Machine Learning-Based Audio Interface Model for Sign Language Recognition

Akash Rai, Sujata Kadu, Satish Salunkhe

Abstract**:** ***Due to the fact that most offices and educational institutions now operate from home, the work-from-home and study-from-home cultures have made it difficult to interact with persons who are deaf or hard of hearing. These people communicate within their society using sign language, which is not widely understood by others. Most of the time, as a result of this, they miss out on the opportunity to express their point in front of every one since they are ignored/passed over without receiving the necessary attention. In real-time, having an independent translator that can process photos and interpret signs quickly at the speed of streaming images is critical. We'll utilize TensorFlow Object Detection and Python to bridge the gap by creating an end-to-end bespoke object detection model that not only translates sign language in real time but also speaks it to others.***

Keywords**:** ***Automated Sign Language Recognition, Object Detection, Sign Language Detection***

I. INTRODUCTION

Because only a few signs are recognized by most people, speech and/or hearing-impaired members of society and abroad have a tough time communicating with non-sign people using their natural signing [1]. Around 7% of the global population speaks sign language as their first language. However, most ordinary people, with the exception of their trainers, do not grasp it. As a result, there is a need for the development of a system that will allow persons who are deaf or hard of hearing to communicate with others using natural signing [2]. Building a system that can translate sign language into written language is also a means to enhance and respect each country's and humanity's overall dialectal heritage. And inward, user-friendly, and resilient technologies based on computer vision and image processing are proposed as a solution to the sign language recognition challenge. Despite some progress in this sector, the challenges of computer vision-based sign language recognition systems present a competitive and intriguing opportunity to achieve resilience, preciseness, and strong computational power. As a result, there is a growing demand for computer vision-based real-time continuous sign recognition research. Today, achieving meaningful sign language recognition studies in lifelike circumstances with occlusions, lighting changes, and a crowded background is a major priority [3] [4].

Through this system of real-time sign language identification, we're studying and performing several techniques to distinguish diverse hand movements and patterns formed by human hands to communicate with deaf/dumb people in the digital world. This technology will attempt to close the gap by removing the need for a third-party translation and allowing deaf-dumb people to converse directly. The hearing-impaired population has advanced its very own subculture and verbal exchange strategies with each other and with most people with the aid of using sign gestures. Sign gestures are a form of nonverbal visual communication that differs from spoken language yet serves the same function. It can be challenging for members of the hearing-impaired population to communicate their ideas and inventive approaches to non-hearing people. The goal is to design an end-to-end custom sign recognition model that enables real-time sign language translation as well as a real-time auto speech generator based on the person's actions/hand gestures using TensorFlow Object Detection [5] and Python. SSD (Single Shot Detector) [6] or DNN (Deep Neural Network) [7] [8] will be utilized as the object detection model. SSD is a multi-box technique for real-life object detection. Multiple images within a single input image will be detected with just one shot [9]. After passing through neural networks, DNN utilizes a search selection method to locate the regions where objects can be detected. The discovered object is predicted using the inputted photos

II. LITERATURE SURVEY

A. Real-Time Translation of Indian Sign Language using LSTM

A gesture-sensing glove that can translate Indian Sign Language into speech in real-time has been developed. Flex sensors, gyroscopes, and accelerometers are used in conjunction with the glove to read data relating to both static and dynamic hand gestures. Data is transferred from the glove to the processing equipment using Bluetooth protocols. LSTM networks are then used to classify the received gesture data into appropriate text and audio outputs. [10]

Manuscript received on 26 November 2022 | Revised Manuscript received on 11 December 2022 | Manuscript Accepted on 15 December 2022 | Manuscript published on 30 December 2022.

