AN EXPLAINABLE DEEP LEARNING FRAMEWORK FOR DEEPFAKE FACE DETECTION USING DENSENET201 AND FORENSIC ANALYSIS

[1] Dr. S. Suresh B. Tech, M. Tech, Ph.D. Professor & HOD, Department of CSE, Eluru College of Engineering and technology Duggirala (v), Pedavegi (m),Eluru-534004.

[2] P. Prem Aravind, M.Tech, Department of CSE, Eluru College of Engineering and technology Duggirala (v), Pedavegi (m),Eluru-534004.

Abstract: The rapid advancement of Artificial Intelligence and image synthesis technologies has enabled the creation of highly realistic deep fake facial images, posing serious challenges to digital security, identity verification, and online trust. This project presents an intelligent Deep Fake Face Detection System that combines deep learning, forensic feature analysis, and Explainable Artificial Intelligence (XAI) techniques to accurately distinguish authentic facial images from manipulated or AI-generated images. The proposed system utilizes the DenseNet201 architecture as a deep feature extractor to learn complex facial characteristics, including facial structure, texture patterns, and visual inconsistencies. In addition to deep learning-based analysis, forensic feature analysis is performed using noise analysis, texture analysis, and edge analysis to identify image artifacts and manipulation traces that may indicate synthetic content. To improve transparency and trustworthiness, Explainable AI techniques using Gradient-weighted Class Activation Mapping (Grad-CAM) are incorporated to visualize the facial regions that contribute most to the model's prediction. This enables users to understand the reasoning behind the classification process and provides visual evidence supporting the detection result. The system is trained and evaluated using a publicly available dataset containing both authentic and deep fake facial images. Performance is assessed using evaluation metrics including Accuracy, Precision, and Recall, F1-Score, and Confusion Matrix analysis. Experimental results demonstrate the effectiveness of the proposed approach in detecting manipulated facial images while providing interpretable and user-friendly explanations. The proposed framework has potential applications in digital forensics, cyber security, biometric authentication, social media content verification, and identity fraud prevention.

1. INTRODUCTION

The rapid advancement of Deep Learning (DL) and Generative Artificial Intelligence (GenAI) has significantly improved the quality of synthetic media generation, particularly Deep fake images and videos. Deep fakes are AI-generated or manipulated visual contents that replace or modify a person's facial features with remarkable realism using deep neural networks such as Generative Adversarial Networks (GANs), Variation Auto encoders (VAEs), and diffusion models. While these technologies have beneficial applications in entertainment, education, virtual reality, and film production, they also pose serious threats by facilitating identity theft, misinformation, political manipulation, cybercrime, financial fraud, and privacy violations. Consequently, developing robust and reliable deep fake detection systems has become a critical research area in computer vision and digital forensics. Traditional deep fake detection methods primarily rely on handcrafted features, image inconsistencies, or conventional machine learning algorithms. These approaches often fail to generalize across different deep fake generation techniques because modern synthesis models continuously improve the visual quality of manipulated images. Recent advances in deep learning have enabled automatic extraction of discriminative features from facial images, significantly improving detection accuracy. Among various Convolutional neural network architectures, DenseNet201 has emerged as a highly effective model due to its dense connectivity, efficient feature reuse, mitigation of the vanishing gradient problem, and superior representation learning capability. Its deep architecture captures both low-level texture information and high-level semantic features essential for distinguishing authentic and manipulated facial images. Despite achieving impressive detection performance, most existing deep learning models operate as black-box systems, providing little or no explanation for their predictions. This lack of transparency limits user trust, makes forensic investigations difficult, and hinders the adoption of AI-based detection systems in security-sensitive domains such as law enforcement, digital media authentication, and judicial evidence analysis. To address this limitation, Explainable Artificial Intelligence (XAI) techniques have been introduced to interpret model decisions by highlighting the facial regions and forensic artifacts that contribute to the classification. Visualization methods such as Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and SHAP provide meaningful explanations, allowing analysts to understand why an image has been identified as genuine or fake. Furthermore, integrating digital forensic analysis with deep learning enhances the robustness of deep fake detection by examining subtle manipulation traces such as compression artifacts, blending inconsistencies, illumination mismatches, facial boundary irregularities, frequency-domain anomalies, and texture distortions that are often invisible to the human eye. The combination of forensic evidence and deep feature representations enables more accurate identification of sophisticated deep fake manipulations while improving the interpretability of the detection process. This project proposes An Explainable Deep Learning Framework for Deep fake Face Detection Using DenseNet201 and Forensic Analysis, which combines the powerful feature extraction capability of DenseNet201 with forensic feature analysis and explainable AI techniques to develop a transparent, accurate, and reliable deep fake detection

