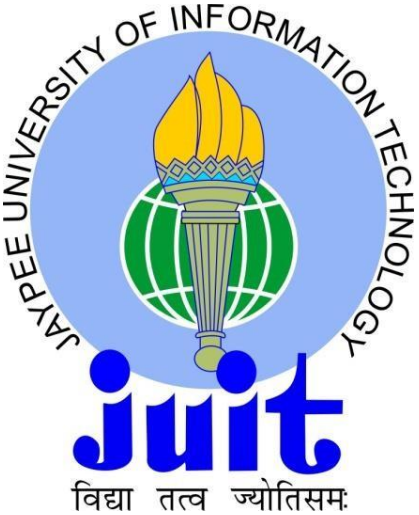
IMAGE AND VIDEO FORENSICS USING DEEP LEARNING

Thesis submitted in fulfillment for the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

SURJEET SINGH



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT, H.P.

May, 2024

Dedicated to
My
Beloved Parents

CONTENTS

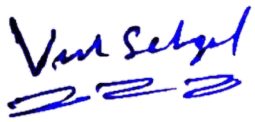
DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled, "**IMAGE AND VIDEO FORENSICS USING DEEP LEARNING**", submitted at **Jaypee University of Information Technology, Waknaghat, Solan (HP), India** is an authentic record of my work carried out under the supervision **Prof. Dr. Vivek Kumar Sehgal**, Jaypee University of Information Technology, Solan, (HP) India. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. thesis.

Surjeet Singh
Enrolment No.: 186215
Computer Science and Engineering
Jaypee University of Information Technology, Waknaghat, Solan
(HP), India
May 2024

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled "**IMAGE AND VIDEO FORENSICS USING DEEP LEARNING**" submitted by **Surjeet Singh** at **Jaypee University of Information Technology, Waknaghat, Solan (HP), India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.



Prof. Dr. Vivek Kumar Sehgal
Computer Science and Engineering
Jaypee University of Information
Technology, Waknaghat,
Solan, (HP), India
May 2024

ACKNOWLEDGMENTS

I extend my heartfelt gratitude to the individuals who supported me during my tenure as a Ph.D. Candidate at Jaypee University of Information Technology, Waknaghat Solan (H.P). This page serves as a token of appreciation for those who played a pivotal role in my journey. First and foremost, I express my sincerest thanks to God for bestowing me with this invaluable opportunity. My deepest appreciation goes to my supervisor, Prof. Dr. Vivek Kumar Sehgal, whose unwavering support and guidance have been instrumental throughout this endeavor. Their profound knowledge and invaluable advice have been integral to the successful completion of this thesis. Without their mentorship, my academic and professional growth would not have been possible.

I am indebted to my parents, Shri Nihal Singh and Smt. Jayanti Devi, for their unwavering belief in me and unwavering support in pursuit of my career goals. Their encouragement has been my guiding light. I am immensely grateful to Dr. Hemraj Saini, Dean DIT University Dehradun, Dr. Pankaj Kumar HOD Civil Dept, National Institute of Technology Nagaland, Dr Saurabh Rawat Associate Professor JUIT, Dr Arvind Dhaka Associate Professor Manipal University Jaipur, Mr. Kaushal Kumar Civil Department JUIT and Mr. Chandra Pal Civil Department JUIT for his continuous encouragement, unwavering support, and invaluable guidance. His expertise in the field has been an immense asset to me. I am thankful to all individuals, both directly and indirectly involved, who contributed to the completion of my thesis. Your support has been invaluable.

Lastly, I express my gratitude to all members of the Computer Science and Engineering Department for providing the necessary resources for the successful culmination of my thesis.

Warm regards,

Surjeet Singh (186215)

Date: 24 May 2024

SPECIAL THANKS TO

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT, SOLAN, HIMACHAL PRADESH, INDIA

ABSTRACT

Digital forensics is a critical research branch that focuses on identifying the original source and verifying the authenticity of digital data, particularly concerning visual information, which poses a major challenge for forensic experts in the present digital era. This field extensively employs image and video data to validate the accuracy of information during the picture gathering phase and to detect tampering throughout the digital image processing pipeline. However, image and video forensics stand as the most significant challenge in the current digital landscape due to the continuous modification of digital content using freely available software and editing tools. Ensuring the authenticity of images and videos involves employing various forensic methods from image acquisition to storage. Despite the complexity of this task, two major hurdles faced by researchers in digital forensics include identifying the source of acquired images and videos to establish their origins with specific devices, models, or brands, which holds particular importance when examining historical context. Additionally, detecting image forgery is essential to preserve the integrity of digital data for legal purposes, as forged images distort original content and disseminate false information. With the ready availability of software editing tools, the risk of modifying captured images with deceptive or defamatory content has increased, leading to the spread of fake information on social media platforms.

In this thesis, we aim to enhance the accuracy of source identification for images and videos using a deep learning framework. Additionally, we focus on image forgery detection to prevent the misuse of falsified content. Our proposed approach involves utilizing the Twin CNN Architecture (TCA) for image source identification, where the initial DnCNN (Denoising Convolutional Neural Network) is used to remove noise from the original dataset, followed by the second CNN architecture to classify images based on extracted features from various convolutional layers. This approach improves the effectiveness of class prediction and efficiency in identifying original source. Furthermore, we introduce a CNN-based architecture for accurately classifying forgery in given images, detecting unseen forgeries through feature extraction from multiple convolutional layers, and employing an SVM classifier for precise labeling. Lastly, our deep learning-based CNN Multi-Modal Camera Model Identification improves video source identification accuracy through the use of CNNs.

Keywords: *Image Forensic; Forgery; CNN; Video Forensic; Image Denoising*

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 *Preface*

In the current digital era, researchers face the challenging task of validating the authenticity and trustworthiness of images captured with widely accessible digital devices. With digital images being an integral part of everyday life, the ability to manipulate them using advanced digitization and image analysis tools raises ethical concerns [1]. Addressing this issue, two key aspects are emphasized: determining the imaging system responsible for capturing the image and detecting potential forgeries. In contemporary society, the value of images and videos as pivotal evidence in legal proceedings and daily conflicts cannot be overstated. The imperative to ascertain the device employed for image capture, particularly in instances of video surveillance or covert recordings, underscores its indispensable role as crucial evidence within the judicial system [2]. To address these obstacles, digital image and video forensics concentrate on discerning and scrutinizing fundamental aspects within images and videos. Key objectives include origin recognition without prior image analysis or registration, extracting hidden information, and detecting manipulations [3]. To trace the origins of collected images or videos, a comprehensive examination of the digital image processing pipeline is conducted, spanning from the initial acquisition phase to the storage phase depicted in Figure 1.1.

To trace the initial origin of the image acquired is approachable in the following way:

- Identifying the source camera model that captured the original image.
- Was this image captured uniquely by a single device, or does it seem to be a composite of multiple images?

Image forensics operates on the principle that a digital image carries inherent evidence from its creation to subsequent stages in its life cycle. Collecting and analyzing these digital footprints helps understand how digital content evolves. Identifying the image source involves two main methods: one follows the image processing pipeline's traces traditionally, while the other relies on feature extraction techniques to determine the image's original source. In conventional image forensics, source determination is carried out throughout the

image acquisition, compilation, and final editing phases based on intrinsic artifacts or footprints [3]. Now a day's image source identification done based on data driven approach to extract features and match similarity index with original image. In recent years, Image source recognition done by different machine learning approaches to improve the accuracy over the conventional approaches. In which multiple features are extracted from image patch dataset and train the model to classify the given input image. Due to noise in image dataset apply different denoising filters to restore the original image then apply classifiers to detect original source and enhance the effectiveness of prediction. In recent scenario, deep learning-based model is used for image source identification give high accuracy over traditional approaches. In deep learning framework, first collect the dataset according to develop model and then train model on dataset and test the accuracy on given input image.

