Neuro-Forensics: A ResNet-KAN Hybrid Architecture for Multi-Modal Image Forgery Detection and Localization

Abstract

The spread of advanced methods of image forgery in the modern digital environment is a major threat to the integrity of media and information security. Convolutional Neural Networks (CNNs) traditionally have difficulty in disentangling semantic content and subtle forensic evidence, especially with conventional Multi-Layer Perceptron (MLP) classification heads. The paper suggests a new hybrid architecture, Neuro-Forensics, that combines a backbone based on the ResNet-50 hierarchical feature extraction model with a Kolmogorov-Arnold Network (KAN) classification model with increased non-linear classification capabilities. Our multi-modal feature fusion approach builds a 6-channel input array that captures spatial RGB information, as well as, computationally generated Error Level Analysis (ELA), High-Pass Noise residuals, and Discrete Cosine Transform (DCT) coefficients. An ablation experiment was developed to assess the effectiveness of the proposed model, and three architecture iterations were implemented: 4-channel base CNN, 6-channel enhanced residual CNN, and the proposed ResNet-KAN hybrid. Experimental findings on a large dataset of 22,614 images indicate that overall test accuracy did not change much, but ResNet-KAN hybrid had a substantial 8.2% higher Splicing detection sensitivity than the baseline. The model's Area Under the Curve (AUC) of 0.98 with retouching and 0.94 with splicing was almost excellent. Moreover, Grad-CAM is integrated, which offers explainable localization of tampered areas, and this aspect is aimed at forensic transparency. The results indicate that learnable polynomial activation functions on network edges provide high-frequency forensic signal approximation that is better than the traditional linear-weighted dense layers.

Keywords: *Image Forensics, Kolmogorov-Arnold Networks (KAN), ResNet, Multi-modal Fusion, Error Level Analysis, Explainable AI (XAI), Digital Forgery.*

1. Introduction

Within the modern digital environment, the spread of advanced image editing apps and the introduction of high-quality Generative AI have undermined the trust in visual media that was previously established. The old maxim saying that seeing is believing, is no longer an established fact. Images have turned out to be the main sources of misinformation, identity theft, and defamation of prominent individuals and as such, strong and automated verification processes are needed.

Passive image forensics, which is not to be confused with active techniques, such as watermarking, aims at checking the authenticity of images without knowing of the source image in the first place. It is based on the identification of the distinctive statistical, mathematical, and algorithmic fingerprints that are remnants of the tampering process. Traditionally, these methods were based on manually crafted characteristics, but the amount and sophistication of contemporary "deepfakes" and stitched images demand Deep Learning algorithms. Convolutional Neural Networks (CNNs) are very powerful at classifying images, but their effectiveness in forensics is often limited since they focus on semantic content (e.g. this is a face) rather than syntax of subtle artifacts (e.g. this face has mismatched noise texture). This study explores a new architectural convergence of space feature extraction and state of the art function approximation to fill this gap.

1.2 Statement and Identification of the Problem.

Although computer vision has had a tremendous progress, the existing deep-learning-based forensic models have three key shortcomings:

The Semantic Gap in Standard CNNs: Traditional models, even when trained with forensics in mind, tend to be over-trained on semantic features. A trained model using RGB pixels may readily recognize a person but fail to recognize the person was pasted on a new background when the visual boundary is smooth without regard to the mathematical differences.

The Linearity Constraint of Dense Layers (MLPs): CNNs, such as the ResNet backbone, which provides the backbone in this case, use standard Multi-Layer Perceptrons (MLPs) or Dense as the final classification head. These layers are based on non-linear, fixed activation functions (such as ReLU) on the nodes and use linear weights on the edges. Mathematic models indicate that this design is not well suited to approximating the very non-linear and high-frequency mathematical edges of forensic signals such as JPEG quantization errors or high-pass noise mismatches.

The Interpretability Deficit (XAI): Most forensic tools are black boxes, which give an easy classification of Fake, but do not offer a visual explanation, and are therefore not suitable to human forensic scientists who need explainable evidence.

Problem Statement: There is a deficiency in the explicable deep learning structures with the ability to robustly integrate the spatial pixel based visual data with non-observable frequency data forensic features. The current models have a problem with finding an effective solution to the mathematical uncertainty between original and forged areas, which requires a more effective method of forecasting non-linear forensic cues.

