Selective Unlearning in Face Recognition: Forgetting Faces without Sacrificing Accuracy

Adithyan M Nair
adithyan146@gmail.com
Department of Computer Science
Amrita School of Engineering
Amritapuri, India

Akshit Sudheer Kumar
akshitsk16@gmail.com
Department of Computer Science
Amrita School of Engineering
Amritapuri, India

Devakrishna Sanil Kumar
devakrishnasanilkumar@gmail.com
Department of Computer Science
Amrita School of Engineering
Amritapuri, India

Anjali T
anjalit@am.amrita.edu
Department of Computer Science
Amrita School of Engineering
Amritapuri, India

***Abstract*—Facial recognition technology is increasingly being used in current applications ranging from smartphone unlocking to security and surveillance systems. This extensive adoption, however, has generated severe concerns about individual privacy and data policies. In response to these concerns, this study introduces a novel methodology that gives people more control over their data in facial recognition systems. This study's main contribution is the proposal of a strategy for selectively unlearning certain faces from trained facial recognition algorithms. An iterative model adaption approach is used to achieve this selective unlearning process. It becomes feasible to empower individuals to control the presence of their facial data in these systems by iteratively fine-tuning the model. This not only improves individual privacy but also conforms with ethical technology development standards. Furthermore, the methodology presented in this study is an important step towards improving privacy in the age of ubiquitous facial recognition. It tackles increasing concerns about surveillance, tracking, and unauthorized use of personal data by adapting and personalizing facial recognition algorithms to accommodate individual tastes. This study provides an innovative answer to the ethical and privacy concerns raised by facial recognition technologies. It empowers users to exercise greater control over their privacy and data in the face of increasingly prevalent facial recognition technologies by allowing them to selectively unlearn their data from training algorithms.**

***Index Terms*—Face Recognition, Selective Unlearning, Privacy, Model Adaptation, and Recognition Accuracy.**

I. Introduction

Concerns about privacy and individual control over personal data have taken the front stage in an era where face recognition technology has become a vital component of current security and authentication systems. The pervasiveness of this technology has resulted in substantial advances in convenience and security, with applications ranging from smartphone unlocking to access control and personnel management systems. However, as more businesses and organizations implement face recognition systems, a fundamental concern arises: how can an individual's face be selectively and securely removed from a face recognition model once their association with a given system has ended? [1]

This research intends to investigate approaches for the deliberate and precise "forgetting" of an individual's facial features from a face recognition model to meet the multifaceted issues provided by the usage of facial recognition technology. The capacity to remove a person's facial data from such a model is not only a question of individual privacy, but it is also a legal and ethical need for organizations that handle sensitive personal information. Data protection laws in various areas, such as the General Data Protection Regulation (GDPR) in Europe, emphasize how crucial it is to give people control over their data, including their facial data.

This study aims to develop strategies for the regulated and successful removal of an individual's face identity from a model while protecting the model's overall performance and the identification accuracy of all other persons. It is critical to strike this delicate balance to guarantee that the system respects an individual's right to be forgotten without jeopardizing its fundamental tasks.

The Facenet face recognition model developed by Microsoft serves as a base model for the proposed technique. Facenet is a well-known model noted for its high accuracy and robust performance. To assist selective unlearning of individuals from the model, the model's last layer is deleted and a fully connected layer is introduced in its place, to perform classification. This adaptation enables the model to be fine-tuned with a specific dataset by training it for a few epochs, which is an important step in ensuring that the model can recognize and unlearn individuals as needed. [2]

This paper's primary purpose is to provide a thorough overview of the procedures and methodologies involved in enabling individuals to selectively and securely remove their facial data from face recognition algorithms. This work has consequences not only for individual privacy but also for organizations attempting to meet legal and ethical data man-

agement duties. Furthermore, the study provides practical insights into the technical aspects of model adaption and fine-tuning, which can be used to construct ethical and privacy-conscious face recognition systems.

The methodologies and strategies used will be covered in depth in the parts that follow, along with a step-by-step description of the selective unlearning procedure. This paper also addresses the broader ethical and legal issues surrounding this technology, emphasizing the significance of these discoveries in light of contemporary privacy concerns and data protection rules.