***** Correspondence Author (s)

Akash Rai, Department of Information Technology, Terna Engineering College, Nerul, Navi Mumbai (Maharashtra), India. Email: akashrai932@gmail.com, ORCID ID: https://orcid.org/0000-0003-0751-8032

Sujata Kadu*, Department of Information Technology Terna Engineering College, Nerul, Navi Mumbai (Maharashtra), India. Email: sujatakadu@ternaengg.ac.in, ORCID ID: https://orcid.org/0000-0001-6346-2104

Satish Salunkhe, Department of Computer Engineering, Terna Engineering College, Nerul, Navi Mumbai (Maharashtra), India. Email: satishsalukhe@ternaengg.ac.in, ORCID ID: https://orcid.org/0000-0002-8395-7101

Retrieval Number: 100.1/ijitee.A93741212122
DOI: 10.35940/ijitee.A9374.1212122
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

B. Real-time Indian Sign Language (ISL) Recognition

A system that uses grid-based features to recognize hand positions and motions from Indian Sign Language (ISL) in real-time. It can recognize 33 hand positions and a few ISL movements. A smartphone camera records sign language, and the frames are sent to a distant server for processing. Hand detection and tracking employ methods including Face detection, Object stabilization, and Skin Color Segmentation. A Grid-based Feature Extraction approach is then used on the image to represent the hand's pose as a Feature Vector [11].

C. Hand Gesture Recognition Software Based on Indian Sign Language

Using skin color segmentation and a neural network model, a straightforward sign language recognition system may be created. We divide the data into various training and testing datasets while using the procedure to train a dataset with a multi-class CNN. The outcomes and applications we produce can be used and deployed as visuals on mobile or other devices. The problem was discovered for alphabets that entail motion, like M and Z, and it is advised to address them using numerous secondary templates [12].

D. Real-Time Sign Language Recognition Based on Video Stream

A Chinese sign language dataset is developed, serving as the foundation for a set of sign language recognition algorithms, and a real-time sign language system based on RGB video streams and 3D-CNN is proposed. Combining RGB video streams with TV-L1 optical flow calculations and 3D-CNN feature extraction can be used to recognize sign language. The frame difference approach is a straightforward and efficient motion detection algorithm that has little impact on the system's real-time performance [13].

III. OBJECTIVE OF THE PROPOSED RESEARCH

The purpose of the study is to those who are deaf or have impaired hearing should not be isolated from their peers when it comes to communication. The system will have the efficiency of making communication easier for persons who are deaf or hard of hearing. The photographs acquired with the webcam are compared, and the results of the comparison are displayed simultaneously. As a result, this feature of the system enables communication very easily and without delay. The objective of this project is

- To provide the real-time output of sign language detected in audio form.
- To make the system signer independent.

To provide better accuracy than previous related research concerning the accuracy

IV. PROPOSED SYSTEM

The block diagram in Figure 1. shows the overall architecture and idea of the system. The gesture image is captured using a webcam from every aspect/angle.

Image acquisition, fragmentation, preprocessing, characteristic extraction, training, and recognition are all part of a sign language recognition system (SLR) [1].

A. The SLR's input method is the acquisition function. For such acquisition of signers' postures or gestures, various types of cameras, sensors, and gloves are usually employed. There are two types of methods: glove-based or sensor-based [14], which uses a specific glove to extract hand posture or movements, and vision-based, which uses a charged-coupled device or web camera to collect image sequences of postures as well as gestures. [15]

B. The process of removing objects or indications from the backdrop of a procured image is referred to as segmentation. Background subtraction, skin-color exposure [16] , and edge identification are all used within the segmentation method. The movement and position of the hand need to be detected and segmented so that it will recognize gestures [17]. Predefined properties such as structure, frame, geometrical feature (viewpoint, range, etc.), histogram, color particularities, and others are fetched from the pre-processed images and used afterward for sign recognition . The system is trained using extracted features by collecting each sign feature's related sign classes before classification.

We recorded each sign, then used background-subtraction methods to remove the backgrounds from each image to create our dataset. The dataset was first divided into two halves for training and validation, and the validation accuracy displayed a high level. However, the validation accuracy significantly dropped when we used datasets from two distinct sources, i.e., training on ours and testing on the prepared and vice versa. We utilized the prepared dataset for the various gestures to train the network because results from training on one dataset and validating on another were not as reliable

C. Classifiers are the tools or methods which are used to classify the signs. General classifiers which classify or recognize sign language include the Hidden Markov Model (HMM), Principle Component Analysis (PCA), Support Vector Machine (SVM) [18], Artificial Neural Network (ANN), and K-Nearest Neighbor (KNN) classifiers. Furthermore, there were two types of sign language phrase recognition: isolated sign recognition, in which each sign has a start and stop point, and continuous sign recognition, in which there are no defined borders between signals. Sequential segmentation is required for the automatic recognition of natural continuous gestures. Often, the start and endpoints of a gesture must be specified in terms of mobility, in both time and space.

