National College of Ireland

Detection of AI-Generated Images Using Multimodal Approach

MSc Research Project
Data Analytics

Nancy Saini
Student ID: x22236040

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nancy Saini
Student ID:	x22236040
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Vikas Tomer
Submission Due Date:	12/08/2024
Project Title:	Detection of AI-Generated Images Using Multimodal Approach
Word Count:	7685
Page Count:	22

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:	Nancy Saini
Date:	26th August 2024

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):	

Detection of AI-Generated Images Using Multimodal Approach

Nancy Saini
x22236040

Abstract

Generative Adversarial Networks have come with great challenges in image forensics, making it increasingly hard to distinguish an AI-generated image from an authentic one. In this work, therefore, a multimodal approach using Histogram of Oriented Gradients, Local Binary Patterns, Convolutional Neural Network with Support Vector Machines, and Logistic Regression is proposed for improving classification accuracy. The methodology combines various techniques of feature extraction, which are applied in a unique way to address the deficiencies of single-feature models in detection. On rigorous experimentation, while the SVM model delivered an accuracy of 81.12%, Logistic Regression went a notch higher, with an accuracy of 83.52%, thus outperforming several other existing models. The results were driven by an emphasis on the effectiveness of feature integration in capturing wide arrays of image artifacts for improving accuracy in detection. It points out the requirement of more diverse datasets and sophisticated feature extraction methodologies to further make these detection systems even more robust. Even though this research was focused on images produced by StyleGAN, future work shall be organized with datasets from several GAN architectures in order to increase generalizability and adaptiveness for detection models. Future studies should also aim at increasing the breadth of the dataset used and the adoption of hybrid methodologies so that more adaptability and applicability of the models to the real world would be very possible.

1 Introduction

1.1 Background

Generative Adversarial Networks (GANs) have become advanced in generating Synthetic Image Creation, produces near-real images. (Goodfellow et al.; 2014). proposed it way back in 2014. Applications of GANs range from image synthesis and data augmentation to artistic creation. While there is much to the credit of GANs, they have serious concerns attached to them about digital image authenticity since they can be used to mislead people with realistic fake images in scenarios such as misinformation or even identity theft. GANs involve two neural networks: one for generating synthetic images and the other for ascertaining the genuineness of the images generated by the generator. The adversarial process is bound to generate very realistic images that may turn out to be pretty challenging for the present detection systems. Because of these attributes, GANs have already started finding several applications besides image synthesis in medical

imaging and the entertainment sector. However, such possible misapplications of GANs underline the requirement for robust methods of detection.

1.2 Importance

It has become very critical to ensure digital content authenticity to maintain trust in digital media, social networks, and other online platforms. Therefore, detecting GAN-generated images becomes very important in digital content integrity, forensic analysis, and security improvement. Although, many detection methods have been explored. However, there is still a scope for some hybrid approaches that can be robust and will have high accuracy, and flexibility across a wide range of GAN architectures. Missing detection of GAN images can eventually turn into social media misinformation, false evidence in courts of law, and an infringement of journalistic ethics. In this regard, developing effective techniques for detection is important. Most conventional image forensic techniques leverage handcrafted features coupled with statistical analyses that are weak against modern GAN techniques. The rapid development of GAN architectures enforces continuous adaption of the detection methods to include new types of synthetic images. Besides, the high quality of images generated by GAN especially GANs like StyleGAN requires sophisticated techniques of detection that would allow differentiating these from real photos.

1.3 Research Question and Objectives

The primary research question guiding this study is:

> *How effective is a multi-modal approach combining Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM) in detecting GAN-generated images?*

This paper focuses on the integration of various feature extraction methods to enable the detection of highly realistic AI-generated images. The research is guided by the objective of harnessing the strengths of each method in improving the accuracy and resilience of detection. Specific to this study are the following objectives:

1. **Assess the contributions of LBP, HOG, and CNN-based features in GAN-generated image detection. Check their effectiveness in the identification of synthetic imagery.**
2. **Assess the performance improvement from combining these features, evaluating whether a combined feature set captures a wider range of image artifacts than any single method.**
3. **Evaluate the performance of the model on a dataset of images generated using a GAN. The effectiveness of the model can be defined by accuracy, precision, recall, or any other relevant measure.**

1.3.1 Novelty and Contribution

It proposes a novel method for detecting GAN-generated images with the help of an integrated SVM-based model driven by different feature extraction techniques: LBP,

HOG, and CNN. The novelty lies in the fact that this technique is built to leverage very different strengths coming from these different methods of feature extraction in a bid to improve both accuracy and resilience in detection. These features are concatenated to include all varieties of image artifacts. As a result, this method does not face the limitations of any one technique. This work will enrich the current methodologies and provide with a more potent solution for detecting synthetic images. It will also provide insight into how robust detection systems in the area of digital media forensics can be built.

1.4 Limitations

This study has some limitations to its scope. First, the quality and diversity of images generated by GANs significantly affect the performance of the model. Computational resources come in at a second place, which is heavily required during the training of complex models like CNNs. The current work mainly focuses on the images that well-known GAN architectures like StyleGAN generate. Findings might need to be generalized or validated across other variations of GANs. It is possible that publicly available datasets do not hold all the varieties of GAN images existing in real life. While the ensembled approach provides improved detection capabilities, the integration of several feature extraction methods and classifiers adds computational overhead, which potentially limits its applicability in every use. This fast-paced technology of GANs is hard to match. With new architectures being brought out into the open, they might beat the current detection techniques. It is by these challenges that a continuous search in the improvement of detection models ensues, and ethics considerations ensure responsible use of the technology.

This report is structured as follows:

- **Related Work**: Discussion of existing methods for detecting GAN-generated images and their strengths and limitations.
- **Methodology**: Describes the data collection, preprocessing steps, feature extraction techniques, and model training process.
- **Design Specification**: Outlines the design constraints and quality attributes of the system.
- **Implementation**: This section describes how the multi-modal approach is implemented.
- **Evaluation**: Presents the experimental setting, evaluation metrics, and results of the model.
- **Conclusion**: This part will summarize the main contributions and state the recommendations for the future work.

Literature review critically evaluates any previous work and provides a context for the current study. The methodology section explains the design and techniques that shall be employed in the research. The design specification involves the identification of constraints and attributes that help guide the development of the system. The implementation shows the practical application of the proposed methods. In the evaluation section, results are analyzed together with key findings summed up and suggestions for further work.

2 Related Work

This section puts the current study in the academic literature by providing a critical review of similar work. The review entails the critical analysis of strengths and weaknesses associated with different approaches towards detecting GAN-generated images. In order to do this, nearly 25 papers were reviewed; however, 14 have been discussed in this chapter.

The breakthrough of GANs in image synthesis has set enormous challenges for image forensics. Accordingly, with the main emphasis of this paper being detection of images generated by GANs, many methods have been proposed by researchers to tackle this problem, and they have succeeded to some level.

2.1 Foundational and Traditional Approaches

The concept of GANs was first coined in Goodfellow et al. (2014), which established a foundation for this rapid development in the field (Goodfellow et al.; 2014). Within the framework of the GANs, there are two neural networks: a generator and a discriminator, pitted against each other in a game-theoretic scenario. In this seminal piece of work, many follow-ups and applications on image synthesis, data augmentation, and so forth have been derived.

