International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 15, No. 2, June 2026, pp. 869~879
ISSN: 2252-8776, DOI: 10.11591/ijict.v15i2.pp869-879

Real-time emotion prediction system using big data analytics

Manpreet Kaur Dhaliwal, Rohini Sharma, Rajbinder Kaur
Department of Computer Science and Applications, Panjab University, Chandigarh, India

Article Info

Article history:

Received Oct 16, 2024
Revised Nov 12, 2025
Accepted Dec 14, 2025

Keywords:

Big data
Cloud
GSR
Outliers
Real-time

ABSTRACT

Emotions are an inseparable part of human existence. Emotions have a big impact on the success and failure of the human race. Comprehending human emotions could prove beneficial in creating improved systems for education, security, market sales, production, healthcare and other areas. Big data analytics applied to streamlined real time emotion sensor's data can give new insights to anticipate stress before it arises and help in making significant choices that improve people's quality of life. This work proposes a framework for big data management and analysis of GSR sensor's data in real-time for predicting emotions in human participants. Supervised learning techniques, ensemble boosted tree, neural network, Naïve Bayes, support vector machine, decision tree, K-nearest neighbor, and quadratic discriminant analysis are applied to the collected data. Two distinct criteria have been utilized for testing on real-time data one is trained on all participant data, resulting in a generalized system, while the other is trained on participant-specific data, resulting in a personalized system. Hence, the personalized system achieves an accuracy of up to 80.64% across all classes and 100% for binary classes as compare to generalized system achieves 78.12% accuracy. It is concluded that for the purpose of predicting emotions, the personalized model performs better than the generalized model.

Corresponding Author:

Rohini Sharma
Department of Computer Science and Applications, Panjab University
Chandigarh, India
Email: rohini@pu.ac.in

1. INTRODUCTION

The healthcare industry uses a wide range of technologies, including big data, cloud computing, machine learning, artificial intelligence, and the internet of things (IoT) for remote health monitoring, epidemic detection, and forecasting, among other applications. According to projections, the value of the worldwide healthcare artificial intelligence industry is expected to reach almost 188 billion dollars by 2030, growing at a compound annual growth rate of 37% between 2022 and 2030 [1]. According to a forecast by the international data corporation (IDC), there will be 41.6 billion IoT devices by 2025, with a data generation capacity of 79.4 zettabytes (ZB) [2]. The IoT will produce brontobyte-sized data in the future [3]. Scalable diverse data that expands exponentially is referred to as "big data". The definition of big data comprises six terms: Volume, which denotes the massive quantity of data generated or gathered; Velocity, which describes the speed at which the data is gathered and continuously updated in real-time; Variety, which describes the variety of sources from which the data is obtained, such as structured, unstructured, or data acquired from sensors; Variability, which describes the variable nature of the data, Veracity that involves the errors and noise in the data; and Value, which describes the usefulness of the data for any application that makes use of the concept of big data. Big data provides numerous opportunities in the fields of health informatics, medical informatics, bioinformatics, or clinical informatics for researchers but there are

several challenges to data management like cleaning, noise, balancing, privacy, and analysis [4]. Big data, from the perspective of many, still refers to a high volume of data [5] but it is not about only quantity variety, speed, and value are also important factors that contribute to Big data. Our study focuses on the variety, velocity, value, and veracity characteristics of big data.

Real-time data collection and processing is essential in today's world because delayed information may lead to loss of lives. The conjunction of cloud and big data techniques are useful in real-time data storage, analysis, and continuous monitoring of health parameters. Various frameworks have been suggested in literature based on IoT-cloud systems, Dhaliwal *et al.* [6], proposed a five-layered hybrid architecture using cloud and wearable sensors. Sensor data is used for real-time analysis, prediction, and sending alerts to caretakers if an emergency arises. Sareen *et al.* [7], suggested a wireless body area network and RFID-based architecture for real-time remote patient monitoring via cloud computing to identify and track Ebola patients. Otoom *et al.* [8], proposed an architecture based on the Internet of Things and the cloud that would gather symptoms in real-time from wearable technology and apply a machine learning algorithm to forecast and track COVID-19 instances. Healthcare professionals utilize these anticipated data to react to patients promptly. In a similar vein, [9]-[14] suggested sensor-cloud-based systems for illness prediction and alerting caregivers. In the literature, the proposed frameworks lack real-time analysis of real data and cloud storage. Real-time sensor data analysis can improve the quality of information, and anticipate emergencies like heart attacks, cardiovascular disease, epidemics, and track infected patients. In this paper, a framework has been proposed and a proof of concept has been presented based on big data analytics and management specifically to handle wearable sensor data in real-time scenarios. Experiments have been performed to identify the predictive models that are most accurate at predicting the condition. The projected outcomes undergo scrutiny, assessment, and illustration. The major contribution of this article is as follows:

- A framework is proposed for big data management and analysis specifically to handle wearable sensor's data in real-time scenarios.
- The utilization of open source IoT platform tool i.e. ThingSpeak for data storage, processing, and analysis. ThingSpeak cloud ensures big storage capacity, security, and real-time analysis of data using the MATLAB tool.
- Real-time descriptive analysis on the big sensors data for detecting outliers.
- Multiclassification of emotions i.e. sad, scary, happy, hold, deep breath, and normal on GSR data and Binary classification is performed to classify emotions by selecting the different subjects randomly.
- Predictive analysis is performed on real-time data for emotion prediction of the subject.

In the rest of the paper, related studies and motivation are discussed in section 2. In section 3, methodology, and in section 4 experimental results and discussion are given. Section 5 covers the conclusion and future directions.

