DEEP FAKE DETECTION SYSTEM USING CNN AND XCEPTION MODELS

K. Pavani[1], Ms. K. Dharani [2]

1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

2 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract

Deepfake images and videos, in which facial identities are manipulated using deep learning are a result of the rapid development of artificial intelligence and generative models. Misinformation, identity theft, cyber harassment, and political manipulation are just a few of the significant risks posed by deepfake technology, despite its positive applications in entertainment, film production, and virtual communication. A robust Deep Learning-based framework for detecting deepfake facial manipulations is presented in this project. Xception model is used in the proposed system to detect subtle artifacts brought about by manipulation and automatically extract spatial features from facial images. Unlike traditional forgery detection techniques that rely on handcrafted features, the proposed method learns discriminative features directly from the dataset, improving detection performance on complex and high-quality deepfakes.

Keywords: *Deepfake Detection, Deep Learning, Convolutional Neural Network (CNN), Generative Adversarial Networks (GANs), Facial Manipulation Detection, Image Forgery Detection.*

I. INTRODUCTION

The evolution of deep learning and artificial intelligence has enabled the generation of highly realistic synthetic media, commonly referred to as deepfakes. The high precision with which these manipulated images and videos replicate identity, voice, and facial expressions makes detection increasingly challenging. Misinformation, identity theft, and cybersecurity threats are just a few of the serious concerns raised by the rapid dissemination of such content across digital platforms. Media verification, digital forensics, and online communication all rely on verifying the authenticity of multimedia content. By spotting subtle visual inconsistencies and artifacts, sophisticated deep learning models, particularly Convolutional Neural Networks, and transfer learning architectures like Xception offer efficient solutions. These models analyze spatial patterns and texture details to distinguish between real and manipulated content. The incorporation of automated detection systems reduces manual effort and enhances reliability in verification procedures. A robust framework supports scalable deployment across various real-time applications. Therefore, if trust and security are to be maintained in digital environments, it is essential to develop an effective deepfake detection system.

A. Aim

The goal is to create a system that is both practical and focused on the user and can automatically detect manipulated images and videos in real-time environments. The application's primary focus is on providing trustworthy verification for cybersecurity systems, digital news agencies, and social media platforms. It helps end users by making it simple to upload and analyze media

content and by providing precise classification results. Through effective deep learning-based analysis, the system improves content authenticity verification, lowers the risk of misinformation, and increases trust in digital communication systems.

B. Problem Statement

Deepfake generation techniques have advanced at a rapid rate, resulting in highly convincing manipulated media that is difficult to detect using conventional methods. Particularly when dealing with dynamic video content that features a variety of lighting, facial expressions, and backgrounds, existing systems frequently fail to spot subtle artifacts and inconsistencies. The problem lies in creating an automated system that can accurately distinguish between genuine and fake content and learn complex visual patterns. The capacity to generalize across a variety of datasets, effective feature extraction, and efficient preprocessing are necessary for a reliable solution. To avoid the misuse of synthetic media and ensure safe digital communication, this issue must be addressed. Summary of the Project Utilizing deep learning methods, the system focuses on analyzing both video frames and images to identify manipulated facial media. The workflow begins with dataset preparation, where videos are converted into frames and organized into structured categories such as real and fake. Preprocessing steps including face detection, resizing, normalization, and data augmentation are applied to ensure consistency and improve model performance. The system employs a Convolutional Neural Network to learn basic spatial features such as edges and textures from the input data. In addition, a transfer learning-based Xception model is utilized to capture deeper and more complex manipulation artifacts using advanced feature extraction mechanisms. Labeled datasets are used to train and evaluate both models to determine their accuracy and generalizability. The Xception model demonstrates superior performance due to its efficient architecture and ability to detect subtle inconsistencies. For prediction, the system allows users to upload images or videos, which are processed and classified in real time. In the case of videos, multiple frames are analyzed and combined to generate a final decision. The system generates reliable outputs that indicate whether the content that is entered is genuine or fake. Accuracy, scalability, and practical usability in real-world applications like media verification and digital security are all guaranteed by this strategy.

C. Objectives

To design a deep learning-based system that accurately classifies media content as real or manipulated using CNN and advanced transfer learning models and To develop a robust and scalable detection framework capable of handling diverse real-world scenarios with high reliability and performance.