II. Literature Review

G. Bae *et al.* [3] introduces DigiFace-1M, a new large-scale synthetic face dataset for face recognition research. The dataset contains 1.2 million photo-realistic rendered face images across 110,000 identities, offering greater diversity and scale compared to previous synthetic datasets. A key contribution is the fully synthetic generation process, avoiding ethical issues associated with collecting real face images without consent. Experiments demonstrate state-of-the-art accuracy among synthetic datasets on standard benchmarks, closing much of the gap to models trained on millions of real images. When combined with even a small number of real images, performance is boosted further. The dataset enables the development of face recognition models with fewer real images, mitigating privacy and consent concerns [4].

G. Singh and A. K. Goel *et al.* [5] present a face detection and recognition system using digital image processing techniques. It reviews various approaches for face detection including feature-based methods like Active Shape Models and low-level analysis of attributes like skin color, as well as image-based methods like neural networks. For face recognition, it discusses geometric vs photo-metric approaches and key algorithms like Eigenfaces, Fisherfaces, and Principal Component Analysis. A core contribution is developing a complete system integrating face detection, pre-processing, feature extraction, and recognition modules. While limitations exist in terms of accuracy and real-world robustness, the system provides a good foundation for further research and improvement in automated face analysis using digital image processing [6].

Vikram S. Chundawat *et al.* [7] introduces the novel problem of zero-shot machine unlearning, where no training data is available to update the model for forgetting. Two solutions are proposed - error minimizing-maximizing noise and gated knowledge transfer. The error-minimizing noise acts as proxy data for retaining classes, while error-maximizing noise forgets the target classes. In gated knowledge transfer, a student model learns from a teacher via generated pseudo-data, with a bandpass filter blocking information on forgotten classes. Experiments on vision datasets demonstrate promising unlearning performance in the zero-shot setting, without using any original training data. The paper formally defines zero-shot unlearning, proposes solutions, and analyzes model inversion and membership inference attacks for robustness. Overall, it opens up a new and efficient direction in machine unlearning research [8].

Huawei Lin *et al.* [9] introduces a novel framework for GBDT's (gradient-boosting decision trees) machine unlearning. It provides improvements including incremental updates, random split candidates, and random layers to enable effective subtree retraining during unlearning and clearly characterizes the unlearning problem for GBDT. Studies show that the framework can successfully unlearn data without a complete model retraining, with no effect on the accuracy of the remaining data. After unlearning, it also succeeds in eliminating backdoors and passes membership inference tests. Overall, this is the first study to address the difficult issue of machine unlearning in GBDT models, and empirical findings support it as a practical and effective solution. [10].

This work suggests a novel process that involves initial training with enhanced data, then substituting photos of the identity that will be forgotten with noise and retraining the model to dissociate that particular face. This makes it possible to tweak the model's parameters to "unlearn" one face while keeping the accuracy of others. The suggested method is contrasted with a baseline that, before fine-tuning the model, reidentifies the to-be-forgotten face as an unknown. The following sections provide a detailed explanation of the methodology's main steps.

III. Methodology

To address the task of forgetting a face from a trained face recognition model while maintaining recognition accuracy for other faces, This study proposes a novel approach that integrates the process of selective unlearning during model training. The proposed methodology consists of the following key steps:

A. Novel Approach

- **Model Training with Augmented Data:** Begin with the initial training of the face recognition model, wherein it is trained to recognize a set of target faces [11] [12]. Importantly, the training dataset includes additional facial images that do not belong to the set of target faces, labeled as a generalized class or category that encompasses all faces outside the target set [13]. The model architecture is illustrated in Fig.1.
- **Unlearning Process:** For the specific identity that needs to be forgotten, the proposed method deletes the images belonging to it, to replicate a real-world scenario and replace them with noisy images to maintain the network structure.
- **Unlearning Training:** Retrain the model on the entire dataset, which does not include the original target faces. During this phase, the model focuses on adjusting its recognition parameters to dissociate the specific face from the target faces, and the model should learn the noisy images over the information of the identity to be forgotten, which would maintain the accuracy of the images that need not be forgotten.

