Bibliography

1. N. Balacheff, S. Ludvigsen, T. De Jong, A. Lazonder, S. A. Barnes, and L. Montandon, *Technology-Enhanced Learning*. Berlin, Germany: Springer, 2009.
2. J. Bourdeau and N. Balacheff, "Technology-Enhanced Learning: From Thesaurus and Dictionary to Ontology," *Technological and Social Environments for Interactive Learning*, pp. 1–33, 2014.
3. Instructure, "Edtech Top 40," Accessed: Nov. 7, 2024. [Online]. Available: https://www.instructure.com/edtech-top40
4. V. Singh and A. Thurman, "How many ways can we define online learning? A systematic literature review of definitions of online learning (1988-2018)," *Am. J. Distance Educ.*, vol. 33, no. 4, pp. 289–306, 2019.
5. L. Daniela, D. Kalniņa, and R. Strods, "An overview on effectiveness of technology enhanced learning (TEL)," *Int. J. Knowl. Soc. Res.*, vol. 8, no. 1, pp. 79–91, 2017.
6. G. Ghinea, W. Lin, S. R. Gulliver, and C. Timmerer, "Mulsemedia," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, no. 1s, pp. 1–23, 2014.
7. Z. Yuan, S. Chen, G. Ghinea, and G. M. Muntean, "User quality of experience of mulsemedia applications," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 1s, pp. 1–19, 2014.
8. E. B. Saleme, A. Covaci, G. Mesfin, C. A. Santos, and G. Ghinea, "Mulsemedia DIY: A survey of devices and a tutorial for building your own mulsemedia environment," *ACM Comput. Surv.*, vol. 52, no. 3, pp. 1–29, 2019.
9. S. Cairncross and M. Mannion, "Interactive multimedia and learning: Realizing the benefits," *Innov. Educ. Teach. Int.*, vol. 38, no. 2, pp. 156–164, 2001.
10. R. E. Mayer, "Multimedia learning," in *Psychology of Learning and Motivation*, vol. 41, Academic Press, 2002, pp. 85–139.
11. R. K. Kushwaha, M. K. Yadav, J. T. Sulaimon, and S. Ahmad, "Mulsemedia in Special Education: A Novel Teaching Approach for the Next Generation," *International Journal of Multidisciplinary Educational Research and Innovation*, vol. 1, no. 4, pp. 85–92, 2023.
12. R. Jain, "Quality of experience," *IEEE Multimedia*, vol. 11, no. 1, pp. 96–95, 2004.

13. Z. Yuan, G. Ghinea, and G. M. Muntean, "Quality of experience study for multiple sensorial media delivery," in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug. 2014, pp. 1142–1146.
14. S. Yu, A. Androsov, H. Yan, and Y. Chen, "Bridging computer and education sciences: a systematic review of automated emotion recognition in online learning environments," *Computers & Education*, vol. 105111, 2024.
15. J. X. Y. Lek and J. Teo, "Academic emotion classification using FER: A systematic review," *Human Behavior and Emerging Technologies*, vol. 2023, no. 1, pp. 9790005, 2023.
16. A. Mehrabian, *Communication without words*, in *Communication Theory*, Routledge, 2017, pp. 193-200.
17. P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.
18. P. Ekman and E. L. Rosenberg, Eds., *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
19. A. Baskar and T. G. Kumar, "Facial expression classification using machine learning approach: A review," in *Data Engineering and Intelligent Computing: Proceedings of IC3T 2016*, 2018, pp. 337-345.
20. M. Mohana and P. Subashini, "Facial Expression Recognition Using Machine Learning and Deep Learning Techniques: A Systematic Review," *SN Computer Science*, Springer, vol. 5, p. 432, 2024.
21. A. A. Pise, M. A. Alqahtani, P. Verma, P. K., D. A. Karras, and A. Halifa, "Methods for facial expression recognition with applications in challenging situations," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 9261438, 2022.
22. M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. A. Cheikh, M. Hijji, and J. J. Rodrigues, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817-840, 2023.
23. M. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," *Smart Learning Environments*, vol. 6, no. 1, pp. 1-20, 2019.

24. M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, and D. Moher, "Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement," *J. Clin. Epidemiol.* vol. 134, pp. 103–112, 2021.
25. R. E. Mayer, "Cognitive theory and the design of multimedia instruction: an example of the two- way street between cognition and instruction," *New Directions for Teaching and Learning*, vol. 2002, no. 89, pp. 55-71, 2002.
26. I. Ghergulescu and C. H. Muntean, "Measurement and analysis of learner's motivation in game-based e-learning," in *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, New York, NY: Springer New York, 2012, pp. 355-378.
27. V. Andonova, F. Reinoso-Carvalho, M. A. Jimenez Ramirez, and D. Carrasquilla, "Does multisensory stimulation with virtual reality (VR) and smell improve learning? An educational experience in recall and creativity," *Frontiers in Psychology*, vol. 14, pp. 1176697, 2023.
28. A. Covaci, E. B. Saleme, G. Mesfin, I. S. Comsa, R. Trestian, C. A. Santos, and G. Ghinea, "Multisensory 360 videos under varying resolution levels enhance presence," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 4, pp. 2093-2101, 2022.
29. A. Tijou, E. Richard, and P. Richard, "Using olfactive virtual environments for learning organic molecules," in *International Conference on Technologies for E-Learning and Digital Entertainment*, Berlin, Heidelberg: Springer Berlin Heidelberg, Apr. 2006, pp. 1223-1233.
30. L. F. Cuturi, G. Cappagli, N. Yiannoutsou, S. Price, and M. Gori, "Informing the design of a multisensory learning environment for elementary mathematics learning," *Journal on Multimodal User Interfaces*, pp. 1-17, 2022.
31. L. Zou, I. Tal, A. Covaci, E. Ibarrola, G. Ghinea, and G. M. Muntean, "Can multisensorial media improve learner experience?" in *Proceedings of the 8th ACM on Multimedia Systems Conference*, Jun. 2017, pp. 315-320.
32. I. Tal, L. Zou, M. Farren, and G. M. Muntean, "Improving learner experience, motivation and knowledge gain when using mulsemedia-based technology enhanced learning," in *Computer Supported Education: 12th International Conference, CSEDU*

2020, Virtual Event, May 2–4, 2020, Revised Selected Papers 12, Springer International Publishing, 2021, pp. 146-161.

33. A. A. Alkasasbeh and G. Ghinea, "Using olfactory media cues in e-learning–perspectives from an empirical investigation," *Multimedia Tools and Applications*, vol. 79, pp. 19265-19287, 2020.
34. I. Tal, L. Zou, A. Covaci, E. Ibarrola, M. Bratu, G. Ghinea, and G. M. Muntean, "Mulsemedia in telecommunication and networking education: a novel teaching approach that improves the learning process," *IEEE Communications Magazine*, vol. 57, no. 11, pp. 60-66, 2019.
35. T. Bi, R. Lyons, G. Fox, and G. M. Muntean, "Improving student learning satisfaction by using an innovative dash-based multiple sensorial media delivery solution," *IEEE Transactions on Multimedia*, vol. 23, pp. 3494-3505, 2020.
36. G. Mesfin, N. Hussain, E. Kani-Zabihi, A. Covaci, E. B. Saleme, and G. Ghinea, "QoE of cross-modally mapped Mulsemedia: an assessment using eye gaze and heart rate," *Multimedia Tools and Applications*, vol. 79, pp. 7987-8009, 2020.
37. A. Raheel, M. Majid, and S. M. Anwar, "DEAR-MULSEMEDIA: Dataset for emotion analysis and recognition in response to multiple sensorial media," *Information Fusion*, vol. 65, pp. 37-49, 2021.
38. A. Sun, Y. J. Li, Y. M. Huang, and Q. Li, "Using facial expression to detect emotion in e-learning system: A deep learning method," in *Emerging Technologies for Education: Second International Symposium, SETE 2017, Held in Conjunction with ICWL 2017, Cape Town, South Africa, September 20–22, 2017, Revised Selected Papers 2*, Springer International Publishing, 2017, pp. 446-455.
39. A. Pise, H. Vadapalli, and I. Sanders, "Facial emotion recognition using temporal relational network: an application to E-learning," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 26633-26653, 2022.
40. X. Zhu and Z. Chen, "Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network," *The Visual Computer*, vol. 36, no. 4, pp. 743-755, 2020.
41. A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132-2143, 2022.

42. K. P. Rao and M. C. S. Rao, "Recognition of learners' cognitive states using facial expressions in e-learning environments," *Journal of University of Shanghai for Science and Technology*, vol. 22, no. 12, pp. 93-103, 2020.
43. Y. Du, R. G. Crespo, and O. S. Martínez, "Human emotion recognition for enhanced performance evaluation in e-learning," *Progress in Artificial Intelligence*, vol. 12, no. 2, pp. 199-211, 2023.
44. B. E. Zakka and H. Vadapalli, "Detecting learning affect in e-learning platform using facial emotion expression," in *Proceedings of the 11th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2019)*, vol. 11, Springer International Publishing, 2021, pp. 217-225.
45. S. Park, K. Lee, J. A. Lim, H. Ko, T. Kim, J. I. Lee, and E. C. Lee, "Differences in facial expressions between spontaneous and posed smiles: Automated method by action units and three-dimensional facial landmarks," *Sensors*, vol. 20, no. 4, pp. 1199, 2020.
46. M. J. Lyons, "Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset," *arXiv preprint arXiv:2107.13998*, 2021.
47. D. Lundqvist, A. Flykt, and A. Öhman, "Karolinska directed emotional faces," *PsycTESTS Dataset*, vol. 91, p. 630, 1998.
48. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, pp. 94-101.
49. M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, Jul. 2005, pp. 5-pp.
50. R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807-813, 2010.
51. G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607-619, 2011.
52. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, ... and Y. Bengio, "Challenges in representation learning: A report on three machine learning

contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, Springer Berlin Heidelberg, 2013, pp. 117-124.

53. F. Benitez-Quiroz, C., R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562-5570.
54. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2017.
55. Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, ... and W. Zhang, "Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20922-20931.
56. A. Miolla, M. Cardaioli, and C. Scarpazza, "Padova Emotional Dataset of Facial Expressions (PEDFE): A unique dataset of genuine and posed emotional facial expressions," *Behavior Research Methods*, vol. 55, no. 5, pp. 2559-2574, 2023.
57. M. Mascaró-Oliver, R. Mas-Sansó, E. Amengual-Alcover, and M. F. Roig-Maimó, "UIBVFED-Mask: A dataset for comparing facial expressions with and without face masks," *Data*, vol. 8, no. 1, p. 17, 2023.
58. H. Aung, A. V. Bobkov, and N. L. Tun, "Face detection in real time live video using YOLO algorithm based on VGG16 convolutional neural network," in *2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, May 2021, pp. 697-702.
59. X. Li, S. Lai, and X. Qian, "DBCFace: Towards pure convolutional neural network face detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1792-1804, 2021.
60. S. Tao, Y. Li, Y. Huang, and X. Lan, "Face detection algorithm based on deep residual network," in *Journal of Physics: Conference Series*, vol. 1802, no. 3, p. 032142, Mar. 2021.
61. H. Yan, Y. Liu, X. Wang, M. Li, and H. Li, "A face detection method based on skin color features and AdaBoost algorithm," in *Journal of Physics: Conference Series*, vol. 1748, no. 4, p. 042015, 2021.

62. N. Zhang, J. Luo, and W. Gao, "Research on face detection technology based on MTCNN," in *2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, Sep. 2020, pp. 154-158.
63. W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-face: A real-time face detector," *The Visual Computer*, vol. 37, pp. 805-813, 2021.
64. M. H. Robin, M. M. U. Rahman, A. M. Taief, and Q. N. Eity, "Improvement of face and eye detection performance by using multi-task cascaded convolutional networks," in *2020 IEEE Region 10 Symposium (TENSYMP)*, Jun. 2020, pp. 977-980.
65. D. Garg, P. Goel, S. Pandya, A. Ganatra, and K. Kotecha, "A Deep Learning Approach for Face Detection Using YOLO," in *2018 IEEE PuneCon*, Nov. 2018, pp. 1–4.
66. F. Jiang, J. Zhang, L. Yan, Y. Xia, and S. Shan, "A three-category face detector with contextual information on finding tiny faces," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct. 2018, pp. 2680-2684.
67. A. Priadana and M. Habibi, "Face detection using Haar cascades to filter selfie face image on Instagram," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, Mar. 2019, pp. 6-9.
68. J. Deng, J. Guo, and S. Zafeiriou, "Single-stage joint face detection and alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0-0.
69. C. Cuimei, Q. Zhiliang, J. Nan, and W. Jianhua, "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers," in *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, Oct. 2017, pp. 483-487.
70. Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, "Detecting faces using region-based fully convolutional networks," *arXiv preprint arXiv:1709.05256*, 2017.
71. A. M. A. Hossen, R. A. A. Ogla, and M. M. Ali, "Face detection by using OpenCV's Viola-Jones algorithm based on coding eyes," *Iraqi Journal of Science*, vol. 58, no. 2A, pp. 735-745, 2017.
72. I. Kalinovskii and V. Spitsyn, "Compact convolutional neural network cascade for face detection," *arXiv preprint arXiv:1508.01292*, 2015.
73. B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," Advances in Face Detection and Facial Image Analysis, pp. 63-100, 2016

74. B. U. H. Sheikh and A. Zafar, "RRFMDS: Rapid real-time face mask detection system for effective COVID-19 monitoring," *SN Computer Science*, vol. 4, no. 3, p. 288, 2023.
75. O. A. Mohammed and J. M. Al-Tuwaijari, "Analysis of challenges and methods for face detection systems: A survey," *International Journal of Nonlinear Analysis and Applications*, vol. 13, no. 1, pp. 3997-4015, 2022.
76. J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade, "Feature-point tracking by optical flow discriminates subtle differences in facial expression," in *Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition*, Apr. 1998, pp. 396-401.
77. S. K. A. Kamarol, M. H. Jaward, H. Kälviäinen, J. Parkkinen, and R. Parthiban, "Joint facial expression recognition and intensity estimation based on weighted votes of image sequences," *Pattern Recognition Letters*, vol. 92, pp. 25-32, 2017.
78. C. Liu, K. Hirota, and Y. Dai, "Patch attention convolutional vision transformer for facial expression recognition with occlusion," *Information Sciences*, vol. 619, pp. 781-794, 2023.
79. K. Roshan, A. Zafar, and S. B. U. Haque, "Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system," *Computer Communications*, vol. 218, pp. 97-113, 2024.
80. Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz, "Measuring the performance of face localization systems," *Image and Vision Computing*, vol. 24, no. 8, pp. 882-893, 2006.
81. R. T. Hasan and A. B. Sallow, "Face Detection and Recognition Using OpenCV," *Journal of Soft Computing and Data Mining*, vol. 2, no. 2, pp. 86–97, 2021.
82. A. Sharifara, M. S. M. Rahim, and Y. Anisi, "A General Review of Human Face Detection Including a Study of Neural Networks and Haar Feature-Based Cascade Classifier in Face Detection," in *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, Aug. 2014, pp. 73–78.
83. W. Yang and J. Ziachun, "Real-time face detection based on YOLO," in *Proc. 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII)*, Jul. 2018, pp. 221-224.

84. H. Y. Patil, S. V. Bharambe, A. G. Kothari, and K. M. Bhurchandi, "Face localization and its implementation on embedded platform," in *Proc. 3rd IEEE International Advance Computing Conference (IACC)*, Feb. 2013, pp. 741-745.
85. D. Luo, G. Wen, D. Li, Y. Hu, and E. Huan, "Deep-learning-based face detection using iterative bounding-box regression," *Multimedia Tools and Applications*, vol. 77, pp. 24663-24680, 2018.
86. L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, pp. 413-426, Berlin Heidelberg: Springer, 2008.
87. R. Lim and M. J. T. Reinders, "Facial landmark detection using a Gabor filter representation and a genetic search algorithm," in *Proc. Annual Conference of the Advanced School for Computing and Imaging (ASCI 2000)*, Lommel, pp. 72-78.
88. B. Johnston and P. D. Chazal, "A review of image-based automatic facial landmark identification techniques," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 86, 2018.
89. K. T. Talele and S. Kadam, "Face detection and geometric face normalization," in *Proc. TENCON 2009–2009 IEEE Region 10 Conference*, Jan. 2009, pp. 1-6.
90. S. T. Chaudhari and A. Kale, "Face normalization: Enhancing face recognition," in *Proc. 3rd International Conference on Emerging Trends in Engineering and Technology*, 2010, pp. 520-525.
91. V. Pali, S. Goswami, and L. P. Bhaiya, "An extensive survey on feature extraction techniques for facial image processing," in *Proc. 2014 International Conference on Computational Intelligence and Communication Networks*, Nov. 2014, pp. 142-148.
92. A. Mikołajczyk and M. Grochowski, "Data Augmentation for Improving Deep Learning in Image Classification Problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 2018, pp. 117–122.
93. X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion classification with data augmentation using generative adversarial networks," in *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III*, pp. 349-360, Springer International Publishing, 2018.

94. E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113-123, 2019.
95. M. Jaderberg et al., "Population based training of neural networks," *arXiv preprint arXiv:1711.09846*, 2017.
96. M. M. Htay, "Feature Extraction and Classification Methods of Facial Expression: A Survey," *Computer Science and Information Technologies*, vol. 2, no. 1, pp. 26–32, 2021.
97. L. Zhang, H. Ai, S. Xin, C. Huang, S. Tsukiji, and S. Lao, "Robust face alignment based on local texture classifiers," in *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. II-354, Sep. 2005.
98. R. Sharma and M. S. Patterh, "Face recognition using face alignment and PCA techniques: a literature survey," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 17, no. 4, pp. 17-30, 2015.
99. A. Boughida, M. N. Kouahla, and Y. Lafifi, "A novel approach for facial expression recognition based on Gabor filters and genetic algorithm," *Evolving Systems*, vol. 13, no. 2, pp. 331-345, 2022.
100. N. Rathee, A. Vaish, and S. Gupta, "Adaptive system to learn and recognize emotional state of mind," in *Proc. 2016 International Conference on Computing, Communication and Automation (ICCCA)*, Apr. 2016, pp. 32-36.
101. H. I. Dino and M. B. Abdulrazzaq, "Facial expression classification based on SVM, KNN and MLP classifiers," in *Proc. 2019 International Conference on Advanced Science and Engineering (ICOASE)*, Apr. 2019, pp. 70-75.
102. H. Lu and F. Yang, "Active shape model and its application to face alignment," in *Subspace Methods for Pattern Recognition in Intelligent Environment*, Berlin Heidelberg: Springer, pp. 1-31, 2014.
103. H. Sikkandar and R. Thiyagarajan, "Deep learning based facial expression recognition using improved Cat Swarm Optimization," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 3037-3053, 2021.
104. I. Dagher, E. Dahdah, and M. Al Shakik, "Facial expression recognition using three-stage support vector machines," *Visual Computing for Industry, Biomedicine, and Art*, vol. 2, pp. 1-9, 2019.

105. A. Boughida, M. N. Kouahla, and Y. Lafifi, "A novel approach for facial expression recognition based on Gabor filters and genetic algorithm," Evolving Systems, vol. 13, no. 2, pp. 331-345, 2022.
106. K. S. Yadav and J. Singha, "Facial expression recognition using modified Viola-John's algorithm and KNN classifier," *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 13089-13107, 2020
107. M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local Learning with Deep and Handcrafted Features for Facial Expression Recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.
108. U. Ayvaz, H. Gürüler, and M. O. Devrim, "Use of facial emotion recognition in e-learning systems," 2017.
109. S. Gupta, "Facial emotion recognition in real-time and static images," in *Proc. 2nd International Conference on Inventive Systems and Control (ICISC)*, Jan. 2018, pp. 553-560.
110. N. Perveen, N. Ahmad, M. A. Q. B. Khan, R. Khalid, and S. Qadri, "Facial expression recognition through machine learning," *International Journal of Scientific & Technology Research*, vol. 5, no. 3, 2016.
111. T. Brosch, G. Pourtois, and D. Sander, "The Perception and Categorisation of Emotional Stimuli: A Review," *Cognition and Emotion*, pp. 76–108, 2010.
112. S. Anwar and M. Milanova, "Real time face expression recognition of children with autism," *International Academy of Engineering and Medical Research*, vol. 1, no. 1, pp. 1-8, 2016.
113. E. M. Bouhabba, A. A. Shafie, and R. Akmeliawati, "Support vector machine for face emotion detection on real-time basis," in *Proc. 4th International Conference on Mechatronics (ICOM)*, May 2011, pp. 1-6.
114. M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, 2021.
115. R. Qayyum *et al.*, "Android based emotion detection using convolutional neural networks," in *Proc. 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Mar. 2021, pp. 360-365.

116. I. K. Choi, H. E. Ahn, and J. Yoo, "Facial expression classification using deep convolutional neural network," *Journal of Electrical Engineering and Technology*, vol. 13, no. 1, pp. 485-492, 2018.
117. E. Pranav, S. Kamal, C. S. Chandran, and M. H. Supriya, "Facial emotion recognition using deep convolutional neural network," in *Proc. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2020, pp. 317-320.
118. T. S. Gunawan *et al.*, "Development of video-based emotion recognition using deep learning with Google Colab," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 5, pp. 2463-2471, 2020.
119. R. Jadhav, J. Bhuke, and N. Patil, "Facial Emotion Detection using Convolutional Neural Network," *International Research Journal of Engineering and Technology*, 2019.
120. P. A. Riyantoko and K. M. Hindrayani, "Facial emotion detection using Haar-cascade classifier and convolutional neural networks," *Journal of Physics: Conference Series*, vol. 1844, no. 1, p. 012004, Mar. 2021.
121. P. Babajee, G. Suddul, S. Armoogum, and R. Foogooa, "Identifying human emotions from facial expressions with deep learning," in *Proc. 2020 Zooming Innovation in Consumer Technologies Conference (ZINC)*, May 2020, pp. 36-39.
122. A. Jaiswal, A. K. Raju, and S. Deb, "Facial emotion detection using deep learning," in *Proc. 2020 International Conference for Emerging Technology (INCET)*, Jun. 2020, pp. 1-5.
123. A. Lopez-Rincon, "Emotion recognition using facial expressions in children using the NAO Robot," in *Proc. 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Feb. 2019, pp. 146-153.
124. I. Lasri, A. R. Solh, and M. El Belkacemi, "Facial emotion recognition of students using convolutional neural network," in *Proc. 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, Oct. 2019, pp. 1-6.
125. W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Facial emotion recognition from videos using deep convolutional neural networks," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, pp. 14-19, 2019.

126. A. Verma, P. Singh, and J. S. R. Alex, "Modified convolutional neural network architecture analysis for facial emotion recognition," in *Proc. 2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Jun. 2019, pp. 169-173.
127. D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognition Letters*, vol. 120, pp. 69-74, 2019.
128. O. El Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El Kafi, and A. El Hore, "Emotion recognition in e-learning systems," in *Proc. 2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, May 2018, pp. 1-6.
129. V. Tümen, Ö. F. Söylemez, and B. Ergen, "Facial emotion recognition on a dataset using convolutional neural network," in *Proc. 2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Sep. 2017, pp. 1-5.
130. G. R. Kumar, R. K. Kumar, and G. Sanyal, "Facial emotion analysis using deep convolution neural network," in *Proc. 2017 International Conference on Signal Processing and Communication (ICSPC)*, Jul. 2017, pp. 369-374
131. L. Yang *et al.*, "A novel feature separation model exchange-GAN for facial expression recognition," *Knowledge-Based Systems*, vol. 204, p. 106217, 2020.
132. L. Sun, Z. Lian, B. Liu, and J. Tao, "MAE-DFER: Efficient Masked Autoencoder for Self-Supervised Dynamic Facial Expression Recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, Oct. 2023, pp. 6110–6121
133. Y. Ming, H. Qian, and G. Guangyuan, "CNN-LSTM facial expression recognition method fused with two-layer attention mechanism," *Computational Intelligence and Neuroscience*, vol. 2022, p. 102384, 2022.
134. X. Zhang, F. Zhang, and C. Xu, "Joint expression synthesis and representation learning for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1681-1695, 2021
135. G. Ali *et al.*, "Artificial neural network-based ensemble approach for multicultural facial expressions analysis," *IEEE Access*, vol. 8, pp. 134950-134963, 2020.
136. M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, "Flepnet: feature level ensemble parallel network for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2058-2070, 2022

137. Y. Tang, X. Zhang, X. Hu, S. Wang, and H. Wang, "Facial expression recognition using frequency neural network," *IEEE Transactions on Image Processing*, vol. 30, pp. 444-457, 2020
138. Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-MobileNet: An approach of facial expression recognition," *Alexandria Engineering Journal*, vol. 61, no. 6, pp. 4435-4444, 2022
139. M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomedical Signal Processing and Control*, vol. 78, p. 103970, 2022
140. A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1-10.
141. G. del Castillo Torres, M. F. Roig-Maimó, M. Mascaró-Oliver, E. Amengual-Alcover, and R. Mas-Sansó, "Understanding how CNNs recognize facial expressions: a case study with LIME and CEM," *Sensors*, vol. 23, no. 1, p. 131, 2022.
142. C. Manresa-Yee, S. Ramis, and J. M. Buades, "Analysis of Gender Differences in Facial Expression Recognition Based on Deep Learning Using Explainable Artificial Intelligence," unpublished, 2023.
143. M. Deramgozin, S. Jovanovic, H. Rabah, and N. Ramzan, "A hybrid explainable AI framework applied to global and local facial expression recognition," in *Proc. 2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, Aug. 2021, pp. 1-5.
144. A. A. Kandeel, H. M. Abbas, and H. S. Hassanein, "Explainable model selection of a convolutional neural network for driver's facial emotion identification," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI*, Springer, pp. 699-713, 2021.
145. S. Ramis Guarinos, C. Manresa Yee, J. M. Buades Rubio, and F. X. Gaya-Morey, "Explainable Facial Expression Recognition for People with Intellectual Disabilities," in *Proc. XXIII International Conference on Human Computer Interaction*, Sep. 2023, pp. 1-7.
146. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, ... and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies,

opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020.

147. A. Pise, H. Vadapalli, and I. Sanders, "Estimation of learning effects experienced by learners: an approach using relational reasoning and adaptive mapping," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 8808283, 2022.
148. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001.
149. E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.
150. S. Kumar and D. S. Vidyadharan, "Face detection and localization of facial features in still and video images," in *Proceedings of the 2008 First International Conference on Emerging Trends in Engineering and Technology*, pp. 95–99, July 2008.
151. L. Zhang and P. Lenders, "Knowledge-based eye detection for human face recognition," in *KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings (Cat. No. 00TH8516)*, vol. 1, pp. 117-120, Aug. 2000.
152. P. Bose and S. Bandyopadhyay, "Human Face and Facial Parts Detection using Template Matching Technique," *International Journal of Engineering and Advanced Technology (IJE)*, vol. 9, no. 4, 2020.
153. H. A. Hosni Mahmoud and H. A. Mengash, "A novel technique for automated concealed face detection in surveillance videos," *Personal and Ubiquitous Computing*, vol. 25, pp. 129-140, 2021.
154. S. Soleymani, B. Chaudhary, A. Dabouei, J. Dawson, and N. M. Nasrabadi, "Differential morphed face detection using deep siamese networks," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI*, Cham: Springer International Publishing, Feb. 2021, pp. 560-572.
155. N. Dalal, "Histogram of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886-893.

156. Y. Wu and X. Ai, "Face detection in color images using AdaBoost algorithm based on skin color information," in *Proceedings of the First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008)*, 2008, pp. 339-342.
157. R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz, "A novel database of children's spontaneous facial expressions (LIRIS-CSE)," *Image and Vision Computing*, vol. 83, pp. 61-69, 2019.
158. F. Marini and B. Walczak, "Particle swarm optimization (PSO): A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 153-165, 2015.
159. S. Ginosar, K. Rakelly, S. Sachs, B. Yin, and A.A. Efros, "A century of portraits: A visual historical record of American high school yearbooks," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2015, pp. 1–7.
160. S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2009.
161. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248-255.
162. S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5525-5533.
163. M. J. Flores, J. M. Armingol, and A. de la Escalera, "Real-time warning system for driver drowsiness detection using visual information," *J. Intell. Robotic Syst.*, vol. 59, pp. 103-125, 2010.
164. B. Fatima, A. R. Shahid, S. Ziauddin, A. A. Safi, and H. Ramzan, "Driver fatigue detection using Viola-Jones and principal component analysis," *Appl. Artif. Intell.*, vol. 34, no. 6, pp. 456–483, 2020.
165. A. Dasgupta, A. George, S. L. Happy, and A. A. Routray, "Vision-based system for monitoring the loss of attention in automotive drivers," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1825–1838, 2013.
166. M. Asim, Z. Ming, and M. Y. Javed, "CNN-based spatiotemporal feature extraction for face anti-spoofing," in *2017 2nd Int. Conf. Image, Vision, and Computing (ICIVC)*, IEEE, 2017, pp. 234–238.

167. A. A. Elngar, M. Arafa, A. Fathy, B. Moustafa, O. Mahmoud, M. Shaban, and N. Fawzy, "Image classification based on CNN: A survey," *J. Cybersecurity Inf. Manag.*, vol. 6, no. 1, pp. 18–50, 2021.
168. T. H. S. Li, P. H. Kuo, T. N. Tsai, and P. C. Luan, "CNN and LSTM based facial expression analysis model for a humanoid robot," *IEEE Access*, vol. 7, pp. 93998–94011, 2019.
169. D. Kollias and S. Zafeiriou, "Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 595–606, 2020.
170. K. O'Shea and R. Nash, "An introduction to convolutional neural networks," arXiv preprint arXiv:1511.08458, 2015.
171. V. Mayya, R. M. Pai, and M. M. Pai, "Automatic facial expression recognition using DCNN," *Procedia Comput. Sci.*, vol. 93, pp. 453–461, 2016.
172. D. A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," arXiv preprint arXiv:1511.07289, 2015.
173. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
174. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
175. R. C. Staudemeyer and E. R. Morris, "Understanding LSTM — A tutorial into long short-term memory recurrent neural networks," arXiv preprint arXiv:1909.09586, 2019.
176. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Machine Learning*, PMLR, 2015, pp. 448–456.
177. S. Dereich and A. Jentzen, "Convergence Rates for the Adam Optimizer," *arXiv preprint arXiv:2407.21078*, 2024.
178. H. S. Lee and B. Y. Kang, "Continuous Emotion Estimation of Facial Expressions on JAFFE and CK+ Datasets for Human–Robot Interaction," *Intelligent Service Robotics*, vol. 13, pp. 15–27, 2020.

179. L. Lu, Y. Zhou, K. Panetta, and S. Agaian, "Comparative study of histogram equalization algorithms for image enhancement," in *Proc. Mobile Multimedia/Image Processing, Security, and Applications*, vol. 7708, SPIE, 2010, pp. 337–347.
180. M. Zhang and B. K. Gunturk, "Multiresolution bilateral filtering for image denoising," *IEEE Trans. Image Process.*, vol. 17, no. 12, pp. 2324–2333, 2008.
181. C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
182. M. S. Bilkhu, S. Gupta, and V. K. Srivastava, "Emotion classification from facial expressions using cascaded regression trees and SVM," in *Computational Intelligence: Theories, Applications and Future Directions - Volume II*, Singapore: Springer, 2019, pp. 585–594.
183. M. Peter, J. L. Minoi, and I. H. M. Hipiny, "3D face recognition using kernel based PCA approach," in *Computational Science and Technology*, Singapore: Springer, 2019, pp. 77–86.
184. K. Shan, J. Guo, W. You, D. Lu, and R. Bie, "Automatic facial expression recognition based on a deep convolutional-neural-network structure," in *2017 IEEE 15th Int. Conf. Software Engineering Research, Management, and Applications (SERA)*, IEEE, 2017, pp. 123–128.
185. I. Buciu and I. Pitas, "Application of non-negative and local nonnegative matrix factorization to facial expression recognition," in *Proc. 17th Int. Conf. Pattern Recognition, 2004. ICPR 2004*, vol. 1, IEEE, 2004, pp. 288–291.
186. D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial Emotion Recognition (FER) Using Convolutional Neural Network (CNN)," *Procedia Computer Science*, vol. 235, pp. 2079–2089, 2024.
187. J. Wang and L. Yin, "Static topographic modeling for facial expression recognition and analysis," *Comput. Vis. Image Underst.*, vol. 108, no. 1–2, pp. 19–34, 2007.
188. H. Tan, Y. Zhang, H. Cheri, Y. Zhao, and W. Wang, "Person-independent expression recognition based on person-similarity weighted expression feature," *J. Syst. Eng. Electron.*, vol. 21, no. 1, pp. 118–126, 2010.
189. M. Sert and N. Aksoy, "Recognizing facial expressions of emotion using action unit specific decision thresholds," in *Proc. 2nd Workshop Advancements in Social Signal Processing for Multimodal Interaction*, Association for Computing Machinery, 2016, pp. 16–21.

190. R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that," arXiv preprint arXiv:1611.07450, 2016.
191. I. O. Lopes, D. Zou, I. H. Abdulqadder, F. A. Ruambo, B. Yuan, and H. Jin, "Effective Network Intrusion Detection via Representation Learning: A Denoising AutoEncoder Approach," *Computer Communications*, vol. 194, pp. 55–65, 2022.
192. U. Michelucci, "An introduction to autoencoders," *arXiv preprint* arXiv:2201.03898, 2022.
193. J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in *2018 IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, IEEE, pp. 415–419, 2018.
194. P. K. Mallick, S. H. Ryu, S. K. Satapathy, S. Mishra, G. N. Nguyen, and P. Tiwari, "Brain MRI image classification for cancer detection using deep wavelet autoencoder-based deep neural network," *IEEE Access*, vol. 7, pp. 46278–46287, 2019.
195. S. Zhou, Z. Xue, and P. Du, "Semisupervised stacked autoencoder with cotraining for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3813–3826, 2019.
196. P. Liang, W. Shi, and X. Zhang, "Remote sensing image classification based on stacked denoising autoencoder," *Remote Sens.*, vol. 10, no. 1, p. 16, 2017.
197. M. Mohana and P. Subashini, "Emotion Recognition using Autoencoders: A Systematic Review," in *2023 Int. Conf. Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pp. 438–443, 2023.
198. S. Chatterjee, A. K. Das, J. Nayak, and D. Pelusi, "Improving Facial Emotion Recognition Using Residual Autoencoder Coupled Affinity-Based Overlapping Reduction," *Mathematics*, vol. 10, no. 3, p. 406, 2022.
199. D. Lakshmi and R. Ponnusamy, "Facial emotion recognition using modified HOG and LBP features with deep stacked autoencoders," *Microprocess. Microsyst.*, vol. 82, p. 103834, 2021.
200. Z. Sun, H. Zhang, J. Bai, M. Liu, and Z. Hu, "A discriminatively deep fusion approach with improved conditional GAN (im-cGAN) for facial expression recognition," *Pattern Recognit.*, vol. 135, p. 109157, 2023.
201. M. Mohana, P. Subashini, and D. Shukla, "Revisiting Face Detection: Supercharging Viola-Jones with Particle Swarm Optimization for Enhanced Performance," *J. Intell. Fuzzy Syst.*, vol. 46, no. 4, pp. 10727–10741, 2024.

202. M. Lyons, M. Kamachi, and J. Gyoba, "The Japanese Female Facial Expression (JAFFE) Dataset," *Zenodo*, 1998.
203. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th Int. Conf., ICONIP 2013, Daegu, Korea*, 2013, pp. 117–124.
204. M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 593–600, 2007.
205. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint* arXiv:1409.1556, 2014.
206. X. Lei, H. Pan, and X. Huang, "A dilated CNN model for image classification," *IEEE Access*, vol. 7, pp. 124087–124095, 2019.
207. M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep feature learning for medical image analysis with convolutional autoencoder neural network," *IEEE Trans. Big Data*, vol. 7, no. 4, pp. 750–758, 2017.
208. M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *arXiv preprint* arXiv:1812.05069, 2018.
209. S. Chen and W. Guo, "Auto-Encoders in Deep Learning—A Review with New Perspectives," *Mathematics*, vol. 11, no. 8, p. 1777, 2023.
210. N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
211. A. Ng, "Sparse autoencoder," *CS294A Lecture Notes*, vol. 72, pp. 1–19, 2011.
212. A. Asperti, "Sparsity in variational autoencoders," *arXiv preprint* arXiv:1812.07238, 2018.
213. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint* arXiv:1412.6980, 2014.
214. J. Zhou, X. Jia, L. Shen, Z. Wen, and Z. Ming, "Improved softmax loss for deep learning- based face and expression recognition," *Cogn. Comput. Syst.*, vol. 1, no. 4, pp. 97–102, 2019.
215. A. Gulli and S. Pal, *Deep Learning with Keras*. Packt Publishing Ltd., 2017.

216. D. H. Lee and J. H. Yoo, "CNN Learning Strategy for Recognizing Facial Expressions," *IEEE Access*, 2023.
217. C. T. Yen and K. H. Li, "Discussions of different deep transfer learning models for emotion recognitions," *IEEE Access*, vol. 10, pp. 102860–102875, 2022.
218. Y. Li, Y. Lu, B. Chen, Z. Zhang, J. Li, G. Lu, and D. Zhang, "Learning Informative and Discriminative Features for Facial Expression Recognition in the Wild," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3178–3189, 2021.
219. J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *2015 IEEE Int. Conf. Imaging Systems and Techniques (IST)*, IEEE, pp. 1–6, 2015.
220. P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on FER-2013," in *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications*, 2018, pp. 1–16.
221. M. Regina, M. S. Josephine, and V. Jeyabalraja, "Performance Comparisons Of Facial Expression Recognition In Jaffe, CK+ And ISED Data Base Using Neural Network," *Webology*, vol. 19, no. 2, 2022.
222. P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1805–1812, 2014.
223. N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, 2018.
224. J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018)*, IEEE, pp. 302–309, 2018.
225. H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2168–2177, 2018.
226. J. H. Kim, B. G. Kim, P. P. Roy, and D. M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE Access*, vol. 7, pp. 41273–41285, 2019.

