DEEPFAKE DETECTION VIA ENSEMBLE TECHNIQUES

Project Reference No.: *47S_BE_3598*

College ***:*** *Vidyavardhaka College of Engineering, Mysuru*
Branch ***:*** *Department of Artificial Intelligence & Machine Learning*
Guide(s) ***:*** *Prof. Anjali R*
Student(S) ***:*** *Mr. Bhuvan J*
Ms. Manogna C G
Ms. Medha M
Ms. Pragya S Babu

Keywords:

Convolution Neural Network (CNN), Ensemble Model, Glitch detection, lip movement analysis, Deepfake detection challenge, Celeb datasets, Kaggle datasets

Introduction:

The rise of deepfake technology has brought about unprecedented challenges concerning the authenticity and reliability of digital content. Deepfakes, characterized by their convincingly realistic nature, present considerable risks across various sectors, including politics, entertainment, and cybersecurity. As the sophistication of deepfake creation tools continues to increase, there is a pressing need for robust detection mechanisms capable of discerning between genuine and manipulated media. This study aims to address this imperative by proposing an integrated framework for deepfake detection, encompassing both image and video-based methodologies. By leveraging advanced machine learning algorithms and innovative pre-processing strategies, our framework endeavours to offer a dependable solution for identifying and mitigating the proliferation of deepfake content.

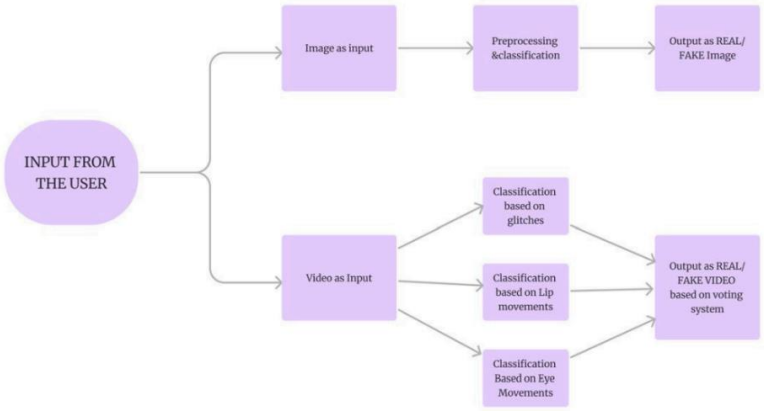
In the domain of deepfake detection, researchers have explored diverse techniques and methodologies to tackle the pervasive challenge of identifying manipulated media. Studies have investigated specialized tools and frameworks tailored to distinct aspects of deepfake detection, including facial recognition and gaze analysis.

For instance, Nicolo Bonettini's work highlights the efficacy of ensembles of Convolutional Neural Networks (CNNs) for video based deepfake detection, leveraging pre-trained networks such as EfficientNetB4 and incorporating attention mechanisms and Siamese training. These ensembles have demonstrated impressive accuracy in identifying manipulated videos, although challenges persist, such as susceptibility to specific manipulation techniques and limitations with low-quality videos. Meanwhile, approaches like Gaze Forensics, introduced by Oscar de Lime, focus on gaze analysis to detect manipulation artifacts such as shifting iris color or unnatural reflections. Despite its promise, this methodology may overlook manipulations crafted using alternative techniques, suggesting the need for continued exploration and refinement in deepfake detection methodologies.

Objectives:

- Develop a robust deep learning-based system for deepfake detection.
- Achieve a high level of accuracy in identifying and classifying deepfakes.
- Ensure consistency in results across different users and settings.
- Improve accessibility to deepfake detection by creating a user-friendly and widely applicable tool.

Methodology:



The methodology begins with the acquisition of a diverse dataset sourced from reputable platforms like Kaggle and the DFDC dataset, ensuring a comprehensive mix of real and synthetic content essential for effective model training. Development proceeds with the creation of individual deepfake detection models, each tailored with convolutional neural network architectures and trained using binary cross-entropy loss and the Adam optimizer. Techniques such as data augmentation, dropout regularization, and early stopping are systematically applied to bolster model robustness and mitigate overfitting risks. Integration of predictions from these models culminates in the construction of an ensemble model, meticulously optimized through iterative training on distinct training and validation sets, refining ensemble composition and combination strategies along the way.

Evaluation encompasses a meticulous partitioning of the dataset into training, validation, and test sets, with a suite of standard classification metrics deployed to gauge model performance comprehensively, including accuracy, precision, recall, F1 score, and AUC-ROC. Robustness testing follows suit, scrutinizing model performance across varied datasets, deepfake generation techniques, and compression formats to ensure its resilience under diverse conditions. Frontend development harnesses React.js for crafting interactive user interfaces, complemented by Flask for backend logic, facilitating seamless communication through RESTful API creation. The user interface design prioritizes simplicity and usability, fostering easy upload and analysis of images or videos, with visualization tools providing nuanced insights into detection results and detected anomalies. Deployment enables researchers and end-users alike to access the platform for deepfake analysis and validation, with a commitment to continuous improvement

guided by user feedback and ongoing research advancements, refining model performance to meet evolving challenges head-on.

