VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belgaum-590018

A PROJECT REPORT **(15CSP85)** ON

"Monitoring System Using Multimedia for Smart Healthcare"

Submitted in Partial fulfillment of the Requirements for the VIII Semester of the Degree of

Bachelor of Engineering in Computer Science & Engineering

By

JAILEKHA C (1CR16CS059)

T SRUJANA (1CR16CS171)

GOPINATH G(1CR16CS408)

NITHIN S A (1CR16CS418)

Under the Guidance of,

Mrs. S Aarthi

Assistant Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **"Monitoring System Using Multimedia for Smart Healthcare"** carried out by **Ms**. **JAILEKHA C** bearing USN **1CR16CS059;Ms**. **T SRUJANA** bearing USN **1CR16CS171;Mr**. **GOPINATH G** bearing USN **1CR16CS408; Mr**. **NITHIN S A** bearing USN **1CR16C418,** bonafide students of CMR Institute of Technology, in partial fulfillment for the award of Bachelor **of Engineering** in **Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the **year 2019-2020**. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library.

The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

________________	________________	________________
Mrs. S Aarthi	**Dr. Prem Kumar Ramesh**	**Dr. Sanjay Jain**
Assistant Professor	**Professor & Head**	**Principal**
Dept. of CSE, CMRIT	**Dept. of CSE, CMRIT**	**CMRIT**

External Viva

Name of the examiners — Signature with date

1. ________________

2. ________________

DECLARATION

We, the students of 8th semester of Computer Science and Engineering, CMR Institute of Technology, Bangalore declare that the work entitled **“Monitoring System Using Multimedia for Smart Healthcare”**has been successfully completed under the guidance of Mrs. S. Aarthi, Computer Science and Engineering Department, CMR Institute of technology, Bangalore. This dissertation work is submitted in partial fulfillment of the requirements for the award of Degree of Bachelor of Engineering in Computer Science and Engineering during the academic year 2019 - 2020. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to anyuniversity.

Place:

Date:

Team members:

JAILEKHA C (1CR16CS059) ____________________

T SRUJANA (1CR16CS171) ____________________

GOPINATH G (1CR16CS408) ____________________

NITHIN S A (1CR16CS418) ____________________

ABSTRACT

The use of multimodal inputs in a smart healthcare framework is promising due to the increase in accuracy of the systems involved in the framework. In this study, we propose a user satisfaction detection system using two multimedia contents, namely, speech and image. The three classes of satisfaction are satisfied, not satisfied, and indifferent. In the proposed system, speech and facial image of the user are captured, transmitted to a cloud, and then analyzed. A decision on the satisfaction is then delivered to the appropriate stakeholders. Several features from these two inputs are extracted from the cloud. For speech, directional derivatives of a spectrogram are used as features, whereas for image, a local binary pattern of the image is used to extract features. These features are combined and input to Random forest and Extra trees. It is shown that the proposed system achieves up to 84% accuracy in detecting satisfaction.

ACKNOWLEDGEMENT

We take this opportunity to express our sincere gratitude and respect to **CMR Institute of Technology, Bengaluru** for providing us a platform to pursue our studies and carry out our final year project

We have a great pleasure in expressing our deep sense of gratitude to **Dr. Sanjay Jain,** Principal, CMRIT, Bangalore, for his constant encouragement.

We would like to thank **Dr. Prem Kumar Ramesh,** HOD, Department of Computer Science and Engineering, CMRIT, Bangalore, who has been a constant support and encouragement throughout the course of this project.

We consider it a privilege and honor to express our sincere gratitude to our guide **Mrs. S Aarthi, Assistant professor,** Department of Computer Science and Engineering, for the valuable guidance throughout the tenure of this review.

We also extend our thanks to all the faculty of Computer Science and Engineering who directly or indirectly encouraged us.

Finally, we would like to thank our family and friends for all their moral support they have given us during the completion of this work.

JAILEKHA.C (1CR16CS059)
T.SRUJANA (1CR16CS171)
GOPINATH.G (1CR16CS408)
NITHIN.S.A (1CR16CS418)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	MEANING
CS-LBP	Center Asymmetric Local Binary Pattern
FDR	False Discovery Rate
GMM	Gaussian Mixture Model
LBP	Local Binary Pattern
PCA	Principle Component Analysis
SVM	Support Vector Machine
WLD	Weber Local Descriptor

CHAPTER 1

INTRODUCTION

With the invention of low-cost processing and storage, several smart solutions are gaining attraction in improving the quality of human life. Particularly, smart healthcare is in great demand because of the increase in population and decrease in doctor-to-people ratio, and some people become busy to travel to a specialized hospital for treatment.

The smart healthcare business is estimated to be more than several billion dollars in next few years. A successful smart healthcare framework requires several parameters, including ease of use of the medical sensors, low cost, high accuracy, ubiquitous nature of the framework, and less delay in making decision. These parameters may not be achieved in a single framework, although efforts have been made for the last several years.

The ease of use of sensors depends on their invasiveness; low-cost depends on the complexity of the devices in acquiring signals and installation; high accuracy depends on the precision of the sensors and the algorithms embedded in the software; and delay depends on the number of features of the signals.

Numerous smart healthcare frameworks have been proposed in the literature from different perspectives. Some frameworks attempt to solve the problems in diabetes, smart cities, voice pathology assessment, emotion recognition, and patient’s state recognition. A software-defined network was proposed to improve the performance of smart healthcare frameworks.

Customer satisfaction (of users and patients) is an important goal for smart healthcare business. A service provider can obtain feedback on customer satisfaction by using a survey conducted electronically or paper-based. The issue in conducting this type of survey is that sometimes the users do not want to participate.

An automatic satisfaction reading from the face, speech, or gesture of the users can greatly solve this problem. In this study, we propose a user satisfaction detection system as part of a smart healthcare framework. A multimedia-based technique is utilized to capture the signals from the users. Speech and image are the two signals captured for the accuracy of the proposed system. These signals are processed in a cloud server. Cloud manager then sends the result to the stakeholder.

A smart home is equipped with multimedia sensors that can capture different signals.

These signals come from the expressions of the user. Subsequently, the signals are transmitted to the cloud for processing. The result is then sent to the hospital, doctors, and caregivers, who analyze the satisfaction result for future quality improvement .

1.1 Relevance of the Project

A user satisfaction detection system using speech and image for a smart healthcare framework was proposed. For the speech signal, we used the directional derivative features from the Mel spectrogram, whereas for the image signal, we used the LBP features. SVM was used as the classifier, and several experiments were performed.

The best accuracy was obtained by combining the features from the speech and image signals.

Table 1.1 Summary of the Approaches

Fourier Transform	In theserver. Thespeechsignalsframed and is Calculated Using Hamming window.	The frame length is 40 ms, and the frame shift is 20 ms. Each windowed frame is transformed into a frequency domainrepresentation (spectrum)using Fourier transform,	Such that the time domainsignal is converted into a frequencydomain signal.
Band-Pass Filters	Twenty- four band-pass filters are passed through the frequency domainsignal tomimic The hearing perception ofthe user.	The Centre frequencies of the filters are distributed on a Mel scale. The bandwidths of the filters correspond to the critical bandwidth	

DirectionalDerivative	The Mel spectrogram Is passes through directional derivatives in four directions toobtainthe relative progress of the signal along four directions,	That is, 0, 45, 90, and 135, whichcorrespond to time, increasing time frequency, and decreasing\time frequency, respectively.	
Face	Thiskey frame is termed as the imagesignals.A face detection algorithm Extracts the face areaof the image signal.	This process is performed in the local processor to decrease the Transmission cost of the video.	
LBP	An LBP is applied to the face image once it is transmitted to the cloud server to obtain An LBP images.		
SVM	An SVM based classifier is used in the cloud server for Classification.	The main idea of the SVM is to maximize the distance of a linear separator from two classes of samples.	Polynomial kernel and a radial basis function (RBF) kernel are the two most common Kernels used in many applications.

1.2 Problem Statement

To understand the mental status of the user with the help of a combination of speech and visual signals. The use of multiple modalities helps in getting rid of any ambiguities present in the recognition.The issue in conducting this type of survey is that sometimes the users do not want to participate. An automatic satisfaction reading from the face, speech, or gesture of the users can greatly solve this problem.In the previous system speech and video were processed separately, but we are proposing a system that combines both speech and video and gives the output.

1.3 Objectives

The use of multimodal inputs in a smart healthcare framework is promising due to the increase inaccuracy of the systems involved in the framework. In this study, we propose a user satisfaction detection system using two multimedia contents, namely, speech and image. The three classes of satisfaction are satisfied, not satisfied, and indifferent. In the proposed system, speech and facial image of the user are captured, transmitted to a cloud, and then analyzed. A decision on the satisfaction is then delivered to the appropriate stakeholders. Several features from these two inputs are extracted from the cloud. For speech, directional derivatives of a spectrogram are used as features, whereas for image, a local binary pattern of the image is used to extract features. These features are combined and input to a support vector machine-based classifier. It is shown that the proposed system achieves up to 80% accuracy in detecting satisfaction.

1.4 Scope of the Project

In future works, we intend to use highly sophisticated classifying approach, such as active learning, which has been successfully used in emotion recognition. In MPEG-7 audio features were effectively used in an audio–visual emotion recognition. We may use such features and include other input modalities to enhance the accuracy of the proposed system.We might try to create a front-end device and deploy it on. If this becomes a success, we can use it directly from home and monitor the feedback cloud.

1.5 Methodology

Table 2.2 Agile Methodology

Story ID	Requirement description	User stories/Task	Description
Requirement	Gathering projectIdeas To work on.	Find if there already exist Implementations Of the chosen project idea. Ifimplementation exists, studythe existing Implementation. Based on the study, arrive at the Missingnecessary Featureswe can build.	Collect data and ideas from various sources to find a potential project.
Planning	Prepare feature list. Technologiesto be used to estimate effort.	Decide on the important Functionalities we can add to our implementation.	Planning is process which embraces a no of steps to be taken.
Development	High level API class design. Cloud based machines,AWS, EC2 services, Google Collab.	Decide and arrive at an overview of modules and classes. Using Collab we made the code much more readable.	Coding basic python, Sci-Kit learning library Was majorly usedand also an instance of AWS.