Without the use of any additional gadgets, the computer vision-based solution for SLR focuses on a natural connection between people and computers. To construct vision-based Sign Language Recognition systems, this technology employs machine vision and image processing techniques. Computer vision-based systems must be adjusted to satisfy more precise and resilient requirements. The vision-based analysis method is similar to how humans perceive their surroundings and make decisions based on the information they get. Here, the only input device for collecting the data on hand gestures is a webcam.

D. The camera's recorded input video is fragmented into a range of features, which are then used for sign language training and recognition. To eliminate noise and focus on essential parts of the image, some type of filtering or preprocessing may be applied to the frames. Different postures are recognized by single hands, while gestures are recognized by a succession of postures interconnected by sustained motions. The gestures that have been detected can be employed to determine specific signs in a given sign language.

As shown in Fig 1. instead of utilizing different types of sensors or gloves, signers perform gestures in front of the webcam, and the system records 2D image streams from the camera. To segment the objects of interest, the collected images are transmitted through a gesture recognition and tracking module, further collecting the data for the dataset and accompanied by a hand area segmentation module. The feature extraction technique receives each image frame after it has been pre-processed. The retrieved features are utilized to train the system, which will then be used to classify or recognize sign languages. We propose a technique for classifying images of the letters, numerals, and words used in sign language using deep convolutional networks. Convolutional neural networks (CNNs) [19] use a technique known as pooling and subsampling layers, nonlinear layers, fully connected layers, and convolution layers to represent the characteristics that will be learned. The new representation will include intricate non-linear feature interactions as well as other image attributes. To identify indications, a SoftMax layer will be utilized [20]

The image thus captured is sent to the computer which segments the image based on the texts they are mapped and processes to extract the gestures and recognize in real-time and gives the output in form of text.

By using our collection of sign datasets to train the network, we used a straightforward supervised learning strategy. We aim to categorize the letters and the digits, 0-9, in ASL using deep convolutional neural networks. The inputs were 100 by 100 fixed-size high-pixel photographs.

