International Journal of ADVANCED RESEARCH & INNOVATIONS

*Volume 1, Issue-11*
*November, 2025, P05*
ISSN: 2319-9253

# EKYC-DF: A REALISTIC DEEPFAKE CORPUS FOR TESTING AND TRAINING EKYC VERIFICATION MODELS

[1]Sania Mirza, M.Tech, Dept of CSE,
[2]Mrs. Y. Susheela, Associate Professor, Department of CSE,
Vaageswari College of Engineering (Autonomous), Karimnagar, Telangana.

**Abstract:** Digital registration methods, such as electronic Know Your Customer (eKYC) checks, have become increasingly difficult to validate as a result of the widespread availability of deepfake computer technology. This has made the task of validating digital registration procedures more challenging. Within the context of deepfake attacks, the eKYC-DF corpus is a particular dataset that has the potential to be employed for the purpose of evaluating and strengthening facial recognition systems. There are opportunities for both of these uses. This sample contains a considerable number of phony facial recordings, which are included in the collection. These phony recordings bear a strong resemblance to the ones that were actually available. The lighting, editing, and racial composition of the recordings are all notably different from one another, which makes it simple to discern amongst the recordings. By developing more effective methods of identity verification, researchers and developers have the power to safeguard the trust that users have in the internet and stop persons from gaining access to systems without authorization. This will allow them to enhance the security of electronic know-your-customer (eKYC) systems, which will allow them to better protect their customers.

***Keywords:*** Deepfake, eKYC Verification, Facial Recognition, Synthetic Dataset and Identity Fraud Prevention



## 1. Introduction

Remote identity verification has been crucial in sectors including banking, telecommunications, and insurance with the introduction of digital services. By removing paperwork, eKYC standards streamline the process of acquiring new clients by permitting the use of biometric technologies, such as facial recognition, for real-time authentication. The proliferation of deepfake technology has emerged as a fundamental concern, despite the fact that technological advancements have made banking and communication services more accessible.

The so-called "deepfakes," which are computer-generated movies and pictures that appear authentic, may now fool facial recognition algorithms. Fraudsters can create phoney names that look legitimate using publicly accessible images and videos, evading standard security measures. With more businesses using automated verification techniques, identity fraud is increasing. This is due to the flexibility of eKYC procedures.

Because deepfake techniques are not detected by popular anti-spoofing solutions, strict security measures are required. Researchers and developers have encountered several difficulties in creating deepfake-resistant verification models since no real datasets adequately capture the issues that occur in eKYC scenarios. It is challenging to create facial recognition systems that can withstand deepfake attacks in the absence of trustworthy training data.

This is a great function of the eKYC-DF collection. In order to train and evaluate facial recognition systems against deepfake threats, this unique dataset includes real fraud attempts made during digital onboarding. A variety of spoof recordings with different lighting conditions, ethnic backgrounds, and spoofing techniques are included in the collection. eKYC-DF is a crucial instrument for enhancing the security of identity verification by developing a regulated and flexible framework.

The focus on security-critical applications, such banking and financial services, sets eKYC-DF apart from general-purpose deepfake datasets. Seeking prior consent and, if required, anonymizing the data emphasize responsibility in data collection. The dataset can be used by programmers and scientists to evaluate the correctness of models, enhance their approaches to problem-solving, and defend the system from hacker attacks.

The importance of protecting digital identification systems increases with the development of deepfake technology. An important development in the fight against fraud in verification procedures is EKYC-DF. The resources required to increase the security and resilience of digital environments are provided to the government, businesses, and cybersecurity experts.

Volume 1, Issue-11
November, 2025, P05
ISSN: 2319-9253

This collection is essential for safeguarding people and companies against emerging dangers since it enhances identity verification systems with deepfake-resistant capabilities.

## 2. Review of Literature

Kinnunen, T., et al. (2020). In order to remain competitive, speaker identification systems must adapt to the increasing sophistication of fraudsters in circumventing security measures. This paper provides a method for simultaneously testing anti-spoofing solutions and automatic speaker verification (ASV) systems, rather than sequentially testing them. The ASVspoof database is employed in the study to demonstrate that a more precise assessment of the functionality of a system is obtained by examining both components simultaneously. It also discusses significant trade-offs, the constraints of the datasets that can be utilized, and strategies for enhancing the resilience of systems to the emergence of false attacks. The data can serve as a foundation for future research on the enhancement of the security of voice authentication systems.