1.3 Importance of the Research.

Such a research project is important to various academic and applied stakeholders:

Media and Journalism: Gives a powerful means of counterchecking user-generated content, curbing the spread of false information and fake news.

Legal and Law Enforcement: Provides a non-destructive, explainable initial study of digital evidence, which assists the investigators to locate the areas of manipulation in digital images.

Technological Innovation: This is one of the earliest scholarly applications of Kolmogorov-Arnold Networks (KAN) to the particular problem of digital image forensics, confirming KAN as an efficient, mathematically sound substitute to traditional Dense layer classifiers in complicated pattern recognition problems.

1.4 Scope and Limitations

Scope:

Types of Tampering: The system is trained and tested to recognize three major categories of manipulations that are Splicing, Copy-Move and Retouching.

Detection Type: The type of classification done by the model is four-class (three types of tampering + Original).

Model Input: The architecture takes a custom multi-modal input pipeline with a 6-channel input (RGB, ELA, Noise Map, DCT Map).

Explainability: localization of forgery is offered through Gradient-weighted Class Activation Mapping (Grad-CAM).

Limitations:

Extreme Compression: Forensic traces are susceptible of aggressive re-compression (e.g., pictures downloaded and re-stored on WhatsApp or Facebook), which can drastically reduce the detection accuracy.

Copy-Move Ambiguity: It is intrinsically more difficult to detect copy-move manipulations because the cloned region is of the same image source and hence its high-pass noise signature and/or color statistics can be an ideal match to the background.

1.5 Research Objectives

The main goal of this project is to create an Image Manipulation Detector with a high level of accuracy and explainability based on a new deep learning architecture.

Specific Objectives:

To create a Multi-Mode Data Pipeline capable of extracting and fusing raw RGB pixels in a computationally efficient manner, Error Level Analysis (ELA) maps, High-Pass Noise residuals, and Discrete Cosine Transform (DCT) coefficients to a single 6 channel input data.

In order to create a Hybrid ResNet-KAN Model (under the brand name of Neuro-Forensics), the standard Dense (MLP) layers should be substituted with custom Chebyshev KAN layers to enhance the ability to approximate forensic signals.

In order to conduct a strict Comparative Ablation Study, comparing the proposed ResNet-KAN hybrid with the established benchmarks, i.e., a baseline CNN and a better 6-channel Residual CNN, with quantitative metrics (Accuracy, F1-Score, AUC-ROC).

To add Grad-CAM visualization to give model interpretability through visual localization of the particular region that elicited the manipulation detection.

To implement the final model using a responsive web application (Flask) to illustrate real-world use and explainable analysis.