The techniques applied in detecting GAN-generated images were originally based on traditional forensic techniques. Deep learning has changed this dramatically. That is, in the framework of a full review, techniques for the detection of images generated by GANs have been shown by way of proving the effectiveness of CNN combined with Benford's Law to assure an accuracy rate of 98.95% to 99.99% on different datasets for GANs (Kit et al.; 2023). Although these results are very promising, the threat from newer GAN architectures is ever-present. Hence, there is a continuous updating of the training datasets for the sake of robustness. This exposes an important limitation of the traditional methods due to their inability to keep up with the evolving models of GAN—a core challenge that the research tries to bridge by putting in place adaptive techniques.

Transitioning from traditional approaches, researchers like Sharma et al. (2023) have recently set up traditional forensic techniques against state-of-the-art deep learning models. Their study underscores the computational intensity of deep learning models but also highlights their superior accuracy. This creates the need for lightweight models that would sustain a high accuracy at a reduced computational cost (Sharma et al.; 2023). The insight is going to be very important for the research, intended to balance accuracy and efficiency with a multi modal approach.

In a novel approach, Monkam et al. (2023) proposed the Generative Joint Bayesian Optimal Detector GAN (G-JOB GAN), encoding a Bayesian framework with a joint optimization strategy. Although this model reached an accuracy of 95.70%, the model's complexity and computational requirements showed that designs more tractable are still required (Monkam et al.; 2023). This complexity suggests that scalable solutions are urgent; The multimodal approach that this author tries to solve is the detection process with hybrid techniques that simplify it and make it more efficient.

Exploring some deep learning approaches, Nataraj et al. (2019) combined co-occurrence matrices on RGB channels with a deep CNN framework, achieving high accuracy rates for CycleGAN and StarGAN. But, their method is sensitive to image manipulations like compression and resizing which higlights the necessity for robust preprocessing techniques

(Nataraj et al.; 2019). This points out the importance of prepossessing in detection systems, a factor that this research integrates into its methodology to enhance resilience to manipulations.

Building on this, Mandelli et al. (2022) have examined orthogonal training of multiple CNNs, with patch-based score aggregation. This new strategy reached practically perfect accuracy under some settings, but it also emphasized the growing demand for more efficient training strategies to master the accruing complexity and resource consumption (Mandelli et al.; 2022). Their work therefore highlights the need for efficient training strategies, which this study wants to improve by resorting to simpler models combined to achieve full detection capacity.

2.2 Advanced and Hybrid Approaches

In the realm of multi-spectral satellite images, Abady et al. (2024) utilized a Vector Quantized Variational Autoencoder 2 (VQ-VAE 2) trained on pristine images to detect GAN manipulations. Their method achieved a very high precision of 0.93, underpinning the strength of one-class classifiers but also their low generalizability and large computational costs (Abady et al.; 2024). That is the search for generalizing classifiers across domains to which this research aspired to make its contribution in cross-domain detection.

Further enhancing detection techniques, Tan et al. (2023) proposed the Learning on Gradients (LGrad) framework, converting images to gradients using a pre-trained CNN model. While this method showed high precision and recall, its computational intensity necessitates further optimization (Tan et al.; 2023). The need for computational efficiency is a critical consideration for my research, which aims to streamline detection without sacrificing accuracy.

Such hybrid innovative approaches, like that by Fu et al. (2022), integrated LBP for texture analysis and SPAM for sensor noise within an SVM classifier. This hybrid approach reached as high as 97.60% in accuracy, although it underlined high-quality sensor data requirements and high computational intensity(Fu et al.; 2022). The potential of this technique in combining texture and noise analyses is clear, and the multimodal approach proposed by this author incorporates to improve detection accuracy and reduce dependency on data quality.

Arora and Arora (2022) explored the possibility of using GANs for the generation of synthetic medical data that mimics trends and characteristics from real patient data. Although there are significant benefits related to privacy that synthetic data brings—ethical considerations aside—the large computational requirements call for robust metrics that will capture both fidelity and anonymity (Arora and Arora; 2022). This research will extend the scope of the GAN detection technique by applying it to a more general and diversified methodological framework.

Adding to recent advancements, Chi Liu et al. (2023) proposed a methodologically different approach in their work, "Towards Robust GAN-Generated Image Detection: A Multi-View Completion Representation." Their new framework is then powered by a multi-view completion strategy that confers a great deal of boosted robustness to the detection systems against advanced GANs like StyleGAN, which typically generate highly realistic and high-quality images. Their approach has the advantages of both high accuracy and a solution scalable for new GAN architectures by fusing multiple data views and using a complex completion task that exploits discrepancies in the image generation process(Liu et al.; 2023). This confirms the problem of adaptability in detection systems

and underlines the appropriateness of the multimodal strategy in this research.

It continued with the contribution in this area by Wang et al. in (2022), with the use of an innovative method in augmenting GAN-generated image data in the fingerprint domain: "General GAN-Generated Image Detection by Data Augmentation in Fingerprint Domain." This means that fingerprint perturbation of a fingerprint domain can increase the generalization ability of GAN detectors, and cross-GAN detection performance could be enhanced considerably with adjustment by different GAN fingerprints. Our results showed significant improvements in mean accuracy and average precision over existing state-of-the-art methods, which further gave credence to the potential of the proposed approach in solving the restrictions introduced by the unseen GAN models (Wang et al.; 2022). This is in line to improve the generalizability of detection in research with an emphasis on novel GAN architectures.

Besides, a transfer learning-based framework was proposed by Zhang et al. (2023), which enhances detection through interleaved parallel gradient transmission between two neural networks. This provided considerable performance improvement and reached the best metric of 99.04% accuracy, attesting to its excellent generalization capabilities(Zhang et al.; 2023). The use of transfer learning for enhancing generalization will be very relevant to this research goal since adaptive models play a major role in learning.

Combining these innovative methods with traditional approaches, researchers have achieved substantial progress. Martin-Rodriguez et al. (2023) further showed that pixel-wise feature extraction with PRNU and ELA, combined with CNNs, makes it possible to accomplish high accuracy and precision in the detection of AI-created images (Martin-Rodriguez et al.; 2023). This further supports the power of hybrid feature extraction techniques central to this research's multimodal detection strategy.

2.3 Novelty and Contribution

While most conventional forensic techniques and state-of-the-art deep learning models have already been applied to GAN-generated image detection, they mostly bear high computational costs, poor generalizability, or even high sensitivity to image manipulations. Most of them cannot keep pace with the rapid development of GAN technology.

The present research envisions a new fusion of LBP, HOG, CNN, and SVM. In view of the above, this new combination is regarded as filling the gaps, by discussing hybrid techniques that seldom have been pursued in the literature. Their synergistic application has greatly improved the detection efficacy, extended generalizability, and reduced susceptibility to common image manipulations in the respective field.

Beyond improving the accuracy in detection, this project has enormous impacts on digital image forensics through computational performance optimization. This work is the basis for future innovation in detecting synthetic media—an area of development that is critical, as GAN technology is never good enough and always improving.