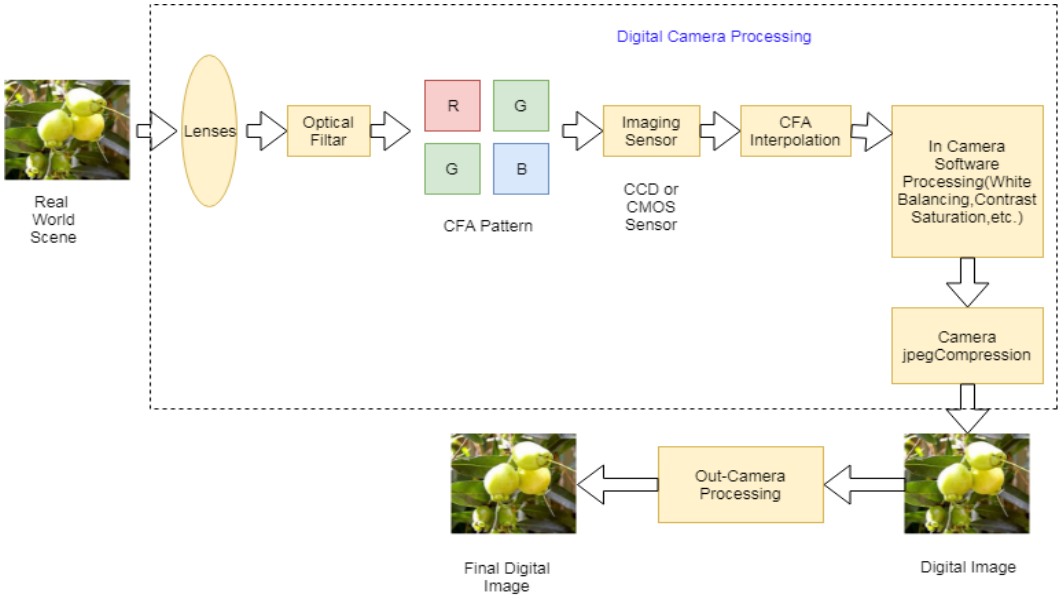


Figure 1.1. Image processing cycle in digital device

Researchers must deal with a number of challenges while developing a video camera model detection system that they do not have to deal with for image manipulation-based systems. Do changing different frame types and used during video encoding, for instance, have an impact on the forensic traces used to identify camera models? If that's the case, how should this be considered while creating and implementing a video-based recognition system? On the basis of a single M x M picture patch, many image-based systems are capable of making decisions about the source model that are reliable. Can this be done with videos, or will the accuracy that can be achieved with just one patch be too low? Where in a movie should these patches be obtained from if forensic data from numerous patches is required? Digital videos are large, making their utilization computationally expensive [4].

The ease of falsifying digital images through readily accessible software tools on digital devices has led to an increase in manipulated content. Device-oriented forgery can be either innocuous or perilous, depending on its intent. Deliberate use to disseminate false information poses a significant threat to society. Often, image manipulation serves as a tool wielded by malicious individuals to tarnish the social or financial standing of public figures. Typically, device forgery involves benign alterations such as contrast adjustments or brightness modifications. Moreover, built-in filters in smart devices facilitate effortless modification of original images, allowing easy sharing. The repercussions of image forgery extend to influencing public sentiments and perpetuating false information within society [5]. Images find extensive use across various domains such as image forensics investigations, legal proceedings, surveillance systems, smart detection technologies, and medical imaging. Presently, researchers face the formidable challenge of detecting and notifying users about forgeries, providing authentic documentation. Traditional methods of image forgery, like copy-move and image manipulation techniques, pose significant threats to image integrity and the preservation of valuable information [6]. The proliferation of mobile applications has facilitated the rapid dissemination of misinformation on social media platforms, leading users to assume the accuracy of all available information associated with specific users. To detect user-intended forgery, examination involves scrutinizing the following aspects:

- Has the provided image been altered or remains unaltered?
- Has the provided image been edited to enhance features or modify specific elements?
- Has the provided image been created by amalgamating two or more images using an intelligent system?

In our endeavor, our primary objective is to precisely detect and identify these alterations, particularly focusing on image forgery, as illustrated in Figure 1.2.

Figure 1.2. Image Forgery Example

Video Source camera identification, according to [7], [8], is a significant topic that

concentrates on several issues related to source class, such as model, brand, and sensor type. The procedure of determining the authenticity of information comes from the claimed source is known as source validation. Video forgery example is shown in given Figure 1.3.

Figure 1.3. Video Forgery Example

1.2 *Problem Statement*

In conventional picture forensics, source determination is based on intrinsic artifacts or imprints that are traces during the image acquisition phase, compression, and final editing phases. Researchers concentrate on the features of the lens, sensors, and CFA (color filter Array) interpolation techniques throughout the image acquisition phase. The light rays that reach the sensor array pattern and are reflected by the lens during the picture acquisition phase are transformed into continuous signals. Each camera model has a distinct lens system, various types of sensors, and demosaicing methods for color filter arrays. The lens creates several types of artifacts and leaves distinctive traces to identify the camera model during the production process. Different digital gadget parts leave distinct traces in the photos that are taken. Investigators discover a link between hardware artifacts and acquired image artifacts based on these traces [9]. The artifacts found in modified images, such as object shape removal, contrast value alteration, sensor pattern noise, and application of connection with altered picture information and genuine image, are the focus of traditional image forgery detection techniques [10], [11]. The forgery detection cycle shown in Figure 1.4.

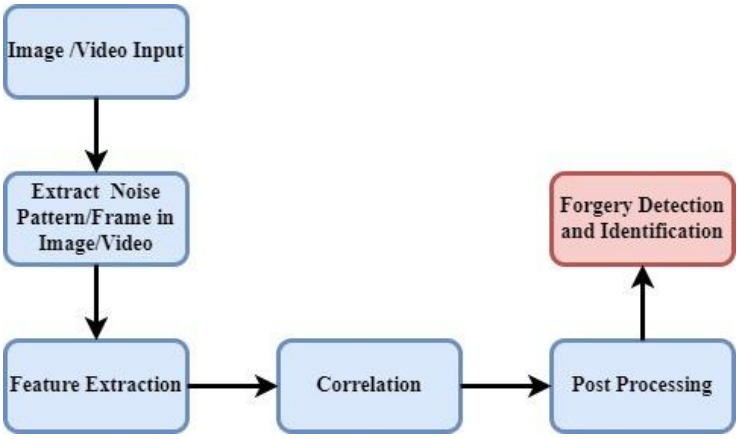


Figure 1.4. Image/Video Forgery General Cycle

By using a data-driven approach, CNN has become adept at handling computer vision tasks in recent years. CNN is a data-driven methodology to extract traits from faked images and forecast how the original image will change based on intrinsic attributes. Because of greater pixel value correlation, the CNN-based technique was successful. The CNN model learns characteristics from a trained dataset, identifies significant forgery artifacts, and accurately classifies the forgery [12], [13] . As malware's influence has grown, it is now easier than ever for anyone to post, download, and distribute files online, including audio, image, and video content. This has led to an increase in the amount of video forgeries online. Adobe Photoshop and Video Editor are among the multimedia tools commonly utilized to modify media files. Furthermore, a common method of harmful video forgery involves manipulating a video sequence by adding or removing items within the frame [14].