system. The proposed framework not only classifies facial images as real or fake with high precision but also provides visual explanations highlighting manipulated regions, thereby enhancing user confidence and supporting digital forensic investigations. By improving both detection performance and interpretability, the proposed system contributes to combating the growing threat of AI-generated synthetic media and strengthening trust in digital content authentication.

II. LITERATURE SURVEY

Deepfake detection has become an important research area due to the rapid development of artificial intelligence-based face manipulation techniques. Researchers have proposed various machine learning, deep learning, and forensic analysis methods to distinguish manipulated facial images and videos from authentic media. The following studies present significant contributions to the field. [1] Y. Li and S. Lyu proposed one of the earliest deepfake detection techniques based on identifying inconsistencies in human eye blinking. Since early GAN-generated videos often failed to reproduce natural blinking behavior, the authors utilized Long Short-Term Memory (LSTM) networks to model temporal eye movement patterns. Their approach demonstrated promising performance for early-generation deepfakes but was less effective against recent synthesis models that accurately simulate facial dynamics. [2] D. Güera and E. J. Delp developed a deepfake video detection framework that combines Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs). The CNN extracts spatial facial features, while the RNN captures temporal inconsistencies across video frames. Experimental results showed improved detection accuracy over traditional handcrafted feature-based methods. However, the framework required computationally intensive video processing and struggled with compressed videos. [3] B. Dolhansky et al. introduced the FaceForensics++ benchmark dataset, providing thousands of manipulated facial videos generated using multiple face manipulation techniques. The dataset enabled standardized evaluation of deepfake detection algorithms and demonstrated that deep learning models significantly outperform traditional machine learning approaches. Nevertheless, detection performance decreased considerably under heavy video compression and unseen manipulation methods. [4] A. Rossler et al. presented an extensive evaluation of CNN architectures for deepfake detection using the FaceForensics++ dataset. The study compared XceptionNet, MesoNet, and other deep neural networks, concluding that XceptionNet achieved superior performance because of its ability to learn discriminative facial artifacts. Despite its high accuracy, the model lacked interpretability and functioned as a black-box classifier. [5] H. H. Nguyen, J. Yamagishi, and I. Echizen proposed Capsule Networks for detecting facial manipulations. Their model effectively preserved spatial relationships between facial components and achieved competitive detection accuracy against several CNN-based methods. However, Capsule Networks required higher computational resources and longer training times compared to conventional CNN architectures. [6] S. Afchar, V. Nozick, J. Yamagishi, and I. Echizen developed MesoNet, a lightweight CNN architecture specifically designed for deepfake detection. The network focused on mesoscopic image features and achieved fast inference with relatively low computational complexity. Although suitable for real-time applications, its detection performance declined when confronted with high-quality GAN-generated images. [7] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva investigated forensic trace analysis for detecting manipulated facial images. Their approach utilized frequency-domain artifacts and image residual analysis to identify subtle manipulation traces that are difficult to perceive visually. Experimental results demonstrated that forensic features substantially improve robustness against sophisticated image manipulations. [8] Recent research has focused on Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM, SHAP, and Layer-wise Relevance Propagation (LRP) to improve transparency in deepfake detection systems. These methods generate visual explanations by highlighting manipulated facial regions responsible for classification decisions. Explainable models increase user trust and support forensic investigations; however, balancing interpretability with high detection accuracy remains an open research challenge. The literature indicates that CNN-based architectures have significantly improved deepfake detection accuracy, while forensic analysis provides complementary evidence for identifying manipulation artifacts. Furthermore, Explainable AI enhances model transparency by enabling visualization of decision-making processes. However, existing methods often suffer from limited generalization across emerging deepfake generation techniques, reduced robustness against compressed media, and insufficient interpretability. Therefore, an integrated framework combining DenseNet201, forensic feature analysis, and Explainable AI offers a promising solution for achieving accurate, robust, and trustworthy deepfake face detection.

