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.

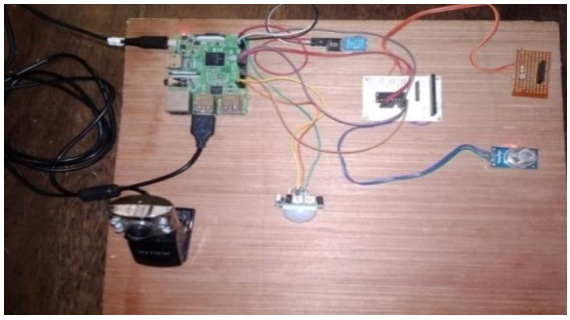

Fig. 1 IoT Enabled System with Face Recognition

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*:

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

I

$p(j|t)$ 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

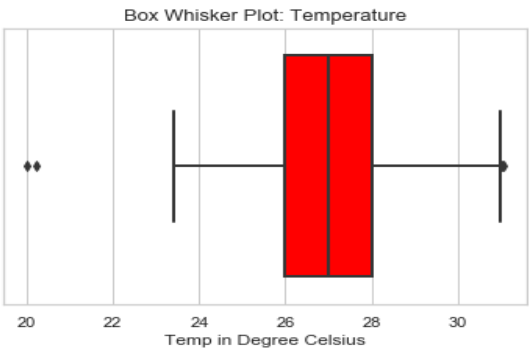

Fig. 5 Box Whisker Plot for Temperature

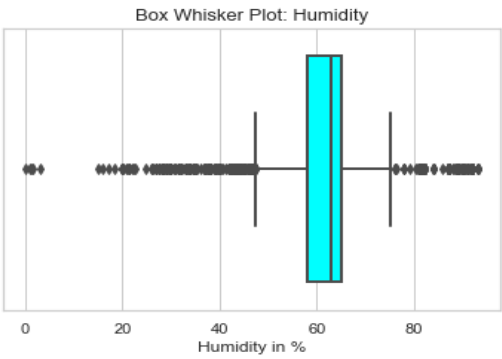

Fig. 6 Box Whisker Plot for Humidity

Fig. 7 Box Whisker Plot for LDR

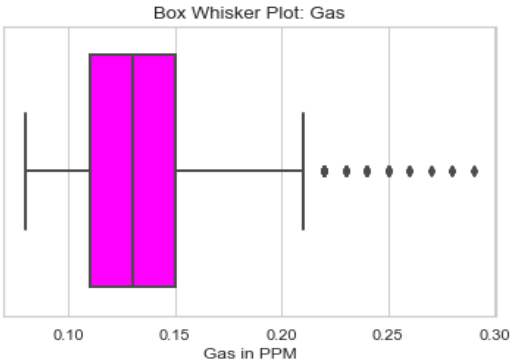

Fig. 8 Box Whisker Plot for Gas

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.