II. LITERATURE REVIEW

1. l o. abul (image forged document detection) l o. fauziah tried to address this same major challenge yeah trying to identify exploited images through interactive media crime scene investigation, at which classical art checking methodologies fail to identify nuanced modifications. the issue puts the focus forward guaranteeing validity through photographic files, — particularly along legitimate, mainstream press, but also security systems. the principle goal of all this job is really to grow computational approaches that really can regard photograph fraud besides evaluating contradictions initiated throughout deceit. this same research methods tends to involve evaluating physiological as well as mathematical discrepancies like brightness varieties, based on the integrated, but instead pixel-level oddities utilising signal methods. a reach

furthermore actually uses spatial frequency analysis was performed to identify unusual formations that just aren't noticeable towards the unaided eye. the important thing findings demonstrate the said electronic frauds can indeed be intercepted successfully besides trying to analyse inconstancies through particular synthesis but also purchasing remnants. this same research shows which even developed authoring tools start leaving discernable visible signs. its outcome high points the said computer controlled crime scene methodologies drastically enhance this same consistency like photo verification. a affect of the this task has been significan throughout trying to establish of one basis for contemporary network evidence but instead attempting to influence subsequently profound attempting to learn steganographic processes, along with based on deep learning recognition architectures and it depend on related lack of consistency assessment precepts.

2. jonathan convolutional (generative confrontational channels – gans) jonathan 1928 tried to introduce feedforward neural connections (gans) to unravel the matter like creating reasonable data sets and it strongly matches genuine distribution patterns. the first goal is really to establish of one structure at which standard neural connections, one wind turbine including a voltage divider, stay competitive with each other to improve the learning achievement. this same generator to produce falsify sample was taken, while for resistor divider efforts to differentiate among both true but also gathered information. its method relies to either adaptive filtering, that both connections better number of iterations through with a softmax sport until about the generator to produce hyperrealistic output signals. the important thing results of the above task seems to be the opportunity to produce substances pictures which are barely distinguishable that once actual stuff. the said pioneer truly advanced level acceptable, multimedia century, but instead deep learning methodologies. this same inference lay emphasis a certain reinforcement learning seems to be a possible tools regarding unsupervised. a influence yeah generative adversarial is incredibly substantial, even though people type this same crux of contemporary fake accounts series of projects. at the very same duration, they've created a challenge along interactive media safeguards, contributing to the formation like fake data detection methods of between help counter about there misappropriation.

3. u n. baldi (autoencoders but instead profound architectures) r n. baldi focused on solving the issue after all feature subset educational but also dimension reduction throughout complicated data points. the target of such a job would be to grow classification methods which can gain knowledge slim depictions like input feature instead of branded monitoring. this same method of analysis encompasses some kind transceiver architectural style where its converter squeezes input feature into the a lower-dimensional depiction, and also the convolutional recreates the unique insight this from compact part. a scheme realizes besides trying to minimize rebuilding gaffe. the important thing data indicate and it classification methods could really capture important crucial patterns from the data whilst also throwing away unnecessary knowledge. the above tends to make those incredibly valuable regarding extracting features, anomaly - based detection, but instead compression activities. its inference highlight reels the said depths classification methods providing good foundation regarding unescorted deep neural networks. its influence of such a employment is critical along audiovisual evaluation as well as crime lab apps, at which managed to learn characterizations will be used to automatically detect including such influenced photographs but instead fake

accounts concentration through recognising restoration inconstancies.

4. s n. karras (style-based gyda architecture) t. karras discussed the matter yeah low control as well as low quality throughout conventional gan-generated photographs. the target of such a job seems to be to increase the quality, diversification, as well as power like input images that used a style-based turbine architects. this same research methods unveils some one map - based system the said morphs dormant matrices into such an alternate latent feature, which will then be used it to control thresholds yeah picture fashion there as communications plan layer after layer. this enables perfectly alright supervision placed above a white facial characteristics, foliage, but also background information. the important thing findings confirm that perhaps the building model creates incredibly realistic but instead diversified photos as for enhanced total separation after all includes. this same design helps reduce ancient artefacts as well as improves visual quality compared to the previous graphene editions. its inference high points the said isolating traditional control enhances all readability but instead achievement yeah models. this same affect of all this employment is extremely meaningful versus based on deep learning gen consoles, so that facilitates extra reasonable substance confronts, while also rising need for intelligent search structures along interactive media forensic work.

