National College of Ireland

Localized Deepfake Detection: Enhancing Authenticity Verification in Irish Media

MSc Research Project
Data Analytics

Simi Silvester Correia
Student ID: X23320818

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Simi Silvester Correia
Student ID:	X23320818
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Teerath Kumar Menghwar
Submission Due Date:	11/08/2025
Project Title:	Localized Deepfake Detection: Enhancing Authenticity Verification in Irish Media
Word Count:	4260
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Simi Silvester Correia
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Localized Deepfake Detection: Enhancing Authenticity Verification in Irish Media

Simi Silvester Correia
X23320818

Abstract

Deepfakes are becoming harder to detect and more convincing than ever, threatening the trust we have in news, public figures, and digital communication. While many detection models already exist, most are trained on large, global datasets that do not reflect regional differences in faces, speech, or media context. This makes them less effective when applied to local content. In this research, the researcher focuses on building a region-specific deepfake detection system for Irish media. The project was carried out in three phases. In the first phase, real and fake images from Irish news sources were collected and used to build a proof-of-concept model. In the second phase, fake images were generated using the Stable Diffusion model, while realistic non-manipulated images were taken from This Person Does Not Exist because of their close resemblance to Irish features. In the third phase, the researcher tested cross-dataset performance by training on the global DFDC dataset and testing on Irish data, where accuracy dropped to 50%, showing that global models fail to generalize well to local content. Finally, by mixing both datasets and emphasizing the Irish samples, the region-specific approach produced much stronger results. Overall, the global dataset achieved 50% accuracy (F1: 33%, ROC-AUC: 46.9%), while the region specific dataset reached 74% accuracy (F1: 76%, ROC-AUC:%). These findings suggest that localized training not only improves accuracy but also strengthens the reliability of deepfake detection in Ireland, offering a model that can be adapted by other regions to safeguard media integrity.

Keywords: ***Deepfake Detection, Region Specific, Stable Diffusion, EfficientNet-B3, Irish Media***

1 Introduction

We are currently in the middle of 2025, a generation defined by rapid technological evolution, where artificial intelligence (AI) is seamlessly integrated into nearly every aspect of daily life. While AI offers transformative advantages across industries, it also raises ethical concerns. With just a few lines of code and access to publicly available tools, it is now possible to generate hyper-realistic videos of individuals saying or doing things they never did, a phenomenon widely known as deepfakes (Chesney and Citron; 2019). This poses a severe challenge to trust in digital media, particularly when fake videos and edited news segments propagate disinformation at an increasing rate.

In 2024, RTÉ news presenter Sharon Tobin, a well-known and trusted Irish journalist, became the target of a deepfake scam that falsely used her likeness to promote a fake investment scheme (Irish Independent; 2024). This incident is a clear reminder of how

deepfake technology can be used to mislead the public and undermine trust in reputable media houses. Although deepfake detection has advanced from early machine learning models to more complex multimodal systems (Agarwal et al.; 2019), challenges remain. Most existing algorithms rely on generic datasets, often composed of celebrity images, which limits their ability to generalise across diverse populations. Variations in facial structure, expressions, and cultural context across regions make it challenging to apply a universal detection approach. As a result, this gap highlights the need to develop region-specific models, which is critical to enhancing accuracy, relevance, and novelty in the research. Even the European Union has acknowledged the growing threat of deepfakes and is actively working on raising awareness and developing guidance to combat them (European Commission; 2024).

A. Research Question

How can a region-specific deepfake detection model trained on Irish media outperform general-purpose models in terms of detection accuracy, contextual relevance, and computational efficiency? In other words, can a model built with Irish-specific datasets better recognise subtle manipulations that global models might overlook, while also being optimised to run efficiently on limited hardware

B. Proposed Solution

The author of this research proposes a four-phase approach to developing a region-specific deepfake detection model tailored for Irish media.

Phase 1 focused on building a proof-of-concept model using only 100 Irish images (50 real, 50 fake) sourced from publicly available news and entertainment outlets, capturing the unique cultural and facial characteristics of Ireland.

Phase 2 expanded the dataset to 500 images by incorporating synthetic deepfake samples generated using the Stable Diffusion model Rombach et al. (2022), while realistic non-manipulated images were obtained from *This Person Does Not Exist* This Person Does Not Exist (2024), carefully selected for their close resemblance to Irish features. This expansion allowed the model to better capture region-specific visual traits while reducing overfitting.

Phase 3 tested the generalization ability of global datasets by training exclusively on the DFDC dataset and then evaluating the model on Irish images. This experiment revealed the limitations of global-only training, with accuracy falling to 50%, confirming that models trained on large international datasets fail to generalize effectively to Irish media.

Finally, Phase 4 adopted a mixed-dataset strategy, combining Irish samples with DFDC data to balance global diversity and local specificity. This approach ensured that training included authentic Irish features while also leveraging the variety of conditions present in global datasets, significantly improving overall performance and reducing bias.

By following these four phases, the proposed solution demonstrates that region-specific training, when combined with global diversity, strengthens the reliability of deepfake detection systems. This contribution addresses a major gap in existing research that is largely dependent on generic, globally sourced datasets, while aligning with ongoing European efforts to combat misinformation at the national and regional levels.

C. Structure of the Paper

This paper is structured into six sections. **Section 1** introduces the topic, outlines the problem, states the research question, and presents a novel solution to the problem. **Section 2** reviews existing literature on deepfake detection, highlighting key findings, commonly used datasets and models, and identifying gaps in current research. **Section 3** outlines the methodology, including data sources, model architecture, and the evaluation and comparison of different models. **Section 4** discusses the ethical considerations relevant to the development of deepfake detection systems. **Section 5** concludes the paper and provides suggestions for future research directions.

2 Related Work

This literature review examines the progression of deepfake detection methods, starting from early approaches that relied solely on visual or audio input, to more sophisticated models that integrate both facial and audio features. The goal is to evaluate the effectiveness of these techniques and identify gaps that are relevant to the focus of this study. The review is organised into sections that reflect key developments in deepfake research, helping the author both identify research gaps and determine suitable techniques and datasets for the proposed model. By reviewing existing literature carefully, the author intends to pinpoint opportunities to create a more effective and practically relevant deepfake detection system tailored specifically to Ireland.

To ensure credibility and relevance, the author selected academic sources from well-established platforms such as Google Scholar, the NCI Library, IEEE Xplore, and ScienceDirect. The selected papers were chosen for their clarity, citation strength, and overall contribution to the field.