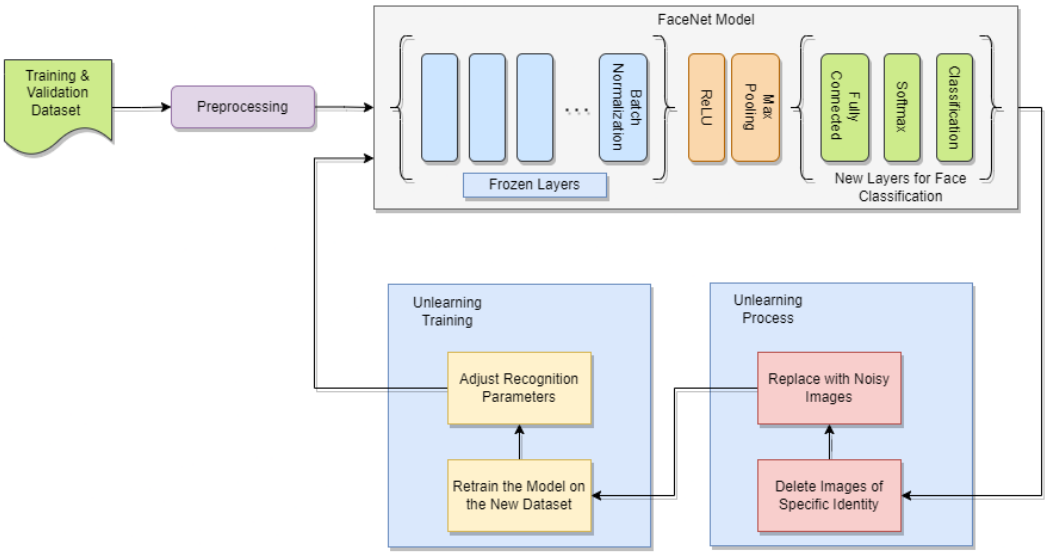

Fig. 1. Process Flow of Novel Approach.

B. Comparison Method

This study uses a method of pointing all the images of the identity to be forgotten to a set containing random unknown faces and retraining the model, to act as a comparison method.

- **Model Training with Augmented Data:** Begin with the initial training of the face recognition model, wherein it is trained to recognize a set of target faces [11] [12]. Importantly, the training dataset includes additional facial images that do not belong to the set of target faces, labeled as a generalized class or category that encompasses all faces outside the target set [13].
- **Unlearning Process:** For the specific face that needs to be forgotten, modify the model's training label to point to the generalized class instead of the individual's identity. This step essentially "relabels" the to-be-forgotten face as an unfamiliar or generalized. [14]
- **Unlearning Training:** Train the model for a few epochs using only the to-be-forgotten face which points to the modified labels. During this phase, the model focuses on adjusting its recognition parameters to dissociate the specific face from the target faces and assign it to the generalized category. [15]
- **Fine-tuning Training:** Subsequently, fine-tune the model on the entire dataset, including both the original target faces and the augmented dataset with the reclassified to-be-forgotten face. This fine-tuning step helps the model maintain recognition accuracy for all other individuals while solidifying the unlearning process.

The above process is illustrated in Fig.2.

IV. Use Case: Selective Unlearning of Facial Data in an Enterprise Authentication System

Use Case Overview: This use case describes the scenario where an organization implements a facial recognition system for employee access control and authentication, and needs to ensure that individuals can selectively and securely remove their facial data from the system when their association with the organization ends. The organization is committed to respecting privacy, complying with data protection regulations, and ethical data handling.

Actors:

- **Employee**: The individual whose facial data is stored in the organization's facial recognition system.
- **System Administrator**: Responsible for managing and maintaining the facial recognition system.
- **Facial Recognition System**: The technology responsible for recognizing and authenticating employees based on their facial features.

Use Case Steps:

1) **User Account Creation:**
   - The system administrator creates user accounts for employees in the facial recognition system.
   - During account creation, the employee's facial data is collected and stored in the system.
2) **Regular Authentication:**
   - Employees use the facial recognition system for daily access control, attendance tracking, or other authentication purposes.
   - The system successfully recognizes and authenticates employees based on their facial features.
3) **End of Association:**
   - When an employee's association with the organization ends (e.g., due to resignation, termination, or