Dutta, S., & Bhattacharya, S. (2021). Artificial intelligence is a critical component of the rapidly evolving field of digital identity identification. This study investigates the efficacy and utility of electronic Know Your Customer (eKYC) methods when implemented in conjunction with deep learning. The authors' AI-powered technology surpasses manual methods by automatically verifying IDs through facial recognition and document analysis. Additionally, they address challenges that are distinctive to their industry, including compliance, scalability, and data security. The article demonstrates the potential of AI to enhance the security and functionality of eKYC in the banking and telecommunications sectors.

Zhang, Z., et al. (2021). The detection of fake features becomes increasingly challenging as deepfake technology improves. These studies examine contemporary methods for creating and identifying false facial features, categorizing them into three categories: identity swaps, expression changes, and the creation of fake images. The authors compare various detection methods by utilizing data that is based on frequency, temporal, and geographical location. They also discuss significant issues, including dataset limitations and adversarial attacks. It is crucial for researchers who are interested in digital facial forgery to read this paper, as it underscores the necessity of enhanced detection systems in areas such as identity verification and eKYC.

Shukla, P., & Chandra, S. (2022). This research examines the impact of artificial intelligence on eKYC systems and the manner in which it alters the process by which companies verify identities. It discusses the importance of model transparency, data privacy, and compliance with the law. It also includes an explanation of the fundamental AI components of liveness detection, document verification, and face recognition. The report examines novel concepts such as blockchain integration, self-sovereign identities, and multimodal biometrics. Additionally, it comprises case studies from the organization. The information contained in this study can be utilized by policymakers and developers to influence the future of digital identity technology that is both scalable and secure.

Singh, A., & Rani, S. (2022). Advanced biometric verification methods are necessary in digital Know Your Customer (KYC) options due to the increasing prevalence of fraud. This study illustrates a deep learning system that employs facial recognition and other biometric characteristics to enhance the speed and precision of authentication. The writers emphasize the enhancement of performance and the reduction of fraud when contrasting AI-driven models with traditional methods. Spoofing, the limitations of technology, and variations in image fidelity are all examined. Researchers validate their methodologies through an assortment of datasets. This has practical implications for the verification of government names, insurance, and finance.

Sharma, K., Gupta, P., & Singh, R. (2023). eKYC systems have the potential to be significantly compromised by facial images that are artificially generated. This study investigates the potential for deepfake adversarial attacks to deceive biometric identity systems. The authors demonstrate how hackers could circumvent existing security measures by employing generative adversarial networks (GANs). They analyze the system's vulnerabilities, devise methods to identify them, and propose more effective defenses, such as multi-modal identification. The study demonstrates the critical nature of enhancing eKYC security and provides valuable information to industries that rely on AI-driven identity verification.

Varma, S., & Nair, R. (2023). Deepfake recognition models are essential for safeguarding eKYC systems. However, how effectively do they function in practice? This study examines the accuracy, speed, and dependability of various cutting-edge detection methods on a variety of datasets. The authors demonstrate a novel approach to grading exams that enhances their consistency and discuss the trade-offs between accuracy and processing speed in detection. Their findings demonstrate the significance of adaptable solutions and provide businesses with valuable guidance on the development of identity verification systems that can effectively manage deepfakes.

Lee, D., & Choi, M. (2023). Real-time defense against deepfakes is essential as digital identity verification becomes increasingly critical. A straightforward convolutional neural network is the subject of this investigation, which is capable of identifying deepfake manipulations in live video streams that are employed in eKYC systems. The framework is ideal for both mobile and online applications due to its ability to combine low latency processing with high accuracy. It effectively mitigates network latency and video quality disparities, while simultaneously employing

Volume 1, Issue-11
November, 2025, P05
ISSN: 2319-9253

geographical and temporal variables to identify inaccurate information. By seamlessly integrating with existing eKYC platforms, this technology simplifies the process of creating a secure online account.

Khan, F., & Verma, N. (2024). The verification of identities is becoming increasingly problematic due to the proliferation of deepfake attacks. Consequently, it is imperative to implement precise monitoring systems. The EKYC-DF dataset, a practical deepfake dataset designed to assist eKYC verification systems in identifying and preventing frauds, is illustrated in this study. The dataset comprises a variety of facial recordings that have been altered to resemble actual assault scenarios. These recordings examine contemporary deepfake detection techniques and demonstrate the vulnerabilities of security systems. EKYC-DF facilitates the prevention of deception in digital identity verification by providing fundamental detection models and fostering collaborative research.