227. C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.
228. W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognit.*, vol. 45, no. 1, pp. 80–91, 2012.
229. T. Jabid, M. H. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI J.*, vol. 32, no. 5, pp. 784–794, 2010.
230. X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Tech. Rev.*, vol. 32, no. 5, pp. 347–355, 2015.
231. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.
232. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 618–626, 2017.
233. X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, 2016.
234. N. Guan, J. Song, and D. Li, "On the Advantages of Computer Multimedia-aided English Teaching," *Procedia Computer Science*, vol. 131, pp. 727–732, 2018.
235. G. Mesfin, E. B. Saleme, O. A. Ademoye, E. Kani-Zabihi, C. A. Santos, and G. Ghinea, "Less Is (Just as Good as) More: An Investigation of Odor Intensity and Hedonic Valence in Mulsemedia QoE Using Heart Rate and Eye Tracking," *IEEE Transactions on Multimedia*, vol. 23, pp. 1095–1105, 2020.
236. A. Y. Kapi, N. Osman, R. Z. Ramli, and J. M. Taib, "Multimedia education tools for effective teaching and learning," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 2-8, pp. 143-146, 2017.
237. E. Alemdag and K. Cagiltay, "A systematic review of eye tracking research on multimedia learning," *Computers & Education*, vol. 125, pp. 413–428, 2018.
238. T. Neo and M. Neo, "Classroom innovation: engaging students in interactive multimedia learning," *Campus-Wide Information Systems*, vol. 21, no. 3, pp. 118–124, 2004.

239. M. D. Abdulrahaman *et al.*, “Multimedia tools in the teaching and learning processes: A systematic review,” *Heliyon*, vol. 6, no. 11, p. e05312, 2020.
240. P. Kumar, C. Saxena, and H. Baber, "Learner-content interaction in e-learning—the moderating role of perceived harm of COVID-19 in assessing the satisfaction of learners," *Smart Learn. Environ.*, vol. 8, pp. 1–15, 2021.
241. Z. Yuan, G.-M. Muntean, G. Ghinea, and S. Chen, “User Quality of Experience of Mulsemedia Applications,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, no. 1s, pp. 1–19, 2014
242. A. Raheel, M. Majid, and S. M. Anwar, “DEAR-MULSEMEDIA: Dataset for emotion analysis and recognition in response to multiple sensorial media,” *Information Fusion*, vol. 65, pp. 37–49, 2020.
243. E. Duval, M. Sharples, and R. Sutherland, “*Technology Enhanced Learning”*. New York: Springer, 2017.
244. H. J. Broadbent et al., "Incidental Learning in a Multisensory Environment Across Childhood," *Developmental Science*, vol. 21, no. 2, p. e12554, 2018.
245. L. Jalal and M. Murroni, "Enhancing TV Broadcasting Services: A Survey on Mulsemedia Quality of Experience," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, June 2017, pp. 1–7.
246. T. Bi, G. Ghinea, G.-M. Muntean, and F. Silva, “Improving Learning Experience by Employing DASH-Based Mulsemedia Delivery,” Aug. 2018.
247. S. Möller and A. Raake, “Quality of Experience: Terminology, Methods and Applications,” *PIK - Praxis der Informationsverarbeitung und Kommunikation*, vol. 37, no. 4, 2014.
248. T. Ebrahimi, "Quality of Multimedia Experience: Past, Present and Future," in *Proc. 17th ACM Int. Conf. on Multimedia*, 2009, pp. 3-4.
249. T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is Not Enough!" in *2011 Third International Workshop on Quality of Multimedia Experience*, IEEE, 2011, pp. 131-136.
250. E. B. Saleme, C. A. Santos, and G. Ghinea, "A Mulsemedia Framework for Delivering Sensory Effects to Heterogeneous Systems," *Multimedia Systems*, vol. 25, pp. 421–447, 2019.

251. How to Use the System Usability Scale (SUS) to Evaluate the Usability of Your Website," Usability Geek, Aug. 2, 2016. [Online]. Available: https://usabilitygeek.com/how-to-use-the-system-usability-scale-sus-to-evaluate-the-usability-of-your-website/. [Accessed: Nov. 3, 2024]
252. R. J. Marzano and J. S. Kendall, Eds., *The New Taxonomy of Educational Objectives*. Thousand Oaks, CA: Corwin Press, 2006.
253. N. Athanassiou, J. M. McNett, and C. Harvey, "Critical Thinking in the Management Classroom: Bloom's Taxonomy as a Learning Tool," *Journal of Management Education*, vol. 27, no. 5, pp. 533-555, 2003.
254. C. A. Mertler, R. A. Vannatta, and K. N. LaVenia, *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation*. New York: Routledge, 2021.
255. S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimed. Tools Appl.*, vol. 82, pp. 11365–11394, 2023.

Annexure I

Rosemary Pre- and Post-Assessment Questionnaire Based on Bloom's Taxonomy

1. What is the scientific name of the rosemary plant? [Knowledge (Remembering)]

A. Rosmarinus officinalis
B. Lavandula angustifolia
C. Mentha piperita
D. Thymus vulgaris

2. Which family of herbs does rosemary belong to? [Knowledge (Remembering)]

A. Basil family
B. Mint family
C. Parsley family
D. Thyme family

3. Rosemary plant grows up to --------- [Knowledge (Remembering)]

A. 5-6 feet
B. 3-4 feet
C. 4-5 feet
D. 3.5 to 5 feet

4. What are the medicinal benefits of rosemary? [Comprehension (Understanding)]

A. Cell Renewal
B. Boosting the immune system
C. Improving digestion
D. All the Above

5. Rosemary stimulates ---------- system. [Comprehension (Understanding)]

A. Nervous
B. Memory
C. Nervous and Memory
D. Nervous and Digestive

6. Where did the name Rosemary come from? [Knowledge (Remembering)]

A. Greek
B. Latin
C. French
D. None of the above

7. Is 'Dew of the sea' the meaning of the Latin word for rosemary? [Comprehension (Understanding)]

A. Yes
B. No

8. Which color flowers are typically seen on rosemary plants? [Knowledge (Remembering)]
    A. Red
    B. Pale Yellow
    C. Purple
    D. White

9. Which of the following options accurately describes the relationship between rosemary oil and essential oil? [Application (Applying)]
    A. Rosemary oil is an essential oil derived exclusively from rosemary plants.
    B. Rosemary oil is a type of essential oil commonly used in aromatherapy.
    C. Rosemary oil is not considered an essential oil but shares some similar properties.
    D. Rosemary oil is a separate category of oil and should not be confused with essential oils.

10. Which of the following options best describes the concentration of rosemary oil? [Application (Applying)]
    A. High concentration
    B. Low concentration
    C. Moderate concentration
    D. **Concentration varies depending on the brand or source**.

Thunder and Lightning Pre- and Post-Assessment Questionnaire Based on Bloom's Taxonomy

1. What is the main cause of thunder during a thunderstorm? [Knowledge (Remembering)]
    A. Rainfall
    B. **Lightning**
    C. Wind
    D. Hailstorms

2. What is lightning? [Knowledge (Remembering)]
    A. Magnetic Spark Through Air
    B. Water Flow Through Air
    C. **Giant Spark of Electricity**
    D. All the above

3. What temperature can the air around a lightning strike reach? [Knowledge (Remembering)]
    A. 5,000°C (9,000°F)
    B. 10,000°C (18,000°F)
    C. 20,000°C (36,000°F)
    D. **30,000°C (54,000°F)**

4. How far can the sound of thunder be heard from a lightning strike? [Knowledge (Remembering)]

A. 5 miles B. **10 miles** C. 15 miles D. 20 miles

5. Top Part of the clouds Develops __ and the bottom part of the clouds develops _____during Thunderstorms and Lightning. [Comprehension (Understanding)]

A. Negative charge, Negative charge
B. Positive charge, Positive charge
C. Negative charge, Positive charge
D. **Positive charge, Negative charge**

6. What causes the separation of positive and negative charges in a thunderstorm cloud? [Comprehension (Understanding)]

A. **Collision of ice crystals**
B. Rainfall
C. Wind patterns
D. Temperature changes

7. What is the speed of lights? [Knowledge (Remembering)]

A. **299792458 m/s** B. 399792458 m/s C. 299792459 m/s D. 409792459 m/s

8. Why do we see lightning before we hear the thunder? [Comprehension (Understanding)]

A. **Lightning travels faster than sound.**
B. Thunder travels faster than light.
C. Lightning and thunder have the same speed.
D. Lightning is brighter than thunder.

9. What is the flow of negative charge calling those rushes toward the Earth? [Knowledge (Remembering)]

A. **Stepped leader.**
B. Proton charge
C. Return stroke.
D. Lightning bolt

10. What factor affects the speed of thunder? [Comprehension (Understanding)]

A. The color of the lightning
B. The size of the thunderstorm cloud
C. **The temperature of the air**
D. The distance between the lightning strike and the observer

Rosemary Informational Recall: Pre- and Post-Test Assessment

	Group	Test_Type_Ros	CR_Q1	CR_Q2	CR_Q3	CR_Q4	CR_Q5	CR_Q6	CR_Q7	CR_Q8	CR_Q9	CR_Q10	Mean_CG_Ros_Pre	Mean_CG_Ros_post	Mean_EG_Ros_Pre	Mean_EG_Ros_Pos
1	11	1	1	1	0	1	0	1	1	0	0	1	60	60	50	80
2	11	1	1	0	0	1	0	0	1	0	0	0	30	70	50	70
3	11	1	1	0	0	1	0	1	0	0	0	1	40	80	70	70
4	11	1	1	0	0	1	0	1	0	0	0	0	30	60	30	80
5	11	1	1	0	0	1	1	1	0	1	0	0	50	60	40	80
6	11	1	1	1	0	0	0	1	0	1	0	0	40	60	50	80
7	11	1	1	0	0	1	0	1	1	0	0	1	50	30	60	80
8	11	1	1	0	1	1	0	0	1	0	0	0	40	50	30	70
9	11	1	1	0	0	0	0	1	0	0	0	0	20	80	60	70
10	11	1	1	0	0	1	0	0	1	0	0	0	30	70	50	70
11	11	1	0	0	0	0	0	0	0	1	0	0	10	50	20	70
12	11	1	1	0	1	0	1	1	1	0	0	1	60	80	50	80
13	11	1	1	0	0	1	0	1	0	0	0	0	30	30	20	80
14	11	1	1	0	0	0	0	1	1	0	0	1	40	50	40	90
15	11	1	1	0	0	1	1	1	1	0	0	1	60	60	50	70
16	11	1	1	1	1	0	0	1	1	0	0	0	50	50	20	90
17	11	1	1	1	0	0	0	1	1	0	0	0	40	50	40	80
18	11	1	1	0	0	0	1	1	1	0	0	1	50	50	50	90
19	11	1	1	0	0	1	0	0	1	0	0	1	40	60	40	90
20	11	1	1	0	1	1	1	1	1	0	1	0	70	50	50	80
21	11	1	1	0	0	0	1	0	1	0	0	0	30	70	60	100
22	11	1	1	0	1	1	1	1	1	0	0	1	70	70	20	80
23	11	1	1	0	0	1	1	0	0	1	0	0	40	70	30	80
24	11	1	0	1	0	1	1	1	0	1	0	0	50	40	30	70
25	11	1	1	1	0	0	1	1	1	1	0	0	60	60	50	90
26	11	1	1	0	0	0	1	1	1	0	1	1	60	70	50	90

*** Right answers are marked as '1' and wrong answers as '0,' similar to the Rosemary assessment, and are then calculated**

Thunder and Lightning Informational Recall Assessment of Pre and Post-Test

Test_Type...	CL_Q1	CL_Q2	CL_Q3	CL_Q4	CL_Q5	CL_Q6	CL_Q7	CL_Q8	CL_Q9	CL_Q10	Mean_CG_TL_Pre	Mean_CG_TL_Pos	Mean_EG_TL_Pre	Mean_EG_TL_Pos
1	1	0	0	0	1	0	0	1	0	0	30	50	60	90
1	1	1	0	0	0	1	0	1	0	0	40	70	50	80
1	1	0	0	1	0	0	0	1	0	1	40	80	60	100
1	1	0	0	1	1	0	0	1	0	0	40	50	30	60
1	1	0	0	0	0	0	0	1	0	1	30	60	30	80
1	1	1	0	1	1	0	0	1	0	0	50	80	40	90
1	1	1	0	0	0	0	1	1	0	1	50	60	30	100
1	1	1	1	1	1	1	0	1	0	0	70	90	40	80
1	1	0	0	0	1	1	0	0	0	0	30	60	60	100
1	1	0	0	1	0	1	1	0	0	0	40	40	40	70
1	1	0	0	1	0	1	0	1	0	1	50	60	40	70
1	0	0	0	1	0	0	0	0	0	0	10	50	30	70
1	1	0	0	1	0	0	0	1	0	0	30	60	30	100
1	0	1	0	1	1	0	0	1	0	1	50	70	40	100
1	1	1	0	0	1	0	0	0	0	0	30	50	50	90
1	1	0	1	1	0	0	0	1	0	0	40	40	20	100
1	1	0	0	0	0	0	1	1	0	0	30	50	10	90
1	0	1	0	0	0	0	1	1	0	0	30	30	30	80
1	0	1	0	1	1	0	0	1	0	0	40	50	40	100
1	0	0	0	0	1	0	1	1	0	1	40	60	50	100
1	0	0	0	1	0	0	0	0	1	0	20	70	30	100
1	0	0	0	0	1	0	1	1	0	0	30	70	40	80
1	0	1	0	0	0	0	1	1	0	0	30	50	40	100
1	0	0	0	0	0	1	0	1	0	1	30	80	20	70
1	1	0	0	1	1	0	1	1	0	0	50	70	10	90
1	0	0	0	0	0	0	0	1	0	0	10	60	20	100

***Right answers are marked as '1' and wrong answers as '0,' similar to the Thunder and lightning assessment, and are then calculated**

Learners' QoE Responses to Mulsemedia Effects (Olfactory, Airflow, and Vibration)

ER_Olf_Q11	ER_Olf_Q12	ER_Olf_Q13	ER_Olf_Q14	EL_Air_Q11	EL_Air_Q12	EL_Air_Q13	EL_Air_Q14	EL_Vib_Q15	EL_Vib_Q16	EL_Vib_Q17	EL_Vib_Q18
5	2	1	5	5	2	2	5	4	2	2	4
5	2	2	5	5	2	2	5	5	2	2	5
5	1	2	5	4	2	2	5	4	2	2	5
5	1	1	5	5	2	1	5	5	1	1	5
5	1	2	5	5	1	2	5	4	1	2	5
3	1	1	4	4	3	1	4	3	2	2	4
5	1	1	5	5	2	1	4	5	2	1	4
4	2	2	4	4	2	2	4	3	3	3	3
5	1	1	5	5	3	2	5	5	2	2	5
3	1	1	5	5	1	2	4	5	1	1	5
5	2	2	5	5	2	2	5	5	2	2	5
4	2	1	5	5	1	1	5	5	1	1	5
5	1	1	5	5	1	1	5	5	1	1	5
5	1	1	5	5	1	1	5	5	1	1	5
5	1	1	5	5	1	1	5	5	1	1	5
5	1	1	5	5	1	1	5	5	1	1	5
5	1	1	5	5	1	1	5	5	1	1	5
5	1	1	5	5	1	1	5	5	1	1	5
5	1	1	5	5	1	1	5	5	1	1	5
5	2	2	5	5	1	2	5	4	1	1	5
5	2	1	5	5	2	1	5	3	2	3	4
5	2	2	4	2	4	4	3	3	4	5	2
5	1	1	5	5	1	1	5	3	1	1	4
5	1	1	5	5	1	1	5	3	3	2	3
5	1	1	5	5	1	1	5	5	1	1	5
5	1	1	5	5	1	1	5	5	1	1	5

*** Likert Scale Analysis: 1 - Strongly Disagree, 5 - Strongly Agree**

Control Group QoE Responses

Group_1	Q11_QoE	Q12_QoE	Q13_QoE	Q14_QoE	Q15_QoE	Q16_QoE	Q17_QoE	Q18_QoE	Q19_QoE	Q20_QoE
11	4	4	4	3	4	3	4	3	4	4
11	5	2	5	1	5	1	4	5	1	4
11	4	1	5	1	5	1	3	5	1	4
11	4	4	4	2	4	2	4	4	2	4
11	4	3	4	2	3	3	3	3	2	4
11	4	4	4	2	4	2	3	4	2	4
11	4	2	5	2	5	2	3	4	4	3
11	4	2	4	2	4	2	2	4	2	3
11	4	3	4	2	4	2	2	4	2	3
11	5	4	4	1	4	2	3	3	3	3
11	4	3	3	3	3	3	4	3	3	2
11	4	3	4	2	3	4	4	3	3	4
11	4	3	4	2	3	2	5	3	3	4
11	4	3	4	2	4	2	2	5	1	4
11	4	4	4	2	4	2	2	4	2	5
11	4	4	3	2	4	2	2	4	2	4
11	4	4	3	2	4	2	2	4	2	5
11	4	4	3	3	3	4	4	4	4	3
11	4	4	4	2	4	2	4	4	2	4
11	3	4	4	4	2	4	2	3	4	2
11	4	3	4	3	3	3	4	3	4	4
11	3	3	4	2	3	3	4	4	3	3
11	4	2	4	2	4	2	4	4	2	4
11	4	4	4	2	4	2	4	3	3	3
11	4	3	4	2	4	2	4	4	2	4

***Likert Scale Analysis: 1 - Strongly Disagree, 5 - Strongly Agree**

Experimental Group QoE Responses

Group_1	Q11_QoE	Q12_QoE	Q13_QoE	Q14_QoE	Q15_QoE	Q16_QoE	Q17_QoE	Q18_QoE	Q19_QoE	Q20_QoE
22	5	2	5	1	5	1	1	5	1	5
22	5	2	5	1	5	2	1	5	2	5
22	5	1	5	1	5	2	1	5	1	5
22	5	1	5	1	5	1	1	5	1	5
22	4	2	4	2	4	2	2	4	1	4
22	5	2	5	1	5	2	1	5	1	5
22	4	2	4	2	4	2	2	4	2	4
22	5	2	5	2	5	1	2	5	1	5
22	5	3	5	1	4	1	1	5	1	5
22	5	1	5	1	5	1	1	5	1	5
22	5	1	5	1	5	1	1	5	1	5
22	5	3	5	1	5	1	1	5	1	5
22	5	2	5	1	5	1	1	5	1	5
22	5	2	5	1	5	1	1	5	1	5
22	5	1	5	2	5	1	1	5	1	5
22	5	1	5	1	5	1	1	5	1	5
22	5	2	5	1	5	1	1	5	1	5
22	5	1	5	1	5	1	1	5	1	5
22	5	2	5	2	5	1	1	5	1	5
22	5	1	5	1	5	2	2	5	1	5
22	4	2	4	2	4	2	2	5	2	4
22	5	1	5	1	5	1	1	5	1	5
22	5	1	5	1	5	1	1	5	1	5
22	5	1	5	1	5	1	1	5	1	5
22	5	1	4	4	5	1	1	5	1	5
22	5	1	5	1	5	1	1	5	2	5

***Likert Scale Analysis: 1 - Strongly Disagree, 5 - Strongly Agree**

Annexure II

Control Group FER Analysis Without Mulsemedia Effects on Learning Content

Control Group						Control Group				
Participants	**Time Step**	**Dominant Emotion**	**Emotion Percentage**	**Engagement**		**Participants**	**Time Step**	**Dominant Emotion**	**Emotion Percentage**	**Engagement**
2	0	angry	54.44	Disengaged		3	0	angry	75.83	Disengaged
2	1	sad	52.75	Disengaged		3	1	neutral	35.93	Disengaged
2	2	neutral	87.33	Highly Engaged		3	2	neutral	70.2	Engaged
2	3	angry	33.53	Disengaged		3	3	neutral	72.03	Engaged
2	4	neutral	71.76	Engaged		3	4	sad	50.96	Disengaged
2	5	neutral	39.89	Disengaged		3	5	sad	62.33	Disengaged
2	6	angry	47.5	Disengaged		3	6	angry	51.22	Disengaged
2	7	sad	79.66	Disengaged		3	7	angry	65.67	Disengaged
2	8	angry	89.97	Disengaged		3	8	angry	48.36	Disengaged
2	9	neutral	61.71	Engaged		3	9	neutral	87.6	Highly Engage
2	10	angry	61.65	Disengaged		3	10	angry	60.2	Disengaged
2	11	angry	55.44	Disengaged		3	11	angry	89.08	Disengaged
2	12	sad	48.32	Disengaged		3	12	neutral	87.55	Highly Engage
2	13	sad	35.76	Disengaged		3	13	sad	42.36	Disengaged
2	14	sad	67.49	Disengaged		3	14	sad	41.57	Disengaged
2	15	sad	69.08	Disengaged		3	15	angry	49.03	Disengaged
2	16	neutral	86.43	Highly Engaged		3	16	sad	55.15	Disengaged
2	17	sad	80.44	Disengaged		3	17	sad	54.8	Disengaged
2	18	sad	37.19	Disengaged		3	18	angry	52.09	Disengaged
2	19	neutral	88.99	Highly Engaged		3	19	sad	65.55	Disengaged
2	20	neutral	81.13	Highly Engaged		3	20	sad	39.88	Disengaged
2	21	sad	61.01	Disengaged		3	21	sad	41.93	Disengaged
2	22	neutral	36.41	Disengaged		3	22	angry	78.97	Disengaged
2	23	neutral	30.94	Disengaged		3	23	sad	54.8	Disengaged
2	24	angry	67.2	Disengaged		3	24	sad	54.8	Disengaged
2	25	neutral	58.96	Engaged		3	25	sad	54.8	Disengaged
2	26	sad	34.68	Disengaged		3	26	sad	64.8	Disengaged
2	27	sad	34.68	Disengaged		3	27	sad	45.7	Disengaged
2	28	sad	34.68	Disengaged		3	28	sad	34.6	Disengaged
2	29	sad	34.68	Disengaged		3	29	sad	67.8	Disengaged
2	30	sad	34.68	Disengaged		3	30	sad	85.7	Disengaged

Experimental Group FER Analysis with Mulsemedia Effects on Learning Content

Experimental Group						Experimental Group				
Participants	**Time Step**	**Dominant Emotion**	**Emotion Percentage**	**Engagement**		**Participants**	**Time Step**	**Dominant Emotion**	**Emotion Percentage**	**Engagement**
8	0	Angry	78.01	Disengaged		9	0	Angry	47.04	Disengaged
8	1	Sad	84.62	Disengaged		9	1	Fear	33.57	Disengaged
8	2	Surprise	77.48	Engaged		9	2	Fear	66.33	Disengaged
8	3	Surprise	48.88	Disengaged		9	3	Happy	59.78	Highly Engaged
8	4	Surprise	82.36	Highly Engaged		9	4	Surprise	39.47	Disengaged
8	5	Happy	89.46	Highly Engaged		9	5	Angry	86.25	Disengaged
8	6	Angry	70.71	Disengaged		9	6	Neutral	89.59	Highly Engaged
8	7	Happy	23.36	Engaged		9	7	Sad	52.53	Disengaged
8	8	Neutral	32.1	Disengaged		9	8	Fear	59.85	Disengaged
8	9	Fear	67.18	Disengaged		9	9	Surprise	37.78	Disengaged
8	10	Neutral	68.42	Engaged		9	10	Happy	69.72	Highly Engaged
8	11	Neutral	49.61	Disengaged		9	11	Fear	85.3	Disengaged
8	12	Sad	20.55	Disengaged		9	12	Fear	82.66	Disengaged
8	13	Sad	35.6	Disengaged		9	13	Neutral	86.96	Highly Engaged
8	14	Angry	25.84	Disengaged		9	14	Sad	32.14	Disengaged
8	15	Fear	52.69	Disengaged		9	15	Neutral	25.71	Disengaged
8	16	Sad	45.04	Disengaged		9	16	Surprise	20.85	Disengaged
8	17	Surprise	65.48	Engaged		9	17	Happy	44.78	Engaged
8	18	Happy	84.72	Highly Engaged		9	18	Surprise	41.99	Disengaged
8	19	Fear	24.05	Disengaged		9	19	Angry	76.73	Disengaged
8	20	Happy	48.65	Engaged		9	20	Happy	79.46	Highly Engaged
8	21	Happy	29.98	Engaged		9	21	Happy	81.33	Highly Engaged
8	22	Happy	86.98	Highly Engaged		9	22	Sad	62.68	Disengaged
8	23	Sad	71.17	Disengaged		9	23	Angry	46.25	Disengaged
8	24	Angry	48.53	Disengaged		9	24	Fear	28.74	Disengaged
8	25	Neutral	89.86	Highly Engaged		9	25	Happy	47.05	Engaged
8	26	Neutral	56.63	Disengaged		9	26	Neutral	44.78	Disengaged
8	27	Fear	22.6	Disengaged		9	27	Fear	70.68	Disengaged
8	28	Surprise	72.53	Engaged		9	28	Neutral	41.23	Disengaged
8	29	Sad	47.15	Disengaged		9	29	Neutral	60.51	Engaged
8	30	Neutral	79.71	Engaged		9	30	Happy	79.48	Highly Engaged

Annexure III

Human Ethical Clearance Certificate

INSTITUTIONAL HUMAN ETHICS COMMITTEE

Avinashilingam

Institute for Home Science and Higher Education for Women
(Deemed to be University under Category 'A' by MHRD, Estd. u/s 3 of UGC Act 1956) Re-accredited with 'A++' Grade by NAAC.
Recognised by UGC Under Section 12 B
Coimbatore-641 043, Tamil Nadu, India

23rd March 2022

To
Ms.Mohana.M
Department of Computer Science
Avinashilingam Institute for Home Science and
Higher Education for Women
Coimbatore – 641 043

Dear Mohana.M,

Ref: Your proposal No. IHEC/21-22/CS-03 entitled "Development of Novel Adaptive Learning Management System based on Primary School Children Facial Emotion Recognition from Video Sequence using Artificial Intelligence Algorithms" resubmitted for approval to IHEC on 15.03.2021.

The Institutional Human Ethics Committee of our University hereby grants approval to your research proposal No. IHEC/21-22/CS-03 entitled "Development of Novel Adaptive Learning Management System based on Primary School Children Facial Emotion Recognition from Video Sequence using Artificial Intelligence Algorithms"resubmitted by you. The Approval number for the same is AUW/IHEC/CS-21-22/XMT-03.

We wish you all the best in your research endeavours.

Regards,

Dr.S.Uma Mageshwari
Member Secretary

Institutional Human Ethics Committee

INSTITUTIONAL HUMAN ETHICS COMMITTEE

Avinashilingam
Institute for Home Science and Higher Education for Women
(Deemed to be university under Category 'A' by MHRD, Estd. u/s 3 of UGC Act 1956) Re-accredited with 'A^{++}' Grade by NAAC.
Recognised by UGC Under Section 12 B
Coimbatore- 641043, Tamil Nadu, India

18.05.2023

To
Dr. P. Subhashini
Department of Computer Science
Avinashilingam Institute for Home Science and
Higher Education for Women
Coimbatore- 641043

Dear Dr. P. Subhashini

Ref: Your proposal No. IHEC/22-23/CS-07 entitled "Affective Learning Using Mulsemedia" submitted for approval of IHEC on 16.05.2023.

The Institutional Human Ethics Committee of our University hereby grants approval to your research proposal No. IHEC/22-23/CS-07 entitled "Affective Learning Using Mulsemedia". The Approval number for the same is AUW/IHEC/CS-22-23/XPD-07

We wish you all the best in your research endeavours.

Regards

Dr. A Thirumani Devi
Member Secretary

List of Publications

Book

1. P. Subashini, P. Lalitha, R. Janani, M. B. Jennifer Susan, and M. Mohana, *Python for Chemistry*. Chennai, India: Notion Press, 2024. [Online]. Available: https://notionpress.com/read/python-for-chemistry

Book Chapters

1. M. Mohana, A. C. da Silveira, P. Subashini, G. Ghinea, and C. A. S. Santos, "Technology Enhanced Mulsemedia Learning Through Design for Learning Disabilities," in *Envisioning the Future of Education Through Design*, R. Huang, D. Liu, M. A. Adarkwah, H. Wang, and B. Shehata, Eds., Lecture Notes in Educational Technology. Singapore: Springer, 2024.

2. M. Mohana, K. Nandhini, and P. Subashini, "Review on Artificial Intelligence and Robots in STEAM Education for Early Childhood Development: The State-of-the-Art Tools and Applications," in *Handbook of Research on Innovative Approaches to Early Childhood Development and School Readiness*, IGI Global, 2022, pp. 468-498.

International Journals

1. M. Mohana, P. Subashini, and M. Krishnaveni, "Emotion Recognition from Facial Expression Using Hybrid CNN–LSTM Network," *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, vol. 37, no. 8, p. 2356008, 2023. (Scopus and SCIE) *(Impact Factor: 1.5)*

2. M. Mohana and P. Subashini, "Facial Expression Recognition Using Machine Learning and Deep Learning Techniques: A Systematic Review," *SN Computer Science*, Springer, vol. 5, p. 432, 2024. (Scopus)

3. M. Mohana, P. Subashini, and D. Shukla, "Revisiting Face Detection: Supercharging Viola-Jones with Particle Swarm Optimization for Enhanced Performance," *Journal of Intelligent & Fuzzy Systems*, IOS Press, vol. 46, no. 4, pp. 10727-10741, 2024. (Scopus and SCIE) (*Impact Factor: 2.0)*

4. M. Mohana and P. Subashini, "Analysing the Performance of Viola-Jones and Multi-Task Convolution Neural Network (MTCNN) Face Detection Algorithms using Video Sequences," *International Journal of Computational Vision and Robotics*, Inderscience, 2024. (Accepted) (Scopus).

5. S. Divyasri, M. Mohana, T. T. Dhivyaprabha, and P. Subashini, "Empowering Young Learners: M-Learning Application with Adaptive Learning and CCI Standards," *International Journal of Technology Enhanced Learning*, Inderscience, 2024. (Accepted) *(Scopus and ESCI) (Impact Factor: 1.0)*

6. M. Mohana and P. Subashini, "Performance Analysis of Autoencoder Using Facial Expression Recognition," *Neural Processing Letters*, Springer, 2024, (Under Second Review) *(Scopus and SCIE) (Impact Factor: 3.1)*

7. M. Mohana, A. C. da Silveira, P. Subashini, C. A. S. Santos, and G. Ghinea, "Towards Enhancing STEM Learner Experience by Leveraging Mulsemedia Technology," *Multimedia Tools and Applications*, Springer, 2024. (Communicated) *(Scopus and SCIE) (Impact Factor: 3.0)*

8. M. Mohana, P. Subashini, and G. Ghinea, "XAI: Explainable AI-based Deep Semi-Supervised Convolutional Sparse Autoencoder for Facial Expression Recognition," *Signal, Image and Video Processing*, Springer, 2024. (Under Second Review) *(Scopus and SCIE) (Impact Factor: 2.0)*

Conferences

1. M. Mohana and P. Subashini, "Comparison of Viola-Jones and Multi-Task Convolution Neural Network (MTCNN) for Face Detection from Children Facial Expression Using Video Sequences," in *Proc. Int. Conf. on Artificial Intelligence – Multidisciplinary Perspectives on Emerging Challenges, Research, and Opportunities (ICAI-2022)*, Centre for Machine Learning and Intelligence, 2021, Abstract Only, ISBN: 979-8886677430.

2. M. Mohana and P. Subashini, "Emotion Recognition Using Autoencoders: A Systematic Review," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, IEEE, 2023, pp. 438-443.

3. M. Mohana and P. Subashini, "Emotion Recognition Using Deep Stacked Autoencoder with Softmax Classifier," in *Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, IEEE, 2023, pp. 864-872.

4. M. Mohana and P. Subashini, "Convolutional Sparse Autoencoder for Emotion Recognition," in *the 3rd International Conference on Artificial Intelligence and Computer Vision (AICV2023)*, A. E. Hassanien et al., Eds., Lecture Notes on Data Engineering and Communications Technologies, vol. 164, Springer, Cham, 2023.

5. M. Mohana, N. Valliammal, V. Suvetha, M. Krishnaveni, P. Subashini, and G. Ghinea, "A Study on Technology-Enhanced Mulsemedia Learning for Enhancing Learner's Experience in E-Learning," in *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, IEEE, 2023, pp. 01-06.

6. M. Mohana, A. C. da Silveira, P. Subashini, C. A. S. Santos, and G. Ghinea, "Technology Enhanced Mulsemedia Learning: Insights of an Evaluation," in *Computer-Human Interaction Research and Applications (CHIRA 2023)*, H. P. da Silva and P. Cipresso, Eds., Communications in Computer and Information Science, vol. 1997, Springer, Cham, 2023.

7. P. Subashini, M. Krishnaveni, T. T. Dhivyaprabha, M. Mohana, and S. Divyasri, "Artificial Intelligence-based Totbot Application for Primary School Children," in *NCMRSI 2023: Proc. Nat. Conf. on Multidisciplinary Research for Sustainable Innovations*, Coimbatore, 2023, pp. 126-137, ISBN: 978-81-931101-6-4

Workshop

1. M. Mohana, A. C. Da Silveira, V. Suvetha, P. Subashini, G. Ghinea, and C. A. S. Santos, "Technology Enhanced Mulsemedia Learning (TEML) for Learners with Dyslexia," in *Proc. 2023 ACM Int. Conf. on Interactive Media Experiences Workshops*, 2023, pp. 62-65.

Avinashilingam Institute for HomeScience and Higher Education for Women
(Deemed to be University Estd. u/s 3 of UGC Act 1956, Category 'A' by MHRD
Re-accredited with A++ Grade by NAAC. CGPA 3.65/4, Category I by UGC
Coimbatore - 641 043, Tamil Nadu, India

Appendix L2
(Item No 5 of Check List)
Details of Research Publications

S.No	Article	Journal	Other Details Vol/No/Page No/ Year	Published in UGC-CARE / Scopus Indexed/ Web of Science
1.	Emotion Recognition from Facial Expression Using Hybrid CNN–LSTM Network	International Journal of Pattern Recognition and Artificial Intelligence, World Scientific	2023, 37(08), 2356008 (27 pages)	Scopus and Web of Science
2	Facial Expression Recognition Using Machine Learning and Deep Learning Techniques: A Systematic Review	SN Computer Science, Springer	2024, 5, 432	Scopus
3	Revisiting face detection: Supercharging Viola-Jones with particle swarm optimization for enhanced performance	Journal of Intelligent & Fuzzy Systems, IOS Press	2024, 46(4), 10727 – 10741	Scopus and Web of Science
4	Analysing the Performance of Viola-Jones and Multi-Task Convolution Neural Network (MTCNN) Face Detection Algorithms using Real-Time Video Sequences	International Journal of Computational Vision and Robotics (IJCVR), Inderscience	IJCVR V 15 N 3 2025	Scopus

*Proof of list of Journals from the Internet to be attached and copies of reprints.

Scholar : M. Mohana

Supervisor : 24/6/2024

Checked By: 25/6/2024
HoD/Dean of Respective School

The scholar Miss. Mohana, M (20PHCSF001) has published her articles in the following journals:
1. International Journal of Pattern Recognition and Artificial Intelligence - indexed in Scopus,
2. SN Computer Science - indexed in Scopus.
This may be considered.

24.6.24.

International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 37, No. 8 (2023) 2356008 (27 pages)

DOI: 10.1142/S0218001423560086

World Scientific
www.worldscientific.com

Emotion Recognition from Facial Expression Using Hybrid CNN–LSTM Network

M. Mohana*, P. Subashini† and M. Krishnaveni‡

Department of Computer Science
Centre for Machine Learning and Intelligence (CMLI)
Avinashilingam Institute (Deemed University)
Coimbatore, India
**mohana_cs@avinuty.ac.in*
†*subashini_cs@avinuty.ac.in*
‡*krishnaveni_cs@avinuty.ac.in*

Received 16 September 2022
Accepted 29 January 2023
Published 7 July 2023

Facial Expression Recognition (FER) is a prominent research area in Computer Vision and Artificial Intelligence that has been playing a crucial role in human–computer interaction. The existing FER system focuses on spatial features for identifying the emotion, which suffers when recognizing emotions from a dynamic sequence of facial expressions in real time. Deep learning techniques based on the fusion of convolutional neural networks (CNN) and long short-term memory (LSTM) are presented in this paper for recognizing emotion and identifying the relationship between the sequence of facial expressions. In this approach, a hyperparameter tweaked VGG-19 skeleton is employed to extract the spatial features automatically from a sequence of images, which avoids the shortcoming of the conventional feature extraction methods. Second, these features are given into bidirectional LSTM (Bi-LSTM) for extracting spatiotemporal features of time series in two directions, which recognize emotion from a sequence of expressions. The proposed method's performance is evaluated using the CK+ benchmark as well as an in-house dataset captured from the designed IoT kit. Finally, this approach has been verified through hold-out cross-validation techniques. The proposed techniques show an accuracy of 0.92% on CK+, and 0.84% on the in-house dataset. The experimental results reveal that the proposed method outperforms compared to baseline methods and state-of-the-art approaches. Furthermore, precision, recall, $F1$-score, and ROC curve metrics have been used to evaluate the performance of the proposed system.

Keywords: Artificial intelligence; Bi-LSTM; computer vision; CNN–LSTM; facial expression recognition; human–computer interaction; spatiotemporal.

1. Introduction

Emotion recognition plays a prominent role in human–computer interaction. Humans can express emotions in different ways such as, through speech, facial

*,† Corresponding authors.

expressions, and body language. Facial expression analysis is the most significant and active research area in recent decades among those concerned with emotion recognition. Mehrabian's[29] works show that 55% of the message can be obtained through feelings and attitudes in facial expressions, 7% is expressed through speech, and the rest by paralinguistic. Ekman and Friesen[11,12] proposed six basic emotions: happy, sad, surprise, fear, anger, and disgust. Sometimes researchers add neutral and contempt in this category. These are known as universal expressions or primary emotions. In the fields of computer vision, deep learning, and pattern recognition, the study of facial expression recognition (FER) is given more consideration. It has also been extensively used in a variety of fields, including education, human–robot interaction, the healthcare system, autism spectrum disorder, and sophisticated driver assistance systems.