Test Cases	Accuracy of random forest and extra tress are compared with the existing system i.e.SVM. Also, comparable with other algorithms with respect to accuracy.	Implement different features for the UI and make sure they are working properly.	Information and researchon the algorithms Used forclustering and classification of data

CHAPTER 2

LITERATURE SURVEY

1. Emotion-Aware Connected Healthcare Big Data towards5G.

The following ideas are taken from the above paper:

An emotion-aware connected healthcare system using a powerful emotion detection module. Different IoT devices are used to capture speech and image signals of a patient in a smart home scenario.

Speech and image signals are processed separately, and classification scores using these signals are fused to produce a final score to take a decision about the emotion.

The system can detect an emotion of a patient by using two input modalities: speech and image. The objective of the system is to detect whether the patient expresses a painful emotion or not. If the system detects that the patient feels pain, the framework notifies about this situation to the concerned entities such as registered medical doctors, caregivers, hospitals, and health centers.

A fast Fourier transform is applied to the frames to convert the signal from a time-domain to a frequency-domain signal. 24 band-pass filters (BPFs) divide the spectrum (frequency-domain signal) into 24 frequency bands. The Centre frequencies of the BPFs are spaced on a Mel scale, and the bandwidths of the filters follow a critical bandwidth of human auditory perception. After this process, we get a Mel-spectrogram, which can be viewed as an image.

We apply a powerful local texture descriptor in the form of the LBP.

The original LBP is calculated in a 3 × 3 window. If the grey-scale intensity level of a neighboring pixel is greater than that of the Center pixel of the rectangular window, a ‘1’ is assigned to the location of that neighboring pixel; otherwise, a ‘0’ is assigned to that location. The sequence of ‘1’ and ‘0’ is collected clockwise or anti-clockwise to form an 8-bit binary number, which is then converted into a denary number.

In our proposed emotion recognition system, we utilize a Centre-symmetric LBP (CSLBP). In the CS-LBP, instead of comparing the pixel intensities between a neighboring pixel and the Centre pixel, they are compared between two pixels located symmetrically around the Centre pixel.

The feature set is fed to a support vector machine (SVM) based classifier. The SVM is a powerful binary classifier, which tries to find an optimal separator between the samples of two classes. As it is very difficult to separate the samples of two classes by a line, often the samples are projected to a high dimensional space by a kernel so that they can be separated by a hyperplane. In our proposed system, we use a radial basis function (RBF) kernel to project the features in a high dimensional space.

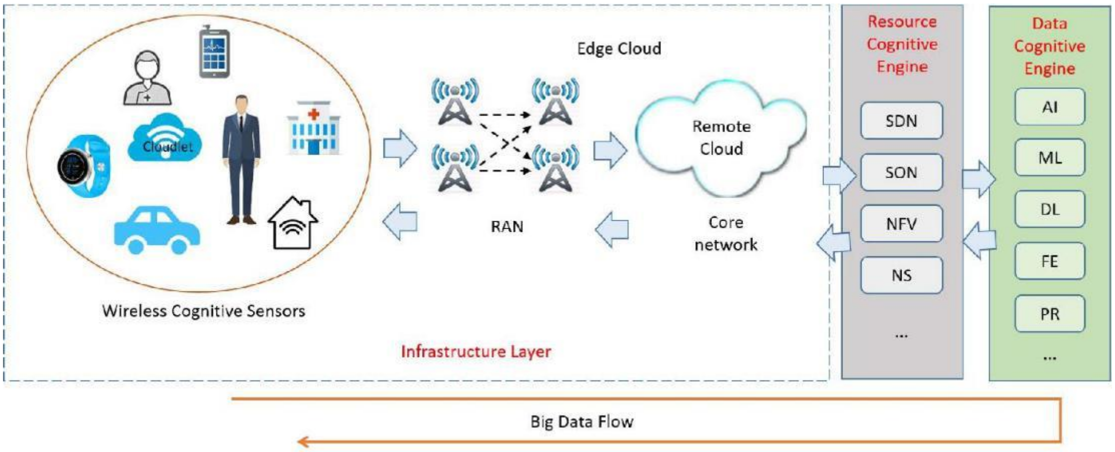


Fig2.1 5G enabled emotion aware connected healthcare bigdata frame work.

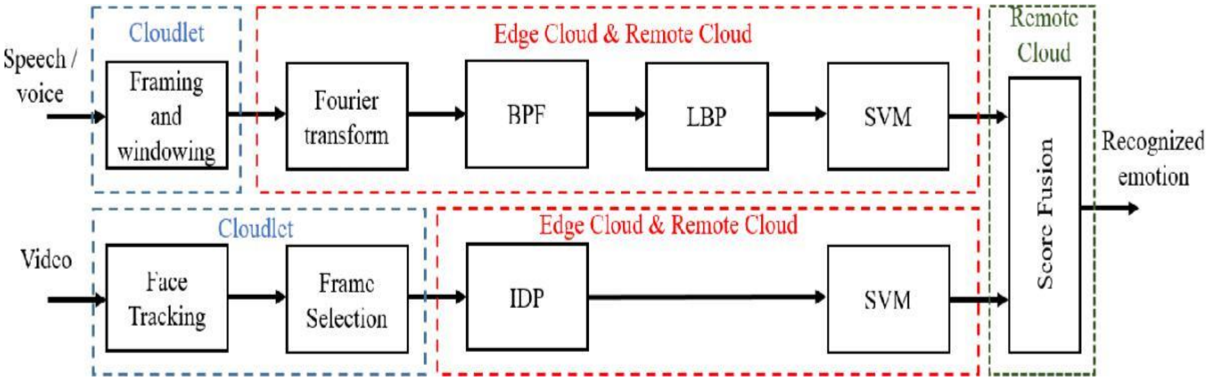


Fig2.2 Emotion recognition system.

2. A Facial-Expression Monitoring System for Improved Healthcare in smartCities.

The following ideas are taken from the above paper:

Human facial expressions change with different states of health; therefore, a facial-expression recognition system can be beneficial to a healthcare framework. In this paper, a facial-expression recognition system is proposed to improve the service of the healthcare in a smart city.

We use two classifiers: a Gaussian mixture model (GMM) and a support vector machine (SVM).

The facial-expression system proposed in this paper differs from those reported earlier in the sense that the proposed system is specifically designed for a healthcare framework in Smart Cities. The contributions of our present work are as follows: (i) a bandlet transform and a local binary pattern (LBP) were used to extract features from facial images; (ii) the spatial information of facial expressions is preserved by using a block-based, center-symmetric LBP (CS-LBP); and (iii) the scores of two classifiers, namely the Gaussian mixture model (GMM) and the support vector machine (SVM), are fused with a confidence score.

A smart video or a smart camera constantly takes images of the patient in the smart home.Facial expressions have many types of geometrical structures, which are very important to recognize a definite expression. In traditional wavelet transforms, these geometrical structures cannot be properly encoded.

In the bandlet transform, the geometrical structures are represented by some orthogonal bandlet bases. To accurately represent the geometric flow, the image is divided into small blocks, where a block can contain only one contour. Normally, smaller blocks can capture the geometrical flows more accurately than can larger blocks. First, the image is decomposed into sub-bands of different scales using wavelet bases, and then the wavelet bases are replaced by the orthogonal bandlet bases.

The next step is to divide each bandlet sub-band image into blocks. The CS-LBP is applied to each block of the sub-band. The LBP is a powerful yet efficient texture descriptor that has been applied to many image-processing applications; however, the length of the LBP histogram is very long. Also, the LBP is not robust against noise. To avoid these issues, the CS-LBP was proposed, in which the center-symmetric pixels are compared based on their grayscale intensities.

In the proposed system, two classifiers are used: the GMM and the SVM. The GMM is a stochastic method of modeling, frequently used in multiclass problems including speech/speaker recognition, emotion recognition, and environment recognition. The SVM is a powerful binary classifier that is also used in many image-processing image- processing applications. In the proposed system, we take the advantages of both classifiers by combining their likelihood scores using a weight coefficient.

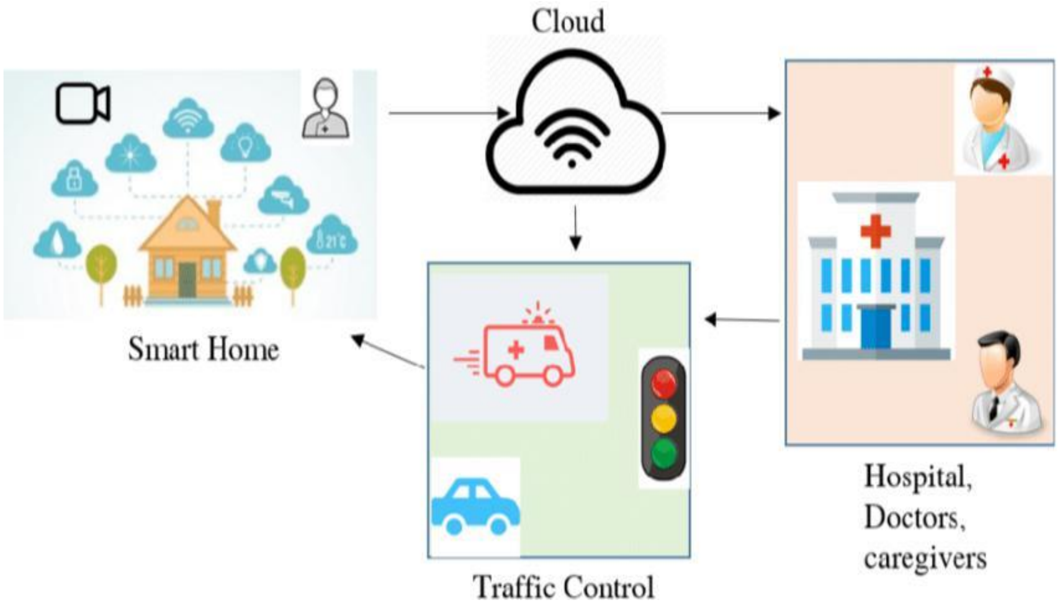


Fig 2.3 A framework of a smart city for a smart health care.

