2008.

[7] R. Galaska, D. Makowiec, A. Dudkowska, A. Koprowski, K. Chlebus, J. Wdowczyk-Szulc, *et al.*, "Comparison of wavelet transform modulus maxima and Multi-fractal detrended fluctuation analysis of heart rate in patients with systolic dysfunction of left ventricle," *Annals of Noninvasive Electrocardiology,* vol. 13, pp. 155-164, 2008.

[8] S. Gopinath and P. Prince, "Multi-fractal characteristics of magnetospheric dynamics and their relationship with sunspot cycle," *Advances in Space Research,* 2017.

[9] P. Oświęcimka, J. Kwapień, and S. Drożdż, "Wavelet versus detrended fluctuation analysis of Multi-fractal structures," *Physical Review E,* vol. 74, p. 016103, 2006.

[10] E. Ihlen, "Introduction to Multi-fractal detrended fluctuation analysis in Matlab," *Fractal Anal,* vol. 3, p. 97, 2012.

[11] D. Pancheva and P. Mukhtarov, "Wavelet analysis on transient behaviour of tidal amplitude fluctuations observed by meteor radar in the lower thermosphere above Bulgaria," in *Annales Geophysicae*, 2000, pp. 316-331.

[12] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley, "Multi-fractal detrended fluctuation analysis of nonstationary time series," *Physica A: Statistical Mechanics and its Applications,* vol. 316, pp. 87-114, 2002.

[13] E. Chandrasekhar, S. S. Prabhudesai, G. K. Seemala, and N. Shenvi, "Multi-fractal detrended fluctuation analysis of ionospheric total electron content data during solar minimum and maximum," *Journal of Atmospheric and Solar-Terrestrial Physics,* vol. 149, pp. 31-39, 2016.

[14] C. Barman, H. Chaudhuri, D. Ghose, A. Deb, and B. Sinha, "Multi-fractal Detrended Fluctuation Analysis of Seismic Induced Radon-222 Time Series."

[15] J.R.K. Kumar Dabbakuti and D. V. Ratnam, "Performance evaluation of Linear Time-Series Ionospheric Total Electron Content Model over Low latitude Indian GPS Stations," *Advances in Space Research,* 2017.

[16] D. V. Ratnam, G. Sivavaraprasad, and N. L. Devi, "Analysis of ionosphere variability over low-latitude GNSS stations during 24th solar maximum period," *Advances in Space Research,* 2016.

[17] S. Raghunath and D. V. Ratnam, "Detection of low-latitude ionospheric irregularities from GNSS observations," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 8, pp. 5171-5176, 2015.

Gait Recognition Using Normal Distance Map and Sparse Multilinear Laplacian Discriminant Analysis

Risil Chhatrala, Shailaja Patil and Dattatray V. Jadhav

Abstract In visual surveillance applications, gait is the preferred candidate for recognition of the identity of the subject under consideration. Gait is a behavioral biometric that has a large amount of redundancy, complex pattern distribution and very large variability, when multiple covariate exist. This demands robust representation and computationally efficient statistical processing approaches for improved performance. In this paper, a robust representation approach called Normal Distance Map and multilinear statistical discriminant analysis called Sparse Multilinear Discriminant Analysis is applied for improving robustness against covariate variation and increase recognition accuracy. Normal Distance Map captures geometry and shape of silhouettes so as to make representation robust and Sparse Multilinear Discriminant Analysis obtains projection matrices to preserve discrimination.

1 Introduction

Automated identification and recognition of the individual person in a surveillance environment with natural setting, is so hard problem that not a single gait recognition system has been reported to be working in challenging real world conditions. A large portion of the literature is dedicated to important aspects of gait recognition, so as to make it a realizable solution. The aspects like segmentation, pre-processing, gait representation schemes and pattern recognition algorithms are widely studied to understand diverse aspect requirement for gait processing.

The main contribution is as follows.