Researchers have been working on FER using machine learning algorithms for the past two decades. The objective of FER is to distinguish and categorize the significant movements of various facial musculature into meaningful diverse emotions. Numerous research on human–computer interaction has been conducted.[27] However, the FER system comprises four main steps: face detection, pre-processing, feature extraction, and emotion classification. First, face detection is a serious processing step to identify or locate whether the images/video frames contain a face or not. Second, pre-processing methods are utilized to highlight the features of the image. The third stage involves extracting appropriate features from the detected face to identify the considerable emotion from facial expressions. Finally, the classifier has been trained to categorize emotions based on the target label.

Conventional FER commonly adopts various handcrafted feature extraction techniques (e.g. HOG, SIFT, LBP) to extract the feature from facial images.[14,37,51] Later, geometric-based and appearance-based feature extraction approaches were used to automate the FER system. According to certain studies,[2,26,30] the geometric-based feature extraction approach retrieves geometric characteristics connected to face action units. Both the active appearance model (AAM) and the active shape model (ASM) extract geometric features depending on the form and location of the facial expression.[36] The appearance technique is more resistant to noise and feature extraction than geometric-based methods. In a few situations, hybrid-based feature extraction approaches were applied, resulting in higher detection performance. These approaches are appropriate only in a clinical situation.

In real time, detecting emotion on the face has several challenges such as complex backgrounds, occlusion, illumination, and spontaneous expression.[26] As people vary by culture, this spontaneous expression differs from the normal expression by subtle expression. In such instances, typical feature extraction algorithms cause significant computing costs, learning time, and poor real-time performance. Furthermore, the complex image necessitates strong memory power and the investigation of discriminative visual features to distinguish facial emotions and connect with a person's related emotional state.

Deep learning has gained prominence in the computer vision sector in recent years because of increased computational power such as graphics processing unit (GPU) and tensor processing unit (TPU) support for massive data training. It showed significant improvement in image recognition and detection. Mainly, the convolutional neural network (CNN)[32] has attained remarkable achievements in the FER system due to its strong expressive power for the feature extraction method. It is widely used in static images which extract features based on the appearance of the images. However, these features could not define emotion entirely because static images lacking in dynamic sequences related to facial expressions.[1] Many two-dimensional (2D) CNN models fail to recognize temporal features in images. As a result, numerous studies integrated both feature extraction approaches, such as CNN with long short-term memory (LSTM)[17] and CNN–recurrent neural network (RNN).[7,19] These techniques are used to extract the dynamic sequence of features from images.

Inspired by different CNN–LSTM-based algorithms, this paper presents fusion feature extraction techniques from a sequence of facial expression images. This study proposes (1) A CNN–LSTM fusion model for the analysis of the sequence of facial expressions, (2) Data augmentation techniques used for generating various illumination, noise, and different angles of the facial image for improving the emotion recognition power of the system in the real-time scenario, (3) A hyperparameter tweaked skeleton of VGG-19 used for extracting spatial feature, which overcomes the shortcoming of conventional feature method, (4) The designed CNN–LSTM-based Bi-LSTM method is used to extract spatiotemporal features which classify each emotion accuracy based on feature vector sequence, and (5) The proposed system's performance is compared to the benchmark dataset and state-of-the-art methods.

This paper is presented as follows: Section 2 describes the related work regarding deep learning techniques for FER systems. Section 3 explains the dataset collection and pre-processing methods. Section 4 describes the proposed CNN with the LSTM network. Section 5 explains the effectiveness of the proposed approach. Section 6 presents the discussions and limitations of this study. Finally, Sec. 7 presents the efficiency of the proposed method.

2. Background and Related Work

2.1. *Facial expression recognition*

The majority of previous research looked at facial expressions from static images, which provide spatial information based on appearance. Facial expressions, on the other hand, are produced by the relaxation and contraction of the facial musculature. Therefore, it is efficient for extracting both spatial and temporal features in facial expressions. This paper analyzed some existing works based on FER using CNN and its limitations.

2.2. *Deep neural network*

In order to address the difficulties in facial expression detection, many researchers have proposed various deep learning techniques based on CNN and the integration of two or three networks in recent decades. The representation of CNN's architecture and the use of a specialized temporal feature extraction approach while processing visual sequences are the key differences between each network.

Mehendale[28] proposed new techniques for facial emotion detection by combining the two-CNN network. First, CNN removed the background of the image. The second one concentrated on facial descriptor vector extraction which works based on different emotions and achieved 96% accuracy on CMU and NIST datasets. Zadeh *et al.* proposed deep CNN with Gabor feature extraction techniques.[53] It increased the speed and accuracy of the FER system. Chowdary *et al.* used transfer learning ResNet-50,[16] VGG-19,[43] Inception V3,[46] and MobileNet to compare performance with the CK+ dataset.[8] Usually, transfer learning achieves better performance on popular benchmark datasets over state-of-the-art techniques. Haddad *et al.* introduced the three-dimensional (3D)-CNN network over a sequence of frames.[15] This proposed method improved the results when tested with CK+ and Oulu-CASIA datasets. A few studies used 3D-CNN in facial emotion recognition. A hybrid CNN–RNN network was suggested by Bai and Goecke and used to retrieve spatiotemporal information from video sequences.[3] Also, this method combines transfer learning ResNet-50[16] for training and used the VGGFace2[54] dataset to enhance the usefulness of the suggested approach. The Grad-CAM heat maps visualization has been used here to visualize the before and after training samples. Sepas-Moghaddam *et al.* proposed VGG-16 with a Bi-LSTM network for extracting spatiotemporal features.[39] The spatial descriptor from image sequences is first extracted by the VGG-16. Then, spatial–angular features are learned using the Bi-LSTM RNN. This method experimented with the IST-EURECOM Light Field Face database. A temporal relational network (TRN) was proposed by Pise *et al.* for recognizing changes in emotions on a student's face in an online learning environment.[34] In addition, MLP has been used as a base classifier for emotion recognition. The proposed framework achieved better results on DISFA+ datasets. Jaiswal and Valstar designed an e-learning system using the CNN network. This method has been mainly proposed for with and without learning disabilities students.[20] It achieved better results in CK+ and JAFFE datasets. In order to jointly learn spatial features and temporal features for FER, Liang *et al.* presented a Bi-LSTM network.[23] This method combined deep spatial network (DSN) and deep temporal network (DTN) for extracting spatial features from image sequences. After that those features are combined and given into the Bi-LSTM network. This combined method achieved better performance in CK+, Oulu-CASIA, and MMI datasets. The CNN–LSTM network was introduced by Li *et al.* to extract spatial–temporal features.[22] Finally, transfer learning has been used to boost the performance of the suggested approach.

Bargal *et al.* proposed CNN with support vector machine (SVM) for emotion classification from a video sequence.[4] After a fully connected layer, SVM was associated to classify the emotions. Xiao *et al.* proposed a malware classification framework based on a CNN–SVM network which is employed to extract the features automatically.[50] Finally, the SVM classifier is used to classify malware according to features with an accuracy of 0.997%. Ruiz-Garcia *et al.* have presented CNN–SVM facial emotion recognition framework for socially assistive robots.[38] This approach combines self-supervised feature extraction methods and SVM for emotion classification. The author achieved 96.26% of accuracy in KDEF datasets. In addition, feature extraction effectiveness was compared using CNN and Gabor filters. Donahue *et al.* presented a long-term RNN to learn spatiotemporal features from long video sequences for activity recognition.[10] Fan *et al.* proposed a hybrid network of RNNs with C3D.[13] First, RNN extracts the features of the image sequence based on appearance while the 3D convolution network combines features map of both image and audio for emotion classification. Jaiswal and Valstar proposed a combination of CNN with Bi-LSTM to learn spatiotemporal features from an image sequence.[20] This approach achieved a great performance in FERA 2015 dataset. Zahara *et al.* proposed an IoT-based facial emotion recognition using a CNN with the help of the OpenCV library.[54] It was mainly proposed to detect micro-expression on the face in real-time facial expression. This method has used FER-2013 for training a neural network and achieved 65.97%.

In this study, CNN is employed to extract spatial information and is trained on labeled facial image sequences. The extracted features are then fed into the designed Bi-LSTM network, which captures the facial expression's temporal and contextual information. Finally, each emotion is classified by the softmax layer. Furthermore, the performance is compared to benchmark datasets and state-of-the-art methods.

3. Methods and Materials

The entire proposed system of numerous stages of the FER is depicted in Fig. 1. The pre-processing pipeline runs on raw video frames of the facial images. In the pre-processing pipelines, histogram equalization, bilateral filtering, flipping, rotating, and normalization are performed for enhancing features and increasing dataset size. The pre-processed dataset consists of three parts: training validation and testing, which consists of different samples. Following that, the CNN–LSTM (Bi-LSTM) network is trained and employed to extract spatiotemporal features from the dynamic sequence for classifying facial expressions. Furthermore, hold-out cross-validation techniques are used to compute the model's accuracy and loss for each epoch. The performance of the proposed approach is measured using the following metrics: confusion matrix, accuracy, precision, recall, and $F1$-score. In addition, the proposed model is evaluated by state-of-the-art techniques.

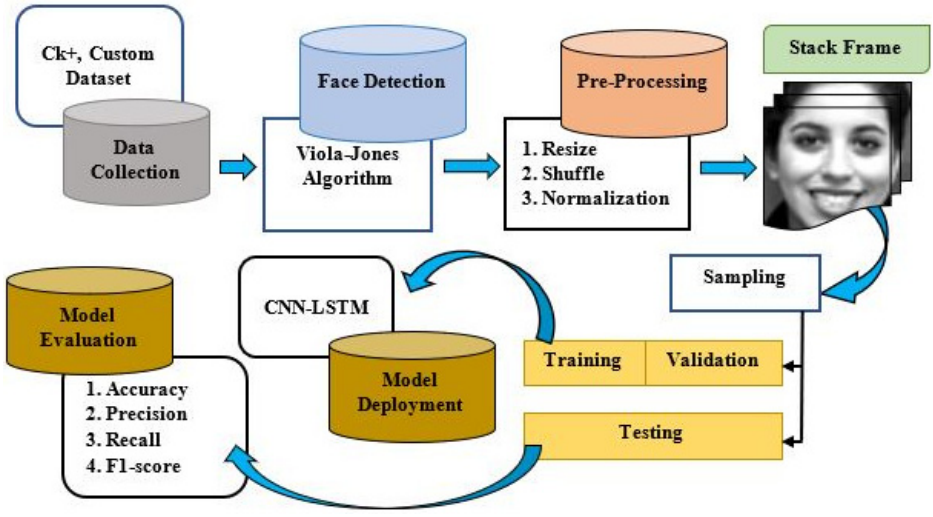


Fig. 1. Overall design of the proposed CNN–LSTM FER system.

3.1. *Dataset collection methods and description*

3.1.1. *CK+ dataset*

A facial expression dataset is a set of static images and short video clips with a range of emotions. The CK+ dataset[25] has 593 image sequences from 123 different people. Out of 593 images, 327 have been labeled for seven facial expressions such as anger, happy, sad, surprise, fear, contempt, and disgust. Figure 2 shows the sample facial expression of the CK+ dataset. Each expression in this database begins with a neutral expression and ends with a peak expression. For this study, the last three frames have been selected for incorporating the spatiotemporal features. Out of seven

Fig. 2. Sample CK+ facial expression images[26]: (a) surprise, (b) disgust, (c) happy.

emotions, only five emotions are considered for this study. The expression of contempt and disgust is less for training and testing. For training, 80% of data has been taken and the remaining 20% has been used for testing where peak images have been validated for the evaluation process.

3.1.2. *In-house dataset*

For analysis, the effectiveness of the proposed system, with a sequence of facial expressions in which in-house data has been captured in a lab setting for evaluating the spatiotemporal features. For that, Raspberry Pi and RGB camera modules are utilized to capture real-time spontaneous facial expressions from subjects in an unconstrained environment. The Raspberry Pi is a credit card-sized computer that usually connects to a TV and monitor. The camera port connects the camera with the Raspberry Pi for image and video processing in the computer vision field. Moreover, a 5 MP camera with Raspberry Pi 3/4 Model B has been employed in video capturing for this investigation. The goal of developing this dataset is to distinguish subjects' emotions during learning, which include happy, surprise, sleepy, and neutral. These emotions are recorded in a controlled laboratory setting with a free hand and head movements. The subject consent form is obtained before the study, and participants are asked to make subjective judgments of their feelings after their facial expressions are recorded. It is often called a self-annotation technique.

In detail, there are 40 female subjects aged from 21 to 26 years, with varying amounts of light, occlusion, and positions. This in-house dataset contains a sequence of facial expressions of individuals who consented to the use of their facial expressions for research purposes. Figure 3 depicts the experimental setup. The entire procedure takes place in the Centre for Machine Learning and Intelligence (CMLI). On the other hand, obtaining facial expressions is a difficult task. The video sequence is more useful for emotion recognition rather than static facial expression which does not contain temporal information. In a total of 1600 video frames, four emotions are captured. Happy represents pride and delight when studying, neutral denotes an inactive state of learning, sleepy signals low energy while learning, and surprise denotes a more extreme sense of enjoyment. Each clip stretched for 10–15 s, and the frame rate is 10 frames per second. Furthermore, the frames range from three

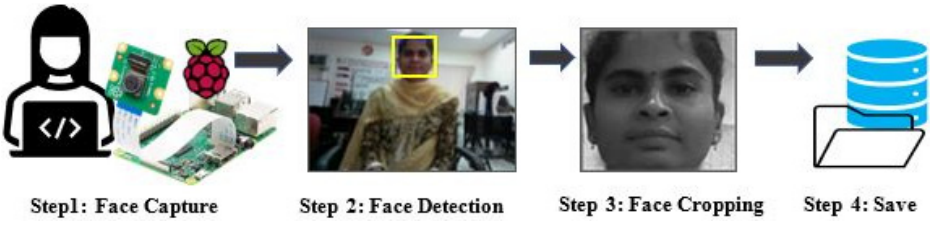


Fig. 3. Framework for facial emotion dataset collection using IoT kit.

peak levels. The algorithm for the facial expression dataset is collected using the following steps:

Step 1. Start the Raspberry camera to capture the subject's spontaneous facial expression.

Step 2. Viola–Jones[48] face detection algorithms are employed to detect the subject's facial expression. This approach is made up of four methods: Haar features, Integral images, Ada-boost, and Cascade classifier. Haar features are utilized to extract face information using line, edge, and four rectangular kernels. Integral images are employed to speed up the Haar feature extraction process, and an Ada-boosting classifier is used to build a strong feature to detect face and nonface in video frames. Finally, a Cascade classifier is utilized to eliminate the nonface region from the video frame.

Step 3. Captured video frames are converted to grayscale and scaled to 48×48 pixels during image pre-processing. The resized frames' probability density function is determined as

$$P(G_M) = \frac{N_M}{N}, \tag{1}$$

where G_M is the number of grayscale video frames in one emotion, N_M denotes the number of frames that occur, and N is the total times of pixels in one frame.

Step 4. Finally, 10 video frames of four emotions are saved in the appropriate folder.

3.2. *Pre-processing*

The term "image pre-processing" describes the transformation of the images before sending them to machine learning and deep learning techniques. The collected facial expression frames are made up of varied lighting conditions and image noise. In addition, grayscale images are widely used rather than color scales which contain less information about facial expressions. Therefore, using RGB photos is not required. So, histogram equalization[24] is done across grayscale video frames to equalize visual contrast. Figure 4 depicts the results. The blue color represents the original pictures,

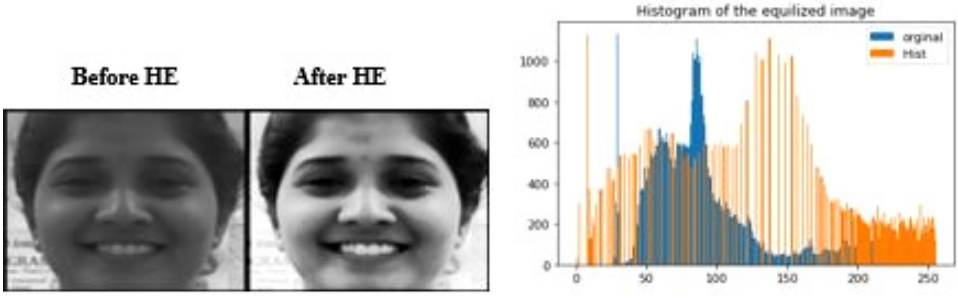


Fig. 4. Histogram equalization.

whereas the orange represents the results after HE. A bilateral filter[55] is used to eliminate noise from video frames. It is one of the image-smoothing filters that is nonlinear, edge-preserving, and noise-reducing. The bilateral filter reduced noise in a video frame more effectively than other filters such as the median filter and Gaussian blur filter. Figure 5 depicts the results of the filter, and the final facial expression after all the processing procedures. Finally, each pixel is divided by 255 to achieve normalization.

3.2.1. *Data augmentation*

Data augmentation[42] is an efficient technique used in deep learning for increasing the quantity of a dataset to increase the model performance. Deep learning-based data augmentation has been the main focus of recent image processing research.[31] The primary goal of this is to prevent overfitting issues caused by insufficient training data. In addition, it helps to generate possibilities of data in a real-world environment. To address these challenges, deep learning will employ data augmentation. The facial images are captured with a straight neck and upward direction, as illustrated in Fig. 5. However, the FER system takes these factors into account when it is built, because the facial position may change in real time depending on the camera's position or the person's posture. There are several techniques such as flipping, rotating, color space, cropping, translation, and noise injection available in data augmentation. Moreover, there are no particular rules for applying data augmentation techniques, it will differ depending on the kind of dataset.[42] For this study, the following technique is applied to the training dataset. Initially, the video frames of the left and right faces are horizontally flipped. Second, each emotion's video frame is rotated from -20° to $+20^\circ$, which is the safest rotation for facial images.[42]

Fig. 5. Collected emotion databases are happy (row 1), neutral (row 2), sleepy (row 3), surprise (row 4).

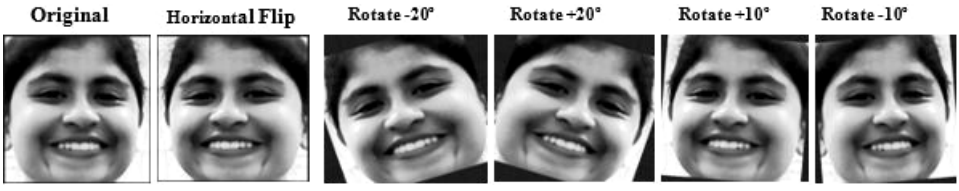


Fig. 6. After data augmentation.

The rotated frames are then horizontally flipped. The data size is expanded by generating synthetic frames by rotating and horizontal flipping, which improved the effectiveness of the proposed model. Figure 6 depicts the outcomes of enhancement procedures.

3.3. *Proposed model combining CNN and LSTM*

The proposed technique aims to discover the relationship between the sequence of facial expressions from their corresponding labels. As previously mentioned, the combination of the contraction and relaxation of one or more facial musculature produces facial expressions; hence, this study has focused on both spatial and temporal features.

3.3.1. *CNN model*

CNN is typically employed to extract spatial information and provides state-of-the-art performance for various computer vision tasks. It has four basic layers: a convolution layer, a pooling layer, a rectified linear unit (ReLU) layer, and a fully connected layer. Figure 7 depicts CNN's core architecture. CNN has been used in a variety of applications[27,32] and has performed admirably in areas such as medical image analysis, image segmentation, object detection, and image classification. The aim of CNN is to extract local descriptors from the top layer and transmit them to the lower level to extract complicated descriptors.

The convolution layer is made up of filters that determine the tensor of each convolution block's feature map. It extracts unique attributes from the given input images.

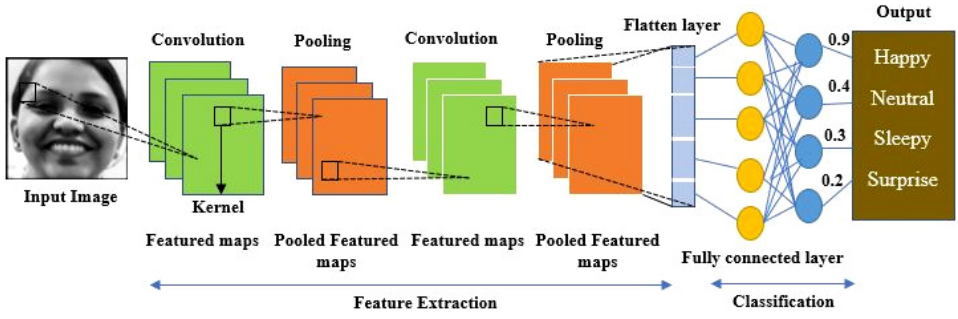


Fig. 7. Basic architecture of CNN.

The kernels (filters) are applied over the input images and use up "stride(s)" such that the result volume size comes to be a numeric matrix. The output of the image dimension is minimized after the striding procedure. Therefore, padding is necessary to pad an input volume with zero while maintaining the size of input images with low-level features. The convolution layer's mathematical process is as follows:

$$F(i,j) = (I * K)(i,j) = \sum I(i+m, j+n)K(m,n), \tag{2}$$

where i denotes the input matrix, K denotes the kernel size $m \times n$, and F denotes the feature map output. I^*K implies the operation between the convolution layer and the kernel. ReLU activation layer is commonly applied to reduce nonlinearity in the output of CNN feature maps. The exponential linear unit (Elu)[9] activation function is utilized in this experiment instead of ReLU as it overcomes the dying problem of the ReLU activation function. It is stated numerically as defined

$$\mathrm{Elu}(x) = \begin{cases} x, & x > 0, \\ \alpha(e^x - 1), & x < 0. \end{cases} \tag{3}$$

The pooling layer down samples incoming data to minimize the dimensionality of feature maps. It decreases the dimension of features to learn as well as the computation time performed for each block of the convolution operation. The most frequent approach is max pooling, which creates the largest value in an input area. Next, the pooling layer, a purposeful dropout layer[44] is employed to generalize the network. It aids in preventing the model from overfitting. Finally, the fully connected layer serves as a classifier, making a judgment based on the basic information gathered from the convolution and pooling layers.

3.3.2. *LSTM model*

LSTM is an extended version of the RNN algorithm that has difficulties of learning high-dimensional data. The conventional RNN architecture has been giving promising results in a shorter length of image sequences, whereas it is giving a poor performance in longer sequences of images due to the vanishing/exploding gradient issues.[17] However, unlike typical RNN units, LSTM incorporates a memory block and was designed to tackle such an issue by offering memory for retaining and forgetting past information over a lengthy period of time. Figure 8 depicts the LSTM's basic structure. It is made up of memory units and three control gates: forget gate, input gate, and output gate, where x_t represents the current input and C_t and C_{t-1} represent the new and prior cell states. Furthermore, h_t and h_{t-1} refer to the new and previous outputs, respectively.[45]

The implementation of various gates aids in the retention of earlier information depending on the network's dependencies. The following diagram depicts the LSTM input gate principle. The input gate (i_t) stores and updates new information about the current state.

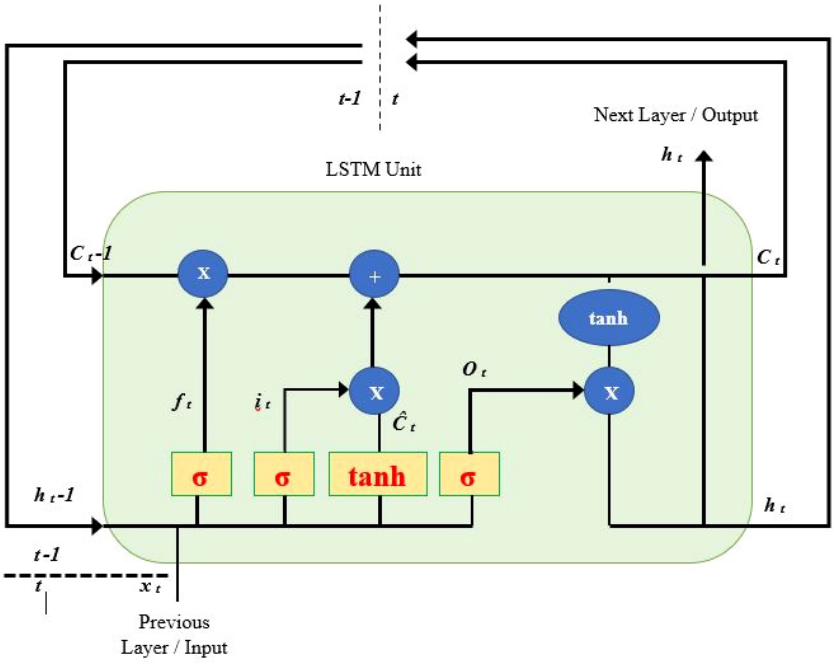


Fig. 8. Basic architecture of LSTM.

$$i_t = \sigma(W_i \cdot [h_{t-1} \cdot x_t] + b_i), \tag{4}$$

$$\hat{C}_t = \tanh(W_C \cdot [h_{t-1} \cdot x_t] + b_C), \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t, \tag{6}$$

where (4) is employed to pass the h_{t-1} and x_t via the sigmoid activation layer to determine which portion of the data needs to be added. Following that, (5) is utilized to acquire the new knowledge after h_{t-1} and x_t, which is then sent by the tanh layer. In LSTM units, information transmission is determined by the sigmoid function and the dot product. It values from 0 to 1. If the value of the sigmoid is 1, then information must be transferred. Otherwise, it may not be transferred. $\hat{C}_t$, C_{t-1} refer the long-term memory information and current moment information are joined in (6), where W_C denotes the sigmoid output and $\hat{C}_t$ denotes tanh output. In addition, W_C denotes weight matrix, and input gate bias of LSTM is b_i. The forget gate (f_t) is used in (4) to forget or keep previous information based on network dependencies, where W_f is the weight matrix and b_f is the offset.

$$f_t = \sigma(W_f \cdot [h_{t-1} \cdot x_t] + b_f). \tag{7}$$

The results are generated via the output gate (O_t). Using (7), this identifies the states that must be continued by the h_{t-1} and x_t inputs (8). The final values are derived by

passing the new information, C_t, through the tanh layer using a state decision vector.

$$O_t = \sigma(W_O \cdot [h_{t-1} \cdot x_t] + b_O), \tag{8}$$

$$h_t = O_t * \tanh(C_t), \tag{9}$$

where b_O and W_O are the weighted matrix and LSTM bias output gates, respectively.

3.3.3. *Fusion of CNN–LSTM model*

After analysis, the fusion of CNN and LSTM is feasible for extracting the spatiotemporal features from a sequence of facial images. The next step is to develop a CNN–LSTM combined approach (see Fig. 9) and improve efficiency by making the network stronger. To avoid the complexity and sensitivity of conventional feature extraction approaches, CNN layer is employed to extract the sequence of spatial features from corresponding labels based on frame sequence. After that, extracted feature vector sequence is fed into the designed Bi-LSTM model, which is incorporated with a forward and backward pass to extract spatiotemporal features. The function of LSTM is that it obtains information about the cells preceding a given cell. However, it is unable to access the data from the preceding cell. As a result, the model Bi-LSTM could perform better when processing time series data. Finally, the softmax layer is used to classify each emotion.

Figure 9 illustrates the proposed network of CNN–LSTM. Initially, increasing the number of layers for improving the efficiency of the network, the VGG-19 skeleton is used for extracting the spatial features from a sequence of images. VGG-19, on the other hand, relied on its deeper network structure and feature extraction ability, which reduced FER and caused overfitting issues due to the layer's size. In addition, the vanishing gradient problem slows down the process. subsequently, the layer has been fine-tuned based on the dataset and dimension of the image sequences. The proposed network consists of 21 layers: 8 TimeDistributed layers, 2 dropout layers, 2 batch normalization, 1 flatten layer, 4 pooling layers, 1 FC layer, and 2 LSTM layers with softmax function for categorical classification. Each convolution block is followed by a dropout layer or batch-normalization layer. The 3×3 kernel convolutional layer is utilized for feature extraction and is activated by the Elu activation function. The max-pooling layer is 2×2 in size and is used to minimize the size of the input dimension. The batch-normalization layer[18] is utilized instead of the dropout layer to normalize the output of the activation map, increasing network performance even more. The feature map vector is assigned to the Bi-LSTM layer after the architecture to extract time sequence information. It combines forward and backward LSTM to obtain hidden information from the past and future. The current input points' sequence of features is calculated by the first LSTM, and the reverse sequence features are read and added by the second LSTM. The forward and backward propagation of neurons updates the interaction between neurons in the two states. Therefore, it improves the ability to extract spatiotemporal features. The input form for the LSTM layer became (4,691,903), and Table 1 gives a

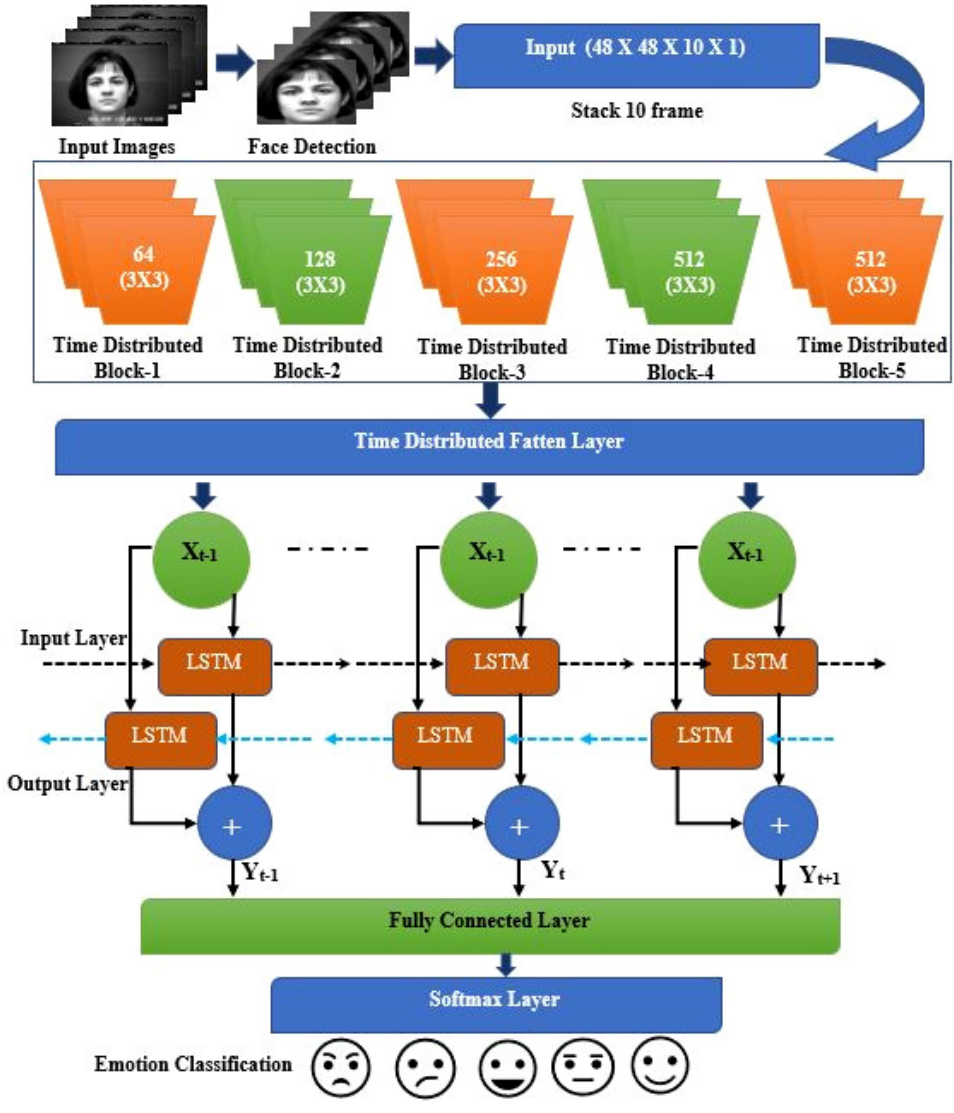


Fig. 9. Enhanced CNN–LSTM (Bi-LSTM) network.

summary of the proposed techniques. The softmax layer then uses labels from the input and the feature map that it has learned to classify and predict each expression. Softmax has been used to efficiently classify the nonlinear function for multi-class classification. It does, however, boost model generality. The softmax layer equation has given in (10). To avoid overfitting issues,[52] regularization techniques, such as dropout, an early stopping method, and kernel_initializer, are applied.

$$P(x) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}. \tag{10}$$

Table 1. Summary of CNN–LSTM (Bi-LSTM) network.

#	Type	Kernel Size	Stride	Kernel/Size	Kernel_Initializer	Parameter
	Input layer					$3 \times 48 \times 48 \times 1$
1	TimeDistributed	3×3	1	64	he_normal	640
2	TimeDistributed	3×3	1	64	he_normal	36,926
3	Pool	2×2	2	—		0
4	Dropout	—	—	0.45		0
5	TimeDistributed	3×3	1	128	he_normal	73,856
6	TimeDistributed	3×3	1	128	he_normal	147,584
7	Pool	2×2	2	—		0
8	BatchNormalization	—	—	—		128
9	TimeDistributed	3×3	1	256	he_normal	2,195,168
10	TimeDistributed	3×3	1	256	he_normal	590,048
11	Pool	2×2	2	—		0
12	Dropout	—	—	0.45		0
13	TimeDistributed	3×3	1	512	he_normal	1,180,160
14	TimeDistributed	3×3	1	512	he_normal	2,359,808
15	Pool	2×2	2	—		0
16	BatchNormalization	—	—	—		2048
17	TimeDistributed	—	—	—		4,691,903
18	LSTM	—	—	512		656,384
19	LSTM			64		164,352
20	FC	—	—	128		16,512
21	Output	—	—	5		645

The parameters setting of the proposed CNN with the LSTM (Bi-LSTM) network is shown in Table 2. The dataset is divided into a training set and a test set, each in an 8:2 ratio, for the aim of obtaining features from the sequence of images. For each round of training, the key model parameters namely the batch size, input size, hidden units, and dropout are monitored regularly. The effectiveness of FER is significantly influenced by the number of hidden units. So, the early stopping method, batch normalization, and dropout are utilized to avoid overfitting and generalize the proposed network. Due to the limited computing capability, the proposed model is trained with 100 iterations and 10 batch sizes. Furthermore, the Adam optimizer[21] incorporates the features of the AdaGrad and RMSProp algorithms to solve the

Table 2. Optimized parameter for proposed method.

Parameter Name	Value
Input size	$3 \times 48 \times 48 \times 1$
Activation function	Elu
Kernel size	3×3
Learning rate	0.001
Batch size	10
Dropout rate	0.3
Iteration	100
Optimizer	Adam

network's sparse gradient and noise challenges. As a result, the Adam optimizer is used to train this network, with a learning rate of 0.001.

4. Results and Analysis

4.1. *Experimental setup*

The FER datasets are partitioned into training and testing parts of 80% and 20%, respectively, for this experiment. The hold-out cross-validation method is employed to obtain the performance data. As can be seen in Table 1, the experimentally used presented network has nine TimeDistributed convolution layers with learning rates of 0.001 and 100 epochs. In order to avoid the overfitting problem, an early stopping method has been adopted in this experiment. The CNN, LSTM, and CNN with LSTM networks, as well as the proposed networks, have been evaluated on Google Colab environments using Python and Kera's package with TensorFlow backends. Furthermore, the studies were carried out using GPU) equipped with an Intel(R) Core (TM) i5-8400 CPU @ 2.80GHz 2.81 GHz Dell desktop.

4.2. *Performance evaluations*

The following performance metric is used to gauge the proposed system's effectiveness. TP indicates the True Positive of accurately predicted emotions. FP denotes the False Positive of misclassified classes. The True Negative of correctly recognized emotions is denoted by TN, whereas the False Negative of the FER system is denoted by FN, which is misclassified emotions. The following equations are used to calculate the accuracy, precision, recall, and $F1$-measure:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{FP} + \text{FN} + \text{TP} + \text{TN}), \tag{11}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}), \tag{12}$$

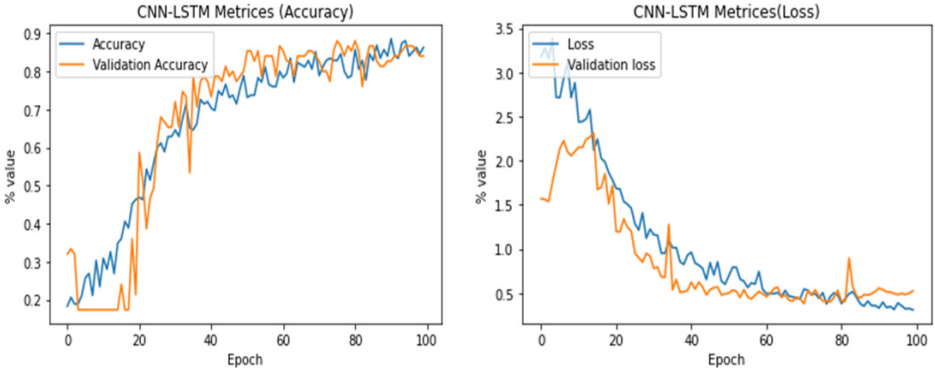
$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}), \tag{13}$$

$$F1\text{-measure} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}). \tag{14}$$

Accuracy is defined in this case as the ratio of correctly classified test frames to total frames. Precision is defined as the ratio of the number of correct frames in the test set for that emotion to the number of correct frames in the emotion recognition results. Recall refers to the ratio of the number of correctly identified frames in that emotion to the total number of frames in those emotions. Finally, the $F1$-score provides the proposed system's average accuracy and recall measures.