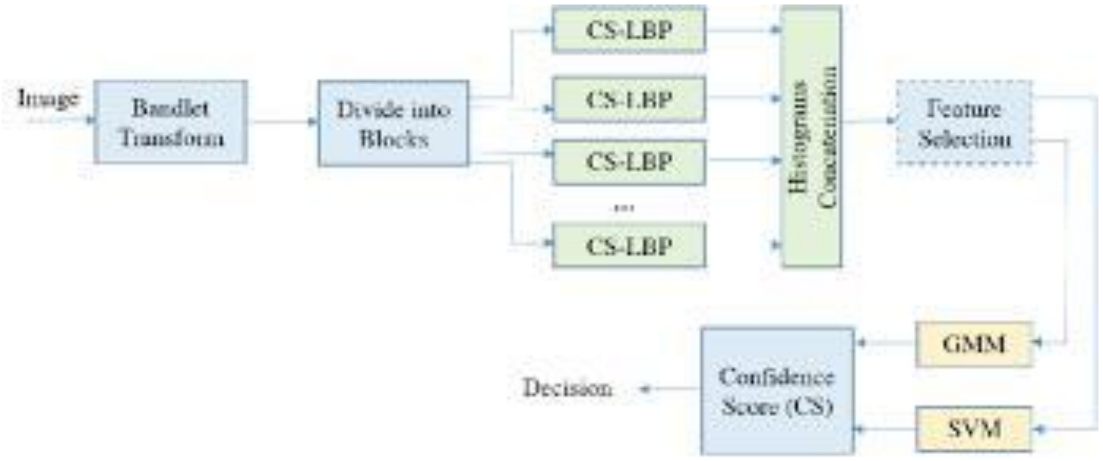


Fig2.4 Block diagram of the proposed facial expression recognition system.

3. Automatic facial emotion recognition using weber local descriptor For e-Healthcare system.

The following ideas are taken from the above paper:

For automatic face recognition WEB LOCAL DESCRIPTORS are used.

A static facial image is subdivided into many blocks. A multiscale WLD is applied to each of the blocks to obtain a WLD histogram for the image. The face descriptors are then input to a support vector machine classifies to recognition the emotion.

In the proposed Healthcare framework, a patient or a registered user sends his or her medical data through smart devices. The smart device also takes facial images of the user. The face images along with the medical data are transmitted to a cloud for further processing. In the cloud, a cloud manager first authenticates the user, and then sends the face images data to different servers for emotion detection.

The servers can be feature extraction server, classification server, etc. Once the emotion is detected, the cloud manager sends the emotion information to appropriate healthcare professionals. Input image is cropped to remove unnecessary regions (for example, hair, ears) of a face. The cropped face image is fed to the system as an input.

Multi-scale WLD is applied to the blocks of the face image, and a WLD histogram is obtained. After applying FDR, the optimum bins are selected as features. WLD is a powerful texture descriptor proposed it is grounded on Weber's law, which states that any noticeable changes is constantly relational to the background.

Ifp is the amount of change in illumination and p is the original background illumination, then their ratio is a constant value, say *C*, as stated. WLD is less sensitive to noise and illumination changes. WLD has two constituents, which are differential excitation (DE) and gradient orientation (GO).

In DE, the change between the Centre pixel, pc, of a neighborhood and the neighboring pixels, pi, is calculated T is the number of quantized orientations, and t∈ $[0, T-1]$. All T sub-histograms for a particular m sub-histogram are calculated to form a histogram Hm. The final feature set is attained by fusing the sub-histograms Hm.

In the proposed system, WLD histograms are calculated for each block of the face image and then concatenated to construct the final WLD histogram of the whole cropped image. FDR is applied to select the most significant bins from the WLD histogram.

FDR calculates the ratio between the square of the mean difference and the variance difference between two classes. If μ1i and μ2i are the average values of classes 1 and 2, respectively, for a particular bin *i* and their corresponding variances are σ1i and σ2i, the fisher ratio for this bin is calculated by using.

A high value of FDR indicates that the corresponding bin is a discriminative bin between the classes and thereby can be chosen for subsequent classifier. The features are arranged in descending order based on the FDR value. The first N features are chosen in the process SVM is a widely used powerful classifier.

It maps a lower dimensional data to a higher dimension using a kernel function, and searches for an optimal hyper plane that maximizes the margin between the samples of two classes. In the experiments, we investigated three different kernels, which are linear kernel, polynomial kernel, and radial basis function (RBF) kernel.

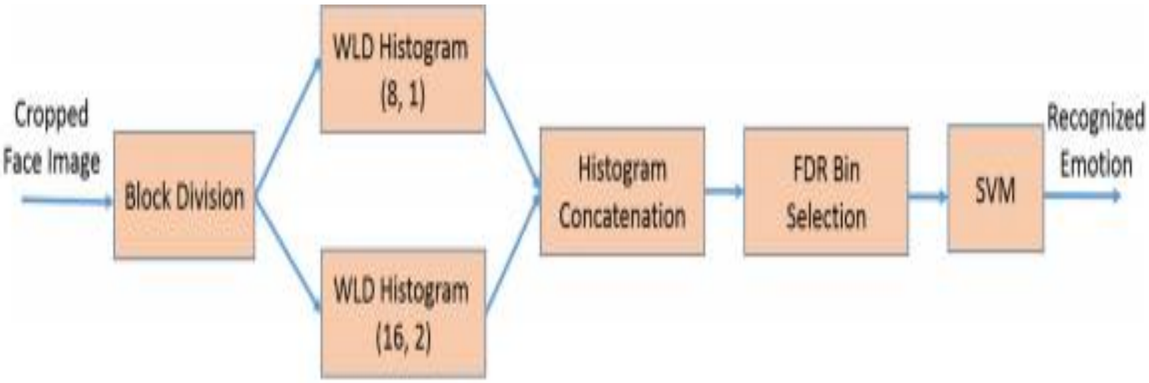


Fig2.5 Block diagram of the proposed face emotion recognition.

4. A New Hybrid PSO Assisted Biogeography-Based Optimization for Emotion and Stress Recognition from Speech Signal.

The following ideas are taken from the above paper:

To identify different emotional states and multi-style of speech from speech signals, their salient features need to be extracted. Based on the energy present in the frame, the unvoiced portions (the frame with less energy) had been removed before the feature extraction process. The frame with lesser energy is removed by setting a threshold value. The threshold value is determined for each database separately.

Then, the first order pre-emphasis filter was used to spectrally flatten the speech waveforms. To identify different emotional states and multi-style of speech from speech signals, their salient features need to be extracted. All speech signals were down sampled to 8 kHz since at the recorded signals of the database have different sampling rates.

The speech signals were segmented into non overlapping frames with 256 samples (32 ms). Based on the energy present in the frame, the unvoiced portions (the frame with less energy) had been removed before the feature extraction process. The frame with lesser energy is removed by setting a threshold value. The threshold value is determined for each database separately.

The remaining voiced frames were concatenated and glottal waveforms were extracted by applying inverse filtering and linear predictive analysis method. Then, the first order pre-emphasis filter was used to spectrally flatten the speech waveforms and glottal waveforms. The filtered signals were segmented into frames with an overlap of50%.

Later, each frame was windowed by applying hamming window technique which reduced the signal discontinuity and spectral distortion. Bispectral and Bicoherence features for each frame were extracted and they were averaged for all frames. The spectral representation of higher order cumulants of a random process is defined as Higher Order Spectra.

The third order cumulant spectra are called bispectrum or bi-spectral. The bi-spectrum is the 2D – Fourier transform of the third order cumulant function. The bi-spectrum is a function of two frequencies unlike the power spectrum which is a function of only one frequency variable.

The normalized bi-spectrum is called bicoherence representation of the signal. These features were derived from each frame. The number of voiced portions varies for each speech signal, since the recording duration of speech signal varies.

The features from each frame were extracted first and were averaged for overall features from all the frames. Biogeography based optimization is an algorithm based on geographical distribution of a group of biological organisms in its isolated environment. Organism in BBO is called species and these species can migrate from one island to other and this migration is called habitat.

Each habitat has a Habitat Suitability Index (HSI) which is similar to the fitness in general optimization algorithms. Suitability Index Variable (SIV) suggests the habitability of the habitant. The habitatwith good HSI will move to other island in order to create a good population for the next generation. The emigration (I) and the immigration (I) of the habitats are controlled by the fitness.

Particle Swarm Optimization (PSO) is a population based stochastic optimization technique proposed. This technique emerged from the behavior of non-guided animals in a group or swarm as in bird flocking and fish schooling. Each individual in a population or a solution in PSO has its own velocity and position.

In a n-dimensional search area, each particle flies to its best position depending on 1) *pBest*, the best solution achieved so far; 2) *gBest*, global best value. The algorithm updates its velocity and position when the best solution is reached. PSO has recently been applied for improving emotion recognition in speech and glottal signals.

Further, the PSO is employed to optimize communication networks, engines, motors, entertainment and metallurgy applications. The proposed PSOBBO enhances the basic BBO. The basic BBO is good at exploiting current population information with its migration operation, but it is slow in exploring the global search space.

In order to improve its exploring capacity, a modified PS velocity and position update of the particles are incorporated and are applied for worst half of the population. Further, the proposed method helps in increasing the diversity among the population.

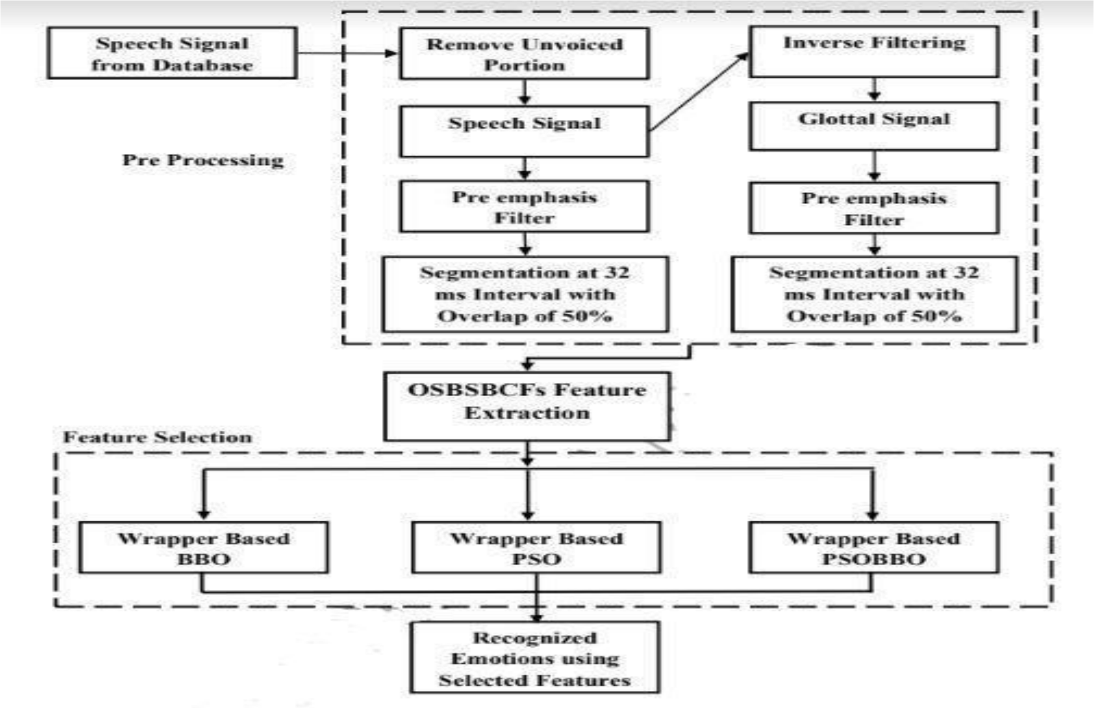

Speech Signal from Database
Pre Processing
Remove Unvoiced Portion
Speech Signal
Pre emphasis Filter
Segmentation at 32 ms Interval with Overlap of 50%
Inverse Filtering
Glottal Signal
Pre emphasis Filter
Segmentation at 32 ms Interval with Overlap of 50%
OSBSBCFs Feature Extraction
Feature Selection
Wrapper Based BBO
Wrapper Based PSO
Wrapper Based PSOBBO
Recognized Emotions using Selected Features

Fig2.6 Block diagram of the proposed speech recognition system.