A. Machine Learning based approach

Traditional machine learning classifiers have been explored for deepfake detection due to their interpretability and lower computational requirements. Mishra et al. (2023) investigates the effectiveness of traditional machine learning algorithms, including SVM, Decision Trees, Naïve Bayes (NB) and k-Nearest Neighbors (KNN), for accurately distinguishing real human images from deepfake images. Their findings indicated that SVM achieved the highest accuracy at 99.5%, while NB recorded the lowest at 96%. However, the study's reliance on accuracy as the sole evaluation metric limits the comprehensiveness of the performance evaluation. Even though the paper achieved very high accuracy, there were some notable gaps identified. The data size was moderate, with 2,848 training and 518 testing videos, limiting model generalization in diverse scenarios. The videos used had controlled conditions like perfect lighting and good quality, but results might differ with more realistic videos. Machine learning models are extensively used for extracting features from the Celeb-DF(v2) and FaceForensics++ datasets, but these models can perform well only if the videos are taken in clear formats Padmashree and Karunakar (2023). From the gaps in the paper, the author studied that use of more diverse and realistic images, particularly screenshots from Irish news and media will aim to create a robust, accurate deepfake detection model.

B. Transition to Deep Learning based models While traditional machine learning

models such as SVM and Decision Trees have shown strong performance in deepfake detection under controlled conditions, they often fall short when dealing with complex data representations or unstructured content Padmashree and Karunakar (2023). To overcome these limitations, researchers have turned towards deep learning approaches, which can automatically learn intricate visual patterns without manual feature extraction. One such study introduced a hybrid model called Deepfake Predictor (DFP), which integrates the pre-trained VGG16 model with custom Convolutional Neural Networks (CNNs) to detect manipulated images Raza et al. (2022). They specifically chose VGG16 due to its proven effectiveness in extracting detailed image features and integrated it with CNN layers to further modify and identify deepfake-specific visual patterns. The DFP model was then tested on the publicly available Celeb-DF dataset, consisting of 590 real and 5,639 fake images, split into 80% training data and 20% testing data. In their experiments, the proposed DFP model achieved an overall detection accuracy of 94%, surpassing other advanced models evaluated for comparison: Xception (89%), NAS-Net (91%), and MobileNet (89%). Additionally, the DFP model obtained good values for evaluation metrics, including a precision of 95%, recall of 95%, and an F1-score of 95%. These exact numbers clearly indicate the proposed method had strong capability in identifying manipulated or fake images, showcasing its efficiency over existing transfer-learning techniques. Despite the promising results, there are several gaps that reduce its real-world applicability. For instance, the authors did not report the training time or resource usage of the model, which is critical since combining VGG16 with additional CNN layers likely demands significant computing power. This makes it unclear how practical the model would be for use cases requiring fast or real-time detection, such as in social media platforms or live news broadcasting. Additionally, the study focused solely on visual input, ignoring the potential of audio-based cues, which can be equally valuable in identifying deepfakes. The paper shows that combining VGG16 with custom CNN layers can boost detection performance. The author also highlights the need for cloud support to improve speed and scalability and suggests that using both image and audio data could lead to more reliable, real-world results.

C. Advancement in Multimodal Datasets

Following the development of deep learning approaches, researchers began to recognize the limitations of using only visual or audio data in isolation. To address this, Khalid et al. (2022) introduced FakeAVCeleb, a novel multimodal dataset that includes both fake videos and perfectly lip-synced cloned audio. This dataset created four types of combinations—real or fake audio with real or fake video resulting in 20,000 manipulated samples. This approach helped tackle common issues in deepfake research, such as data imbalance and lack of diversity. Although the study made significant progress, it did not include speaker verification or speech spoofing techniques, which could further improve fake audio detection. Similarly, Yan et al. (2024) presented DF40, a large-scale benchmark dataset featuring 40 different deepfake techniques, ten times more than the commonly used FaceForensics++ dataset. The researchers tested popular models like ViT, Swin Transformer, EfficientNet, and Xception. While these models did well on known deepfakes (over 90% AUC), their performance dropped on new, unseen deepfakes. For example, ViT dropped to 68.5% AUC and Swin to 65.1%. But when trained on the full DF40 dataset, ViT's performance improved by 14.1%, showing that using more varied data helps models do better Yan et al. (2024). These studies strongly influenced the direction of the present study. Inspired by their work, this research aims to build a localized Irish deepfake data-

set using real-world content from Irish news and entertainment platforms. The goal is to develop a more complete and context-aware deepfake detection system suitable for regional use.

D. Evolution from Unimodal to Multimodal Deepfake Detection Techniques

Expanding on the importance of diverse and realistic datasets, recent research has also focused on how deepfake detection techniques have shifted from unimodal to multimodal approaches. Liu et al. (2024) explored how deepfake detection has evolved toward multimodal approaches that combine both audio and video inputs. Their survey reviewed over 120 models and datasets, showing that using multiple data types together improves detection accuracy. For example, on the FakeAVCeleb dataset, the multimodal model MIS-AVoiDD reached 92.5% accuracy, compared to only 82.9% using audio alone and 85.7% with just images. Similarly, when applied to the DFDC dataset, the AUC score increased from 85.5% (video only) and 79.2% (audio only) to 91.7% when both were combined. The paper also highlighted that newer transformer-based models like Swin and ViViT outperform older CNN-based architectures in multimodal settings. Although multimodal models outperform single-modality approaches, the author identifies several challenges. Merging video with perfectly synced audio remains difficult, and models may underperform when either audio or video data is missing in certain cases. Additionally, multimodal systems are computationally demanding due to larger model sizes and the need to process both video and audio streams simultaneously. These models can also struggle when exposed to data in unfamiliar languages or faces that have not been encountered during training Katamneni and Rattani (2023). The current study aims to develop a model capable of processing various input types, including audio, images, and video. The model will be trained using Irish-specific media content to ensure its effectiveness in European-based applications, particularly within local news and entertainment contexts. This focused strategy is expected to help the model better understand regional characteristics, such as accents, facial features, and cultural patterns, thereby improving its accuracy and contextual relevance.

E. Importance of using Localized dataset

Continuing the focus on tailoring models to specific populations, Kwon et al. (2021) introduced KoDF, a deepfake detection dataset built specifically for the Korean population, containing over 62,000 real and fake videos using Korean faces, voices, and media content. Models trained on KoDF achieved an AUC of 94.3% on Korean data but dropped by up to 14.4% when tested on non-Korean datasets. This shows that region-specific training improves accuracy for local content. Since the model is trained for local Korean subjects, it might not work well for global deepfakes. This paper inspired the current study to create an Ireland-specific dataset using local news and entertainment clips, proving that a location-based model performs better than a generalized one. The dataset creation process outlined in KoDF also provides a valuable structure that can be adapted to develop a similar dataset tailored for Irish media. The literature shows that deep learning models and larger datasets have improved results, but most are built on global or celebrity data, which limits their effectiveness in regional settings. Studies like FakeAVCeleb and KoDF highlight how combining modalities and using local data improves detection. This reinforces the need for an Irish-specific approach, which this study aims to explore.