2. RELATED STUDIES

Ayata *et al.* [15] used time domain, wavelet, and empirical mode decomposition aspects of GSR signals to identify emotions. Using K-nearest neighbor (KNN), decision tree (DT), random forest (RF), and support vector machine (SVM) algorithms emotions are classified. The accuracy rates for arousal and valence were 81.81% and 89.29%, respectively using the RF classifier. Santamaria-Granados *et al.* [16] proposed a method to classify emotions on the AMIGOS dataset. In this study, 40 participants are considered and ECG, GSR features are used. Deep convolutional networks achieved 71% and 75% accuracy for arousal and valence classes. Komuro *et al.* [17] proposed a system to recognize individual emotions using environmental, skin temperature and pulse sensors. Three machine learning (ML) models SVM, KNN, and RF are used with an accuracy of 74.7%, 82.1%, and 86.7% respectively. Sen and Pal [18] conducted a study using EMG and GSR sensors for pain assessment on the BioVid database. Three ML techniques DT, linear support vector machine (LSVM), and logistic regression (LR) are implemented to estimate the level of pain. LR achieved 80% accuracy while LSVM and DT achieved 73.8% and 76.2% respectively. Chueh *et al.* [19] analyzed the data related to ECG, skin temperature variation (SKT), and GSR using MANOVA and six ML methods KNN, Bayesian network learning, SVM, logistic model, DT of C4.5, and Naive Bayesian classification are used to predict emotions. Logistic model achieved 74.76% accuracy. For automatic emotion recognition Kim and Andr´ [20] considered physiological features i.e., ECG, EMG, Skin conductance, and respiration. After the extraction of features using pseudoinverse LDA (pLDA), an emotion-specific multilevel dichotomous classification (EMDC) scheme is used to classify four musical emotions. The EMDC scheme achieved 70% accuracy for all subjects. Soleymani *et al.* [21] used the ANOVA test to analyze the emotions during movie scenes using physiological signals captured from ECG, EMG, temperature, respiration and GSR. The author concluded that MANOVA and logistic regression significantly improve the

performance of the system. For the safety of women and children Alisha *et al.* [22], proposed an IoT-based wearable device using LM35, EDA, and a 3-axis accelerometer. The data is collected from 50 subjects representing two classes 'stressed' and 'relaxed'. ThingSpeak cloud receives and monitors the sensors' data continuously. If subjects are stressed, activity is recognized and an alert is sent to the caretaker in real-time. ML is applied only to activity data but activities are only recognized when a person is stressed. Mohamad *et al.* [23] proposed an IoT-cloud-based system that employed principal component analysis (PCA) algorithm for classifying heart signals received from sensors. Thomas *et al.* [24] used LM35 sensor to monitor the temperature of the body connecting to the Arduino board. Saha *et al.* [25] designed a prototype using a DS18B20 temperature sensor that is connected to a board using Zigbee technology to provide vital health information to the person in real-time after analyzing the data. The results obtained from the prototype are satisfactory as the researcher validated it using the traditional thermometer to compare the results. Anandh and Indirani [26] used LM35 and Ad8232 sensors to collect the data, collected signals are then transferred to the ThingSpeak cloud using an ESP8266 Wi-Fi module. Rahimoon *et al.* [27] used 2 temperature sensors (LM35 and MLX-90614) for continuous monitoring of body temperature in real life. Nduka *et al.* [28] designed a system for monitoring vital signs of the human body in real-time. The temperature, Heart rate, and respiration of a person are displayed on the user's as well as the physician's system to provide emergency treatment. The temperature received from the sensor is also compared with the temperature readings taken from Omron digital temperature. The results prove that the proposed system provides the readings accurately. As the LM35 sensor has been used successfully. This sensor is used in this study too for real-time analysis on the cloud.

In this study, Jameil and Al-Raweshidy [29] employed a state-of-the-art technology called a digital twin to track and identify a person's diabetic status through cloud computing. Additionally, it incorporates anomaly detection and secure data transfer techniques. Using an ESP8266 module linked to MAX30102 and MLX90614 sensors, body temperature (BT), SPO2, and HR feature data were sent to the cloud. Overall, the study's accuracy was 98.9%, its real-time accuracy was 95.4%, and its F1 score on real-time analysis was 98.4%. Papri *et al.* [30] created a non-invasive diabetes detection system using AWS for real-time photoplethysmography (PPG) signal data analysis. The XGboost model analyzes PPG signals on the AWS cloud platform to forecast diabetes illness in real time. Here, ML models were trained in Python using two distinct datasets, Mazandaran and PPG-BP. The models were then deployed in the cloud for the binary classification of diabetic or non-diabetic classes, and a web application was created using the Flask API to show a person's diabetic condition. By using XGboost, 87.88% accuracy and 93.75% recall were attained on the PPG-BP dataset and 90% accuracy and 75% recall were attained on the Mazandaran dataset.

A real-time cloud-based method for diagnosing atherosclerosis was created by the Babu *et al.* [31]. KNN and K-means clustering (KMC) were used to train the suggested system after the dataset from 300 patients was gathered using an Android application. Here, KNN performs better than KMC, with 91% specificity and 87% accuracy. This proposed system continuously gathers health data using cloud computing technologies and NoSQL databases. In the app step, subjects can utilize the mobile app to view their usage reports. Hamzehei *et al.* [32] study's main topics include ML techniques and Parkinson's disease. Motor is the most significant predictor of the total unified Parkinson's disease Rating Scale, according to the analysis, indicating the importance of movement-related symptoms in the assessment of Parkinson's disease. The study used a dataset on Parkinson's disease from the UC Irvine ML repository, which was curated by University of Oxford academics, and applied ML techniques to it. The optimum coefficient of determination was found using the Adam optimization algorithm, which was one of four linear regression techniques used: least linear regression, conjugate gradient, ridge regression, and Adam optimization. The cloud-based Python code execution tool is Google Colab.

Wang *et al.* [33] suggest a HealthAIoT framework that uses proven ML algorithms, such as random forest and logistic regression, to detect diabetes. The CDC's Diabetes Health Indicators Dataset is used in the study. There are 253,680 participant records in the dataset. Predictive analysis and ongoing healthcare monitoring are supported by advanced ML algorithms. On test data, the diabetes prediction model's overall accuracy is 78.30%. It deploys assessments using CloudAIBus on the Google Cloud Platform. An AI-based testbed called CloudAIBus was created to enhance resource distribution in cloud settings. In terms of execution time and latency, it performs better than FedSDM and HealthFog. Compared to HealthFog, the framework uses 20.11% less power.