CHAPTER 3

SYSTEMREQUIREMENTS SPECIFICATION

3.1 Functional Requirements

In software engineering, a functional requirement defines a function of the software system or its components. A function is described as a set of inputs, the behavior and outputs. Functional requirements maybe calculations, technical details, data manipulation and processing and other specific functionality that define what a system must accomplish. A functional requirement defines a function of a software system or its components. It captures the intended behavior of the system. This behavior may be expressed in terms of services, tasks or functions that the system has to perform.

3.2Non-Functional Requirements

Non-Functional Requirements specify the criteria that can be used to judge the operation of a system, rather than specific behaviors. They are the metrics that are considered to measure the performance of the developed system.

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. Three key considerations involved in the feasibility analysis are

Usability: Simple is the key here. The system must be simple thatpeople like to use it, but not so complex that people avoid using it. The user must be familiar with the user interfaces and should not have problems in migrating to a new system with a new environment. The menus, buttons and dialog boxes should be named in a manner that they provide clear understanding of the functionality.

Reliability: The system should be trustworthy and reliable inproviding the functionalities.

Performance: The system is going to be used by many peoplesimultaneously. Performance becomes a major concern. The system should not succumb when many users would be using it simultaneously. It should allow fast accessibility to all of its users

Scalability: The system should be scalable enough to add newfunctionalities at a later stage. There should be a common channel, which can accommodate the new functionalities

Maintainability: The system monitoring and maintenance shouldbe simple and objective in its approach.

Portability: The system should be easily portable to anothersystem.

Reusability: The system should be divided into such modules thatit could be used as a part of another system without requiring much of work.

Security: Security is a major concern. This system must not allowunauthorized users to access the information of other users.

3.3 HARDWARE REQUIREMENTS

The hardware requirements listed below are almost in a significantly higher level which represents the ideal situations to run the system. Following are the system hardware requirements used:

Processor	:	Intel(R) Core (TM) i3-8250U CPU @ 1.60GHz 1.80GHz
Speed	:	1.1 GHz
RAM	:	256 MB (min)
Hard Disk	:	20 GB
Key Board	:	Standard Keyboard
Mouse	:	Two or Three Button Mouse
System type	:	64-bit Operating System, x64-basedprocessor

3.4 SOFTWARE REQUIREMENTS

A major element in building a system is a section of compatible software since the software in the market is experiencing in geometric progression. Selected software should be acceptable by the firm and the user as well as it should be feasible for the system. This document gives the detailed description of the software requirements specification. The study of requirement specification is focused specially on the functioning of the system. It allows the analyst to understand the system, functions to be carried out and the performance level which has to be maintained including the interfaces established.

Operating System	- Windows
Back End	- ML Implementation
Google colaboratory	- MFCC implementation
Database	- Dataset(https://github.com/CheyneyComputerScience/CREMA-D)

CHAPTER 4

SYSTEM ANALYSIS AND DESIGN

4.1SYSTEM ARCHITECTURE

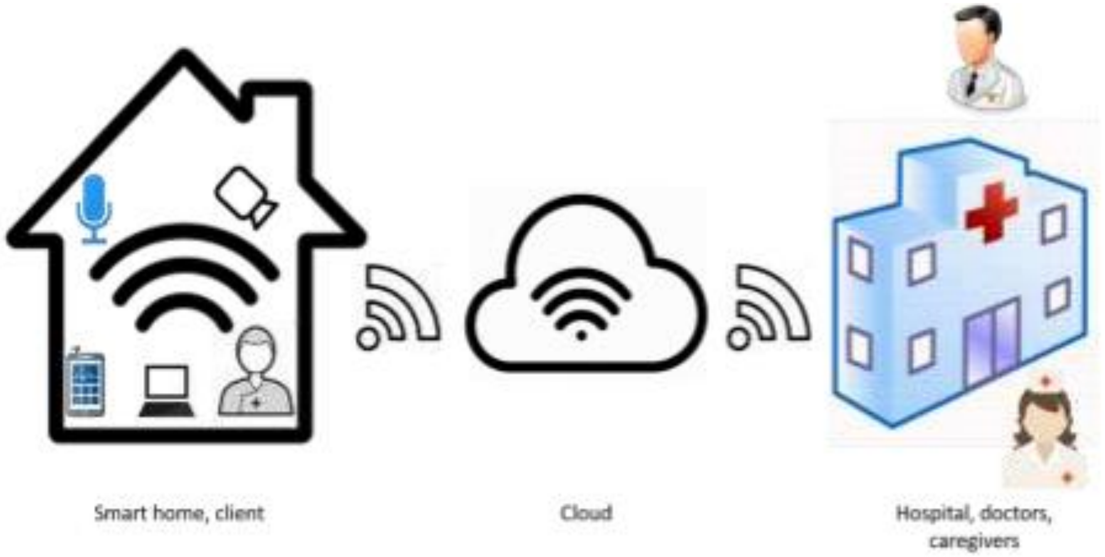


Fig4.1 Architecture of proposed system.

The emotion or human’s mental status can be recognized using speech only, image only, and their combination. An emotion recognition system using nonlinear features from speech was proposed. An optimal set of features was selected by a particle swarm optimization algorithm, which achieved 99.47% accuracy in the Emo-DB database. The same database was used in other works. For example, a support vector machine (SVM) with spectral and prosody features was used, which achieved 94.9% accuracy. A deep neural network was used several selected acoustic features were fed into the network, which obtained accuracy of 81.9%. A hidden Markov model-based classification was utilized, which achieved approximately 73% accuracy. Wavelet packet energy with entropy was used as the input to an extreme learning machine-based classifier, which obtained accuracy of 97.24%. Human facial expressions have been automatically recognized using images or videos in several studies. The most commonly used database in these works is the Cohn–Kanade (CK) database.

The most prominent features of images were wavelet and geometric features, and texture pattern. The accuracies of the systems varied between 94% and 97% using the CK database. Monitoring systems of patients' expressions were proposed a healthcare framework using big data of emotions and deep learning model was developed. Another framework for smart cities was introduced. A center symmetric local binary pattern (LBP) with bandlet transformation was also realized. A combination of speech and facial features were employed to monitor patients' status. A facial expression recognition for an e-healthcare system was proposed based on Weber local descriptor. The recognition accuracy reached up to 98%. A biologically-inspired multimedia management system was proposed.

4.2Existing System Architecture

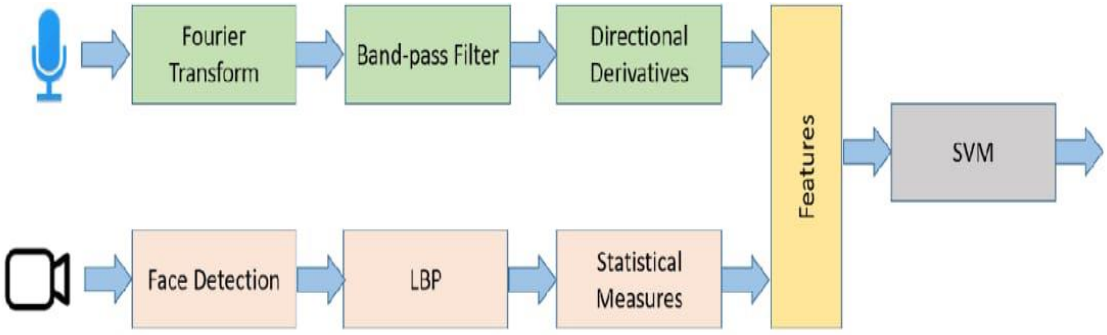


FIGURE 2. **Block diagram of the proposed user satisfaction detection system.**

Fig4.2 Existing system user satisfaction detection system.

Processing of speech signal:

1. **Mel-frequency cepstral coefficients** (MFCC): The speech signal is framed in 40ms window time frame.
2. Each windowed time frame is transformed into Frequency domain (spectrum representation).
3. 24 bandpass filters passed through frequency domain signals. The center frequencies of Bandpass filters are distributed in Mel scale.
4. The frequency domain signal passed through directional derivatives.
5. The derivatives used are a linear regression, where the window size is 3 frames before and 3 frames after the current frame.
6. These derivatives are useful for emotion prediction.
7. Traditionally it was used for detecting emotions of speech. In our case we combine these features with image processing to determine the emotion at higher accuracy.

Processing of image signal:

1. This key frame is termed as the image signal. A face detection algorithm extracts the face area of the image signal. This process is performed in the local processor to decrease the transmission cost of the video. An LBP is applied to the face image once it is transmitted to the cloud server to obtain an LBP image.
2. The extracted features are the average (mean), standard deviation, skewness, and kurtosis. These are well-known statistical features, which are successfully used in many applications.
3. The extracted features are the average (mean), standard deviation, skewness, and kurtosis. These are well-known statistical features, which are successfully used in many applications.

SVM Classifier:

1. An SVM-based classifier is used in the cloud server for classification.
2. The SVM is a simple powerful binary classifier, and it has successfully been applied in many signal processing applications such as speaker recognition, image classification, image forgery detection, and electrocardiogram signal classification.
3. The main idea of the SVM is to maximize the distance of a linear separator from two classes of samples.

