International Journal of Engineering Research and Science & Technology

CIFAKE: Explainable Image Classification and Identification of AI-Generated Synthetic Images

K.pavani[1], Y.Naga Malleswarao[2],N.Teja Sree[3]

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

#2 Assistant Professor in the Department of MCA,SRK Institute of Technology, Vijayawada

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract: Humans now find it more difficult to distinguish between artificial intelligence (AI)-generated and real-world pictures due to recent developments in synthetic data synthesis. By offering a technique to improve the identification of AI-generated pictures using computer vision algorithms, this study meets the crucial demand for data authenticity and trustworthiness. The study creates a set of pictures for comparison with actual photos using a synthetic dataset that was constructed using latent diffusion and patterned after the CIFAR-10 dataset. The classification task is presented as a binary problem that entails differentiating between images created by artificial intelligence and those that are real. For this, a CNN2D model—which has 32 neurons and produces the best results—is used. This structure consists of two convolutional layers, two MaxPooling2D layers, and two dense layers. Furthermore, the Grad-CAM (Gradient Class Activation Mapping) method identifies elements that help CNN differentiate between authentic and fraudulent pictures. In contrast to the suggested CNN2D's 94.98% accuracy, a modified version that was optimized without the addition of extra layers like dropout or global average pooling obtained 95.94%. Accuracy, precision, recall, F1-score, and a confusion matrix are used to assess the performance.

Index terms *- ai-generated images, synthetic image detection, cnn2d, grad-cam, image classification, latent diffusion, explainable ai, cifar-10, deep learning, binary classification, fake image detection, computer vision, performance evaluation, neural networks*

1. INTRODUCTION

The area of artificial intelligence (AI)-generated synthetic photos has advanced quickly in recent years, and identifying AI-generated images is increasingly essential to guaranteeing the accuracy of image data. We now confront the potential of AI models producing high-fidelity and photorealistic photos in a matter of seconds, in contrast to the recent past when generative technologies frequently created images with serious visual flaws that were visible to the human eye. Because of their growing quality, AI-generated images may now compete with humans and win art contests. Generative models based on the

Latent Diffusion Model (LDM) have become a potent technique for creating artificial pictures. Our conceptions of truth, authenticity, and originality have been drastically changed by these recent advancements. This has led to the availability of consumer-level technology that is easily abused for fraud and privacy violations. At the cutting edge of modern technology, these philosophical and sociological ramifications pose important queries regarding the nature of reality and reliability. Recent technical developments have made it possible to create photographs of such high quality that people are unable to distinguish between a real-life snapshot and one that is only a hallucination of the weights and biases of an artificial neural network.

2. LITERATURE SURVEY

2.1 High-Resolution Image Synthesis with Latent Diffusion Models:

ABSTRACT: To produce state-of-the-art synthesis outcomes on picture data and beyond, diffusion models (DMs) split the image generation process into a sequential application of denoising autoencoders. Furthermore, a directing mechanism to regulate the picture-generating process without retraining is made possible by its formulation. However, because these models sometimes work directly in pixel space, inference is costly owing to sequential assessments, and tuning strong DMs occasionally takes hundreds of GPU days. We use them in the latent space of strong pretrained autoencoders to enable DM training on constrained processing resources while maintaining their flexibility and quality. In contrast to earlier research, training diffusion models on such a representation makes it possible to achieve a near-optimal balance between reducing complexity and maintaining detail for the first time, significantly enhancing visual fidelity. We transform diffusion models into reliable and adaptable generators for broad conditioning inputs, such as text or bounding boxes, by including cross-attention layers into the model design. This enables convolutional high-resolution synthesis. With significantly less computing power than pixel-based DMs, our latent diffusion models (LDMs) achieve highly competitive performance on a variety of tasks, such as unconditional image generation, text-to-image synthesis, and super-resolution, as well as new state-of-the-art scores for image inpainting and class-conditional image synthesis.

2.2 Predicting image credibility in fake news over social media using multi-modal approach:

ABSTRACT: The majority of fake photos are disseminated on social media. Fakes are images that have been manipulated using software or other techniques to modify their meaning. False visuals are the source of misinformation and divisiveness spread through Twitter. In order to fight false photographs on social media, identity verification is crucial. Phony visuals are frequently linked to text. Thus, textual and visual feature learning is incorporated into a multi-modal framework. There aren't many multi-modal frameworks, and those that do need additional study to identify the relationships across modalities. An effective multi-modal method for identifying phony photographs on microblogging platforms is presented in this study. No more subcomponents are required. The proposed approach makes use of an explicit convolution neural network model and a sentence transformer for text processing. EfficientNetB0 for images. Fake photographs are predicted by combining deep layers of linguistic and

visual feature embeddings. When the model is tested on Weibo and MediaEval (Twitter), the accuracy forecasts are 81.2% and 85.3%, respectively. The most recent Twitter dataset, which includes images of significant events in India in 2020, is used to assess the model. According to experimental data, the suggested model performs better than state-of-the-art multi-modal frameworks.

