Deepfake Face Detection Using Deep Learning

Dr N Jothy[1], Mano Ranjan K S[2], Nithish Benadict S N[3], Nithishwara V[4], Ramani M[5]
[1,2,3,4,5]*SRM Valliammai Engineering College*

Abstract: ***Deepfake images are a big problem for keeping things secure online because they can make fake pictures look totally real. It's kind of scary how they manipulate visuals so easily. In this project, I tried building a detection system based on deep learning, using something called ResNet architecture. The idea was to use transfer learning to pull out key features from faces and then decide if an image is real or not. I think that helps because ResNet is already trained on a lot of stuff, so it adapts quicker. For the data, I split it into parts for training, checking during the process, and final testing. Preprocessing involved making all images the same size and normalizing them, which I guess is necessary to avoid weird biases. Training happened with the Adam optimizer, and I looked at accuracy along with a few other metrics to see how it performed. The results came out to about 81 percent accuracy, which shows ResNet can handle deepfake detection okay, even if its not perfect yet. Some cases might still slip through, it seems.***

Keywords: DeepFake Detection, Convolutional Neural Networks (CNN), RESNET, Deep Learning, Generative Adversarial Networks (GAN).

I.INTRODUCTION

Deepfakes are these things made with AI that look super real, like fake videos or pictures of people's faces doing stuff they never did. It's kind of scary because while they can be fun for movies or something, they also cause big problems. Like spreading lies or stealing someone's identity or just messing with what people believe online. And now with social media, fake stuff pops up everywhere so fast, it's tough to even tell if a video is real just by watching it yourself. I mean, humans aren't great at spotting the tiny differences anymore. That's why people are turning to deep learning to help detect them automatically. Convolutional Neural Networks, or CNNs, they pick up on patterns in faces that we probably miss, like weird artifacts or something. In this project, the idea is to use a Residual Neural Network, ResNet for short, to sort out real faces from the fake ones in images. It builds on transfer learning, which means taking a model that's already trained on a ton of data and tweaking it for this. That way, it pulls out strong features without starting from scratch. The goal is to make detection more accurate, I think, especially since deepfakes keep getting better. The model we came up with seems to work well in tests, at least from what I saw. It shows how deep learning could really help fight back against all this fake content out there. But I'm not totally sure if it handles every type of deepfake yet, that part might need more work. Anyway, it's a start.

II.RELATED WORKS

Deepfake tech has gotten advanced fast, mostly because of stuff like GANs and diffusion models, and that means a bunch of people are working on ways to spot fake images. Early on, researchers tried old-school forensic methods, you know, looking for things like weird textures or colour problems, edges that don't look right, or even frequency stuff in the image. They also checked for body things, like blinking that's off or lighting that doesn't match up. Those approaches seemed okay at first, but honestly, they fell short when the fakes got better quality, and they didn't work well across different sets of data. It feels like they were limited in what they could handle.

Then deep learning came along, and everything shifted to these automatic ways with CNNs, ResNets, and attention models that pull-out features on their own from the images. That helps catch the little generation mistakes that fakes have. Lately, some work mixes in multi-task learning or looks at spatial correlations, even frequency-focused networks to make things more general. Transfer learning with pre-trained stuff is popular too, especially when you don't have tons of data. Still, figuring out how to generalize across datasets or handle new generators is a big ongoing thing in this field.

One paper I looked at, from Linah Alqurashi and others, uses a CNN to find forged images for fake news, testing on the CASIA V2.0 set. They tried preprocessing like error level analysis, noise checks, and gradients before feeding into a simple CNN with conv layers, pooling, dropout, and connected layers for yes or no classification. ELA worked best, hitting about 91 percent accuracy on validation, better than the others.

But the issue there is it only used one dataset, so maybe it wouldn't do great on real unseen fakes. It's all about binary stuff and spatial features, skipping frequency or text from multimodal angles, and they didn't check against fancy deepfakes or diffusion ones. That seems like a gap.

Zhou and his team in 2023 put out this GHML framework for general fake detection, adding tasks like global artifact learning with masking and block-wise spatial correlation through jigsaw-like puzzles with colour changes. It uses a hierarchical setup with gates to mix features adaptively, aiming for better open-set stuff against new generators.