4.3Proposed System Architecture

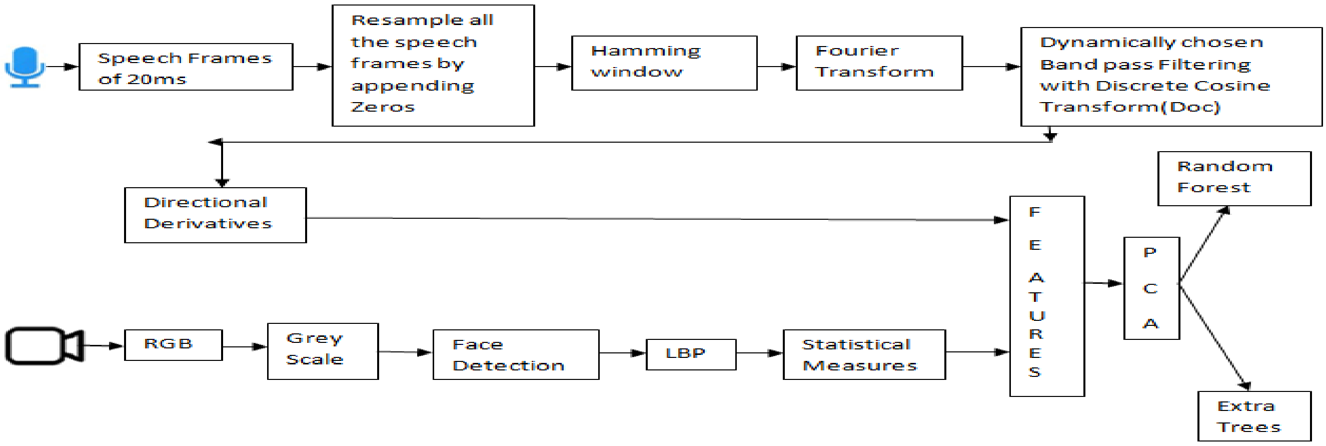


Fig 4.3 Proposed system user satisfaction detection system.

Processing of speech signal

1. **Mel-frequency cepstral coefficients** (MFCC): The speech signal is framed in 40ms window time frame.
2. Each windowed time frame is transformed into Frequency domain (spectrum representation).
3. 24 bandpass filters passed through frequency domain signals. The center frequencies of Bandpass filters are distributed in Mel scale.
4. The frequency domain signal passed through directional derivatives.
5. The derivatives used are a linear regression, where the window size is 3 frames before and 3 frames after the current frame.
6. These derivatives are useful for emotion prediction.
7. Traditionally it was used for detecting emotions of speech. In our case we combine these features with image processing to determine the emotion at higher accuracy.

Speech signal is transmitted to the cloud, where a server extracts and classifies the features. In the server, the speech signal is framed and is calculated using Hamming window. The frame length is 40ms, and the frame shift is 20 ms.

Eachwindowed frame is transformed into a frequency domain representation (spectrum),using Fourier transform, such that the time domain signal is converted into a frequency domain signal.

24-band-pass filters are passed through the frequency domain signal to mimic the hearing perception of the user.

The Centre frequencies of the filters are distributed on a Mel scale.

The bandwidths of the filters correspond to the critical bandwidth. The result of this step is a Mel spectrogram.

The Mel spectrogram is passed through directional derivatives in four directions to obtain the relative progress of the signal along four directions, that is, 0□, 45□, 90□, and 135□, which correspond to time, increasing time frequency, frequency, and decreasing time frequency, respectively.

The derivatives used are a linear regression, where the window size is three frames before and three frames after the current frame.

The following equation shows the calculation of the linear regression along time (0□), where *S*n,f corresponds to the Mel spectrogram at frame *n* and filter *f*.

The directional derivatives are then processed via a discrete cosine transform for compression and de-correlation. Subsequently, 48 features per frame are available for the speech signal.

Processing of image signal

First, a key image frame per one-second video is selected. The key frame is determined by a histogram comparison of the frames.

It is selected when the minimal distance between a frame and its previous and next frames is achieved. This key frame is termed as the image signal.

A face detection algorithm extracts the face area of the image signal.

This processis performed in the local processor to decrease the transmission cost of the video.

An LBP is applied to the face image once it is transmitted to the cloud server to obtain an LBP image. LBP is a powerful texture descriptor and is computationally efficient.

In a rectangular LBP, a window size of 3 x 3 pixels is selected. The intensity of the middle pixel is set as a threshold of the window. If the intensity value of a neighboring pixel is larger than the threshold, then “1” is

assigned to the location of this neighboring pixel; otherwise, “0” is assigned.

The arrangement of “1” and “0” of location of the eight neighboring pixels is concatenated to form an 8-bit binary number, which is then transformed to a denary value.

The denary value is the LBP of the middle pixel. The window is slid by one pixel, and the process is repeated. A histogram is formed from the LBP image. Several features are calculated from the histogram to describe the facial image.

Here Four Features are selected on each LBP image.

1. Mean
2. Standard Deviation
3. Skewness
4. kurtosis

These are well-known statisticalfeatures, which are successfully used in many applications.

Features:

The total sound features are extracted at directional derivatives which is 5976 and also the video features are extracted at statistical measures which is 296.These two extracted sizes are combined at features and resampled. Now, the actual length or shape of features is 7132.These extracted features are only used at the classifiers Random Forest and Extra Trees which helps us give maximum accuracy.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning. PCA makes maximum variability in the dataset more visible by rotating the axes.

PCA is used to identify the components with the maximum variance, and the contribution of each variable to a component is based on its magnitude of variance. It is best practice to normalize the data before conducting a PCA as unscaled data with different measurement units can distort the relative comparison of variance across features.

The optimal number of principal components is determined by looking at the cumulative explained variance ratio as a function of the number of components. The choice of PCs is entirely dependent on the tradeoff between dimensionality reduction and information loss.

The neural net with no PCA application shows a large divergence between training and validation loss metrics, indicating significant overfitting. As the feature space is reduced through PCA, the loss metrics start to converge without significant impact on the accuracy measurement. Dimensionality reduction has increased the model performance and efficiency.PCA is a handy addition to the data scientist toolkit and will improve model performance in most scenarios.

Classifiers:

Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Extra Trees

Adding one further step of randomization yields extremely randomized trees, or Extra Trees. While similar to ordinary random forests in that they are an ensemble of individual trees, there are two main differences: first, each tree is trained using the whole learning sample (rather than a bootstrap sample), and second, the top-down splitting in the tree learner is randomized.Instead of computing the locally optimal cut-point for each feature under consideration, a random cut-point is selected. This value is selected from a uniform distribution within the feature's empirical range (in the tree's training set). Then, of all the randomly generated splits, the split that yields the highest score is chosen to split the node. Similar to ordinary random forests, the number of randomly selected features to be considered at each node can be specified. Default values for this parameter are for classification and for regression, where is the number of features in the model.

CHAPTER 5

IMPLEMENTATION

5.1ALGORITHMS

Algorithms used in the implementation of the project:

Support Vector Machine

SVM algorithm was proposed by Boser, Guyon, and Vapn*ik***.** It was very well used for both classification and regression problem. SVM maps all the data points to a higher dimensional plane to make the data points linear separable. The plane which divides data points is known as hyper plane. It can be used for small dataset to give an optimal solution. SVM cannot be more effective for noisy data. SVM model tries to find out the churn and non-churn customer. In order to divide the dataset into churner and non-churner group, first it will take all the data points in *n-dimensional* plane and divide the data points into churner and non-churner group based on maximum marginal hyper plane. Based on the maximum marginal hyper plane it will divide the data points into churner and non-churner group. Here *n* represents the number of predictor variable associated with the dataset.

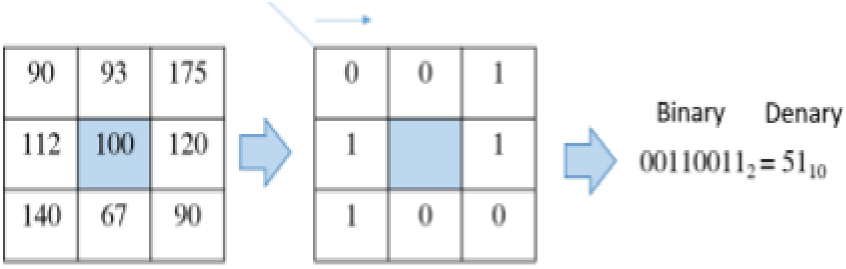
Random Forest

Random forest works based on the random subspace method. The designed strategy used in Random Forest is divide and conquer. It forms number of Decision Trees and each Decision Tree is trained by selecting any random subset of attribute from the whole predictor attribute set. Each tree will grow up to maximum extent based on the attribute present in the subset. Then after, based on average or weighted average method, the final Decision Tree will be constructed for the prediction of the test dataset. Random forest runs efficiently in large dataset. It can handle thousands of input variables without variable deletion. It also handles the missing values inside the dataset for training the model. It is difficult to handle the unbalanced dataset by using Random Forest.

LBP(Local Binary Pattern)

The local binary pattern operator is an image operator which transforms an image into an array or image of integer labels describing small-scale appearance (textures) of the image.

These labels directly or their statistics are used for further analysis.



90	93	175
112	100	120
140	67	90

0	0	1
1		1
1	0	0

Figure 3. Graphical representation of LBP calculation.

Fig5.1 LBP Calculation

An LBP is applied to the face image once it is transmitted to the cloud server to obtain an LBP image. LBP is a powerful texture descriptor and is computationally efficient [13]. In a rectangular LBP, a window size of 3 x 3 pixels is selected. The intensity of the middle pixel is set as a threshold of the window. If the intensity value of a neighboring pixel is larger than the threshold, then “1” is assigned to the location of this neighboring pixel; otherwise, “0” is assigned. The arrangement of “1” and “0” of location of the eight neighboring pixels is concatenated to form an 8-bit binary number, which is then transformed to a denary value. The denary value is the LBP of the middle pixel. The window is slid by one pixel, and the process is repeated. Figure 3 shows an illustration of the LBP calculation. A histogram is formed from the LBP image. Several features are calculated from the histogram to describe the facialimage. The extracted features are the average (mean), standard deviation, skewness, and kurtosis. These are well-known statistical features, which are successfully used in many applications. Other texture descriptors are used; however, the LBP is used in the present work for its simplicity.