2.4 Summary of Reviewed Studies

The following table summarizes the methodologies, findings, and future directions from the studies presented, which helps in quickly referring to the state of the art in GAN-generated image detection research.

Table 1: Consolidated Overview of Studies on GAN-Generated Image Detection

Authors	**Year**	**Datasets Used**	**Methodology**	**Model Used**	**Metrics**	**Value**	**Limitations**	**Future Work**
K. S. Kit et al.	2023	Various GAN images	CNNs and Benford's Law	CNNs	Accuracy	**98.95% to 99.99%**	Struggles with advanced GAN models	Include advanced GAN architectures
Preeti Sharma et al.	2023	Diverse forgery datasets	Forensic and deep learning methods	Forensic methods, CNNs	Accuracy	**98.95% to 99.99%**	High computational demand	Develop efficient models
G. Monkam et al.	2023	CelebA, 200,000 images	Bayesian optimization in G-JOB GAN	G-JOB GAN	Accuracy	**95.70%**	Needs scalability	Optimize for real-time applications
L. Nataraj et al.	2019	CycleGAN, StarGAN images	Co-occurrence matrices with CNN	CNN	Accuracy	74.5% (high quality image)	Sensitive to image modifications and resizing	Improve robustness to image changes
S. Mandelli et al.	2022	Real and synthetic images from StyleGAN	Orthogonal CNN training	EfficientNet-B4	AUC	**Up to 0.9999**	Orthogonality issues	Improve training strategies
L. Abady et al.	2024	Multi-spectral satellite images	VQ-VAE 2 with loss	One-Class Classifier	Precision	**0.93**	Limited to specific image types only	Extend to other image types
C. Tan et al.	2023	Mixed datasets including ProGAN, StyleGAN	Gradient conversion using CNNs	Gradient-based CNNs	Precision	**0.92**	Very High computational intensity and uses single extraction.	Optimize gradient computation
T. Fu et al.	2022	100,000 images, real and GAN-generated	LBP and SPAM via SVM	Hybrid SVM	Accuracy	**Up to 97.60%**	Requires high-quality data for good performance	Combine with deep learning
A. Arora and A. Arora	2022	Various types of medical images	GANs for synthetic data	GANs (StyleGAN2-ADA)	Not applicable	N/A	Ethical and privacy concerns	Develop quantitative metrics
D. Gragnaniello et al.	2021	LSUN, ImageNet, COCO	Systematic experimental study	Pre-trained CNNs (Xception, Inception)	AUC Accuracy	$>$ 0.9 for low-resolution images, $>$ 90% for high-resolution images	Depends on image quality	Improve generalization to new GANs
F. Martín-Rodríguez et al.	2023	459 AI-generated and real photographs	Pixel-wise feature extraction with CNNs	CNNs using PRNU and ELA	Accuracy	**PRNU:0.95 ELA:0.98**	Issues with non-JPEG images	Include other image engines
Chi Liu et al.	2023	ProGAN, CramerGAN, SNGAN, MMDGAN, StyleGAN, StyleGAN2	Multi-view completion strategy	Multi-View Completion Classification Learning (MCLL)	Accuracy	**100%**	Overfitting on Unstable Features	Robustness Against Perturbations Simplify model for broader deployment
Huaming Wang et al.	2022	ForenSynths dataset with images from ProGAN, StyleGAN	Data augmentation in fingerprint domain	ResNet50	Accuracy	**Improved mean Accuracy by 7.0%**	Dependency on perturbation strategy	Explore other domains like video GANs

3 Methodology

This section provides a comprehensive account of the methodology employed in this research, detailing the systematic procedures used to detect AI-generated images using a multi-modal approach. The methodology is structured to ensure replicability and verifiability by other researchers.

3.1 Research Setup and Environment

The research was conducted on a personally used laptop running a 13th Gen Intel(R) Core(TM) i7-13700H processor at 2.40 GHz, 16 GB RAM, and a 64-bit operating system having a x64-based processor, running Windows 11 Home, Version 23H2. All experiments were conducted within Jupyter Notebook, which allowed iterative development and testing due to its interactive environment that supports real-time code execution with visualization. This codebase was developed in Python 3.8, and for the image processing part, OpenCV was used; for machine learning, it used Scikit-learn; for deep learning, TensorFlow/Keras; for data manipulation, NumPy and Pandas; and lastly, for the visualization part, Matplotlib and Seaborn. All this setup helped much to deal efficiently with the data and run complex algorithms needed for the goals of the study

3.2 Data Collection and Preparation

The study used Kaggle dataset of 140,000 images, where half of them were real and half of them were fake. Originally, the real images came from Nvidia and were collected from the Flickr dataset consisting of images of human faces in different conditions. The fake images were generated through Bojan's StyleGAN model, very similar to real human faces. Such a balanced dataset contributed to very important training in enabling the model for generalization onto unseen data, and it also helps in distinguishing between real and synthetic images using a robust, efficient system. The dataset was managed through CSV files that has metadata for each image, which supported data management across different subsets such as training, validation, and test phases. These files contained key attributes including the **Original Path** (the location of the image before preprocessing), **ID** (a unique identifier for each image), **Label** (a binary indicator, where '0' represents fake and '1' represents real, to signify image authenticity), **Label_Str** (the string representation of the label, either 'real' or 'fake'), and **Path** (updated post-processing path reflecting changes due to preprocessing or restructuring). Programmatically updated to match with the directory structure of the local setup, the path column facilitated efficient data loading and preprocessing, ensuring seamless access and processing of images during training and evaluation phases.

First, **stratified sampling** was done on the data to create subsets for training, validation, and testing. Each subset contained equal numbers of classes so as to be completely representative of the sample population. First, it used a small subset of 5,000 images (2,500 real, 2,500 fake) to test the pipeline and model performance, and this smaller set allowed making changes and iterations quickly. Then, it used a test set of 10,000 images, with 5,000 real and 5,000 fake, for a more accurate evaluation. In splitting, Python's stratified sampling ensured that this would be random and consistent concerning class balance.

3.3 Image Preprocessing

Preprocessing is one of the most important steps that help standardize the dataset for feature extraction. All images were resized to a size of 128 x 128 pixels to provide reduced computational complexity while retaining all the facial features. Conversion into grayscale, which helped concentrate on intensity variation, reduces dimensionality, helping in adopting some of these techniques in feature extraction. The pixel values were further normalized by dividing them by 255 to put the data in the range [0, 1]. This makes the data homogeneous and also provides improved convergence for the data during training. Further, a Gaussian blur with a kernel size of (3,3) was applied on every image to reduce noise and showing main features so that robust feature extraction is possible.