1. A new robust gait representation scheme that make use of boundary curvature information over a complete gait cycle called as Normal Distance Map is proposed.

R. Chhatrala (✉) · S. Patil
Rajarshi Sahu College of Engineering,
Savitribai Phule Pune University, Pune, Maharashtra, India
e-mail: therisil@gmail.com

D. V. Jadhav
Directorate of Technical Education, Mumbai, Maharashtra, India

D. Pandian et al. (eds.), *Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering 2018 (ISMAC-CVB)*, Lecture Notes in Computational Vision and Biomechanics 30,
https://doi.org/10.1007/978-3-030-00665-5_14

2. The representation preserves the natural tensorial structure along with spatio-temporal and structural information of gait.
3. The Extracted features provide a new feature space that addresses covariates and is found to be robust for gait recognition.
4. Sparse Multilinear Laplacian Discriminant Analysis for tensor objects is used to improve discrimination capability and increase recognition rate.

This paper is structured as follows. After reviewing the literature work in Sect. 2, the gait representation scheme based on boundary curvature information over a complete gait cycle called as Normal Distance Map is presented in Sect. 3. Section 4 presents feature extraction and pattern recognition using Sparse Multilinear Laplacian Discriminant Analysis followed by experimentation in Sect. 5. Sections 6 and 7 gives Discussion and Conclusion respectively.

2 Review of Literature

Comprehensive review of the published techniques and strategies for gait as a biometric can be found in the work of Makihara et al. [1], Sivarathinabala et al. [2], Zhang et al. [3], Boulgouris et al. [4] and Wang et al. [5]. The widely researched areas of the gait recognition system is gait representation, feature dimensionality reduction and classification.

The pioneer approaches for gait descriptor are Gait Energy Image (GEI) [6], Shifted Energy Image (SEI) [7], Gait Entropy Image (GEnI) [8], Gait flow image [9], Frequency-Domain Features [10], Depth Gradient Histogram Energy Image (DGHEI) [11] and Histogram of boundary normal vector (HoNV) using local Gauss Maps [12].

3 Gait Representation

In most recent work; it is observed that, the spatio-temporal variation exhibited by gait is mainly represented by shapes and kinematic variation averaged over gait period. The work of Tang et al. [13] and El-Alfy et al. [12, 14] inspired us to use local curvatures of a silhouette contour obtained from the geometry of the silhouettes and distance transform, together to capture boundary and area information. This feature descriptor is called Normal Distance Map (NDM). In all further discussion, Normal Distance Map (NDM) as suggested by El-Alfy et al. [14] is employed as gait feature representation. Following subsection gives a brief review of work from El-Alfy et al. [12, 14].

3.1 Histogram of Boundary Normal Vector

The histograms of boundary normal vectors were introduced by El-Alfy et al. [12]. The method focuses on curvature of contours extracted from the geometry of the silhouettes. Local Gauss maps are used to link Unit vectors normal to a surface to its curvature [15, 16]. The key advantage of doing this is to come to conclusion that, "If the normal vectors magnitude is fixed and scaled to unity, then contours extracted from 'parallel' geometric curvatures need to have same histograms." [12]. This causes the descriptor to be robust to covariate that affect gait itself.

3.2 Gauss Maps

A Gauss map g is a mapping function that maps each point p from a surface $M \in R^3$ to the unit sphere S^2, with the unit vector n_p normal to M at p.

$$\begin{aligned} g : M &\rightarrow S^2 \\ p &\mapsto n_p \end{aligned} \tag{1}$$

The normal vectors to flat surface M has no variation between them and is always parallel to each other. On the other hand, variation is seen for an "overly" curved surface. Hence, the mapping g is used to model the curvature of the surface. The Gaussian curvature k is defined as:

$$k = \lim_{\Omega \to 0} \frac{\text{area of } g\,(\Omega)}{\text{area of } \Omega} \tag{2}$$

The total curvature of Ω is defined by:

$$\text{total curvature of } \Omega = \int_{\Omega} k dA \tag{3}$$

where dA is a surface element on M.