KURTOSIS

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

Let $x_1, x_2, ..., x_n$ be n observatio ns. Then,

$$\text{Kurtosis} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

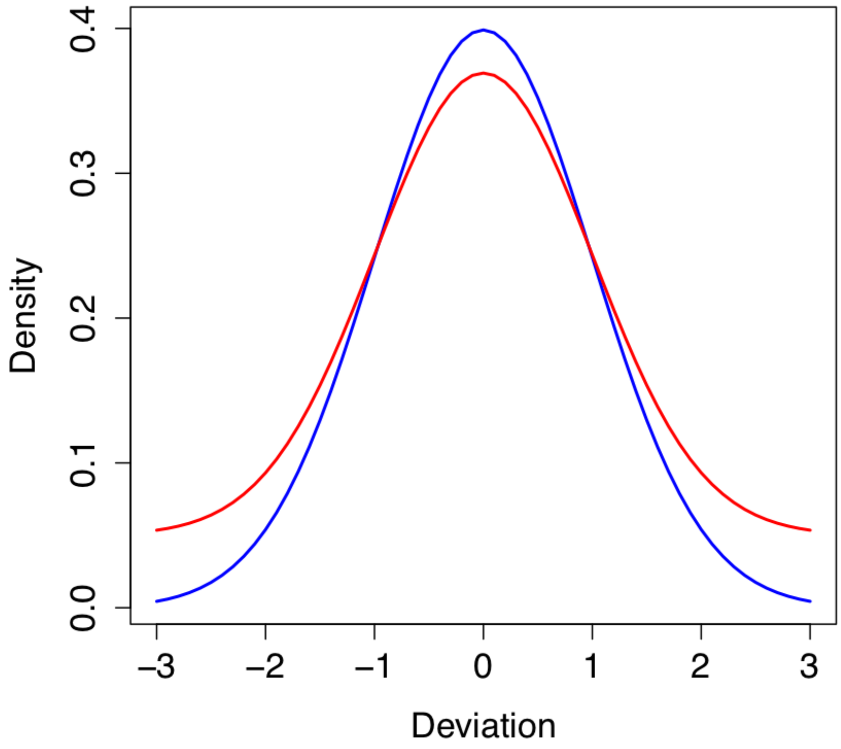


Fig5.2 Kurtosis graph representation.

Kurtosis relates to the relative flatness or peakedness of a distribution. A standard normal distribution (blue line: µ = 0; s = 1) has kurtosis = 0. A distribution like that illustrated with the red curve has kurtosis > 0 with a lower peak relative to its tails.

SKEWNESS

The skewness characterizes the degree of asymmetry of a distributionaround its mean. While the mean, standard deviation, and average deviation are dimensional quantities, that is, have the same units as the measured quantities xj, the skewness is conventionally defined in such a way as to make it non-dimensional. It is a pure number that characterizes only the shape of the distribution.

Measures of asymmetry of data

Positive or right skewed: Longer right tail

Negative or left skewed: Longer left tail

Let $x_1, x_2, ...x_n$ be n observatio ns. Then,

$$\text{Skewness} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

Mel frequency cepstral coefficients (MFCC)

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale.

Speech recognition mainly focuses on training the system to recognize an individual’s unique voice characteristics. The most popular feature extraction technique is the Mel Frequency Cepstral Coefficients called MFCC as it is less complex in implementation and more effective and robust under various conditions.

MFCC is designed using the knowledge of human auditory system. It is a standard method for feature extraction in speech recognition. Steps involved in MFCC are Pre-emphasis, Framing, Windowing, FFT, Mel filter bank, computing DCT.

Speech signals are naturally occurring signals and hence, are random signals. These informationcarrying signals are functions of an independent variable called time. Speech recognition is the process of automatically recognizing certain word which is spoken by a particular speaker based on some information included in voice sample. It conveys information about words, expression, style of speech, accent, emotion, speaker identity, gender, age, the state of health of the speaker etc. There has been a lot of advancement in speech recognition technology, but still it has huge scope. Speech based devices find their applications in our daily lives and have huge benefits especially for those people who are suffering from some kind of disabilities. We can say that such people are restricted to show their hidden talent and creativity. We can also use these speech based devices for security measures to reduce cases of fraud and theft.

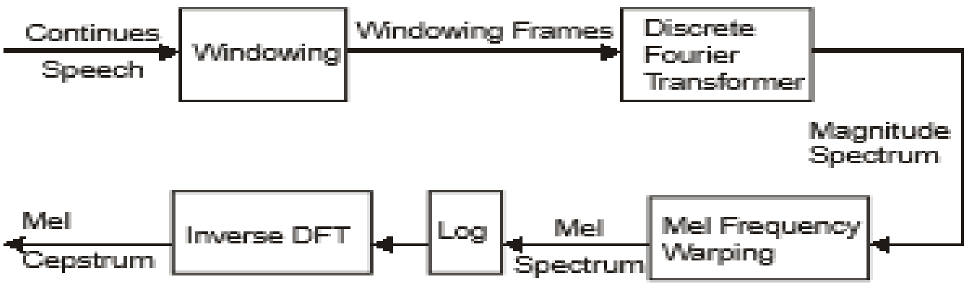


Fig5.3 Complete pipeline for MFCC.

Histogram of Oriented Gradients (HOG)

First used for application of person detection.To increase the efficiency of the face recognition, histogram based facial recognition is chosen, where a face region is fragmented into a number of regions and histogram values are extracted and they are linked together in to single vector.

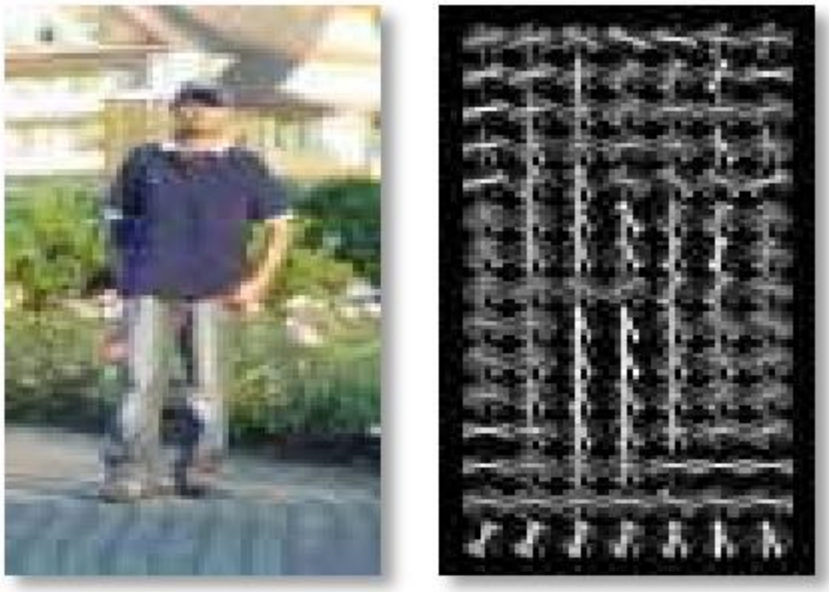


Fig5.4 HOG representation.

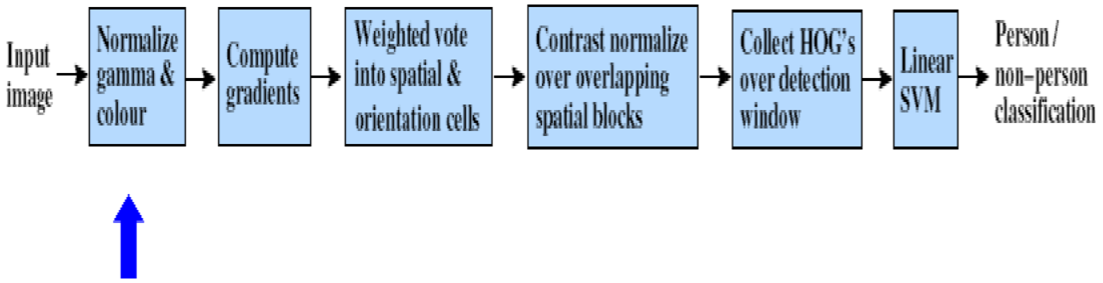

Input image
Normalize gamma & colour
Compute gradients
Weighted vote into spatial & orientation cells
Contrast normalize over overlapping spatial blocks
Collect HOG's over detection window
Linear SVM
Person / non-person classification

Fig5.5 Block diagram of HOG

Tested with – RGB – LAB – Grayscale

Gamma Normalization and Compression – Square root – Log

Extremely Randomized Trees Classifier(Extra Trees Classifier)

It is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result.

Extra Tree has more number of branches and it gives better results than svm.

When the features are huge extra trees are used.

COMBINING THE AUDIO AND VIDEO FEATURES:

import pickle as pk

import numpy as np

importos

fromsklearn.model_selection import train_test_split

from collections import Counter

fromimblearn.over_sampling import SMOTE

fromsklearn.decomposition import PCA

import pickle

```
 #-------------- PCA ----------------------------------------#

 _pca = PCA(n_components=100)

    _pca.fit(X_train)

 pickle.dump(_pca, open("pca_sm.pkl", "wb"))

 X_train = _pca.transform(X_train)

 X_test = _pca.transform(X_test)

 pickle.dump(X_train, open("data_RF_train.pkl", "wb"))

 pickle.dump(X_test, open("data_RF_test.pkl", "wb"))

 pickle.dump(y_train, open("lab_RF_train.pkl", "wb"))

 pickle.dump(y_test, open("lab_RF_test.pkl", "wb"))

 else:

 X_train = pickle.load(open("data_RF_train.pkl", "rb"))

 X_test = pickle.load(open("data_RF_test.pkl", "rb"))

 y_train = pickle.load(open("lab_RF_train.pkl", "rb"))

 y_test = pickle.load(open("lab_RF_test.pkl", "rb"))

 #-------------- Random Forest --------------------------------#

 fromsklearn.ensemble import RandomForestClassifier

 print("\n#--------------- Random Forest ----------------------#\n")

 n_estimators = 100

 rf_mod = RandomForestClassifier(criterion='gini', n_estimators=n_estimators, max_depth=50,
random_state=1, min_samples_split=30)

 rf_mod.fit(X_train,y_train)

rf_trainscore=rf_mod.score(X_train,y_train)
```

```
 print("RF training score:",rf_trainscore)
rf_testscore=rf_mod.score(X_test,y_test)
 print(y_test)
 for yemo in y_test[:20]:
     if(yemo==-1):
         print('neutral')
     elif(yemo==0):
         print('Angry')
     elif(yemo==1):
         print('Happy')
 print("RF testing score:",rf_testscore)
#-------------- Extra Forest --------------------------------#
 from sklearn.ensemble import ExtraTreesClassifier
print("\n#--------------- Extra Trees ----------------------#\n")
n_estimators = 100
 ext_mod = ExtraTreesClassifier(criterion='gini', n_estimators=n_estimators, max_depth=50, random_state=1, min_samples_split=30)
 ext_mod.fit(X_train,y_train)
 ext_mod.predict(X_test)
 ext_testscore=ext_mod.score(X_test,y_test)
print(y_test)
 for yemo in y_test[:20]:
     if(yemo==-1):
       print('neutral')
    elif(yemo==0):
       print('Angry')
    elif(yemo==1):
       print('Happy')
       pickle.dump(ext_mod, open("ext_mod.pkl", "wb"))
 ext_trainscore=ext_mod.score(X_train,y_train)
 print("ExtraTrees training score:",ext_trainscore)
ext_testscore=ext_mod.score(X_test,y_test)
 print("ExtraTrees testing score:",ext_testscore)
```