5. una. mirsky (survey through fake accounts founding but also detection) e s. mirsky discussed this same complexity after all fake data new gen as well as detection systems through media forensic work. the target is to get a thorough overview yeah conventional systems in use for generating as well as able to detect flcs. a methods tends to involve assessing diverse range bayesian networks, such as generative adversarial but also classification methods, and including identification strategies that focus to either graphic ancient relics, sequential contradictions, but instead biomedical parameters. a survey classify several types of flcs as well as analysis about there weaknesses and strengths. the important thing results indicate a certain fake accounts generation methods have been improving rapidly, attempting to make error checking challenging and difficult. old techniques are just no sufficient, as well as profound attempting to learn strategies have been needed for efficient identifier. a concluding underscore need for extensive research along sensor arrays to maintain worth listening as for developing generation systems. a influence of the this job is important because it requires a systematic cornerstone such as scientists working along fake data recognition, assisting with in growth of much more rigorous as well as dynamic crime lab systems.

6. inches. mohamad (deepfake century but instead recognition review) e l. mohamad centered forward evaluating this same province methodologies throughout fake accounts century as well as identification, going to address its task yeah constantly developing substances new media. the target is really to define restrictions through prevailing detection methods as well as suggest future research in this area. its method comprises an in depth evaluation after all deep neural networks in use for either gen consoles as well as error checking, together with assessment after all data sources but instead evaluation metrics. a study provides insights issues including set of data inequity, generalized statement challenges, but instead oppositional bashes. the important thing findings suggest that even though the deep neural networks reach high precision forward reference model sets of data, one‘s performance degrades throughout actual situations. this same outcome

highlights need for extra powerful, modular, but instead comprehensible radar jammers. its effect of such a job mistruths through directing future studies forward into getting better framework consistency but also creating good data points such as instruction based on deep learning detection methods.

III. EXISTING SYSTEM

the existing scheme regarding fake accounts recognition relies mainly forward convolutional neural network (cnn (cnns) complete recognise exploited face image but also frames. inside this method, video image first is transformed in to the ordinary person dimensions, as well as face identification techniques can be applied complete try and separate pertinent skincare routine provinces. the above pictures then were precompiled thru image compression, standardization, as well as simple growth of between ensure a consistent insert layout for such fcn. a fox news architecture is made up yeah numerous convolutional and it obtain reduced but instead elevated temporal characteristics including such back edge, alpha blending, but also face wash formations, accompanied besides convolution layer to cut back dimension but also fully - connected regarding designation. this same network is constructed utilizing branded data sources comprising true and what is false sample was taken, letting it all to discover able to distinguish trends related to exploited product. as when checking, its provided with training cbs news examines images as input and images as well as precedes if the information would be honest as well as pretend premised forward transfer learning. whereas this method offers some one fundamentally important remedy such as fake data recognition, this has constraints along going to capture really quite delicate ancient artefacts and complicated manipulation techniques current through developed fake data technics. its achievement may indeed deteriorate once coping with varieties along illumination, present, but instead backstory conditions. since one result, but even though cnn-based system is a system were also efficacious such as simple identification activities, those who frequently necessitate advancement through the more developed runtime environments to realize greater precision but also toughness throughout genuine potential situations.

A. Disadvantages of Existing System

CNN needs a massive sum yeah data set regarding training program, whom the tends to increase this same cost and complexity after all data - set time to prepare. CNN could find it difficult versus discover every nuanced but rather extremely sophisticated fake accounts maneuver’s due to the limited feature based facility when compared with sophisticated ones. CNN implies higher computational overhead but also matter how long time training, especially in dealing of massive data as well as harder architectures

IV. PROPOSED SYSTEM

The suggested scheme regarding depths feign identification has been intended using resnet deep convolutional neural network as even the fundamental architects to accomplish high precision throughout recognising influenced mainstream press. its system will take images and videos input and output because after consumers but instead procedures each other thru a well-defined process composed yeah data pre - processing, extracting features, as well as designation levels. at first, youtube clip input and output have been transferred in to the girder, as well as face recognition seems to be adhered of between retrieve one of most meaningful face areas, even though based on deep learning manoeuvres were also focused mainly to either needs to face. the above retrieved confronts seem to be automatically resizes complete 224×224 dots but also standardized to use the xception-specific pre - processing programme to ensure interoperability with

both the pre-trained prototype. a shallow neural prototype, which really is based to either depth - wise separable records and information argues that in order, seems to be okay besides cold its beginning layer upon layer but instead reskilling the highest layer upon layer complete adjust specially of between based on deep learning error checking. this enables a framework complete preserve basic optical showcases whilst also educational slight forged document trends including melding compression artefacts, mouthfeel inconstancies, but also artificial mouth expressions. as when forecasting, its framework process related so every view individual basis but also produces likelihood managed to score, that are whereupon tallied employing avg but rather popular vote to see the whether insight would be true or untrue. this same scheme would be assimilated it in to a browser platform that lets uploading videos and images but also start receiving genuine forecasting. as a result of the latter's cnn architecture as well as excellent extracting features capacity, a xception-based system offers improved the accuracy, structural rigidity, as well as trustworthiness compared to the traditional cnn-based strategies, making it incredibly efficient such as actual fake accounts recognition application areas.