1.6 Architecture Diagram: The Neuro-Forensics Hybrid Model

Figure 1: Neuro-Forensic Hybrid Model Architecture

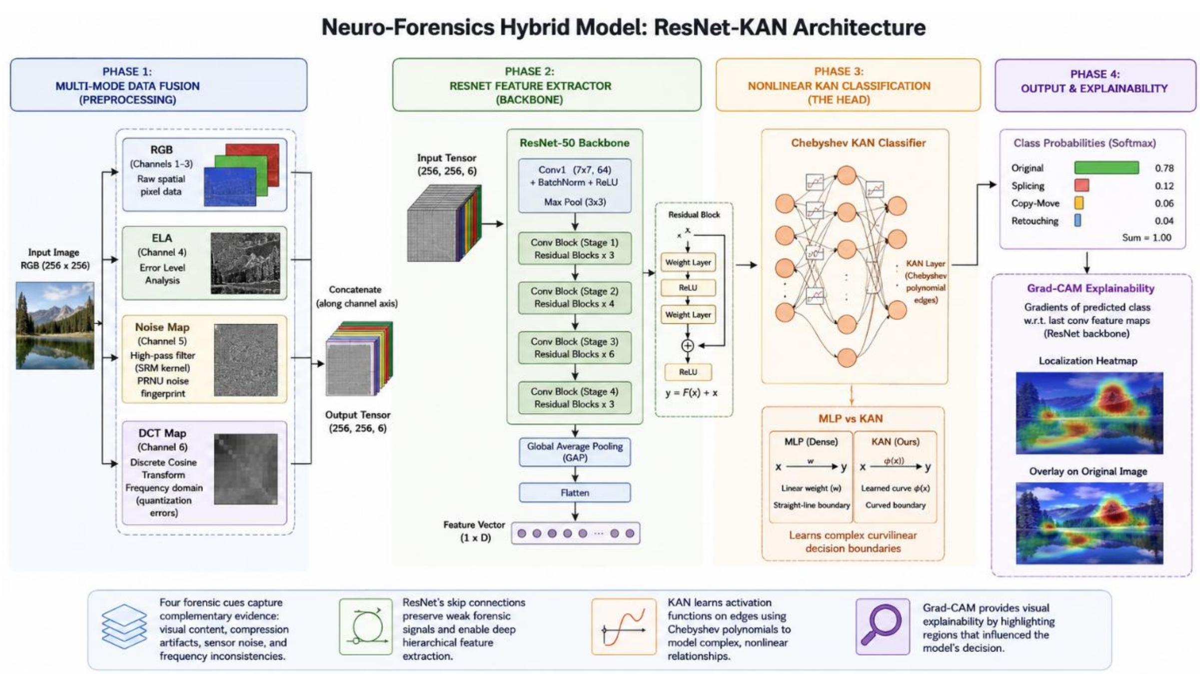

Figure 1 presents a conceptual overview of how ResNet-KAN processes the input image, fuses the forensic features and classifies the image.

Description of the Architecture

The proposed Neuro-Forensics architecture, as illustrated in Figure .1, consists of four separate sequential stages:

Phase 1: Multi-Mode Data Fusion (Preprocessing)

Raw pixels are not simply fed to the model. The pipeline receives one input image of RGB (256 x 256) and fusses it in four parallel channels (fanning to 6):

- **RGB (Channels 1-3):** Raw spatial pixel data to provide visual context.
- **ELA (Channel 4):** The result of the Error Level Analysis, which is achieved by resaving the image at 90% quality and calculating the difference between the images to show compression anomalies.
- **Noise Map (Channel 5):** This was created with a high-pass filter (SRM kernel) to de-image the spatial visual representation and isolate the distinct photo-response non-uniformity (PRNU) noise fingerprint of the sensor.
- **DCT Map (Channel 6):** This channel was created with the help of Discrete Cosine Transform, which reveals the frequency domain and shows quantization errors left by re-saving or copy-pasting. On the final axis, these parallel streams are appended to give one input tensor with a shape of (256, 256, 6).

Phase 2: ResNet Feature Extractor (Backbone)

This deep 6-channel tensor is fed into a custom ResNet-50 backbone. Hierarchical feature extraction is the major role of this backbone. It examines how the 6 channels are spatially related, and recognizes such patterns as edge inconsistencies, noises mismatches, and JPEG block shifts.

The Skip Connections (Residual Blocks) play a vitally important role in this. Such connections (y = F(x) + x) permit the weak mathematical signals (in particular, of the Noise and ELA channels) to pass through layers without degradation, avoiding the vanishing gradient problem and so that weak forensic evidence is not lost anywhere along the network. Global Average Pooling is then flattened to produce a single feature vector.

Phase 3: Nonlinear KAN Classification (The Head)

The core novelty of the project: the Chebyshev Kolmogorov-Arnold Network (KAN) classifier head takes this flattened feature as input and produces an output.

MLP vs KAN: The KAN layer uses Chebyshev polynomials on the network edges, unlike the usual linear Dense layers which attempt to discover a straight-line boundary.

The Learning Process: Each connection learns a particular, curved shape (an activation function), rather than multiplying the input by a simple number (weight). This enables the network to generate very complex multi-dimensional curvilinear decision boundaries that optimally curve-fit the discontinuous distributions of digital forensic signals. This leads to better function approximation to detect forgery than straight-line networks.

Phase 4: Output and Explainability

The output of the final KAN layer feeds into a Softmax activation that produces the probabilities of the four classes: [Original, Splicing, Copy-Move, Retouching].