Table 1: Summary of Reviewed Studies

Category	Author and Year	Gaps Identified	Key Findings
ML-Based Approach	Mishra et al. (2023)	Model was overfitted. Only Accuracy was used as performance metrics. Dataset in controlled conditions	SVM best (99.5%), NB lowest (96%)
	Padmashree and Karunakar (2023)	Performance drops on low-quality	ML models effective on clean data; need for realistic region-specific datasets
	?Laurence et al. (2022); Matern et al. (2019); Afchar et al. (2018)	All the four papers struggled with unseen fake dataset. ML models overfitted for smaller datasets.	Good accuracy on FaceForensics++ with low compute cost
Transition to DL	Say et al. (2025)	Used static images only	Identified GAN-specific residual patterns invisible to humans
	Raza et al. (2022)	No training time/resources reported	94% accuracy of EfficientNet and Xceptionet
	Dang et al. (2020)		Showed augmentation boosts generalisation
Multi-modal Deepfake detection	Liu et al. (2024)	Audio–video sync issues; high compute	Multimodal performed better than unimodal
	Katamneni and Rattani (2023)	High complexity and very high training time	Noted strong potential but operational difficulties
	Yan et al. (2024)	Drop in performance on unknown fakes	ViT improved 14.1% AUC with diverse training
	Khalid et al. (2022)	No speaker verification / speech spoofing	Balanced multimodal dataset of 20k samples
Region Specific approach	Kwon et al. (2021)	Accuracy drops by 14% on other datasets	AUC 94.3% on Korean data
	Choi et al. (2020)	Limited global applicability	Region-specific features improved local detection
	Present Study (2025)	Small Irish-specific visuals (400 images)	EfficientNet-B3 best performance

3 Methodology

3.1 Research Method

This study adopts a supervised learning approach to develop a unimodal deepfake detection model tailored to Irish media content. The model is designed to analyse visuals to detect manipulated media with higher contextual relevance and accuracy, particularly for local news and entertainment contexts. To support this, a custom dataset has been created by collecting real and fake media samples from publicly available Irish content like RTÉ Libraries (RTÉ Libraries and Archives, 2025), National Library of Ireland, Digital Repository of Ireland (DRI) etc. Pre-trained models like Restnet18 and Efficient Net B3 are used for visual features. Grad-CAM visualisation is applied to interpret the model's decisions and verify whether it focuses on relevant facial features when identifying fakes. This research method directly addresses the core question of whether a region-specific, unimodal model can outperform generic models, offering a targeted solution to deepfake detection in Irish media.

3.2 Data Collection

In Phase 1 of the dataset collection, real images were obtained from RTÉ Libraries & Archives (2025), the National Library of Ireland (2025), and the IFTN – Irish Film &

Television Network. Fake images were sourced from TheJournal.ie, CorkBeo, The Guardian, and Independent.ie, where they were explicitly labelled as fakes in published reports. A total of 50 real and 50 fake images were collected for this phase.

Table 2: Summary of Data Collection for Phase 1 of Analysis

Source Type	**Description**	**Author/Reference**
Real Images	News broadcasts from RTÉ Libraries and Archives	RTÉ Libraries and Archives (2025)
Real Images	Historic and new age photographs from the National Library of Ireland	National Library of Ireland (2025)
Real Images	Visual samples featuring Irish actors via a film & TV actors database	IFTN – Irish Film & Television Network
Fake Media	Deepfake scam ad featuring Sharon Tobin, Brendan Gleeson, and Luke O'Neill; AI deepfake of Michael O'Leary in scam ad	TheJournal.ie (2025)
Fake Media	AI-generated images circulation report	CorkBeo (2025)
Fake Media	Deepfake video of Nigel Farage playing Minecraft	The Guardian (2024)
Fake Media	Deepfake scam ad uses video of Leo Varadkar and Colette Fitzpatrick	Independent.ie (2024)

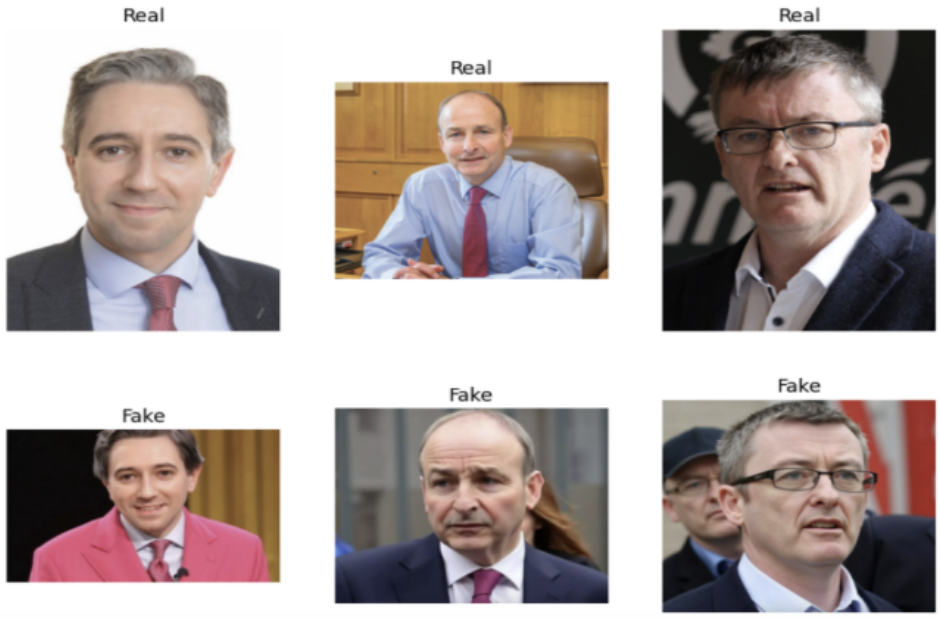

Figure 1: Images from News article

In Phase 2, fake images were generated using Stable Diffusion v1.5 with targeted Irish prompts such as "Deepfake of an Irish man, realistic, DSLR, natural light" and "Deepfake of an Irish woman, realistic, DSLR, studio lighting". Real images for Phase 2 were generated using *This Person Does Not Exist*, selected to closely match Irish ethnicity. This produced 200 real and 200 fake images.