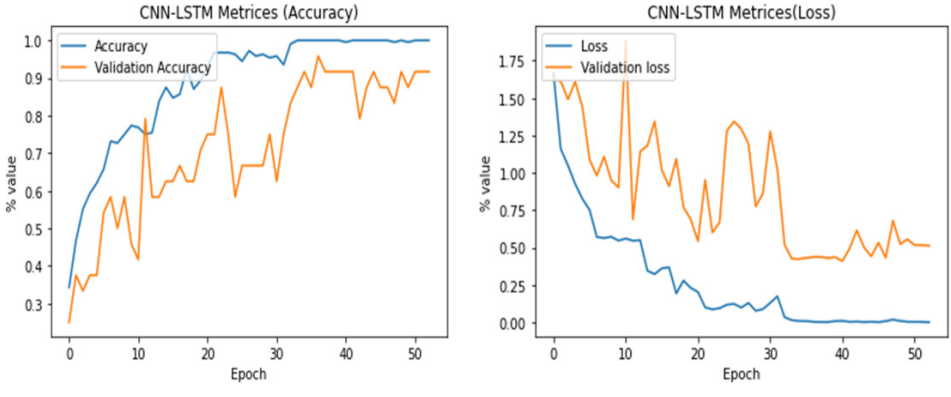
4.3. *Results on proposed network*

Figures 10 and 11 show the accuracy and cross-entropy (loss) performance evaluations of the proposed model on the CK+ and in-house datasets during the training and testing phases. In the CK+ dataset, at epoch 100, the training and testing accuracy are, respectively, 0.91% and 0.84%. Similar to this, the in-house dataset's



(a) (b)

Fig. 10. Timing process of CNN–LSTM in CK+ dataset (a) accuracy (b) loss.

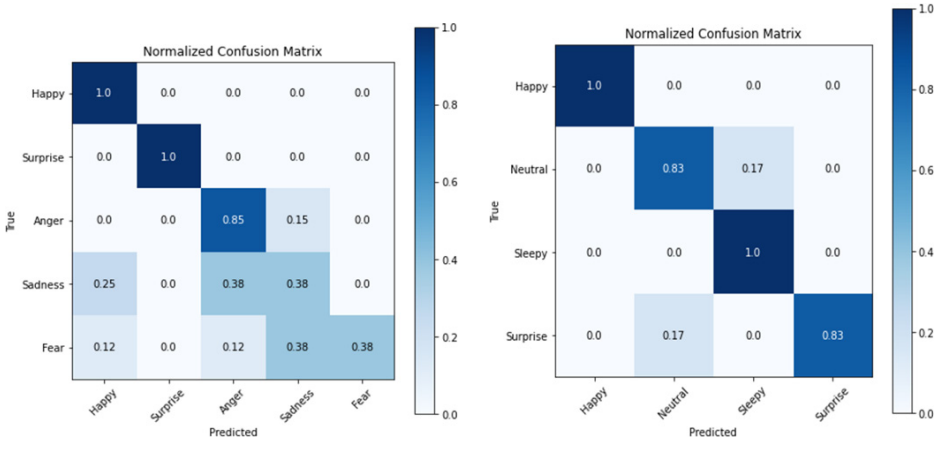


(a) (b)

Fig. 11. Timing process of CNN–LSTM in in-house dataset (a) accuracy (b) loss.

training and testing accuracies are 0.99% and 0.92%, respectively, at epoch 50. The training's initial stage's slower convergence speed could be seen, but it then picked up and converged at epochs 90–100 (see Fig. 10(a)), proving that the model's learning efficiency is satisfactory. The slope of the training curve is reduced during model training. Same as in Fig. 11(a), training accuracy gradually increased with slight fluctuation. Furthermore, the CK+ and in-house datasets have training and testing losses of 0.84 and 0.61, respectively.

Figure 12 illustrates the confusion matrix of the proposed model on both datasets. Figure 12(a) shows the performance on CK+ dataset. Among 750 image sequences, 21 were misclassified due to the similarity of facial appearances, with five emotions. From Fig. 12(a), happy and surprise achieved superior performance due to the sequence of facial features and time series information, whereas, anger, sad, and fear



(a) (b)

Fig. 12. (a) Confusion matrix of CK+. (b) Confusion matrix of in-house dataset.

are misclassified as other emotions slightly. Anger is falsely classified as sad, sad is falsely classified as happy and anger. Fear is falsely classified as sad, anger, and happy. The precision of four emotions varied from 0.83 to 1.0, the recall ranged from 0.83 to 1.0, and the $F1$-score ranged from 0.83 to 1.0, respectively, while the average precision and recall were 0.92 and 0.91, respectively (see Fig. 13). From Fig. 12(b), it is seen that happy and sleepy achieved a greater performance compared to neutral and surprise. Similarly, neutral overlapped sleepy expression, and surprise slightly overlapped with a neutral expression. The precision of five emotions varied from 0.38 to 1.0, the recall ranged from 0.38 to 1.0, and the $F1$-score ranged from 0.38 to 1.0, respectively, while the average precision and recall were 0.85 and 0.84, respectively (see Fig. 14). Due to similarities in the shape and appearance of the facial features and individual variations of the same facial expression, an expression is typically easily confused with another expression. The proposed CNN with LSTM yields notable results, more reliable true positive and true negative values, fewer false negative and false positive values, and more consistent true positive and true negative values. As a result, the proposed model is able to efficiently categorize the sequence of emotions.

Additionally, the receiver operating characteristic (ROC) curves between true positive rate and false positive rate are presented in Fig. 15 to assess the overall performance. The ROC curve clearly reveals that the suggested model performance is determined to be 0.92 in the in-house dataset and 0.84 in the CK+ dataset. The primary goal of this research is to discover the relationship between sequences of images of human facial expressions and the labels that correspond to them. In addition, it aims to extract the spatiotemporal features using a CNN architecture with LSTM (Bi-LSTM). From Fig. 15(a), happy and sleepy have been correctly identified, however, neutral and surprise have been slightly misclassified with 0.17%. From Fig. 15(b), happy and surprise are classified with a higher rate, fear and anger are

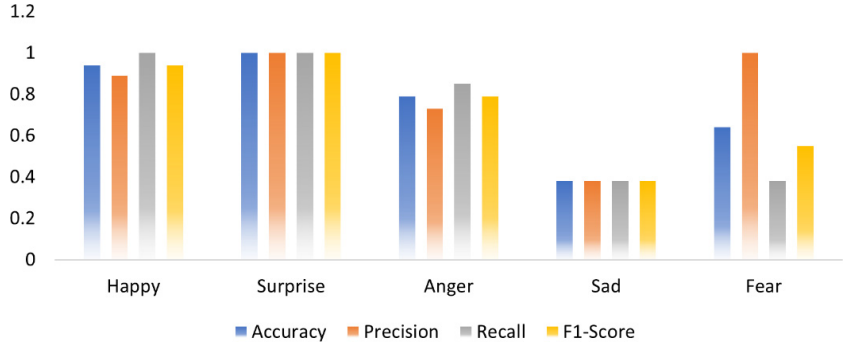


Fig. 13. Recognization accuracy of proposed model on CK+ dataset.

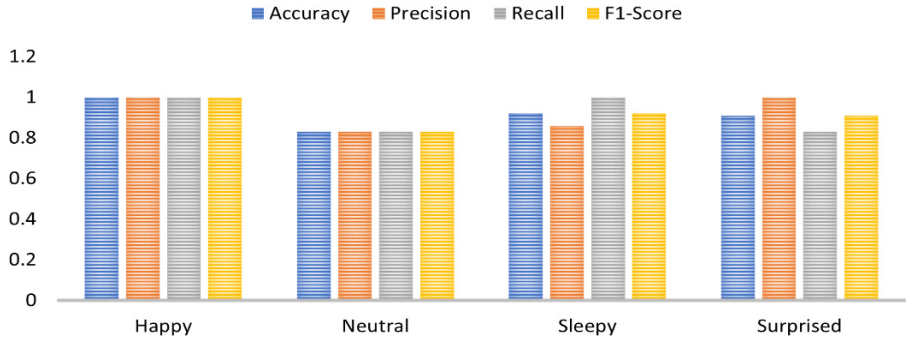
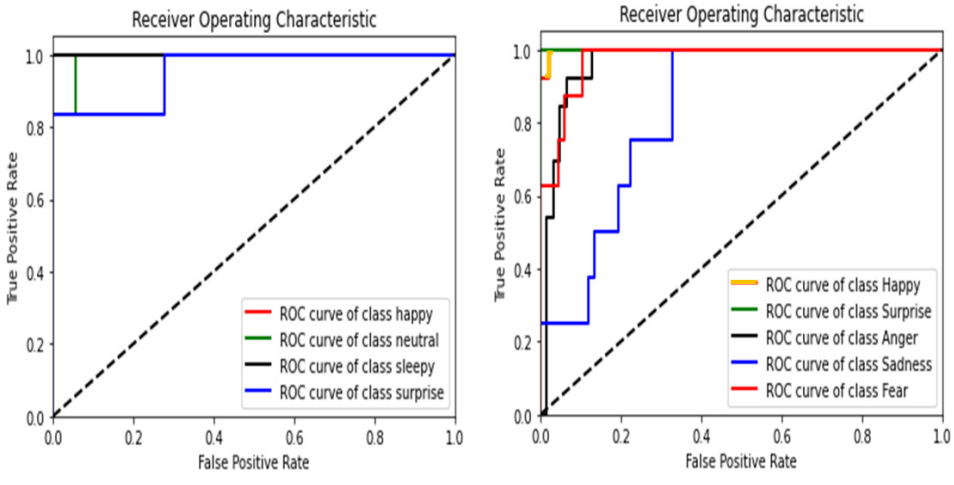


Fig. 14. Recognization accuracy of proposed model on in-house dataset.



(a) (b)

Fig. 15. ROC curve of the proposed model (a) in-house dataset (b) CK+ dataset.

given satisfiable results, and sad is given little lower results in the CK+ dataset. Furthermore, the primary goal of this research is to achieve satisfactory results in recognizing facial emotion from a dynamic sequence of facial expressions. The results of the experiment show that the proposed method outperformed the competitive state-of-the-art methods (see Tables 4 and 5).

4.4. *Comparison of baseline network and state-of-the-art FER models*

To evaluate the performance and generalization of the proposed model, hold-out cross-validation techniques are used to compare it to another baseline network such as CNN, LSTM, and CNN–LSTM. The above-mentioned model which has the same structure as the proposed model has been designed and validated. Table 3 summarizes the training parameters for each model. All models have the same learning rate, loss function, epoch, batch size, optimizer, and activation function; these parameters influence model performance while training. The gradient calculation is heavily influenced by the loss function and optimizer. The Learning rate is used to regulate the parameters which are updated with respect to the gradient. In most cases, CNN and LSTM networks use the ReLU activation function. In this study, the Elu activation function has been used instead of ReLU to address the dying ReLU problem. Moreover, Elu outperformed LeakyReLU. When compared to the other architectures in Table 3, the proposed model has the fewest parameters. The proposed model and all other models have been trained and tested with CK+ and an in-house dataset, shown in Table 4. The proposed model performance clearly boosted from 69.85% to 84.87% on CK+, and 74.56–92.84% on the in-house dataset due to time series features. From Table 4, CNN architecture is given poor performance with time sequence information whereas CNN–LSTM (only 78.34 on CK+, 84.32% on in-house) has given slightly superior results compared to general CNN (only 69.84 on CK+, 74.56% on in-house) and LSTM (only 50.78 on CK+, 79.38% on in-house) network. In addition, the computational time is also calculated for each model. Compared to

Table 3. Parameter settings of each baseline model.

Methods	Learning Rate	Epoch	Batch Size	Loss Function	Optimizers	Parameters
CNN	0.001	100	10	Cross-entropy	Adam	76 M
LSTM	0.001	100	10	Cross-entropy	Adam	57 M
CNN + LSTM	0.001	100	10	Cross-entropy	Adam	49 M
Proposed model	0.001	100	10	Cross-entropy	Adam	46 M

Table 4. Comparison between different models on CK+ and in-house dataset.

Methods	Accuracy CK+ (%)	Time	Accuracy in-House Dataset (%)	Time
CNN	69.84	8 min	74.56	12 min
LSTM	50.78	10 min	79.38	16 min
CNN + LSTM	78.34	7 min	84.32	10 min
Proposed model	84.87	5.5 min	92.84	7.5 min

Table 5. Performance analysis of the proposed model with state-of-the-art models.

Approach	Dataset	Accuracy (%)
SVM + NN[47]	CK+	80.00
PCA + KNN[48]	CK+	77.29
CNN + KNN[49]	CK+	80.30
	JAFFE	76.74
PCA + NMF + LNMF[50]	CK+	81.40
CNN[51]	In-house	78.04
Topographic Context + LDA[52]	CK+	82.68
HOSVD[53]	CK+/JAFFE	73.30
AAM + AUs + SVM[54]	CK+	54.47
Proposed method	CK+	**84.87**
	In-house	**92.84**

other baseline models proposed model took less time for training the model. The amount of the dataset and the complexity of the image sequences always affect the model's training time and accuracy. Additionally, the proposed model has been examined using current state-of-the-art methodologies, as shown in Table 5. From Table 5, it is found that some of the existing systems[5,6,33,35,40,41,47,49] obtained slightly lower accuracy in the range of 54.47–76.74%. The moderately reasonable accuracy of 77.29%, 78.04%, 80.00%, 80.30%, 81.40%, and 82.68% are found in Refs. 5, 33, 35, 41, and 49, respectively. The proposed model's 2.19% accuracy is increased compared to the state-of-the-art methods.

5. Discussion

The proposed model has been compared with baseline models such as CNN, LSTM, and CNN with LSTM (VGG-19) network. The collected in-house datasets have various challenging conditions such as illumination, partial occlusion, and noise. The proposed hyperparameter-tuned CNN with LSTM performed efficiently in this dataset as well as the benchmark CK+ dataset. This integrated model has been investigated experimentally on CNN and LSTM networks individually before being combined for extracting spatiotemporal features from a sequence of images. The majority of the facial expression datasets were created using a high-quality camera in a controlled setting. In this work, an emotion dataset has been captured by a low-quality 5 MP camera module attached to a Raspberry Pi to analyze the FER system's performance with challenging images. First, a FER has been trained with CNN, which yielded unsatisfactory performance testing results because CNN is incapable of extracting spatiotemporal properties from image sequences.[1] Nevertheless, this network has been working effectively in 2D images. Second, the LSTM model has been used to evaluate the performance of emotion recognition from the feature vector from the image sequence. This network did not give efficient results in images, because, when flattening the image sequence $48 \times 48 \times 3 \times 1$, it could not handle the massive size of the features. As a result, the performance of LSTM is insufficient for

developing a FER system model. Finally, features have been extracted using the VGG-19 architecture and fed into the LSTM network. Due to the number of layers, the FER model suffered from the vanishing gradient problem.[16] As a result, this model has been optimized by neuron size, layer count, batch normalization, dropout, and learning rate.

Furthermore, the proposed model is examined with pre-processing, which boosted the recognition rate to 0.92, and without pre-processing, which dropped the recognition rate to 0.80. When CNN and LSTM models are evaluated individually, the performance suffered due to overfitting and misclassification of sleepy and neutral emotions. Similarly, the VGG-19 architecture model has been used to extract spatiotemporal features from the sequence of images without tweaking the layer size, and activation function before entering the LSTM network. This combined network has given better results compared to the general CNN and LSTM network, even though, a hypermeter-tuned CNN with an LSTM model improved the recognition accuracy and classification outcomes rather than the VGG-19 architecture. Table 4 depicts the effectiveness of the proposed system with different baseline architectures. In addition, the proposed model marginally suffered in generalization when training due to insufficient sample size. Moreover, it does not possess the ability to recognize unseen emotions which did not fall in training. In the future, more subjects' facial samples will be collected for each expression from neutral to peak for analyzing the performance of the proposed system. However, when compared to the other state-of-the-art techniques, the proposed model outperforms them.

6. Conclusion

In the last two decades, emotion recognition is an active research area due to its application in many fields. The contraction and relaxation of some facial muscles produce continuous facial expressions. Hence, it is necessary to consider the dynamic and temporal features when recognizing facial emotion from facial expressions. The proposed model extracts spatiotemporal features from a sequence of frames, whereas existing CNN–LSTM works on FER with single images for emotion recognition. Initially, CNN is used to extract spatial features, while Bi-LSTM is used to identify time information in order to classify emotions based on labels. For this study, the in-house dataset has been collected with a small 5 MP camera with Raspberry Pi. The collected dataset consisted of high illumination and noise. Those challenges have been tried to minimize by pre-processing before entering the CNN model. This proposed system is trained from scratch; moreover, data augmentation, batch normalization, and dropout improve the proposed system's efficiency. In addition, the proposed model evaluated by the benchmark dataset achieved 0.84% and 0.92% on the in-house dataset. Furthermore, baseline model efficiency on the FER dataset is also compared with the proposed CNN–LSTM (Bi-LSTM) which efficiently learns the relationship between sequences of facial expression for identifying emotions. This

study will be expanded in the future to recognize accurate facial emotion recognition while combining it with other multimodal aspects.

Acknowledgment

This work has been supported by the CMLI funded by the Department of Science and Technology (DST-CURIE).

References

1. M. Asim, Z. Ming and M. Y. Javed, CNN-based spatiotemporal feature extraction for face anti-spoofing, in *2017 2nd Int. Conf. Image, Vision, and Computing (ICIVC)* (IEEE, 2017), pp. 234–238.
2. M. A. Azizan *et al.*, Development of real-time emotion recognition system based on machine learning algorithm, in *Human-Centered Technology for a Better Tomorrow* (Springer, Singapore, 2022), pp. 101–114, https://doi.org/10.1007/978-981-16-4115-2_8.
3. M. Bai and R. Goecke, Investigating LSTM for micro-expression recognition, in *Companion Publication of the 2020 Int. Conf. Multimodal Interaction* (Association for Computing Machinery, 2020), pp. 7–11, https://doi.org/10.1145/3395035.3425248.
4. S. A. Bargal, E. Barsoum, C. C. Ferrer and C. Zhang, Emotion recognition in the wild from videos using images, in *Proc. 18th ACM Int. Conf. Multimodal Interaction* (Association for Computing Machinery, 2016), pp. 433–436, https://doi.org/10.1145/2993148.2997627.
5. M. S. Bilkhu, S. Gupta and V. K. Srivastava, Emotion classification from facial expressions using cascaded regression trees and SVM, in *Computational Intelligence: Theories, Applications and Future Directions - Volume II* (Springer, Singapore, 2019), pp. 585–594.
6. I. Buciu and I. Pitas, Application of non-negative and local nonnegative matrix factorization to facial expression recognition, in *Proc. 17th Int. Conf. Pattern Recognition, 2004. ICPR 2004*, Vol. 1 (IEEE, 2004), pp. 288–291.
7. Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, VGGFace2: A dataset for recognising faces across pose and age, in *2018 13th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018)* (IEEE, 2018), pp. 67–74, https://doi.org/10.1109/FG.2018.00020.
8. M. K. Chowdary, T. N. Nguyen and D. J. Hemanth, Deep learning-based facial emotion recognition for human–computer interaction applications, *Neural Comput. Appl.* (2021) 1–18, https://doi.org/10.1007/s00521-021-06012-8.
9. D. A. Clevert, T. Unterthiner and S. Hochreiter, Fast and accurate deep network learning by exponential linear units (ELUs), preprint (2015), arXiv:1511.07289.
10. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko and T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2015), pp. 2625–2634.
11. P. Ekman and W. V. Friesen, Constants across cultures in the face and emotion, *J. Pers. Soc. Psychol.* **17**(2) (1971) 124.
12. P. Ekman and W. V. Friesen, Facial action coding system, *Environ. Psychol. Nonverbal Behav.* (1978).
13. Y. Fan, X. Lu, D. Li and Y. Liu, Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in *Proc. 18th ACM Int. Conf. Multimodal Interaction* (Association

for Computing Machinery, 2016), pp. 445–450, https://doi.org/10.1145/2993148.2997632.
14. X. Feng, M. Pietikäinen and A. Hadid, Facial expression recognition based on local binary patterns, *Pattern Recognit. Image Anal.* **17**(4) (2007) 592–598, https://doi.org/10.1134/S1054661807040190.
15. J. Haddad, O. Lézoray and P. Hamel, 3D-CNN for facial emotion recognition in videos, in *Int. Symp. Visual Computing* (Springer, Cham, 2020), pp. 298–309, https://doi.org/10.1007/978-3-030-64559-5_23.
16. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
17. S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.* **9**(8) (1997) 1735–1780.
18. S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in *Int. Conf. Machine Learning* (PMLR, 2015), pp. 448–456.
19. N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali and M. Zareapoor, Hybrid deep neural networks for face emotion recognition, *Pattern Recognit. Lett.* **115** (2018) 101–106.
20. S. Jaiswal and M. Valstar, Deep learning the dynamic appearance and shape of facial action units, in *2016 IEEE Winter Conf. Applications of Computer Vision (WACV)* (IEEE, 2016), pp. 1–8, https://doi.org/10.1109/WACV.2016.7477625.
21. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, preprint (2014), arXiv:1412.6980.
22. T. H. S. Li, P. H. Kuo, T. N. Tsai and P. C. Luan, CNN and LSTM based facial expression analysis model for a humanoid robot, *IEEE Access* **7** (2019) 93998–94011.
23. D. Liang, H. Liang, Z. Yu and Y. Zhang, Deep convolutional BiLSTM fusion network for facial expression recognition, *Vis. Comput.* **36**(3) (2020) 499–508.
24. L. Lu, Y. Zhou, K. Panetta and S. Agaian, Comparative study of histogram equalization algorithms for image enhancement, *Proc. Mobile Multimedia/Image Processing, Security, and Applications*, Vol. 7708 (SPIE, 2010), pp. 337–347.
25. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, in *2010 IEEE Computer Society Conf. Computer Vision and Pattern Recognition — Workshops* (IEEE, 2010), pp. 94–101.
26. B. Martinez and M. F. Valstar, Advances, challenges, and opportunities in automatic facial expression recognition, in *Advances in Face Detection and Facial Image Analysis* (Springer, Cham, 2016), pp. 63–100, https://doi.org/10.1007/978-3-319-25958-1_4.
27. V. Mayya, R. M. Pai and M. M. Pai, Automatic facial expression recognition using DCNN, *Procedia Comput. Sci.* **93** (2016) 453–461.
28. N. Mehendale, Facial emotion recognition using convolutional neural networks (FERC), *SN Appl. Sci.* **2**(3) (2020) 1–8.
29. A. Mehrabian, Communication without words, in *Communication Theory* (Routledge, 2017), pp. 193–200.
30. P. Michel and R. El Kaliouby, Real-time facial expression recognition in video using support vector machines, in *Proc. 5th Int. Conf. Multimodal Interfaces* (Association for Computing Machinery, 2003), pp. 258–264, https://doi.org/10.1145/958432.958479.
31. A. Mikołajczyk and M. Grochowski, Data augmentation for improving deep learning in image classification problem, in *2018 Int. Interdisciplinary PhD Workshop (IIPhDW)* (IEEE, 2018), pp. 117–122, https://doi.org/10.1109/IIPHDW.2018.8388338.
32. K. O'Shea and R. Nash, An introduction to convolutional neural networks, preprint (2015), arXiv:1511.08458.

33. M. Peter, J. L. Minoi and I. H. M. Hipiny, 3D face recognition using kernel based PCA approach, in *Computational Science and Technology* (Springer, Singapore, 2019), pp. 77–86.
34. A. Pise, H. Vadapalli and I. Sanders, Facial emotion recognition using temporal relational network: An application to E-learning, *Multimedia Tools Appl.* **81** (2022) 26633–26653, https://doi.org/10.1007/s11042-020-10133-y.
35. E. Pranav, S. Kamal, C. S. Chandran and M. H. Supriya, Facial emotion recognition using deep convolutional neural network, in *2020 6th Int. Conf. Advanced Computing and Communication Systems (ICACCS)* (IEEE, 2020), pp. 317–320.
36. M. S. Ratliff and E. Patterson, Emotion recognition using facial expressions with active appearance models, in *Proc. HRI* (Citeseer, 2008), pp. 1–6.
37. F. Ren and Z. Huang, Facial expression recognition based on AAM–SIFT and adaptive regional weighting, *IEEJ Trans. Electr. Electron. Eng.* **10**(6) (2015) 713–722, https://doi.org/10.1002/tee.22151.
38. A. Ruiz-Garcia, M. Elshaw, A. Altahhan and V. Palade, A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots, *Neural Comput. Appl.* **29**(7) (2018) 359–373.
39. A. Sepas-Moghaddam, A. Etemad, F. Pereira and P. L. Correia, Facial emotion recognition using light field images with deep attention-based bidirectional LSTM, in *2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020), pp. 3367–3371, https://doi.org/10.1109/ICASSP40776.2020.9053919.
40. M. Sert and N. Aksoy, Recognizing facial expressions of emotion using action unit specific decision thresholds, in *Proc. 2nd Workshop Advancements in Social Signal Processing for Multimodal Interaction* (Association for Computing Machinery, 2016), pp. 16–21.
41. K. Shan, J. Guo, W. You, D. Lu and R. Bie, Automatic facial expression recognition based on a deep convolutional-neural-network structure, in *2017 IEEE 15th Int. Conf. Software Engineering Research, Management, and Applications (SERA)* (IEEE, 2017), pp. 123–128.
42. C. Shorten and T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data* **6**(1) (2019) 1–48.
43. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint (2014), arXiv:1409.1556.
44. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* **15**(1) (2014) 1929–1958.
45. R. C. Staudemeyer and E. R. Morris, Understanding LSTM — A tutorial into long short-term memory recurrent neural networks, preprint (2019), arXiv:1909.09586.
46. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the inception architecture for computer vision, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 2818–2826.
47. H. Tan, Y. Zhang, H. Cheri, Y. Zhao and W. Wang, Person-independent expression recognition based on person-similarity weighted expression feature, *J. Syst. Eng. Electron.* **21**(1) (2010) 118–126.
48. P. Viola and M. J. Jones, Robust real-time face detection, *Int. J. Comput. Vis.* **57**(2) (2004) 137–154.
49. J. Wang and L. Yin, Static topographic modeling for facial expression recognition and analysis, *Comput. Vis. Image Underst.* **108**(1–2) (2007) 19–34.

50. G. Xiao, J. Li, Y. Chen and K. Li, MalFCS: An effective malware classification framework with automated feature extraction based on deep convolutional neural networks, *J. Parallel Distrib. Comput.* **141** (2020) 49–58.
51. X. Xu, C. Quan and F. Ren, Facial expression recognition based on Gabor wavelet transform and histogram of oriented gradients, in *2015 IEEE Int. Conf. Mechatronics and Automation (ICMA)* (IEEE, 2015), pp. 2117–2122.
52. X. Ying, An overview of overfitting and its solutions, *J. Phys., Conf. Ser.* **1168**(2) (2019) 022022.
53. M. M. T. Zadeh, M. Imani and B. Majidi, Fast facial emotion recognition using convolutional neural networks and Gabor filters, in *2019 5th Conf. Knowledge Based Engineering and Innovation (KBEI)* (IEEE, 2019), pp. 577–581, https://doi.org/10.1109/KBEI.2019.8734943.
54. L. Zahara, P. Musa, E. P. Wibowo, I. Karim and S. B. Musa, The facial emotion recognition (FER-2013) dataset for prediction system of the micro-expressions face using the convolutional neural network (CNN) algorithm-based Raspberry Pi, in *2020 Fifth Int. Conf. Informatics and Computing (ICIC)* (IEEE, 2020), pp. 1–9.
55. M. Zhang and B. K. Gunturk, Multiresolution bilateral filtering for image denoising, *IEEE Trans. Image Process.* **17**(12) (2008) 2324–2333.

M. Mohana is Ph.D. candidate in the Department of Computer Science, CMLI at the Avinashilingam Institute, Coimbatore, India. She is now working in the field of Mulsemedia and Affective Computing using Deep Learning Techniques. She has qualified for the UGC NET (National Eligibility Test for Assistant Professor) with JRF (Junior Research Fellowship) in June 2020. She received her M.Phil., MCA, and B.Sc. degree in Computer Science from the Bharathiar University in 2020, 2018, and 2015, respectively. She has an experience of one year in teaching in programming languages. Her research has spanned Artificial Intelligence, Facial Emotion Recognition, Computer Vision, Machine Learning, and Deep Learning. She also extended her contribution towards various international collaborations with universities from USA and UK.

P. Subashini is working for the Department of Computer Science, the Avinashilingam University for Women, Tamil Nadu, India, since 1994. She is also the Coordinator of the CMLI sanctioned by the Department of Science and Technology. Her research has spanned many disciplines like Image analysis, Pattern Recognition, Neural Networks, and Computational Intelligence. She has authored and co-authored four books, four book chapters, one monograph, and 145 research papers both at international and national levels. She has 10 sponsored research projects worth more than 2.54 crores from various government funding agencies. She also extended her contribution towards various international collaborations with universities from USA, Germany, and Morocco.

M. Krishnaveni is Assistant Professor in the Department of Computer Science, the Avinashilingam University for Women, Coimbatore, Tamil Nadu, India. She has research experience in Defense projects and worked in disciplines like IoT, Image Processing, Speech Processing, Data Mining, and Computational Intelligence. She has published four books, six book chapters, one monograph, and 86 research papers at both national and international levels. She has research projects under various funding agencies and acts as an active member of the CMLI and coordinates the AI Start-up program (Product Development Lab) for the student. She has received awards such as the best young teacher award IASTE 2017, the best NSS program officer award, NYLP 2016, Government of India.

SN Computer Science (2024) 5:432
https://doi.org/10.1007/s42979-024-02792-7

REVIEW ARTICLE

Facial Expression Recognition Using Machine Learning and Deep Learning Techniques: A Systematic Review

M. Mohana[1] · **P. Subashini**[1]

Received: 2 August 2023 / Accepted: 14 March 2024

Abstract
In the contemporary era, Facial Expression Recognition (FER) plays a pivotal role in numerous fields due to its vast application areas, such as e-learning, healthcare, marketing, and psychology, to name a few examples. Several research studies have been conducted on FER, and many reviews are available. The existing FER review paper focused on presenting a standard pipeline for FER to predict basic expressions. However, previous studies have not given an adequate amount of importance to FER datasets and their influence on affecting FER system performance. In this systematic review, 105 papers retrieved papers from IEEE, ACM, Science Direct, Scopus, Web of Science, and Springer from the years 2002 to 2023, following systematic review guidelines. Review protocol and research questions are also developed for the analysis of study results. The review identified that the accuracy of the FER system in controlled and spontaneous facial expression datasets is being affected, along with other challenges such as illumination, pose, and scale variation. Furthermore, this paper comparatively analyzed the FER model in both machine and deep learning techniques, including face detection, pre-processing, handcrafted feature extraction techniques, and emotion classifiers. In addition, we discussed some unresolved issues in FER and suggested solutions to enhance FER system performance further. In the future, multimodal FER systems need to be developed for real-time scenarios, considering the computational efficiency of model performance when integrating more than one model and dataset to achieve promising accuracy and reduce error rates.

Keywords Facial Expression Recognition (FER) · Machine learning (ML) · Deep learning (DL) · Face detection · Facial emotion · Survey

Introduction

Facial Expression Recognition (FER) is being actively explored in the fields of computer vision (CV) [141], machine learning (ML) [18], deep learning (DL) [79], Affective computing [22, 44, 116] and pattern recognition (PR) [18]. It is a powerful non-verbal expression that helps understand human internal states of emotions, social communication, and intent [97]. Face expressions convey 55 percent of communications about feelings and attitudes, with just 7% of those words being uttered and the rest being employed as non-linguistic [1, 3]. In recent years, numerous studies have been conducted on automatic facial expression recognition due to its wide application and practical importance in Human–Computer Interaction (HCI) [12], Human-Robotic Interaction (HRI) [36], and Psychology [36], particularly in the context of social robots, driver fatigue analysis, e-learning [45], medical treatment, and candidate interview processes [36].

In the twentieth century, Ekman and Friesen [42] defined six fundamental universal expressions: happiness, sadness, anger, surprise, fear, and disgust, based on cross-cultural studies. Also, 'contempt' was often included as one of the basic expressions in research studies [96]. Frustration, perplexity, boredom, and other advanced emotions are examples of secondary emotions [43]. However, some neuroscientists argued that these basic emotions are culturally specific not universal [62]. Additionally, various databases can be employed to recognize facial emotions. Also, Emotion recognition and expression recognition are similar but not quite

✉ M. Mohana
mohana_cs@avinuty.ac.in

P. Subashini
subashini_cs@avinuty.ac.in

[1] Centre for Machine Learning and Intelligence, Department of Computer Science, Avinashilingam Institute, Coimbatore, India

Published online: 13 April 2024

the same [17]. Facial expression recognition is a biometric-based technique used to identify emotions portrayed in facial images [62]. Facial Action Coding System [48, 147] and Valence-Arousal [25, 128] space are the two popular methods used to recognize emotions in the last decades of studying FER [49]. Furthermore, the automation of FER analysis in ML is still challenging [129].

FER system can be categorized based on feature extraction into two groups: (1) the static method [1, 56], which involves extracting spatial features from single images, and (2) the dynamic method [2, 8, 10], which focuses on extracting temporal features from a sequence of images/frames. Emotion recognition based on this approach encompasses audio, video, and physiological signals. However, static facial expression images yield purer expression results compared to other modalities, as they provide complementary information when involving high-level frames.

Facial expression analysis using ML and DL techniques is the focus of this systematic review. Generally, FER follows four basic steps: face detection and pre-processing, facial feature extraction, and emotion classification or recognition as shown in Fig. 1. In the conventional FER approach, the first step involves face detection. Faces are identified and highlighted with a bounding box within desired regions of interest (ROIs) using various approaches. However, it remains a challenging task to detect faces in real-time environments, especially under conditions of high illumination, occlusion, and extreme pose variation. The second step encompasses normalization, face alignment, and various pre-processing techniques such as data augmentation [139], noise removal, cropping, and feature enhancement.

The majority of ML techniques employ handcrafted feature extraction techniques such as histogram-oriented gradients (HOG) [1], scale-invariant feature transformation (SIFT) [115], local binary pattern (LBP), and principal component analysis (PCA) [38]. In the final stage, classifier algorithms like support vector machine (SVM) [20], k-nearest neighbor (KNN), decision tree (DT) [38], and ensemble methods are commonly utilized to classify emotions. However, since 2012, emotion recognition competitions such as FER-2013 [52] and a wild dataset [102], released a sufficient number of facial expressions collected from challenging real-time scenarios. Furthermore, with advancements in chip processing capabilities and well-designed

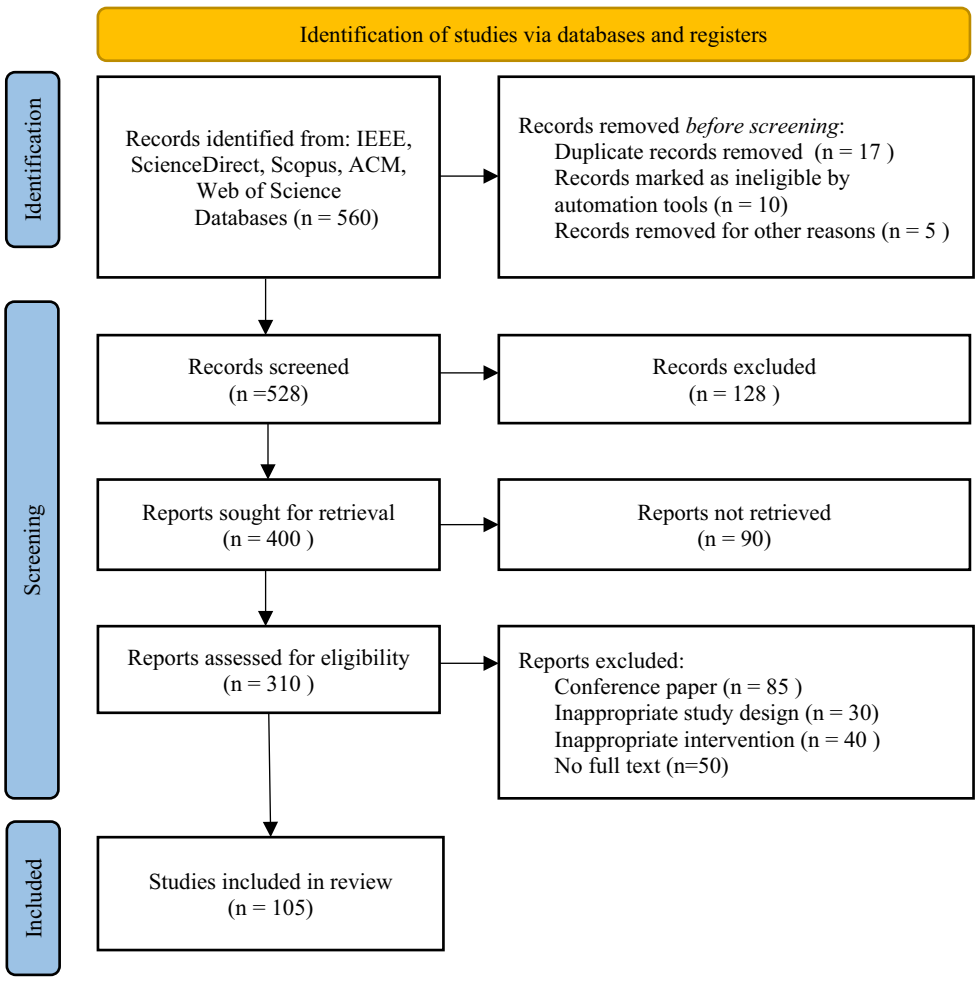


Fig. 1 PRISMA [108] flow diagram for paper selection for this study

neural network architectures, studies in the FER approach have started to incorporate deep learning methods. This shift aims to achieve state-of-the-art accuracy and surpass existing results due to the automatic feature extraction capabilities inherent in deep learning [13, 155]. In the realm of deep learning methods for emotion recognition, convolutional neural networks (CNN) [78], CNN-LSTM [101], 3D-CNN [103], recurrent neural networks (RNN) [120], autoencoders [26], generative adversarial networks (GAN) [157, 162], and reinforcement learning are commonly applied.