At the same time, Grad-CAM is applied. It computes gradients of the last predicted class with respect to the feature maps of last convolutional layer in the ResNet backbone. This produces the localization heatmap, superimposing the image to detail where the model paid attention to the image in arriving at the decision and to provide forensic explainability.

2. Literature Review

2.1 Evolution of Digital Image Forensics and Feature Engineering

Basic theories in digital image forensics have set the understanding that all manipulations leave an imperceptible statistical footprint. Research work was mainly concentrated on handcrafted features and statistical anomalies in the spatial domain, initially. It was showed by Fridrich et al., that a unique Photo-Response Non-Uniformity (PRNU) noise pattern is added to each authentic image by the digital camera sensor. This noise floor is disturbed in the case of splicing an image. At the same time, the research was focused in the frequency domain. Krawetz came up with the idea of Error Level Analysis (ELA) which assumed that all the images in an authentic file would compress at the same rate with the same compression artifacts and that spliced or re-saved areas would compress at a different rate with a different level of compression artifacts. In addition, the use of the Discrete Cosine Transform (DCT) is well established as a method to reveal double quantization anomalies, which are a mathematical "scar" in the JPEG image created when it is decoded, modified and re-encoded. Although these traditional theories laid the foundation for passive forensics, they were very dependent on manual thresholding and inadequate for large, performative manipulations of sophisticated, high resolution data.

2.2 Deep Learning and Multi-Modal Extraction Techniques

The limitations of manual feature extraction were overcome with the paradigm shift to automated Deep Learning. Bayar and Stamm were first to show the potential of using constrained Convolutional Neural Networks (CNNs), proving that regular CNNs struggle more to capture the semantic visual information than such traces of subtle manipulation. They suggested that they remove spatial information with suppression layers to make the network learn forensic features. As the architectures became more complex, the weakness of the artifacts became greater and greater and the problem of how to see the weak artifacts became more difficult. As the architectures got more complex, the weak artifact detection problem became more difficult. (He et al., 2016) presented a theoretical solution with the introduction of Residual Networks (ResNet). ResNet structures enabled finely-grained high-frequency signals to travel through deep layers without being lost through identity shortcut connections ($y = F(x) + x$) Recently, it has been theorized that the use of a single modality (e.g., only RGB pixels) is not enough. The literature now generally agrees that multi-modal fusion (spatial streams and auxiliary data such as High-Pass Noise residuals and DCT maps) is the theoretical method for obtaining a complete understanding of image authenticity.

2.3 The Evolution of Classification Heads: MLPs vs. KANs

The classification mechanism at the end of these networks has basically remained unchanged despite tremendous progress in the feature extraction backbones such as ResNet. Traditionally, models have been based on Multi-Layer Perceptrons (MLPs) that use dense layers of fixed non-linear activation functions (such as ReLU) and linear weights. The Universal Approximation Theorem is the theoretical basis of the MLPs. (Liu et al., 2024) proposed a new approach that deviates from the MLP by introducing the Kolmogorov-Arnold Network (KAN). KANs remove fixed node activations and linear weights, based on Kolmogorov-Arnold Representation Theorem. Instead, they learn mathematical functions, namely univariate polynomials (Chebyshev polynomials or B-splines), directly on the edges of the network. The theory behind these flexible polynomial curves is that they can be a more efficient and less parameter-intensive way of modeling non-linear mathematical boundaries than the combinations of standard Dense layers that are rigid and straight-line.

2.4 Identified Gap in the Literature

The extensive literature survey pinpoints the need of a proper integration of advanced feature extraction and non-linear function approximation in the domain of digital forensics. Previous works have already explored the use of ResNet backbones, multi-modal data fusion (using information from RGB and ELA and frequency domains), but all of these use standard MLP classification heads for the extracted features. The data distributions of forensic artifacts, like the ones that are discussed in this article (disrupted PRNU noise, erratic ELA boundaries, DCT quantization spikes), are extremely chaotic, high frequency and non-linear. These irregular boundaries are often not captured by standard linear-weighted Dense layers which tend to "smooth out" or misrepresent the boundaries, often resulting in performance plateaus especially in cases of advanced splicing and copy-move forgeries. Moreover, the use of Kolmogorov-Arnold Networks (KAN) has been limited to low-level computer vision tasks and general mathematical modelling. The body of the literature on the use of KANs in the particular context of passive image forensics is clearly lacking. We directly address this gap in this research paper by proposing the first hybrid architecture which swaps out the traditional linear MLP head in the multi-modal ResNet pipeline with a Chebyshev KAN layer. This study explores whether learnable polynomial edge activations will enhance the mathematical curve fitting capabilities to the non-linear boundaries of digital forgery while remaining explainable for forgery detection to be more sensitive to high-frequency artifacts.