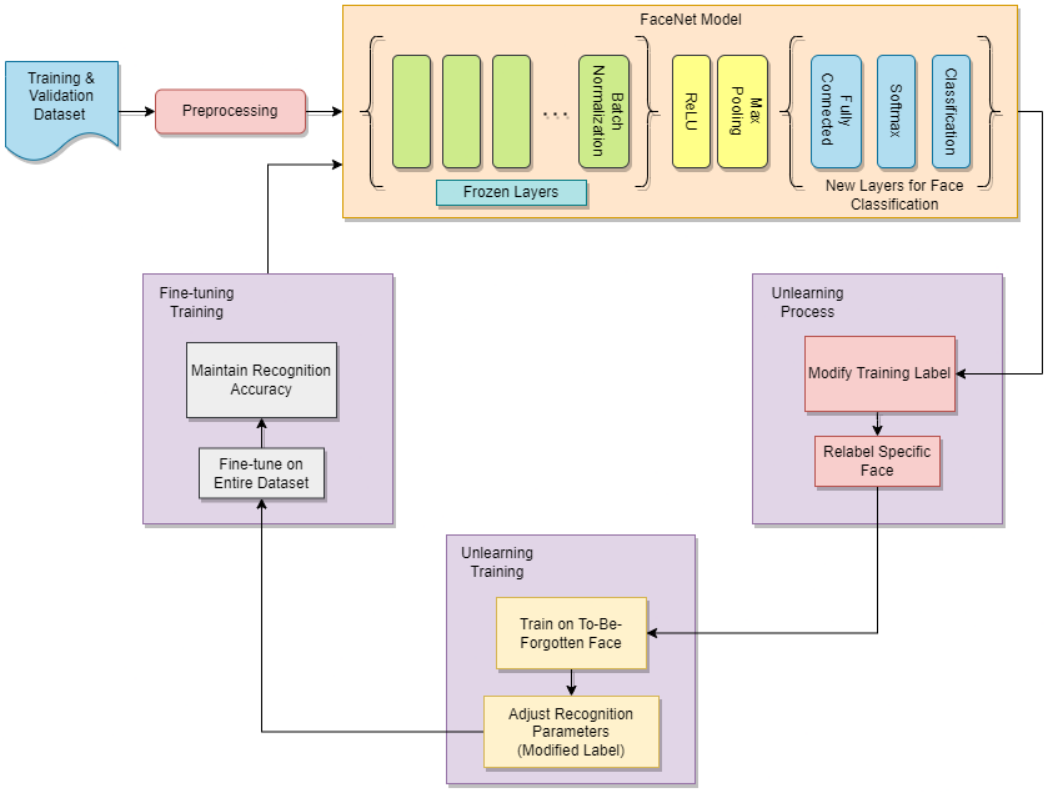

Fig. 2. Process flow of Comparison Model.

change in responsibilities), the employee or the system administrator initiates the process of removing the individual's facial data from the system.

4) **Selective Unlearning Request:**
   - The employee informs the system administrator about their decision to remove their facial data from the system.
   - Alternatively, the system administrator may initiate this process upon employee separation, following organization policies.
5) **Model Adaptation:**
   - The system administrator uses the methodology outlined in the research paper, adapting the facial recognition model to selectively unlearn the facial data of the departing employee.
   - This process involves removing the employee's facial features from the model while preserving recognition accuracy for all other employees.
6) **Verification and Testing:**
   - The system administrator thoroughly tests the adapted model to ensure that it no longer recognizes the departing employee, while still accurately recognizing all other authorized users.
7) **Secure Data Deletion:**
   - Once the model adaptation is deemed successful, the employee's facial data is securely deleted from the system, complying with data protection regulations and ensuring privacy.
8) **Post-Removal Verification:**
   - The system administrator confirms that the employee's facial data is no longer stored in the system and that their facial features are no longer recognized.
9) **Ongoing System Use:**
   - The facial recognition system continues to operate for remaining employees, recognizing them accurately without any compromise in performance.
   - Periodic model updates and retraining are conducted to maintain the system's overall accuracy and performance.

This use case outlines the practical implementation of the selective unlearning methodology in an enterprise environment, where privacy, data control, and ethical data handling are paramount. It demonstrates how organizations can responsibly manage facial recognition data, providing individuals with the ability to control the use of their facial data while maintaining the system's functionality and accuracy for ongoing employees.

V. Results

Fig. 3. Model fine-tuning performance graph.

A. Phase 1: Face Recognition

The Face Recognition model used for the unlearning experiment is the FaceNet model [2], with a fully connected layer added in place of the last layer of the model [16] [17], which is then fine-tuned to recognize the selected faces from the DigiFace Dataset [3]. The model's performance during fine-tuning epochs is shown in Fig. 3. The model with no unlearning implemented performance is shown in Fig. 4.