Recent strides in image and video forensics owe their progress to the ongoing advancements in deep learning, computer vision, and signal processing techniques. Innovations like convolutional neural networks (CNNs) [15], recurrent neural networks (RNNs), and generative adversarial networks (GANs) have been explored by researchers, aiming for enhanced accuracy and efficiency in forgery detection. These technologies have reshaped the landscape, offering promising avenues for more precise and effective detection methods. CNNs have proven valuable in image forensics, extracting intricate features and patterns to identify manipulations or alterations. RNNs excel in analyzing temporal relationships in video data, enhancing deepfake and video forgery detection. Meanwhile, GANs [16], known for generating realistic fake content, are now applied adversarially in forensic contexts, challenging existing methods and refining overall accuracy. Beyond deep learning advancements, blockchain technology's integration offers a promising solution to ensure the integrity and traceability of digital media. By timestamping and storing forensic analyses on

a tamper-resistant blockchain, investigators maintain permanent records of manipulations or alterations, bolstering the trustworthiness of presented evidence [17]. These collective developments empower image and video forensics to address the growing challenges of detecting digital forgeries, preserving the integrity of digital content, and maintaining the trustworthiness of evidence in an ever-evolving digital landscape.

1.3 *Contributions*

This dissertation presents significant contributions to the field of image and video forensics by leveraging deep learning models. In this study, we made use of the recently created VISION dataset, which contains over 35,000 images and videos collected from 35 various portable devices made by 11 major manufacturers. It is noteworthy that many existing datasets intended for picture forensics do not contain photographs taken at various times and with diverse levels of quality. This gap is filled by the VISION dataset, which offers an extensive and varied collection of multimedia content recorded using a variety of cameras and recording settings. As it includes a wide range of real-world scenarios, this special dataset enables more thorough and realistic evaluation of image source identification techniques. Researchers can use it to assess how well deep learning-based methods handle varying image qualities and temporal factors. Conventional image forgery detection centers on altered image artifacts like object tampering, contrast changes, sensor pattern noise, and correlating forged and authentic image data. These are key areas scrutinized in detecting falsified images. By using a data-driven approach, CNN has become adept at handling computer vision tasks in recent years. CNN uses a data-driven methodology to extract characteristics from fake images and forecast how the genuine image will change based on intrinsic parameters. Because of greater pixel value correlation, the CNN-based technique was successful. The CNN model finds significant artifacts of a forged image, learns features from a trained dataset, and accurately classifies the forged image. The major findings mentioned above align with the following research objectives.

i. The image source identification/detection from small cluster capturing devices has been studied but images obtained from large cluster of capturing devices with high dynamic range images has not been analyses yet.
ii. mage forgery detection in shared information using a deep convolutional neural networks model.

iii. Technique to analyze the video capturing source and enhancement in accuracy of the video capturing device using deep learning model.

1.4 *Thesis Outline*

This thesis is structured into chapters to facilitate comprehension. Here's a concise overview of each chapter's content for clarity and coherence.

CHAPTER 1: Provides an introductory overview of key concepts related to digital forensics, image processing pipelines, image forgery, and image/video source identification. Furthermore, it delves into the application of deep learning models to enhance accuracy in identifying image forgeries and their original sources, particularly focusing on Convolutional Neural Networks (CNNs). Finally, the chapter outlines the problem statement and articulates the research objectives of the thesis.

CHAPTER 2: Give insight a comprehensive review of the literature, offering insights into the various aspects of digital image forensics, image processing, and the use of different machine learning and deep learning models. It also lays the groundwork for the research conducted in this thesis by identifying research gaps and emphasizing the role of CNNs in image and video analysis. The knowledge gained from this review serves as a valuable foundation for the subsequent chapters, which will delve into the methodology, experimentation, and contributions of the research.

CHAPTER 3: Presents a novel strategy for image source identification by introducing a Twin Convolutional Neural Network Architecture (TCA) designed to enhance the accuracy of source identification. Within the TCA framework, the initial CNN architecture, referred to as DnCNN, is utilized to remove the unknown level noise from the original dataset, creating 256x256 patches for the training and testing phases. Subsequently, the second CNN architecture is engaged to classify images by leveraging features extracted from multiple convolutional layers using a 3x3 filter, thereby enhancing the efficiency of predictions.

CHAPTER 4: Proposes novel approach utilizes a CNN-based architecture to classify image forgeries, demonstrating a unique capability to identify even previously unencountered forgeries by extracting features from various convolution layers. The integration of an SVM classifier ensures high-precision labeling of forged images. This methodology not only enhances the accuracy of image forgery detection but also extends its applicability to emerging types of forgeries.

CHAPTER 5: Proposes Two unique camera model recognition techniques were developed

using Convolutional Neural Networks (CNNs) for deployment within an enhanced multi-modal framework. The developed multi-modal approaches amalgamate both audio and visual data to tackle the identification challenge of original device, demonstrating clear superiority over mono-modal methods, which solely rely on either visual or audio cues from the examined video sequence to perform the identification.

CHAPTER 6: It outlines the thesis's conclusion and forecasts its future scope, encapsulating the significant achievements and delineating potential avenues for further exploration and development.

CHAPTER 2

PRELIMINARIES AND BACKGROUND

2.1 *Introduction*

Digital image forensics has emerged as a pivotal discipline in recent times, primarily driven by the ubiquitous adoption of digital images and the accessibility of sophisticated image editing software. The fundamental objective of image forensics revolves around the precise identification, comprehensive analysis, and conclusive authentication of digital images to safeguard their integrity and establish their veracity. This comprehensive literature survey delves into the essential technical methodologies, advanced techniques, and formidable challenges encountered in the realm of image forensics. In the digital age, images and videos have taken over as the primary information bearers. Visual media are increasingly being used to transmit information, even rational knowledge, due to its expressive power and simplicity of acquisition, delivery, and preservation. As a result, pictures and videos are now often used as evidence in both court cases and disagreements in daily life [2]. The primary subject areas of image source model identification, digital image forgery detection, and video frame forensics analysis are covered in this introduction to the developing discipline of digital image forensics. In source camera identification, we aim to pinpoint the specific camera model—or the precise camera—that captured the captured picture. Establishing a picture's validity or revealing any possible manipulation with the image is the aim of forgery detection [18].

2.2 *Related Work*

This study focuses on scrutinizing methodologies within image and video forensics, particularly exploring the innate characteristics within digital images throughout their life cycle. It delves into the transition from traditional image processing to the cognitive image processing paradigm, aiming to analyze these evolving approaches comprehensively. The investigation aims to uncover the distinct footprints left by these methodologies as they evolve and adapt to the shifting landscape of image processing techniques.

2.2.1 *Digital Image Formation Pipeline*

An image is described mathematically as a function f[x, y], with x and y as spatial coordinates and f[x, y] as the intensity value at that position. In digital form, the image becomes an array of integers, a 2-dimensional array (f(x, y)) depicted in Figure 2.1. Here, x ranges from 0 to the image's height (h-1), and y ranges from 0 to the image's width (w-1). The intensity values within this digital image fall within the range of f(x, y) $\in$ [0, L-1], where L represents the maximum intensity value. This representation as an array allows the discrete handling of the image's visual information, aiding in its processing and analysis, essential in various image-related tasks. The discrete nature of digital images enables computational methods to manipulate and interpret visual content efficiently, paving the way for extensive applications in image processing, computer vision, and other domains reliant on visual data analysis, where L-1 is equal to 255 for an 8-bit image, indicating the maximum intensity level [19].