III. EXISTING SYSTEM

Most deep fake detection systems use image processing techniques and basic machine learning to spot fake facial

images. They look for mistakes like weird facial edges, color issues, lighting problems and strange image marks. These methods can catch fakes but they are not good at detecting modern AI-made images that look very real. Some systems use machine learning methods like Support Vector Machines, Decision Trees and Random Forest classifiers. These methods need people to manually pick out features. They often struggle to understand complex patterns in deep fake images. Many current solutions just give a yes or no answer without explaining why which makes users not trust them and not understand whats going on. Deep fake detection systems need to be better, at spotting images and telling users why they made that decision. These systems should be able to handle complex deep fake images.

DISADVANTAGES OF EXISTING SYSTEM

• Limited Generalization.
• Sensitivity to Compression.
• High Computational Complexity.

IV. PROPOSED SYSTEM

The proposed system uses a way to detect fake faces. It is based on a type of computer vision called DenseNet201, Forensic Feature Analysis and Explainable Artificial Intelligence. The system checks if a face image is real or fake. It does this by learning from real and fake images. DenseNet201 is good at finding features in images. It helps the system classify images as real or fake. The system also looks at the images details. It checks for things, like texture, noise and edges. These details help the system decide if the image is real or fake. The system wants to be transparent. So it uses a tool called Grad-CAM. Grad CAM shows which parts of the face the system used to make its decision. The system has a web interface. Users can upload images. See the results. They can see how sure the system is and what it found in the image. The system uses Streamlit to make the interface. The system provides a lot of information. It helps users understand its decisions. The system is based on Deep Fake Face Detection. Deep Fake Face Detection is a way to check if a face image is real or fake. The system uses Forensic Feature Analysis. Forensic Feature Analysis helps the system check the images details. The system provides Explainable Artificial Intelligence. Explainable Artificial Intelligence helps users understand its decisions.

ADVANTAGES

- High Detection Accuracy
- Explainable Predictions
- Enhanced Forensic Analysis

SYSTEM ARCHITECTURE

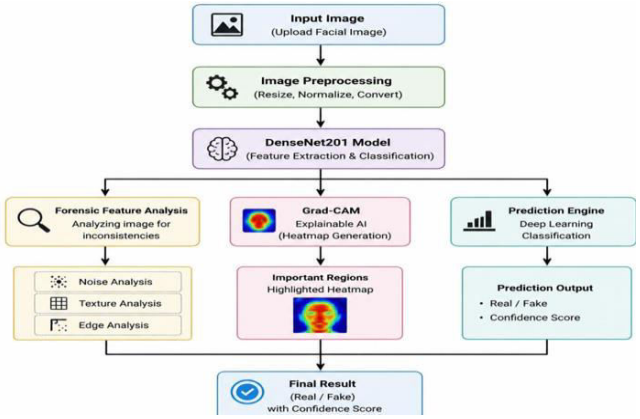

Fig 1: System Architecture

V. UML DIAGRAMS

1. USE CASE DIAGRAM

A Use Case Diagram is one of the most important behavioral diagrams in the Unified Modeling Language (UML). It is used to represent the functional requirements of a system from the user's perspective. A use case diagram illustrates the interaction between users (actors) and the system, showing the services or functionalities provided by the system. The primary purpose of a use case diagram is to identify the various activities that users can perform within the system and how the system responds to those activities. It provides a simple and understandable representation of system behavior without focusing on implementation details.