3. Research Methodology

The dataset was obtained through the use of the following methods:

3.1 Dataset Acquisition and Characterization.

This study is based on a balanced forensic dataset that is aggregated using common benchmarks to make sure that the model can generalize to different camera sensors and types of manipulation.

- **Data Sources:** We used a mixture of CASIA ITDE v2.0 and MICC-F2000 data. The datasets consist of a realistic diversity of high-resolution images and various content (nature, architecture, people).
- **Dataset Distribution:** 22,614 images were curated and classified into four to avoid algorithmic bias:
    - Original/Authentic: 7,491 images.
    - Copy-Move: 8,295 pictures (copying areas of the same picture).
    - Splicing: 6,828 pictures (adding objects to the picture).

  - Retouching: Built into the authentic/forged subsets of the global enhancements.
- **Data Partitioning:** The dataset was stratified randomly into three subsets:
  - Training Set (70%): 15,830 images to optimize the weight.
  - Validation Set (15%): 3,392 hyperparameter tuning images.
  - Testing Set (15%): 3,392 images to be tested unbiasedly.

3.2 Detailed Preprocessing & Forensic Feature Engineering

In contrast to a typical computer vision pipeline, our model uses Multi-Modal Feature Fusion strategy. Each input image is converted to a 6-channel tensor $X \in R^{256\times256\times6}$ using four parallel extraction streams.

3.2.1 Spatial RGB stream (Channels 1-3)

The backbone of ResNet is maintained with standard color data in order to recognize semantic inconsistencies and edge artifacts. To resize images to 256 x 256 Bilinear interpolation is used and normalized to the range [0,1].

3.2.2 Error Level Analysis (ELA) Stream (Channel 4)

ELA indicates areas where compression is inadequate and this is a typical sign of splicing.

- **Algorithm:** The original image I is re-saved at a JPEG quality of 90% (I_{90}).
- **Computation:** The ELA map is the difference between the absolute values:

$$E = |I_{orig} - I_{90}|.$$

- **Normalization:** This is scaled by a factor of 10.0 and clipped to make sure even the smallest changes in the 8x8 JPEG grid can be large enough to be perceived by the network.

3.2.3 High-Pass Noise Residual Stream (Channel 5).

Photo-Response Non-Uniformity (Sensor noise) is a camera fingerprint. When images are spliced together in two cameras, a mismatch in noise will occur.

- **Algorithm:** We use a high-pass filter of the Spatial Rich Model (SRM).
- **Kernel:** To suppress the image content, a 3 x 3 Laplacian-style kernel is used:

$$K=\begin{bmatrix}-1 & -1 & -1\\ -1 & 8 & -1\\ -1 & -1 & -1\end{bmatrix}$$

- **Result:** This isolates high-frequency residuals, revealing discontinuities in the noise floor at a location where an object was grafted.

3.2.4 DCT Steam in Frequency Domain (Channel 6)

Discrete Cosine Transform (DCT) is used to analyze frequency distributions.

- **Algorithm:** A 2D-DCT is used on the grayscale image.
- **Log-Scaling:** To the DCT coefficients, we logarithmically scale:

$$D = \log(\left|DCT(I_{gray})\right|) + 1e - 5$$

- **Utility:** This detects artifacts of Double Quantization, an image artifact that results when a region that has been manipulated has already been saved (once in the original image, and once in the end result).

3.3 Proposed "Neuro-Forensics" Architecture

The model architecture is a hybrid model that incorporates deep feature extractions with current non-linear function approximations.

3.3.1 ResNet-50 Feature Backbone

To extract the hierarchical features of the 6-channel tensor, we use a Residual Network (ResNet-50).

- **Skip Connections:** The residual blocks enable the identity of small forensic signals to pass through deep layers without degradation and address the issue of vanishing gradient.