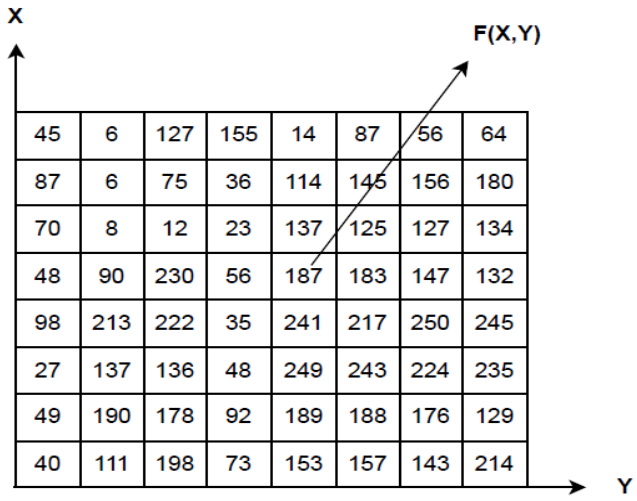


Figure 2.1. Image Pixel Array Representation

Image processing involves the acquisition of real-world scenes, which are then stored in a compressed digital format on various devices. However, a major challenge in this process lies in preserving the integrity of the image throughout the post-processing cycles Figure 2.2. Ensuring the image's integrity becomes crucial, as multiple operations are applied to the image during post-processing, and any degradation or alteration can impact the accuracy and reliability of the visual information [20].In the context of digital image processing, the acquired information undergoes conversion by diverse digital image processing units to be represented in digital form. Silicon-based sensors play a crucial role in this process, where

they convert the incident light intensity into analog signals. The widely adopted CFA pattern in digital image capture devices is the Green-Red-Green-Blue (GRGB) Bayer pattern. This mosaic arranges pixels with varying intensities of red, green, and blue in a specific pattern [21]. Since each pixel in the Bayer pattern only captures one of the three primary colors, Digital Image Processing (DIP) utilizes various interpolation algorithms, known as demosaicing, to reconstruct a full-color image. Other than the prevalent Bayer pattern, alternative Color Filter Array (CFA) configurations like Cyan-Yellow-Green-Magenta (CYGM), and Cyan-Magenta-Yellow (CMY) can function as alternatives. In addition to demosaicing, Digital Image Processing (DIP) integrates supplementary methods to elevate image quality. These techniques encompass advanced white balancing of captured image, profound noise level reduction, implied different matrix manipulation, image intensity sharpening, aperture correction, and fine-tuning gamma correction. These processes collectively refine image attributes, ensuring higher quality and improved visual fidelity in the final output. These enhancements play a crucial role in addressing various image imperfections and enhancing overall image quality, contributing significantly to the realm of digital image processing and its diverse applications across industries. [22].

Overall, this intricate process of image transformation through sensor capture, CFA utilization, and advanced digital image processing techniques ensures the production of high-quality and visually accurate digital images.

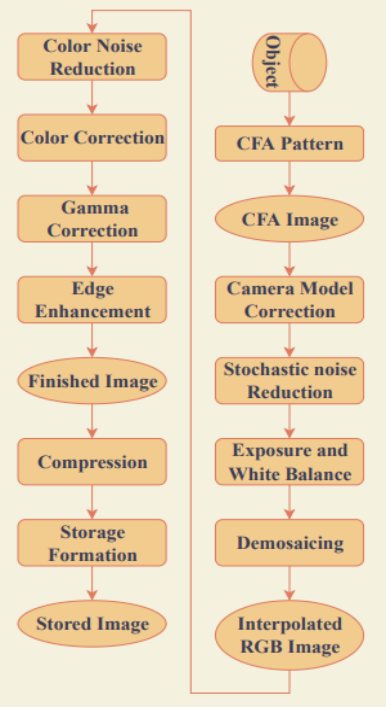


Figure 2.2. Digital Image Processing Cycle

2.3 *Image Processing Paradigm*

In the 18th century, people started finding ways to turn real scenes into digital or similar signal formats through what became known as Traditional Image Processing. A major moment came in the early 1920s when Bartland achieved a groundbreaking feat. They transmitted the very first digital image through an undersea cable, sending it from London to New York. At the other end, the image was reconstructed, marking a huge leap forward for handling images with computers. This breakthrough opened the door for computational image processing methods. These included various techniques like Image Enhancement, Restoration, Color Modification, Contrast Adjustment, and more. They were all designed to make images look better and clearer. Fast forward to the 1940s, computers began to emerge, allowing for manipulation of algorithms and new ways to improve image quality. This was also when images started to be stored as arrays of bits, changing how we saved and handled visual data. The 1960s saw another leap with the progress of satellite imaging. Now, satellites could capture images of landscapes and even identify objects from space. By the 1970s, digital image processing found extensive applications in the medical field, significantly impacting medical imaging. During the 1980s, image processing found its way into many different fields like creating artistic effects, visualizing medical data, inspecting products in industries, and aiding law enforcement [2][3]. As the 20th century went on, image processing kept growing in areas like automotive technology, computer vision, and specialized industries, showing how flexible and useful it could be. In Figure 2.3, you can see the shift from the old-school way of handling images to a more advanced cognitive image processing system. This change represents how image processing techniques and methods have evolved and improved over time. This transition has played a pivotal role in unlocking new possibilities and applications in the field of image processing [23].

2.3.1 *Traditional Image Processing*

In the early 1700s, image capture began with the invention of the camera obscura, utilizing a pinhole to project inverted real object images. Advancements like the Optical Camera Obscura expanded this by duplicating images through mirrors [24]. Soon after, scientists experimented with metal plates coated in various chemicals to capture scenes. In the mid-1700s, the Calotypes method emerged, utilizing white shells to refine image quality, yet motion picture capture remained a challenge [25]. Early image processing involved time-consuming conversion of light intensity into analog form and storage in electronic devices.

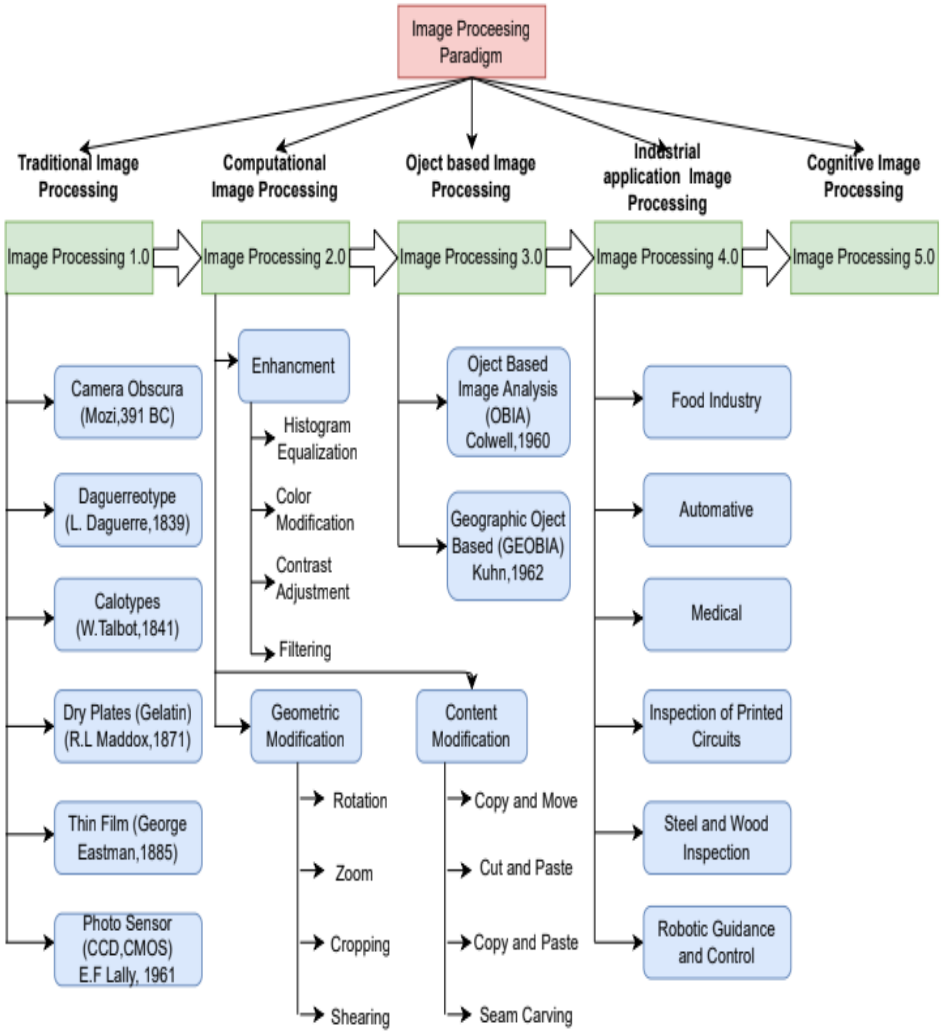


Figure 2.3. Image Processing Taxonomy.