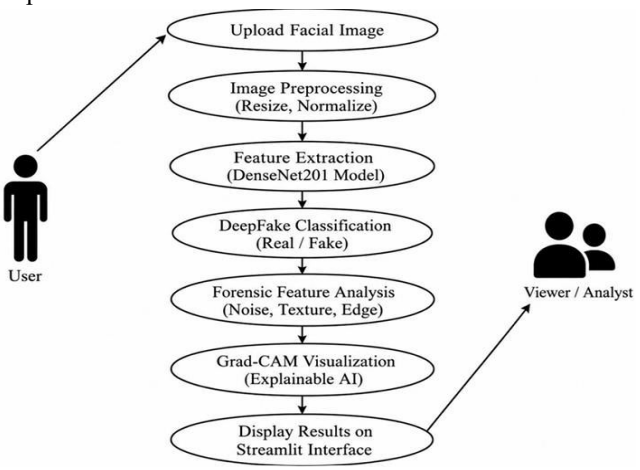

Fig 5.1 shows the Activity diagram

2. CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

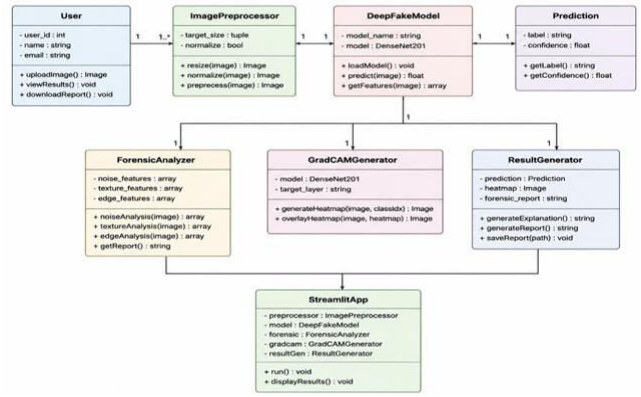

Fig 5.2 Shows the Use case Diagram

3. SEQUENCE DIAGRAM:

A sequence diagram represents the interaction between different objects in the system. The important aspect of a sequence diagram is that it is time-ordered. This means that the exact sequence of the interactions between the objects is represented step by step. Different objects in the sequence diagram interact with each other by passing "messages".