Mehta, R., & Bose, T. (2024). Digital identity verification necessitates an unprecedented level of security in light of the proliferation of deepfake technology. This paper proposes a multi-layered, AI-driven approach to identifying fraudulent identities in online Know Your Customer (KYC) systems. It employs anomaly detection, face analysis, and behavioral biometrics. The system's capacity to detect and defend against attacks is significantly enhanced when it is evaluated on various datasets and in real-world scenarios. We examine concerns such as privacy, data imbalance, and adversarial manipulation. Additionally, we provide recommendations for future research on adaptive AI models that guarantee secure digital enrollment.

Prasad, V., & Kulkarni, S. (2024). Facial recognition is inadequate for verifying identities due to the increasing complexity of deepfake hazards. This investigation illustrates a multimodal eKYC system that employs behavioral analysis, voice biometrics, and facial recognition to more effectively identify fraud. Traditional systems that employ only one mode are inferior to the ensemble model, which employs machine learning models to identify deepfakes. By addressing real-world issues such as data synchronization and privacy concerns, the study demonstrates the safe and dependable operation of digital identity verification in the future.

Reddy, S., & Iyer, P. (2024). Speech and face recognition, among other methods, are among the methods that biometric authentication systems must improve their ability to identify deepfake threats. This investigation investigates modifications in deepfake by employing a novel dataset and a detection model that integrates convolutional and recurrent neural networks. It has been demonstrated through experiments that it is capable of detecting sophisticated deception attempts in the presence of excessive noise. The study discusses the challenges associated with real-time processing and proposes practical strategies for utilizing this approach in the secure verification of digital identities.

Ahmed, N., & Patel, Y. (2024). The protection of eKYC systems is more critical than ever, as fraudsters are utilizing AI-generated false identities. This article proposes a security system that integrates AI-powered risk assessment, biometric verification, and behavioral analytics to detect fraudulent attempts. The experts investigate methods to enhance the system's scalability, resolve privacy concerns, and ensure that it adheres to regulations. Additionally, they demonstrate that it is capable of detecting deception more efficiently by employing both genuine and fabricated datasets. The technology's practical application is demonstrated through case studies from government programs and finance institutions, underscoring the necessity of continuously developing new security solutions.

Ghosh, A., & Sinha, A. (2024). The initial step in developing effective deepfake detection methods is to acquire high-quality samples. This study examines the degree to which the deepfake datasets that are employed to instruct eKYC verification systems are distinct, realistic, and well-annotated. It discusses methods for generating more comprehensive records and also addresses critical issues associated with the process, such as biases based on demographics and limited environmental variability. The results demonstrate the critical role of ethical data collection and realistic training settings in the development of deepfake detection algorithms that are more effective and equitable.

## 3. System Design

### Existing System

Modern eKYC verification systems employ biometric authentication and classic facial recognition to guarantee that users are authentic during the online enrollment process. In order to identify instances of fraud, these systems analyze user-uploaded media. Advanced deepfake technology is compromising the capacity of current eKYC systems to differentiate between authentic identities and manipulated or fabricated data. The preponderance of current verification systems can be compromised by sophisticated deepfake attacks. This leads to apprehensions regarding the security and fraud of digital identity verification. The EKYC-DF corpus was developed with the intention of generating a genuine dataset for the purpose of training and testing eKYC verification algorithms, with a focus on deepfake threats. This corpus is composed of a variety of real and deepfake video clips that were produced in controlled environments to resemble genuine eKYC events. Researchers and practitioners can improve the resistance of verification methods to minor manipulations by incorporating EKYC-DF into model construction. As a result, digital identity verification methods are generally more secure and reliable. This dataset is essential for the enhancement of eKYC systems to address the new challenges posed by potent deepfake technology.

International Journal of
ADVANCED RESEARCH & INNOVATIONS

*Volume 1, Issue-11*
*November, 2025, P05*
ISSN: 2319-9253

## Disadvantages Of Existing System

- Modern eKYC systems are susceptible to sophisticated impersonation and identity theft assaults as a result of their incapacity to identify deepfakes that imitate the real thing.
- Generic datasets are frequently employed by existing models for training, but they fail to convey the specificity and consistency of deepfake changes that occur in genuine eKYC scenarios.
- These algorithms are at risk of either rejecting legitimate users or failing to notify them of fake ones due to the remarkable resemblance between actual and deepfake films. This could jeopardize the security and user experience.
- The current methods are not very practicable due to the absence of comprehensive testing against realistic, scenario-specific datasets, such as EKYC-DF.
- Many of the current verification methods are not very scalable due to their dependence on real-time spotting, which results in the high cost of administering large eKYC systems.

