**Informatics Institute of Technology**

**In Collaboration With**

**University of Westminster**

# Deepfake Low Resource Image Detection with Explainable Reporting

A Dissertation By

**Mr. Abdul Baasith**

W1790108 / 2019566

Supervised By: Guhanathan Poravi

Department: Computer Science

**May 2023**

This Final Thesis is submitted in partial fulfillment of the requirements for the
BSc (Hons) Computer Science degree at the
University of Westminster

## ABSTRACT

Deepfake technology is widely used, which has led to serious worries about its potential for malicious usage, particularly in the context of profile images on social media platforms. This can result in misleading information, impersonation, and other harmful activities, making it crucial to develop robust deepfake detection methods.

In this research, the author proposes a solution for deepfake detection in profile images using deep learning approaches, including transfer learning and hybrid models, along with Explainable AI (XAI). The proposed solution leverages transfer learning to fine-tune a pre-trained deep neural network and alter the architecture of models to improve the overall performance. Additionally, XAI is employed to increase the interpretability and transparency of the decision-making process, enabling the understanding of why a particular image was classified as fake or real. The proposed solution was evaluated on a on various testing matrix and was able achieved high accuracy, robustness, and interpretability. These results demonstrate the potential of transfer learning, hybrid models, and XAI for effective deepfake detection in profile images.

The author of this work has trained a InspetionResnetV2 model as the backbone and have made a custom hybrid architecture for the feature extraction which achieved an accuracy of 0.96 for the testing data and loss of 0.1388. To interpret the results of the model, the author used a combination of LIME and Integrated Gradient XAI techniques. LIME XAI helps to provide an explanation of how the model arrived at its predictions highlighting the regions that are used for the prediction, In the integrated gradient it’s shows the interpretation of the pixels that are used for the arrival of the decision

**Keywords:** Deepfake, CNN, Optimization, Computer Vision, face2face Digital Media Manipulation, XAI

### ACM Subject Descriptions

CCS >> Theory of computation >>Design and analysis of algorithms >> Parallel algorithms

Computing methodologies >> Machine learning >> Machine learning approaches >> Neural networks

**PUBLICATIONS**

1. **International Journal of Innovative Science and Research Technology (IJISRT)**
   **Status** – **Accepted Pending Payment**
   **Reference**: **APPENDIX**

2. **ETLTC - Summer2023 (4th Summer International Conference on Careers in Applied Sciences)**
   **Event Name -** 4th Summer International Conference on Careers in Applied Sciences
   **Status** – **Accepted & Reviewed**
   **Reference**: **APPENDIX**

3. **4th International Conference on Innovations in Info-business & Technology (ICIIT 23)**
   **Status** – Manuscript Completed and submitted (Review)
   **Reference**: **APPENDIX**