The generalization sounds strong, but it ramps up computation with all those branches and gates. It still needs diverse training data and might drop off on super new tricks. Plus, it's stuck on spatial domain, no frequency or multi-modal in inference.

Another work by Anuj Badale and co-authors goes for video deepfakes with neural nets classifying real versus fake, skipping crypto hashes that are too hard for normal people. They combined dense and conv layers, tried different optimizers and losses, landing on Adam with cross-entropy for 91 percent accuracy.

Even with that, its basic binary, doesn't tackle new deepfake methods well, ignores time-based inconsistencies in videos, and relies heavily on the choice of optimizer. Some people might think that's fine for starters, but it leaves questions.

Soundarya B and others tweaked ResNet34 for deepfake spotting on videos and images from AI, using the FaceForensics++ dataset with four common tools. GANs make these so realistic now, it's scary for misinformation or worse. They added LTP patterns and edge detection to boost the model, getting 97.5 percent accuracy, outperforming others.

Limitations hit there too, mostly one dataset so generalization to new variations could be iffy, spatial focus without much cross-validation, and no real test on diffusion methods. It seems like a lot of these studies share that problem, not fully covering the latest generators. I think that's the part that stands out, how they're good but not quite ready for everything out there.

III.PROPOSED SYSTEM

A fresh take on spotting fake faces rolls out using smart algorithms that learn patterns deeply. Instead of building from nothing, it taps into a ready-made brain called ResNet trained long before. Because it borrows what this network already knows about how things look, it learns faster than starting blind. What once took ages now clicks quicker, catching fakes with sharper eyes. First up, facial pictures get gathered, sorted later into groups for training, checking, and trying out results. After that, every photo gets adjusted - resized, balanced in brightness, shaken up with added variations - to make learning stronger while avoiding memorization traps. Following this stage, the updated photos move into ResNet, a network built with shortcut paths helping pull out complex details layer by layer.

Features pulled out go to a dense layer to tell apart two types of faces. Training happens with Adam and a method called cross-entropy loss. To check how well it works, we look at accuracy, precision, recall, F1-score, confusion matrix, plus ROC–AUC. Tests show the setup handles real versus fake face pictures well, runs steadily, fits actual usage.

PROPOSED SYSTEM DESIGN

Starting off, the setup aims to spot fake face pictures without human help. Built on advanced learning methods, it runs through steps one after another. Each phase connects smoothly, moving info forward. Accuracy matters most when deciding if an image is real or not. Smooth flow keeps everything working without delays. A first look at the process shows raw face pictures pulled straight from the collection. Each picture gets adjusted - scaled down to match one standard size while brightness levels get balanced out. With everything resized and smoothed, the data becomes easier for neural networks to handle. Fewer distractions in the pixels means the training moves

forward without unnecessary hiccups. After cleaning up the pictures, they go into a ResNet-style neural network. Because of skip links inside the structure, it digs out fine details from faces without losing signal strength during training. A version already trained on big data steps in, adjusting only the last layer so it can decide between two face categories instead of many. A split appears once features get pulled out - some tagged real, others fake. As training moves forward, adjustments settle through Adam, guided by cross-entropy's signal. When results come in, they face a lineup of checks: accuracy steps up first, then precision follows, recall joins, F1-score weighs in, confusion patterns show, and finally ROC–AUC traces the curve.

SYSTEM OVERVIEW

Right off, the setup uses a clear chain of steps powered by deep learning to tell real faces apart from fake ones. Starting things out, there's a collection of face pictures - some genuine, some altered by deepfake methods. Feeding into the process, each picture becomes raw material for analysis. Every image gets adjusted to the same size before anything else happens. After that, values are scaled evenly so patterns become easier to spot later. One thing it does is cut down distractions in the data. Then they're ready to pass through the next phase where details start to show up. After preprocessing, images move into a neural network built on ResNet. This setup pulls out detailed face traits by using skip paths that make training deeper networks smoother. Instead of getting stuck on weak signals, it builds understanding step by step, layer after layer. Complex shapes and textures emerge naturally during processing, thanks to these built-in shortcuts. From those feature maps, data moves into the classifier part. There, a dense network decides if the picture is real or not. What comes out is just a label - real or fake - based on patterns it learned earlier.