The growing demand for the development of efficient automatic FER systems in various applications is evident. Previous FER review works [4, 23, 47] have presented standard FER processes, encompassing handcrafted feature extraction techniques and machine learning (ML) classification techniques. Similarly, various deep learning (DL) techniques in the FER system are discussed. However, these discussions have not given adequate importance to the dataset and its influences on FER system accuracy, which serves as the motivation for this work. As a result, this systematic review primarily focuses on various efficient and robust FER techniques and their limitations. It also delves into face detection techniques and their associated challenges. Furthermore, we discuss recently released FER datasets and how they impact FER performance in a real-time environment. Additionally, we focus on the shortcomings in DL techniques when applied to FER datasets. Also, DL techniques necessitate a large amount of training data to overcome overfitting issues. However, existing datasets are insufficient to train a well-designed DL model to achieve promising results in a real-time environment.

A systematic literature review was conducted to gather FER studies utilizing ML and DL techniques spanning from the year 2002 to 2023. Additionally, the authors formulated and evaluated a significant research question to identify potential challenges in FER research when employing ML and DL techniques. This preliminary step precedes the analysis and retrieval of FER papers. The approach not only highlights potential research gaps in the specified problem area but also offers clear guidance to researchers, practitioners, and industries interested in undertaking new research in these domains. All associated papers were retrieved from various reputable databases, integrated, and analyzed to address the research question outlined in "Selection Criteria". Also, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram shows the screening process of paper selection including the inclusion and exclusion reasons of FER papers for this study. Furthermore, this review provides valuable insights for new researchers, aiding in a better understanding of the state-of-the-art FER processes.

The contribution to this Systematic review included as following:

- The primary objective is to analyze and provide a fundamental understanding of existing Machine Learning and Deep Learning techniques within the FER approach. This serves to assist new researchers in comprehending the basic FER process and gaining insights into current work and helps to focus and develop new solutions for existing challenges in FER.
- We present a standard FER database commonly employed in the FER process, encompassing facial expressions of both adults and children. The database includes 3D, video sequence, illumination, pose variation, and occlusion datasets.
- Face detection techniques and their methods are discussed, alongside significant challenges in the face detection process and its performance measured in terms of accuracy.
- The FER process is assessed through two methods, revealing that the Machine Learning (ML) approach is less effective in capturing subtle facial expressions, such as micro-expressions and compound expressions. In contrast, Deep Learning (DL) achieves superior accuracy on FER datasets but requires more time and a larger dataset to prevent overfitting. Additionally, DL FER models necessitate GPUs for training.
- Potential challenges in FER are discussed, and solutions are suggested to enhance the accuracy of the FER process further.

This systematic review is organized as follows: "Selection Criteria" presents the paper selection strategy, review paper selection, and procedures. "Facial Expression Databases" provides a detailed discussion of publicly available both children and adult FER datasets. "Facial Expression Recognition Techniques" describes a brief review of conventional machine learning and deep learning techniques in FER. "Performance Evaluation Mechanisms of FER" presents performance metrics commonly applied in FER analysis. Likewise, "Unresolved FER Challenges and Future Directions" discusses the challenges and opportunities in FER systems. Finally, "Conclusion" summarizes the overview of this systematic study analysis of this work, its findings, and the future scope of FER.

Selection Criteria

Numerous studies on FER have been published in the literature over the last two decades. In conducting this analysis, the Systematic Literature Review (SLR) adhered to ensuring the selection of papers that enhance the quality of reporting, identify existing research gaps on specific problems, and optimize the efficiency of the review paper for both FER

researchers and practitioners interested in addressing significant research problems in a particular area.

For this critical analysis, initially, the SLR review protocol was prepared to collect the literature paper, which guidelines were published in 2007 [75, 108]. After, the review protocol is evaluated as it was recursive. This protocol helped to minimize potential bias in publication. First, research questions (RQ) were formulated by researchers. Four research questions are formulated to guide this SLR:

RQ1: *What is the most significant process in FER using ML and DL techniques?*
RQ2: *What challenges are associated with FER databases?*
RQ3: *How many facial expressions are the main focus of the FER system?*
RQ4: *What are the accuracy and limitations of existing FER studies when using ML and DL techniques?*

The search query and domain list are frequently adjusted until the search results yield significant findings for each target paper based on the RQ. Next, the publications were identified and chosen by searching available databases during paper selection. The paper extraction process is done through authors' details, publication types and years, and other details that were asked in the research question. After this step, data synthesis was made to present an overview of the related studies published till 2023. The review was conducted at the last stage by reporting and answering the search questions. The review should be thoroughly examined in adequate detail for researchers and practitioners to evaluate the comprehensive search. Also, the unfiltered research results have been stored for further analysis as required in the future.

The papers were retrieved from six different standard databases such as IEEE, Web of Science, Scopus, ACM, Springer, and Science Direct. The initial search query was "Facial expression recognition" AND "Facial emotion recognition". Open Access papers, journals, conference materials, and manuscripts were used in the search method. The final search query is as follows: (("Facial Expression Recognition" OR "Facial Emotion Recognition" OR "Facial Expression Analysis") AND ("Emotion Recognition" OR "Emotion Detection") AND ("machine learning" OR "deep learning")). There are many keywords available to search the FER manuscript, but search techniques have used limited keywords for selecting more appropriate papers. Also, the search query was additionally filtered with the years 2002 to 2023, open access papers and the language is English.

All selected papers are carefully validated to exclude the inappropriate papers for this review analysis, resulting in a total of 560 non-duplicated papers. First, 180 publications were selected after a rapid scan of all the titles and abstracts to weed out research that did not apply to the subject of this systematic review or did not match the selection criteria. The remaining studies that did not fit the selection criteria were then removed by carefully reading the entire texts of the remaining publications, leaving 310 papers. The following listed exclusion criteria (EC) were used to exclude irrelevant studies:

EC 1: Theoretical studies and improper FER results in conference papers.
EC 2: Papers are not related computer science field.
EC 3: Survey and short communication papers.
EC 4: Duplicate publication of multiple sources.
EC 5: Older version of FER research techniques before 2002.

Following the listed EC criteria, finally, 105 papers were left for further review. To answer the RQ accordingly, the data from the database was extracted and synthesized. The systematic review's search and exclusion procedures are summarized in Fig. 2. It is significant to note that some included publications were excluded from "Facial Expression Recognition Techniques" because they did not disclose results, were inconclusive, or deviated from the accepted research paradigm. Nevertheless, such studies still contributed to the other parts.

Furthermore, despite the extensive history of endeavours in the field of FER, there have been some notable FER comparisons using conventional machine learning (ML) and deep learning (DL) techniques. Review and survey papers were one of the exclusion criteria during the analysis of retrieved papers. The following review paper has been omitted such as Li et al. [81], presented a detailed review of DL techniques applied in the FER process. Huang et al. [61] compared ML techniques and focused little on DL techniques in FER. Canedo et al. [23] mainly focused on FER using conventional techniques, such as the Convolutional Neural Networks (CNN) approach. Among these, some notable systematic reviews have not focused extensively on face detection techniques and other significant challenges in the FER system, particularly during training with DL and ML techniques using broader facial expression datasets and testing the FER model in real-time environments. To address this gap, this paper systematically analyzes ML and DL techniques, examines existing challenges in emotion recognition and face detection, and explores the challenges in the FER system when utilizing wider facial expression datasets and validating the FER model in real-time environments.

Facial Expression Databases

This section provides some existing popular databases that are frequently used in FER in both the training and testing phases. The performance of the system is affected when

training is not done with sufficient datasets. Most of the facial expression dataset contains six basic emotions plus neutral, which consists only of the frontal face with some challenges like pose variation and illumination. Table 1 shows some popular publicly available databases used in the Facial Expression Recognition (FER) system in recent decades, and it also contains recently released facial expression datasets. In general, FER datasets have been categorized under two conditions: (1) Spontaneous and (2) Posed. Spontaneous datasets are captured naturally, presenting more realistic expressions, and are often considered genuine and authentic. Posed expressions, on the other hand, are deliberately created and controlled by subjects. Early research exploring facial expressions primarily used posed expressions, where respondents were instructed to exhibit or reproduce each basic six expressions, which expression may not always accurately identify participants' emotions on their faces [71, 112]. Most FER systems are still trained using posed expressions along with spontaneous ones due to a lack of a sufficient count of spontaneous expressions in the available dataset. Figure 2 shows the sample facial expression of spontaneous vs. posed.

JAFFE [93]: The Japanese Female Facial Expression (JAFFE) database consists of 213 photographic images capturing the posed facial expressions of 10 female Japanese individuals. Each image resolution is 256×256 pixels. It includes six primary emotions along with neutral expressions. This database is openly accessible for non-commercial research purposes. **KDEF** [91]: The Karolinska Directed Emotional Faces (KDEF) consists of a total of 4900 images of facial expressions. It includes 17 subjects, 7 different facial expressions of 5 different angles. **CK+** [90]: The Extended Cohn-Kanade (CK+) database comprises 593 sequences obtained from 123 subjects. The sequences do not have fixed lengths, and their durations range from 10 to 15 frames. It contains 6 primary expressions besides neutral along with facial landmark location. Out of 593 videos, only 309 were labeled as six basic emotions. It is also available to all kinds of researchers. The images in the database have a pixel resolution of 640×480 and 640×490 pixels, while their grey levels are represented in an 8-bit precision format. **MMI** [111]: The Multimedia Interface (MMI) database includes 740 images and 2900 videos from 32 subjects. A total of 213 images are labeled with six primary emotions. The pixel of the image size is 720×576 resolution. **Multi-PIE**[53]: The Multi-PIE database contains 7,50,000 photography pictures of 337 subjects, which includes 15 viewpoints, 19 illumination conditions, and 5 different facial expressions. **Oulu-CASIA** [168]: It contains six facial expressions from 80 subjects. The size of the frame is 320×420-pixel resolution of 25 frames per second. Also, the camera distance face is about 60 cm. **FER-2013** [52]: The FER-2013 dataset is comprised of 35,887 grayscale images, each with a resolution of 48×48 pixels. The dataset is divided into three sets, with 28,709 images allocated for training, 3589 images for validation, and another 3589 images for testing purposes. Researchers can access and download this dataset for research purposes. **EmotioNet** [46]: EmotioNet is an extensive database that encompasses a collection of one million human faces sourced from the internet. This database is accompanied by annotations for each expression captured in the images. Approximately, 9,50,000 photo images were annotated through Action Units (AUs) with the help of the intensity of their emotions. The remaining 25,000 images were annotated manually by 11 action units. **AffectNet** [102]: The AffectNet database includes more than one million human faces collected from the World Wide Web. Those images are manually labeled for 8 facial expressions (happy, sad, neutral, angry, fear, surprise, disgust, contempt) based on the strength of valence and arousal space. Researchers can access this dataset by email request. **Ferv39k** [153]: The Ferv39K is a large-scale video sequence database, which has been collected from real-time video clips from various real-world contexts such as scenes, movies, TV, live shows, and official events. It contains 38,935 video clips of 7 labeled facial expressions. **PEDFE** [99]: The Padova Emotional Dataset of Facial Expression (PEDFE) newly created dataset with six universal expressions. It contains 1458 facial images of 56 participants. **UIBVFED-Mask** [95]: The UIBVFED-Mask dataset is an extended version of the UIBVFED [107] dataset. Recognising facial expressions from occlusion is a challenging task due to the loss of significant facial expression information. This dataset contains 640 images of 32 facial expressions of 20 participants.

NIMH-ChEFS [41]: The NIMH-ChEFS (National Institute of Mental Health Child Emotional Faces Picture Set) comprises a set of 482 stimuli showcasing angry, fearful, neutral, and sad facial expressions exhibited by children. This dataset is useful for affective and neuroscience research. **RaFD** [77]: The Radboud Faces Database (RaFD) consists of a comprehensive collection of images featuring individuals of different age groups, including adults and children. This database encompasses six fundamental emotions, in addition to neutral and contempt expressions. Each emotion has been created with five different camera angles and three various gaze directions. **DDCF** [35]: The Dartmouth Database of Children's datasets contains 40 female and 40 male photographs in a black background and aged between 6 and 16 years. This photograph was collected with 5 various camera angles (60° right,30° right,0°,30° left, 60° left) and 8 facial expressions (neutral, content, happy, sad surprised, afraid, angry, and disgusted). **CAFE** [87]: The Child Affective Facial Expression (CAFÉ) set contains 1192 photographs taken from 2 to 8-year-old children with 154 subjects. This database includes six basic expressions along

Table 1 Commonly used existing Facial Expressions Datasets in FER systems

Name of the database	Dataset	Subjects	No. of expression	Condition	Age group	Posed/spontaneous	Created/released year
JAFFE [93]	213 images	10	7	Lab	Adults	P	1998
KDEF [91]	49,000 images	140	7	Lab	20–30	P	1998
CK + [90]	593 images	123	7	Lab	Adults	P & S	2010
MMI [111]	740 images, 2900 videos	32	7	Lab	Adults	P & S	2005/2010
Multi-PIE [53]	755,370 images	337	6	Lab	20–30	P	2010
Oulu-CASIA [168]	2,880 images	80	6	Lab	23–58	P	2011
FER-2013 [52]	35,887 images	N/A	7	Internet	20–60	P & S	2013
EmotioNet [46]	10,00,000 images	N/A	23	Internet	N/A	P & S	2016
AffectNet [102]	4,50,0000 images	N/A	7	Internet	N/A	S	2017
Ferv39k [153]	38,935 videos	N/A	7	Internet videos	N/A	S	2021
PEDFE [99]	1458	56	6	Lab	20–30	P	2022
UIBVFED-Mask [95]	640 images	20	32	Lab	20–80	P	2023
Children facial emotion dataset							
NIMH-ChEFS[41]	482 images	59	4	Instruction	10–17	P	2004
RaFD [77]	5880 images	10	8	Instruction	N/A	P	2010
DDCF [35]	Unspecified	123	8	Imagination	6–16	P	2013
CAFE [87]	1192 images	154	6	Imitation	2–8	P	2014
CEPS [126]	237 images	18	7	Mimic a photo	6–11	P	2015
EmoReact [106]	1102 Audio, Video	63	17	YouTube, Facebook	8–12	Unspecified	2016
DuckEES [51]	142	37	8	Lab	8–18	Unspecified	2017
LIRIS-CSE[74]	208 videos	12	6	Lab	6–12	S	2019
ChildEFES[105]	1985 images and Videos	124	8	Lab	4–6	P & S	2021

P posed; *S* spontaneous

Fig. 2 Sample facial expression of posed (Ck+) [90] vs spontaneous (AffectNet) [102]

with neutral emotions. It is freely available for researchers. **CEPS** [126]: The Child Emotions Picture Set (CEPS) contains six basic emotions of 17 children. Those emotions are created with three divergent intensities (low, medium, and high). **EmoReact** [106]: EmoReact is a comprehensive multimodal emotions database specifically designed for children aged 4–14 years old. It comprises 1102 audio-visual clips, each meticulously annotated with 17 distinct emotions. These emotions encompass the six primary emotions, neutral expression, and valence, as well as nine complex emotions, including curiosity, uncertainty, and frustration. **DuckEES** [51]: The University of Oregon ("Duck") Emotional Expression Stimulus (DuckEES) database consists of 142 dynamic video sequences featuring children between the ages of 8 and 18 years. This database encompasses a range of facial expressions, including negative emotions such as disgust, embarrassment, fear, and sadness, as well as positive emotions like happiness and pride. Additionally, neutral expressions are also included in the dataset. **LIRIS-CSE** [74]: The Children's Spontaneous Facial Expression (LIRIS-CSE) contains 208 video sequences from 12 ethnically diverse children. It includes spontaneous (natural) six universal facial expressions (happy, sad, anger, surprise, disgust, and fear) along with videos created in the open environment (no restriction for head and hand moments) conditions. **ChildEFES** [105]: The Child Emotional Facial Expression Set (ChildEFES) dataset encompasses a video sequence consisting of 1,985 instances featuring children aged between 4 and 6 years. Within this dataset, there are varying numbers of stimuli for each emotion, including 87 neutral, 363 happiness, 170 disgust, 140 surprise, 152 fear, 144 sadness, 157 anger, and 183 contempt expressions. Researchers can freely access and utilize this dataset for non-commercial research purposes.

Furthermore, existing FER datasets have some limitations, such as a lack of large-scale expressions, varied sizes, inconsistent image quality, and susceptibility to indoor and outdoor conditions. However, numerous solutions have been proposed to overcome these challenges. In cases of low image quality, a diverse range of cleaning and enhancing pre-processing techniques is required to improve FER accuracy. Generally, FER models are trained on a limited number of facial expressions. However, deep learning techniques necessitate a large amount of data for effective training and to attain promising results. Therefore, data augmentation techniques can be applied in this context.

Facial Expression Recognition Techniques

The conventional FER approach using ML techniques comprises four key methods: face detection, pre-processing, feature extraction, and emotion classification or recognition, as illustrated in Fig. 3.

Machine Learning-Based FER Approach

Face Detection

Face detection holds a significant role as the initial step in many areas including face alignment, face recognition [15], face verification, and recognizing emotions from face images or video sequences [17, 57]. It serves as a crucial component in the process of emotion recognition. The primary objective of face detection is to ascertain the presence of a face within an image, as this region is vital for emotion detection as well as facial recognition of individuals, as depicted in Fig. 4. The accurate detection of faces lays the foundation for subsequent steps in the process. In addition, face movement has been tracked in the video sequence. It is part of object detection used in many places like security, biometrics, personal safety, and so on. Face detection and facial recognition are not the same but are interrelated [16]. Face detection allows a system to identify the presence of a human face in an image or video sequence, whereas face recognition can indicate the name of the person in that image.

Different Approaches of Face Detection

Different face detection methods, including knowledge-based, feature invariant, template matching, and appearance-based techniques, are employed for detecting faces, as depicted in Fig. 5. These methods encompass a range

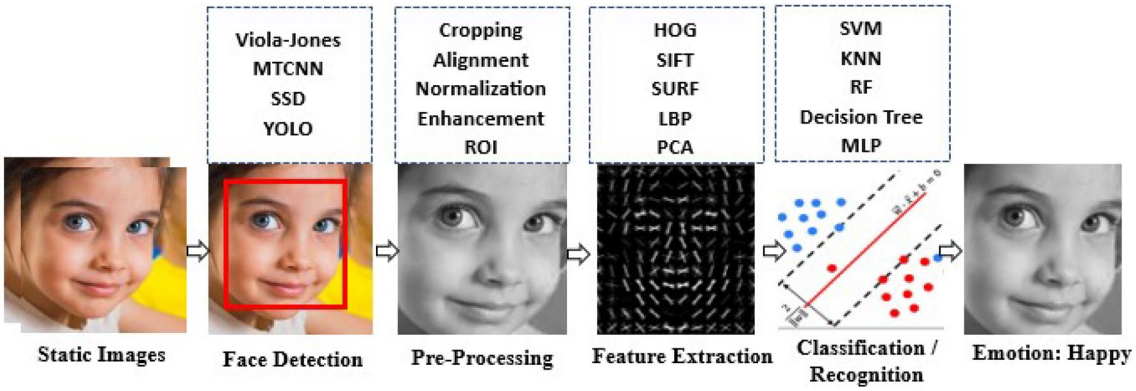


Fig. 3 Conventional FER in machine learning

(a) Original Image (b) Detected Face Image

Fig. 4 Face detection from digital images

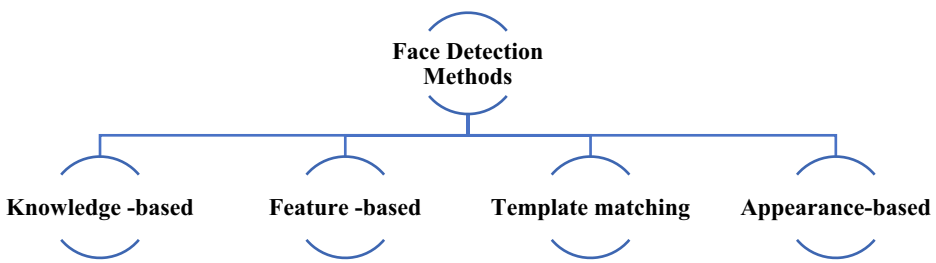
of approaches used to accurately identify and locate faces within images or video sequences.

Knowledge-Based The knowledge-based method follows a set of rules developed by humans [16]. Firstly, facial features such as the nose, eyes, mouth, shape, size, texture, etc., are obtained from images. Secondly, this type of face detection process relies solely on predefined rules, which are performed based on prior knowledge of facial geometry created by human-crafted knowledge [57]. However, these methods prove to be less effective in real-world scenarios, such as faces with illumination, pose variation, and diverse facial features. Nevertheless, this method was found to be useful for front-face images and in well-controlled environments.

Feature-Based Feature-based methods extract structural features from images [144]. These structural features include skin color, shape, texture, and facial local features. Other features such as eyes, mouth, eyebrows, and nose are extracted with the help of filters. According to some studies, skin color is considered one of the best features for detecting a face in images [160]. For instance, methods like HOG [1] and Viola-Jones [60, 151] are employed for feature extraction to identify faces or objects in images. However, these features are sometimes corrupted due to factors such as high illumination, face orientation, occlusion, and noise, as they heavily rely on the visibility of specific features Furthermore, these methods may fail to capture the full complexity of facial patterns in diverse datasets. Also, feature boundaries can be weakened for the face, and shades can cause strong edges, collectively making existing feature extraction methods less fruitful in certain scenarios.

Template Matching The template matching method is a straightforward approach that detects faces through the correlation between pre-determined face templates and input images using a predefined or parameterized face template [100]. The edge detection model and filters are employed to construct edges in images. However, this method has some limitations. Accuracy is affected by real-world scenarios with diverse conditions, and computation time increases when searching for faces in larger images. Additionally, it

Fig. 5 Face detection methods



suffers from overlapping faces with complex backgrounds, leading to a higher false-positive rate. Furthermore, selecting an efficient template for various conditions is less robust, as different scenarios may require different templates. Moreover, illumination on the face image also impacts the performance of the template matching method, resulting in non-face region detection.

Appearance-Based The appearance-based method identifies relevant patterns and features from facial images using a machine learning algorithm instead of explicit templates. Commonly used techniques in face detection include Haar-like features [144], PCA [160], SVM, Eigenface, Hidden Markov Model, Naïve Bayes Classifier, and CNN [82]. Furthermore, this method is considered robust compared to other face detection techniques, especially in challenging real-world environments. Nowadays, most face detection techniques are developed based on this approach. However, it often requires a large and diverse set of annotated face and non-face images for training; otherwise, it may suffer from overfitting. Additionally, it may struggle with low-quality image resolution, partial occlusion, noise, blurriness, or compression, and the computational cost is high when developing a face detection model with large real-world images using deep learning techniques.

Table 2 shows the comprehensive study of the existing face detection methods that have been widely used in recent years. This analysis considers the factors related to the type of techniques used in both pre-processing and face detection, the name of the datasets used for training and testing the model, the platform environment to develop the face detection model, and finally permeance of the developed face detection algorithm. During the last decades, several face detection approaches have been proposed and tested in various environments with many conditions. To begin with, You Only Look Once (YOLO) [11, 29, 163] is a real-time object detection neural network-based algorithm used in face detection. It contains three techniques namely, residual block, bounding box regression, and intersection over union (IOU). For face detection, YOLO is trained with a larger dataset for face detection like FDDB, Wider face, and Celeb face benchmarks. The VGG-16 and DarkNet-53 model has been used to extract features from the images before applying the YOLO algorithm. YOLO algorithm detects the various positions, illumination, and different skin complexations in real-time. However, it suffers from the precise location of small and multiple faces, and different scales of face image on the real-time scenario

Similarly, Dual-Branch Center Face detector (DBCFace) [82], and ResNet-50 [146] are convolution neural network-based algorithms designed to reduce problems in non-maximum suppression (NMS). It has trained with various databases like AFW, PASCAL face, FDDB, and WIDER FACE. It detects the occulted face with higher accuracy. Adaboost [38] is also known as an adaptive boosting machine learning-based ensemble method which is used to find out the strong features and to identify the face in the YCbCr color model. Multi-task Convolution neural networks [69, 165]

Table 2 Existing face detection techniques from the year 2001 to 2023

Ref	Techniques	Dataset	Accuracy/result
[11]	You Only Look Once (YOLO), Vgg-16	FDDB, Real-Time Live Video	Achieved 95% average precision
[82]	Dual-Branch Center Face detector (DBC-Face)	AFW, PASCAL face, FDDB, WIDER FACE	Achieved 90.34% accuracy
[146]	Soft-NMS, Resnet-50	FDDB	Achieved 94.2% accuracy
[160]	Improved AdaBoost	Real-Time Data	Reduced false detection rate
[165]	MTCNN	WIDER FACE	Obtained 85.7% results
[29]	YOLOv3, Darknet-53	WIDER FACE, FDDB,	Achieved 93.57% accuracy
[123]	Improved MTCNN	MIT, Casia, NICE-II	Achieved 98% accuracy
[163]	You Only Look Once (YOLO)	WIDER FACE, Celeb Faces, FDDB	Achieved efficient face detection time than the traditional algorithm
[67]	Three-category face detector, Fast R-CNN	WIDER FACE, FDDB	Obtained 97% true positive rates
[119]	Haar Cascade	Instagram Selfie Images	Achieved 71.48% accuracy
[37]	Single-Stage Joint	WIDER FACE	Achieved 56.66% average precision
[33]	Haar Cascade	Open internet images	Achieved a Positive Prediction Value (PPV) of 98.01%
[152]	Region-based Fully Convolutional Networks (R-FCN),	WIDER FACE dataset, FDDB dataset	Achieved True Positive Rate 98.99%
[60]	Viola-Jones	MIT, FERET	Achieved 98.97% accuracy
[69]	Compact CNN	FDDB	Achieved high speed compared with traditional GPUs and CPUs

and Viola-Jones [33, 60] algorithms are used to solve the numerous challenges in face detection methods. MTCNN consists of three networks namely the Proposal Network (P-Net) gives more false-positive predictions, the Refine Network (R-Net) which uses the NMS method to reduce the false positive rate, and the Output Network(O-Net) which gives a more accurate face position with five landmark locations of eyes, nose, and mouth corner [123, 165]. Those networks are not connected directly but the result of one network is given to the input of another network.

Furthermore, Viola-Jones [60] is a conventional face detection algorithm widely used to figure out a front face in digital images. This method follows the four main steps namely Haar-like feature, integral images, AdaBoost, and cascade classifier. The Haar-like feature is used to extract features from the images [33]. These features are transformed into pixel values with the help of integral images. AdaBoost algorithm selects the important features from whole extracted features [160]. A final cascade classifier is used to discard the non-face in an image which speeds up the face detection process. Viola-Jones algorithm gives a more false-positive rate when the face angle is over 45^{o} and above. Single-stage joint face detection [37] is a fully convolution neural network. It contains three components: feature pyramid network gets input face and outputs five scale feature map; context head module calculates multi-task loss from feature map and cascade multi-task loss predict the bounding box from the regular anchor. Region-based Fully Convolutional Networks (R-FCN) serve as an object detection framework [67, 152]. This framework incorporates a ResNet with over 101 layers to extract features from facial images, resulting in enhanced accuracy for face detection in benchmark datasets such as WIDER and FDDB. Moreover, R-FCN has demonstrated superior performance in accurately detecting faces compared to other methods.

In summary, ML face-based face detection methods are less robust in real-time environments because this approach relies solely on human-crafted features. They also suffer from challenges such as high illumination, pose variation, and face orientation. On the other hand, the neural network-based approach has demonstrated significant face detection performance in real-time scenarios, especially in challenging environments. CNNs, in particular, are capable of learning relevant patterns and representations for face detection without human intervention for feature identification. However, DL methods, especially those involving large neural networks, may have higher computational requirements during both training and inference compared to traditional ML methods.

Figure 6 illustrates the accuracy rates of various face detection techniques. The x-axis represents the names of the face detection methods, while the y-axis indicates the corresponding accuracy levels obtained from Table 2. The graph provides a visual representation of the performance of each face detection technique in terms of accuracy.

Challenges in Face Detection

The advancement of computer vision over the last few decades has made research more efficient. The challenges in face detection/face recognition affect the quality of the outcome [94]. The face detection process is difficult due to the various sizes of the face, pose variation, occlusion [138], aging, noise, low resolution, and illumination [100]. Figure 7 shows the challenges that occur in face detection and recognition.

Illumination and Pose Variation To begin with, illumination refers to lighting conditions and the presence of shadows on the face images that cause a cluttered background

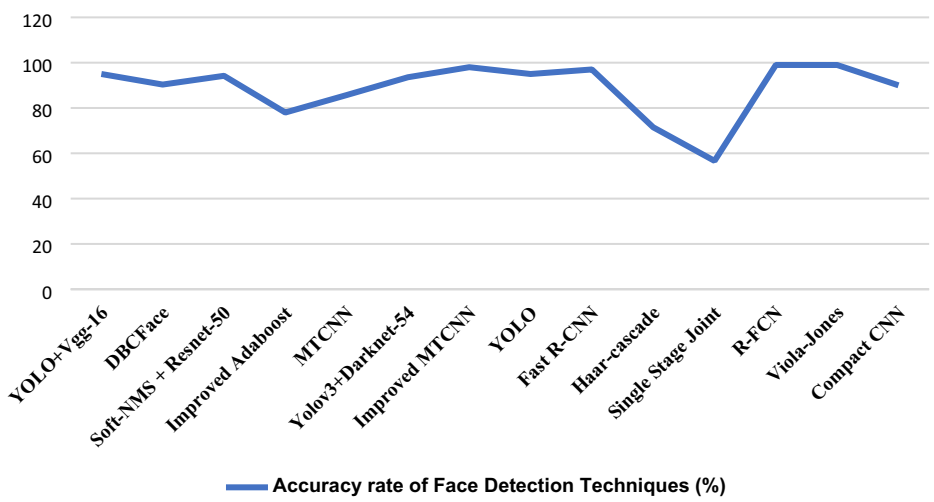


Fig. 6 Accuracy rate of face detection techniques

for expression images, which can pose challenges in facial detection and recognition. Secondly, the pose of a face can significantly varies due to head movements and changes in viewing angles, leading to inaccuracies or failures in face recognition or detection. Facial expressions constantly change at both macro and micro levels. In 2D facial images, to reduce computational costs, the extraction of spatial features is extremely difficult. Similarly, in 3D images, a side view could affect the system's performance. However, most existing facial expression datasets were collected in a controlled environment, where expressions are static and captured by both professional and non-professional actors. Consequently, the performance of FER is degraded in the real world, where spontaneous and sequential facial images are common.

Aging Similarly, aging involves changes in facial appearance and texture over time, presenting a significant hurdle in accurate detection and recognition due to alterations in features, shapes, lines, and other aspects of the face. To support this claim, studies suggest that when investigating the performance of optical flow and high gradient detection on infants, the developed algorithm showed lower performance on infant facial images than on adults [31]. The reason is that infant skin texture, fatty tissues, and the absence of transient furrows reduce the algorithm's performance. This claim is further emphasized by Ref. [70], indicating that different physical appearances, such as skin texture, can affect FER performance. This is also the main reason for not combining multiple facial expression datasets when training the FER model.

Partial Occlusion A partial occlusion occurs when certain parts of the face are blocked, resulting in incomplete input images where the entire face is not available for detection. These factors collectively contribute to the complexities and challenges faced in facial detection and recognition [84]. These challenges extend to natural occurrences such as beards, wearing glasses, hijabs, mustaches, cosmetics, and headscarves. Moreover, in recent times, many studies have been designed to detect faces even when individuals are wearing face masks [127, 134–138], and this technique could prove fruitful in overcoming these kinds of challenges.

Pre-processing

Pre-processing plays a crucial role in enhancing the efficacy of FER systems before the facial feature extraction process. It involves a series of steps aimed at refining, reducing redundant information, and optimizing the input data features, ultimately improving the accuracy and reliability of the FER system. This process is essential in both conventional machine learning and deep learning methods. Moreover, this phase consists of different types of processes such as face localization, facial landmark detection, normalization, and augmentation. Additionally, it includes various image-enhancing techniques like scaling, contrast adjustment, improving image clarity, histogram equalization, gamma correction, pixel brightness transformation, Fourier transform, and filtering.

Face Localization Face Localization is generally used to detect the region and size of the human face in images or video sequences [125]. This approach removes unrelated background information that can affect prediction accuracy. Moreover, to detect the face from the input images, Viola-Jones [60], Haar features [33], and the AdaBoost algorithm [163] have been actively used for decades. These algorithms were optimized for speed with integral images. However, Viola-Jones has some shortcomings, including non-robustness in partial occlusion and pose variation. In addition, Region of Interest (ROI) segmentation is one of the most

Fig. 7 Typical challenges in face detection

important functions used in face localization to identify and mark the facial organs [113]. Nevertheless, the method fails to implement bounding box regression. This study [92] has explored the issues in bounding box regression using a CNN model to determine if the bounding box accurately fits on the face. The authors applied these steps iteratively until they achieved a significant face location in face images or a sequence of face frames.

Face Alignment On the other hand, face landmark detection (face alignment) is used to mark facial features such as eyebrows, mouth corners, eyes, and lips as shown in Fig. 8. This process is performed after the face detection step. It serves as another pre-processing method for determining the geometrical model of the human face [54, 83]. This landmark improves the FER system performance, for instance, the SIFT algorithm is often employed to identify facial features. Subsequently, all facial expressions are aligned using related reference images. The facial landmarks visualize parts of the human face such as the eyes, mouth, nose, eyebrows, and jawlines, as shown in Fig. 8. Normally, it is used for images or video sequences to detect faces and objects [68, 156, 167]. Also, a comprehensive review is available in this face mark localization for readers [68]. After, the advancement of deep learning methods, face landmark detection became easier and more proved its superiority over existing machine learning-based methods.

Face Normalization Face normalization is an important pre-processing technique employed to alleviate the impact of irrelevant and redundant information, such as background, hair, and clothing, in order to streamline the detection process [142, 144]. By removing these non-essential elements, face normalization aims to enhance the effectiveness and efficiency of the recognition process, focusing solely on the facial region of interest. In layman's terms, normalization is a method of rotating a non-frontal facial expression to a frontal pose in terms of improving face recognition. For normalization, Euclidean points are being used to measure the position between facial features [27]. Furthermore, several pre-processing methods have been used but ROI and histogram equalization are widely applied in FER pre-processes. Also, cropping and scaling were decided to apply to the face images, with the nose of the facial parts chosen to take as the central axis as well as other points physically involved. Noise reduction is to reduce the noise from facial images as median filter (MF), adaptive median filter (AMF), gaussian filter (GF), and bilateral filter (BF) are often used as filters in FER systems. Gaussian filter is used to resize the image which provides the smoothness of the images. Similarly, Histogram Equalization (EQ) is a pre-processing method used to enhance color contrast in image histograms which separates the most frequency intensity pixel values [110], which is commonly used in FER.

Fig. 8 Facial landmark detection (ChildEFES dataset [105])

Data Augmentation Data augmentation is a technique used to increase the dataset size through computational manipulation, including flipping, cropping, rotation, zooming, scaling, and many more. This approach helps improve the performance of machine learning models, particularly deep learning models. Data augmentation can be implemented in two ways: (1) *offline approach* is employed when data should be stored in a separate folder after augmentation, and (2) the *online approach* dynamically augments data during training. This approach is widely applied in all deep learning (DL) techniques. However, some existing literature and Facial Expression Recognition (FER) papers have suggested [139, 169] that automatic augmentation can introduce possible biases through a random selection of samples and incorrect augmentation policies. Recently, the AutoAugment approach, using reinforcement learning, has been introduced [32], but it is computationally expensive. Similarly, population-based augmentation (PBA) [58] and population-based training (PBT) [63] have been presented, but they have not achieved significant results. In computer vision, robust research on data augmentation is still open to researchers.

Facial Feature Extraction Methods in ML

The feature extraction method of the FER system is the subsequent phase after the pre-processing stage. This stage involves extracting and highlighting useful information from unstructured data [3, 16]. This approach helps reduce potential biases in recognizing facial expressions from a vast number of features, playing a crucial role in Computer Vision. This section comprehensively discusses the existing machine learning feature extraction techniques as follows:

Global and local feature extractors are the two types of feature extractors commonly utilized in digital photographs [110, 164]. For image retrieval, object detection, and classification, global descriptors are used. Local descriptors, on the other hand, are utilized in object detection and identification. Moreover, PCA is a dimensionality-reduction approach for extracting local and global level dimensional information [38, 133]. Through multi-channel observation, Independent

component analysis (ICA) retrieves local features. The feature extraction approach stepwise linear discriminant analysis (SWLDA) features extracted from forward and backward linear regression. It is determined by the estimated class label F-test values for regression models. The local curvelet transform (LCT) serves as a geometric feature descriptor that effectively captures geometrical features through a wrapping mechanism. This technique extracts features such as mean, median, and standard deviation. Additionally, energy and kurtosis characteristics are obtained by utilizing three-stage directional pyramid representations. These extracted features contribute to a comprehensive understanding of the geometric properties of the analyzed data. The Gabor Filter serves as a texture descriptor utilized for feature extraction, encompassing both magnitude and phase parameters [19]. The magnitude feature provides limited information regarding the arrangement of facial image components, while the phase feature complements it by providing a more comprehensive description. Together, these features offer a comprehensive representation of the texture characteristics present in the face image. LBP is a texture descriptor that is used to retrieve features from images [121]. It generates binary code, which can be obtained by differing the threshold levels between both the center and locality pixel resolution.

The HOG feature descriptor is a window-based technique that leverages gradient filters to extract features from images [38]. Specifically, it focuses on the edge information derived from authorized facial expression images. The HOG descriptor captures visual characteristics, such as smile expressions characterized by curved-shaped eyes. By analyzing the gradients and orientations of local image regions, HOG effectively captures key facial expression features. Similarly, the active shape model (ASM) [40] is a mathematical prototype model that is frequently used to extract feature marks from a facial expression by incorporating local texture characteristics [40, 89]. To handle the high-dimensional nature of extracted features, various dimensionality reduction techniques, such as PCA and linear discriminant analysis (LDA), are employed. These techniques aim to reduce the dimensionality of the feature vectors while retaining the most important information. Additionally, different algorithms, such as Viola-Jones and similar ones, are utilized to select the most relevant features, further enhancing the efficiency and effectiveness of the overall facial expression recognition process.