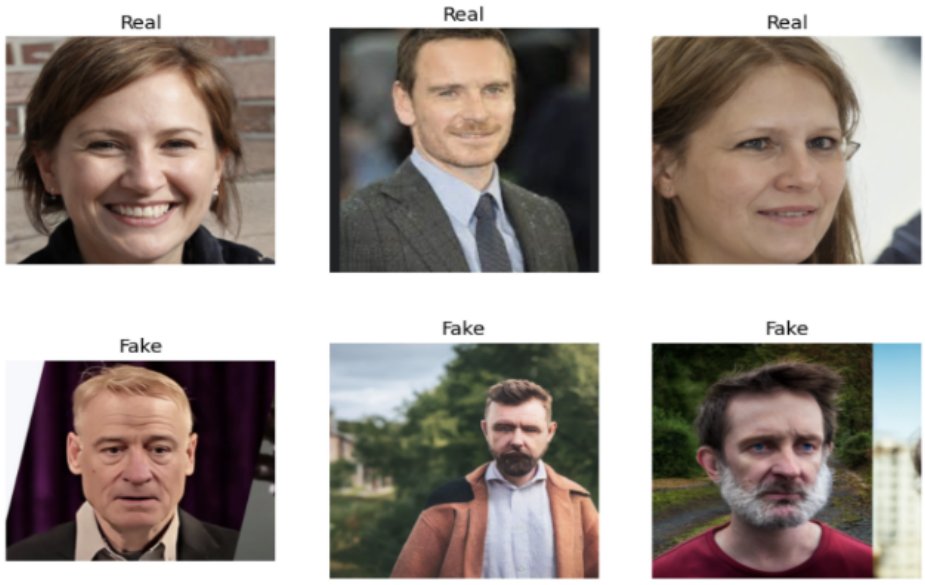

Figure 2: Synthetic Images

In Phase 3, To evaluate generalization, the DeepFake Detection Challenge (DFDC) dataset **?** was incorporated. DFDC contains large-scale, globally sourced deepfake and real video samples, making it a widely used benchmark. In this study, DFDC data was used in two ways: (i) training models on DFDC and testing them on Irish data to measure how well global models transfer to regional contexts, and (ii) mixing DFDC with Irish samples to test whether hybrid training improves detection across both domains. This integration ensured that the evaluation captured both global and regional performance.

Figure 3: DFDC Images.

All images were sourced from publicly available, non-private materials, with terms of use reviewed to ensure compliance for academic purposes. No personally identifiable

information was stored, and all real and fake images were clearly labelled and stored in separate directories.

3.3 Data Preprocessing

Resizing the data: Once the data is collected the next step is data pre-processing. The data was processed so that it was ready for model training. Firstly, all the images were resized to 224 × 224 pixels. This size is a common choice for computer vision models like ResNet and EfficientNet and ensures that all images are the same size and can be processed consistently.

Convert to Tensor: Next, the images were converted into tensors with three channels (red, green, blue). This gave each image a shape of (3, 224, 224). After that, the pixel values were normalised, meaning they were scaled so they were on a similar range. For this project, the images ranged between -1.36 and 2.64.

Data Augmentation: To help the model generalise to different types of images in real life, data augmentation was used. This creates extra training examples by making small, controlled changes to the original images without altering their meaning. In this project, two main techniques were applied: horizontal flipping and rotation using `RandomRotation(10)`.

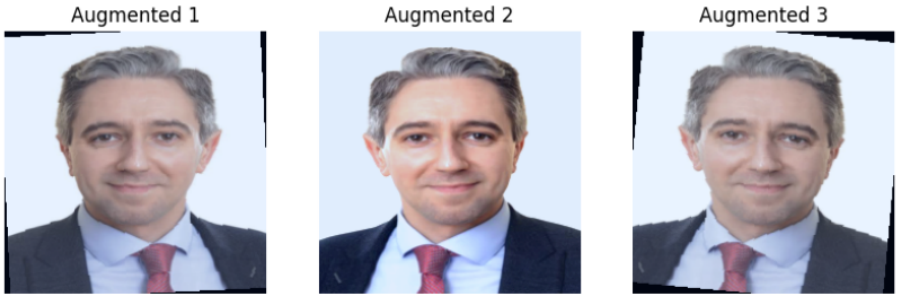

Figure 4: Augmented, resized, and normalised dataset samples

Train and Validation Split: The data was split into training (80%) and validation (20%) datasets.

Data Loaders: Images were wrapped in data loaders to feed the model in mini-batches. The training loader shuffles the data each epoch, so the model does not memorise the order, while the validation loader does not shuffle.

4 Design Specification

Two convolutional neural network architectures were selected for this study: ResNet-18 and EfficientNet-B3. ResNet-18 is a lightweight residual network with skip connections designed to address vanishing gradient problems, making it efficient for small datasets

and fast to train while maintaining strong classification accuracy. EfficientNet-B3 employs compound scaling to balance depth, width, and resolution, allowing it to capture fine-grained details and subtle artefacts in facial images. Although more computationally intensive than ResNet-18, EfficientNet-B3 has the potential for higher accuracy in detecting deepfakes. Both models were pre-trained on ImageNet and fine-tuned on the Irish deepfake dataset to adapt to facial features, lighting, and styles specific to Irish media content.

The system workflow begins with input images, which undergo preprocessing steps including resizing, tensor conversion, normalisation, and augmentation. These processed images are then fed into the chosen CNN backbone (ResNet-18 or EfficientNet-B3) for feature extraction. A classification layer outputs the prediction as "real" or "fake". Grad-CAM visualisation is applied to verify whether the model is attending to relevant facial regions during prediction.

The design required a balanced dataset for each phase, computation within the capabilities of Google Colab Pro (T4 GPU) and CPU environments, and compliance with GDPR and ethical guidelines, ensuring that only publicly available, non-private data was used.