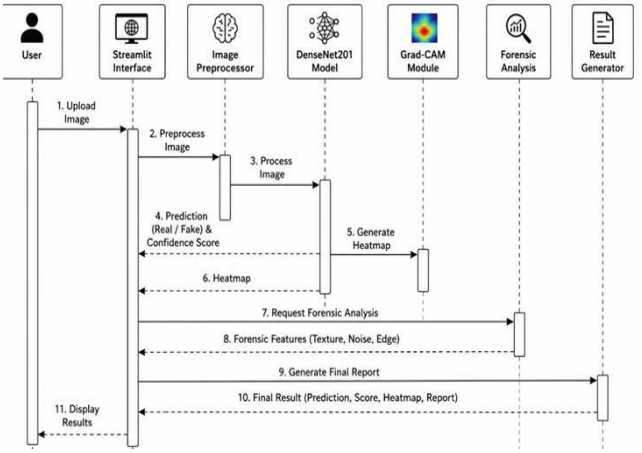

Fig 5.3 Shows the Sequence Diagram

VI. RESULTS

6.1 Output Screens

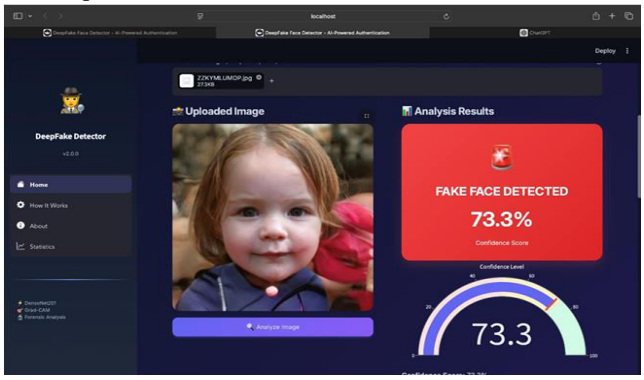

Fig 6.1 Deep fake Face Detected

In above shows the Deep fake Face Detected.

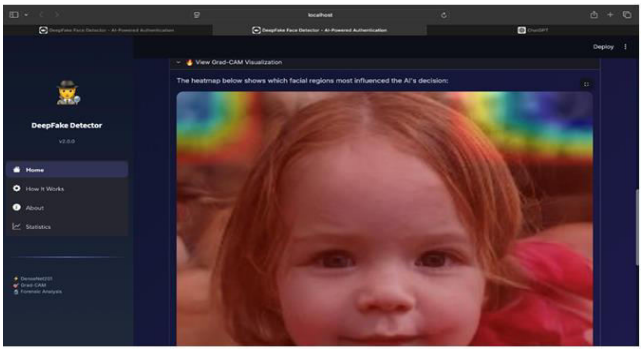

Fig 6.2 Deep fake Face Detected Grad-Cam Visualization

In above screen shows the deep fake Face Detected Grad-Cam Visualization.

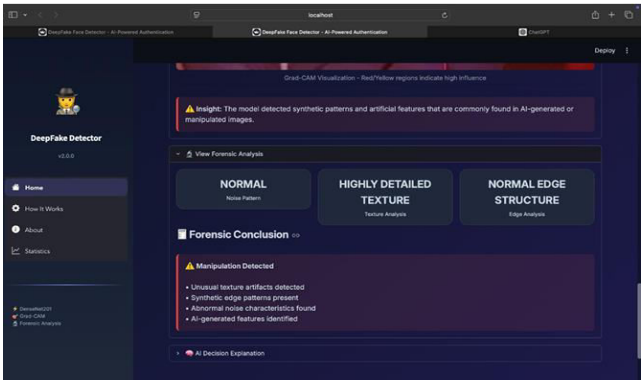

Fig 6.3 Forensic Conclusion

In above screen shows the Forensic Conclusion

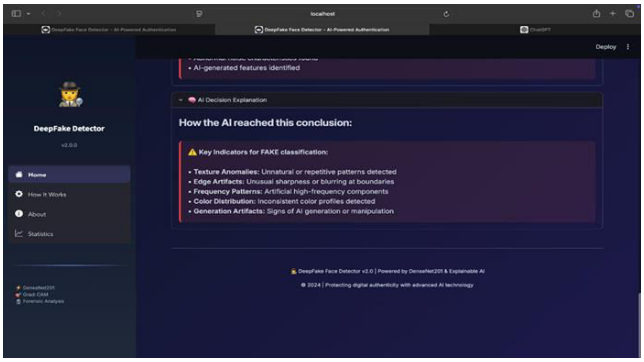

Fig 6.4 AI Decision Explanation

In above screen shows the AI Decision Explanation.

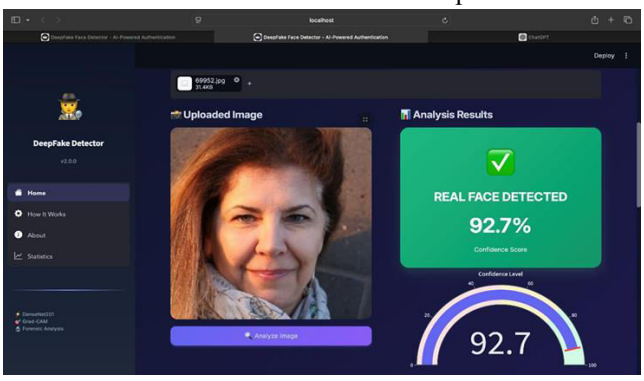

Fig 6.5 Real Face Detected

In above screen shows the Real Face Detection.