Swain *et al.* [34] suggested utilizing cloud computing technologies and sensors (humidity, heart rate, temperature, and light-dependent resistors) to create a digital twin for emotional well-being. Connected to these sensors, a microcontroller unit (MCU), like the ESP32 or NODEMCU, serves as the central location for gathering data. The MCU combines the sensor data and prepares it for transmission to the cloud. An AWS EC2 instance acting as a server receives the data safely. Heart rate and environmental data are analyzed by ML models such as ridge regression, linear regression, and random forest regressor. For both indoor and outdoor settings, each model was trained independently. Sensor data from temperature, humidity, and

brightness in relation to heart rate are included in the dataset. It is concluded that in indoor temperature and in outdoor humidity is the most influential factor in heart rate predictions.

As shown in Table 1, in most of the studies, own dataset is collected for experiments in contrast to [15], [16], [18], [21], [33]. Worked on cloud storage and real-time analytics. performed statistical analysis using the ANOVA parametric test [22], [23], [29], [30], [31], [33], [34]. Ayata *et al.* [15] achieved the highest accuracy of 89.29% using the RF classifier for emotion dataset, whereas [29] achieved 95.4% accuracy with MLP and XGboost on the Diabetic dataset in real-time. As compare to [29] our study is achieving 100% accuracy on binary classification and on multiclassification 80.64% is achieved. In [16], only two emotions are classified and 81% accuracy is achieved for only the arousal class of ECG sensor data. Otherwise for GSR in arousal and valence, it is achieving 71% and 75% accuracy using DCNN. A benchmark dataset of big size is used [15], [16], [18], [33]. Komuro *et al.* [17], arousal and valence are divided into four classes, and using RF it is achieving 86.7% accuracy. The study classified four classes using only two features and environmental factors. However, the environmental factors might not be the characteristic features for predicting emotions. It is concluded from Table 1 that most of the studies are performing binary classification on datasets whereas in this study, binary as well as multiclass classification with six classes is performed with a good accuracy.

Table 1. Comparison of existing studies

Ref.	Sensors used	Dataset used	Machine learning models	Highest accuracy	Cloud /big data analytics	No of subjects	No. of classes
[15]	GSR	DEAP	SVM, KNN, RF, DT	89.29%	No	32	2
[16]	ECG, GSR	AMIGOS	NB, KNN, LDA, Multilayer perceptron, Adaboost, RF, DCNN	81%	No	40	2
[17]	Pulse, Temp, environmental sensors	Own dataset	SVM, KNN, RF	86.7%	No	10	4
[18]	EMG, GSR	Biovid Pain database	LSVM, LR, DT	80%	No	90	2
[19]	SKT, GSR, ECG	Own dataset	C4.5, Bayesian Net, NB, SVM, KNN, Logistic	74.76%	No	10	3
[20]	SC ECG, EMG, RSP	Own dataset	EMDC, pLDA	70%	No	3	4
[21]	ECG, GSR, EMG Respiration, Temp	Own dataset	Statistical Analysis ANOVA	Significance value less than 0.005	No	8	2
[22]	LM35, EDA, 3-axis Accelerometer	Own dataset	-	-	Yes	50	-
[29]	BT, SPO2 and HR	MIMIC-III Dataset and own data	MLP and XGBoost	95.4%	Yes	20	2
[30]	PPG	PPG-BP, Mazandar dataset and own data	SVM, RF, XGBoost and LR	90% on Mazandan dataset	Yes	-	2
[31]	-	Own dataset	KNN, and K-means clustering	87%	Yes	300	2
[33]	-	CDC's Diabetes Health Indicators	RF, LR	78.30%.	Yes	253,680	2
[34]	Humidity, HR, light dependent resistors, Temp,	Own dataset	RF regressor, ridge, linear regression,	-	Yes	12,377 rows	2
Our Study	GSR, LM-35	Own dataset	Ensemble boosted Tree, KNN, NN, NB, SVM, DT, And QDA	80.64% 100%	Yes	12	6 2

In the current study, a framework has been proposed for classifying emotion using big sensor data. The evaluation has been performed to classify six classes of emotions. An accuracy of 80.64% has been achieved on multiclassification and 100% has been achieved on binary classification. The research in the field of real-time storage and analysis of big emotions data is still in a nascent stage. However, the applications are huge including identification of stress and anxiety in students, sports persons, disabled or visually impaired persons and elderly people, and human-computer interaction.

3. RESEARCH METHOD

The proposed framework has two layers: Big data management layer and Big data Analytics layer. The framework is shown in Figure 1. In the proposed design, the wearables (GSR and LM35 sensors) data are transferred to the cloud using a WiFi module. The cloud data are used by the suggested framework for processing and classifying in real-time. The big data outliers are eliminated as part of the preprocessing and the data are then imported into MATLAB for further processing.

3.1. Big data management

In this layer, sensor's data are read and transferred to the cloud for storage. The layer consists of sensors, modules, microcontrollers, and storage units. The galvanic skin response (GSR) sensor is a device that gauges the skin's electrical conductance. It tracks alterations in the autonomic nerve system, which encompasses levels of stress, mental state, and physical excitement. The temperature sensor LM35 has a temperature measuring range of -50°C to 150°C and an operating voltage of 4 V to 30 V.