Since, Gaussian curvature of a surface $M \in R^3$ is invariant under local isometries, globally computed surface cannot always be used to discriminate. Hence, Gauss maps are defined locally, in the form of patches or cells and for each small surface or cell, local curvatures are computed.

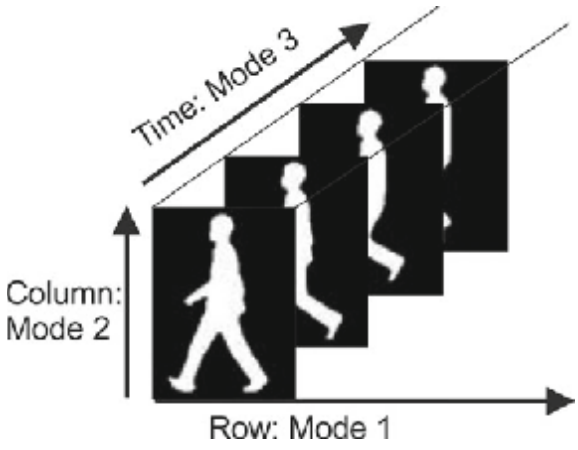


Fig. 1 Third order tensorial binary silhouette video sequence (GSV)

3.3 Normal Distance Map as Tensor

In order to compute NDM, the approach of El-Alfy et al. [14] is used. In this paper, the final feature descriptor is represented in the form of matrix for each frame. The gait cycle consists of multiple frames exhibiting inherent variation in pixel distribution due to kinematic motion, it is represented as third order binary gait silhouette volume (GSV) as shown in Fig. 1. Once, NDM for each frame from GSV is computed separately, it is repeated for all frames in GSV and as a result third order tensor representation based on NDM is obtained. It is then further processed by using statistical tensor based dimensionality reduction technique called as SMLDA [17]. Following section gives a brief overview of SMLDA.

4 Sparse Multilinear Laplacian Discriminant Analysis

It is a multilinear discriminant subspace learning derived from previous work [17]. The weighted Laplacian scatter difference along with sparsity constraint derives, sparse matrices $\mathbf{W}_k \in R^{m_k \times D_k}, m_k < D_k, k = 1, 2, 3$ that map $\mathcal{X}$ in to reduced tensor space.

$$\mathcal{Y} = \mathcal{X} \times_1 \mathbf{W}_1^T \times_2 \mathbf{W}_2^T \times_3 \mathbf{W}_3^T \in R^{m_1 \times m_2 \times m_3} \tag{4}$$

The key idea for SMLDA is to provide discrimination; when samples are from different classes and minimize variation, when samples are from the same class. It has sparsity constraints in the form of L_1 and L_2 norms penalty. The objective function is given by

$$J\left(\mathbf{W}_k\right)\Big|_{k=1}^{N} = \min tr\left(\mathbf{W}_k^T\left(\mathbf{LS}_w^{(k)} - \lambda_k * \mathbf{LS}_B^{(k)}\right)\mathbf{W}_k\right) + \alpha_k \left\|W_k\right\|^2 + \sum_j \beta_{kj}\left|w_{kj}\right| \text{ Subject to } W_k^T W_k = I_k \quad (5)$$

where $|\cdot|$ and $\|\bullet\|$denote L_1 and L_2 norm respectively, λ_k is weight parameter.

The projection matrices are computed from the gallery NDM sequence by training. Locality constrained group sparse representation (LGSR) classifier [18] is used to classify the projected low dimension features. Recognition accuracies are computed with Rank 1 ("R1") correct classification rate.

5 Experimental Evaluation

The proposed approach is validated using USF benchmark [19] and OU-ISIR data sets [20]. The recognition rate in terms of absolute as well as relative correct classification rates (CCR) are reported.