Conclusion:

We've achieved a remarkable milestone by developing a comprehensive deepfake detection model capable of discerning manipulated content in both images and videos. This achievement is particularly notable as it addresses a longstanding gap in the field, where previous models often specialized in either images or videos, but not both. Our ensemble approach for video classification, integrating multiple models for glitch detection, lip movement analysis, and eye tracking, has significantly bolstered the model's resilience and effectiveness. Specifically, our image classification model stands out with an impressive accuracy rate of approximately 91% across diverse datasets and for video - glitch detection shows 98% , lip movement shows 94% and eye movement shows 93% accuracy. So, on average we have around 95% accuracy for our ensemble model. This high accuracy ensures dependable identification of deepfake content within static images, thereby enhancing the overall efficiency of our model. Similarly, our ensemble video model, which incorporates glitch detection and facial movement analysis, attains remarkable accuracy in discerning manipulated videos. These achievements underscore the robustness and reliability of our deepfake detection model across different media formats. Despite these notable successes, challenges persist due to the scarcity of high-quality datasets for both images and videos. The absence of such datasets limits our model's exposure to extreme cases of deepfakes, potentially affecting its accuracy in detecting highly sophisticated manipulations. Thus, while our model demonstrates impressive accuracy overall, addressing the shortage of quality data remains a critical priority to further refine and enhance its performance, especially in the face of evolving deepfake technologies.

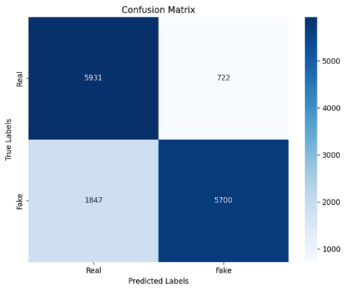


Fig1. Confusion Matrix of Glitch Detection Model

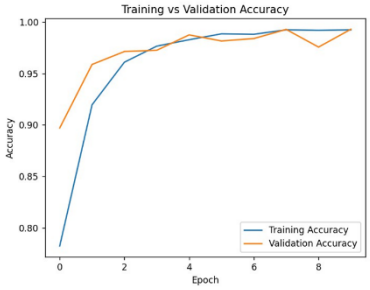


Fig2. Training v/s Validation Graph of Glitch Detection Model

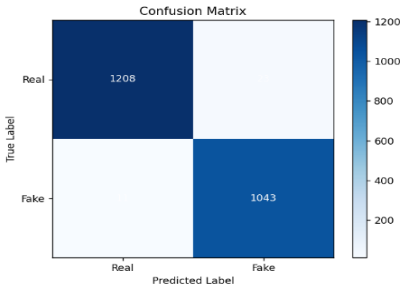


Fig3. Confusion Matrix of Lip Movement Detection Model

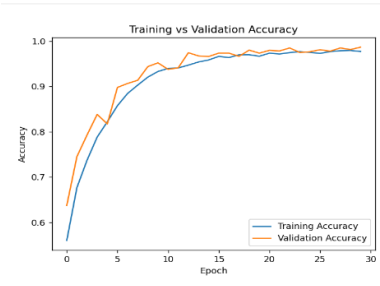


Fig4. Training v/s Validation Graph of Lip Movement Detection Model

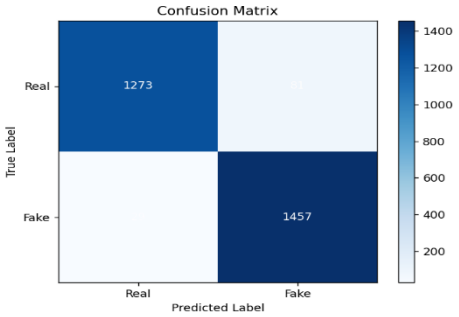


Fig5. Confusion Matrix of Eye Movement Detection Model

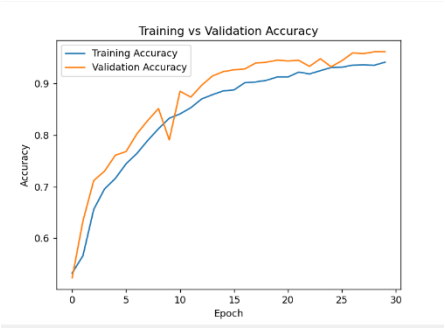


Fig6. Confusion Matrix of Glitch Detection Model

In conclusion, our project represents a significant step forward in the ongoing battle against the proliferation of deepfake content in the digital landscape. In response to the urgent need for robust detection mechanisms, we have successfully developed a comprehensive deepfake detection system capable of accurately identifying manipulated media in both images and videos. Leveraging advanced techniques from computer vision and machine learning, our model achieves remarkable accuracy rates, with our image classification model boasting an impressive accuracy rate of approximately 91% across diverse datasets. Similarly, our ensemble video model, incorporating glitch detection and facial movement analysis, attains remarkable accuracy in discerning manipulated videos. By preserving trust in digital media and mitigating the spread of misinformation, we aim to foster a safer and more trustworthy online environment for all users. Moving forward, we remain committed to advancing our deepfake detection system, incorporating feedback and insights gained from real-world applications to combating the ever-evolving threat of deepfake content

Scope for future work:

In the pursuit of advancing our deepfake detection model, several avenues for future enhancement present themselves. Chief among these is the urgent need to address the scarcity of high-quality datasets, a challenge that has limited the model's exposure to highly accurate deepfake content. To tackle this issue, future efforts will focus on sourcing and curating diverse datasets containing authentic and manipulated media of the highest fidelity, facilitated through collaborations with research institutions, industry partners, and data providers. By fostering partnerships with experts from diverse backgrounds and promoting open access to research findings, datasets, and methodologies, we can accelerate progress towards more effective deepfake detection mechanisms, ultimately contributing to a safer and more trustworthy digital environment for all users.