2.3 On the use of Benford's law to detect GAN-generated images:

ABSTRACT: Anyone may create realistic synthetic pictures thanks to Generative Adversarial Network (GAN) designs. Malevolent GAN-generated pictures might lead to false news, opinion formation, and other social and political repercussions. Therefore, in order to prevent fraudulent pictures from spreading widely, detection-capable technologies are required. This study investigates if Benford's rule can differentiate between actual and GAN-generated images. The most important digit distribution of quantized DCT coefficients is described by Benford's law. We demonstrate that a compact feature vector may be extracted from an image by extending and generalizing this feature. To identify GAN-generated images, this feature vector may be passed into a basic classifier.

2.4 Zero-Shot Text-to-Image Generation:

ABSTRACT: Finding improved modeling assumptions for training on a certain dataset has always been the main goal of text-to-image generation. These presumptions might contain intricate structures, extra losses, or unintentional data from training, such segmentation masks or object component labels. For this, we provide a straightforward method based on a transformer that represents the text and picture tokens as a single stream of data in an autoregressive manner. When tested in a zero-shot fashion, our method is competitive with earlier domain-specific models when there is enough data and scalability.

2.5 Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding:

ABSTRACT: We introduce Imagen, a text-to-image diffusion model with deep language comprehension and unparalleled photorealism. Imagen leverages the strength of diffusion models to create high-fidelity visuals with the power of massive transformer language models for text comprehension. Our main conclusion is that, when pretrained on text-only corpora, generic large language models (like T5) are surprisingly good at encoding text for image synthesis: expanding the language model in Imagen significantly improves sample fidelity and image-text alignment more than expanding the image diffusion model. Without ever training on the COCO dataset, Imagen obtains a new state-of-the-art FID score of 7.27, and human raters judged Imagen samples to be on par with the COCO data itself in terms of image-text alignment. We provide DrawBench, a thorough and exacting benchmark for text-to-image models, to evaluate them more thoroughly. Using DrawBench, we contrast Imagen with more current techniques such as VQ-GAN+CLIP, Latent Diffusion Models, and DALL-E 2. In side-by-side comparisons, we discover that when it comes to sample quality and image-text alignment, human raters favor Imagen over other models.

3. METHODOLOGY

i) Proposed Work:

The enlarged proposed method enhances the previous CNN2D model to better classify real and AI-generated images by adding new deep learning layers including Global Average Pooling and Dropout. These enhancements focus on improving the model's ability to generalize and strengthening feature extraction. By lowering dimensionality while preserving crucial spatial information, the Global Average Pooling layer makes sure the model concentrates only on the most significant features. Additionally, by randomly deactivating neurons during training, the Dropout layer prevents overfitting by encouraging the model to learn more robust and diversified representations.

Grad-CAM increases confidence in AI assessments and makes the system more understandable and user-friendly by graphically highlighting important image regions that influence the model's predictions. Additionally, a Flask-based web application is created that allows users to upload images and receive real-time heatmap-based classification results. This comprehensive and understandable methodology not only ensures that the model functions with more accuracy and reliability, but it also makes it possible for real-world deployment by allowing users to understand the rationale behind each decision.

ii) System Architecture:

The system architecture is designed as a streamlined pipeline that seamlessly integrates picture processing, classification, and explainability. The first stage is input picture collection, in which users submit real or artificial intelligence-generated photos using a Flask-based web interface. These preprocessed images are fed into the Modified CNN2D model, which consists of two convolutional layers, two MaxPooling layers, Global Average Pooling, and Dropout layers. The model extracts deep properties from an image and then does binary classification to determine if it is real or fabricated.

Following classification, the Grad-CAM (Gradient-weighted Class Activation Mapping) technique creates heatmaps that highlight the specific regions of the image that influenced the model's decision. The classification result is shown in the user interface with these heatmaps, which offer visual explanations to enhance confidence and transparency. The system also includes performance assessment modules that compute accuracy, precision, recall, F1-score, and confusion matrix values in order to test and enhance the model. This modular and interpretable architecture ensures high accuracy and end-user usability.