The GSR sensor is placed in the first two fingers of the right hand of the subject as shown in Figure 2 and the temperature sensor is touched by left hand fingers while collecting data. Esp8266 is a system-on-chip Wi-Fi module that enables microcontrollers to connect to 2.4 GHz Wi-Fi, using IEEE 802.11 bgn. Certain functions are carried out by it, like gathering sensor data, managing other devices, and sending and receiving data via the internet. For connecting and managing other electronic parts like sensors, LEDs, or motors, it features general purpose input/output (GPIO) pins [35]. An analytics platform built on the Internet of Things, ThingSpeak Cloud can store, process, and analyze data in real time. Additionally, it offers data visualization from live streams [36]. For this investigation, a channel is created and two fields for temperature and GSR sensor data are generated. Real-time data of sensors and analysed results are shown in Figures 3 and 4.

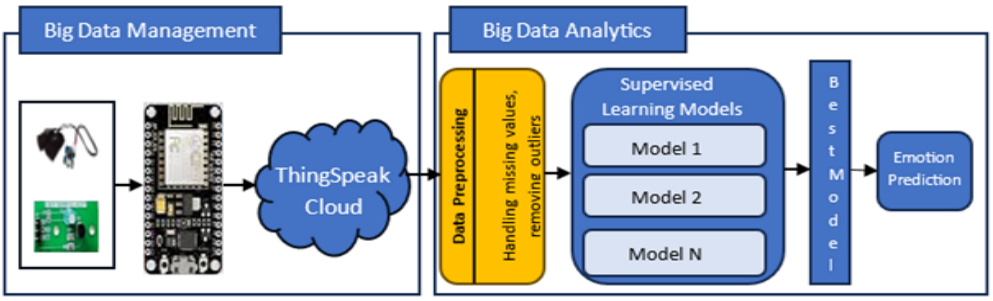


Figure 1. Proposed framework for big data transmission and analysis

3.2. Big data analytics

Big data analytics refers to the different analytics techniques that are used to identify various patterns, knowledge for decision-making, association between data, statistical analysis, ML analysis, and visualization of data. This layer is divided into subparts i.e. Preprocessing of data and MATLAB analysis. In Preprocessing outliers are detected in data coming from sensors in real time for LM35 and GSR sensors values and cleaned data is used for prediction of emotions.

3.2.1. Preprocessing of data

Exploring ML with big data requires preprocessing of the data, which is an essential step. As big data is streamed and comes from a variety of sources, its quality needs improvement. The accuracy and functionality of machine-learning models can be impacted by noise, outliers, and missing values in sensor-derived data [37]. Bad decisions and inefficient use of data are caused by low-quality data [38], [39]. Thus, in this study, we employed the technique of identifying and eliminating outliers for further analysis.

Outlier detection: An outlier is an unwanted data point that is far away from the normal data points and generated because of noise, loss of signal in sensors, and inappropriate readings of sensor signals [40]. There are various applications in real life where outlier detection techniques are used, like fraudulent detection, fault diagnosis, abnormal sensor readings, and intrusion detection in networks and medical informatics. There are three types of outliers: Point, Contextual, and Collective. In point outliers, one point is deviated from the rest of the dataset. Contextual outliers are data points' values that substantially differ from the other data points in the same setting. Collective outliers are a group of data points that are different from the complete dataset [41]. Various visualization, statistical, and ML techniques are available to identify outliers like box plots, Z-score, mean, median, DBSCAN, and isolation forest.

Steps to detect outliers in this study:

- Read the GSR and temperature sensor signals.
- Transfer the data to the ThingSpeak cloud for storage using the ESP8266 module.
- Realtime Analysis in MATLAB.
    1. Read the field data from the Channel.
    2. Calculate the 3-scale median for outlier detection.

        3-*scale median = a*median(abs(G-median(G)))*, *where a=-1/(sqrt(2)*erfcinv(3/2))* and G represents Sensor value

        where erfcinv() represents the inverse complementary error function
    3. Visualise the complete data.

Outlier detection results: Real-time visualized results based on GSR and temperature are shown in Figures 3(a)-3(c) and Figures 4(a)-4(c). The real-time analysis provides two types of data: clean and outlier data. Figure 3(a) represents the real-time readings of GSR signals. The data are processed in real-time, and instant cleaning is essential for further processing to get an accurate analysis.

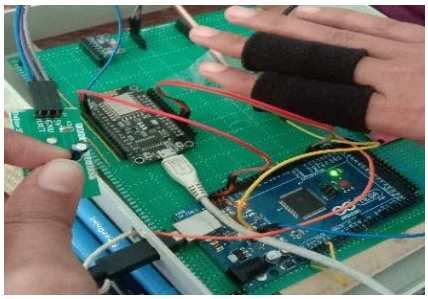
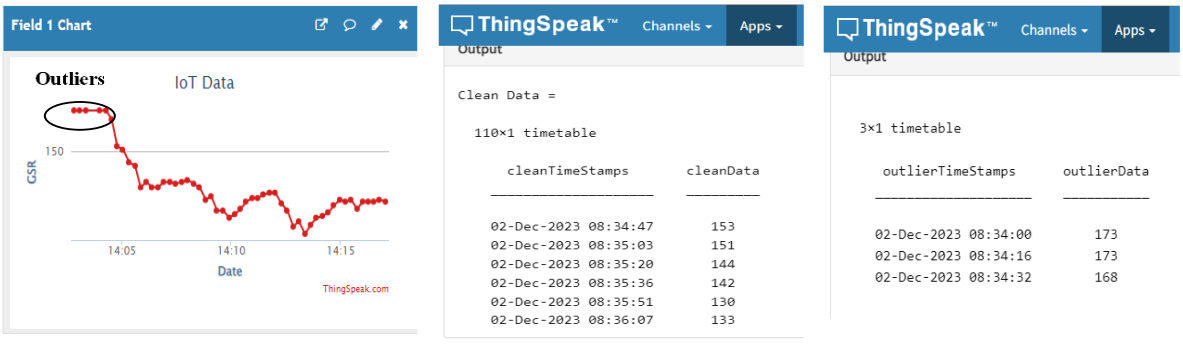
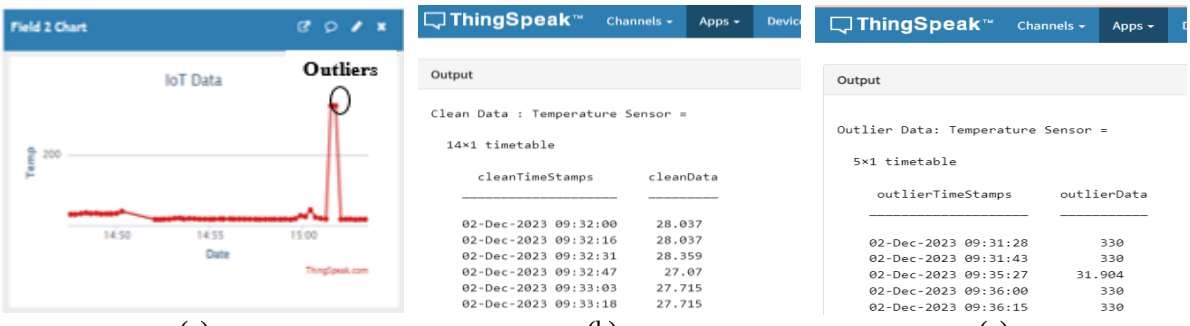