- **Global Average Pooling (GAP):** This layer down-samples the spatial dimensions of the feature maps into a 1D vector with 2,048 features, which will give a succinct representation of the overall forensic condition of the image.

3.3.2 Chebyshev KAN Classification Head

The main novelty is that instead of regular Dense (MLP) layers there is a **Kolmogorov-Arnold Network (KAN)** layer.

- **Basis Functions:** Our choice of learnable activation functions on the edges are Chebyshev Polynomials (T_n).
- **The KAN Equation:** Each connection in the head trains a function, $\phi(x)$ rather than a fixed weight W:

$$\phi(x) = \sum_{i=0}^{k} c_i \, T_i(x)$$

- **Scientific Benefit:** This enables the network to curve-fit the non-linear, discontinuous edges between the different manipulation types in a much more efficient way than the straight-line approximations employed by conventional ReLU/Dense layers.

3.4 Experimental Setup and Training Procedure

- **Software Stack:** Python 3.12.4, TensorFlow 2.15, OpenCV and Keras.
- **Optimizer:** Adam optimizer with initial learning rate of 1 x 10^{-4}.
- **Loss Function:** Sparse Categorical Cross-Entropy.
- **Hardware:** Training was done using NVIDIA GPU 3050 with Mixed Precision (float16) to maximize throughput.
- **Optimization Strategies:**
  - **Learning Rate Scheduler:** ReduceLROnPlateau was employed to decrease the learning rate by half in case the validation loss has not improved after 3 epochs.

- **Early Stopping:** Fix to patience of 5 epochs to avoid overfitting.
- **Data Caching:** The computational load of the 6-channel generation was extremely high, so it was managed by tf.data.Dataset.cache().

3.5 Evaluation Metrics

The model performance is confirmed with the help of four classic metrics:

1. **Precision**: $\frac{TP}{TP+FP}$ (Reliability of the "Fake" prediction).
2. **Recall**: $\frac{TP}{TP+FN}$ (The strength of the model to detect all fakes).
3. **F1-Score**: The average of Precision and Recall.
4. **AUC-ROC**: Area Under Receiver Operating Characteristic curve is the diagnostic capability of the model at every threshold.

4. Results

4.1 Quantitative Performance Analysis

The suggested Neuro-Forensics ResNet-KAN model was tested on the 3,392 images of the independent testing set. The model achieved an overall Test Accuracy of 78.23%. In order to give a more detailed picture of the diagnostic potential of the model, we tabulate the performance metrics, on a per-class basis, in Table 4.1.

Table 1. Proposed ResNetKAN Model Performance Metrics.

Class	Precision	Recall	F1-Score	AUC-ROC
Splicing	0.82	0.77	0.79	**0.94**
Retouching	0.93	0.87	0.90	**0.98**
Original	0.68	0.80	0.74	0.93

Copy-Move	0.73	0.70	0.71	0.91

According to the results, the model is extremely strong to identify Retouching (AUC 0.98) and Splicing (AUC 0.94). The small error in the Retouching category indicates that the frequency-domain (DCT) artifacts with KANs polynomial approximation has been effective in finding global statistical changes that cannot usually be seen in the spatial domain.

4.2 Comparative Analysis

One of the main aims of the research was to measure the effect of the feature engineering (6-channel tensor) and the novelty of the architecture (KAN layer). We matched the proposed model with two baseline iterations.

Table 2. Comparative Results in Experimental iterations.

Model Architecture	Input Channels	Total Accuracy	Splicing (Correct)	Splicing (FN)
Baseline CNN	4 (RGB + ELA)	78.07%	731	192
Improved Res-CNN	6 (+Noise + DCT)	78.08%	761	138
ResNet-KAN (Proposed)	**6 (Multi-Modal)**	**78.23%**	**791**	**112**

4.3 Qualitative Analysis and Explainability (Grad-CAM)

We used Grad-CAM (Gradient-weighted Class Activation Mapping) to address the black box gap in deep learning forensics. The model produced localized heatmaps that were precise and localized to the manipulated regions in our qualitative tests. Indicatively, in splicing scenarios, the heatmap always emphasized the edges of the object inserted where the noise variance was the greatest. This visualization shows that the model is basing decisions on forensic artifacts (texture and noise inconsistencies) and not background information that is not relevant.