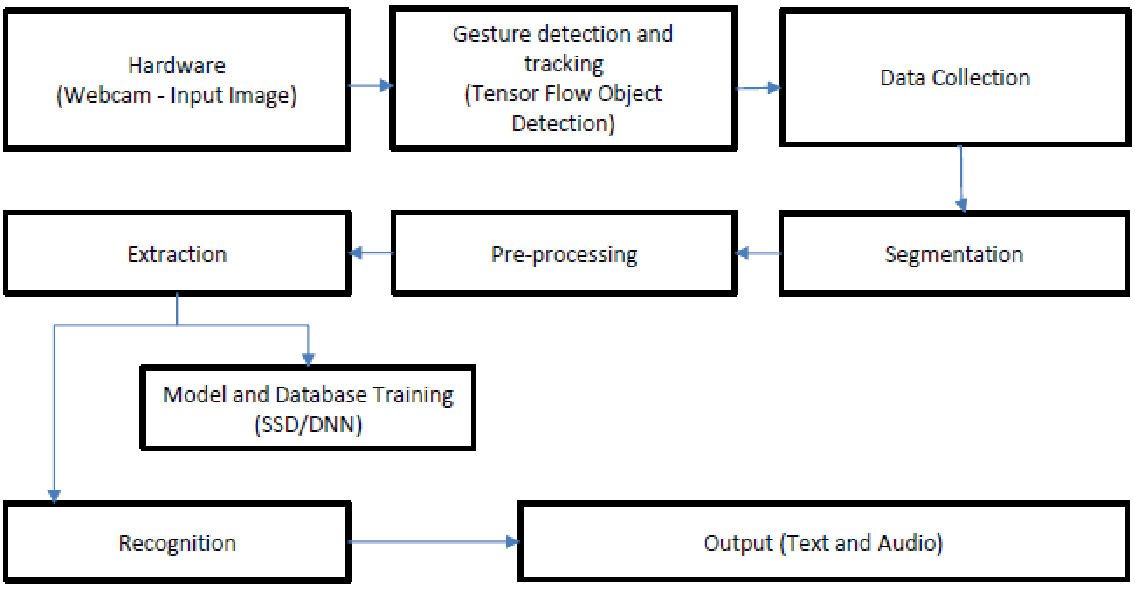

Fig. 1.: Proposed System Flowchart.

Tensor Flowjs and React.js are used to create a system-independent web app that can run on any machine and allow users to easily communicate with one another using sign language on any platform

V. RESULT & ANALYSIS

Two distinct tests serve as the foundation for evaluating the effectiveness of the Indian Sign Language recognition system. First, the parameters that were used to train the model are tweaked, changing the number of layers, filters, and optimizers. The trained model's performance is assessed in the second experiment using both the color and the grayscale image datasets. Additionally, the ISL recognition system's average precision, recall, F-measure, and accuracy have been calculated [21].

A measure of precision counts how many correctly positive forecasts were made. It is determined by dividing the total number of correctly anticipated positive examples by the ratio of correctly predicted positive examples. The ratio of true positives to total true positives and false positives is used to calculate accuracy.

Precision = True Positives/(True Positives +False Positives)

A recall is a metric that measures the proportion of accurate positive predictions among all possible positive predictions.

A recall is determined by dividing the total number of true positives by the sum of true positives and false negatives. An outcome is a number that ranges from 0.0 for no recall to 1.0 for uniform or perfect recall.

Recall = TruePositives / (TruePositives + FalseNegatives)

Precision and recall can be combined into one metric using F-Measure, which covers both characteristics.

Precision and memory don't give the complete tale on their own. We can have absolute precision but poor recall, or vice versa, decent precision but incredible recall. A way to communicate both concerns with a single score is offered by the F-measure.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

For a binary or multiclass classification task, precision and recall can be determined, and the two scores can then be combined to obtain the F-Measure.
F-Measure = (2 * Precision * Recall) / (Precision + Recall)
Table 1 displays the classification results for a few of the grayscale sign samples, including precision, recall, and F measure.

Table- I: Performance Classification into Precision, Recall & F-measure

Sign	Precision	Recall	F-measure
A	0.98	0.92	0.95
B	1	0.95	0.97
C	0.96	0.96	0.96
D	0.79	0.97	0.87
E	0.98	0.9	0.94
F	1	1	1.00
G	0.97	0.97	0.97
H	1	1	1.00
I	1	1	1.00
J	0.92	0.96	0.94
K	1	1	1.00
L	1	1	1.00
M	0.97	0.79	0.87
N	1	1	1.00
O	0.96	0.97	0.96
P	1	0.97	0.98
Q	0.97	1	0.98
R	1	0.98	0.99
S	0.95	1	0.97
T	0.99	1	0.99
U	1	0.99	0.99
V	1	1	1.00
W	0.97	0.96	0.96
X	1	1	1.00
Y	1	1	1.00
Z	0.99	1	0.99
0	0.96	0.88	0.92
1	1	0.96	0.98
2	1	1	1.00
3	0.98	1	0.99
4	0.97	0.92	0.94
5	0.79	0.94	0.86
6	0.9	0.99	0.94
7	1	0.8	0.89
8	0.94	1	0.97
9	1	1	1.00

VI. CONCLUSION

The goal of this project is to predict alphanumeric hand gestures in real time. This will can be solved with better accuracy than most of the systems present by training various models and using more data collection. By means of making use of depth-based segmentation, we eliminate the overheads of the dynamic background. Also, the voice command will be invoked as soon as it guesses the sign/gesture.

It also aims to focus on the different techniques and applications of object detection available and used and to identify which one is more effective and reliable. Also, the system is independent of the hands used, irrespective of whether the hand used to make gestures is the left hand or right hand.