Figure 2. Prototype used for realtime data transmission to the cloud



(a) (b) (c)

Figure 3. Real time visualization of GSR data: (a) real-time readings of GSR signal, (b) clean data in GSR signals, and (c) outliers data in GSR signal



(a) (b) (c)

Figure 4. Real time visualization of Body's temperature data: (a) real-time readings of temp signal, (b) clean data in temp signals, and (c) outliers data in temp signal

3.2.2. Supervised learning analysis

Pre-processed data is used for real-time testing for classifying emotions. During the experimentation, ML, deep learning, and boosting methods viz. DT, NB, QDA, SVM, KNN, NN, and Ensemble Boosted Tree respectively are used. The models are trained on PIF v1 Dataset [42], [43].

4. RESULTS AND DISCUSSION

For the experimental evaluation of the data collected and managed in real time, trained model is required. Therefore, the models were first trained and tested on PIF v1 dataset [42], [43]. This dataset contains the data related to six emotions Scary, Sad, Hold, Normal, Deep Breath, and Happy. The data are divided into 75:25 for training/testing and 10-fold cross-validation is used for experimentation. All the experiments are performed using MATLAB analysis. The trained models are then used for making predictions on real time data.

4.1. Training and testing on PIF v1 dataset

Experiments are performed on combined data of all subjects, and results are shown in Table 2. It is noticed that highest accuracy achieved is 46.88% accuracy using SVM. However, training the model subject-wise gives better results. Six participants are chosen randomly and the models are trained subject-wise as shown in Figure 5. For PID 1, the DT outperforms with an accuracy of 65.79%. In the case of PID 4, SVM outperforms other techniques with an accuracy of 69.02%. Ensemble boosted tree, NN, and DT achieved 80.64% accuracy for PID 6 and for PID 7 NN and NB achieved 79.42% accuracy. For PID 9 and 10 NN achieved higher accuracy as shown in Figure 5.

In generalized system, where data from all combined subjects and all emotional states are considered, the overall performance of all the models are considerably low, with accuracy is falling below 50% as shown in Table 2. This concludes that how emotions vary from persons to person, making it difficult to predict. Additionally, the analysis indicates that models perform better when there is clear emotional variability in the subject's responses across activities. In contrast, for individuals who exhibit minimal emotional variation or respond similarly in diverse situations, make it harder for models to distinguish between states, resulting in lower accuracy. These findings emphasize the importance of emotional variability in improving the performance of emotion recognition systems.

Table 2. Classification results of six emotional states of all subjects

Models	Train	Test
Decision tree	43.85	43.8
Naïve Bayes	31.93	31.73
QDA	31.93	31.73
NN	46.84	46.6
SVM	47.1	46.88
Ensemble Boosted Tree	45.33	44.92
KNN	37.76	37.5

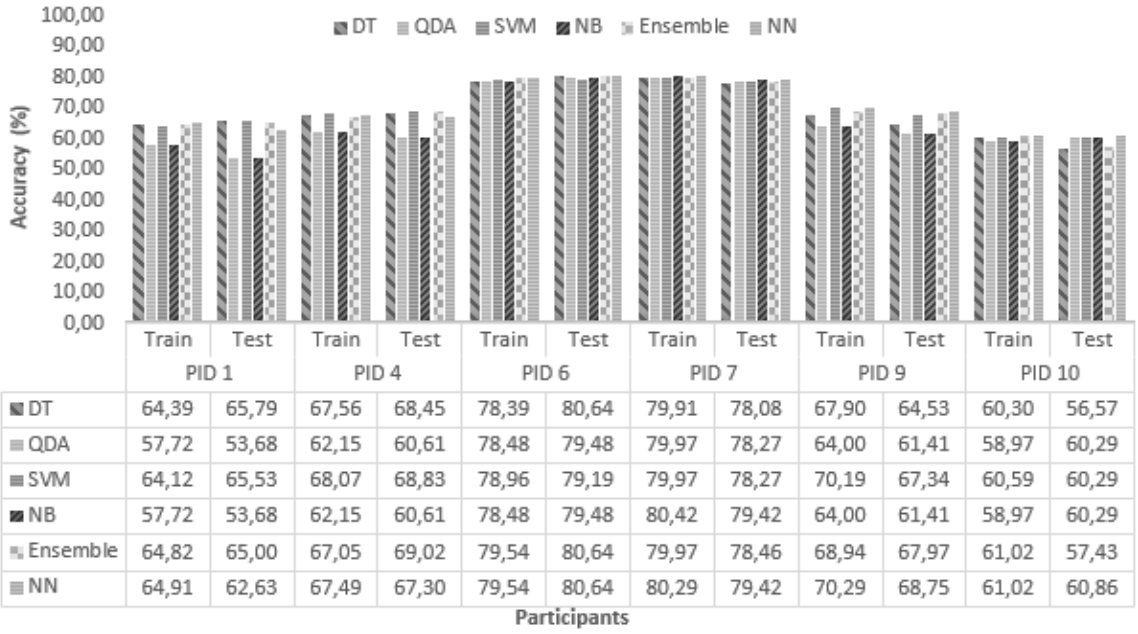
But still NN and SVM do better that other models, representing these models can handle variability and noise in data very well. On the other hand, when models are made for subject specific data that is a personalized approach is followed the results are much better as shown in Figure 5.