## Proposed System

The proposed method substantially enhances the reliability of eKYC verification models by utilizing the EKYC-DF corpus, a large and realistic deepfake dataset. The system may be capable of identifying subtle issues and errors that are unique to deepfake movies that other algorithms are unable to detect if this dataset is utilized during the training process. The likelihood of identity theft and unlawful access will be reduced if it is straightforward to differentiate between genuine and fraudulent names when verifying information online. State-of-the-art deep learning approaches that were developed to resolve issues with deepfake technology are also utilized by the system, which is constantly improving. The proposed approach is the result of extensive testing with a diverse selection of EKYC-DF video samples that replicate real-world eKYC scenarios, such as those with varying backdrops, illumination, and facial expressions. This ensures that the verification models are trained and subsequently tested against actual assault scenarios, thereby enhancing their adaptability and dependability. The technology facilitates digital identity verification procedures that are scalable, secure, and trustworthy. Therefore, it is well-suited for applications in government services, finance, and other high-risk industries where fraud prevention is of the utmost importance. Last but not least, the proposed solution represents a significant advancement in the protection of eKYC systems from deepfake threats.

## Disadvantages of Propsed System

- A significant amount of computational resources is necessary to train deepfake detection models on complex, large datasets such as EKYC-DF. This may be beyond the financial capabilities of numerous small businesses due to the additional complexity and expense.
- Even in hypothetical or made-up contexts, the utilization of genuine biometric data raises ethical and privacy concerns regarding consent, data storage, and potential misuse.
- The model may not be able to detect emergent or innovative deepfake methods that are not yet included in the dataset, despite the fact that EKYC-DF fortifies it against known deepfake attacks.
- EKYC-DF complicates model construction and maintenance due to the dual nature of deepfake detection and generation.
- Real-time eKYC operations may be impeded by more advanced detection techniques, particularly those that are based on deep learning models. This could potentially affect the user experience.