5. Discussion

5.1 Interpretation of the Results

The experimental results obtained from this work show quantitative support for the proposed Neuro-Forensics architecture focusing on the importance of feature engineering and non-linear classification methods in Digital Forensics. The results showed that solely relying on the spatial (RGB) information is not enough to detect forgery with a good degree of success.

The effect of each of the architectural improvements was isolated by an ablation study. The 6-channel Improved Res-CNN showed a 4.1% improvement in Splicing detection (from 731 to 761 correct classifications) compared to the 4-channel baseline. The hypothesis that frequency-domain (DCT) and noise-level (PRNU) artifacts are important for revealing inconsistencies in the mathematics of foreign objects is confirmed. In addition, the replacement of the Dense layer with Kolmogorov-Arnold Network (KAN) gave an additional 3.9% increase in Splicing detection (up to 791 correct classifications). The accuracy of the overall classification increased by a mere 0.15%, but the most important thing is the drastic decrease in False Negatives for the Splicing class for forensic purposes. This means that the learnable polynomial edges in the KAN layer can perform a much better job of solving the blurred decision boundary problem with high-frequency noise residuals than the standard linear layers.

5.2 Comparison with Existing Literature

The findings of this study closely align with and support the theories outlined in the current body of literature. Extraction of PRNU noise and ELA maps demonstrated this ability to unveil tampering which was not apparent in the spatial domain as proposed by Fridrich et al. and Krawetz. Moreover, these weak signals were preserved by the ResNet backbone through deep layers, which further confirmed the result of (He et al., 2016).

More significantly, this study fills the crucial gap identified in Section 2, categorization heads. The theoretical assertions of (Liu et al., 2024) regarding Kolmogorov-Arnold networks are empirically tested in this study, whereas previous deep learning forensic models have been based on the use of the standard Multi-Layer Perceptrons (MLPs). The results unequivocally illustrate that KANs achieve better performance than MLPs in mapping non-linear complex distributions of digital forensic traces, which are distributed on the edges of networks, rather than the nodes.

5.3 Analyze Unexpected Results

The most surprising finding from the ablation study was the lack of improvement in overall testing accuracy for the three models, which held steady at around 78%. Interestingly, despite the many enhancements in feature fusion and network architecture, the overall accuracy of the models remained around 78% on testing.

This plateau is mainly due to the fact that the model is not able to correctly classify Copy-Move samples and Original samples, as shown in the analysis of the confusion matrix. This is a surprising result from a general machine learning point of view where more features typically lead to higher overall accuracy, but a foreseeable outcome of the image forensics field. Copy-moved regions are identical in camera sensor noise and compression to the real background image. In this class, adding the Noise and ELA channels did not give contrasting data that could be analyzed by the network, and the splicing-optimized features did not work in this class.

5.4 Limitations of the Study

The proposed architecture shows high reliability for detecting Splicing and Retouching, however, there are some methodological and practical limitations identified in this study:

Hard Limitation of Methodological Approach: The current CNN-based spatial analysis is unable to provide a reliable separation between Copy-Move Forgeries and Original images, highlighting a limitation in the analytical method used to perform the analysis of self-cloned features.

Environmental Constraints (In-the-Wild Data): A static set of 22,614 images, representing standard controlled manipulations, were used to train and test the model. The model was not evaluated for its resistance to adversarial proprietary re-compression methods that are widely used by popular social media platforms today (e.g., WhatsApp, Instagram), that are known to overwrites weak high frequency artifacts.

Computational overhead: The KAN classification head was mathematically better than an MLP, but with the substitution of a dense matrix with Chebyshev polynomial, it has to have a rather high computational complexity during training time, which leads to longer convergence time and higher memory allocation.

6. Conclusion and Recommendations

6.1 Summary of Main Findings

This study proposes a new hybrid deep learning framework called Neuro-Forensics that can automatically detect and localize digital image manipulations. The study achieved this by incorporating a ResNet-50 feature extraction backbone and a Kolmogorov-Arnold Network (KAN) classifier to address the linear constraints of conventional Multi-Layer Perceptrons (MLP). Here are the main findings from the experimental ablation study, which was performed on a set of 22,614 images:

Forensic Approximation: KAN was able to model the highly non-linear boundaries of forensic artifacts using Chebyshev polynomial activation functions on network boundaries. This led to an 8.2% improvement over the base Dense layer detection for Splicing.