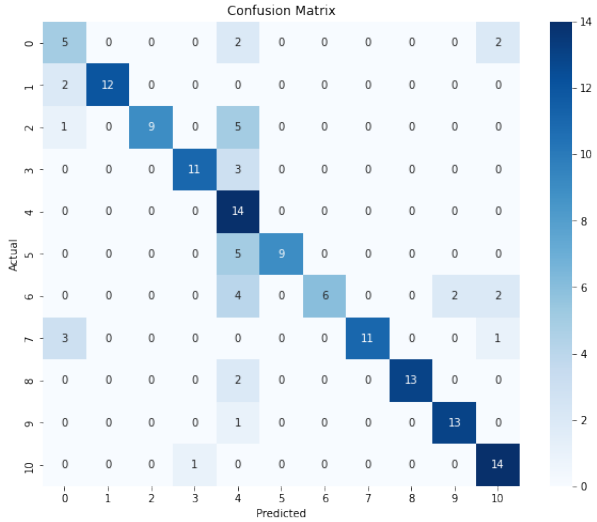

Fig. 4. Confusion matrix of the model after fine-tuning.

B. Phase 2: Unlearning

Now the model has been fine-tuned and trained on the selected 10 identities as well as a few random faces to act as unknown identities under the 0th identity, The work moves on to forgetting or unlearning a selected identity, in this case, number 5 from the DigiFace dataset [3].

To achieve this, the proposed method is to delete the images marked under identity 5 and populate identity 5 with grainy images, then retrain the model for a few epochs. It can be observed from the confusion matrix in Fig. 5 that the model has started to confuse images with the 5th identity's face as other identities which implies that the 5th identity is slowly being forgotten by the model.

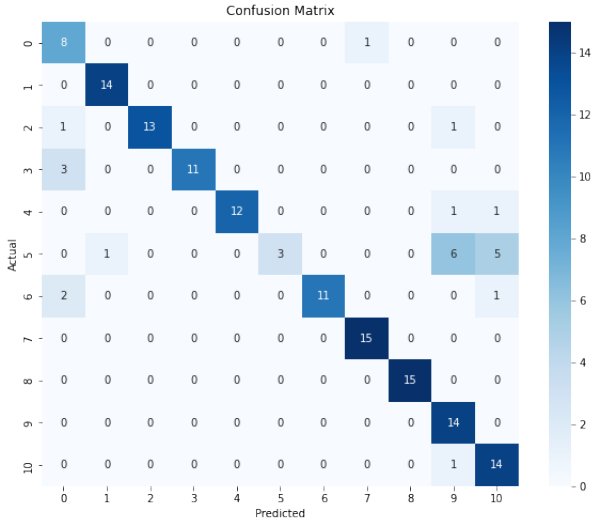

Fig. 5. Confusion matrix of the model after novel method unlearning.

For contrast, refer the included results of the comparison method mentioned above which simply relabels identity 5 as part of identity 0 in Fig. 6.

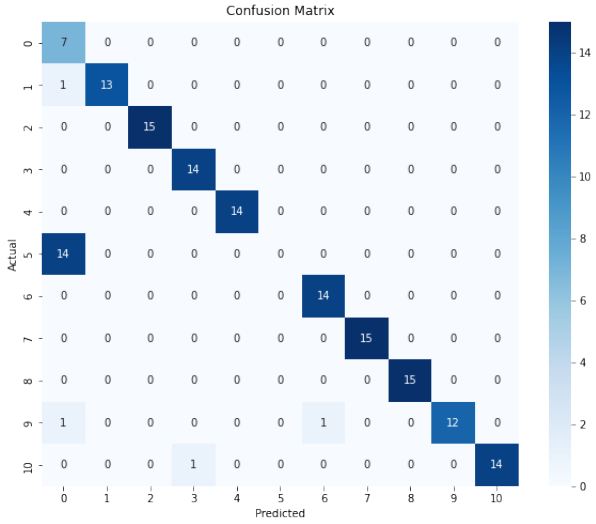

Fig. 6. Confusion matrix of the model after comparison method unlearning.

Conclusion

This paper proposed and evaluated a novel methodology for selective unlearning of faces from trained face recognition models. Experimental results on the DigiFace dataset [3] demonstrated the viability of the proposed approach.