The experiments are also performed using binary classes that are normal and deep breath, sad and happy, and scary and happy on different subjects. The results of various models are given in Table 3. The results prove that the accuracy is far greater in binary classification as compared to multiclass classification. Yet, SVM and NN perform well across both binary and multiclass classification reinforcing their robustness and effectiveness in emotion classification tasks. Lastly, it is observed that QDA consistently underperforms in all scenarios. Whether using generalized or subject specific, or working with binary or multiclass classification, QDA consistently performs poorly. This means QDA probably is not a good choice for emotion detection especially in healthcare, where emotional data can be very complicated and unpredictable.

4.2. Testing on real-time data

For real-time testing, the prototype is worn by the user and data are transferred in real-time to cloud for MATLAB analysis using Supervised Learning Models as shown in Figure 1. Two models have been employed for testing on the real time data; one that is trained on complete data comprising of all the participants leading to a generalized system and the other one that is trained on the data pertaining to the specific participant leading to personalized system. The results are shown in Table 4. The personalized

training and testing give 90.62% accuracy as compared to a generalized system that gives up to 78.12% accuracy. It is pertinent to mention that data collected in real time is very small in size. The testing has been performed to understand the viability of the system. It has also been observed and concluded that the range of GSR values is different for different individuals. For instance, 265–310 raw GSR value denotes one person's scary readings, while for another, this value denotes the normal or sad class. So, personalized systems outperform generalized systems for prediction of emotions.



	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
	PID 1		PID 4		PID 6		PID 7		PID 9		PID 10	
DT	64,39	65,79	67,56	68,45	78,39	80,64	79,91	78,08	67,90	64,53	60,30	56,57
QDA	57,72	53,68	62,15	60,61	78,48	79,48	79,97	78,27	64,00	61,41	58,97	60,29
SVM	64,12	65,53	68,07	68,83	78,96	79,19	79,97	78,27	70,19	67,34	60,59	60,29
NB	57,72	53,68	62,15	60,61	78,48	79,48	80,42	79,42	64,00	61,41	58,97	60,29
Ensemble	64,82	65,00	67,05	69,02	79,54	80,64	79,97	78,46	68,94	67,97	61,02	57,43
NN	64,91	62,63	67,49	67,30	79,54	80,64	80,29	79,42	70,29	68,75	61,02	60,86

Figure 5. Performance comparison of various classifiers with six classes of emotions

Table 3. Performance of different classifiers for binary classification of emotions

Model		DT	QDA	SVM	NB	KNN	NN
PID 1	Normal/DB	88.35	80.58	88.35	80.58	87.38	88.35
	Happy/Sad	100	100	100	100	100	100
	Scary/Happy	100	100	100	100	100	100
PID 4	Normal/DB	78.79	80.30	80.30	80.30	81.82	81.82
	Happy/Sad	97.38	97.00	97.00	97.00	97.38	97.38
	Scary/Happy	93.56	89.77	93.56	89.77	93.56	93.56
PID 6	Normal/DB	96.67	93.33	96.67	93.33	96.67	96.67
	Happy/Sad	96.28	96.28	96.28	96.28	96.28	96.28
	Scary/Happy	98.51	97.01	99.25	97.01	98.51	98.51
PID 7	Normal/DB	100	97.96	100	97.96	100	100
	Happy/Sad	94.20	94.88	94.20	94.88	94.88	94.20
	Scary/Happy	99.04	99.04	99.04	99.04	99.04	99.04
PID 9	Normal/DB	100	100	100	100	100	100
	Happy/Sad	100	100	100	100	100	100
	Scary/Happy	96.15	96.15	96.15	96.15	94.76	96.15
PID 10	Normal/DB	95.10	88.24	95.10	88.24	95.10	95.10
	Happy/Sad	100	100	100	100	100	100
	Scary/Happy	71.39	71.39	72.73	71.39	71.39	72.73

Table 4. Prediction of emotions on real time data

Models	Model trained with PID 6 data (acc in %)	Model trained with all subjects data (acc in %)
DT	90.62%	78.12%
SVM	90.62%	78.12%
NN	90.62%	62.50%
Ensemble	90.62%	78.12%

5. CONCLUSION

This study introduces big data framework comprising of two layers for predicting emotions on real-time data using supervised learning techniques. In this framework, the open-source platform ThingSpeak Cloud is used for the storage, processing, and analysis of real-time data. Seven supervised learning methods DT, Gaussian Naïve Bayes, NN, SVM, QDA, Ensemble, and KNN are used to training and testing. The models are trained and tested using PIFv1 dataset giving accuracy of up to 46.88% for generalized system and 80.64% for personalized training and testing. The trained models are applied on real-time data that gives an accuracy of up to 78.12% for using generalized model and 90.62% for personalized model. It is concluded that personalized model outperforms generalized model for prediction of emotions. In future work, techniques will further be explored to make the generalized model more reliable and dependable. Other sensors like ECG, EMG, and Heart Rate. will also be included to widen the horizon of the study. It is further observed that processing speed for analytics is slow on cloud. So, edge network will be considered in future work.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Manpreet Kaur Dhaliwal	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Rohini Sharma		✓				✓		✓		✓		✓		
Rajbinder Kaur	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Mendeley at https://data.mendeley.com/datasets/phb9y6cp5c/1 [42].