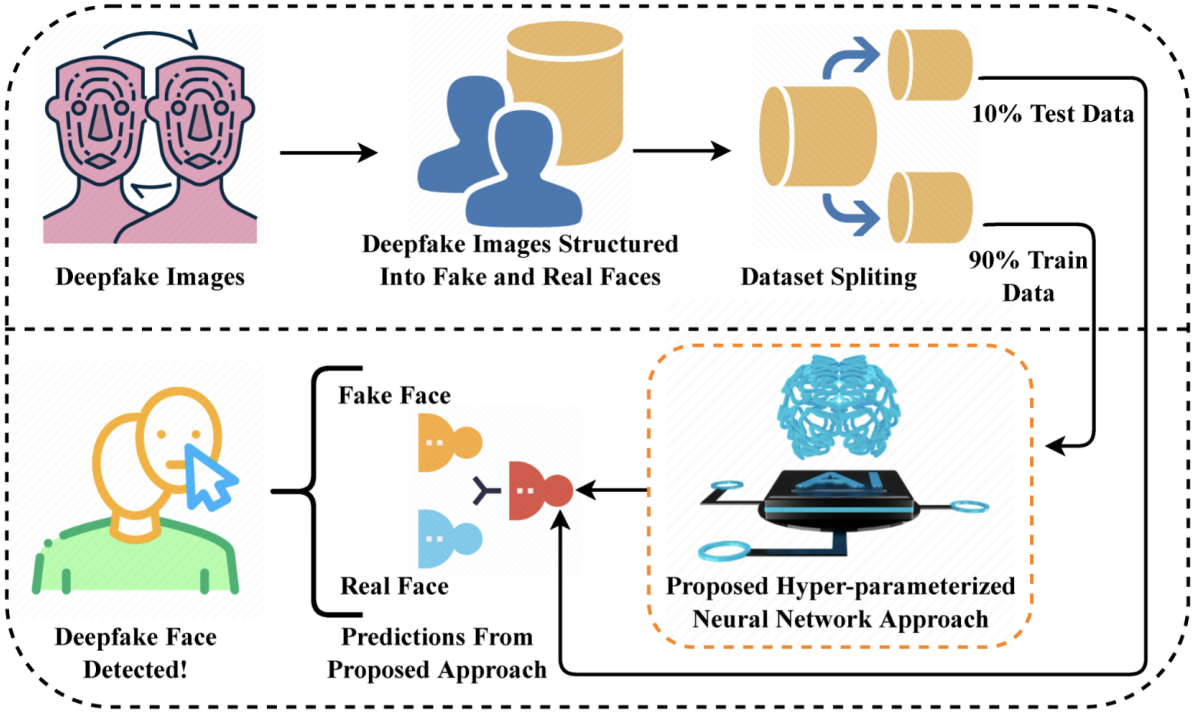

Figure 5: Proposed unimodal deepfake detection architecture

5 Implementation

The project was implemented in a Google Colab environment using a T4 GPU for accelerated computation, with CPU fallback for smaller tests. The implementation utilised PyTorch and `torchvision` for model training, and Grad-CAM libraries for interpretability. Synthetic fake images were generated using Stable Diffusion v1.5, while synthetic real images were generated from *This Person Does Not Exist*. All images were stored in

a structured Google Drive directory, separated into `/real` and `/fake` folders.

1. **Irish-only evaluation:** Training and testing were carried out on 100 Irish images (50 real, 50 fake) to build a proof-of-concept model.
2. **Expanded Irish dataset:** Training and testing on 500 images, including real/fake samples from Irish media, Stable Diffusion, and *This Person Does Not Exist*.
3. **Cross-dataset evaluation:** Training on 1,600 images from the DFDC dataset and testing on 100 Irish samples to evaluate generalization from global to local data.
4. **Mixed dataset training:** Combining DFDC with Irish datasets for training and testing, ensuring that the model is exposed to region-specific features while retaining global diversity.

Both models trained efficiently within Colab's constraints: ResNet-18 provided faster training, while EfficientNet-B3 required more computation but delivered stronger results. Outputs included trained models, confusion matrices, ROC curves, and Grad-CAM heatmaps, providing both performance and interpretability for the proposed unimodal detection approach.

6 Evaluation

The purpose of this section is to provide a comprehensive analysis of the results and main findings of the study, as well as the implications of these findings from both academic and practitioner perspectives. Only the most relevant results that support the research question and objectives are presented here. An in-depth and rigorous analysis of the results is provided, using statistical measures such as accuracy, precision, recall, F1-score, and ROC-AUC to assess performance. Visual aids such as confusion matrices, ROC curves, and Grad-CAM visualisations are also used to illustrate the results.

To assess model performance, standard classification metrics were used. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. In the context of this study, a 'real' image refers to authentic media, and a 'fake' image refers to a manipulated or AI-generated sample.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Accuracy measures the overall proportion of images (real and fake) that the model classified correctly.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Precision measures how many of the images predicted as 'fake' were actually fake.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Recall measures how many of the actual fake images were detected by the model as fake.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

F1-score balances precision and recall.

In addition, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was used as an overall measure of model performance. It shows how well the model can separate real images from fake images across different thresholds. A higher ROC-AUC means that the model is good at detecting fake images as 'fake' and real images as 'real' even when the decision boundary changes.

$$\text{PR-AUC} = \int_0^1 p(r)\, dr \tag{5}$$

Precision–Recall AUC (PR-AUC) measures the area under the precision–recall curve. It is particularly useful for imbalanced datasets, as it focuses on how well the model identifies the positive class (fake images) without being biased on real images. A higher PR-AUC for fake images indicates the model is more reliable in detecting fakes.

The results of this study are presented in four experimental phases, each designed to test the effectiveness of region-specific training compared to global datasets. The findings highlight how different data sets influence the detection accuracy, generalization, and overall robustness of the models.

6.1 Experiment 1: Proof of Concept (Phase 1)

In the first phase, the ResNet-18 model was trained on a small dataset of 50 real and 50 fake images, all sourced from Irish media. This stage served as a proof of concept to test whether a model could learn to distinguish between authentic and manipulated content in a region-specific context. Training was carried out for five epochs. Performance in the initial epoch was modest (Train: 57.5%, Val: 45.0%), as the model began to learn facial patterns. By the fifth epoch, accuracy improved to 97.5% on the training set and 80.0% on the validation set.

The 17% difference between training and validation accuracy indicates overfitting, expected given the small dataset size. However, these results demonstrated that the model could distinguish features, while also highlighting the need for a larger and more diverse dataset to improve generalization.

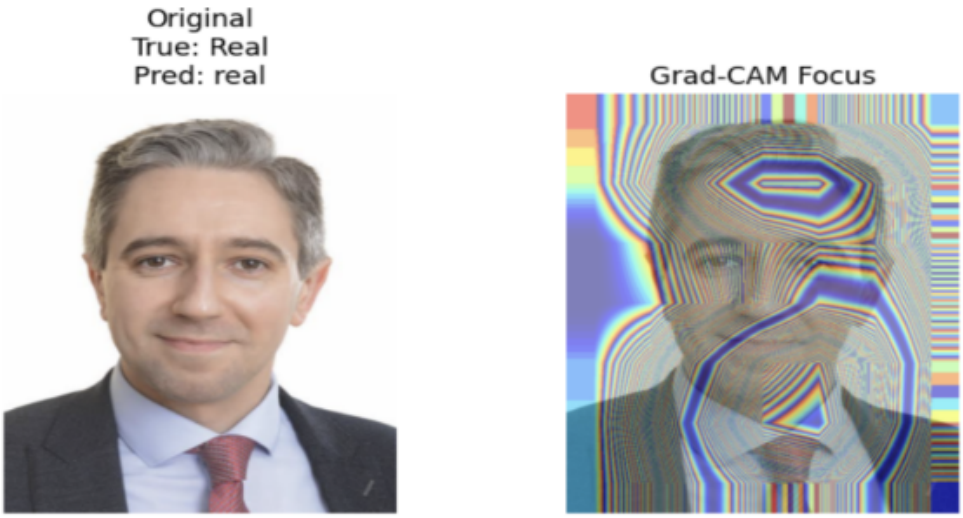

Figure 6: Grad CAM Visualization for Phase 1: RestNet-18 Model

From the Grad CAM Visualization we can clearly state that the model focuses on the facial dimensions of the image like the forehead, but it also focusses on the tie and neck area of the person. So, the model performed decently but it can be improvised further.