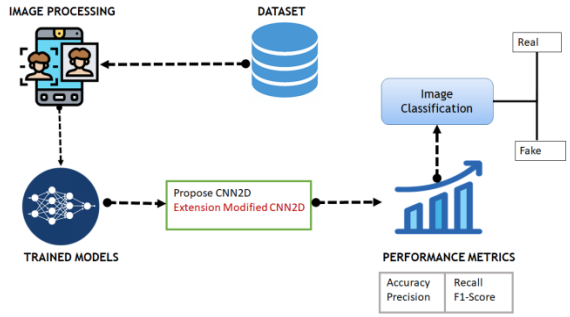

Fig.1. Proposed Architecture

iii) MODULES:

- **Data loading:** We will import the dataset using this module.
- **Image Processing:** In order to improve the quality of photos and extract useful information,

image processing entails modifying and evaluating them. Images must be prepared for efficient categorisation and analysis using methods including resizing, shuffling, normalising, and dividing datasets into training and testing groups.

- **Model generation:** Model construction: CNN2D proposal and extension CNN2D was altered. Each algorithm's performance assessment metrics are computed.
- **Admin login:** This module allows the administrator to log in.
- **Drug Side Effect Prediction:** This module allows users to upload test data.
- **Prediction:** The final prediction was shown.

iv) ALGORITHMS:

CNN2D: CNN2D is a convolutional neural network that uses many convolutional and pooling layers to automatically extract information from pictures for image classification applications. CNN2D is used in this study to differentiate between artificial intelligence (AI)-generated and non-AI-generated pictures. The model finds visual patterns that distinguish the two groups by examining the set of photos. The basic model is this algorithm that consistently detects and classifies the photos with high accuracy.

Modified CNN2D: By adding more layers like Global Average Pooling and Dropout layers, the Modified CNN2D algorithm improves the original CNN2D architecture. By concentrating on the most important characteristics and eliminating the less important ones, this improvement seeks to enhance feature extraction and lessen overfitting. The project's enhanced model provides more resilience and classification accuracy by examining the same dataset of actual and artificial intelligence-generated photos. The adjustments guarantee that the model is well tuned for detecting minute variations among picture classes.

4. EXPERIMENTAL RESULTS

Real pictures from the CIFAR-10 dataset and artificial images created with latent diffusion models were used to test the suggested method. To differentiate between genuine and artificial intelligence-generated pictures, the CNN2D and Modified CNN2D architectures were trained and evaluated on this binary classification task. With an accuracy of 95.94% as opposed to the original CNN2D's 94.98%, the Modified CNN2D model fared better than the basic version. The robustness of the model was confirmed by key performance measures including accuracy, recall, and F1-score, which also showed great efficacy in properly identifying both classes.

Grad-CAM was utilized to generate heatmaps that visually indicated the areas in each image that most influenced the classification decision in order to verify the interpretability of the model. This made it easier to confirm that the model was concentrating on pertinent characteristics rather than background noise. The model's accuracy was further supported by the confusion matrix, which revealed a low rate of misclassification. Overall, the experimental findings demonstrated that the combination of explainability tools with feature-optimizing layers produced a highly accurate, comprehensible, and effective

system for identifying artificial intelligence-generated synthetic pictures.

Accuracy: A test's accuracy is determined by its capacity to distinguish between healthy and ill cases. To gauge the accuracy of the test, find the percentage of examined instances that had true positives and true negatives. According to the computations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Precision is the number of affirmative cases or the classification's accuracy rate. The following formula is applied to assess accuracy:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: A model's ability to recognise every instance of a pertinent machine learning class is measured by its recall. The ratio of accurately predicted positive observations to the total number of positives indicates how well a model can identify class instances.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: Mean Average Precision is one ranking quality metric (MAP). It considers the number of relevant recommendations and their position on the list. MAP at K is calculated as the arithmetic mean of the Average Precision (AP) at K for each user or query.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: An accurate machine learning model is indicated by a high F1 score. combining precision and recall to increase model correctness. The accuracy statistic quantifies the frequency with which a model correctly predicts a dataset.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