REFERENCES

[1] C. Stewart, "Artificial intelligence (AI) in healthcare market size worldwide from 2021 to 2030," [Online]. Available: https://www.statista.com/statistics/1334826/ai-in-healthcare-market-size-worldwide/. [Accessed: 03-Dec-2023].
[2] D. Technologies, "Internet of Things and data placement," [Online]. Available: https://infohub.delltechnologies.com/l/edge-to-core-and-the-internet-of-things-2/internet-of-things-and-data-placement/#:~:text=The Internet of Things (IoT,zettabytes (ZB) of data. [Accessed: 03-Dec-2023].
[3] Michael.Walker, "Internet of things and data science transforms life, business and government," [Online]. Available: https://www.datascienceassn.org/content/internet-things-and-data-science-transforms-life-business-and-government. [Accessed: 03-Dec-2023].
[4] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018, doi: 10.1016/j.jksuci.2017.06.001.
[5] D. Che, M. Safran, and Z. Peng, "From big data to big data mining: challenges, issues, and opportunities BT - database systems for advanced applications," 2013, pp. 1–15.
[6] M. K. Dhaliwal, R. Sharma, and N. Bindra, "Role of internet of things (IoT) in preventing and controlling disease outbreak: a snapshot of existing scenario," in *Proceedings of the International Conference on Intelligent Computing, Communication and Information Security. ICICCIS 2022. Algorithms for Intelligent Systems.*, 2023, pp. 359–373.
[7] S. Sareen, S. K. Sood, and S. K. Gupta, "IoT-based cloud framework to control Ebola virus outbreak," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 3, pp. 459–476, 2018, doi: 10.1007/s12652-016-0427-7.

[8] M. Otoom, N. Otoum, M. A. Alzubaidi, Y. Etoom, and R. Banihani, "An IoT-based framework for early identification and monitoring of COVID-19 cases," *Biomedical Signal Processing and Control*, vol. 62, p. 102149, Sep. 2020, doi: 10.1016/J.BSPC.2020.102149.
[9] P. Thakur and S. Kaur, "An intelligent system for predicting and preventing Chikungunya virus," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017*, Aug. 2018, pp. 3483–3492, doi: 10.1109/ICECDS.2017.8390109.
[10] S. K. Sood and I. Mahajan, "Wearable IoT sensor based healthcare system for identifying and controlling chikungunya virus," *Computers in Industry*, vol. 91, pp. 33–44, 2017, doi: 10.1016/j.compind.2017.05.006.
[11] B. Mahalakshmi and G. Suseendran, "Zika virus: A secure system usinf nbn classifier for predicting and preventing zika in cloud," *International Journal of Recent Technology and Engineering*, vol. 7, no. 4, pp. 28–32, 2018.
[12] S. Sareen, S. K. Sood, and S. K. Gupta, "Secure internet of things-based cloud framework to control zika virus outbreak," *International Journal of Technology Assessment in Health Care*, vol. 33, no. 1, pp. 11–18, 2017, doi: 10.1017/S0266462317000113.
[13] S. Sareen, S. K. Gupta, and S. K. Sood, "An intelligent and secure system for predicting and preventing Zika virus outbreak using Fog computing," *Enterprise Information Systems*, vol. 11, no. 9, pp. 1436–1456, 2017, doi: 10.1080/17517575.2016.1277558.
[14] R. Sandhu, H. K. Gill, and S. K. Sood, "Smart monitoring and controlling of Pandemic Influenza A (H1N1) using social network analysis and cloud computing," *Journal of Computational Science*, vol. 12, pp. 11–22, 2016, doi: 10.1016/j.jocs.2015.11.001.
[15] D. Ayata, Y. Yaslan, and M. Kamasak, "Emotion recognition via galvanic skin response: Comparison of machine learning algorithms and feature extraction methods," *Istanbul University - Journal of Electrical and Electronics Engineering*, vol. 17, no. 1, pp. 3129–3136, 2017.
[16] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2019, doi: 10.1109/ACCESS.2018.2883213.
[17] N. Komuro, T. Hashiguchi, K. Hirai, and M. Ichikawa, "Predicting individual emotion from perception-based non-contact sensor big data," *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021, doi: 10.1038/s41598-021-81958-2.
[18] K. Sen and S. Pal, "Alternative method for pain assessment using EMG and GSR," *Journal of Mechanics in Medicine and Biology*, vol. 21, no. 6, pp. 1–24, 2021, doi: 10.1142/S0219519421500391.
[19] T. H. Chueh, T. B. Chen, H. H. S. Lu, S. S. Ju, T. H. Tao, and J. H. Shaw, "Statistical prediction of emotional states by physiological signals with manova and machine learning," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 4, pp. 1–18, 2012, doi: 10.1142/S0218001412500085.
[20] J. Kim and E. Andr´, "Detection of affective patterns in physiological signals towards improving automatic emotion recognition," *Handbook of Pattern Recognition and Computer Vision, Fourth Edition*, pp. 415–434, 2009, doi: 10.1142/9789814273398_018.
[21] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on content analysis and physiological changes," *International Journal of Semantic Computing*, vol. 3, no. 2, pp. 235–254, 2009, doi: 10.1142/S1793351X09000744.
[22] R. M. Alisha, P. Vijayalakshmi, A. Jatti, M. Kannan, and S. Sinha, "Design and development of an IOT based wearable device for the safety and security of women and girl children," in *2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings*, May 2017, pp. 1108–1112, doi: 10.1109/RTEICT.2016.7808003.
[23] A. A. H. Mohamad, N. K. Jumaa, and S. H. Majeed, "Thingspeak cloud computing platform based ECG diagnose system," vol. 1, no. 1, 2019.
[24] S. S. Thomas, A. Saraswat, A. Shashwat, and V. Bharti, "Sensing heart beat and body temperature digitally using Arduino," *International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings*, pp. 1721–1724, 2017, doi: 10.1109/SCOPES.2016.7955737.
[25] Ramesh Saha, S. Biswas, S. Sarmah, S. Karmakar, and P. Das, "A working prototype using DS18B20 temperature sensor and arduino for health monitoring," *SN Computer Science*, vol. 2, no. 1, pp. 1–21, 2021, doi: 10.1007/s42979-020-00434-2.
[26] R. Anandh and G. Indirani, "Real time health monitoring system using arduino with cloud technology," *Asian Journal of Computer Science and Technology*, vol. 7, no. S1, pp. 29–32, 2018, doi: 10.51983/ajcst-2018.7.s1.1810.
[27] A. A. Rahimoon, M. N. Abdullah, and I. Taib, "Design of a contactless body temperature measurement system using Arduino," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 3, pp. 1251–1258, 2020, doi: 10.11591/ijeecs.v19.i3.pp1251-1258.
[28] A. Nduka, J. Samual, S. Elango, S. Divakaran, U. Umar, and R. Senthilprabha, "Internet of things based remote health monitoring system using arduino," *Proceedings of the 3rd International Conference on I-SMAC IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2019*, pp. 572–576, 2019, doi: 10.1109/I-SMAC47947.2019.9032438.
[29] A. K. Jameil and H. Al-Raweshidy, "A digital twin framework for real-time healthcare monitoring: leveraging AI and secure systems for enhanced patient outcomes," *Discover Internet of Things*, vol. 5, no. 1, p. 37, Apr. 2025, doi: 10.1007/s43926-025-00135-3.
[30] N. J. Papri, A. Ahmed, and A. Chowdhury, "IoT and cloud-based non-invasive diabetes detection system from photoplethysmogram," *Discover Internet of Things*, vol. 5, no. 1, p. 57, May 2025, doi: 10.1007/s43926-025-00158-w.
[31] K. Babu, A. G. Chandar, and S. Kannadhasan, "Prediction and diagnosis of cardiovascular disease using cloud and machine learning design," *Journal of Cloud Computing*, vol. 14, no. 1, p. 3, Jan. 2025, doi: 10.1186/s13677-024-00720-x.
[32] S. Hamzehei, O. Akbarzadeh, H. Attar, K. Rezaee, N. Fasihihour, and M. R. Khosravi, "Predicting the total unified parkinson's disease rating scale (UPDRS) based on ML techniques and cloud-based update," *Journal of Cloud Computing*, vol. 12, no. 1, pp. 1–16, 2023, doi: 10.1186/s13677-022-00388-1.
[33] H. Wang *et al.*, "HealthAIoT: AIoT-driven smart healthcare system for sustainable cloud computing environments," *Internet of Things*, vol. 31, no. December 2024, p. 101555, May 2025, doi: 10.1016/j.iot.2025.101555.
[34] A. Swain, A. Dalei, and S. Sahoo, "Emotional well-being of a human: a case study of proposing a digital twin solution integrating IoT sensors and cloud computing," in *2025 International Conference on Ambient Intelligence in Health Care (ICAIHC)*, Jan. 2025, pp. 1–7, doi: 10.1109/ICAIHC64101.2025.10956277.
[35] E. Systems, "ESP8266EX datasheet version 6.0," 2020.
[36] "ThingSpeak for IoT," [Online]. Available: https://thingspeak.com/.
[37] M. Roccetti, G. Delnevo, L. Casini, and P. Salomoni, "A cautionary tale for machine learning design: why we still need human-assisted big data analysis," *Mobile Networks and Applications*, vol. 25, no. 3, pp. 1075-1083, 2020, doi: 10.1007/s11036-020-01530-6.
[38] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal*, vol. 14, no. 0, p. 2, May 2015, doi: 10.5334/dsj-2015-002.