5.1 Experimentation Settings

The gait video sequence is preprocessed by background subtraction, binarization, spatial normalization, gait period detection and temporal alignment processes. As suggested in El-Alfy et al. [14], NDM is computed cell wise for each frame and the same procedure is repeated for all frames in one gait cycle. The processing for the gallery and probe sequence is repeated to obtain processed NDM tensor. Once the tensorial NDM sequence is available for both gallery and probe sets, similarity matrix based on tensor coding length is computed. The approach of minimum reconstruction error is used in the locality constrained group sparse representation (LGSR) is used for classification of probe binary gait tensor in to one from a gallery.

5.2 Experiments on OU-ISIR Dataset

OU-ISIR is the worlds largest and widely preferred available gait database with 1872 females and 2135 males totaling more than 4007 subjects with maximum diversity of covariate.

Experiment on OU-ISIR dataset A As suggested by El-Alfy et al. [12] all guidelines for experimental protocol are followed. The entire database is divided in to five sets as A-55 to A-85 and A-All. Table 1 shows the performance of the proposed technique.

Table 1 Recognition rate for OU-ISIR dataset

Dataset	No. of Subjects	GEI	HONV	NDM [14]	NDM SMLDA (TCL)
A-55	3706	84.7	91.6	94.1	95
A-65	3770	86.6	92.1	95	96.2
A-75	3751	86.9	93.3	95.7	97
A-85	3249	85.9	93.0	95.9	96.6
A-All	3141	94.2	97.5	98.1	97.2
Average		87.6	93.5	95.8	96.4

5.3 Experiment on USF Database

The USF dataset by Sarkar et al. [19] has a total of 1870 gait sequences captured in an outdoor environment for 122 subjects. Entire dataset is divided into two categories, one for the gallery and another category as twelve probe sets (A–L) having diverse variations: like shoe, surface, view points, carrying condition and time. Table 2 shows the performance of proposed technique.

Table 2 Rank-1 correct classification rate (%) for USF dataset

Probe set	Probe size	Baseline [19]	CGI Fusion [21]	Gabor-PDF + LGSR [18]	Gabor + RSM + HDF [22]	NDM + SMLDA (TCL)
A	122	73	91	95	100	100
B	54	78	93	93	95	97
C	54	48	78	89	94	91
D	121	32	51	62	73	78
E	60	22	53	62	73	75
F	121	17	35	39	55	60
G	60	17	38	38	64	50
H	120	61	84	94	97	87
I	60	57	78	91	99	93
J	120	36	64	78	94	93
K	33	3	3	21	42	45
L	33	3	9	21	42	45
Average		43	56	65	77	76

6 Discussion

Comparing the results in Tables 1 and 2, the following observations can be drawn:

1. When single-template-based or matrix based gait representation approach is compared with the third-order tensor representation, it outperforms and the average correct recognition rate is much improved. The simplest justification is the preservation of inherent structural information.
2. The covariate leads to partial feature corruption problems. The proposed representation preserves inherent correlation by retaining tensor based structural information. Simply by avoiding vectorization, robustness are improved by restricting corruption of features.
3. Tensor coding length (TCL) approach uses natural correlation of pixels from their spatial locations to reduce contamination thereon.

6.1 Timing Analysis

The timing analysis of the discriminative feature extraction scheme and the iterative projection-method-based optimization procedure of SMLDA is done by measuring the amount of time required for execution. The code of our method is run on a PC with an Intel Core i7 3.5 GHz processor and 16GB RAM. For USF dataset, the training time for the tensor coding length approach is reported as 300 s and query time as 0.42 s.

7 Conclusion

In this paper, a robust representation approach called Normal Distance Map and multilinear statistical discriminant analysis called Sparse Multilinear Discriminant Analysis is applied for improving robustness against covariate variation and increase recognition accuracy. NDM captures geometry and shape of silhouettes so as to make representation robust and SMLDA obtains projection matrices that preserve discrimination.