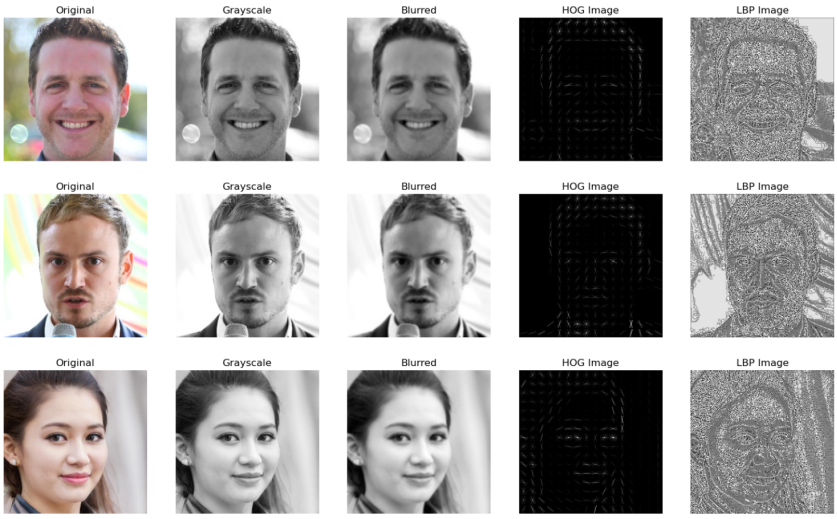

Figure 1: Representation of preprocessing steps applied to an image. From left to right: Original, Grayscale, Blurred, HOG Image, and LBP Image.

These preprocessing steps prepare the data into a consistent format required by the subsequent steps: analysis and modeling. Resizing maintains uniform input size through all the images, hence feature extraction and modeling can be conducted uniformly over the images. Grayscale conversion reduces the computational overhead of color information toward concentration on texture and intensity information, which are critical means of real-versus-fake image classification. Normalization gets pixel values on identical scales in an attempt to get to convergence during model training more quickly.

3.4 Feature Extraction Techniques

Feature extraction can be used to transform raw data into a set of features that can be effectively used by the machine learning algorithm. In this work, three feature extraction techniques will be applied, which are Local Binary Patterns, Histogram of Oriented Gradients, and Convolutional Neural Networks.

LBP is a simple yet efficient texture descriptor, which encodes the patterns of intensity differences around each pixel into a binary string and builds a histogram of those patterns. The intensity of each pixel is compared with its eight surrounding neighbors. These binary values are combined to form an 8-bit number. It builds up a histogram of the frequency of each LBP code within the image, which thereafter gives a strong descriptor of texture.

LBP has high computational efficiency and rotation invariance that can be applied in the analysis of textures within images across different orientations. This has a radius of 3, with points equating to 8 times the radius.

HOG describes the distribution of gradient orientations in small parts of an image, thus emphasizing edge structures. This algorithm first breaks the image into small, connected regions called cells and, for each cell, computes a histogram of gradient directions. The descriptor is just the concatenation of these histograms. Parameters: pixels per cell set at (8, 8) and cells per block at (2, 2). HOG focuses on the shape and structure of facial features, which are very important in differentiating real from fake images. It captures very well the local edge information that is robust to changes in illumination and shadowing.

CNNs are deep learning models that perform automatic feature learning from raw image data at high levels. An already pre-trained VGG16 model was used for the extraction of features from images; high-level convolutional features will then be applied without the classification top layer. The convolutional layers extract a complex pattern and hierarchy of features in a way that projects subtle differences between genuine and fake images. CNNs are used for learning abstract features, hence they are very successful on tasks dealing with image data, mostly due to their capability of capturing spatial hierarchies.

Each image was processed through LBP, HOG, and CNN feature extraction, and all of them were used together for training and evaluation. Combinations of such very diverse and different techniques of feature extraction improve the ability to capture comprehensive image artifacts.

3.5 Model Training and Evaluation

In this research, a well-structured model training and evaluation process was applied to ensure the effectiveness of this approach in classifying real and AI-generated images.

In PART A of the implementation, all features which were extracted using the three extraction techniques were concatenated into one feature vector per image. This approach was applied to obtain the appropriate strengths of each method and, consequently, build an all-inclusive feature set that enables the detection of even minute details across different dimensions in image data. After that, an SVM with a linear kernel was used because this algorithm works very well in very high-dimensional spaces and performs excellently in binary classification tasks. That would enable it to use the combined features set for training an SVM with complementary strengths of LBP, HOG, and CNN features. The main aim was the exploration of the holistic approach to the integration of multiple feature extraction techniques into a robust classification.

Part B investigated each technique in feature extraction on its own so that it would set out different contributions for the task of classification. By doing so, evaluation for each type of feature gave information about LBP, HOG, and CNN features on their own, showing the strengths and weaknesses. The author could assess the contribution of each feature set by training individual SVM classifiers on LBP, HOG, and CNN. This gave information about how good each feature type could be and which of the features' contribution in differentiating real images from fake images is maximum.

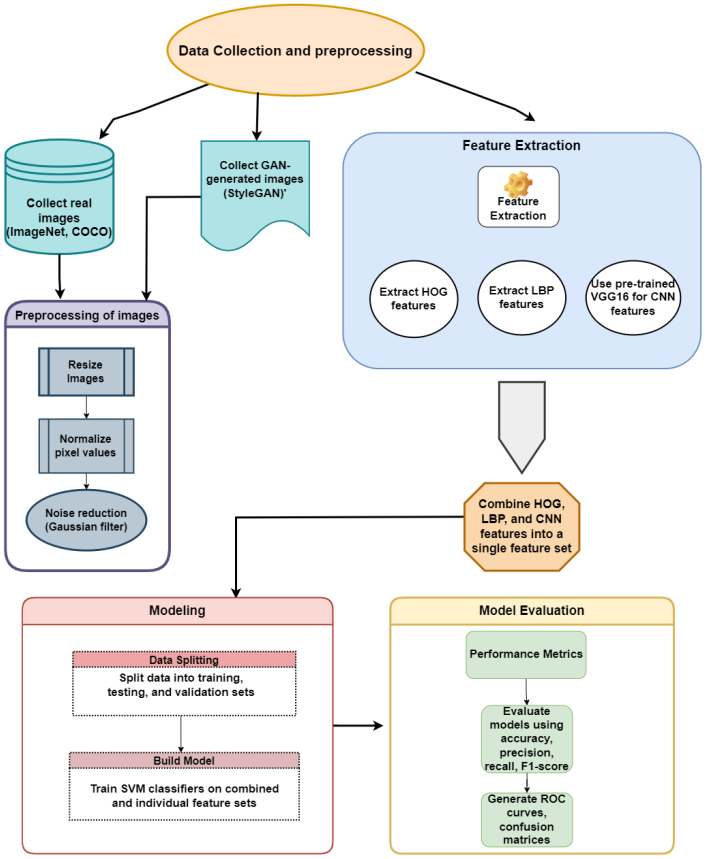

Figure 2: System architecture for GAN-generated image detection, illustrating the full workflow from data collection to evaluation

As an additional exploratory method, logistic regression was applied to the combined feature set to provide a baseline for comparison against the more complex SVM models. The model fitted with 5000 maximum iterations and L2 regularization to avoid overfitting helped in assessing the linearity of the features. It gave a notion of how the SVM model could still be further optimized and hence showed the need for more advanced classification techniques to deal with complex data distributions.

3.6 Evaluation Metrics

Model performance was assessed based on accuracy, precision, recall, the F1-score, and the area under the ROC curve. Accuracy is the proportion of correctly classified instances against total instances, and it gives a measure for model performance. Precision is a measure of the number of true positive results against that of positive results predicted by the classifier; that is, the proportion of actual positives among predicted positive instances. Recall is the ratio of true positives to the number of positives that should have been returned, measuring a model's ability to get all relevant instances. The F1-score is the harmonic mean of precision and recall and hence class-balance-aware. AUC gives a notion of the model's performance in class differentiation at various thresholds and thus gives a feel for the sensitivity-specificity trade-offs.