[39] H. Chen, D. Hailey, N. Wang, and P. Yu, "A review of data quality assessment methods for public health information systems.," *International journal of environmental research and public health*, vol. 11, no. 5, pp. 5170–207, May 2014, doi: 10.3390/ijerph110505170.
[40] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, Oct. 2004, doi: 10.1007/s10462-004-4304-y.
[41] K. Singh and S. Upadhyaya, "Outlier detection: applications and techniques.," *International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 307–323, 2012.
[42] M. K. Dhaliwal, R. Sharma, and R. Kaur, "PIF dataset: a comprehensive dataset with physiological and inertial features of the human body," *Mendeley Data*, 2023.
[43] M. K. Dhaliwal, R. Sharma, and R. Kaur, "PIF dataset: a comprehensive dataset of physiological and inertial features for recognition of human activities," *Multimedia Tools and Applications*, vol. 83, no. 29, pp. 73607–73625, 2024, doi: 10.1007/s11042-024-19285-7.

BIOGRAPHIES OF AUTHORS

Manpreet Kaur Dhaliwal was graduated from Panjab University, Chandigarh, India. She received her Masters of Computer Applications from I.K. Gujral Punjab Technical University, Jalandhar, India. She is currently working as an assistant professor in the Department of Computer Science and Applications at Shree Atam Vallabh Jain College Ludhiana, Punjab. Her research interest includes internet of things (IoT), IoMT, cloud computing, machine learning and big data. She has published 6 research articles in reputed SCI and Scopus indexed journals and conferences. She can be contacted at email: preet2016@pu.ac.in.

Rohini Sharma holds Ph.D. in Computer Science from Panjab University Chandigarh India. She is currently working as an Associate Professor in the Department of Computer Science and Applications, Panjab University, Chandigarh, India since September, 2011. Her research interests include anomaly detection in networks, network traffic data analysis, internet of things, health analytics and natural language processing. She has published 20 research articles in reputed SCI and Scopus indexed journals and conferences. She can be contacted at email: rohini@pu.ac.in.

Rajbinder Kaur was graduated and post graduated from Panjab University, Chandigarh, India. Her research interest includes internet of things (IoT), IoMT, cloud computing, machine learning and big data. She has published 4 research articles in reputed SCI and Scopus indexed journals and conferences. She can be contacted at email: raj1995@pu.ac.in.