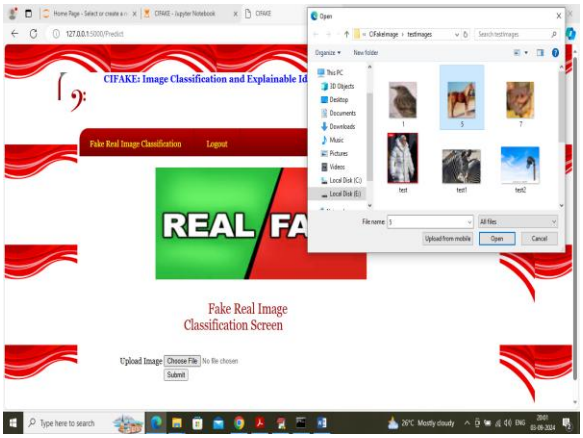

Fig.4. dataset upload

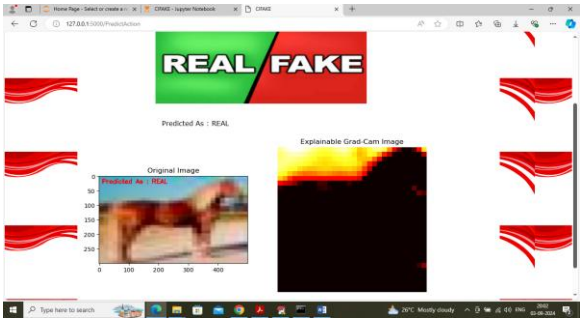

Fig.5. results

	ML Model	Accuracy	Precision	Recall	f1_score
0	Propose CNN2D	94.98	95.02	94.98	94.98
1	Extension- Modified CNN2D	95.94	95.99	95.94	95.94

Fig.6.accuracy table results

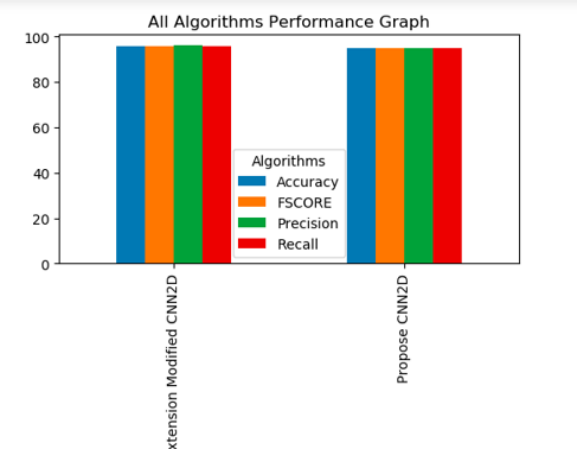

Fig.6. graphical representation

5. CONCLUSION

In conclusion, the study effectively tackles the problem of differentiating between actual and artificial intelligence-generated pictures, a task that has grown increasingly difficult as synthetic data synthesis has progressed. The system demonstrates its capacity to categorize photos based on their validity by achieving an exceptional accuracy rate of 94.98% using a potent CNN2D algorithm. Even better results are obtained with the Modified CNN2D approach, which achieves an astounding accuracy of 95.94%. These findings support the strategy of using both original and modified architectures to improve performance and demonstrate the efficacy of convolutional neural networks in image classification tasks. Explainability approaches, such as Grad-CAM, enhance the model and assist users in understanding which factors influence categorization outcomes by offering insights into the network's decision-making process. All things considered, the project not only achieves high classification accuracy but also advances computer vision and image processing research by providing useful data and useful techniques for identifying AI-generated material in further studies.

6. FUTURE SCOPE

In order to improve the model's performance and provide more techniques for explainable AI, future research may examine attention-based ways for dataset categorization. The breadth of this study might be expanded by updating the dataset with sophisticated synthetic imagery and adding photos from other domains, such as human faces and clinical imaging. In addition to increasing classification accuracy, this change will make the suggested method more applicable in a wider range of contexts, which will ultimately result in a better comprehension of AI-generated content detection.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘‘High-resolution image synthesis with latent diffusion models,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.

[2] B. Singh and D. K. Sharma, ‘‘Predicting image credibility in fake news over social media using multi-modal approach,’’ Neural Comput. Appl., vol. 34, no. 24, pp. 21503–21517, Dec. 2022.

[3] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, ‘‘On the use of Benford’s law to detect GAN-generated images,’’ in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5495–5502.

[4] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I.

Sutskever, ‘‘Zero-shot text-to-image generation,’’ in Proc. Int. Conf. Mach. Learn., 2021, pp. 8821–8831.

[5] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, ‘‘Photorealistic text-to-image diffusion models with deep language understanding,’’ 2022, arXiv:2205.11487.

[6] D. Deb, J. Zhang, and A. K. Jain, ‘‘AdvFaces: Adversarial face synthesis,’’ in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Sep. 2020, pp. 1–10.