6.2 Experiment 2: Expanded Irish Dataset (Phase 2).

In the second phase, the dataset was expanded to include 500 images sourced from Irish media, Stable Diffusion, and *This Person Does Not Exist*. This was to check whether more diverse region-specific samples improved detection.

The first ResNet-18 model, trained with basic parameters, achieved an accuracy of 72%. The confusion matrix showed that it correctly classified 29 out of 35 real images and 25 out of 40 fake images, with recall rates of 83% for real and 62% for fake images. Misclassification of fake images as real remained a key weakness.

A tuned ResNet-18 model, trained using the Adam optimiser and adjusted weights, improved performance to 87% accuracy, correctly identifying 27 real and 38 fake images. Recall for fake images rose to 95%, but recall for real images dropped to 77%, highlighting a trade-off between the two classes.

To address this and further test generalization, EfficientNet-B3 was evaluated. It outperformed both ResNet-18 models, achieving an accuracy of 89%. The model correctly classified 29 out of 35 real images and 38 out of 40 fake images, with precision for real images at 94% and recall for fake images at 95%. The ROC-AUC score of 0.974 indicated excellent discrimination between real and fake images. Grad-CAM visualisations confirmed that EfficientNet-B3 focused primarily on key facial regions such as the eyes, nose, lips, cheeks, and eyebrows, supporting the model's reliability. The results from the 500 images showed that region-specific training works better; however, the dataset size was still too small for deep learning models. In addition, the researcher aimed to test whether global datasets could perform well on region-specific evaluation, which motivated further analysis in the next phase.

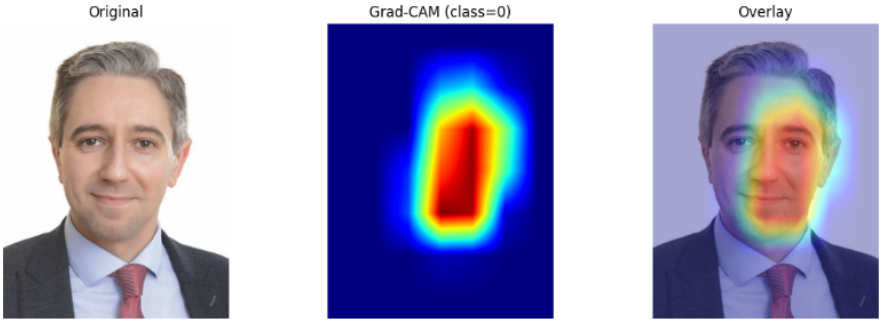

Figure 7: Grad CAM Visualization for Phase 2 EfficientNet B3

6.3 Experiment 3: Cross-Dataset (Phase 3)

One of the most important experiments in this study was to evaluate the effectiveness of global trained datasets on region-specific data. To test this, EfficientNet-B3 was trained solely on the globally available DFDC dataset and evaluated on 100 Irish media images. As expected, the model performed poorly when applied to unseen regional content, achieving only 50% accuracy, a weighted precision of 25%, weighted recall of 50%, F1-score of 33%, and an AUC-ROC of 46.87%. These results highlight the inability of models trained on global datasets to generalize effectively to local contexts.

This finding is consistent with prior research, where models that achieved strong results on familiar datasets (over 90% AUC) showed significant drops on new, unseen data. The cross-dataset evaluation in this study therefore reinforces the importance of including regional diversity in training data.

6.4 Experiment 4: Mixed Dataset Training (Phase 4)

To address the shortcomings observed in the cross-dataset evaluation, a mixed dataset approach was adopted. The DFDC dataset was combined with Irish images, ensuring that the training set captured both the diversity of global manipulations and the cultural and facial features specific to Ireland. This mixture of datasets was then split into training and testing sets to evaluate the benefits of hybrid training.

After retraining, EfficientNet-B3 achieved 74% accuracy, 79% precision, 74% recall, 76% F1-score, and 76% AUC-ROC. Compared to the cross-dataset setup, this represents a significant improvement in performance. The results demonstrate that including region-specific data not only enhances the detection of local deepfakes but also improves the model's ability to generalize across different types of images present in the DFDC dataset. The model does not overfit as it has enough data for training and testing.

In addition, the Precision–Recall AUC (PR-AUC) was calculated to provide further insight, especially given the imbalance between real and fake samples. While the overall PR-AUC for the mixed dataset was lower, analysis of the positive class (fake images) showed strong results: 92% for ResNet-18 and 93% for EfficientNet-B3. This indicates that both models were highly effective at correctly identifying fake images, even under class imbalance.

These findings directly answer the research question, showing that region-specific datasets are essential to improving the accuracy and robustness of deepfake detection systems.