4 Design Specifications

This section provides information about the architecture and framework developed to detect AI-generated images using various feature extraction techniques and their classification. The system has been designed robustly to distinguish real images from synthetically generated ones, keeping in mind the basic factors of scalability and maintainability. Figure 3 illustrates a flowchart showing the summary of the system architecture and process flow.

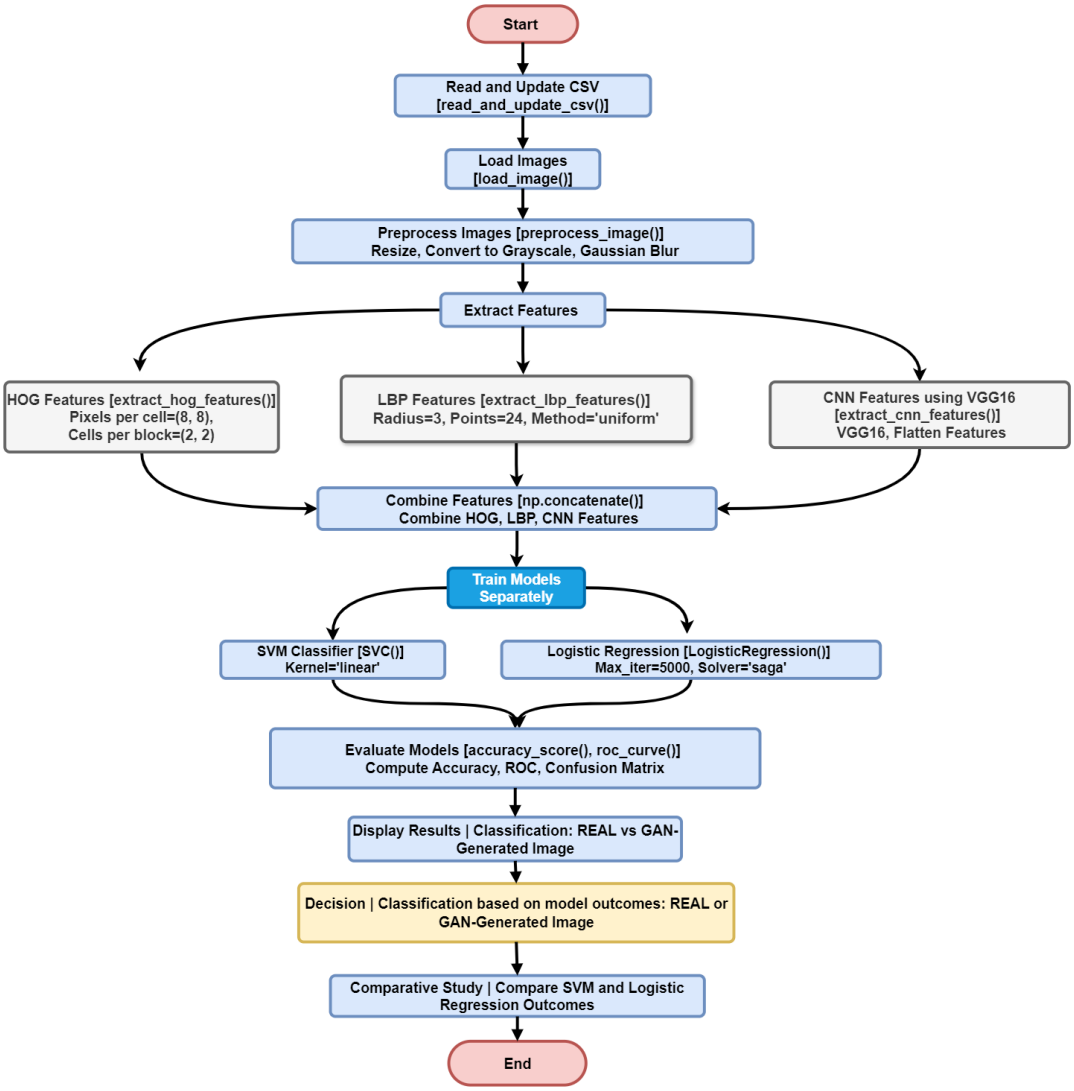

Figure 3: Process flow for feature extraction, model training, and evaluation.

4.1 System Architecture

The system applies a multimodal system that combines HOG, LBP, and CNN for feature extraction, whose features are afterward used with SVM classifiers for image classification. This approach enhances the system's ability to analyze and interpret complex image data effectively. The system automates feature extraction from input images through HOG, LBP, and CNN techniques to capture the essential characteristics of images for classification. After extracting features, an SVM classifier is trained using these features

and tested on its performance with a dataset to ensure its reliability and accuracy.

For high-performance needs, the minimum accuracy rate targeted by the system should be 80% for both real and synthetic image datasets. The design allows the completion of image processing and classification to take just a few seconds per image, facilitates quick throughput, and features optimized resource usage in order not to let peak operations consume more than 80% of the CPU and 4GB RAM. While currently no Graphical User Interface is implemented, it envisions the development of a simple interface for easy image uploading and handling, together with API access to be able to integrate with other digital forensic tools and systems.

4.2 Algorithmic Specifications

The system incorporates several algorithmic specifications:

- **Support Vector Machines (SVM):** Used to classify images based on extracted features, effectively separating high-dimensional data Cortes and Vapnik (1995).

- **Histogram of Oriented Gradients (HOG):** Utilized to detect edges and gradients, providing crucial shape and texture information Dalal and Triggs (2005).

- **Local Binary Patterns (LBP):** Applied to capture texture variations in images, aiding in the distinction between real and synthetic faces Ojala et al. (2002).

- **Convolutional Neural Networks (CNN):** Deployed to automatically extract and learn complex features from images, enhancing the classification process.

These algorithms form the core of the system, facilitating robust feature extraction and classification.

4.3 Design Constraints and Quality Attributes

This system can work on both Windows and Linux environments and follows all rules of data protection strictly, where personal data is kept only for as long as the processing takes. Scalability is also targeted as one of the quality attributes, where architecture supports the easy addition of new methods for feature extraction and model types. Its modular design, with extensive documentation, will also enhance maintainability, thus helping in the continuous updating process and maintenance of the system. It is a detailed design specification for the delivery of a system that will not only be functional and efficient but also adaptable with further developments in image processing and machine learning technologies.

5 Implementation

In the last implementation phase, this research focused on the integration of the developed models with applications into effective ways of differentiating real from GAN-generated images. This phase consisted of a series of tasks done rather carefully in a way that contributed collectively to the success of the project.

5.1 Transformed Data

First, the implementation transformed raw image data in order to prepare it for feature extraction. This was critical to standardize the dataset in order to enrich learning within the model. Thereafter, every image underwent preprocessing steps to make the dataset

homogeneous. First, normalization adjusted pixel values to a consistent range between 0 and 1, which actually helped the model stabilize learning and greatly increased the speed during training. After normalization, the images were resized to a standard size of 128x128 pixels, ensuring that each image has a regular size, which is a precondition for effective feature extraction. Gaussian methods were applied to blur the images, reducing noise and smoothing textures, avoiding model sensitivities for small variations and focusing on key features of the image. This improved the robustness of feature extractors by focusing model attention on the most relevant features.