In summary, ML-based feature extraction techniques are not commonly applied in FER systems these days due to the handcrafted feature extraction approach. Additionally, these techniques are not sufficient for extracting subtle, intricate, and complex patterns in facial expressions. Furthermore, the ML approach faces challenges in handling high-dimensional data in the FER dataset, as well as unseen data in real-time scenarios with variations in illumination, pose, and extreme facial expressions. Moreover, it is a time-consuming process to find relevant features to build robust FER systems.

Classification/Recognition Approach of FER in ML

The final stage of the FER system involves emotion recognition or classification, responsible for predicting specific labels such as happy, neutral, sad, disgust, anger, surprise, fear, boredom, confusion, and frustration, given input images. This is illustrated in Fig. 9. The FER system is typically trained with six primary emotions (happy, sad, anger, fear, surprise, disgust) [43, 44] due to the limited availability of facial expression counts in the FER dataset. Additionally, compound facial expressions are formed through the combination of two basic emotions, resulting in a more nuanced emotional state [39]. In the context of facial expressions, a total of twenty-one emotions are represented, including the six main emotions, along with a neutral expression. Furthermore, there are twelve compound emotions that individuals can exhibit, along with three additional emotions (awed, appalled, and hated). On the other hand, micro-expressions refer to spontaneous and subtle facial muscle movements that occur impulsively [21, 117]. These micro-expressions

Fig. 9 Facial expression in AffectNet [102]

are often unobtrusive and provide valuable insights into an individual's underlying emotions. They are incredibly short, lasting approximately 1/25 to 1/3 of a second [61].

Table 3 presents a comprehensive examination of the most applied machine learning algorithms in FER systems. The study considers factors such as the year, feature extraction techniques, machine learning algorithms for facial expression classification, types of datasets, the number of expressions, and the accuracy of each method in FER.

To begin with, HOG [1], Gabor filter [19], PCA [38], HAAR filters [56], Log-Gabor filter [121], LBP [121], and ASM [7, 89] are feature extraction methods used to extract facial features from pre-processed images. This approach has already been discussed in the previous section. The FER system targets more than five emotions and commonly utilizes FER datasets such as CK, CK+, JAFFE, MMI, and PIE. Viola-Jones [20, 150, 159], Haar features, and the AdaBoost algorithm are employed to identify a face in digital images, as facial emotion recognition requires only the face region for processing and classifying emotions. Ensemble techniques (combining more than one technique) are used to enhance facial emotion recognition efficiency before introducing the deep learning approach.

Upon closer examination, the Support Vector Machine emerged as a powerful supervised learning algorithm commonly employed for classification tasks in FER [1, 19, 140]. SVM works by creating an optimal line or decision boundary, known as a hyperplane, that effectively separates different classes in the data. Through a recursive process, SVM iteratively generates the best possible hyperplane to minimize errors and maximize the margin between data points of different classes. This enables SVM to make accurate predictions and handle complex classification and regression problems effectively. SVM employs the maximum marginal

Table 3 Conventional FER in machine learning

Ref	Feature extraction techniques	Classifier	Dataset	Emotions	Accuracy/result
[140]	Improved Cat Swarm Optimization (ICSO), DCNN,	SVM, Neural network	JAFFE, CK+, PIE, Real-world images	Normal, Happy, Sad, Surprised, Anger, Fear	It gave good results in comparison with the four datasets
[1]	Viola-Jones, HOG	SVM	Own Dataset	Smile and No smile Speech: Crying and No crying	Speech: 85.72% FER: 92.88%
[19]	Gabor filter	SVM	JAFFE, CK, CK+	Angry, Contempt, Disgust, Fear, Happiness, Sadness, Surprise	JAFFE:96.30%, CK:94.20%, CK+:94.26%
[159]	Modified Viola-John's	KNN, SVM	JAFFE, LNMIIT, CK+, MMI	Neutral, Happy, Sad, Fear, Disgust, Surprise, Anger	MMI:97.5, JAFFE: 97.65, LNMIIT:99.77, CK+:98.56
[38]	Viola-Jones-face detection, HOG-Feature extraction, PCA-reduce dimensionality of the feature	SVM, KNN, and MLPNN	CK+	Neutral, Anger, Contempt, Disgust, Fear, Happy, Sadness, Surprise	93% of accuracy
[12]	Viola-Jones	KNN, SVM, RF, CART	N/A	Happy, Surprise, Sad, Anger, Disgust, Fear	98.24% of accuracy
[56]	HAAR filters	SVM	CK, CK+	Neutral, Happy, Sad, Anger, Contempt, Disgust, Fear, Surprise	93.7% of accuracy
[115]	RST-Invariant features and texture features	KNN, SVM, ANN	JAFFE	Happy, Anger, Sadness, Fear, Disgust, Surprise	90% of accuracy
[121]	Viola-Jones, Haar feature, AdaBoost learning, Log-Gabor filters, LBP	SVM	CK+	Anger, Disgust, Fear, Happy, Sad, and Surprise	79% of accuracy
[7]	Active Shape Model (ASM) tracker	SVM	CAFÉ (Child Affective Facial Expression)	Surprise, Anger, Happiness, Sadness, Fear, and Disgust	93% of accuracy
[20]	Haar features, Viola, and Jones, AdaBoost Classier (EmguCV, OpenCV)	SVM	Real-time face Stimuli	Joy, Sad, Surprise, Neutral	87.9% of accuracy

hyperplane (MMH), facilitating the classification of features into six distinct emotions. Similarly, K-Nearest Neighbor is also a machine learning-based supervised algorithm. KNN [115] can be used in both classification and regression but is mostly employed for classification purposes. It saves all available occurrences and categorizes new cases based on image similarity measurement (distance), which is used for better clarification and quick calculation. The training sample is stored in n-dimensional space for analysis.

Random Forest (RF) [2] is an ensemble learning technique that combines multiple classifiers to address complex problems and enhance model performance. It operates in two phases. In the first phase, the random forest is created by combining multiple decision trees. In the second phase, predictions are made by each individual tree generated in the first phase. RF offers several advantages, including reduced training time compared to other algorithms, high prediction accuracy, and efficient performance with larger datasets. Similarly, one popular decision tree algorithm used within RF is classification and regression trees (CART) [12]. CART employs a tree-based structure and utilizes the if-then-else rule to predict outcomes for data points. This algorithm finds applications in both facial emotion recognition and facial expression recognition tasks. By leveraging the inherent decision-making capabilities of decision trees, CART contributes to accurate predictions and the effective recognition of facial emotions and expressions. Moreover, a nature-inspired algorithm was also applied to optimize the feature search techniques. For instance, cat swarm optimization (CSO) is an intelligent optimization technique inspired by the behavior of cats and operates in two distinct modes: seeking mode and tracing mode [140]. In the seeking mode, the cat assumes a relaxed position, while in the tracing mode, it mimics the behavior of a cat searching for prey. This swarm-based approach draws upon the natural instincts and behaviors of cats to efficiently explore and exploit search spaces, leading to effective optimization results. CSO is used in the FER system to select the best features to improve FER recognition accuracy. Additionally, particle swarm optimization (PSO) [85] and ant colony optimization (ACO) [9, 161] have been used in FER systems in the past.

Furthermore, neural network-based techniques have begun to be applied in the FER system. This architecture, consisting of three layers, includes the input layer, which receives the input data; the hidden layer, responsible for processing and transforming the input; and the output layer, which produces the classification results. This neural network architecture enables the model to effectively learn and extract meaningful features from facial expressions, facilitating the accurate classification of various emotions in this network, parameters such as the classification feature vectors, the dimension of the feature vector, and the total number of classes, such as happy, sad, surprise, neutral, angry, and fear, are used. One of the major strengths of feedforward artificial neural network techniques is the multiple-layer perceptron neural network (MLPNN) [38]. The neurons in the MLP are trained with a backpropagation learning algorithm for classification, recognition, approximation, and prediction.

In summary, ML classifiers have struggled to generalize well to diverse FER datasets and real-world scenarios. A classifier trained on one FER dataset may not be effective when tested with unseen facial expressions, considering variations in illumination, pose, and facial expressions. Moreover, these classifiers are designed to handle static images independently, making them insufficient for extracting and classifying temporal facial image sequences. Additionally, ML classifiers require a substantial amount of well-annotated FER datasets and may encounter issues related to class imbalance, leading to reduced classification accuracy in FER. Furthermore, ML classifiers face challenges in accurately recognizing emotions from facial expressions when dealing with variations across different individuals.

Deep Learning-Based FER Approach

Deep learning is a field within machine learning that draws inspiration from the functioning of the human brain, specifically neural networks. It aims to develop algorithms and models that can learn and make predictions by mimicking the complex processes of the brain. There are several kinds of deep learning techniques such as Artificial Neural Networks (ANN), Autoencoders, Recurrent Neural Networks (RNN) [120], and Reinforcement learning are shown in Figs. 10 and 11. Among all, Convolutional Neural Networks (CNN) or ConvNets are frequently adapted in FER image processing, The main advantage of this method is combining the feature extraction and classification parts, which greatly reduces the handcrafted feature extraction process and its challenges. The following subsections present an overview of CNN and transfer learning (TL) [5, 86] techniques that are commonly employed in FER systems. In addition, state-of-the-art techniques like autoencoder, GAN, Bi-LSTM, RNN, reinforcement learning, and ensemble methods have recently been applied in the FER system. These techniques increase the accuracy of emotion recognition when compared to conventional approaches and are also highly efficient for extracting spatiotemporal features in a sequence of facial expressions [34, 139].

Table 4 presents a comprehensive study of a frequently used deep learning algorithm in FER systems. This study considers factors related to deep learning techniques, datasets, the number of emotions, accuracy, and results across various datasets. The conventional FER system typically involves three main methods: face detection, feature extraction, and the classification of emotions such as happy, fear,

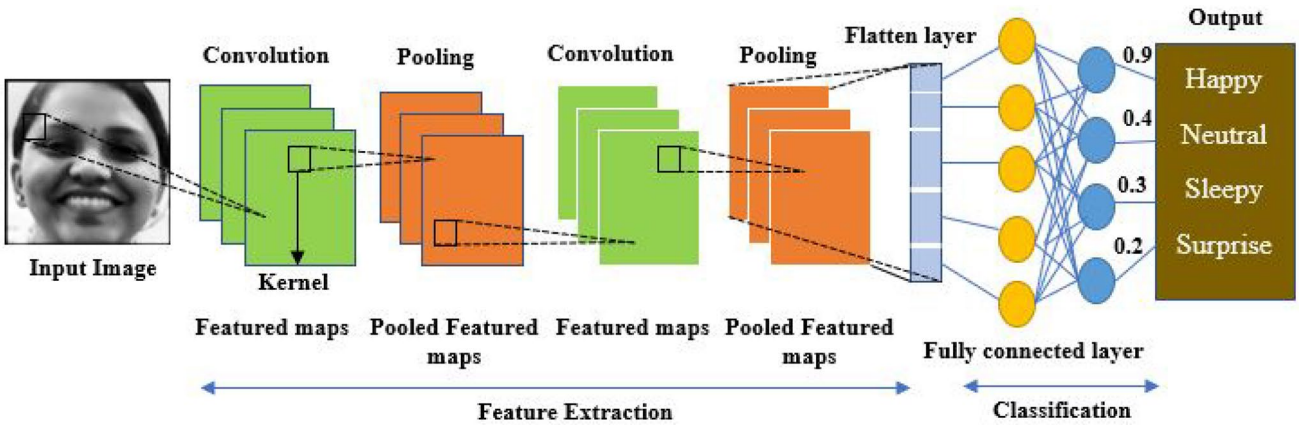


Fig. 10 Training process of FER in CNN models

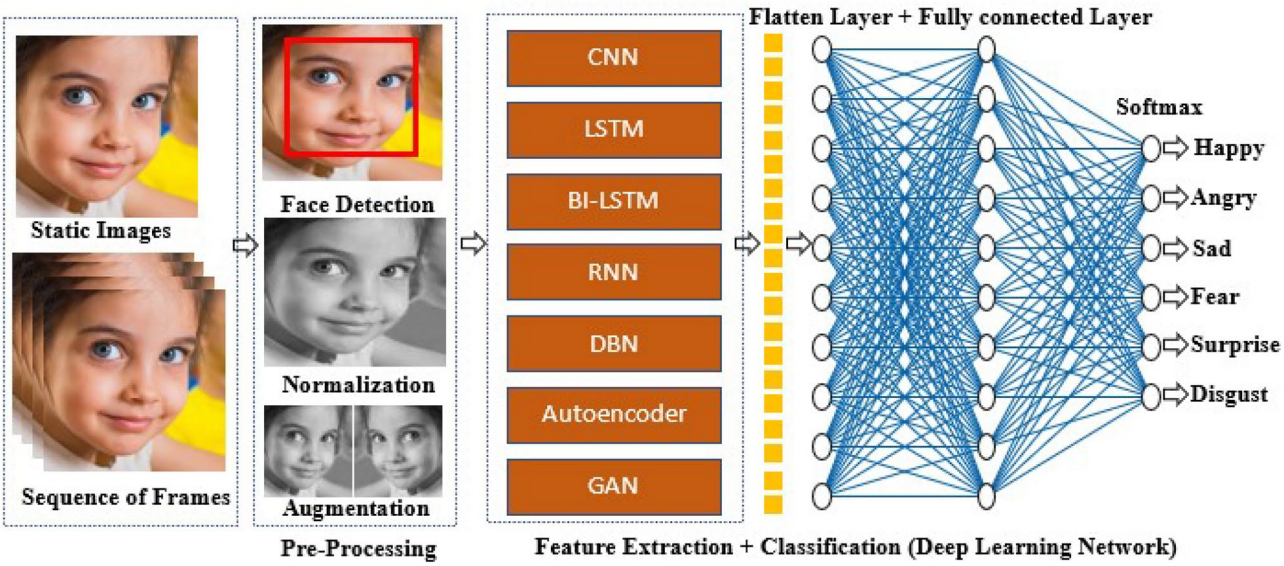


Fig. 11 FER process in deep learning approach

sad, anger, neutral, surprise, and fear. First, pre-processing techniques are applied to facial images, including histogram equalization and conversion to grayscale. Then, face detection techniques such as Viola-Jones, Haar cascade, and multi-task convolutional neural networks are employed to detect a face in images or video frames. Finally, convolutional neural networks extract useful information and classify emotions based on labeled datasets.

Based on analysis, CNN [13, 45, 55, 59, 64, 78] is a subfield of neural networks, comprised of two major blocks: feature extraction and classification. The CNN architecture includes layers such as the convolution layer, pooling layers, dropout, activation function, and fully connected layers, along with batch normalization and regularization, which are employed when CNN encounters an overfitting problem [2, 30, 65]. The convolution layer, the lowest layer, is used to extract various information from the input images using an (MxM) filter, and the result is known as a feature map. Feature maps provide information about the edges and corners of an image. The pooling layer's primary purpose is to reduce the dimensionality of the feature map and computation expense. The fully connected layer, responsible for connecting multiple layers, consists of bias, weight, and neurons. Typically, it is placed before the output layer of the CNN. The dropout layer is used to address the overfitting problem in the CNN architecture. The activation function initiates the connection between layers using neurons. Several commonly used activation functions include Softmax, ReLu, TanH, and Sigmoid. To build the CNN model, the dataset can be divided into three

Table 4 Emotion classification techniques in deep learning

Conventional FER Techniques in Deep Learning				
Ref	Techniques	Dataset	Emotions	Accuracy/result
[5]	DCNN (VGG-16)	KDEF, JAFFE	Afraid, Angry, Disgust, Happy, Neutral, Sad, Surprised	KDEF: 93.47%, JAFFE: 100%
[120]	CNN, RNN, ConvLSTM	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	CNN: 65% RNN:41%
[30]	CNN, OpenCV	FER-2013	Neutral, Happy, Sad, Angry, Surprised, Disgusted	93.95%
[118]	DCNN	Real-time dataset	Angry, Happy, Neutral, Sad, Surprise	78.04%
[55]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Training:97 Testing:57.4
[64]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	62%
[122]	CNN, Haar-Cascade Classifier	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Experimental done with different epochs
[13]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Accuracy 79.8
[66]	CNN	FERC-2013, JAFFE	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	FERC:70.14 JAFFE: 98.65
[88]	CNN, Viola-Jones, Haar feature	NIMH-ChEF, CAFÉ, AM-FED, and EmoReact	Neutral, Happy, Sad, Surprise, Fear, Disgust, and Anger	46.05%
[78]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	70%
[2]	DCNN, OpenCV	ADFES-BIV, WSEFEP	Happy, Sad, Anger, Surprise, Disgust, Fear, Neutral, Pride, Contempt, Embarrassment	95.12%
[149]	CNN, DNNs	Karolinska Directed Emotional (KDEF)	Afraid, Angry, Disgusted, Happy, Neutral, Sad, Surprised	86.73 5
[65]	DNNs	CK+, JAFFE	Sad, Happy, Surprised, Angry, Neutral, Disgust, Fear	JAFFE: 95.23 CK+: 93.24
[45]	CNN	CK+, KDEF	Anger, Disgust, Fear, Happy, Sa, Surprise, Neutral	Testing: 97.53 JAFFE:97.53
[148]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	57.1%
[76]	CNN	FERC-2013, CK+	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Around 90+ accuracy
State-art-of-the Techniques				
[162]	Feature separation model exchange-GAN	Multi-PIE, FACES, Oulu-CASIA	Sad, Happy, Surprised, Angry, Neutral, Disgust, Fear	Multi-PIE: 91.08 FACES: 95.24 Oulu-CASIA:86.33
[26]	Residual Variational Autoencoder	Affectnet	Neutral, Happy, Sad, Surprise, Fear, Disgust, and Anger	98.0%
[98]	CNN-LSTM	FER-2013, CK+	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	FER-2013:78.2 CK+: 99.7
[166]	GAN	Multi-PIE, MMI, RAF-DB	Sad, Happy, Surprised, Angry, Neutral, Disgust, Fear	Multi-PIE: 93.66 MMI: 76.44 RAF-DB: 89.01
[6]	Ensemble Classifier	JAFFE, TFEID, Moroccan, Caucasian	Sad, Happy, Surprised, Angry, Neutral, Fear	JAFFE: 86.67 TFEID: 83.19, Moroccan: 89.47, Caucasian: 86.36

Table 4 (continued)

Conventional FER Techniques in Deep Learning				
Ref	Techniques	Dataset	Emotions	Accuracy/result
[72]	Multi-scale convolutional and residual block based DCNN	FER2013, JAFFE, CK+, KDEF, RAFDB	Neutral, Sad, Happy, Surprised, Angry, Neutral, Fear, Disgust	FER2013: 80 JAFFE:99 CK+:98 KDEF:97 RAFDB:87
[145]	Frequency neural network (FreNet)	CK+ OULU KDEF	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	CK+: 98.91 OULU: 88.33 KDEF: 91.22
[104]	Mobile-Net	FER+ RAF-DB	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	FER+: 88.11 RAF-DB: 84.49
[132]	Bi-LSTM	SAVEE, RAVDESS, and RML	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral, Calm	SAVEE: 99.75 RAVDESS: 94.99 RML: 99.23
[102]	Convolutional 3D	CK+, MMI Oulu-CASIA	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	CK+: 93.45 MMI: 84.53 FERA: 93.45

sets: training, validation, and testing, along with batch size and epochs [76, 148].

Typically, training a neural network requires a massive amount of data to achieve significant accuracy. In deep learning, data augmentation techniques, such as flipping, rotating, scaling, cropping, translation, and adding Gaussian noise, are used to increase the size of the dataset when the DL model suffers from overfitting issues. TensorFlow, Keras, PyTorch, and various Python libraries are commonly employed to develop the CNN architecture. FER utilizes pre-trained deep convolutional neural network (DCNN) models through appropriate transfer learning, such as VGG-16 [5], ResNet, DenseNet, and Inception, which are trained with large datasets (e.g., ImageNet) containing different classes. CNN yields better results when combined with transfer learning as the base model [5]. Also, fine-tuning is a crucial step in transfer learning-based FER systems, and carefully chosen techniques are employed to fine-tune the proposed model for better results. Moreover, optimizers such as Adam, AdaGrad, and RMSProp, along with the learning rate, are used to select relevant features from the available ones in this optimization process. Furthermore, in recent years, state-of-the-art techniques such as Bi-LSTM [101], Autoencoder [26], and GAN [163] have been applied in FER to enhance accuracy and overcome challenges, such as the vanishing gradient problem in the conventional CNN approach [148]. Also, this approach is efficient for the extraction of spatiotemporal features from video sequences.

In this systematic review, the performance evaluation of Facial Expression Recognition (FER) systems is analyzed by comparing various FER datasets and their accuracy, as presented in Tables 3 and 4. The number of facial emotions recognized by different machine learning and deep learning techniques is captured and summarized in Fig. 12. On the x-axis, the number of expressions recognized by FER methods is represented, while the y-axis displays the names of FER methods within the machine and deep learning domains. Based on the analysis conducted in this survey, it is observed that the majority of papers recognize up to eight facial expressions as the maximum number of emotions in their FER systems. This comparison provides insights into the range and diversity of emotions recognized by different FER methods various machine learning and deep learning approaches.

Figure 13 shows the most frequently used datasets, as presented in both Tables 3 and 4. The FER-2013 [52], JAFFE [93], and CK+[90] facial expression datasets are widely applied in the FER system. Moreover, many authors combine more than one dataset for their experiments and test their FER model's accuracy, as analysed in Tables 3 and 4.

In summary, the DL model in FER is capable of learning hierarchical representation features from raw input images and capturing complex patterns of facial expression. Consequently, it reduces the need for manual feature engineering, demonstrating the ability to adopt diverse large datasets and generalize effectively on unseen data. Moreover, it exhibits significant performance in handling challenges such as illumination, pose variation, and extreme facial expressions. However, certain research gaps persist within the FER domain, particularly concerning posed facial expressions that may appear too artificial. These concerns have been addressed in FER datasets. Notably, when training a model using a posed dataset, higher accuracy is achieved. Yet, challenges arise during real-time testing, where the model may fail to accurately recognize facial expressions in dynamic environments. Additionally, issues such as poorly annotated

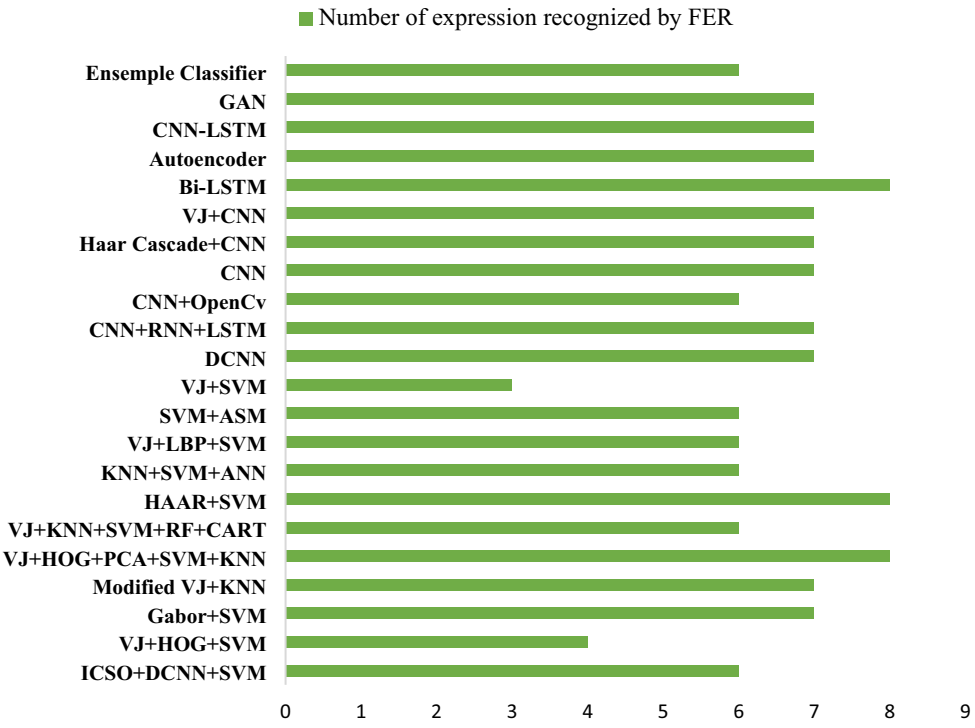


Fig. 12 Number of Facial Expressions commonly used in FER systems

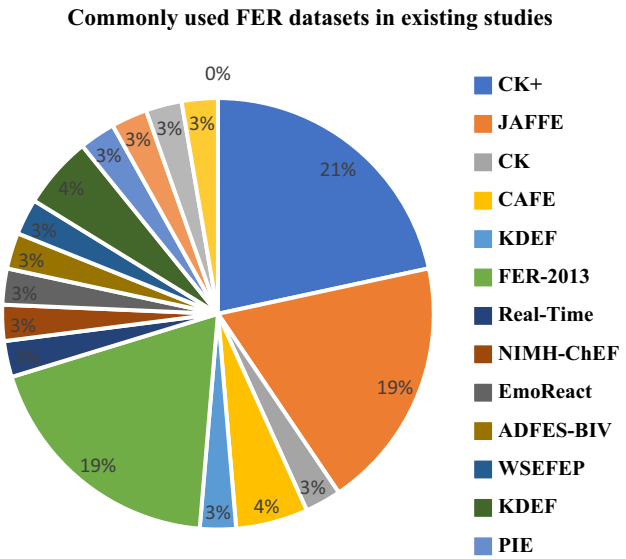


Fig. 13 Most commonly used Facial Expression Datasets

datasets with high ambiguity contribute to the complexity of FER research. Publicly available facial expression datasets are often limited in both quantity and basic expression variety, rendering them insufficient for training models to recognize real-world facial expressions. Furthermore, it is important to note that DL FER models are computationally intensive, necessitating powerful hardware like GPUs or TPUs and considerable time for training. Regularization techniques are also crucial to address pattern generalization and overfitting issues arising from imbalanced datasets.

Performance Evaluation Mechanisms of FER

Performance evaluation is a crucial aspect in the field of FER as it allows for the quantitative comparison of different datasets and the effectiveness of machine learning and deep learning models [154]. By conducting performance evaluations, researchers and practitioners can assess the accuracy, robustness, and overall performance of FER models in recognizing and classifying facial expressions. This evaluation process enables the identification of strengths and weaknesses in different datasets and models, facilitating improvements and advancements in FER technology. There are three different mechanisms commonly used in FER as follows.

Subject-Independent Analysis

Subject-independent analysis typically separates the dataset into three parts, namely, training, validation, and testing (e.g., 60%, 20%, 20%). This process is commonly known as Hold-out cross-validation. Generally, training has more data than testing, also validation data could be separated from training data. This method helps to provide fast computing analysis in FER models. However, it might produce high variance during evaluation, since it solely depends on

which data point ends up training, and testing, also, it heavily depends on the dataset split.

On the other hand, the K-fold cross-validation technique is used to overcome above mentioned issues as overfitting problems in FER and provide insight into model generalization capabilities in unknown and independent datasets. Furthermore, it can be divided into:

- **K-fold cross-validation** [124], which is used to randomly split the entire dataset into k groups. One group for testing and the remaining k-1 group for training. This process will repeat until each group is used for testing. Moreover, it has advantages when using this process in FER, since each group was used for training and tested once during model training. However, it increases computation time when the train model k-times.
- **Leave-p-out cross-validation [**24**]**, which is all possible of p set is used for training and validation. This method leads to robust evaluation of the FER model than k-fold validation, also it could be computationally infeasible depending on the p set. Furthermore, it is efficient for imbalanced facial expression datasets.

Cross-Database

Cross-database validation involves assessing the performance of the FER model on multiple datasets to ensure its generalization across different sources. Some studies [80, 114, 158] have experimented with combining more than one dataset to enhance FER performance, primarily to address the limitation of annotating sufficient training samples. Additionally, the Deep Emo-transfer Network (DETAN) [81] has been introduced to mitigate dataset bias, specifically addressing class imbalance challenges in FER model training. Similarly, adversarial graph representation adaptation (AGRA) [28] has been employed in experiments to evaluate the cross-dataset generalization ability of FER models. However, this approach has the potential for bias when combining FER datasets, such as those from different cultures and age groups, which may impact accurate emotion recognition. Furthermore, variation in annotation styles across datasets can lead to confusion and affect model learning. So, before combining the FER dataset, through analyse is required in each dataset annotation style, expression count, and cultural context.

Evaluation Metrics

The evaluation measure plays a vital role in the training process, as it helps in distinguishing and selecting the best classifier. Choosing the appropriate evaluation metrics is crucial for accurately assessing the performance of a classifier. Commonly used evaluation metrics in the field of FER include accuracy, precision, recall, F1-score, confusion matrix, and the Receiver Operating Characteristic (ROC) curve. The number of precise predictions divided by the total number of predictions is used to calculate accuracy.

$$Accuracy = (TP + FN)/(TP + TN + FP + FN) \tag{1}$$

where TP represents True Positive, FN represents False Negative, TN represents True Negative, and FP represents False Positive.

$$Precision = TP/(TP + FP) \tag{2}$$

The precision metric is useful for conveying more information compared to accuracy. It is calculated as the ratio of true positives (TP) to the sum of true positives and false positives (FP) for each class.

$$Recall = TP/TP + FN \tag{3}$$

The recall metric measures the ability to correctly identify all positive instances. It is calculated as the ratio of true positives (TP) to the sum of true positives and false negatives (FN) and ranges from 0 to 1, with 1 being the optimal value.

$$F1 - Score = (2 * (Precision * Recall))/((Precision + Recall)) \tag{4}$$

The F1-Score metric represents a weighted average of precision and recall, where a score of 0.0 indicates the worst performance and a score of 1.0 signifies the best performance. In addition to the commonly used evaluation metrics mentioned earlier, other techniques are essential for assessing the performance, efficiency, and scalability of classifiers in practical applications. These evaluation techniques go beyond the accuracy and predictive capabilities and consider factors such as execution time, training time, and resource occupancy.

Visualization Techniques

Furthermore, Explainable Artificial Intelligence (XAI) [73, 109] techniques are actively applied to understand and interpret the results of deeper insights into DL model outcomes. These methods aid in comprehending which facial features or regions significantly contribute to FER model decisions. Commonly used techniques for quantifying the significance of different features include Shapley Additive exPlanations (SHAP) [73] and local interpretable model-agnostic explanation (LIME) [73] Similarly, gradient-weighted class activation mapping (Grad-CAM) [131] is employed to highlight regions of facial features that the FER model focuses on when making predictions. Likewise, layer-wise relevance propagation (LRP) [14] decomposes the output of the neural network to understand the contribution of each input feature,

helping trace the feature influence on the final prediction. Moreover, the effectiveness of XAI approaches depends on the specific FER systems and the level of interpretability required for particular applications.

Unresolved FER Challenges and Future Directions

Numerous FER techniques have evolved across both theoretical and practical evaluation during the last few decades. As per the literature review, we further introduced some potential FER issues still not solved efficiently that depend on the deep learning model. The challenges are as follows:

- The FER system is still facing a lot of challenges such as illumination, pose variation, and occlusion [50]. Firstly, changes in lighting effects can lead to variations in the appearance of facial features. Also, poor lighting conditions, shadows, or overexposure can obscure facial landmarks, affecting the system's ability to recognize subtle expressions. Secondly, different head poses alter the spatial feature relationship between facial features, which means extreme head rotations or tilts may result in partial or distorted facial information, making it difficult for the system to recognize expressions accurately. Thirdly, occlusion can lead to the loss of critical facial information, especially around key regions like the eyes and mouth, which are essential for accurate expression recognition. Many researchers have proposed solutions to existing challenges in a real-world setting, but they are still working to improve the facial expression detection system's accuracy. Moreover, a combination of feature engineering, pre-processing techniques, and the application of advanced models can effectively generalise across a variety of facial situations. Furthermore, training FER models that function well in real-world situations requires a broad and representative dataset comprising samples with different illumination, postures, and occlusions.
- On the other hand, the major issue is the smaller amount of high-quality publicly available data for FER systems. The existing FER dataset is mostly captured in a controlled environment by professional and non-professional actors. Also, it consists of the basic expression pulse neutral. As a result, the currently developed FER systems struggle to significantly generalize unseen facial expressions in real-time environments. By incorporating spontaneous datasets with illumination and pose variation, FER systems can be better equipped to handle real-world situations. Moreover, in the domain of deep learning methods, FER requires a sufficient quantity and quality of training samples to build promising models for real-world scenarios. These challenges necessitate efficient data augmentation techniques to overcome the limitation of data insufficiency in FER and to help mitigate the class imbalance problem. Additionally, this study suggests combining different well-annotated FER datasets to enhance the generalizability of FER systems.
- Likewise, the sheer volume of data presents another challenge, with datasets growing into terabytes in size. This poses difficulties in terms of data storage, transmission, and processing for FER systems. Additionally, in real-world environments, there is a growing demand for data compression techniques to optimize the performance of FER systems. Moreover, data compression becomes more critical for FER systems when operating in real-world environments.

Furthermore, based on this systematic review analysis, recognizing emotional states solely from facial expressions, without revealing the exact emotional state of the person, is a challenging task [143]. To address this limitation, multimodal emotion recognition would be beneficial for accurately identifying human emotional states [130]. We also recommend the creation of a FER dataset encompassing different emotions, including both primary and secondary facial expressions. This initiative aims to encourage researchers to explore and develop efficient deep learning architectures capable of recognizing as many as possible facial expressions accurately. Moreover, it's essential to note that the accuracy of deep learning models varies across datasets. Facial expressions are not universally standardized and are culturally specific. Therefore, constructing FER databases and annotations must prioritize label correlation and discrepancies with careful consideration. In the future, young researchers focusing on FER should emphasize multimodal emotion recognition and take into account factors such as ethnicity, race, and cultural behavior [6].

Conclusion

In this paper, we present a detailed comparative study of both machine learning and deep learning techniques in FER systems. To conduct this analysis, we formulated a systematic review protocol and research questions based on the performance of FER systems across diverse datasets and the challenges they face in real-world environments. The conventional FER approach encompasses face detection, pre-processing methods, feature extraction techniques, and emotion classification using ML classifiers. However, these hand-crafted feature extraction techniques prove inefficient in handling complex facial features, and classifiers often struggle to predict accurate emotions in real-time scenarios. On the other hand, DL-based techniques in FER have

demonstrated significant improvements in the extraction of relevant features for facial expression recognition, eliminating the need for manual intervention in the feature extraction process. Consequently, DL methods show robustness in handling complex facial images, including challenges such as illumination, pose variation, and extreme facial expressions. However, it's worth noting that these methods consume more time in both training and testing, necessitating the use of GPU and TPU processors. Moreover, ensemble DL techniques, as well as the use of GANs, LSTM, and Autoencoders, have shown promising accuracy in current FER systems based on this analysis. Moreover, efficient data augmentation techniques, the combination of FER datasets, a hybrid well-designed DL model, and transfer learning with fine-tuning techniques could help to increase FER accuracy in real-time scenarios. However, micro-expression and addressing cultural diversity in different FER expression analyses remain challenging tasks due to the presence of invisible features and varying intensity levels of facial expressions among different groups, respectively.

Additionally, various types of datasets related to FER are examined in two ways: (1) posed and (2) spontaneous. The existing FER systems are mostly trained with posed facial expression datasets due to the limitations of spontaneous facial expressions. However, the available spontaneous dataset is highly complex, encompassing different levels of facial expression intensity, illumination, and pose variation. This analysis is conducted based on the FER approach and its accuracy on various datasets. As a result, there is still a need to design a highly efficient FER system with reduced computation time. This is particularly crucial with the integration of IoT for various applications, such as healthcare for continuous monitoring of patients' pain to customize the treatment plan, e-learning for assessing learners' emotions to suggest content based on their engagement levels, and FER via IoT can help with surveillance in crowded settings by tracking crowd behavior, spotting possible threats, or spotting emergencies based on people's facial expressions.

Acknowledgements The authors sincerely thank the ISO Certified (ISO/IEC 20000-1:2018) Centre for Machine Learning and Intelligence (CMLI), funded by the Department of Science and Technology (DST-CURIE), India, for providing the facility to carry out this research study.

Funding This research study received no external funding.

Data Availability No data availability.

Declarations

Conflict of Interest The author declared no potential conflict of interest concerning the publishing of this article.

Declaration of AI and AI-assisted Technologies in the Writing Process During the preparation of this work, the authors utilized Grammarly assistant tools included in Microsoft Word for grammar checking. After using these tools, the authors reviewed and edited the content as necessary and took full responsibility for the publication's content.