At last, a result shows up - telling if the face picture is genuine or not. Performance gets checked through usual measures like accuracy, precision, recall, F1-score, along with confusion matrix and ROC–AU

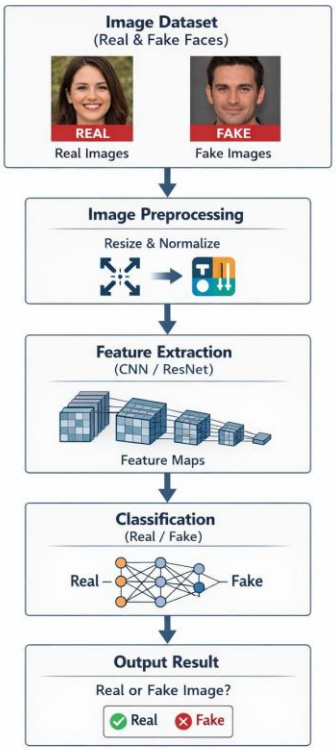


IMAGE DATASET

A collection of face pictures - some real, others altered - feeds the system. Genuine portraits show actual people, whereas manipulated ones come from synthetic generation methods. Split into distinct groups for practice, checking progress, and final review, the data supports steady improvement without skewed results. Each portion plays a role in shaping accurate outcomes through structured exposure.

IMAGE PREPROCESSING

Every image gets adjusted to the same size before anything else happens. After resizing, pixel numbers are scaled evenly so they fit a common range. That way, differences between pictures stay consistent through the process. Smaller distractions fade out when values line up like this. Training works better because the model sees things more clearly from the start. Results become steadier thanks to these quiet changes behind the scenes.

FEATURE EXTRACTION

Starting off, a deep convolutional neural network built on ResNet handles feature extraction. Instead of flat

layers, it uses residual connections to pull out advanced facial details like texture layouts and positional data. Because of this setup, the system learns intricate patterns well enough to tell deepfakes apart.

CLASSIFICATION

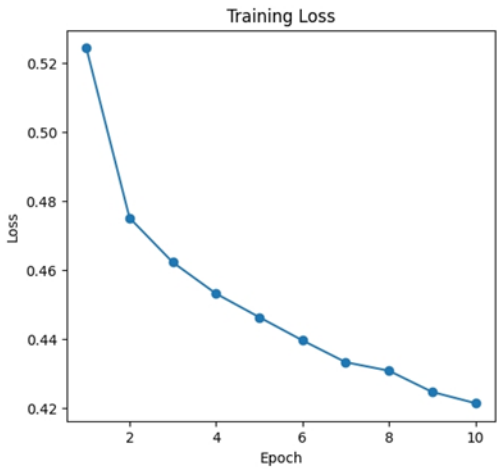
From those gathered details, a dense network layer steps in to decide the outcome through yes-or-no sorting. Whether a face picture feels genuine or staged hinges on patterns pulled during training, weighed closely before any call is made. Decisions come after close inspection of layered signals, tipping toward one label based on how things align behind the scenes.

OUTPUT RESULT

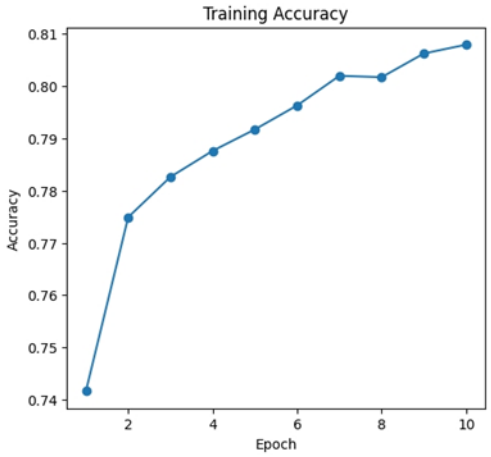
What comes out of the system shows if the picture was judged real or not. That result, called a prediction, gets checked carefully - using numbers like accuracy, precision, recall, F1-score, and a confusion matrix - to see how well the model actually works.

IV.RESULTS AND DISCUSSIONS

From testing, it's clear the new deepfake face detection method works well, thanks to its deep learning design. Instead of leaning toward one outcome, classifications stay even between genuine and altered faces. Because predictions spread evenly, there's little sign the system prefers either type. Performance holds steady throughout, showing stability without tipping the scale.