5.2 Developed Models

In the implementation phase, machine learning models were created due to advanced feature extraction techniques and classification algorithms invented. The LBP captured the patterns of the differences in intensity to identify subtle textural variations between real and synthetic images. The HOG detected edge orientations and geometric structures critical for differentiating shapes and contours. Transfer learning helped leverage the benefits of CNN, especially the pre-trained VGG16 model for abstracting patterns and complex spatial hierarchies within the images through its convolutional layers. Its feature set was trained on an SVM classifier owing to its characteristic capabilities to deal with high-dimensional data and construct a hyperplane that maximizes the margin between classes. Binary classification tasks were used in attempts to optimize models of SVMs, seeking intricacies across multiple dimensions in images. Logistic regression provided a baseline model against which to compare and gave a sense of the linear separability for this dataset. In this way, the dual-model approach helped to perform comparative analysis, showing the virtues and possible deficiencies in the SVM framework.

5.3 Code Implementation

The models were implemented in Python 3.8, which provides a robust environment in dealing with complex data processing tasks. This project used major libraries like TensorFlow and Keras for implementing CNNs that enable efficient deep learning model training and integrate them together. These very pre-trained models, particularly VGG16, could be used with ease in the TensorFlow ecosystem for transferring the learned features to the task at hand, while not being as computationally resource-intensive. As for the SVM and logistic regression models, scikit-learn would provide a full set of tools for the training and evaluation of the models, including their hyperparameter tuning by cross-validation. NumPy and Pandas would handle data handling and preprocessing, therefore playing an important role in efficiently manipulating large datasets and ensuring a smooth flow of data through the processing pipeline. These libraries and tools helped much in completing the project successfully. Using these tools, it was possible to train different models that would tell the difference between real images and GAN-generated images. Thus, these tools and their effectiveness in the detection task are proved by thorough testing and optimization. The actual implementation ended with models far above the set accuracy benchmark, turning out to be adaptable to further fine-tuning and extension if need be.

6 Evaluation

In-depth evaluation of all experiments carried out in this study investigates not only individual feature effectiveness but also the strength of a multimodal approach. This is shown through a detailed model performance metric analysis for each feature in terms of confusion matrices and ROC curves, followed by their comparative analysis with a multimodal approach.

6.1 Experiment 1: Evaluation of Individual Features

In the case of Experiment 1, three different types of features are used independently: Histogram of Oriented Gradients, Local Binary Patterns, and Convolutional Neural Networks, all combined with Support Vector Machine classifiers. This experiment will measure their ability to differentiate between 'Real' and 'Fake' images.

6.1.1 Case A: SVM with HOG Features

The Histogram of Oriented Gradients descriptor ensures that gradient and edge information is mapped, information highly useful in object detection having distinct shapes and outline boundaries. The accuracy obtained on an SVM model trained on HOG features was 78%, with precision and recall both at 78%, thus giving an F1-score of 0.78.

Looking at this confusion matrix, it is obvious that with HOG features, the SVM model is pretty effective: It has correctly classified images that are 'Real' by catching edges and contours uniquely. But 268 'Fake' images have been misclassified as 'Real', thus proving some confusion with less defined features. Moreover, 968 'Fake' images were correctly identified, and 282 'Real' were labeled as 'Fake'; therefore, the model seems to be having trouble identifying the subtle features of real images.

The receiver operating characteristic curve of the HOG features, shown in Figure 5, fetches an area under the curve of 0.78 approximately, thus revealing class-moderate discrimination.

6.1.2 Case B: SVM with LBP Features

The LBP features extract information about the texture by analyzing the patterns of local pixels. In this experiment, LBP features produced an accuracy of 59.2% with precision, recall, and an F1-score all at 0.59.

The confusion matrix at 4 for the SVM model using LBP is given below. In contrast with HOG, it performed poorly. While it rightly identified 718 'Real' images, it wrongly classified 488 'Fake' images as 'Real', which means that LBP, with its features on local texture pattern extraction, has a shortage of features in complex images. In addition, 762 'Fake' images were rightly identified, whereas 532 'Real' images were labeled 'Fake', which simply means that LBP has failed to extract fine patterns needed in classifying 'Real' and 'Fake' images with a high degree of accuracy.

The AUC value for LBP features was approximately 0.59, thus showing only a limited ability to distinguish classes, particularly in the case of complicated textures.

6.1.3 Case C: SVM with CNN Features

CNN features, which are derived from a pre-trained convolutional neural network, provide meaningful hierarchical features that extract both low-level and high-level patterns. Specifically, after the inclusion of CNN features, accuracy was raised to 71.4% with precision, recall, and F1-score at 0.71 for the SVM model.

The confusion matrix indicates that CNN features provide better balancing in classification. For instance, the model managed to classify 867 'Real' images correctly, thus prove the efficiency of CNN in extracting features through its multiple layers. On the negative side, it misclassified 332 'Fake' images as 'Real' and managed to identify 918 'Fake' images correctly. Notwithstanding these strengths, 383 'Real' images were labeled 'Fake', thereby indicating scope for improvement in capturing weak genuine features amidst noise.

The AUC for CNN features as shown in 5 was only around 0.71, which obviously indicates stronger discriminative capability compared to LBP and extremely close performance compared with HOG.

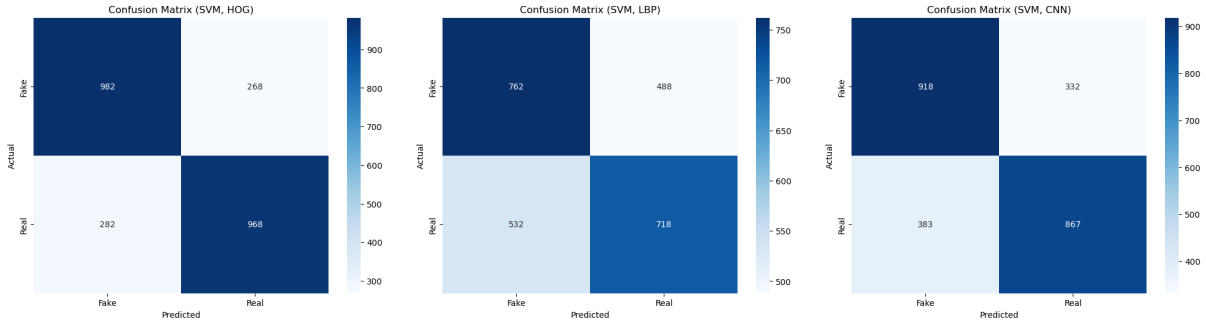

Figure 4: Confusion matrices for SVM with HOG, LBP, and CNN features.

Figure 4 illustrates the confusion matrices for SVM models using each feature type, visually showing classification accuracy and the balance between true and false predictions.