Thin films were eventually employed to expedite capture and streamline commercial image production. A significant advancement occurred in 1961 when photo sensors were capable of converting light intensity into digital form [26]. This marked a pivotal shift in image processing paradigm to provide a comprehensive understanding of the evolution of traditional image processing.

2.3.2 *Computational Image Processing*

Computational image processing represents a significant advancement over traditional methods, leveraging diverse algorithms to construct images of interest. This approach ensures seamless integration between the acquisition phase and computational operators, resulting in enhanced picture quality with higher resolution. Computational processing finds extensive applications in various domains, such as medical imaging, Synthetic Aperture Radar (SAR), seismic imaging, and high dynamic range (HDR) images [27][28][29]

2.3.3 *Object Identification-Based Image Processing*

The essence of identifying objects or targets holds a central position in the image processing 3.0 framework, encompassing various processes of capturing images and videos through diverse devices like smartphones, satellites, and robotic image systems. This approach empowers researchers to implement real-time applications in computer vision, unlocking numerous possibilities across different fields. Object detection has seen substantial utilization in diverse applications, ranging from video surveillance, image captioning, and robot vision to enhancing digital camera positioning, satellite image analysis, drone scene examination, road accident detection, enemy spotting in military operations, autonomous driving, and computer-human interactions [55]. Its broad spectrum of applications signifies the pivotal role of object detection in advancing computer vision technologies and its significant contributions to solving real-world challenges and automation. This task involves analyzing input images to identify specific objects like vehicles, obstacles, aerial entities, animals, or buildings within suggested regions in both images and videos, is performed by the object detection process, which is a crucial technology linked with visual analysis and image processing. Object detection is frequently used to identify static objects, but in recent years, researchers have begun to focus on moving objects by utilizing advanced machine-learning techniques. Object identification is also crucial in biometric methods like iris and face recognition, fingerprint recognition, and locating moving objects in videos. Many industries, including manufacturing, human resources, healthcare, autonomous driving, and others, employ target detection. Figure 2.4 depicts a typical detection of an object scenario that follows:

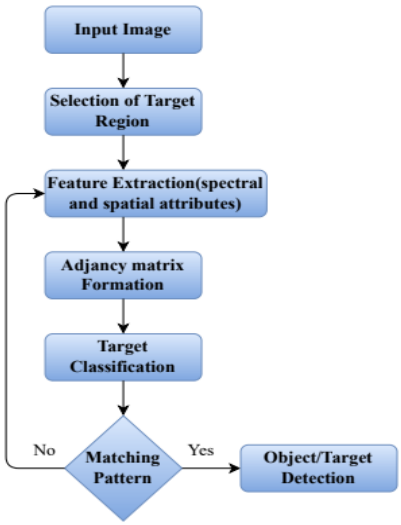


Figure 2.4. Object Detection Analysis

Traditional object identification methods involve the sequential extraction of pixel values from the source image, followed by correlation analysis to identify objects within the image[56]. However, this approach typically yields low accuracy, ranging from 40% to 50%. To significantly improve accuracy, advanced techniques such as machine learning and deep learning have been employed, resulting in substantial enhancements with accuracy levels reaching 80% to 99% [55][57], [58]. Figure 2.5 presents a taxonomy outlining various object detecting methods, providing a structured overview of the different approaches and their respective characteristics.

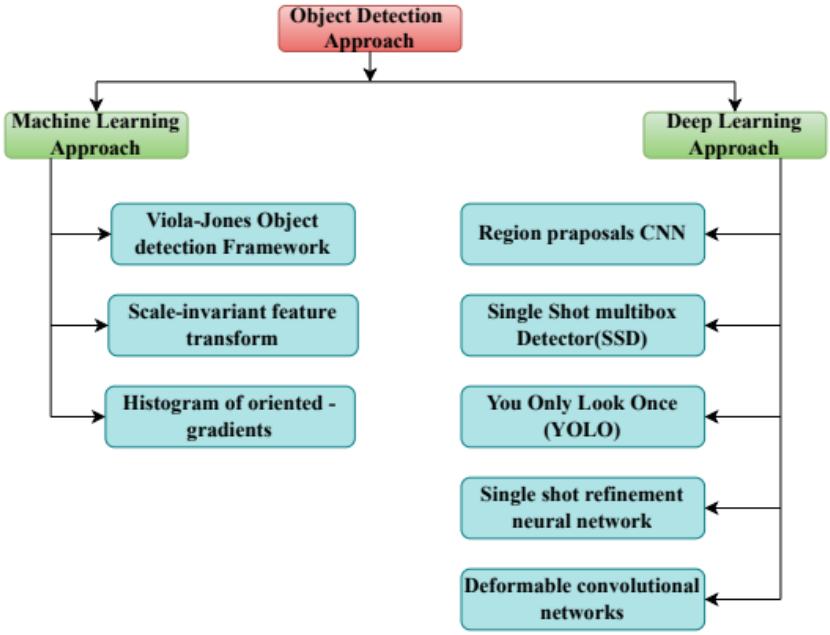


Figure 2.5. Object Detection Techniques

2.3.4 *Industrial Automation Based Image Processing*

In the 20th century, the advancement of image processing has played a pivotal role in driving industries towards automation. Classical image processing techniques are employed to create accurate representations of genuine object pictures using various formation approaches. In Industry 4.0, intelligent systems are integrated with smart vision systems, enabling precise control over production quality, reduced labor requirements, and increased output efficiency. Industrial smart vision systems synergize computational processing with advanced object recognition methods to facilitate automated production inspection, quality control, error reduction, shorter production times, part identification, robotic control implementation, and real-time monitoring of assembly line output [59][60].

2.3.5 *Cognitive Image Processing Paradigm*

In the realm of automotive computer vision systems, image processing assumes a vital role in tackling intricate human challenges. The modern digital landscape witnesses an overwhelming influx of image and video data transmitted across various devices and networks, inundating the online sphere with vast digital information. Yet, existing vision systems encounter difficulties in managing and processing this extensive data flow. Traditional image processing primarily aims to improve image resolution and furnish comprehensive information. However, in legal contexts, distinguishing authentic data from altered information poses challenges due to the application of computational image processing operators tailored to user specifications, potentially modifying original data. On the contrary, cognitive image processing introduces an innovative approach by directly extracting data and insights from readily accessible digital sources online, offering a promising avenue for more astute and precise data analysis [60]. Cognitive image processing has proven its worth across various IT domains, showcasing its effectiveness in tasks like text extraction, image comprehension, spatial analysis, and facial recognition. This technology leverages advanced algorithms and deep learning methodologies to intelligently process and interpret images, enabling accurate and efficient extraction of textual information, comprehensive picture comprehension, spatial analysis of complex scenes, and reliable facial recognition capabilities. Its applicability in these diverse fields highlights the potential for cognitive image processing to revolutionize various industries and contribute to the advancement of artificial intelligence and computer vision technologies.