The project aims to make communication between deaf and dumb people easier by including a computer in the communication chain, allowing sign language to be automatically captured, recognized, and translated into text and voice. Converting from one sign language to another can be done in a variety of ways. Some employ a wired electronic glove, while others rely on a visual method. Electronic gloves are expensive, and one individual cannot use another's glove. Different techniques are utilized in the vision-based approach to recognize and match captured motions with gestures in the database. For the benefit of the hearing impaired, the image obtained must be examined, processed, and transformed to either sign or text and audio. The usage of SSD Mobile Nets and other image processing techniques aids in the system's ability to convert standard sign language from the form of gestures to textual content and voice with greater accuracy. The Web app gives the system more features, enabling it to work with any operating system and any browser.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	According to the authors, the contribution to the paper as follows: Study conception and Design: Akash Rai, Dr. Sujata Kadu; Data collection: Akash Rai; Analysis and interpretation of results: Akash Rai; Draft manuscript preparation: Akash Rai, Dr. Sujata Kadu, Dr. Satish Salunkhe.

REFERENCES

1. Z. M. Malakan and H. A. Albaqami, "Classify, Detect and Tell: Real-Time American Sign Language," In National Computing Colleges Conference (NCCC), 2021, pp. 1-6, DOI: 10.1109/NCCC49330.2021.9428808. [CrossRef]
2. M. Safeel, T. Sukumar, S. K. S, A. M. D, S. R and P. S. B, "Sign Language Recognition Techniques- A Review," In IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-9, DOI: 10.1109/INOCON50539.2020.9298376. [CrossRef]
3. J. Guo, P. Chen, Y. Jiang, H. Yokoi, and S. Togo, "Real-time Object Detection with Deep Learning for Robot Vision on Mixed Reality Device," In IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), 2021, pp. 82-83, DOI: 10.1109/LifeTech52111.2021.9391811. [CrossRef]
4. I. Ahmed, S. Din, G. Jeon, F. Picciallі and G. Fortino, "Towards Collaborative Robotics in Top View Surveillance: A Framework for Multiple Object Tracking by Detection Using Deep Learning," in IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 7, pp. 1253-1270, July 2021, DOI: 10.1109/JAS.2020.1003453. [CrossRef]
5. I. Kilic and G. Aydin, "Traffic Sign Detection And Recognition Using TensorFlow’ s Object Detection API With A New Benchmark Dataset," In International Conference on Electrical Engineering (ICEE), 2020, pp. 1-5, DOI: 10.1109/ICEE49691.2020.9249914. [CrossRef]