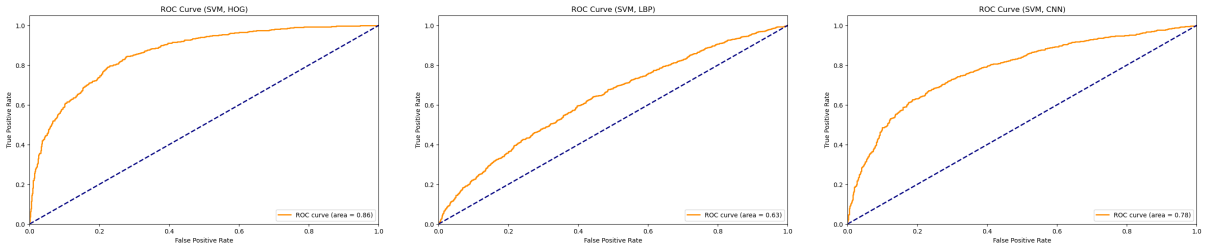

Figure 5: ROC curves comparing SVM performance with HOG, LBP, and CNN features

Figure 5 presents the ROC curves, illustrating each model's true positive rate versus false positive rate across varying thresholds, with CNN features showing a superior curve compared to LBP and comparable to HOG.

6.2 Experiment 2: Multimodal Approach

Experiment 2 measures the performance of a multimodal approach by integrating HOG, LBP, and CNN features within an SVM framework. The idea here is to leverage the

strength of every kind of feature to improve overall classification performance. Moreover, another model is Logistic Regression since it showed great promise in the initial tests.

The SVM model, utilizing the combined feature set, achieved an accuracy of 81.12%. This demonstrates the advantage of leveraging diverse feature sets to improve classification performance. The Receiver Operating Characteristic (ROC) curve, illustrated in Figure 6, shows an Area Under the Curve (AUC) of 0.85 for the SVM model, indicating strong model discrimination between real and synthetic images. The efficacy of the model can be viewed in the results on the confusion matrix, which returned a high degree of correct classifications realized in 1,038 images being classified as 'Real' and 1,007 classified as 'Fake'. However, this came with the presence of 212 false negatives and 243 false positives, proving there is still much needed in the feature extraction process for further accuracy.

The Logistic Regression model,applied to the same combined feature set, returned an even better accuracy of 83.52%, making this technique very promising as a more robust alternative for complex classification tasks. Logistics Regression also returned a better AUC, as visible in the ROC curve in Figure (Figure 6), which means better traceability between sensitivity and specificity concerning the SVM. This goes on to prove that logistic regression not only realizes complementary strengths from combined features but also handles the complexity of a dataset efficiently.

The precision-recall curves (Figure 7) further explain the models' ability in relation to class balance. In this case, both SVM and Logistic Regression models achieved a precision and recall of 0.81, with an F1-Score of 0.81, thus proving to be quite effective in the multimodal approach for the attainment of homogeneous and reliable classification performance. This balance is quite important in applications of image classification where misclassifications may have far-reaching impacts.

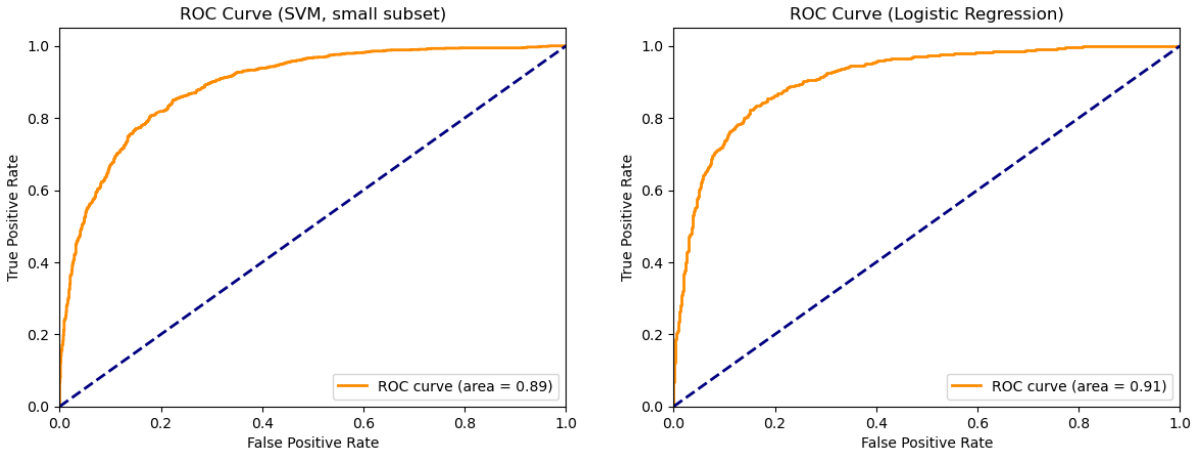

Figure 6: ROC curves for SVM and Logistic Regression with combined features

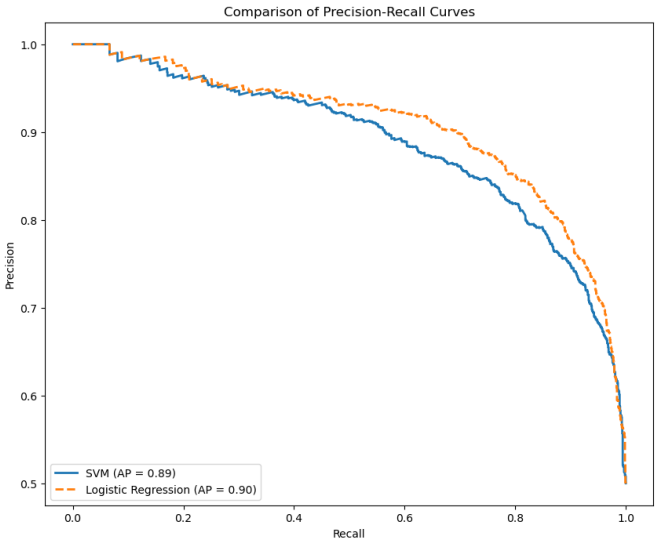

Figure 7: ROC comparison between SVM and Logistic Regression using combined features

6.3 Comparative Analysis

This section makes a comparison of the results obtained from this research with those of selected studies in current literature. Such an analysis is purposed to contextualize the effectiveness of this research's developed multimodal approach: What were its strengths? What were its unique contributions to the field of image classification?.

In this study, the author uses a multimodal approach with Histogram of Oriented Gradients, Local Binary Patterns, and Convolutional Neural Network features. Combining these methods has considerably improved the metrics for performance classification concerning models trained with singular feature sets. The results support the suggestion that including different feature types in one model may explain more comprehensively the diversity of characteristics found in image data, hence improving model accuracy and resilience.