2.4 *Digital Image Forensic*

Digital forensics encompasses diverse scientific methodologies and techniques utilized to trace the original source and guarantee the legitimacy of digital data. In the contemporary digital landscape, ensuring the credibility and trustworthiness of visual information presents substantial hurdles for forensic specialists [61]. These digital forensic methods find extensive use in analyzing image and video data, aiming to validate the authenticity and source of information, starting from the initial image capture phase through storage on the original device and across every step of the digital image processing workflow.[62]. The core objective of digital forensic investigations is to establish the trustworthiness and accuracy of digital evidence in a manner that is admissible in legal proceedings. This is particularly crucial in cases involving cybercrimes, data breaches, intellectual property theft, and other

digital offenses. Forensic experts employ robust methodologies and specialized tools to analyse digital images and videos, ensuring that the evidence can withstand legal scrutiny and maintain its evidentiary value. The first step in digital forensic analysis is often image acquisition, where investigators gather the digital media that is relevant to the investigation. This process involves preserving the integrity of the data, adhering to strict chain of custody procedures to ensure that the evidence remains unaltered throughout the investigation. The acquisition phase also involves recording metadata and identifying crucial information such as timestamps, camera settings, and other pertinent details that may aid in the analysis[63]. Once the images and videos are acquired, forensic experts delve into the various phases of the digital image processing pipeline. This includes examining the original images and their metadata to identify any potential tampering or alterations. Common techniques employed in this phase include noise inconsistency analysis, copy-move forgery detection, and detecting inconsistencies in compression artifacts.

Moreover, digital forensic analysis delves into the examination of image processing operations applied to the images. These operations can include resizing, cropping, filtering, and various enhancement techniques. The goal is to determine if any of these operations have been utilized to manipulate the visual information in any way. Deep learning and machine learning models have become increasingly valuable tools in digital forensics. These models can be trained to identify patterns associated with image manipulation, recognize image sources, and detect specific artifacts indicative of tampering. Such techniques have significantly enhanced the ability to detect forgeries and establish the authenticity of visual evidence. Another crucial aspect of digital forensic analysis is steganography detection. Steganography involves hiding information within images or videos in a way that is imperceptible to the human eye. Forensic experts utilize sophisticated algorithms to detect and extract hidden data, ensuring that no incriminating information remains concealed. To further validate the integrity of the evidence, digital forensic experts also focus on error level analysis and analysing inconsistencies in lighting and shadows within the images. These methods can reveal potential manipulation attempts or digital compositing[64].

In conclusion, digital forensics plays a pivotal role in ensuring the accuracy, authenticity, and integrity of visual information in the current digital era. Through meticulous examination of image and video data, employing advanced techniques, and leveraging cutting-edge technologies like machine learning and deep learning, forensic experts can detect forgeries, establish the original source of the data, and provide crucial evidence in legal proceedings.

The continuous advancement of digital forensic methodologies will remain paramount in upholding the trustworthiness of visual information and safeguarding the integrity of digital evidence in the ever-evolving landscape of digital technologies. The forensic taxonomy obtained from the extensive survey is presented in Figure 2.6 below.

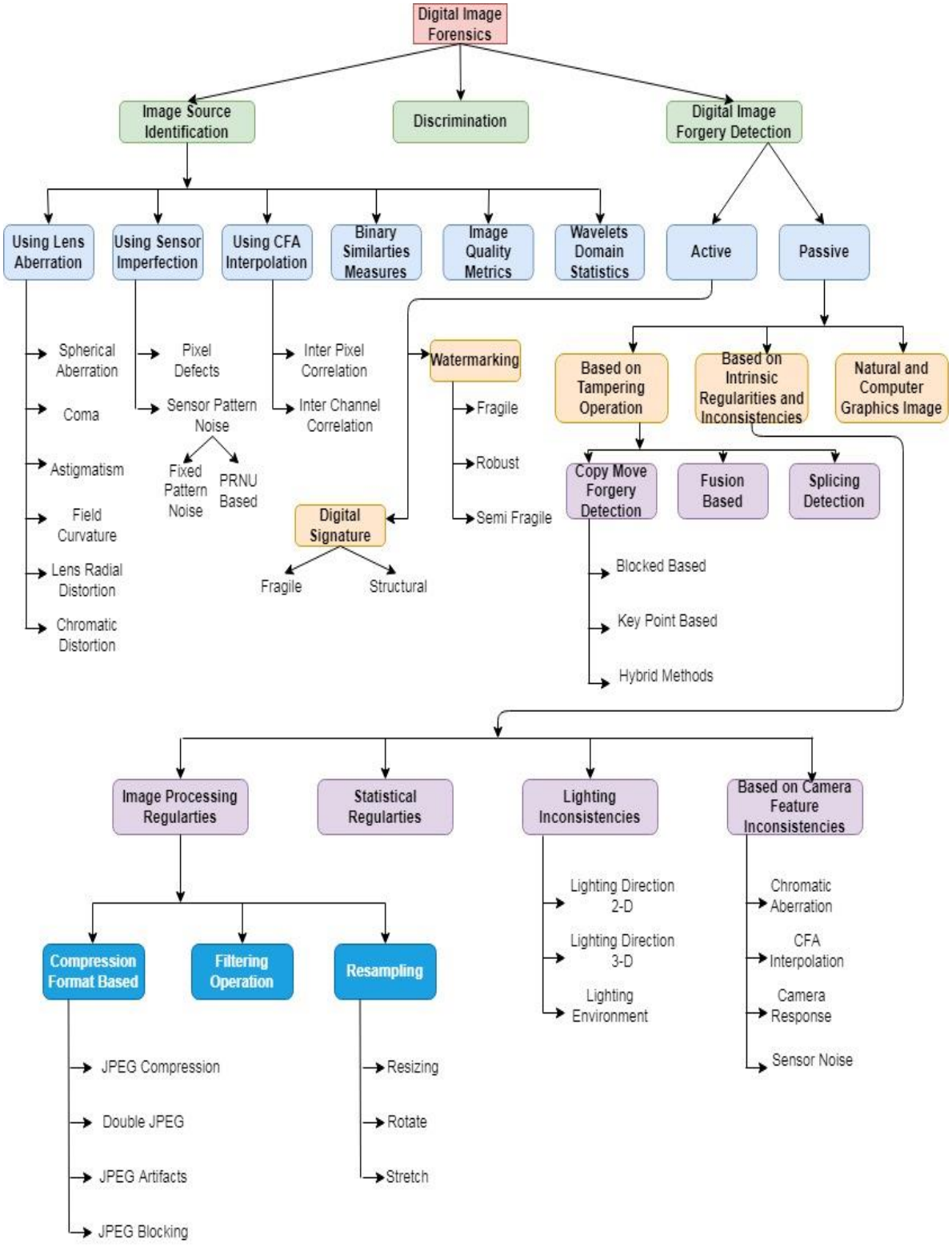


Figure 2.6. Digital Image Forensic Taxonomy

2.4.1 *Image Source Identification Techniques*

Image source identification involves the process of determining the origin of an image through rigorous analysis and techniques given in Figure 2.7. Two fundamental methods are commonly employed for this purpose. The first approach relies on the traditional image processing pipeline tracing. In this method, various stages of the image processing pipeline are analyzed, which includes operations such as compression, filtering, and resizing. By examining the distinct patterns left by these operations, it becomes possible to infer the source from which the image originated. However, this approach may face challenges when images have undergone multiple transformations or have been subjected to significant post-processing. The second approach revolves around feature extraction-based techniques. In this method, distinctive features, such as statistical properties, noise patterns, and camera sensor fingerprints, are extracted from the image. These features serve as unique identifiers for different image sources. Advanced machine learning algorithms are often employed to classify images based on these extracted features, enabling accurate identification of the image source[65]. To improve the effectiveness of image source identification, researchers have developed a diverse range of approaches. These include methods that integrate multiple features, utilize deep learning models for more precise classification, and consider specific imaging devices or platforms. The field of image source identification remains an active area of research, with ongoing efforts to enhance its reliability and applicability in various real-world scenarios, such as forensic investigations, copyright protection, and fake image detection.