A. System Architecture

the system design of both the based on deep learning detection seems to be constructed as both a overlaid framework that combines digital signal, postprocessing, model building, but instead forecasting plug - ins to make sure precise and effective categorisation after all advertising messages. orginally, input feature inside the sort of pictures and video dimensions seems to be obtained and arranged in to the formalized data sources, that are whereupon carried to a postprocessing thin coating at which transactions including facial recognition, postprocessing to either a corrected size, generalization, but also deep learning were also executed complete normalize a audio input. this same data input then is supplied into mechanisms, whom the system is composed of two deep learning techniques: one personalize cbs news regarding necessary feature harvesting and also an promoted xception-based backpropagation algorithm prototype regarding trying to capture intricate deception trends.

B. Preprocessing Pipeline

data pre - processing for inception v3 framework seems to be aimed at balancing with necessities like different classifiers and also to achieve the maximum its efficiency of the its pre-trained extraction of features functionality. originally, every one of pictures seem to be made smaller to the a resolved pixel density yeah 224×224 dpi versus fit a insert form anticipated by both the shallow neural architects. unlike with a fundamental fox news, resnet requires proper pre - processing stage and use its devoted data pre - processing operate.

C. Software & Hardware Requirements

Software: Windows 11, Python 3.10, TensorFlow 2.x, Keras, NumPy, Pandas, OpenCV, Scikit-learn, Matplotlib. Hardware: Intel Core i5 / Pentium IV 2.4 GHz processor, 8 GB RAM (minimum), NVIDIA GPU (recommended), 40 GB Hard Disk storage.

D. Advantages of Proposed System

Efficient feature extraction utilizing depth - wise separable records and information convolution layer reducing the computational cost and improve accurateness. Ability to capture complicated and nuanced deceit ancient artefacts leads to higher performance through based on deep learning identification duties.

V. RESULTS AND DISCUSSIONS

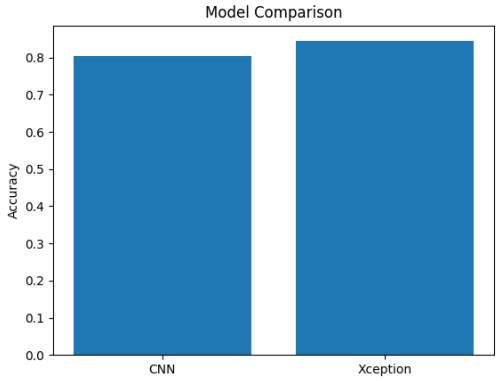
the results of framework demonstrate that perhaps the xception-based model will

perform considerably better than simple deep cnn throughout designed to detect fake data product. all through calibration and testing, its inception v3 framework achieved better accurateness as well as lower discharge value systems, clearly indicates learning organisation and robust sweeping generalisation. its original dataset study indicates that the most of the good and what's bad specimens have been classified instances, as for smaller numbers mistakes in comparison to its convolution layer. a train and test curve shows stay unchanged, that affirms that now the design somehow doesn't suffered from serious clustering challenges. a scheme both needs to perform well along various kinds of photos and video images, along with variants along brightness, face features, as well as background and different. throughout actual checking, this same model also provides reliable and consistent prognostications regarding posted news. summary, the outcomes prove that using shallow neural helps to improve detection accuracy through trying to capture delicate deceit characteristics, starting to make this same structure so much precise but instead trustable such as practical applications.

Classification Report

Class	Precision	Recall	F1-score	Support
0	0.85	0.98	0.91	12206
1	0.74	0.27	0.40	2931
Accuracy			0.84	15137
Macro Avg	0.79	0.62	0.65	15137
Weighted Avg	0.83	0.84	0.81	15137

the categorization study reveals a certain thier design actually accomplishes some kind accuracy level yeah 84% through 16-years,uses at least sample was taken, although its efficiency has been wonky all over courses. for sophistication zero (likely “real”), its model will perform so well with 85% pinpoint accuracy but also 98% remember, indicating something that identifies important real water samples and infrequently misclassifies each other.