[7] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, ‘‘Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system,’’ KSII Trans. Internet Inf. Syst., vol. 15, no. 3, pp. 1100–1118, Mar. 2021.

[8] J. J. Bird, A. Naser, and A. Lotfi, ‘‘Writer-independent signature verification; evaluation of robotic and generative adversarial attacks,’’ Inf. Sci., vol. 633, pp. 170–181, Jul. 2023.

[9] G. Pennycook and D. G. Rand, ‘‘The psychology of fake news,’’ Trends Cogn. Sci., vol. 25, no. 5, pp. 388–402, May 2021.

[10] K. Roose, ‘‘An AI-generated picture won an art prize. Artists aren’t happy,’’ New York Times, vol. 2, p. 2022, Sep. 2022.

[11] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, ‘‘Adapting pretrained vision-language foundational models to medical imaging domains,’’ 2022, arXiv:2210.04133.

[12] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, ‘‘Moûsai: Text-to-music generation with long-context latent diffusion,’’ 2023, arXiv:2301.11757.

[13] F. Schneider, ‘‘ArchiSound: Audio generation with diffusion,’’ M.S. thesis, ETH Zurich, Zürich, Switzerland, 2023.

[14] D. Yi, C. Guo, and T. Bai, ‘‘Exploring painting synthesis with diffusion models,’’ in Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI), Jul. 2021, pp. 332–335.

[15] C. Guo, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, ‘‘ArtVerse: A paradigm for parallel human–machine collaborative painting creation in Metaverses,’’ IEEE Trans. Syst., Man, Cybern., Syst., vol. 53, no. 4, pp. 2200–2208, Apr. 2023.

[16] Z. Sha, Z. Li, N. Yu, and Y. Zhang, ‘‘DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models,’’ 2022, arXiv:2210.06998.

[17] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, ‘‘On the detection of synthetic images generated by diffusion models,’’ 2022, arXiv:2211.00680.

[18] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, ‘‘Deepfake video detection through optical flow based CNN,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 1205–1207.

[19] D. Güera and E. J. Delp, ‘‘Deepfake video detection using recurrent neural networks,’’ in Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS), Nov. 2018, pp. 1–6.

[20] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, ‘‘M2TR: Multi-modal multi-scale transformers for Deepfake detection,’’ in Proc. Int. Conf. Multimedia Retr., Jun. 2022, pp. 615–623.

[21] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, ‘‘A hybrid CNNLSTM model for video Deepfake detection by leveraging optical flow features,’’ in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2022, pp. 1–7.

[22] H. Li, B. Li, S. Tan, and J. Huang, ‘‘Identification of deep network generated images using disparities in color components,’’ Signal Process., vol. 174, Sep. 2020, Art. no. 107616.

[23] S. J. Nightingale, K. A. Wade, and D. G. Watson, ‘‘Can people identify original and manipulated photos of real-world scenes?’’ Cognit. Res., Princ. Implications, vol. 2, no. 1, pp. 1–21, Dec. 2017.

[24] A. Krizhevsky and G. Hinton, ‘‘Learning multiple layers of features from tiny images,’’ 2009.

[25] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, ‘‘LAION-5B: An open large-scale dataset for training next generation image-text models,’’ 2022, arXiv:2210.08402.

[26] Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, ‘‘Recent advances in convolutional neural networks,’’ Pattern Recognit., vol. 77, pp. 354–377, May 2018.

[28] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, ‘‘A survey of convolutional neural networks: Analysis, applications, and prospects,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[29] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, ‘‘XAI—Explainable artificial intelligence,’’ Sci. Robot., vol. 4, no. 37, Dec. 2019, Art. no. eaay7120.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘‘Grad-CAM: Visual explanations from deep networks via gradient-based localization,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.

Author Profiles

Mrs. K. Pavani is working as an Assistant and Head of Department of MCA, in SRK Institute of technology in Vijayawada. She completed her MCA and M.Tech in Computer Science. She has 10 years of teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her areas of interest include AI and ML, etc

Mr.Y.Naga Malleswarao Completed his Masters of Technology from JNTUK,MSC(IS) from ANU,BCA from ANU. He has System Administrator ,Networking Administrator and Oracle Administrator. He also a web developer and python developer, Currently working has an Assistant Professor in the department of MCA at SRK Institute of Technology, Enikepadu, NTR District. His area of interest include Artificial Intelligence and Machine Learning.

Ms.N.Teja Sree is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. He has Completed Degree in B.COM(computers) from Maris Stella College, Vijayawada. Her area of interest are DBMS and Machine Learning with Python.