By augmenting the training data with additional unlabeled faces and selectively removing target faces during retraining, the proposed method showed the promising ability to "forget" a chosen identity from the model. The forgetting process was reinforced through retraining with noisy samples replacing the

removed face images. This helped dissociate the face from the recognized identities without sacrificing overall accuracy.

In future work, the method could be evaluated on larger and more diverse datasets to verify scalability. Additional metrics like area under the ROC curve could provide finer-grained performance analysis. Expanding the approach to other model architectures like VGGFace and exploring alternatives to noisy samples are also directions for future improvement.

Overall, this research contributes toward privacy-respecting facial recognition technology. The ability to selectively and securely remove individuals' data gives users more control over their biometric information and strengthens ethical AI. With further refinement, this work could facilitate the right to be forgotten in real-world face recognition applications.

REFERENCES

[1] A. Majeed and S. O. Hwang, "When AI Meets Information Privacy: The Adversarial Role of AI in Data Sharing Scenario," in IEEE Access, vol. 11, pp. 76177-76195, 2023, doi: 10.1109/ACCESS.2023.3297646.

[2] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 815-823, doi: 10.1109/CVPR.2015.7298682.

[3] G. Bae, M. de La Gorce, T. Baltrusaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen, "DigiFace-1M: 1 Million Digital Face Images for Face Recognition," arXiv preprint arXiv:2210.02579, 2022.

[4] P. S. Kumar and R. Ranjith, "Real Time Personal Data Protection Using Image Processing," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-5, doi: 10.1109/ICCCNT51525.2021.9579719.

[5] G. Singh and A. K. Goel, "Face Detection and Recognition System using Digital Image Processing," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 348-352, doi: 10.1109/ICIMIA48430.2020.9074838.

[6] P. Yaswanthram and B. A. Sabarish, "Face Recognition Using Machine Learning Models - Comparative Analysis and impact of dimensionality reduction," 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC), Bengaluru, India, 2022, pp. 1-4, doi: 10.1109/ICAECC54045.2022.9716590.

[7] Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. 2023. Zero-Shot Machine Unlearning. Trans. Info. For. Sec. 18 (2023), 2345–2354. https://doi.org/10.1109/TIFS.2023.3265506

[8] Y. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," 2015 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 2015, pp. 463-480, doi: 10.1109/SP.2015.35.

[9] Huawei Lin, Jun Woo Chung, Yingjie Lao, and Weijie Zhao. 2023. Machine Unlearning in Gradient Boosting Decision Trees. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, NY, USA, 1374–1383. https://doi.org/10.1145/3580305.3599420

[10] L. Bourtoule et al., "Machine Unlearning," 2021 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2021, pp. 141-159, doi: 10.1109/SP40001.2021.00019.

[11] A. Krishnadas and S. Nithin, "A comparative study of machine learning and deep learning algorithms for recognizing facial emotions," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1506-1512, doi: 10.1109/ICESC51422.2021.9532745.

[12] A. Das and N. N, "Facial Expression Recognition System with Local Binary Features of Neural Network," 2023 International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, 2023, pp. 1-5, doi: 10.1109/ICDSNS58469.2023.10244983.

[13] V. C. R, V. Asha, B. Saju, S. N, T. R. Mrudhula Reddy and S. K. M, "Face Recognition and Identification Using Deep Learning," 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2023, pp. 1-5, doi: 10.1109/ICAECT57570.2023.10118154.

[14] W. Abbasi, P. Mori, A. Saracino and V. Frascolla, "Privacy vs Accuracy Trade-Off in Privacy Aware Face Recognition in Smart Systems," 2022 IEEE Symposium on Computers and Communications (ISCC), Rhodes, Greece, 2022, pp. 1-8, doi: 10.1109/ISCC55528.2022.9912465.

[15] Zhang, H., Nakamura, T., Isohara, T. *et al.* A Review on Machine Unlearning. SN COMPUT. SCI. 4, 337 (2023).

[16] K. Vikram and S. Padmavathi, "Facial parts detection using Viola Jones algorithm," 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2017, pp. 1-4, doi: 10.1109/ICACCS.2017.8014636.

[17] A. Stalin, A. Sha, A. S. Kumar, S. Nandakumar, and G. Gopakumar, "Face Recognition at varying angles from distant CCTV Footage using Siamese Architecture," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-6, doi: 10.1109/INCET54531.2022.9824723.