References

1. Makihara Y, Matovski DS, Nixon MS, Carter JN, Yagi Y (2015) Gait recognition: databases, representations, and applications. Wiley Online Library
2. Sivarathinabala M, Abirami S, Baskaran R (2017) A study on security and surveillance system using gait recognition. In: Intelligent techniques in signal processing for multimedia security. Springer, Berlin, pp 227–252

3. Zhang Z, Hu M, Wang Y (2011) A survey of advances in biometric gait recognition. In: Chinese conference on biometric recognition. Springer, Berlin, pp 150–158
4. Boulgouris NV, Hatzinakos D, Plataniotis KN (2005) Gait recognition: a challenging signal processing technology for biometric identification. IEEE Signal Process Mag 22(6):78–90
5. Wang J, She M, Nahavandi S, Kouzani A (2010) A review of vision-based gait recognition methods for human identification. Digit Image Comput: Tech Appl pp 320–327
6. Han J, Bhanu B (2006) Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 28:316–322
7. Huang X, Boulgouris NV (2012) Gait recognition with shifted energy image and structural feature extraction. IEEE Trans Image Process 21:2256–2268
8. Bashir K, Xiang T, Gong S (2009) Gait recognition using gait entropy image. In: In 3rd international conference on crime detection and protection, London, UK
9. Lam THW, Cheung K, Liu JN (2011) Gait flow image: a silhouette-based gait representation for human identification. Pattern Recogn 44:973–987
10. Makihara Y, Sagawa R, Mukaigawa Y, Echigo T, Yagi Y (2006) Gait recognition using a view transformation model in the frequency domain. In: European conference on computer vision. Springer, Berlin, pp 151–163
11. Hofmann M, Bachmann S, Rigoll G (2012) 2.5 d gait biometrics using the depth gradient histogram energy image. In: 2012 IEEE fifth international conference on biometrics: theory, applications and systems (BTAS). IEEE, New York, pp 399–403
12. El-Alfy H, Mitsugami I, Yagi Y (2014) A new gait-based identification method using local gauss maps. In: Asian conference on computer vision. Springer, Berlin, pp 3–18
13. Tang S, Wang X, Lv X, Han TX, Keller J, He Z, Skubic M, Lao S (2012) Histogram of oriented normal vectors for object recognition with a depth sensor. In: Asian conference on computer vision. Springer, Berlin, pp 525–538
14. El-Alfy H, Mitsugami I, Yagi Y (2017) Gait recognition based on normal distance maps. IEEE Trans Cybern
15. Gauss KF (1902) General investigations of curved surfaces of 1827 and 1825
16. Hazewinkel M (2001) Encyclopaedia of mathematics, vol 13. Springer, Berlin
17. Chhatrala R, Patil S, Lahudkar S, Jadhav DV (2017) Sparse multilinear Laplacian discriminant analysis for gait recognition. Pattern Anal Appl pp 1–14
18. Xu D, Huang Y, Zeng Z, Xu X (2012) Human gait recognition using patch distribution feature and locality-constrained group sparse representation. IEEE Trans Image Process 21(1):316–326
19. Sarkar S, Phillips P, Liu Z, Vega IR, Grother P, Bowyer K (2005) The humanid gait challenge problem: data sets, performance, and analysis. IEEE Trans Pattern Anal Mach Intell 27:166–177
20. Iwama H, Okumura M, Makihara Y, Yagi Y (2012) The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans Inf Forensics Secur 7(5):1511–1521
21. Wang C, Zhang J, Pu J, Yuan X, Wang L (2010) Chrono-gait image: a novel temporal template for gait recognition. In: European conference on computer vision. Springer, Berlin, pp 257–270
22. Guan Y, Li CT, Roli F (2015) On reducing the effect of covariate factors in gait recognition: a classifier ensemble method. IEEE Trans Pattern Anal Mach Intell 37(7):1521–1528