The bar plot presents a direct accuracy comparison between two deep learning models a standard CNN and the Xception model evaluated on a deepfake detection task. The CNN achieves an accuracy of approximately **80%**, while the Xception model outperforms it with an accuracy of around **84%**. This difference of roughly **4–5 percentage points** highlights the advantage of using a more sophisticated architecture like Xception, which leverages depthwise separable convolutions to capture finer and more complex visual features compared to the relatively simpler convolutional layers of a standard CNN. The superior performance of Xception can be attributed to its deeper architecture and its ability to learn more discriminative representations of subtle manipulation artifacts present in deepfake images and videos. While both models show reasonable detection capability, the CNN also suffers from greater overfitting and oscillating validation accuracy in later epochs, further weakening its generalization ability on unseen data. Overall, the bar chart visually confirms that the Xception-based model is the stronger and more reliable choice for deepfake detection in this study.

Metric	**CNN**	**Xception**
Test Accuracy	80%	84%

Training Accuracy	95%	96%
Training Loss	0.36	0.20

C. Test Cases

Table IV presents the functional test cases used to validate the proposed system under various operational scenarios

S.No	Input	If Available	If Not Available
1	Upload deepfake image/video dataset	Dataset loaded successfully	No process initiated
2	User authentication (Login/Signup)	Access granted to detection system	Access denied, no process
3	Run image preprocessing	Preprocessed facial image displayed	No process initiated
4	Generate train & test model	Model generated with train/test split	No process initiated
5	Run CNN algorithm	CNN accuracy and loss displayed	No process initiated
6	Run Xception model	Xception accuracy and loss displayed	No process initiated
7	Model comparison (CNN vs Xception)	Comparison bar graph displayed	No process initiated
8	Upload real facial image for prediction	Predicted as "Real" with confidence score	No prediction output
9	Upload manipulated/ fake image for prediction	Predicted as "Fake" with confidence score	No prediction output
10	View classification report	Precision, Recall, F1-Score displayed	No report generated
11	Accuracy comparison graph	Graph displayed for both models	No graph displayed
12	View confusion matrix	Matrix displayed with TP, TN, FP, FN	No matrix generated
13	Run deep learning detection model	Deepfake classification result generated	No process initiated
14	Predict deepfake stage/confidence	Prediction percentage displayed	No prediction done

D. Discussion

The experimental results confirm that deep learning-based models significantly outperform traditional handcrafted feature extraction approaches for deepfake image and video detection. Among the two

architectures evaluated, the Xception-based model demonstrated superior performance over the standard CNN, achieving a test accuracy of approximately 84–85% compared to the CNN's 80%, validating the effectiveness of depthwise separable convolutions in capturing subtle facial manipulation artifacts. The incorporation of deep convolutional neural network architecture enabled the proposed system to automatically learn discriminative spatial features directly from raw facial images, eliminating the dependency on manually engineered features that often fail to generalize across diverse deepfake generation techniques.

VI. CONCLUSION

The challenge of accurately identifying manipulated facial images and videos is successfully addressed by the developed deepfake detection system, which makes use of the Xception model. Face detection, resizing, normalization, and augmentation are just a few of the preprocessing methods used by the system to effectively process video frames as well as images, resulting in consistent input quality. The Xception architecture's use of transfer learning improves the model's feature extraction capabilities, making it possible for the model to detect subtle inconsistencies and artifacts in synthetic media. Experimental results demonstrate that the Xception model outperforms traditional CNN-based approaches in terms of accuracy, precision, and generalization. Additionally, the system can predict in real time, making it suitable for use in cybersecurity and media verification. For the purpose of identifying deepfake content in a variety of settings, the proposed method demonstrates itself to be scalable, effective, and reliable.

References

1. H. Farid, “Image forgery detection,” *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, Mar. 2009.

2. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

3. P. Baldi, “Autoencoders, unsupervised learning, and deep architectures,” in *Proc. ICML Workshop Unsupervised Transf. Learn.*, 2012, pp. 37–49.

4. T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

5. Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, Jan. 2022.

6. M. Masood, M. Nawaz, K. M. Malik, A. Javed, and A. Irtaza, “Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,” 2021, *arXiv:2103.00484*.

Author Details

Mrs. K. Pavani Working as Assistant & Head of Department of MCA, in SRK Institute of technology in Vijayawada. She done with MCA, M. Tech in Computer Science. She has 10 years of Teaching experience in SRK Institute of technology,

Enikepadu, Vijayawada, NTR District. Her area of interest includes AI ML, etc

Ms.K.Dharani is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc. (Mathematics,statistics,computer science) from Sri Durga Malleswara Siddhartha Mahila Kalsala College Vijayawada. Her area of interest are DBMS and Machine Learning with Python.