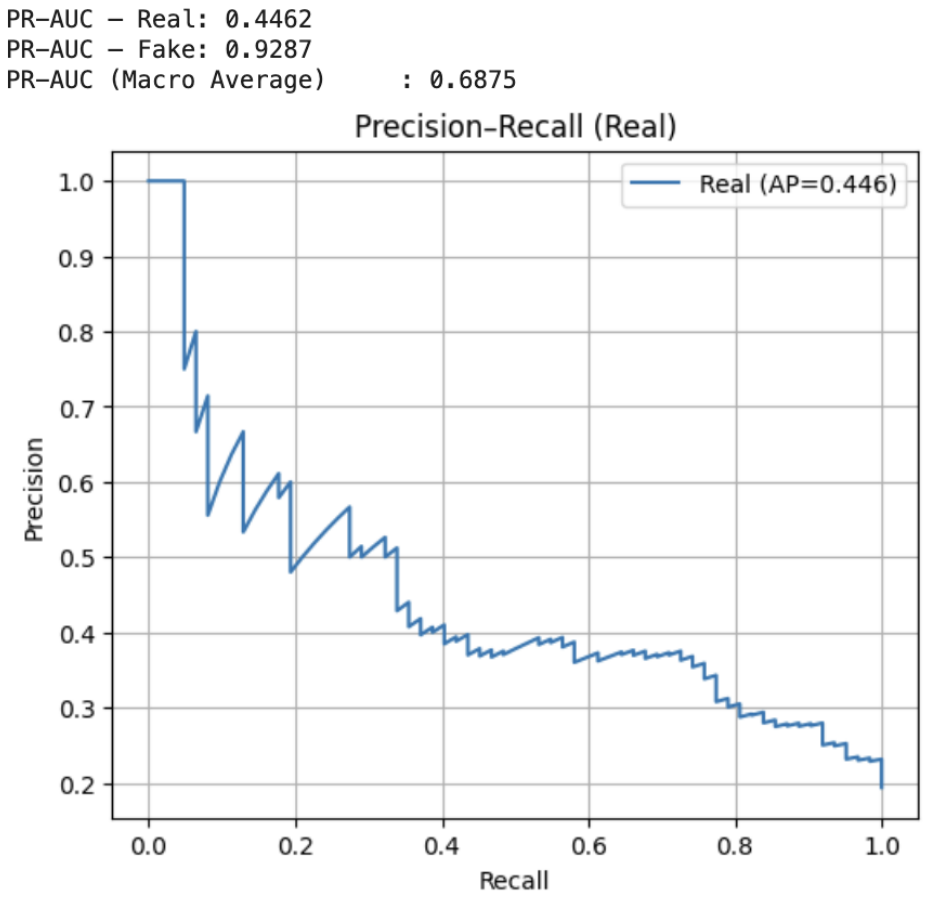

Figure 8: PR AUC for Mixed Dataset

6.5 Discussion

As shown in Table 3, performance varied significantly across the four phases. In Phase 1, ResNet-18 achieved promising results on a small proof-of-concept dataset, but overfitting was evident. Phase 2 demonstrated improvements with EfficientNet-B3, which achieved the highest accuracy (89%) and AUC (97%) on the expanded Irish dataset. Phase 3 highlighted the limitations of global-only training, with accuracy dropping to 50% when DFDC-trained models were tested on Irish data, confirming the lack of generalization. Finally, Phase 4 showed that combining global and Irish data improved generalization, reaching 74% accuracy and 76% AUC, proving the benefit of including region-specific data in training.

The evaluation revealed that **EfficientNet-B3** consistently outperformed both the baseline and optimised versions of **ResNet-18**. While ResNet-18 demonstrated solid performance, especially after optimisation with the Adam optimiser and weight adjustments, EfficientNet-B3 achieved higher accuracy, a stronger ROC-AUC score, and more balanced precision and recall across both real and fake classifications. These findings align with prior research (Raza et al.; 2022; Kwon et al.; 2021), which has shown that EfficientNet architectures, through compound scaling of depth, width, and resolution, are particularly effective for image classification tasks involving subtle visual cues, such as those found in deepfake detection.

One possible explanation for EfficientNet-B3's superior results lies in its architectural

Table 3: Comparison of model performance across phases

Phase	Acc.	P(R)	R(R)	P(F)	R(F)	AUC
Phase 1	72%	66%	83%	81%	62%	88%
Phase 2	89%	94%	83%	86%	95%	97%
Phase 3	50%	25%	50%	25%	50%	46.9%
Phase 4	74%	79%	74%	79%	74%	76%

design. By scaling all network dimensions in a balanced manner, it captures finer spatial details without significantly increasing computational cost. This ability is crucial when distinguishing deepfakes that may only differ from real images in minute facial regions, such as the eyes, mouth, or skin texture. In contrast, ResNet-18, while lightweight and less prone to overfitting on small datasets, may not capture these fine-grained details as effectively, especially when the manipulations are subtle.

Dataset characteristics also played a significant role in model performance. The curated Irish-specific dataset used in this study was balanced across gender (equal male and female representation) and content type (equal real and fake images), which likely contributed to the strong generalisation within this controlled domain. However, as highlighted in previous work (Padmashree and Karunakar; 2023; Mishra et al.; 2023), small datasets can also lead to overfitting, limiting the model's ability to perform reliably on unseen data, particularly if media quality or environmental conditions differ.

From a computational perspective, EfficientNet-B3 required approximately two minutes to train on a T4 GPU (Colab Pro) a relatively low overhead given its superior accuracy. This efficiency suggests it could be viable for deployment in real-world applications, particularly where detection must be performed on large volumes of content. However, when tested on CPU environments, inference speed decreased noticeably, indicating that deployment in low-resource settings would require optimisation or model compression techniques.

While the results are promising, several limitations must be acknowledged. The dataset size (400 images) remains small by deep learning standards, and the controlled nature of image generation (using Stable Diffusion prompts designed for Irish features) may not reflect the diversity and complexity of real-world manipulated media. Additionally, the study focused exclusively on static images, without incorporating temporal or audio features, which have been shown to enhance detection performance in multimodal frameworks (Liu et al.; 2024; Katamneni and Rattani; 2023). Finally, performance may degrade under challenging conditions such as low lighting, heavy compression, or intentional obfuscation.

In summary, this study demonstrates that an image-only, region-specific approach to deepfake detection can achieve high accuracy when supported by a carefully curated dataset. The findings suggest that EfficientNet-B3 is well-suited to this task, offering a strong balance of performance and computational efficiency. Future work should aim to expand the dataset to include more diverse visual conditions, explore multimodal architectures to capture additional cues from audio and temporal dynamics, and test the models on a wider range of real-world media to fully assess robustness.

6.6 Ethical Considerations

All data used in this study was collected in full compliance with the General Data Protection Regulation (GDPR) and broader ethical guidelines for AI research. Real and fake

media samples from Irish news outlets were obtained from publicly accessible articles and reports, where terms of service explicitly permitted reuse for research and educational purposes. Similarly, synthetic images generated using the *This Person Does Not Exist* platform are openly licensed for non-commercial research, and the Stable Diffusion model Rombach et al. (2022) was employed solely to generate artificial images with prompts tailored for this study. At no stage were private or personally identifiable materials collected, ensuring strict adherence to privacy requirements.

Each dataset was created, with images clearly labelled as either 'real' or 'fake' to maintain transparency and traceability. This aligns with best practices for dataset documentation, which emphasise clarity and accountability in AI datasets **?**. In addition, only publicly available, non-private content was included, reflecting a commitment to ethical sourcing that minimizes the risk of harm or misuse.

Finally, this research aligns with the principles of trustworthy AI outlined by the European Commission **?**, which stress transparency, accountability, and fairness. By designing the dataset specifically for Irish media, the study not only improves technical performance but also contributes to responsible AI development that respects cultural and legal contexts. These safeguards ensure that the work upholds both data protection laws and ethical integrity in AI research.

7 Conclusion and Future Work

7.1 Conclusion

The central aim of this research was to develop an Irish-specific dataset for deepfake detection and evaluate how region-specific training compares against global approaches. This objective was achieved through the creation of a carefully balanced dataset sourced from Irish media, synthetic samples from Stable Diffusion v1.5 Rombach et al. (2022), and realistic non-manipulated images from *This Person Does Not Exist*. The dataset was further complemented by the inclusion of the DFDC benchmark for cross-dataset testing, ensuring both local and global perspectives were examined.