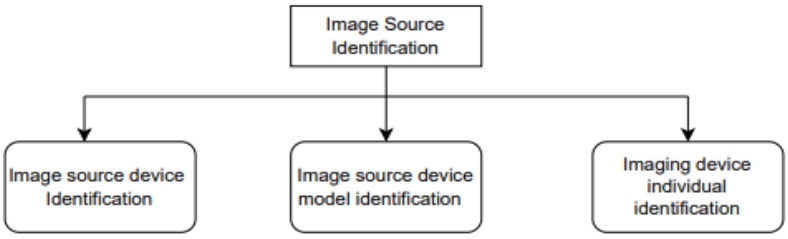


Figure 2.7. Image Source Identification Approaches

The traditional image capture and processing approach involves several distinct steps that culminate in the storage of the image in a compressed format within the device's database. During the digital image life cycle, the image is initially captured by a digital device, and the incident light intensity from the object interacts with the color filter array (CFA). The CFA serves to convert the incident light intensities into specific color values for each channel

(Red, Green, Blue). Subsequently, various interpolation techniques are employed to convert the individual channel data into a single-colored image, where each color's intensity is represented within the range of 0 to 255. This transformation enables the image to be visually interpretable by human observers and facilitates further processing and storage. Figure 8 visually depicts the sequential stages of the traditional image processing pipeline in Figure 2.8 associated with a specific imaging device. These steps encompass image capture, color filtering, interpolation, and the final representation of the image in the device's database [66]. This traditional approach to image processing has served as a fundamental foundation for image source identification and understanding the characteristics of digital images generated by various devices.

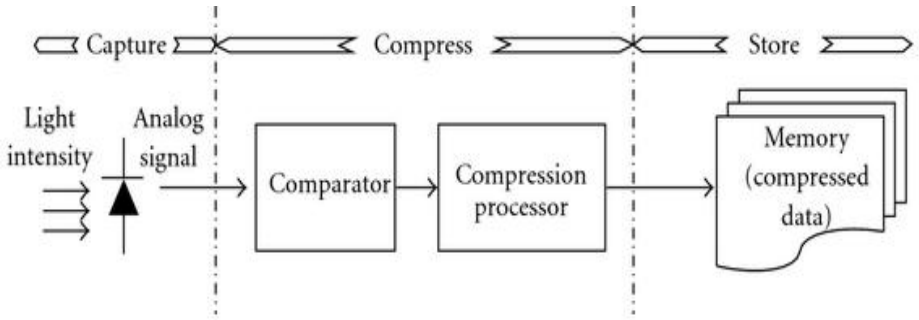


Figure 2.8. Digital Device Internal Processing

2.4.1.1 *Conventional Approach for Image Source Identification*

In traditional image forensics relies on the identification of source-specific footprints and intrinsic artifacts present in images during the acquisition, compression, and editing phases. During the image acquisition phase, researchers closely examine the characteristics of the lens, sensors, and CFA (color filter array) interpolation techniques. In the image acquisition phase, light rays are captured by the lens, and the sensor array pattern converts them into continuous signal form. Unique lens systems, sensor qualities, and color filter array demosaicing techniques are found in each camera model. During the manufacturing process, lenses produce various types of artifacts, leaving behind distinctive traces that enable camera model identification. Different digital device components imprint diverse footprints in captured images, and researchers establish correlations between these device artifacts and the artifacts present in the captured images. Notably, each camera model exhibits lens distortion artifacts, which prove valuable in camera model identification. Radial lens distortion is linked to optical systems and influenced by nonlinear geometrical parameters, such as the lens's focal length and shape. Chromatic aberrations, dependent on the lens dispersion index and

varying wavelengths following Snell's Law, also contribute to the identification process. Vignetting, which causes intensity fall-off towards the corner of the image, likewise leaves discernible traces used for identification purposes. By thoroughly analyzing these footprints and artifacts, researchers in traditional image forensics can effectively trace the origin and history of digital images, aiding in various applications, including authentication, tamper detection, and source attribution [67].

In image source forensics, each camera model exhibits distinct internal footprints caused by sensor defects during the manufacturing process. These sensor defects introduce noise into images, which serve as essential clues for forensics analysis. To identify the camera model, researchers estimate the sensor pattern noise within the image by employing denoising filters. Subsequently, they correlate this noise pattern with the original image.

A commonly utilized feature for image source forensics is Photo Response Non-Uniformity (PRNU) noise. By estimating the residual noise present in the image, researchers can effectively extract the PRNU noise [68]. This involves a systematic process that includes various steps, which are outlined below:

- First, the image is preprocessed to remove any artifacts or irrelevant noise that could potentially interfere with the PRNU estimation given in Eq 2.1.
- Next, a denoising filter is applied to the preprocessed image to suppress noise and enhance the detection of the underlying PRNU noise pattern.
- The denoised image is then correlated with the original image to determine the specific PRNU noise unique to the camera model.
- Statistical techniques and algorithms are often utilized to refine the estimation process, ensuring robust and accurate identification of the PRNU noise.

$$R = I - F(I) \tag{2.1}$$

Here R is the residual noise of the image, I is the original image and $F()$ is the denoising filter applied on the original image and obtain a denoise image using low pass filter or various other denoising techniques. To identify the original source and determine a correlation between the noise pattern and the test image,[69] employed the sensor pattern noise estimation technique. Implement the technique and determine the highest likelihood estimate of PRNU noise in a certain picture in order to find the original source [70]. Because each picture contains varying noise in various places, [71] uses the local information of the image to locate the source and introduces a method to choose the best region of the image.[68]

determine the weighting factor for boosting the SPN to get the SPN magnitude that is inversely proportionate. Using pair-wise magnitude relations of the obtained picture with its residual noise, [72] suggested a unique method of source detection. To determine the original camera model, [73] made advantage of the dust-spot function. Suggested using the Minimum Mean Square Error (MMSE) approach for calculating PRNU using wavelets. To determine SPN elements and determine the image's original source, [74] recommended using principal component analysis. [75] provide a formula for computing inter-channel demosaicing artifacts. A suggested technique by [76] that uses an SPN predictor and content-adaptive interpolation calculates the value of the center pixel from its neighbors' pixels. To extract the SPN from various photos, [77] various filters were utilized. After examining all of the conventional methods, anticipate the accuracy comparisons shown in the accompanying chart Figure 2.9 and Figure 2.10.