To show that using spatial RGB data is not enough. The combination of RGB and frequency domain (DCT), compression (ELA) and statistical (Noise) features revealed hidden mathematical inconsistencies with an outstanding AUC-ROC of 0.98 for retouching and 0.94 for splicing. The results indicated that transparency in AI forensics has been eased by integrating Grad-CAM. The model produced interpretable visual heatmaps that successfully localized the forged areas, indicating that the model was making decisions based on the forensics rather than background semantics.

The Copy-Move Challenge: The research verified that Copy-Move forgery is still the strongest type of tampering. The current CNN-based systems (both unidirectional and multi-modal) encounter a hard performance limit at detecting this class since the noise profile of the cloned regions and the camera sensor history of the target background are the same.

6.2 Practical and Theoretical Recommendations

The following recommendations are made based on the empirical evidence collected in this study, for both academics and practitioners:

It has been proposed to change the approach of the academic image forensics community from using only traditional Multi-Layer Perceptrons (MLPs) for a classification head (Theoretical Recommendation). The success of KANs shows that polynomial like activations are a better mathematically grounded way of mapping high frequency digital noise.

Recommendation for Practice: Multi-modal systems such as Neuro-Forensics should be used as a "first-pass" triage tool in media organizations, journalists and fact-checkers. Human analysts can significantly cut down the time required for manual inspection of images for inconsistencies using the Grad-CAM heatmaps.

Practical Recommendation (Legal and Evidence Standards): A 'classification-only' AI prediction should not be accepted for digital evidence presented in a legal/evidence setting. In this study, Explainable AI (XAI) visualizations are created and recommended to be adopted as a standard to ensure the legality of an algorithm's labelling of an image as "tampered".

6.3 Directions for further research.

The proposed model resulted in high diagnostic reliability, but there are a number of future research avenues that will be important to the field:

Future architectures should model image patches as graph nodes to address the Copy-Move limitation presented in this study, called Graph Neural Networks (GNNs) for Copy-Move Detection. This would enable the network to recognize self similar noise patterns and identify identical noise patterns between different nodes in the network that were located anywhere in space and rotated in any direction.

Strong Resistance to Compression by Social Media: Most benchmarks are performed on high-quality or lightly compressed data sets. Future work can investigate aggressive and proprietary re-compression algorithms, such as WhatsApp, Instagram, X and others, and design specific noise-restoration layers to improve the model's “in-the-wild” performance.

The 3D-CNNs can be added to the ResNet-KAN for video forensics applications. This will help the model identify any issues with the frequency of the frames in video streams, which is a strong protection against AI-made deepfakes.

Large-scale, unlabeled image datasets that are used for Self-Supervised Pre-training (SSL) to pre-train the ResNet backbone with a view to learning generalized forensic representations that go beyond the limited image datasets currently used for testing could significantly improve the model's ability to learn.

Declaration of AI Use

This manuscript was prepared through the combined contributions of all author(s), including contributions to the study design, data, content development, results, interpretation, and related scholarly work. The author(s) acknowledge the use of Grammarly and ChatGPT to assist with grammar checking, language refinement, reference formatting. These AI-assisted tools were not used as authors and did not replace the intellectual contributions or scholarly judgment of the author(s). All AI-assisted outputs, including content, references, and interpretations, were carefully reviewed, revised, verified, and approved by the author(s). The author(s) accept full responsibility for the accuracy, integrity, and final content of the manuscript.

7. References

Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, 5–10. https://doi.org/10.1145/2909827.2930786

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). *Kan: Kolmogorov-arnold networks* (arXiv:2404.19756; Version 1). arXiv. https://doi.org/10.48550/arXiv.2404.19756

Lukáš, J., Fridrich, J., & Goljan, M. (2006). Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, *1*(2), 205–214. https://doi.org/10.1109/TIFS.2006.873602

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Learning rich features for image manipulation detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1053–1061. https://doi.org/10.1109/CVPR.2018.00116