The experiments produced three key insights. First, Phase 1 demonstrated that even a small Irish-only dataset allowed ResNet-18 He et al. (2016) to learn distinguishing features, although overfitting was apparent due to limited sample size. Second, Phase 2 showed the benefit of expansion, with EfficientNet-B3 Tan and Le (2019) achieving 89% accuracy and a ROC-AUC of 0.974, while Grad-CAM visualisations Selvaraju et al. (2017) confirmed the model's focus on meaningful facial regions. Third, Phase 3 revealed the clear limitations of global-only training: models trained exclusively on DFDC dropped to 50% accuracy on Irish test data, confirming that global benchmarks alone cannot generalize to unseen regional content. Finally, Phase 4 highlighted the strength of hybrid training, where mixing Irish and global datasets improved robustness, yielding 74% accuracy and strong PR-AUC scores for fake images.

Overall, the findings show that region-specific datasets substantially enhance the accuracy and trustworthiness of deepfake detection. They also underscore a broader implication: global models, when applied directly, risk underperformance in local contexts. This research not only delivers a proof-of-concept for Ireland but also provides a transferable framework that other regions can adopt to build culturally and geographically relevant detection systems.

7.2 Future Work

Building on the results of this study, several promising directions can be explored to further advance the detection of deepfakes in an Irish and broader European context:

1. **Expand the Dataset:** Increasing the dataset size to several thousand images would likely improve model robustness and generalisation. This expansion should include greater diversity in terms of lighting conditions, facial expressions, camera angles, age groups, and background environments. Collaboration with the Government of Ireland, media organisations, or research institutions could facilitate the acquisition of a larger, ethically sourced dataset. Data augmentation techniques, including rotation, colour jittering, and random cropping, could also be applied to synthetically increase dataset variability.

2. **Explore Multimodal Approaches:** While this project focused on image-only detection, future work could integrate audio-visual deepfake detection methods (Khalid et al.; 2022; Liu et al.; 2024). This would involve combining image-based analysis with audio processing to capture artefacts in speech, tone, and lip synchronisation. Incorporating audio alongside visual cues could significantly enhance detection accuracy, especially for video-based deepfakes where both modalities contain complementary clues.

3. **Enhance Model Architectures:** Future research could investigate transformer-based architectures, such as Vision Transformers (ViT) or hybrid CNN-transformer models, which have shown strong results in other image classification tasks. These models could be compared directly against the CNN-based baselines used in this study to determine whether they offer advantages in recognising subtle deepfake artefacts.

4. **Real-Time and Edge Deployment:** For practical adoption, especially by newsrooms, social media platforms, and content verification agencies, it will be essential to optimise models for real-time performance. Future work could explore quantisation, pruning, and other optimisation techniques to deploy detection systems on lightweight, portable devices without sacrificing accuracy.

5. **Towards a European Media Defence Framework:** Building on the success of this Irish-specific approach, future research could expand to create a pan-European deepfake defence system. This would involve developing a large-scale dataset representing the diverse facial features, languages, and cultural contexts found across the European Union, aligning with ongoing European initiatives on digital trust and security.

By addressing these areas, the work presented here could evolve into a more comprehensive and globally relevant solution, while retaining the cultural specificity that gives the Irish dataset its unique value.

References

Afchar, D., Nozick, V., Yamagishi, J. and Echizen, I. (2018). Mesonet: a compact facial video forgery detection network, *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7. [Accessed 13 Apr. 2025].

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. and Li, H. (2019). Protecting world leaders against deep fakes, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Chesney, R. and Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security, *California Law Review* **107**: 1753–1819.

Choi, S., Lee, H., Kim, H., Yoon, S. and Yoon, K. (2020). Detecting deepfakes with region-aware face discriminator, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2021–2030. [Accessed 13 Apr. 2025].

Dang, H., Liu, F., Stehouwer, J., Liu, X. and Jain, A. (2020). On the detection of digital face manipulation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5781–5790. [Accessed 13 Apr. 2025].

European Commission (2024). 6 tips to outsmart deepfakes, `https://ec.europa.eu/stories/6-tips/`. [Accessed 13 Apr. 2025].

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. [Accessed 13 Apr. 2025].
URL: *https://arxiv.org/abs/1512.03385*

Irish Independent (2024). RtÉ presenter sharon tobin targeted in deepfake investment scam. Available at: `https://www.independent.ie`.

Katamneni, S. and Rattani, A. (2023). A comparative study of fusion strategies for multimodal deepfake detection, *arXiv preprint arXiv:2304.09778* . [Accessed 13 Apr. 2025].

Khalid, H., Tariq, S., Kim, M. and Woo, S. (2022). Fakeavceleb: A novel audio-video multimodal deepfake dataset, *arXiv preprint arXiv:2108.05080v4* . [Accessed 13 Apr. 2025].

Kwon, P., You, J., Nam, G., Park, S. and Chae, G. (2021). Kodf: A large-scale korean deepfake detection dataset, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10778–10787. [Accessed 13 Apr. 2025].

Laurence, M., Patel, K. and Wong, S. (2022). Generalization issues in deepfake detection: A study on cross-dataset performance, *Journal of Visual Communication and Image Representation* **86**: 103534. [Accessed 13 Apr. 2025].

Liu, P., Tao, Q. and Zhou, J. (2024). Evolving from single-modal to multi-modal facial deepfake detection: A survey, *arXiv preprint arXiv:2406.06965* . [Accessed 13 Apr. 2025].

Matern, F., Riess, C. and Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations, *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* pp. 83–92. [Accessed 13 Apr. 2025].

Mishra, A., Soni, D. and Singh, R. (2023). Ensemble of machine learning classifiers for detecting deepfake images, *International Journal of Computer Science* **50**(4): 1–9. [Accessed 13 Apr. 2025].

Padmashree, B. and Karunakar, A. (2023). A comprehensive review on deepfake detection using machine learning and deep learning techniques, *Materials Today: Proceedings* . [Accessed 13 Apr. 2025].

Raza, A., Munir, K. and Almutairi, M. (2022). A novel deep learning approach for deepfake image detection, *Applied Sciences* **12**(19): 9820. [Accessed 13 Apr. 2025].

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695.

Say, J., Nguyen, T. and Park, H. (2025). Gan residual analysis for deepfake detection, *Proceedings of the International Conference on Computer Vision Applications*. [Accessed 13 Apr. 2025].

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. [Accessed 13 Apr. 2025].
URL: *https://arxiv.org/abs/1610.02391*

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks, *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114. [Accessed 13 Apr. 2025].
URL: *https://arxiv.org/abs/1905.11946*

This Person Does Not Exist (2024). Ai-generated human faces, `https://www.thispersondoesnotexist.com/`. [Accessed 13 Apr. 2025].

Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y. and Yuan, L. (2024). Df40: Toward next-generation deepfake detection, *Proceedings of the NeurIPS 2024 Datasets and Benchmarks Track*. [Accessed 13 Apr. 2025].