References

1. Abdul-Hadi MH, Waleed, J. Human speech and facial emotion recognition technique using svm. In: 2020 International Conference on Computer Science and Software Engineering (CSASE). 2020;191–196. IEEE. https://doi.org/10.1186/s42492-019-0034-5
2. Abdulsalam WH, Alhamdani RS, Abdullah MN. Facial emotion recognition from videos using deep convolutional neural networks. Int J Mach Learn Comput. 2019;9(1):14–9. https://doi.org/10.18178/ijmlc.2019.9.1.759.
3. Abiyev RH. Facial feature extraction techniques for face recognition. J Comput Sci. 2014;10(12):2360. https://doi.org/10.3844/jcssp.2014.2360.2365.
4. Adyapady RR, Annappa BA. comprehensive review of facial expression recognition techniques. Multimedia Syst. 2023;29:73–103. https://doi.org/10.1007/s00530-022-00984-w.
5. Akhand MAH, Roy S, Siddique N, Kamal MAS, Shimamura T. Facial emotion recognition using transfer learning in the deep CNN. Electronics. 2021;10(9):1036. https://doi.org/10.3390/electronics10091036.
6. Ali G, Ali A, Ali F, Draz U, Majeed F, Yasin S, Haider N. Artificial neural network-based ensemble approach for multicultural facial expressions analysis. IEEE Access. 2023;8:134950–63.
7. Anwar S, Milanova M. Real-time face expression recognition of children with autism. In: Proc. IAEMR. 2016.
8. Aouani H, Ayed YB. Speech emotion recognition with deep learning. Procedia Comput Sci. 2020;176:251–60. https://doi.org/10.1016/j.procs.2020.08.027.
9. Aro T, Abikoye O, Oladipo I, Awotunde B. Enhanced Gabor features based facial recognition using ant colony optimization algorithm. J Sustain Technol. 2019;10(1):1–28
10. Ashraf A, Gunawan T S, Rahman F D A, Kartiwi M. A Summarization of Image and Video Databases for Emotion Recognition. In: Recent Trends in Mechatronics Towards Industry 4.0, Springer. 2022; 669–680.
11. Aung H, Bobkov AV, Tun NL. Face detection in real-time live video using Yolo algorithm based on Vgg16 convolutional neural network. In: 2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), IEEE. 2021; 697–702.
12. Ayvaz U, Gürüler H, Devrim MO. Use of facial emotion recognition in e-learning systems. 2017.
13. Babajee P, Suddul G, Armoogum S, Foogooa R. Identifying human emotions from facial expressions with deep learning. In: 2020 Zooming Innovation in Consumer Technologies Conference (ZINC). IEEE. 2020; 36–9.
14. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10(7):e0130140.
15. Bagherian E, Rahmat RWO. Facial feature extraction for face recognition: a review. In: 2008 International Symposium on Information Technology, IEEE. 2008;2;1–9.
16. Bakshi U, Singhal R. A survey on face detection methods and feature extraction techniques of face recognition. Int J Emerg Trends Technol Comput Sci (IJETTCS). 2014;3(3):233–7.
17. Basak P, De S, Agarwal M, Malhotra A, Vatsa M, Singh R. Multimodal biometric recognition for toddlers and pre-school

children. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2017; 627–33.
18. Bishop C M, Nasrabadi NM. Pattern recognition and machine learning. New York: springer. 2006;4(4):738.
19. Boughida A, Kouahla MN, Lafifi Y. A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. Evol Syst. 2021. https://doi.org/10.1007/s12530-021-09393-2.
20. Bouhabba EM, Shafie AA, Akmeliawati R. Support vector machine for face emotion detection on a real-time basis. In: 2011 4th International Conference on Mechatronics (ICOM). IEEE. 2011; 1–6. IEEE.
21. Buhari AM, Ooi CP, Baskaran VM, Phan RC, Wong K, Tan WH. FACS-based graph features for real-time micro-expression recognition. J Imaging. 2020;6(12):130.
22. Cambria E, Das D, Bandyopadhyay S, Feraco A. Affective computing and sentiment analysis. In: A practical guide to sentiment analysis. Springer, Cham. 2017; 1–10.
23. Canedo D, Neves AJ. Facial expression recognition using computer vision: a systematic review. Appl Sci. 2019;9(21):4678.
24. Celisse A, Robin S. Nonparametric density estimation by exact leave-p-out cross-validation. Comput Stat Data Anal. 2008;52(5):2350–68.
25. Chang W Y, Hsu S H, Chien J H. FATAUVA-Net: an integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017; 17–25.
26. Chatterjee S, Das AK, Nayak J, Pelusi D. Improving facial emotion recognition using residual autoencoder coupled affinity-based overlapping reduction. Mathematics. 2022;10(3):406.
27. Chaudhari S T, Kale A. Face normalization: enhancing face recognition. In: 2010 3rd International Conference on Emerging Trends in Engineering and Technology. IEEE. 2010; 520–5.
28. Chen T, Pu T, Wu H, Xie Y, Liu L, Lin L. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. IEEE Trans Pattern Anal Mach Intell. 2021;44(12):9887–903.
29. Chen W, Huang H, Peng S, Zhou C, Zhang C. YOLO-face: a real-time face detector. Vis Comput. 2021;37(4):805–13.
30. Choi IK, Ahn HE, Yoo J. Facial expression classification using deep convolutional neural network. J Elect Eng Technol. 2018;13(1):485–92.
31. Cohn J F, Zlochower A J, Lien J J, Kanade T. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In: Proceedings third IEEE international conference on automatic face and gesture recognition. IEEE. 1998; 396–401.
32. Cubuk E D, Zoph B, Mane D, Vasudevan V, Le Q V. Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019; 113–23.
33. Cuimei L, Zhiliang Q, Nan J, Jianhua W. Human face detection algorithm via Haar cascade classifier combined with three additional classifiers. In: 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI). IEEE. 2017; 483–7.
34. Cunningham S, Ridley H, Weinel J, Picking R. Supervised machine learning for audio emotion recognition. Pers Ubiquit Comput. 2021;25(4):637–50.
35. Dalrymple KA, Gomez J, Duchaine B. The dartmouth database of children's faces: acquisition and validation of a new face stimulus set. PLoS One. 2013;8(11): e79131.
36. Dang V T, Do H Q, Vu V V, Yoon B. Facial expression recognition: a survey and its applications. In: 2021 23rd International Conference on Advanced Communication Technology (ICACT), IEEE. 2021; 359–67.
37. Deng J, Guo J, Zafeiriou S. Single-stage joint face detection and alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019.
38. Dino H I, Abdulrazzaq M B. Facial expression classification based on SVM, KNN and MLP classifiers. In: 2019 International Conference on Advanced Science and Engineering (ICOASE). IEEE. 2019; 70–5.
39. Du S, Tao Y, Martinez AM. Compound facial expression of emotion. Proc Natl Acad Sci. 2014;111(15):E1454–62.
40. Edwards G J, Cootes TF, Taylor CJ. Face recognition using active appearance models. In: European conference on computer vision. Springer, Berlin, Heidelberg. 1998; 581–95.
41. Egger HL, Pine DS, Nelson E, Leibenluft E, Ernst M, Towbin KE, Angold A. The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS): a new set of children's facial emotion stimuli. Int J Methods Psychiatr Res. 2011;20(3):145–56.
42. Ekman P, Friesen WV. Constants across cultures in the face and emotion. J Pers Soc Psychol. 1971;17(2):124.
43. Ekman P. An argument for basic emotions. Cogn Emot. 1992;6(3–4):169–200.
44. Ekman P. Darwin, deception, and facial expression. Ann NY Acad Sci. 2003;1000(1):205–21.
45. El Hammoumi O, Benmarrakchi F, Ouherrou N, El Kafi J, El Hore A. Emotion recognition in e-learning systems. In: 2018 6th international conference on multimedia computing and systems (ICMCS). IEEE. 2018; 1–6.
46. Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM. Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; 5562–70.
47. Fasel B, Luettin J. Automatic facial expression analysis: a survey. Pattern Recogn. 2003;36(1):259–75.
48. Friesen E, Ekman P. Facial action coding system: a technique for the measurement of facial movement. Palo Alto. 1978;3(2):5.
49. Gavrilescu M, Vizireanu N. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. Sensors. 2019;19(17):3693.
50. Gehrig T, Ekenel H K. Why is facial expression analysis in the wild challenging? In: Proceedings of 2013 on Emotion recognition in the wild challenge and workshop. 2013; 9–16.
51. Giuliani NR, Flournoy JC, Ivie EJ, Von Hippel A, Pfeifer JH. Presentation and validation of the DuckEES child and adolescent dynamic facial expressions stimulus set. Int J Methods Psychiatr Res. 2017;26(1): e1553.
52. Goodfellow I J, Erhan D, Carrier P L, Courville A, Mirza M, Hamner B, Bengio Y. Challenges in representation learning: a report on three machine learning contests. In: International conference on neural information processing. Springer, Berlin, Heidelberg. 2013;117–24.
53. Gross R, Matthews I, Cohn J, Kanade T, Baker S. Multi-pie. Image Vis Comput. 2010;28(5):807–13.
54. Gu L, Kanade T. A generative shape regularization model for robust face alignment. In: European conference on computer vision. Springer, Berlin, Heidelberg. 2008; 413–26 (2008).
55. Gunawan TS, Ashraf A, Riza BS, Haryanto EV, Rosnelly R, Kartiwi M, Janin Z. Development of video-based emotion recognition using deep learning with Google Colab. Telkomnika. 2020;18(5):2463–71.
56. Gupta S. Facial emotion recognition in real-time and static images. In: 2018 2nd international conference on inventive systems and control (ICISC). IEEE. 2018; 553–60.
57. Hjelmås E, Low BK. Face detection: a survey. Comput Vis Image Underst. 2001;83(3):236–74.
58. Ho D, Liang E, Chen X, Stoica I, Abbeel P. Population-based augmentation: Efficient learning of augmentation policy

schedules. In: International conference on machine learning. 2019; 2731–41. PMLR.
59. Hossain S, Umer S, Rout RK, Tanveer M. Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling. Appl Soft Comput. 2023;134: 109997.
60. Hossen AMA, Ogla RAA, Ali MM. Face detection by using OpenCV's Viola-Jones Algorithm based on coding eyes. Iraqi J Sci. 2017;58(2A):735–45.
61. Huang Y, Chen F, Lv S, Wang X. Facial expression recognition: a survey. Symmetry. 2019;11(10):1189.
62. Jack RE, Garrod OG, Yu H, Caldara R, Schyns PG. Facial expressions of emotion are not culturally universal. Proc Natl Acad Sci. 2012;109(19):7241–4.
63. Jaderberg M, Dalibard V, Osindero S, Czarnecki W M, Donahue J, Razavi A, Kavukcuoglu K. Population based training of neural networks. arXiv preprint arXiv:1711.09846. 2017.
64. Jadhav R, Bhuke J, Patil N. Facial emotion detection using convolutional neural network. Int Res J Eng Technol. e-ISSN, 2395–0056. 2019.
65. Jain DK, Shamsolmoali P, Sehdev P. Extended deep neural network for facial emotion recognition. Pattern Recogn Lett. 2019;120:69–74.
66. Jaiswal A, Raju A K, Deb S. Facial emotion detection using deep learning. In: 2020 International Conference for Emerging Technology (INCET). IEEE. 2020; 1–5.
67. Jiang F, Zhang J, Yan L, Xia Y, Shan S. A three-category face detector with contextual information on finding tiny faces. In: 2018 25th IEEE international conference on image processing (ICIP). IEEE. 2018; 2680–4
68. Johnston B, de Chazal P. A review of image-based automatic facial landmark identification techniques. EURASIP J Image Video Process. 2018;1:1–23.
69. Kalinovskii I, Spitsyn, V. Compact convolutional neural network cascade for face detection. arXiv preprint arXiv:1508.01292. 2015.
70. Kamarol SKA, Jaward MH, Kälviäinen H, Parkkinen J, Parthiban R. Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. Pattern Recogn Lett. 2017;92:25–32.
71. Kanade T, Cohn J F, Tian Y. Comprehensive database for facial expression analysis. In: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580). IEEE. 2000; 46–53.
72. Karnati M, Seal A, Yazidi A, Krejcar O. FLEPNet: feature level ensemble parallel network for facial expression recognition. IEEE Trans Affect Comput. 2022;13(4):2058–70.
73. Kawakura S, Hirafuji M, Ninomiya S, Shibasaki R. Analyses of diverse agricultural worker data with explainable artificial intelligence: Xai based on shap, lime, and lightgbm. Eur J Agric Food Sci. 2022;4(6):11–9.
74. Khan RA, Crenn A, Meyer A, Bouakaz S. A novel database of children's spontaneous facial expressions (LIRIS-CSE). Image Vis Comput. 2019;83:61–9.
75. Kitchenham B. Procedures for performing systematic reviews, vol. 33. Keele: Keele University; 2004. p. 1–26.
76. Kumar GR, Kumar RK, Sanyal G. Facial emotion analysis using deep convolution neural network. In: 2017 International Conference on Signal Processing and Communication (ICSPC). IEEE. 2017; 369–74.
77. Langner O, Dotsch R, Bijlstra G, Wigboldus DH, Hawk ST, Van Knippenberg AD. Presentation and validation of the Radboud Faces Database. Cogn Emot. 2010;24(8):1377–88.
78. Lasri I, Solh A R, El Belkacemi M. Facial emotion recognition of students using a convolutional neural network. In: 2019 third international conference on intelligent computing in data sciences (ICDS), IEEE. 2019;1–6.
79. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
80. Li S, Deng W. Deep emotion transfer network for cross-database facial expression recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE. 2018;3092–9.
81. Li S, Deng W. Deep facial expression recognition: a survey. IEEE Trans Affect Comput. 2020. 13(3);1195–1215.
82. Li X, Lai S, Qian X. DBCFace: towards pure convolutional neural network face detection. IEEE Trans Circ Syst Video Technol. 2021. https://doi.org/10.1109/TCSVT.2021.3082635.
83. Lim R, MJT Reinders T. Facial landmark detection using a Gabor filter representation and a genetic search algorithm. In: Proceeding, (SITIA'2000), Graha Institut Teknologi Sepuluh November. 2000.
84. Liu C, Hirota K, Dai Y. Patch attention convolutional vision transformer for facial expression recognition with occlusion. Inf Sci. 2023;619:781–94.
85. Liu S, Tian Y, Peng C, Li J. Facial expression recognition approaches based on least squares support vector machines with improved particle swarm optimization algorithms. In: 2010 IEEE International Conference on Robotics and Biomimetics. IEEE. 2010; 399–404 (2010).
86. Liu Y, Wang W, Feng C, Zhang H, Chen Z, Zhan Y. Expression snippet transformer for robust video-based facial expression recognition. Pattern Recogn. 2023;138: 109368.
87. LoBue V, Thrasher C. The Child Affective Facial Expression (CAFE) set: validity and reliability from untrained adults. Front Psychol. 2015;5:532.
88. Lopez-Rincon, A. Emotion recognition using facial expressions in children using the NAO Robot. In: 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), IEEE. 2019;146–53.
89. Lu H, Yang F. Active shape model and its application to face alignment. In: Subspace methods for pattern recognition in intelligent environment, Springer. 2014; 1–31.
90. Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE. 2010; 94–10.
91. Lundqvist D, Flykt A, Öhman A. The Karolinska directed emotional faces—KDEF [CD ROM]. Karolinska Institutet, Stockholm. 1998.
92. Luo D, Wen G, Li D, Hu Y, Huan E. Deep-learning-based face detection using iterative bounding-box regression. Multimedia Tools Appl. 2018;77:24663–80.
93. Michael L, Miyuki K, Jiro G. The Japanese Female Facial Expression (JAFFE) Dataset .1998. Zenodo.
94. Martinez B, Valstar MF. Advances, challenges, and opportunities in automatic facial expression recognition. Adv Face Detect Facial Image Anal. 2016; 63–100.
95. Mascaró-Oliver M, Mas-Sansó R, Amengual-Alcover E, Roig-Maimó MF. UIBVFED-mask: a dataset for comparing facial expressions with and without face masks. Data. 2023;8(1):17. https://doi.org/10.3390/data8010017.
96. Matsumoto D. More evidence for the universality of a contempt expression. Motiv Emot. 1992;16:363–8.
97. Mehrabian, A. Nonverbal communication. In: Nebraska symposium on motivation. University of Nebraska Press. 1971.
98. Ming Y, Qian H, Guangyuan L. CNN-LSTM facial expression recognition method fused with two-layer attention mechanism. Comput Intell Neurosci. 2022.
99. Miolla A, Cardaioli M, Scarpazza C. Padova Emotional Dataset of Facial Expressions (PEDFE): a unique

dataset of genuine and posed emotional facial expressions. Behav Res. 2023;55:2559–74. https://doi.org/10.3758/s13428-022-01914-4.
100. Mohammed OA, Al-Tuwaijari JM. Analysis of challenges and methods for face detection systems: a survey. Int J Nonlinear Anal Appl. 2022;13(1):3997–4015.
101. Mohana M, Subashini P, Krishnaveni M. Emotion recognition from facial expression using hybrid CNN–LSTM network. Int J Pattern Recognit Artif Intell. 2023;37(08):2356008.
102. Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter conference on applications of computer vision (WACV), IEEE. 2016;1–10.
103. Mollahosseini A, Hasani B, Mahoor MH. Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput. 2017;10(1):18–31.
104. Nan Y, Ju J, Hua Q, Zhang H, Wang B. A-MobileNet: an approach of facial expression recognition. Alex Eng J. 2022;61(6):4435–44.
105. Negrão JG, Osorio AAC, Siciliano RF, Lederman VRG, Kozasa EH, D'Antino MEF, Schwartzman JS. The child emotion facial expression set: a database for emotion recognition in children. Front Psychol. 2021;12:1352.
106. Nojavanasghari B, Baltrušaitis T, Hughes CE, Morency LP. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In: Proceedings of the 18th acm international conference on multimodal interaction. 2016; 137–44.
107. Oliver MM, Amengual AE. UIBVFED: virtual facial expression dataset. PLoS One. 2020;15(4): e0231266.
108. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Moher D. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. J Clin Epidemiol. 2021;134:103–12.
109. Palacio S, Lucieri A, Munir M, Ahmed S, Hees J, Dengel A. Xai handbook: towards a unified framework for explainable AI. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; 3766–75.
110. Pali V, Goswami S, Bhaiya L P. An extensive survey on feature extraction techniques for facial image processing. In: 2014 International Conference on Computational Intelligence and Communication Networks, IEEE. 2014; 142–148.
111. Pantic M, Valstar M, Rademaker R, Maat L. Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and Expo, IEEE. 2005; 5.
112. Park S, Lee K, Lim JA, Ko H, Kim T, Lee JI, Lee EC. Differences in facial expressions between spontaneous and posed smiles: automated method by action units and three-dimensional facial landmarks. Sensors. 2020;20(4):1199.
113. Patil HY, Bharambe SV, Kothari AG, Bhurchandi KM. Face localization and its implementation on embedded platform. In: 2013 3rd IEEE International Advance Computing Conference (IACC). IEEE. 2013; 741–5
114. Peng P, Xiang T, Wang Y, Pontil M, Gong S, Huang T, Tian Y. Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016;1306–15.
115. Perveen N, Ahmad N, Khan M A Q B, Khalid R, Qadri S. Facial expression recognition through machine learning. Int J Sci Technol Res. 2016;5(03)91–97
116. Picard RW. Affective computing. MIT press; 2000.
117. Polikovsky S, Kameda Y, Ohta Y. Facial micro-expression detection in hi-speed video based on facial action coding system (FACS). IEICE Trans Inf Syst. 2013;96(1):81–92.
118. Pranav E, Kamal S, Chandran C S, Supriya M H. Facial emotion recognition using deep convolutional neural network. In: 2020 6th International conference on advanced computing and communication Systems (ICACCS), IEEE. 2020; 317–20.
119. Priadana A, Habibi M. Face detection using haar cascades to filter selfie face images on instagram. In: 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), IEEE. 2019; 6–9.
120. Qayyum R, Akre V, Hafeez T, Khattak H A, Nawaz A, Ahmed S, ur Rahman, K. Android based Emotion Detection Using Convolutions Neural Networks. In: 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE. 2021; 360–5.
121. Rathee N, Vaish A, Gupta S. Adaptive system to learn and recognize the emotional state of mind. In: 2016 International Conference on Computing, Communication and Automation (ICCCA). IEEE. 2016; 32–6.
122. Riyantoko PA, Hindrayani KM. Facial emotion detection using haar-cascade classifier and convolutional neural networks. J Phys Conf Ser. 2021;1844(1): 012004 (**IOP Publishing**).
123. Robin M H, Rahman M M U, Taief A M, Eity Q N. Improvement of face and eye detection performance by using multi-task cascaded convolutional networks. In: 2020 IEEE Region 10 Symposium (TENSYMP), IEEE. 2020; 977–980
124. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell. 2009;32(3):569–75.
125. Rodriguez Y, Cardinaux F, Bengio S, Mariéthoz J. Measuring the performance of face localization systems. Image Vis Comput. 2006;24(8):882–93.
126. Romani-Sponchiado A, Sanvicente-Vieira B, Mottin C, Hertzog-Fonini D, Arteche A. Child Emotions Picture Set (CEPS): development of a database of children's emotional expressions. Psychol Neurosci. 2015;8(4):467.
127. Roshan K, Zafar A, Haque SBU. Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system. Comput Commun. 2023. https://doi.org/10.1016/j.comcom.2023.09.030.
128. Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39(6):1161.
129. Sajjad M, Ullah FUM, Ullah M, Christodoulou G, Cheikh FA, Hijji M, Rodrigues JJ. A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. Alex Eng J. 2023;68:817–40.
130. Sebe N, Cohen I, Huang TS. Multimodal emotion recognition. In: Handbook of pattern recognition and computer vision. 2005;387–409.
131. Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017; 618–626.
132. Sharafi M, Yazdchi M, Rasti R, Nasimi F. A novel Spatio-temporal convolutional neural framework for multimodal emotion recognition. Biomed Signal Process Control. 2022;78: 103970.
133. Sharma R, Patterh MS. Face recognition using face alignment and PCA techniques: a literature survey. IOSR J Comput Eng (IOSR-JCE). 2015;17(4):17–30.
134. Sheikh BUH, Zafar A. White-box inference attack: compromising the security of deep learning-based COVID-19 diagnosis systems. Int J Inf Tecnol. 2023. https://doi.org/10.1007/s41870-023-01538-7.
135. Sheikh BUH, Zafar A. Unlocking adversarial transferability: a security threat towards deep learning-based surveillance systems via black box inference attack—a case study on face mask surveillance. Multimed Tools Appl. 2023. https://doi.org/10.1007/s11042-023-16439-x.
136. Sheikh BUH, Zafar A. Untargeted white-box adversarial attack to break into deep learning based COVID-19 monitoring face

mask detection system. Multimed Tools Appl. 2023. https://doi.org/10.1007/s11042-023-15405-x.
137. Sheikh B, Zafar A. Beyond accuracy and precision: a robust deep learning framework to enhance the resilience of face mask detection models against adversarial attacks. Evol Syst. 2023. https://doi.org/10.1007/s12530-023-09522-z.
138. Sheikh B, Zafar A. RRFMDS: rapid real-time face mask detection system for effective COVID-19 monitoring. SN Comput Sci. 2023;4:288. https://doi.org/10.1007/s42979-023-01738-9.
139. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):1–48.
140. Sikkandar H, Thiyagarajan R. Deep learning based facial expression recognition using improved Cat Swarm Optimization. J Ambient Intell Humaniz Comput. 2021;12(2):3037–53.
141. Stockman G, Shapiro L G. *Computer vision*. Prentice Hall PTR.
142. Su Y, Liu Z, Ban X. Symmetric face normalization. Symmetry. 2019;11(1):96.
143. Suhaimi NS, Mountstephens J, Teo J. EEG-based emotion recognition: a state-of-the-art review of current trends and opportunities. Comput Intell Neurosci. 2020. https://doi.org/10.1155/2020/8875426.
144. Talele KT, Kadam S. Face detection and geometric face normalization. In: TENCON 2009–2009 IEEE Region 10 Conference. IEEE. 2009; 1–6.
145. Tang Y, Zhang X, Hu X, Wang S, Wang H. Facial expression recognition using frequency neural network. IEEE Trans Image Process. 2020;30:444–57.
146. Tao S, Li Y, Huang Y, Lan X. Face detection algorithm based on deep residual network. J Phys: Conf Ser. 2021;1802(3): 032142 (**IOP Publishing**).
147. Tian YI, Kanade T, Cohn JF. Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell. 2001;23(2):97–115.
148. Tümen V, Söylemez ÖF, Ergen B. Facial emotion recognition on a dataset using convolutional neural network. In: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE. 2017; 1–5
149. Verma A, Singh P, Alex J S R. Modified convolutional neural network architecture analysis for facial emotion recognition. In: 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE. 2019; 169–173.
150. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. IEEE. 2001; 1; I-I.
151. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: a brief review. Comput Intell Neurosci. 2018. https://doi.org/10.1155/2018/7068349.
152. Wang Y, Ji X, Zhou Z, Wang H, Li Z. Detecting faces using region-based fully convolutional networks. arXiv preprint arXiv:1709.05256. 2017.
153. Wang Y, Sun Y, Huang Y, Liu Z, Gao S, Zhang W, Zhang W. Ferv39k: a large-scale multi-scene dataset for facial expression recognition in videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022;20922–31.
154. Wardhani N W S, Rochayani M Y, Iriany A, Sulistyono A D, Lestantyo P. Cross-validation metrics for evaluating classification performance on imbalanced data. In: 2019 international conference on computer, control, informatics and its applications (IC3INA), IEEE. 2019; 14–18.
155. Wen Z, Lin W, Wang T, Xu G. Distract your attention: multi-head cross attention network for facial expression recognition. Biomimetics. 2023;8(2):199.
156. Wu Y, Ji Q. Facial landmark detection: a literature survey. Int J Comput Vis. 2019;127(2):115–42.
157. Wu Y, Zhang L, Gu Z, Lu H, Wan S. Edge-AI-driven framework with efficient mobile network design for facial expression recognition. ACM Trans Embed Comput Syst. 2023;22(3):1–17.
158. Xie Y, Chen T, Pu T, Wu H, Lin L. Adversarial graph representation adaptation for cross-domain facial expression recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. 2020;1255–64.
159. Yadav KS, Singha J. Facial expression recognition using modified Viola-John's algorithm and KNN classifier. Multimed Tools Appl. 2020;79(19):13089–107.
160. Yan H, Liu Y, Wang X, Li M, Li H. A face detection method based on skin color features and AdaBoost algorithm. J Phys: Conf Ser. 2021;1748(4): 042015 (**IOP Publishing**).
161. Yan Z, Yuan C. Ant colony optimization for feature selection in face recognition. In International conference on biometric authentication. Springer, Berlin, Heidelberg. 2004; 221–6.
162. Yang L, Tian Y, Song Y, Yang N, Ma K, Xie L. A novel feature separation model exchange-GAN for facial expression recognition. Knowl-Based Syst. 2020;204: 106217.
163. Yang W, Jiachun Z. Real-time face detection based on YOLO. In: 2018 1st IEEE international conference on knowledge innovation and invention (ICKII), IEEE. 2018; 221–4.
164. Zhang L, Ai H, Xin S, Huang C, Tsukiji S, Lao S. Robust face alignment based on local texture classifiers. In: IEEE International Conference on Image Processing, IEEE. 2005; 2, II-354.
165. Zhang N, Luo J, Gao W. Research on face detection technology based on MTCNN. In: 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), IEEE. 2020; 154–8.
166. Zhang X, Zhang F, Xu C. Joint expression synthesis and representation learning for facial expression recognition. IEEE Trans Circuits Syst Video Technol. 2021;32(3):1681–95.
167. Zhang Z, Luo P, Loy C C, Tang X. Facial landmark detection by deep multi-task learning. In: European conference on computer vision, Springer, Cham. 2014; 94–108. https://doi.org/10.1007/978-3-319-10599-4_7.
168. Zhao G, Huang X, Taini M, Li SZ, PietikäInen M. Facial expression recognition from near-infrared videos. Image Vis Comput. 2011;29(9):607–19.
169. Zhu X, Liu Y, Li J, Wan T, Qin Z. Emotion classification with data augmentation using generative adversarial networks. In: Advances in knowledge discovery and data mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Proceedings, Part III 22. Springer International Publishing. 2018;49–360.https://doi.org/10.1007/978-3-319-93040-4_28

Journal of Intelligent & Fuzzy Systems 46 (2024) 10727–10741
DOI:10.3233/JIFS-238947
IOS Press

Revisiting face detection: Supercharging Viola-Jones with particle swarm optimization for enhanced performance

M. Mohana[a,*], P. Subashini[a] and Diksha Shukla[b]
[a]*Centre for Machine Learning and Intelligence, Department of Computer Science, Avinashilingam Institute, Coimbatore, Tamil Nadu, India*
[b]*Department of Electrical Engineering and Computer Science, University of Wyoming, Laramie, USA*

Abstract. In recent years, face detection has emerged as a prominent research field within Computer Vision (CV) and Deep Learning. Detecting faces in images and video sequences remains a challenging task due to various factors such as pose variation, varying illumination, occlusion, and scale differences. Despite the development of numerous face detection algorithms in deep learning, the Viola-Jones algorithm, with its simple yet effective approach, continues to be widely used in real-time camera applications. The conventional Viola-Jones algorithm employs AdaBoost for classifying faces in images and videos. The challenge lies in working with cluttered real-time facial images. AdaBoost needs to search through all possible thresholds for all samples to find the minimum training error when receiving features from Haar-like detectors. Therefore, this exhaustive search consumes significant time to discover the best threshold values and optimize feature selection to build an efficient classifier for face detection. In this paper, we propose enhancing the conventional Viola-Jones algorithm by incorporating Particle Swarm Optimization (PSO) to improve its predictive accuracy, particularly in complex face images. We leverage PSO in two key areas within the Viola-Jones framework. Firstly, PSO is employed to dynamically select optimal threshold values for feature selection, thereby improving computational efficiency. Secondly, we adapt the feature selection process using AdaBoost within the Viola-Jones algorithm, integrating PSO to identify the most discriminative features for constructing a robust classifier. Our approach significantly reduces the feature selection process time and search complexity compared to the traditional algorithm, particularly in challenging environments. We evaluated our proposed method on a comprehensive face detection benchmark dataset, achieving impressive results, including an average true positive rate of 98.73% and a 2.1% higher average prediction accuracy when compared against both the conventional Viola-Jones approach and contemporary state-of-the-art methods.

Keywords: AdaBoost, Computer Vision (CV), face detection algorithm, particle swarm optimization, Viola-Jones

1. Introduction

Face detection and tracking play pivotal roles in a multitude of computer vision applications, encompassing human-computer interaction (HCI), human-robot interaction (HRI), computer surveillance systems, biometrics, facial recognition, facial expression recognition (FER), and various authentication solutions [22]. Yet, it remains an intricate challenge within the realms of computer vision, image processing, and pattern recognition. Face detection involves the identification of faces in digital images or videos, encompassing tasks such as determining their precise locations, recognizing facial landmarks, and even discerning emotional expres-

*Corresponding author. M. Mohana, Department of Computer Science, Avinashilingam Institute, India. E-mail: mohana_cs@avinuty.ac.in.

ISSN 1064-1246/$35.00

sions [15]. Furthermore, achieving high detection accuracy in complex backgrounds holds paramount importance in real-time scenarios, where factors such as pose variations, varying illumination, occlusions, and scale variations pose significant hurdles [28].

In recent years, researchers have introduced various techniques to address the challenges associated with face detection. These techniques leverage different types of prior knowledge about faces and can be broadly categorized into four distinct approaches: *(i) Knowledge-Based Approach:* This approach, as outlined in [16], relies on predefined rules based on human understanding of facial geometry. These rules dictate the relative distances and positions of facial features. By applying these rules, faces are detected and recognized. A subsequent verification process is often employed to eliminate incorrect detections. *(ii) Template Matching Method:* The template matching method, as described in [20], involves using a predefined face template or a parameterized face model to identify faces within input images. This technique entails analyzing the pixels within an image window using a predefined pattern to determine the presence of a human face. After initial detection, a verification step is typically applied to refine the results. The edge detection method is employed to detect specific facial features such as eyes, nose, and mouth within a face model. This method is utilized both for face detection and facial feature localization. *(iii) Feature-Invariant Approach:* The feature-invariant approach, as discussed in [9], focuses on extracting structural features of the face. Initially, these features are utilized for classifier algorithms that distinguish between faces and non-faces in images or videos. Such features may include skin tone, facial contours, and specific facial elements like eyes, nose, and mouth. *(iv) Appearance-Based Approach:* In the appearance-based approach, detailed in [29, 32], a collection of representative training face images is used to create a face model. This model encapsulates pixel intensities, effectively representing the human face. Machine learning techniques are often employed to identify relevant facial image characteristics.

The Viola-Jones algorithm yields significant results in real-time scenarios. It was introduced by Paul Viola and Michael Jones in 2001 [21]. This algorithm is a general-purpose tool for object detection when trained with datasets of other objects. It comprises four key components: Haar-like features with thresholds, integral images, AdaBoost, and a cascade classifier. Haar features are used to extract a vast number of features for identifying faces in images. These features are designed as black-and-white rectangular regions where the difference in pixel intensities is calculated. If the feature values fall below a threshold, the detection window is classified as positive (indicating a face); otherwise, it's classified as negative (non-face). Integral images expedite the Haar-like Feature extraction process. AdaBoost, a machine-learning algorithm, selects Haar-like features and combines them to build a strong classifier by iteratively selecting the weak features [7]. However, constructing a classifier with a low error rate often requires a significant number of rounds for identifying optimal features. When using a decision stump as a weak classifier, AdaBoost may require more time to identify optimal features. This often leads to a higher false-positive rate, particularly in dynamic environments [31]. Furthermore, AdaBoost must search among over 180,000 possible features, involving a staggering 2.16×10^{13} feature evaluation combinations. The cascade classifier efficiently dismisses non-faced regions in images or video frames. Besides, the algorithm is sensitive to face rotation, potentially leading to missed detections if faces are not upright. Scale variations are another challenge, impacting accuracy for extremely small or large faces. The algorithm's computational complexity during training time is one of its shortcomings, as it necessitates a large number of features due to the exhaustive search mechanism used in the AdaBoost algorithm. Moreover, faces that are partially obscured or hidden by occlusions might not be accurately detected. These challenges stem from the selection of numerous features in AdaBoost and the time required to determine the optimal threshold values for identifying strong features while detecting faces in the search window. However, several studies [5–7, 9] have suggested exploring more homogeneous feature types to enhance detector performance. Nevertheless, expanding the number of features inevitably leads to a larger feature set and increased storage memory requirements. As the feature space grows significantly, it becomes evident that the exhaustive search mechanism employed in the standard AdaBoost algorithm is inadequate for efficiently managing the search process. Consequently, this prolongs the training time, which constitutes one of the primary factors discouraging many approaches from exploring alternative feature types.

On the other hand, the advancement of deep learning approaches, such as YOLO [40, 41], SSD [44], Fast-RCNN [47], and CNN-based face detection [43], offers significant performance in a real-time

environment. These algorithms are general object detection methods designed for real-time processing speed but may not be as specialized for face detection. They could encounter challenges, especially in scenarios with crowded faces, impacting optimal face detection performance. Similarly, Faster R-CNN achieves high accuracy in complex scenarios in images and video sequences but demands a substantial amount of training data and computational resources. This makes it less suitable for simple applications. Notwithstanding, these algorithms require a large amount of data for training and significant computational resources [46] for implementation on small devices such as mobile cameras.

In this paper, we proposed a Particle Swarm Optimization (PSO) algorithm that integrates into the AdaBoost framework and replaces the exhaustive search used in the original AdaBoost for efficient feature selection and finding the optimal threshold values in the decision stump. PSO is used in a wide range of feature screening optimization and computer vision tasks and has given promising results so far [8, 11]. The proposed approach aims to expedite the training process time, minimize the training error, and develop a robust classifier for face detection by selecting discriminative features. The essential contributions of this paper are as follows:

- We have optimized the extraneous feature selection process in AdaBoost with the PSO algorithm,
- Threshold values of the AdaBoost selection process are optimized using PSO,
- The proposed method has reduced the computational time during the feature selection process, and,
- The performance of the proposed method has been compared with the conventional method using related metrics of the face detection algorithm.

This paper is organized into five sections. Section 2 presents related works regarding face detection using evolutionary and heuristic approaches and challenges. Section 3 summarizes background information on conventional methods. Section 4 discusses the proposed Viola-Jones algorithm utilizing PSO. Section 5 presents the experimental results and comparison of other state-of-the-art algorithms. Finally, Section 6 concludes the overall works, findings, and results summary.

2. Related work

2.1. Selection criteria

This section presents existing works in the face detection approach, focusing on efforts to reduce computational time and optimize the feature selection using evolutionary and heuristic approaches like PSO. To gather related works for this analysis, we conducted searches across various databases, including Scopus, Web of Science, IEEE, and Science Direct. We used specific keywords such as "Face Detection", "PSO", "Viola-Jones algorithm", and "Object Detection and PSO" to refine the search. While many keywords were available, we limited our search to find papers related to Viola-Jones and evolutionary algorithms. Fig. 1 illustrates the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols) flow diagram, depicting the paper selection process and the exclusion criteria applied to carry out this research work. Additionally, this paper [2, 27] provides insights into existing face detection methods and highlights current challenges that occur still in the real-time face detection process.