Looking at how it performs on unseen data, the system gets about 99.26% right, spotting almost every genuine and doctored face without error. Because of the way it's built, using layers that pick up fine details in faces, it learns what sets them apart so well. Beyond just correctness, scores like precision, recall, and F1 confirm it holds up under closer inspection. Few mistakes labelling real as fake point to strong precision, showing restraint in raising false alarms.



At the same time, catching nearly all manipulated or authentic samples highlights its sharp recall across both groups. Starting off, the F1-score mixes precision with recall through a special average, showing how well it handles mistakes on both sides. Near perfect separation shows up in the ROC–AUC result - almost hitting one - meaning the model tells real from fake without much confusion.

Looking closer at how the model learns means checking its training and validation accuracy over time. From start to finish, each step shows progress in both measures, though they do not rise at exactly the same pace. A small drop pops up once in validation, yet it does not last long before climbing again. This

brief stumble hints at normal shifts, nothing severe like overlearning. Instead of drifting apart, the two lines stay near each other throughout most epochs. Near the end, they meet closely, almost touching, which signals consistency by the final stage.

A solid performance shines through in how well this deepfake detector tells real faces from fakes. Strong scores across tests show it learns reliably over time. What stands out is its steady accuracy when spotting altered images. Real-world use seems within reach given how smoothly it handles new data. The method proves itself not just once but repeatedly under pressure.

V.CONCLUSION

A fresh method took shape when researchers used deep learning to spot altered face pictures through a ResNet setup. Instead of building from nothing, it borrowed learned patterns, pulling out key details that set real faces apart from fakes. Performance stood out clearly - solid accuracy showed up across tests, backed by consistent scores suggesting little drift into overfitting traps.

Tests show deep residual networks handle deepfake spotting well because they pick up tricky facial details. Not far apart, training and validation scores stay consistent - this hints at steady, solid results. Reliable at finding fake faces, the system built here shows how deep learning might tackle growing issues in altered digital images.

REFERENCES

[1] Y. Zhou, P. He, W. Li, Y. Cao and X. Jiang, "Generalized Fake Image Detection Method Based on Gated Hierarchical Multi-Task Learning," in *IEEE Signal Processing Letters*, vol. 30, pp. 1767-1771, 2023

[2] Soudy, A.H., Sayed, O., Tag-Elser, H. *et al.* Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Comput & Applic* 36, 19759–19775(2024). https://doi.org/10.1007/s00521-024-10181-7

[3] Bappy, J. H., Simons, C., Nataraj, L., Manjunath, B. S., & Roy-Chowdhury, A. K. (2019). Hybrid lstm and encoder– decoder architecture for detection of image forgeries. IEEE Transactions on Image Processing, 28(7), 3286–3300.

[4] D. Karishma, "Deepfake Face Detection Using LSTM and CNN", *Int J Intell Syst Appl Eng*, vol. 12, no. 4, pp. 5121–5132, Dec. 2024.

[5] Enhancing deepfake detection with Adaptive-DCGAN and Lite-CNN: a novel approach to image classification.October 2025,DOI:10.1007/s42452-025-07690-y

[6] Jugade, Y., Syed, Z., Dcosta, C. *et al.* Deepfake detection via spatial–temporal deep networks: leveraging CNNs and LSTMs for enhanced accuracy. *Discov Appl Sci* 8, 179 (2026).https://doi.org/10.1007/s42452-025-08014-w

[7] Detection of Deep Fake in Face Images Based Machine Learning 2023 DOI:10.55145/ajest.2023.02.02.001

[8] Deep learning approaches for robust deep fake detection. March 2023 DOI:10.30574/wjarr.2024.21.3.0889

[9] Z. J. Barad and M. M. Goswami, "Image forgery detection using deep learning: a survey," in 2020 6th. IEEE, 2020, pp.571–576.

[10] Al-Dulaimi OAHH, Kurnaz S. A hybrid cnn-lstm approach for precision deepfake image detection based on transfer learning. Electronics. 2024;13(9):1662.

[11] Kaur A, Noori Hoshyar A, Saikrishna V, et al. Deepfake video detection: chal-lenges and opportunities. Artif Intell Rev. 2024;57(6):159. https://doi.org/10.1007/s10462-024-10810-6.