DeepFake Image Detection Using Transfer Learning and Attention-Enhanced EfficientNet

Sarvesh S. Gharat
MTech Student, D Y Patil Deemed to be University, Navi Mumbai, India

Dr. Puja Padiya
Professor, D Y Patil Deemed to be University, Navi Mumbai, India

Dr. Amarsinh Vidhate
Head of Department (HOD), D Y Patil Deemed to be University, Navi Mumbai, India

Abstract

Deepfake detection involves identifying facial images that have been synthetically generated or manipulated to closely resemble real human faces. With the advancement of generative adversarial networks (GANs) such as StyleGAN2, detecting these highly realistic forgeries has become increasingly challenging for traditional methods. Conventional machine learning algorithms and shallow neural networks struggle to capture the fine-grained pixel-level details and semantic context necessary for accurate classification. To overcome these limitations, this study employs advanced deep learning techniques, particularly convolutional neural networks (CNNs), combined with state-of-the-art pretrained architectures including VGG16, InceptionResNet, Xception, MobileNet, and EfficientNet-B2. Leveraging transfer learning, each model was fine-tuned to perform binary classification (real vs. fake) on a comprehensive dataset comprising authentic and fabricated face images. Performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Notably, EfficientNet-B2 integrated with an attention mechanism delivered the highest accuracy of 83%, demonstrating superior capability in focusing on critical facial features and resisting complex deepfake manipulations. This work presents a robust and scalable framework for real-time deepfake detection, offering enhanced accuracy and reliability. The proposed approach holds significant promise for reinforcing digital media integrity and safeguarding against misinformation in an era of increasing synthetic content proliferation.