6. Z. -Q. Zhao, P. Zheng, S. -T. Xu and X. Wu, "Object Detection with Deep Learning: A Review," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212-3232, Nov. 2019, DOI: 10.1109/TNNLS.2018.2876865 [CrossRef]
7. W. Li, X. Tan and Z. Wang, "Small Object Detection of Table Tennis Based on Deep Learning Network," In International Conference on Computer Science and Management Technology (ICCSMT), 2020, pp. 149-152, DOI: 10.1109/ICCSMT51754.2020.00036. [CrossRef]
8. Y. Wang, L. Li, X. Yang, X. Wang, and H. Liu, "A Camouflaged Object Detection Model Based on Deep Learning," In IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), 2020, pp. 150-153, DOI: 10.1109/ICAIIS49377.2020.9194881. [CrossRef]
9. R. Abiyev, J. B. Idoko, and M. Arslan, "Reconstruction of Convolutional Neural Network for Sign Language Recognition," In International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 2020, pp. 1-5, DOI: 10.1109/ICECCE49384.2020.9179356. [CrossRef]
10. E. Abraham, A. Nayak, and A. Iqbal, "Real-Time Translation of Indian Sign Language using LSTM," 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-5, DOI: 10.1109/GCAT47503.2019.8978343. [CrossRef]
11. K. Shenoy, T. Dastane, V. Rao and D. Vyavaharkar, "Real-time Indian Sign Language (ISL) Recognition," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1-9, DOI: 10.1109/ICCCNT.2018.8493808. [CrossRef]
12. S. Kadam, A. Ghodke and S. Sadhukhan, "Hand Gesture Recognition Software Based on Indian Sign Language," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-6, DOI: 10.1109/ICIICT1.2019.8741512. [CrossRef]
13. K. Zhao, K. Zhang, Y. Zhai, D. Wang, and J. Su, "Real-Time Sign Language Recognition Based on Video Stream," 2020 39th Chinese Control Conference (CCC), 2020, pp. 7469-7474, DOI: 10.23919/CCC50068.2020.9188508. [CrossRef]
14. S. Kumuda and P. K. Mane, "Smart Assistant for Deaf and Dumb Using Flexible Resistive Sensor: Implemented on LabVIEW Platform," In International Conference on Inventive Computation Technologies (ICICT), 2020, pp. 994-1000, DOI: 10.1109/ICICT48043.2020.9112553. [CrossRef]
15. K. Tiku, J. Maloo, A. Ramesh and I. R., "Real-time Conversion of Sign Language to Text and Speech," In Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 346-351, DOI: 10.1109/ICIRCA48905.2020.9182877. [CrossRef]
16. H. Bhavsar and J. Trivedi, "Hand Gesture Recognition for Indian Sign Language using Skin Color Detection and Correlation-Coefficient algorithm with Neuro-Fuzzy Approach," In International Conference on Advances in Computing, Communication and Control (ICAC3), 2019, pp. 1-5, DOI: 10.1109/ICAC347590.2019.9036832. [CrossRef]
17. M. A. Rady, S. M. Youssef and S. F. Fayed, "Smart Gesture-based Control in Human-Computer Interaction Applications for Special-need People," In Novel Intelligent and Leading Emerging Sciences Conference (NILES), 2019, pp. 244-248, DOI: 10.1109/NILES.2019.8909324 [CrossRef]
18. W. Zhang, C. -f. Yang, F. Jiang, X. -z. Gao and X. Zhang, "Safety Helmet Wearing Detection Based on Image Processing and Deep Learning," In International Conference on Communications, Information System and Computer Engineering (CISCE), 2020, pp. 343-347, DOI: 10.1109/CISCE50729.2020.00076. [CrossRef]
19. Y. Zhao, J. Zhao, C. Zhao, W. Xiong, Q. Li, and J. Yang, "Robust Real-Time Object Detection Based on Deep Learning for Very High-Resolution Remote Sensing Images," In IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 1314-1317, DOI: 10.1109/IGARSS.2019.8897976 [CrossRef]
20. R. Bhadra and S. Kar, "Sign Language Detection from Hand Gesture Images using Deep Multi-layered Convolution Neural Network," In IEEE Second International Conference on Control, Measurement and Instrumentation (CMI), 2021, pp. 196-200, DOI: 10.1109/CMI50323.2021.9362897. [CrossRef]
21. H. S. DIKBAYIR and H. İbrahim BÜLBÜL, "Deep Learning-Based Vehicle Detection From Aerial Images," In 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 956-960, DOI: 10.1109/ICMLA51294.2020.00155. [CrossRef]

AUTHORS PROFILE

Akash R. Rai received a bachelor. degree in Information & Technology from the University of Mumbai in 2020. He is currently pursuing his Master's degree in Information Technology from Terna Engineering College. His areas of research include Machine Learning, Data Analysis, and Web Development Frameworks.

Sujata R Kadu received a Ph.D. degree in Electronics and Telecommunication from the University of Mumbai in 2022. She is currently an assistant professor at the University of Mumbai, Terna Engineering College. Her areas of research include Signal Processing, Image Segmentation and Classification, Object-Based Image Analysis, and Multiresolution Segmentation.

Satish S. Salunkhe holds a Ph.D. in Computer Science and Engineering. Currently, he is working as a Professor in the Department of Computer Engineering, at Terna Engineering College, Navi, Mumbai. He has overall 16 years of experience in teaching, research, and administration. His research areas include AI in fuzzy logic, data mining, machine learning, and data warehousing. He has written six journal articles and four conference proceedings published in reputed Scopus and web of science listed journals.