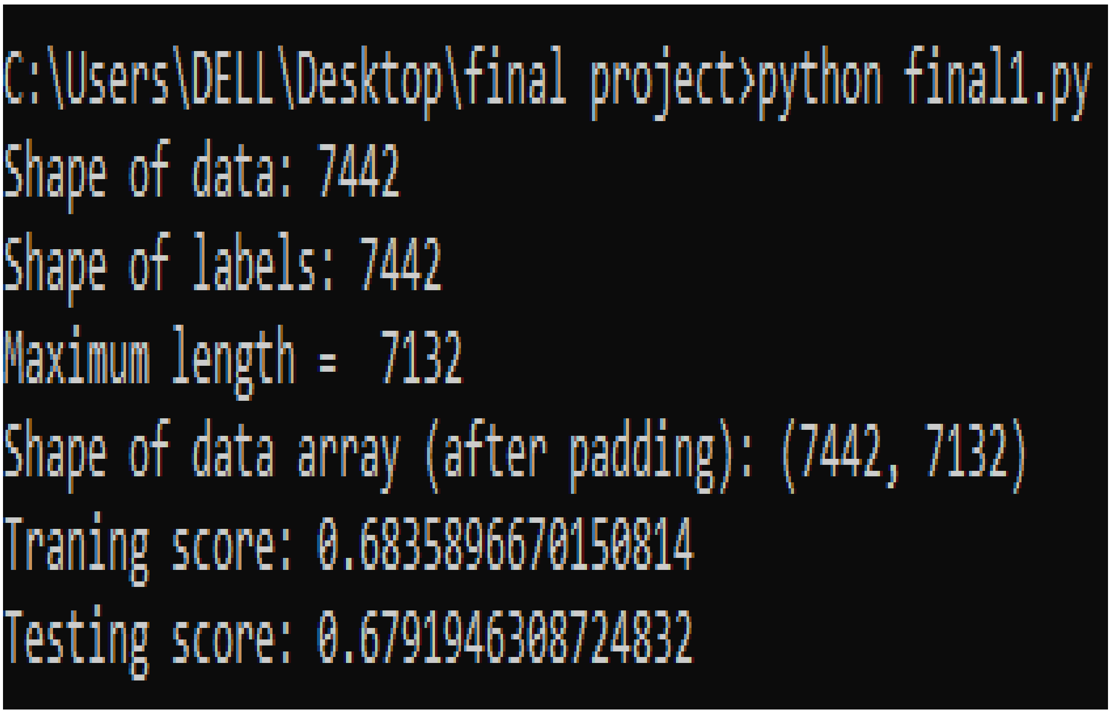
EXPLANATION:

This project monitors a persons emotion and that can be used in healthcare system.This above code is a combination of both audio and video which gives the best accuracy using Random Forest and Extra Trees.We used PCA for feature extraction,it reduces the features size and selects the wanted data.PCA is mainly used to reduce the dimensionality this giving us the features which will represent our image or any application data that will decide the output.PCA is used as it is general for any problem.They keep doing clustering,and will construct centroid wherever they are mostly present and this our first cluster.then for second dense cluster,third dense cluster,they keep on finding the cluster.This first centroid represents the first component and respectively the second and so on.PCA does not harm data is preserved and presented well.Extra trees:Since it contains lot of branches,its extream random forest and gives the best result.Random forest:It has less number of trees and comes up with best tree and gives the result.

CHAPTER 6

RESULTS AND DISCUSSION

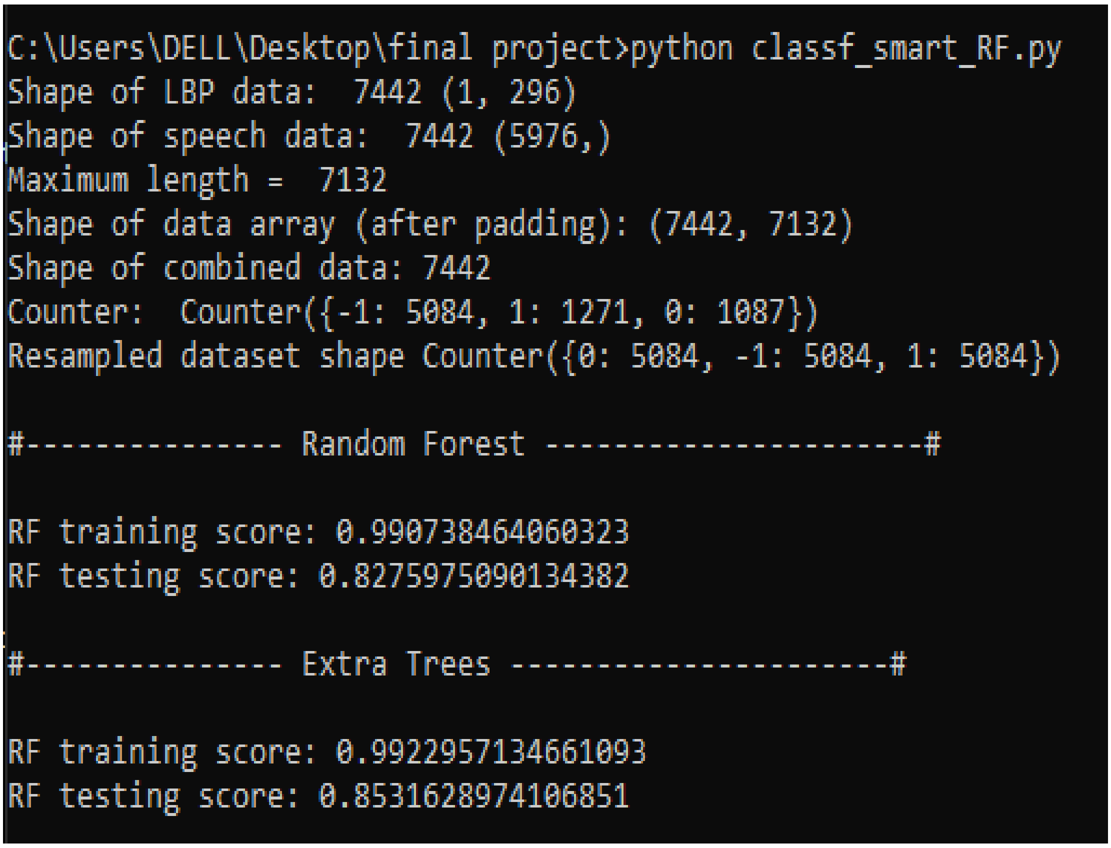
- In fig 6.1 the data set was trained and tested using SVM but the accuracy was same as shown in the paper.
- So to get better accuracy we haven't used SVM for our proposed system.



```
C:\Users\DELL\Desktop\final project>python final1.py
Shape of data: 7442
Shape of labels: 7442
Maximum length =  7132
Shape of data array (after padding): (7442, 7132)
Traning score: 0.6835896670150814
Testing score: 0.6791946308724832
```

Fig 6.1 Output using SVM

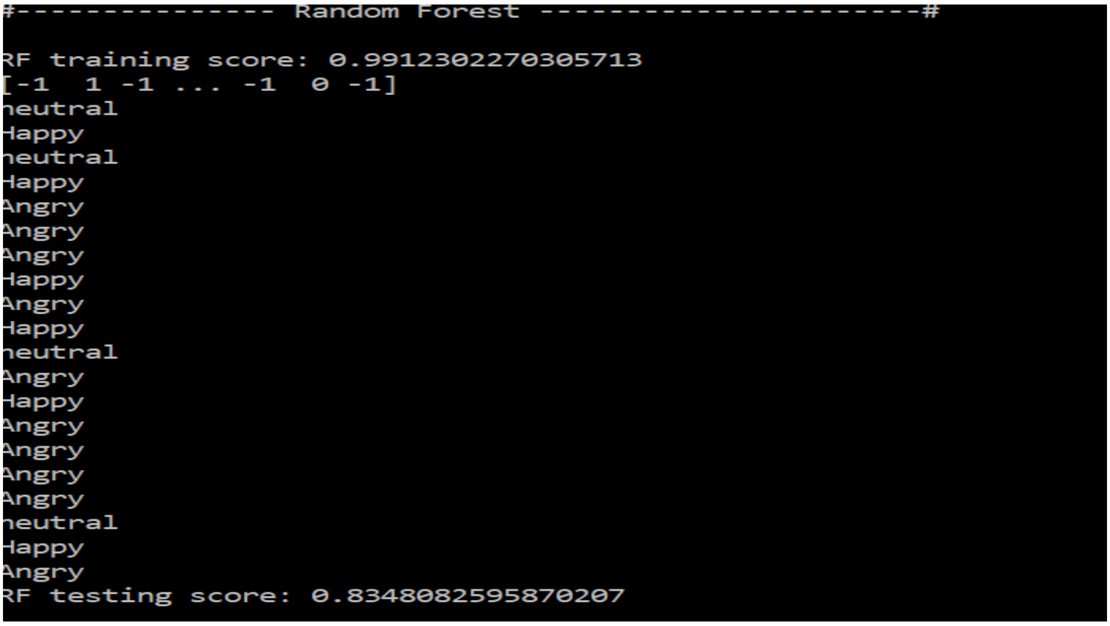
- In fig 6.2 the data set was trained and tested using Random forest and extra tress for best accuracy.
- Better result was shown than SVM so we implemented it using Random forest and Extra trees.



```
C:\Users\DELL\Desktop\final project>python classf_smart_RF.py
Shape of LBP data:  7442 (1, 296)
Shape of speech data:  7442 (5976,)
Maximum length =  7132
Shape of data array (after padding): (7442, 7132)
Shape of combined data: 7442
Counter:  Counter({-1: 5084, 1: 1271, 0: 1087})
Resampled dataset shape Counter({0: 5084, -1: 5084, 1: 5084})

#--------------- Random Forest ----------------------#

RF training score: 0.9907384640603232
RF testing score: 0.8275975090134382

#--------------- Extra Trees ----------------------#

RF training score: 0.9922957134661093
RF testing score: 0.8531628974106851
```