Study/Author	Year	Methodology	Accuracy	Precision	Recall	F1-Score
Other Studies						
G. Monkam et al.	2023	Bayesian optimization in G-JOB GAN	95.70%	-	-	-
L. Natraj et al.	2019	Co-occurrence matrices with CNN	74.5%	-	-	-
F. Martin-Rodriguez et al.	2023	CNNs using PRNU and ELA	-	0.93	0.99	0.95
C. Tan et al.	2023	Gradient-based CNNs	-	0.92	0.88	-
This Research						
Multimodal SVM (HOG, LBP, CNN)	2024	SVM with combined features (HOG, LBP, CNN)	81.12%	0.81	0.81	0.81
Multimodal Logistic Regression	2024	Logistic Regression	83.52%	0.84	0.84	0.84
SVM on HOG	2024	SVM with HOG	78%	0.78	0.78	0.78
SVM on LBP	2024	SVM with LBP	59.2%	0.59	0.59	0.59
SVM on CNN	2024	SVM with CNN	71.4%	0.71	0.71	0.71

Table 2: Comparative Analysis of the Experiments with Relevant Studies

As illustrated in table in section 6.3, the multimodal approach in this research achieved an accuracy of 81.12%, a marked improvement over the unimodal approaches tested and compares favorably to several existing methods in the literature. Notably, the logistic regression model, which emerged as an alternate analysis, showed even higher accuracy at 83.52%, suggesting that the combined features provide rich information that enhances performance beyond the capabilities of SVM alone.

This approach surpasses the performance of models such as those by L. Natraj et al., which achieved an accuracy of 74.5% using co-occurrence matrices with CNNs, underscoring the effectiveness of feature integration over singular methodologies. While the results did not reach the exceptionally high accuracy of 95.70% reported by G. Monkam et al. with GAN optimization, the methodology in this research offers a more accessible framework that balances computational efficiency and accuracy.

These findings support the conclusion that a multimodal feature integration strategy effectively addresses the complexities inherent in image classification tasks, leading to robust model development that aligns with the research community's ongoing efforts to refine classification techniques.

6.4 Discussion

The research given here proves the effectiveness of a multimodal approach in image classification by using Histogram of Oriented Gradients, Local Binary Patterns, and Convolutional Neural Networks. The accuracy for the multimodal approach with SVM was

81.12%, and logistic regression on the same set of features reached an accuracy of 83.52%. These results are better than the single features-based models of Natraj et al. and Tan et al., thus justifying the strength of diversification of features as a means to improve performance. A few limitations were noted notwithstanding the above-mentioned results. Although comprehensive, the dataset was less diverse and thus could further limit its model's generalization across different domains of images. Feature extraction methods may miss some nuances in the images, thus setting improvements through other methods or advanced preprocessing techniques. In future work, testing on more datasets would be very interesting for the improvement of generalizability and the robustness of the model. Generally, in comparison to many of the currently available methods, the multimodal approach within the context of prior studies performed better in this study. The fact that higher accuracy was obtained by Monkam et al. may suggest that sophisticated optimization techniques such as Bayesian optimization could be of benefit. This would involve incorporating these techniques in future studies to achieve further enhanced robustness and accuracy.

Design Considerations and Improvements Though the design of the experiment was very strong for the combination of feature extraction and classification techniques that were followed, there are still some scopes for improvement at different points. Random sampling and inclusion of larger and more varied datasets may reduce sampling biases, especially from advanced architectures like CycleGANs and others or hyperparameter tuning could give more holistic insights.

Contributions to Knowledge It contributes to image classification by validating the efficiency of integrating multiple features and outperforming models with a single feature. This work paves the way for future research focused on multimodal techniques while emphasizing the potential of these techniques in complex image analysis, opening up avenues for commercial applications where strong requirements need to be placed on the performance of image classification.

7 Conclusion and Future Work

This study aimed to enhance the detection of AI-generated images by employing a multimodal approach integrating HOG, LBP, and CNN features with SVM and Logistic Regression classifiers. The findings demonstrate that the multimodal strategy significantly improves classification accuracy, achieving 81.12% with SVM and 83.52% with Logistic Regression, compared to traditional single-feature models. This validates the hypothesis that combining multiple feature types effectively captures complex image artifacts and improves detection capabilities.

Despite these promising results, limitations were identified. The dataset, while extensive, lacked sufficient diversity, which may hinder the model's ability to generalize across different image domains. Additionally, processing large volumes of images was time-consuming, requiring the division of data into subsets to manage computational demands. Addressing these issues can further enhance model robustness and efficiency, paving the way for more advanced detection systems.

Future research should focus on expanding datasets to include a broader variety of GAN-generated images to improve generalization. Incorporating advanced hybrid ap-

proaches and exploring novel feature extraction techniques could further enhance detection accuracy and resilience. Optimizing algorithms for faster processing will also be essential, especially with large-scale datasets. Developing user-friendly interfaces and APIs to integrate with existing digital forensic tools can broaden the practical applications of this research.

References

Abady, L. et al. (2024). A one-class classifier for the detection of gan manipulated multispectral satellite images, *arXiv preprint arXiv:2305.11795* .
URL: *https://arxiv.org/abs/2305.11795*

Arora, A. and Arora, A. (2022). Generative adversarial networks and synthetic patient data, *Future Healthcare Journal* **9**(2): 190–195.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9345230*

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine learning* **20**(3): 273–297.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection, *CVPR (1)*, pp. 886–893.

Fu, T. et al. (2022). Detecting gan-generated face images via hybrid texture and sensor noise based features, *Multimedia Tools and Applications* .

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets, *Advances in Neural Information Processing Systems*, Vol. 27, pp. 2672–2680.
URL: *https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf*

Kit, K. S., Wong, W. K., Chew, I. M., Juwono, F. H. and Sivakumar, S. (2023). A scoping review of gan-generated images detection, *2023 International Conference on Digital Applications, Transformation Economy (ICDATE)*, pp. 1–6.

Liu, Y., Zhu, T., Shen, S. and Zhou, W. (2023). Towards robust gan-generated image detection: a multi-view completion representation, *arXiv preprint arXiv:2306.01364* .
URL: *https://arxiv.org/abs/2306.01364*

Mandelli, S. et al. (2022). Detecting gan-generated images by orthogonal training of multiple cnns, *arXiv preprint arXiv:2203.02246* .
URL: *https://arxiv.org/abs/2203.02246*

Martin-Rodriguez, F. et al. (2023). Detection of ai-created images using pixel-wise feature extraction and convolutional neural networks, *Sensors* **23**(22): 9037.

Monkam, G., Xu, W. and Yan, J. (2023). A gan-based approach to detect ai-generated images, *2023 26th ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, pp. 229–232.

Nataraj, L., Shetty, S., Manjunath, B. S. and Chandrasekaran, S. (2019). Detecting gan generated fake images using co-occurrence matrices, *Electronic Imaging* **2019**(5): 532–1–532–7.

Ojala, T., Pietikainen, M. and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Transactions on pattern analysis and machine intelligence* **24**(7): 971–987.

Sharma, P., Kumar, M. and Sharma, H. (2023). Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches, *Multimedia Tools and Applications* .

Tan, C., Zhao, Y., Wei, S., Gu, G. and Wei, Y. (2023). Learning on gradients: Generalized artifacts representation for gan-generated images detection, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12105–12114.

Wang, H., Fei, J., Dai, Y., Leng, L. and Xia, Z. (2022). General gan-generated image detection by data augmentation in fingerprint domain, *arXiv preprint arXiv:2212.13466* .
URL: *https://arxiv.org/abs/2212.13466*

Zhang, L. et al. (2023). X-transfer: A transfer learning-based framework for gan-generated fake image detection, *arXiv preprint arXiv:2310.04639* .
URL: *https://arxiv.org/abs/2310.04639*