2.2. Face detection using machine learning

Perez and Vallejos [7] proposed using PSO to optimize template-based face detection on frontal faces. This approach yielded significant results, relying on face size and line integral values. Lu and Ming [35] introduced a composite feature-based face detection algorithm to enhance the detection rate of rigid objects on faces. They conducted an experiment using the FDDB benchmark dataset and achieved considerable results compared to conventional approaches. Huang et al. [34] improved the Viola-Jones algorithm by upgrading it with HoloLens and enhancing face detection using Haar-like rectangle features. This approach resulted in a 12% increase in average detection accuracy compared to existing face detection methods. Mohemmed et al. [5] proposed optimizing the AdaBoost feature selection process using the PSO evolutionary algorithm. This method selects the best features and optimizes the threshold values within the search space. Experiments were conducted with a "Wisconsin Breast Cancer" image dataset, achieving an average classification rate of 0.97% and a false-negative rate of 0.02%. Similarly, Zakaria and Suandi [38] combined a neural network with AdaBoost methods for face detection, improving detection per-

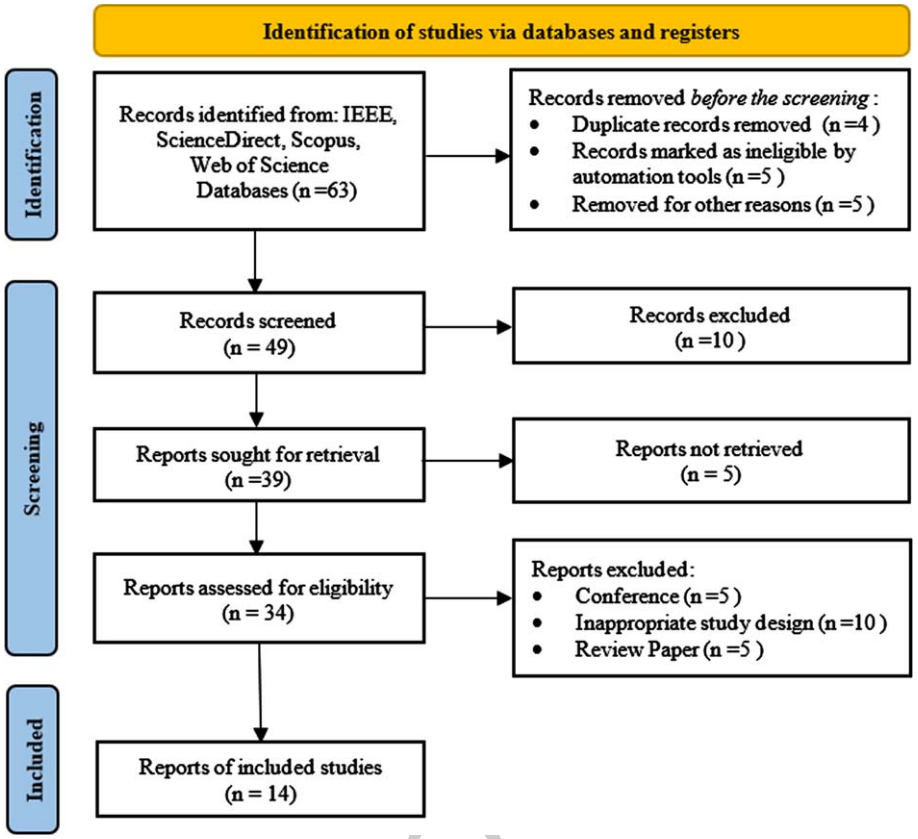


Fig. 1. Prisma flow diagram for paper selection for this study.

formance and creating a robust AdaBoost classifier. However, the method was too complex for rapid face detection. An AdaBoost neural network, developed by Zakaria et al. [39], is hierarchical, with the skin module roughly identifying faces, the AdaBoost filtering non-face regions, and the neural network serving as the primary face recognition tool. Li et al. [19] created a Gaussian model for skin color distribution, identified regions with different shades, and detected skin color areas using a cascade classifier. Additionally, the work of Lee et al. [33] attempted to incorporate a weight adjustment factor into a normalized support vector machine (SVM) as the base learner for AdaBoost.

In addition, Zhang and Ye [14] modified AdaBoost by incorporating two features: used PSO to determine the threshold values corresponding to the optimal solution of these two features, and they formed a strong classifier by combining weak classifiers based on these dual features. Zhang and Fan [12] employed Q-statistic correlation determination in training weak classifiers to reduce commonality among them and eliminate similar rectangular functionalities. In a study conducted by Yang et al. [36], they utilized a neural network and AdaBoost to develop an efficient pedestrian detection algorithm. Krishnan et al. [46] designed face detection for thermal and visible image registration using a saliency map strategy integrated with PSO techniques. The author achieved an average improvement of 16.93% similarity index score and 7.02% image quality index score. Subsequently, Besnassi et al. [47] introduced a dispersed Haar filter and optimized it with PSO, differential evolution, and genetic algorithm. The author achieved significant results when using Haar-differential evolution on frontal face detection on various state-of-the-art face detection datasets. Babu et al. [50] presented a facial expression recognition system based on a Deep belief network and PSO used for feature extraction with PCA. Taherkhania et al. [4] developed a CNN based on AdaBoost to reduce the computational processing time required for component prediction over large training datasets. These approaches either reduce training time or enhance detection rates, but

Fig. 2. Sample Face Detection challenge images [30].

none of them completely address the shortcomings of the Viola-Jones-based face detection algorithm. Similarly, Deep learning-based approaches like the paper [45] have presented a modified version of U-NET for significant face detection and recognition accuracy of face images captured by AI cameras with filters. This study attained reasonable detection and recognition accuracy on AI face images. Ranjan et al. [42] introduced a deep pyramid single-shot face detector method for face verification and identification. However, this study does not focus on face-detection challenging images.

2.3. Face detection challenges

Face detection is the first step of various face-related applications such as face recognition, facial emotions recognition, face tracking, and face analysis. The process of face detection is to identify the location of the face in images or video frames. There are two different types of factors that affect the effectiveness of the face detection algorithm. One is intrinsic factors that affect face appearances through facial hair, sunglasses, age, and cosmetics, and another one is extrinsic factors, which are illumination, pose variation, scale variations, and noise [30]. However, face detection techniques always require an efficient method to detect the face in various challenging conditions [12]. Figure 2 shows examples of typical challenges associated with faces.

The first one is locating, and detecting the face is not an easy task in motion and when it has a complex environment. The second one is the illumination which affects the image visibility by the magnitude of light intensity as well as patterns of shading and shadows of the visible image. The third one is, pose variations with different rotations of the face orientation. It is one of the serious problems for face identification. However, the face detection system can handle rotation of head movement up to 40° in the Viola-Jones algorithm. It becomes more challenging when it goes to a higher angle if the trained image is with a particular angle. The final one is occlusion, which is a blockage on the face. It is one of the hardest challenges in face detection when the whole face is not available as input images.

Hence, there are several face detection algorithms in deep learning and machine learning. Currently, the Viola-Jones algorithm is still widely used in digital cameras and social networking applications. Many authors have attempted to modify the Viola-Jones algorithm using various techniques for general object detection purposes. However, there is still a gap in identifying problems and exploring new approaches for improvement.

3. Methodology

This section briefly explains the fundamental taxonomy of the Viola-Jones and PSO algorithms and their significance in integrating to enhance face detection performance.

3.1. Viola-Jones algorithm

Viola and Jones [21] introduced the Viola-Jones algorithm for general object detection, but it was later trained using face images for face detection.

The Pseudocode for the AdaBoost algorithm is as follows:

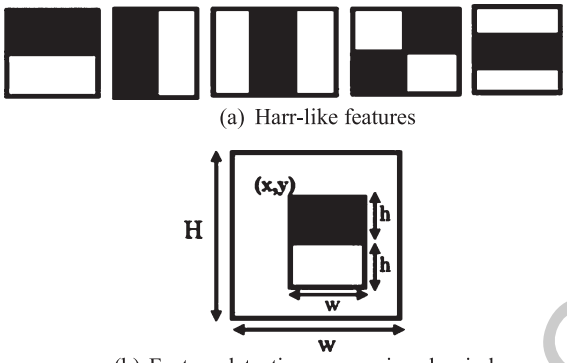
Algorithm 1 Pseudocode for AdaBoost [21]

Given N examples $(x_1, y_1), \ldots (x_i, y_i) .., (x_M, y_M)$ where $y_i \in \{0, 1\}$
Define initial weights $w_{1,i} = 1/2m, 1/2l$ for $y_i = 0, 1$ respectively, where m and l represent the number of negative and positive entries.

for t=1,. . . .,T **do**
(1) For each feature j, train a classifier h_j ()
(2) Evaluate the error of the classifier $\in_t = \sum_{i=0} w_{t,i}.b_i$
(3) Select a classifier $h_t()$ with the minimum error $\in_t$
Update weights: $w_{t+1,i} = w_{t,i}\beta_t^{1-b_i}$ (2)
where b_i=0 if $h_t(x_i) = y_i$, b_i =1 otherwise
with $\beta_t = \in_t/(1- \in_t)$ (3)
end for
Output strong classifier:

$$H(x) = \begin{cases} 1, & if\ sign\ \left(\sum_{t=1}^{T} \alpha_t b_t(I) \geq \theta \sum_{t=1}^{T} \alpha_t\right)\ is\ positive \\ 0, & otherwise \end{cases}$$

With $\alpha_t = log(1/\beta_t)$



(a) Harr-like features

(b) Feature detection process in sub-window

Fig. 3. Haar-like Features for feature extraction.

This system accelerates the detection process by promptly eliminating non-face images upon detection. It employs four fundamental sets of Haar-like features, Integral images, AdaBoost, and Cascade classifier. For identifying facial features, this system utilizes five sets of Haar-wavelet features, which are black-and-white regions subtracted to compute features (refer to Fig. 3(a)). Approximately 1.80 million pixels of features are generated by varying the height, width, and feature position in a 24 × 24 moving window, as depicted in Fig. 3(b). Integral images play a crucial role in rapidly computing these simple features, as defined by Equation (1). To construct a robust classifier, it's important to note the abundance of rectangle attributes associated with sub-windows [35, 13].

$$II(x, y) = \sum_{x^i \leq x, y' \leq y} I\left(x^i, y^i\right) \quad (1)$$

Recalling a substantial number of features selected from rectangles allows for the construction of an effective classifier. The primary objective is to identify the relevant features. The fundamental AdaBoost algorithm is presented in Algorithm 1. This algorithm iteratively trains a weak classifier over T rounds. During training, the algorithm adjusts the weights for samples that were misclassified, increasing the weight for those misidentified and decreasing it for those correctly identified. Likewise, correctly classified samples are less likely to be included in the next iteration, while misclassified samples are given greater consideration. AdaBoost takes a training sample, denoted as $S = (x_1, y_1), \ldots (x_i, y_i) .., (x_m, y_m)$ with a size M, as input. In this context, each sample x_i represents a vector in the domain space X, and y_i represents a label in the label space Y. A weight vector is assigned to each sample and updated during each iteration of the training process (as described in Equation (2)). The error rate for each sample is calculated using Equation (3), as shown AdaBoost algorithm. Based on the weights assigned to the weak classifiers h_1, and h_2, the final strong classifier, $H(x)$, is determined. Therefore, this algorithm is particularly focused on challenging facial samples that are difficult to detect. This research work focuses on binary classification in which $Y = \{0, 1\}$.

3.2. *Particle swarm optimization*

This algorithm works based on a population approach to determine optimal function parameters through a naturally inspired optimization method known as particle swarm optimization (PSO) [5, 18, 25]. PSO, a stochastic gradient technique inspired by the collective behavior of a swarm, was initially proposed by James Kennedy in 1995. In the PSO algorithm, each solution is referred to as a particle

in the search space. These particles have cost values that are optimized by a cost function, and their velocities define their orientation [26, 3].

The PSO process begins with a random population of particles drawn from the original solution space, and their velocities are initialized at irregular intervals within the problem search area. The motion of all particles is guided by the most promising location in the search space, as well as the best position of each particle. The ideal position for each particle is attained by adjusting the particle's speed and acceleration [37]. According to the following equation, the velocity and position of each particle '*i*' are updated at iteration '*t*':

$$V_i(t+1) = V_i(t) + c_1 s_1 (P_i(t) - Q_i(t)) + c_2 s_2 \left(P_i^g(t) - Q_i(t)\right) \quad (4)$$

$$Q_i(t+1) = Q_i(t) + V_i(t+1) \quad (5)$$

Here, s_1 and s_2 represent random values within the range [0,1], while c_1 and c_2 represent cognitive constants, and P and V denote the position and velocity of the particles, respectively. In each iteration, all particles undergo dynamic modifications based on the aforementioned position and velocity, as defined by the algorithm above. Consequently, in most cases, the velocity quickly attains extremely high values, especially for a population distant from its global optimum. PSO consists of two topologies: the Global Neighborhood Topology, which promotes information sharing among particles, and the Ring Topology, which restricts knowledge transfer and prevents a population from converging to a local best solution [3].

4. Combining Viola-Jones with PSO

In this paper, we propose two ways of approaches to optimize the Viola-Jones algorithm for enhancing prediction accuracy and minimizing training errors. First, we employ PSO to select the optimal threshold values in the decision stump for choosing the best features. Second, we optimize AdaBoost to select the optimal features, enhancing face detection accuracy and speed on cluttered real-time images.

The proposed algorithm of Viola-Jones using PSO pseudo-code is given below.

Algorithm 2 Proposed Viola-Jones algorithm using PSO

Input: Original sample images
Output: Identified face will be shown in the bounding box
for i ← 1 to number of different scale images **do**
Load sample image to create $images_i$
Calculate integral image, $images_{i,i}$
Given N labeled samples $(x_1, y_1), \ldots (x_i, y_i) .., (x_N, y_N)$ where $y_i \in \{0, 1\}$ (0,1 class labels)
Initialize $w_{1,i} = 1/2m, 1/2p$ for $y_i = 0, 1$, where m and p represent no. of negative and positive samples respectively.
for t=1,. . . .,T **do**
(1) Initialize the weights: $w_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$ for each feature j, instruct a classifier $h_j()$
(2) **If** t ≤ T/2
Optimize weak classifiers $\left\{h_j()\right\}_{j=1}^{J}$ using PSO algorithm:
$$\{h_t, \in_t\} = \textbf{PSO}\left(\left\{h_j\right\}_{j=1}^{j}, \{x_n, y_n, w_n\}_{n=1}^{N}\right)$$
Evaluate the fitness function for each particle
Select the classifiers $h_t()$ matching to the particles at the best global position
end if
(3) Assess the classifier weights $\alpha_t = log\frac{1-\bar{\text{I}}\mu_t}{\bar{\text{I}}\mu_t}$
(4) Reform weights: $w_{t+1,i} = w_{t,i}\beta_t^{1-b_i}$ where b_i=1 if $h_t(x_i) = y_i$, b_i =–1 otherwise with $\beta_t = \in_t/(1- \in_t)$
end for
Output strong classifier:
$$H(x) = \begin{cases} 1, & if\ sign\ \left(\sum_{t=1}^{T} \alpha_t b_t(I) \geq \theta \sum_{t=1}^{T} \alpha_t\right)\ is\ positive \\ -1, & otherwise \end{cases}$$
With $\alpha_t = log(1/\beta_t)$
If sub-window verified all per-stage checks **then**
Select this sub-window as a face

end if
end for
Optimization Function **PSO ( )** for AdaBoost method
Input arguments $\left\{\{h_i()\}_{j=1}^{j},\ \{x_n, y_n, w_n\}_{n=1}^{N}\right\}$
Define $C_s, C_g = 2, W_{min} = 0.2,\ w = w_{max} = 1.5$
Define random parameters: $r_s, r_g \in [0, 1]$
Define state vector: $X_t^j \in R^D$ *and* $V_t^l \in R^D$ with random values.
for l=1, . . . , L
for i=1, . . . , I
(1) Set a classifier h $(X_t^l; ; x)$ to the training examples using weights Adw_n
(2) Evaluate $\in_t^j = \dfrac{\sum_{n=1}^{N} w_n\ X\ \left|h\left(X_i^l;;x_n\right) - y^n\right|}{\sum_{n=1}^{N} W_n}$ (8)
(3) Updates the particles:
$V_i\,(t+1) = V_i\,(t) + c_s r_s\,(Q_1\,(t) - X_i\,(t)) + c_g r_g(Q_2^g\,(t) - X_i\,(t))$
$X_i\,(t+1) = X_i\,(t) + V_i(t+1)$
(4) Update the personal best point Q_i^s , if necessary
end for
(5) Update the global best point Q^g, if necessary
(6) Update momentum: w $w_{max} - \frac{l}{L}\,(w_{max} - w_{min})$
end for
return $\left\{h_{Q^g}(), \in_{h_{Q^g()}}\right\}$

4.1. Selecting threshold value using PSO

In the proposed method, for selection of threshold values, a weak classifier has significantly improved the computational efficiency of the base algorithm by utilizing a decision tree with two leaves, commonly known as a decision stump instead of exhaustively searching for a multitude of features to construct a weak classifier, we employ PSO to pinpoint the optimal decision stump threshold. When decision stumps are employed as weak classifiers on complex datasets, the algorithm must explore all possible thresholds to minimize training error. Consequently, finding the best threshold values can be time-consuming. In such cases, an evolutionary search strategy PSO is invaluable as a proposed approach, accelerating the training of an AdaBoost classifier. Furthermore, each iteration of the PSO approach is dedicated to learning a new weak classifier, and through numerous runs, it may uncover the ideal set of values that collectively form a strong classifier.

In the PSO algorithm, the cost function is utilized to optimize each particle within the entire solution space. Particles employ thresholding values to categorize the solution space into two classes: 1 (representing 'face') and 0 (representing 'non-face'). In the initial stage, sample values greater than the threshold are classified as 1, while values below the threshold are classified as 0. This classification is reversed in the subsequent stage during the training of weak classifiers. The training loss is computed for each subgroup, and the weak classifier output with the lowest error is selected. For instance, let $S = ((x_1, y_1) \ldots . (x_n, y_n))$ represent a training set of weak classifiers, where the labels $y_i \in \{0, 1\}$. To calculate each particle, a decision stump requires three parameters: the decision limit (+1 or 0), index characteristics (j), and the optimized threshold value to split the solution space. For input examples x, Equation (6) defines the positive cost function, while Equation (7) defines the negative stump.

$$h_{j,\theta} + (x) = \begin{cases} +x, & if\ x(j) \geq\ \theta \\ x, & otherwise \end{cases} \tag{6}$$

$$h_{j,\theta} - (x) = \begin{cases} -x, & if\ x(j) \geq\ \theta \\ x, & otherwise \end{cases} \tag{7}$$

The computational cost of the enhanced Viola-Jones algorithm depends on two factors: the population size (S) and the number of iterations (T). Each step in the boosting procedure optimizes the S × T classifiers. PSO is employed to select the best threshold Haar-like features in the AdaBoost.

4.2. Selecting the best features in the AdaBoost algorithm using PSO

Enhancing the speed of the face detector without compromising classifier accuracy is a crucial objective. However, the exhaustive feature selection process in AdaBoost often leads to increased com-

Fig. 4. Training and testing sample images (a) Face images (Positive Images) [23] (b) non-face images (Negative Images) [24].

plexity. Furthermore, the limited learning capacity of the simple decision stump classifier reduces the efficiency of conventional face detection approach. To address this, we have incorporated the PSO in AdaBoost for the selection of the optimal features for face detection and optimizing the computational processing time. Considering these factors, we propose two improvements to our face detector to reduce the computational burden of feature selection and enhance the selection speed. Firstly, we employ PSO to select optimized threshold values, as discussed in the previous section. Lastly, we combine the PSO technique with the AdaBoost algorithm, enabling rapid exploration of the entire feature space and the selection of the most optimal feature sets, thus expediting the training process and minimise the training error. Algorithm 2 illustrates the proposed Viola-Jones using PSO approach.

In the AdaBoost classifier, exhaustive searches are conducted each time to select relevant features and minimize classification errors. To address the high complexity associated with this exhaustive search, we introduced the use of PSO within the AdaBoost algorithm. PSO is applied to explore potential feature locations, sizes, orientations, and combinations, resulting in the selection of a discriminative feature set. These selected features are then incorporated into AdaBoost to construct an ensemble classifier. The PSO demonstrates efficient search capabilities compared to exhaustive search techniques. In PSO, each particle could explore not only its own space but also the spaces of other particles. Consequently, many particles collectively strive to identify the best possible positions. However, this collaborative approach can lead to a decline in the diversity of selected features as we integrated a random feature selection approach. Specifically, we initially employ PSO to identify the most relevant features at an early stage. As the boosting phase unfolds, our proposed approach transitions to random feature selection to uncover additional discriminative features, thus expanding the pool of candidate features. This adjustment strikes a balance between efficient feature selection and the preservation of feature diversity, enabling us to discover a wider range of optimal features during the boosting process.

5. Experiments and discussion

This section analyzes the performance of the conventional Viola-Jones algorithm and compares it with an improved approach incorporating PSO optimization. In the conventional Viola-Jones algorithm, AdaBoost employs an exhaustive search to build a weak classifier, while in the proposed approach, AdaBoost utilizes an optimized search to select the best features and threshold values. Furthermore, significantly reduced the false positive rate.

5.1. Dataset description

Face images were collected from the Yearbook Dataset of frontal-facing American high-school seniors [23], while non-face images were obtained from the Stanford Background Dataset [24] and ImageNet [10]. These images are used for both training and testing purposes. These databases contain 4,999 different face images and 6,960 non-face images, all with a pixel resolution of 25×25. The positive and negative images are randomly divided into two folders. The training folder comprises 1,200 positive and 1,000 negative grayscale images (See Fig. 4). The test set consists of 750 positive and 658 negative images. This experiment was validated using the Wider Face test benchmark dataset, which includes various real-time facial detection challenge images

[30]. The dataset comprises 32,203 photographs and labels 393,703 faces, covering a wide range of scales, poses, and occlusions.

5.2. Evaluation metrics

The face detection algorithm has two classes: faces and non-faces. The performance of the proposed method is evaluated using Equations (9) and (10). The True Positive Rate (TPR) is used to measure how well the model correctly predicts the positive class. The equation for TPR is given below:

$$\text{TPR} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} \tag{9}$$

False Positive Rate (FPR) is used to measure the outcome of the model that incorrectly predicted the negative classes. This equation is given below:

$$\text{FPR} = \frac{\text{False Positive (FP)}}{\text{False Positive (FP)} + \text{True Negative (TN)}} \tag{10}$$

To construct the weak classifier for selecting the best threshold value, we examined the particle's size and maximum iteration using the ImageNet dataset. The results are displayed in Table 1 showing the performance of various particle sizes and their iterations. The selected optimal threshold value is then applied in the feature selection section of AdaBoost to optimize the features and computation time. Besides, the best PSO parameters were chosen according to Table 1 (Particle size 20 and iteration 100).

5.3. Parameter setting and threshold selection

To analyse the proposed approach reliability, accuracy and time spend of each sample parameter are used. The proposed approach consists of 200 particles and could run for up to 1,000 iterations for constructing a weak classifier. However, it terminates if there is improvements are observed in the feature selection process within the global solution search space. Initially, the population is randomly defined, with the feature selection parameters (x, y, w, h) in the range of [0, 250] and the feature type in the range of [0, 4]. The social value parameters are set to c_1 and c_2, both ranging from 100 to -100. Random values are independently sampled from the range [0, 1], and Q1 and Q2 are both set to 3.05. These experiments were conducted with 1,000 iterations, and the results,

Table 1
The training error of the training dataset

S (No. of particles)	T (Iterations)	Training Error
5	50	0.9034
5	100	0.8912
10	100	0.8713
10	50	0.9613
20	**100**	**0.8531**
20	50	0.1034
30	100	0.8989

Table 2
Number of features needed for detection

	Best	Average	Worst
Viola-Jones	340	340	340
Proposed	120	134	150

including the best, worst, and average, are reported in Table 2. These experiments were run on Google Colab with GPU K80.

5.4. Feature selection using PSO

The proposed approach is analyzed in terms of classifier accuracy and execution time. The experiments were repeated ten times for each algorithm, and the best, average, and worst outcomes are presented in Table 2. Additionally, the analysis indicates the number of features required for face detection process on complex face images. The performance of face detection is influenced by both the population size and the number of PSO iterations. The results demonstrate the effectiveness of PSO for optimal feature selection in this problem when compared to the conventional Viola-Jones algorithm.

This experiment reveals that the proposed method utilizes as few features as possible compared to the conventional algorithm. Table 2 shows the number of features generated for a strong classifier during the training process. Viola-Jones with PSO required only 120 features in the best case, 134 in the average case, and 150 features in the worst case for building the weak classifiers, whereas the conventional Viola-Jones algorithm required 340 features in the best case. Therefore, the proposed method constructs superior classifiers using only 120 features in the best case, which is significantly fewer than the conventional Viola-Jones method.

Table 3 summarizes the overall comparison between the conventional Viola-Jones and the proposed method on the ImageNet test dataset. The proposed approach achieved an average classification

Table 3
Face detection performance

Approach	TPR(Average)	FPR(Average)	Time
Viola-Jones	96.63 %	0.0337	52.5s
Proposed	98.73 %	0.0127	30.6s

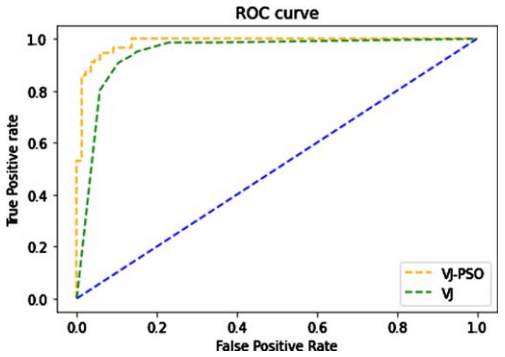


Fig. 5. ROC curves of Conventional-VJ and VJ with PSO.

accuracy rate of 98.73% and a False Positive Rate (FPR) of 1.27% on the face images when testing. In contrast, the conventional Viola-Jones algorithm achieved an average accuracy of 96.63% and a 3.37% FPR on the face images. These results indicate that the proposed method is not only more efficient than the Conventional-VJ algorithm but also more effective in classifying unseen datasets. Furthermore, the proposed method requires less time to test the dataset compared to the conventional algorithm.

The performance of the proposed approach and the conventional Viola-Jones algorithm is depicted using the Receiver Operating Characteristic (ROC) curve (See Fig. 5). The performance of the proposed method is represented in orange color, whereas the conventional Viola-Jones algorithm's performance is shown in green. According to the ROC curve, the proposed approach achieved a 98.73% accuracy on the face and non-face image dataset, whereas the conventional algorithm achieved 96.63%. After a successful testing process, the proposed approach converted as face detection model also saved as .XML file, allowing it to detect faces in various challenging contexts. Finally, a comparison of the performance of the classic machine learning based face detection algorithm and the proposed technique is presented in Table 4. The Viola-Jones algorithm with PSO performed effectively in various face detection complex real-time face images, including scale variation, illumination, pose variation, and occlusion, compared to

Table 4
Performance comparison of the proposed method with another approach in Face Detection

Face detection approach	Accuracy
Viola–Jones, Geometric Distribution [17]	95%
Viola–Jones, Condensation Algorithm (CA), NN, SVM [6]	95%
Skin Color Algorithm, Circular Hough Transform [1]	80%
Kalman Filter, Principal Component Analysis (PCA), Local-Binary-Pattern (LBP), SVM [2]	95%
Proposed (Viola-Jone with PSO)	**98.73%**

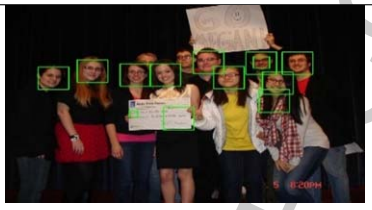
the conventional method. Table 5 shows the results of prosed approach detection performance and conventional approach.

The computational complexity of the optimized Viola-Jones algorithm is determined by two parameters: S, which represents the number of particles, and T, which represents the number of iterations. In contrast, the computational complexity of the AdaBoost algorithm is determined by the parameter N, denoting the number of samples. The time complexity of the proposed algorithm at each stage of the boosting technique is O(S × T), whereas the time complexity in the base model is O(N^2). The basic AdaBoost technique trains a weak classifier in polynomial time, while the improved PSO-based Viola-Jones algorithm's time complexity scales linearly with S and T.

6. Conclusion

In this paper, we propose an efficient and enhanced face detection approach using PSO in Viola-Jones to improve prediction accuracy for complex real-time face images. This research work aims to enhance the optimal feature selection process and global threshold determination in AdaBoost and Haar-like features using PSO. The use of PSO enables a reduction in false-positive rates and computational time significantly. The proposed approach constructs a more efficient weak classifier for face detection in complex face images. Instead of an exhaustive feature search, PSO optimizes the selection process, leading to better performance. The proposed method is validated on the Wider face detection benchmark and demonstrated superior results compared to the conventional algorithm. It achieved an impressive average true positive rate of 98.73% with only a 1.27% false positive rate. Additionally, the proposed approach significantly reduced face detection time

Table 5
Performance comparison of proposed and conventional Viola-Jones algorithm on the Wider benchmark dataset [30]

Challenges in Face Detection	Conventional Viola-Jones	Viola-Jones with PSO
Scale variation		
		
Illumination		
		
Pose variation		
		

(*Continued*)

Table 5
(*Continued*)

Occlusion		
		

on the test samples. Although the proposed approach outperformed the conventional algorithm in terms of true positive rate, longer training time on the dataset. The results suggest that the proposed method can be a promising solution for achieving accurate and rapid face detection in various applications.

Acknowledgment

The authors convey sincere thanks to the ISO Certified (ISO/IEC 20000-1:2018) Centre for Machine Learning and Intelligence (CMLI) funded by the Department of Science and Technology (DST-CURIE), India for providing the facility to carry out this research study.

Conflict of interest

The authors declared that no conflict of interest in this work.

Author's contribution

P. Subashini developed the methodology and design of the manuscript. Diksha Shukla contributed to the text and content of the manuscript, including entire revisions and edits. M. Mohana conducted the experiments, compared the results, created the figures, and drafted the entire manuscript with contributions from the co-authors. All authors were involved in conducting the experiment and analyzing the results. They have reviewed and approved the content of the manuscript and are willing to be held accountable for the work.

Ethics approval and consent to participate

This study was approved by the Avinashilingam Human Ethics Committee, Coimbatore, India. The approval number is AUW/IHEC/CS-21-22/XMT-03.

Data availability

Open-source Wider Face, Yearbook, and ImageNet Datasets.

Fund availability

There is no external funding for this research study.

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors utilize Grammarly assistant tools included in Microsoft Word for grammar checking. After using these tools, the authors reviewed and edited the content as necessary and take full responsibility for the publication's content.

References

[1] A. Dasgupta, A. George, S.L. Happy and A.A. Routray, Vision-based system for monitoring the loss of attention in automotive drivers, *IEEE Transactions Intelligent Transportation Systems* **14**(4) (2013), 1825–1838.

[2] A. Kumar, A. Kaur and M. Kumar, Face detection techniques: A review, *Artificial Intelligence Review* **52** (2019), 927–948.

[3] A. Sharma and N. Singh, Object detection in image using particle swarm optimization, *International Journal of Engineering and Technology* **2**(6) (2010), 419–426.

[4] A. Taherkhani, G. Cosma and T.M. McGinnity, AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing* **404** (2020), 351–366.

[5] A.W. Mohemmed, M. Zhang and M. Johnston, Particle swarm optimization based adaboost for face detection. In *2009 IEEE Congress on Evolutionary Computation* (2009), 2494–2501. IEEE.

[6] B. Fatima, A.R. Shahid, S. Ziauddin, A.A. Safi and H. Ramzan, Driver fatigue detection using viola jones and principal component analysis. *Applied Artificial Intelligence* **34**(6) (2020), 456–483.

[7] C.A. Perez and J.I. Vallejos, Face detection using PSO template selection. In *2006 IEEE International Conference on Systems, Man and Cybernetics* **5** (2006), 4220–4224. IEEE.

[8] F. Marini and B. Walczak, Particle swarm optimization (PSO). A tutorial. *Chemometrics and Intelligent Laboratory Systems* **149** (2015),153–165.

[9] H.A. Hosni Mahmoud and H.A. Mengash, A novel technique for automated concealed face detection in surveillance videos. *Personal and Ubiquitous Computing* **25** (2021), 129–140.

[10] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li and L. Fei-Fei, Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255. IEEE.

[11] J. Kennedy and R. Eberhart, Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks* **4** (1995), 1942–1948.

[12] J.C. Zhang and W. Fan, AdaBoost face detection algorithm based on correlation, *Computer Engineering* **37**(8) (2010), 158–163.

[13] J. Huang, Y. Shang and H. Chen, Improved Viola-Jones face detection algorithm based on HoloLens. *EURASIP Journal on Image and Video Processing* **1** (2019), 1–11.

[14] J. Zhang and Q.W. Ye, Improved AdaBoost face detection algorithm based on dual features, *Wireless Communication Technology* **29**(2) (2020), 23–27.

[15] K.C. Kirana, S. Wibawanto and H.W. Herwanto, Facial emotion recognition based on Viola-Jones algorithm in the learning environment. In *2018 International seminar on application for technology of information and communication* (2018), 406–410. IEEE.

[16] L. Zhang and P. Lenders, Knowledge-based eye detection for human face recognition. In *KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings (Cat. No. 00TH8516)*, **1** (2000), 117–120. IEEE.

[17] M.J. Flores, J.M. Armingol and A.de la Escalera, Real-time warning system for driver drowsiness detection using visual information, *Journal of Intelligent & Robotic Systems* **59** (2010), 103–125.

[18] M. Mohammadpour, M. Ghorbanian and S. Mozaffari, AdaBoost performance improvement using PSO algorithm. In *2016 Eighth international conference on information and knowledge technology (IKT)* (2016), 273–275. IEEE.

[19] P. Li, H. Wang, Y. Li and M. Liu, Analysis of face detection based on skin color characteristic and AdaBoost algorithm, *Journal of Physics: Conference Series* **1601**(5) (2020), 052019.

[20] P. Bose and S. Bandyopadhyay, Human face and facial parts detection using template matching technique, *International Journal of Engineering and Advanced Technology* **9**(4) (2020), 2249-8958.

[21] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR* (2001), 1. IEEE.

[22] R. Belaroussi and M. Milgram, A comparative study on face detection and tracking algorithms, *Expert Systems with Applications* **39**(8) (2012), 7158–7164.

[23] S. Ginosar, K. Rakelly, S. Sachs, B. Yin and A.A. Efros, A century of portraits: A visual historical record of American high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2015), 1–7.

[24] S. Gould, R. Fulton and D. Koller. Decomposing a scene into geometric and semantically consistent regions. *Proceedings of International Conference on Computer Vision (ICCV)* (2009).

[25] S. Ma and T. Du, Improved adaboost face detection. In *2010 International Conference on Measuring Technology and Mechatronics Automation* **2** (2010), 434–437. IEEE.

[26] S. Cagnoni, M. Mordonini and J. Sartori, Particle swarm optimization for object detection and segmentation. In *Workshops on Applications of Evolutionary Computation* (2007), 241–250.

[27] S. Minaee, P. Luo, Z. Lin and K. Bowyer, Going deeper into face detection: A survey. arXiv preprint arXiv:2103.14983 (2021).

[28] S. Singh and S.V.A.V. Prasad, Techniques and challenges of face recognition: A critical review, *Procedia Computer Science* **143** (2018), 536–543.

[29] S. Soleymani, B. Chaudhary, A. Dabouei, J. Dawson and N.M. Nasrabadi, Differential morphed face detection using deep Siamese networks. In *International Conference on Pattern Recognition*. Cham: Springer International Publishing (2021), 560–572.

[30] S. Yang, P. Luo, C.C. Loy and X. Tang, Wider face: a face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 5525–5533.

[31] T. Ephraim, T. Himmelman and K. Siddiqi, Real-time viola-jones face detection in a web browser. In *2009 Canadian Conference on Computer and Robot Vision* (2009), 321–328. IEEE.

[32] T.G. Dietterich, Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Proceedings* **1** (2000), 1–15. Springer Berlin Heidelberg.

[33] W. Lee, C.H. Jun and J.S. Lee, Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification. *Information Sciences*, **381** (2017), 92–103.

[34] W. Kong, L. Zhou, Y. Wang, J. Zhang, J. Liu and S. Gao, A system of driving fatigue detection based on machine vision and its application on smart device, *Journal of Sensors* **2015**(1) (2015), 1-11.

[35] W.Y. Lu and Y.A.N.G. Ming, Face detection based on viola-jones algorithm applying composite features. In *2019 International Conference on Robots & Intelligent System (ICRIS)* (2019), 82–85. IEEE.

[36] X. Yang, N. Liang, W. Zhou and H. Lu, A face detection method based on skin color model and improved AdaBoost algorithm, *Traitement du Signal* **37**(6) (2020), 929-937.

[37] Y. Wu and X. Ai. Face detection in color images using AdaBoost algorithm based on skin color information. In *First International Workshop on Knowledge Discovery and Data Mining (WKDD)* (2008), 339–342. IEEE

[38] Z. Zakaria and S.A. Suandi, Face detection using combination of Neural Network and Adaboost. In *TENCON 2011–2011 IEEE Region 10 Conference* (2011), 335–338, Bali, Indonesia.

[39] Z. Zakaria, S.A. Suandi and J. Mohamad-Saleh, Hierarchical skin-AdaBoost-neural network (H-SKANN) for multi-face detection, *Applied Soft Computing* **68** (2018), 172–190.

[40] D. Qi, W. Tan, Q. Yao and J. Liu, YOLO5Face: Why reinventing a face detector. In *European Conference on Computer Vision*. Cham: Springer Nature Switzerland (2022), 228–244.

[41] W. Chen, H. Huang, S. Peng, C. Zhou and C. Zhang, YOLO-face: A real-time face detector. *The Visual Computer* **37**(2021), 805–813.

[42] R. Ranjan, A. Bansal, J. Zheng, H. Xu, J. Gleason, B. Lu, A. Nanduri, J.C. Chen et al., A fast and accurate system for face detection, identification, and verification, *IEEE Transactions on Biometrics, Behavior, and Identity Science* **1**(2) (2019), 82–96.

[43] D. Mamieva, A.B. Abdusalomov, M. Mukhiddinov and T.K. Whangbo, Improved face detection method via learning small faces on hard images based on a deep learning approach, *Sensors* **23**(1) (2023), 502.

[44] J. Yu, X. Hao and P. He, Single-stage face detection under extremely low-light conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 3523–3532.

[45] P. Hedman, V. Skepetzis, K. Hernandez-Diaz, J. Bigun and F. Alonso-Fernandez, On the effect of selfie beautification filters on face detection and recognition, *Pattern Recognition Letters* **163** (2022), 104–111.

[46] S.E. Whang, and J.G. Lee. Data collection and quality challenges for deep learning, *Proceedings of the VLDB Endowment* **13**(12) (2020), 3429–3432.

[47] V.K. Sharma and R.N. Mir, Saliency guided faster-RCNN (SGFr-RCNN) model for object detection and recognition, *Journal of King Saud University-Computer and Information Sciences* **34**(5) (2022), 1687–1699.

[48] P.T. Krishnan, P. Balasubramanian, V. Jeyakumar, S. Mahadevan and A. Noel Joseph Raj, Intensity matching through saliency maps for thermal and visible image registration for face detection applications, *The Visual Computer* **39**(10) (2023), 4529–4542.

[49] M. Besnassi, N. Neggaz and A. Benyettou, Face detection based on evolutionary Haar filter, *Pattern Analysis and Applications* **23** (2020), 309–330.

[50] K. Babu, C. Kumar and C. Kannaiyaraju, Face recognition system using deep belief network and particle swarm optimization, *Intelligent Automation & Soft Computing* **33**(1) (2022), 317-329.