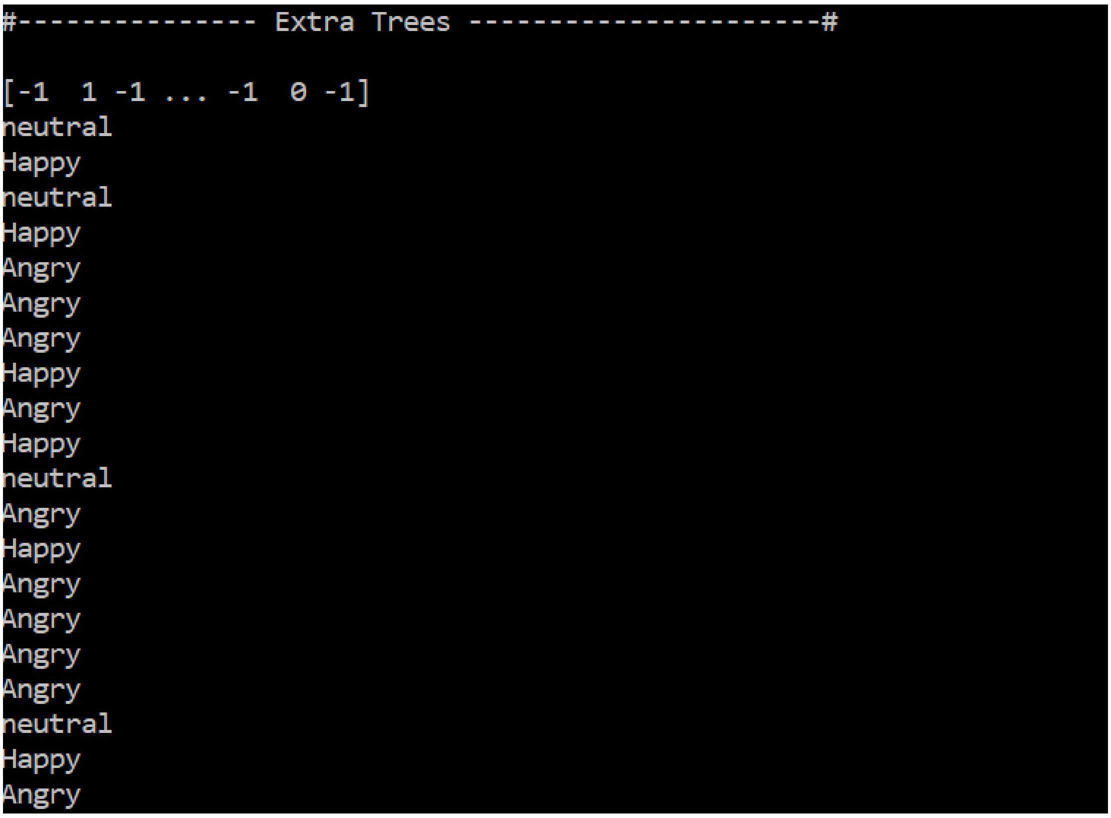
Fig6.2 Output using RF and XTREES

- In fig 6.3 and 6.4 we have used 20 features to test whether the patient is happy, neutral or angry using random forest and extra trees.



```
#--------------- Random Forest ----------------------#

RF training score: 0.9912302270305713
[-1  1 -1 ... -1  0 -1]
neutral
Happy
neutral
Happy
Angry
Angry
Angry
Happy
Angry
Happy
neutral
Angry
Happy
Angry
Angry
Angry
Angry
neutral
Happy
Angry
RF testing score: 0.8348082595870207
```

Fig6.3 Features testing using Random forest.



```
#-------------- Extra Trees ---------------------#

[-1  1 -1 ... -1  0 -1]
neutral
Happy
neutral
Happy
Angry
Angry
Angry
Happy
Angry
Happy
neutral
Angry
Happy
Angry
Angry
Angry
Angry
neutral
Happy
Angry
```

Fig6.4 Features testing using Extra trees.

CHAPTER 7

TESTING

This chapter gives an overview of the various types of testing incorporated during the entire duration of the project.

7.1 Unit Testing:

- Audio Code: Firstly we tested the audio code where mfcc was used in it.Here we checked if the features were extracted or not from the data sample given. Later, we were successful at extracting 5976 features from the data sample of size 7442.We made sure we are getting the same number of extracted features for all the files.

- Video Code: As mentioned above we continue with same process of testing if we are able to extract features or not. Although we got the accuracy similar to the existing system, we were successful in extracting 296 features from the data sample of size 7442. We mainly used SVM to check if the extracted features are working or not.

- The above two were the main unit testing which were carried out and the, we combined both the extracted features and carried out resampling again so that all the subject samples are of same size this was the third unit testing that was done.

- Random forest and Extra Tress algorithms were separately tested to see if they were able to give the accuracy better than the existing system. These both were tested separately so that, we could compare between each other and also SVM.As predicted we were able to achieve 83-84% accuracy.

7.2 Integration Testing: For the integration testing we combined all the unit tested modules which were tested at various instances, to eradicate any errors that may affect the predicted accuracy. After combining we resampled the extracted features i.e. all components of audio and video, we were able to extract a size of 7132 features, we carried out resampling so that there are same number of features. Here the audio/sound code, video code, combination code and the final machine learning algorithms were as a whole tested and we were successful in getting good results i.e. the existed system’s accuracy was 78% and we achieved 83-84% accuracy.Hence, comparatively we have passed the integrity test with better result.

7.3 System Testing: Under System Testing technique, the entire system is tested as per the requirements. It is a Black-box type testing that is based on overall requirement specifications and covers all the combined parts of a system. Here we tested if the end integrated code could run on any system, we saw that the integrated code can run on any system having python version 3.6 or more, and we never faced any error.

7.4 Interface Testing:In our project there was no much of an interface testing as were looking at the console output, so there was not much of interfacing here. Hence the interface testing was quite easy and simple.

7.5 Compatibility Testing: Compatibility was never an issue as we are not dealing with multiple systems or multiple compliers or multiple Id's. In general, we used it on windows and python version greater than 3.6. And we used the libraries like Sci-Kit,numpy,pandas etc. And also, there was no specific requirement like we need to use this particular version of python or Sci-Kit library of particular version, so we worked in a closed loop. We were successful at getting greater efficiency with these.

7.6 Performance Testing:Initially we had some problem with the CPU cycle and all we overcame that by using cloud machines were there were multiple core system and GPU. Given in the present research era, everyone targets 76-78% accuracy so, we tried to be as efficient as possible with the other research on the same given dataset. Though we are using the same dataset as other researchers we tried to get 83-84% accuracy i.e. we got better result so this proves that our performance is better than existing paper. Thus we were able to deliver the expected accuracy.

7.7 Usability Testing:This project could be easy for python and data science programmer,not meant for general purpose. Application is usable for data science engineers to pick the model for future research in the area of emotion recognition and also easy for them to develop monitoring system in CCTV or any electronic appliances etc.

CHAPTER 8

CONCLUSION

A user satisfaction detection system using speech and image for a smart healthcare framework was proposed. For the speech signal, we used the directional derivative features from the Mel-spectrogram, whereas for the image signal, we used the LBP features. SVM, XG boost, random forest were used as the classifier, and several experiments were performed. The best accuracy (75%) will be obtained by combining the features from the speech and image signals.

Managing storage capacity:

From patient, the image and speech data are collected and stored in cloud which requires a massive storage to store the patient details.

Social Challenges:

Accepting and adopting technology that is completely different from the traditional work methods never come easy. Pathologists and the patients nowadays will not be ready to adopt the traditional methods.

Security and privacy of data:

It is important for creating a trusting environment by respecting patient privacy. Confidentially the patient’s data should be maintained.

8.1 FUTURE SCOPE

In future works, we intend to use highly sophisticated classifying approach, such as active learning, which has been successfully used in emotion recognition.

In MPEG-7 audio features were effectively used in an audio– visual emotion recognition. We may use such features and include other input modalities to enhance the accuracy of the proposed system.

We might try to create a front-end device and deploy it on. If this becomes a success, we can use it directly from home and monitor the feedback cloud.

REFERENCES

M. Jampour, V. Lepetit, T. Mauthner, and H. Bischof, “Pose-specific non-linear mappings in feature space towards Multiview facial expression recognition,” Image and Vision Computing, vol. 58, pp. 38-46, February 2017

L. Y. Mano, B. S. Faiçal, L. H.V. Nakamura, P. H. Gomes, G. L. Libralon, R. I. Meneguete, G. P.R. Filho, “Exploiting IoT technologies for enhancing Health Smart Homes through patient identification and emotion recognition,” Computer Communications, vol. 89–90, pp. 178-190, September 2016.

M. S. Hossain and G. Muhammad, “Cloud-assisted Industrial Internet of Things (IoT) - enabled framework for Health Monitoring,” Computer Networks, Vol. 101, No. (2016), pp.192-202, June 2016.

S. Hossain and G. Muhammad, “Audio-visual emotion recognition using multi-directional regression and Ridgelet transform,” J. Multimodal Interfaces, vol. 10, no.4, pp. 325-333, 2016.

Hossain, M.S., Muhammad, G.: Cloud-assisted speech and face recognition framework for health monitoring. Mobile Netw. Appl. **20**(3), 391–399 (2015)

Sun, Y., Wen, G., & Wang, J. (2015). Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomedical Signal Processing and Control,18*, 80-90.

Muhammad, G., Melhem, M.: Pathological voice detection and binary classification using MPEG-7 audio features. Biomed. Signal Process. Controls **11**, 1–9 (2014)

A. Statnikov, I. T. (2005). Gene Expression Model Selector. In.Acharya, U. R., Chua, E. C.-P., Chua, K. C., Min, L. C., & Tamura, T. (2010).Analysis and automatic identification of sleep stages using higher order spectra.*International journal of neural systems, 20*, 509-521.