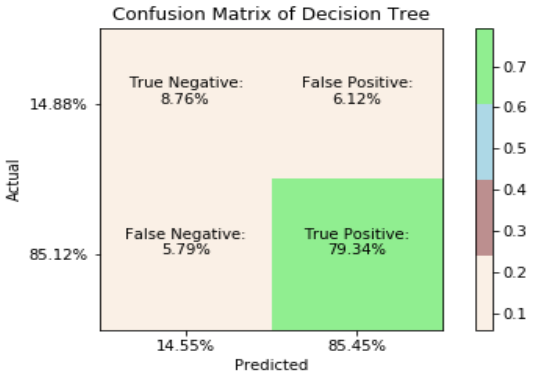

Fig. 9 Confusion Matrix of DT

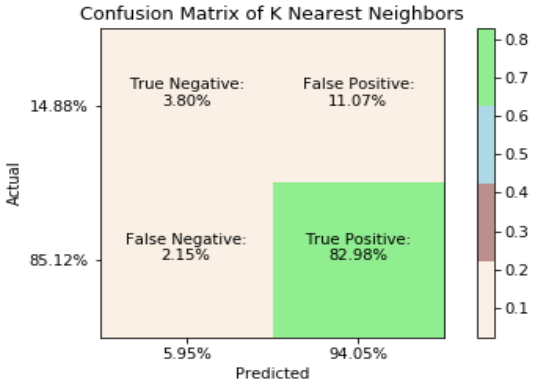

Fig. 10 Confusion Matrix of KNN

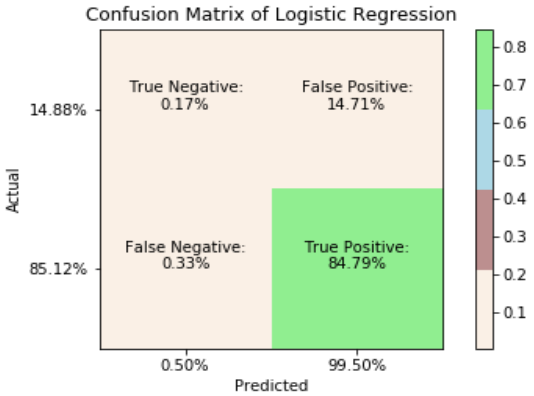

Fig. 11 Confusion Matrix of LR

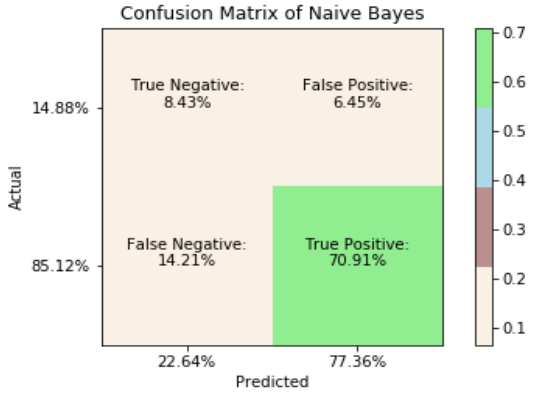

Fig. 12 Confusion Matrix of NB

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

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.

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

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.

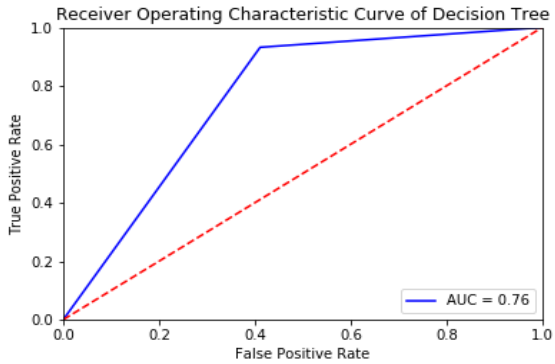

Fig. 13 ROC Curve of DT

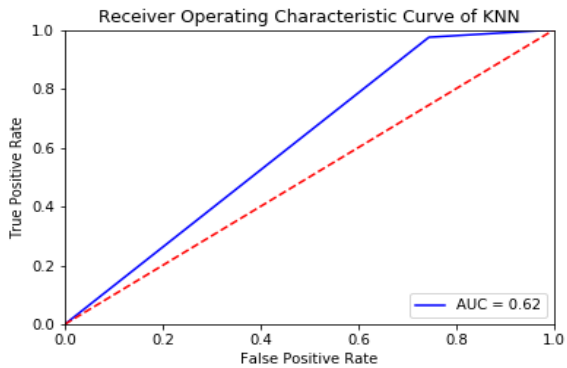

Fig. 14 ROC Curve of KNN

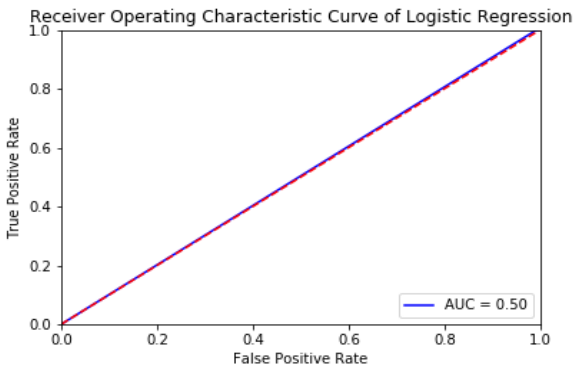

Fig. 15 ROC Curve of LR

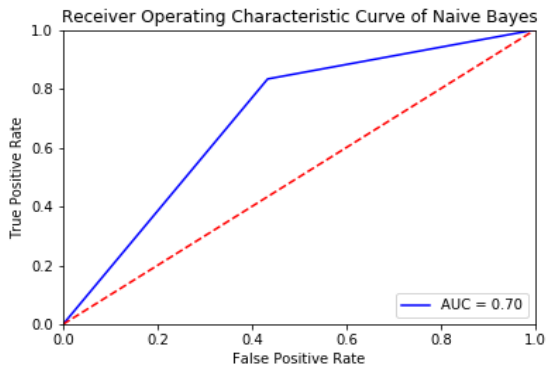

Fig. 16 ROC Curve of NB

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.

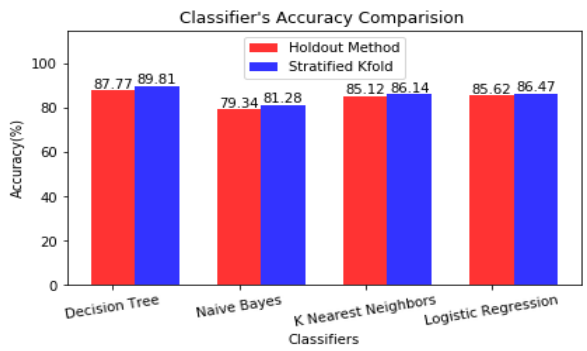

Fig. 17 Accuracy Comparison Chart of Four Classifiers

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.