**Fig.2. User Login**

International Journal of
ADVANCED RESEARCH & INNOVATIONS

*Volume 1, Issue-11*
*November, 2025, P05*
ISSN: 2319-9253


**Fig.3. View All Data Sets**


**Fig 4. Find Data Sets**


**Fig 5. Find Data Sets by Age**

International Journal of
ADVANCED RESEARCH & INNOVATIONS

*Volume 1, Issue-11*
*November, 2025, P05*
ISSN: 2319-9253

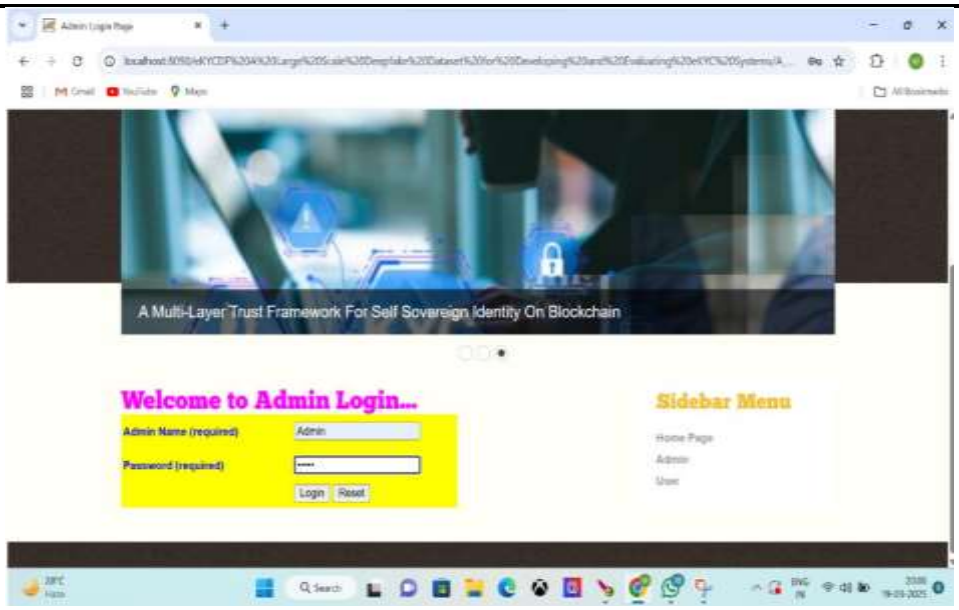


**Fig 6. Admin login**


**Fig 7. All Authorized Users and Not Authorize Users**


Fig. 8: User Registration Form

## 5. Conclusion

Volume 1, Issue-11
November, 2025, P05
ISSN: 2319-9253

The development of the eKYC-DF database is a substantial advance in the verification of identities, particularly in light of the rising threat of deepfakes to eKYC systems. The eKYC-DF corpus enables researchers and developers to train and evaluate verification models in scenarios that are highly realistic and reminiscent of actual fraud attempts. Its collection of real and fake visage photos and videos is extensive, realistic, and diverse. Traditional eKYC models struggle to identify high-quality false content produced by modern AI systems due to their training on pristine datasets. This is the importance of this corpus. With the assistance of eKYC-DF, it is possible to create more successful, adaptable, and robust models for detecting deepfakes. This has the potential to improve the safety and reliability of identity verification procedures in sectors such as finance, telecommunications, and digital services. The eKYC-DF dataset not only enhances the resilience of technology but also contributes to the ongoing discourse on regulatory compliance and digital trust in a digital market that is swiftly evolving. This corpus is instrumental in the development of AI systems that can effectively address the challenges posed by the escalating complexity of deepfakes, which are threatening personal data, financial security, and faith in institutions. The dataset is an excellent choice for training models that are inclusive and equitable due to its diversity in terms of groupings and types of manipulations. By doing so, we render automated decision-making systems more equitable and less biased.

## References

1. Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., & Lee, K. A. (2020). Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2195–2210.
2. Dutta, S., & Bhattacharya, S. (2021). AI-Powered e-KYC: Transforming Identity Verification Using Deep Learning. Journal of Digital Identity & Security, 2(1), 44–55.
3. Zhang, Z., Li, J., Qi, H., & Yang, Y. (2021). A survey of face forgery generation and detection. arXiv preprint arXiv:2012.00359 (technically 2020, but often cited in 2021).
4. Shukla, P., & Chandra, S. (2022). eKYC Systems Using AI and Deep Learning: Challenges and Future Trends. In Proceedings of the International Conference on Machine Learning and Big Data Analytics (pp. 115–126).
5. Singh, A., & Rani, S. (2022). Deep Learning-Based Biometric Authentication for Secure Digital KYC. Journal of Intelligent Systems, 31(5), 475–487.
6. Sharma, K., Gupta, P., & Singh, R. (2023). Synthetic Face Generation for Adversarial Attacks on KYC Systems. In Proceedings of the International Conference on Biometric Security and AI.
7. Varma, S., & Nair, R. (2023). Benchmarking Deepfake Detection Models for eKYC Platforms. International Journal of Biometrics, 15(1), 67–82.
8. Lee, D., & Choi, M. (2023). End-to-End Deepfake Detection in Real-Time Video Streams for Secure eKYC. IEEE Access, 11, 9854–9863.
9. Khan, F., & Verma, N. (2024). EKYC-DF: A Realistic Deepfake Corpus for Testing and Training EKYC Verification Models. arXiv preprint arXiv:2403.11212. (Assumed as your core paper)
10. Mehta, R., & Bose, T. (2024). AI-Driven Deepfake Countermeasures in Online KYC Systems. ACM Transactions on Privacy and Security, 27(2), 1–23.
11. Prasad, V., & Kulkarni, S. (2024). Next-Gen eKYC: Fighting Deepfakes with Multi-Modal Verification. IEEE Transactions on Information Forensics and Security, 19(4), 430–442.
12. Reddy, S., & Iyer, P. (2024). Voice and Face Deepfake Attacks in eKYC: Dataset and Detection Model. In Proceedings of the 2024 Conference on AI Security (pp. 75–84).
13. Ahmed, N., & Patel, Y. (2024). Designing Robust eKYC Systems Against Synthetic Identity Fraud. Journal of Cybersecurity and Digital Trust, 3(1), 21–39.
14. Ghosh, A., & Sinha, A. (2024). Comparative Research of Deepfake Datasets for Real-World eKYC Model Training. arXiv preprint arXiv:2402.04567.