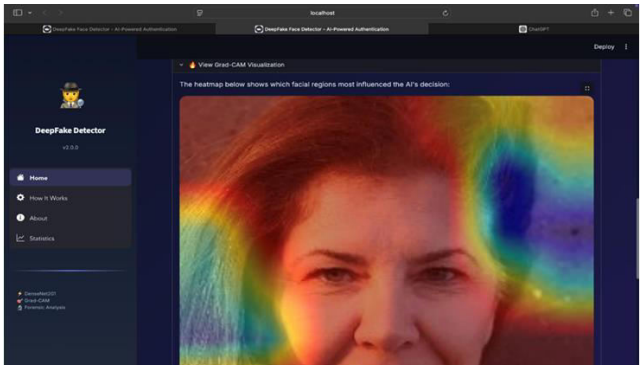

Fig 6.6 Real Face Detected Grad-Cam Visualization

In the above screen shows the Real Face Detected Grad-Cam Visualization

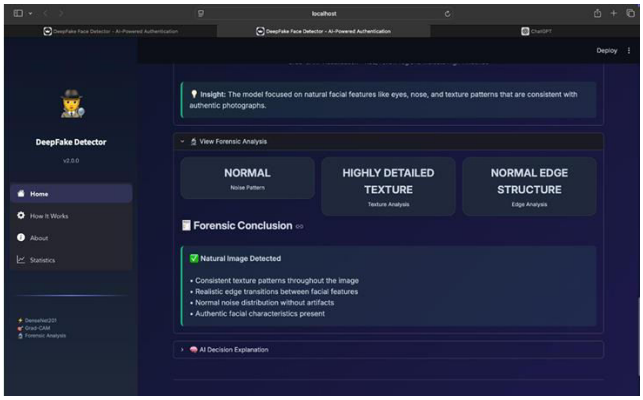

Fig 6.7 Forensic Conclusion

In above screen shows the Forensic Conclusion.

VII. CONCLUSION

The proposed Deep Fake Face Detection System was successfully developed and implemented using DenseNet201 Deep Learning Architecture, Grad-CAM Explainable Artificial Intelligence, and Forensic Feature Analysis techniques. The system performs image preprocessing, feature extraction, deepfake classification, and visual explanation generation to provide reliable and transparent detection results. The project effectively distinguishes between real and manipulated facial images by analyzing facial patterns, texture information, edge characteristics, and deep learning features. Grad-CAM visualization helps users understand the regions that influenced the model's decision, thereby improving the interpretability and trustworthiness of the prediction process. The developed system achieved satisfactory performance in terms of accuracy, precision, recall, and F1-score during testing. The integration of forensic analysis and explainable AI further enhances the credibility of the detection results. Therefore, the proposed solution can serve as an effective tool for identifying deep fake images and supporting digital media authenticity verification.

VIII. REFERENCES

1. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4700–4708.

2. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626.

3. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 1–11.

4. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge Dataset. arXiv Preprint arXiv:2006.07397.

5. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. Information Fusion, 64, 131–148.

6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press, Cambridge, Massachusetts.

7. Chollet, F. (2021). Deep Learning with Python (2nd Edition). Manning Publications.

8. Abadi, M., Barham, P., Chen, J., et al. (2016). TensorFlow: A System for LargeScale Machine Learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 265–283.

9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

10. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 56–61. 41

11. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95.

12. Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

13. Streamlit Inc. (2024). Streamlit Documentation for Machine Learning Applications.

14. TensorFlow Team. (2024). TensorFlow Keras API Documentation.

15. Pillow Development Team. (2024). Pillow Documentation: Python Imaging Library 4