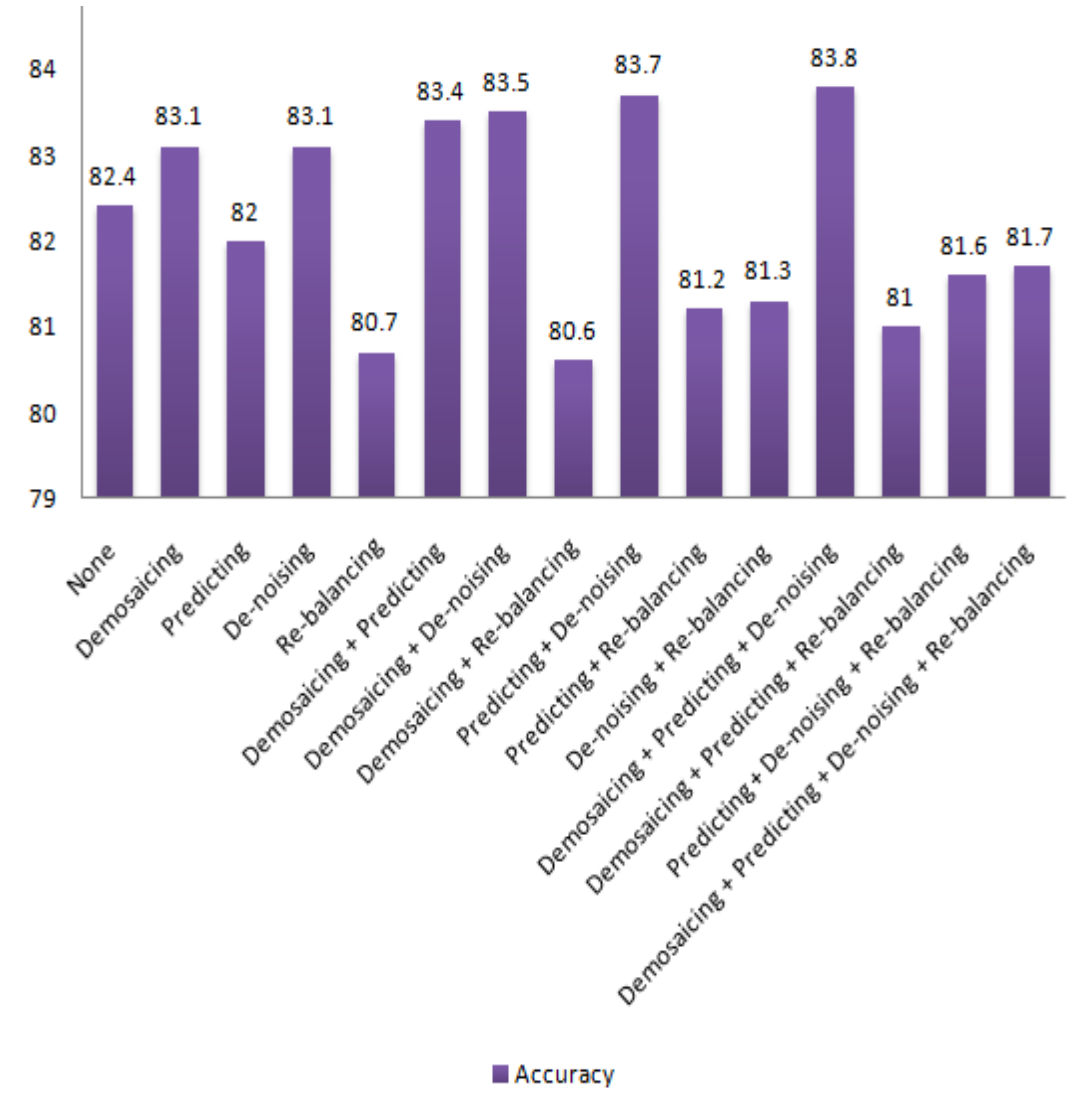


Figure 2.9. Conventional Approach Accuracy Chart (%)

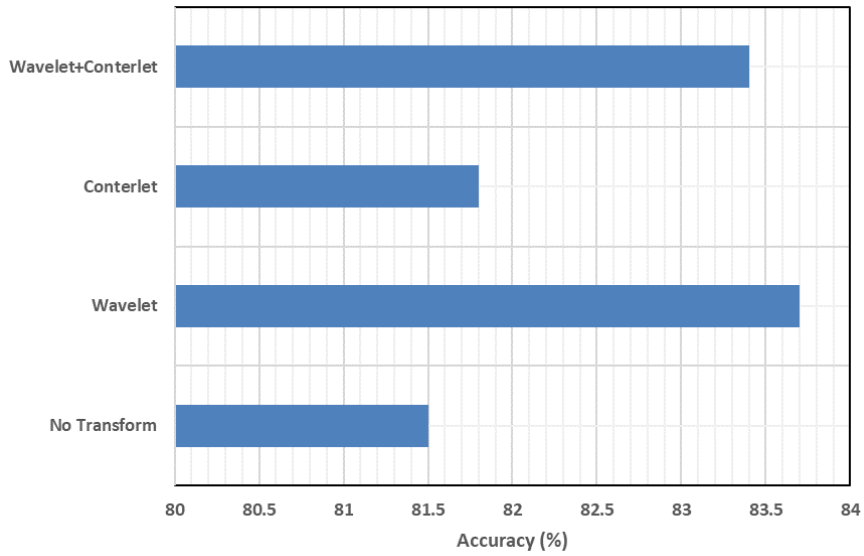


Figure 2.10. Image Transformation-Based Accuracy Prediction

[78] outlined a unique method for information extraction and employed binary similarity measures along with HOWS to identify the source model. In order to determine uniform gray-scale invariant in picture texture owing to device hardware and its interpretation technique causing artifacts, [79] created Local Binary Pattern (LBP). [80] suggested a WLBP operator that combined LBP with various excitations. Use CFA interpolation traces using a minimal mean square estimate to determine the internal footprints [81]. In order to improve robustness, [82] presented a unique method based on CFA interpolation with 1022.

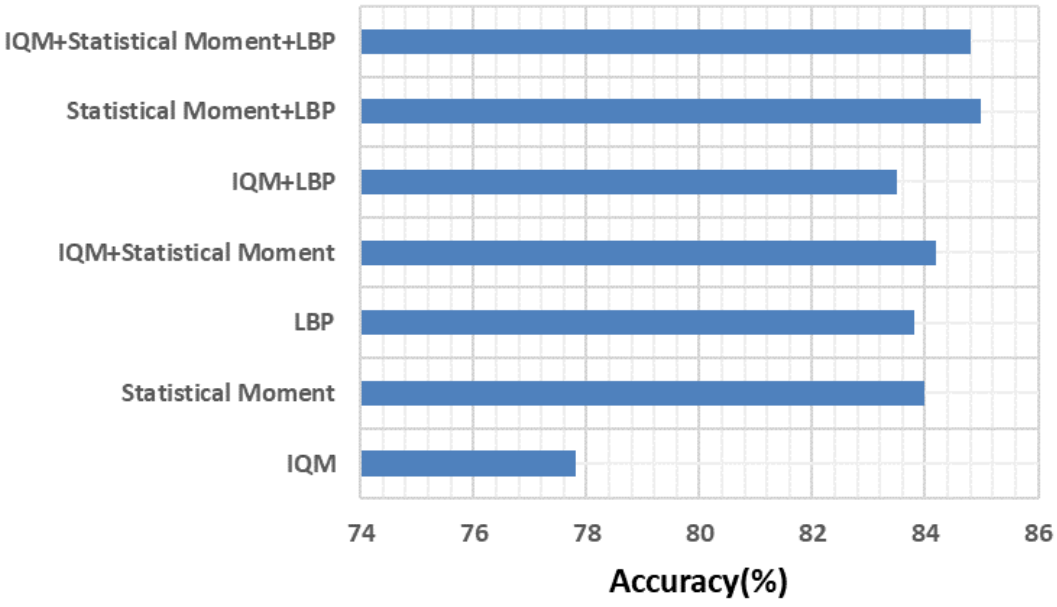


Figure 2.11. Comparative Analysis of Local Image Features

coefficients. Accuracy prediction based on picture pre-processing, image transformation, and local image feature stages after study of inherent artifacts of the image processing pipeline shown on Figure 2.11. Some classifiers are also used to forecast the original source and categorize the picture source based on feature extraction. The following provides an accuracy prediction study of several classifiers in Figure 2.12:

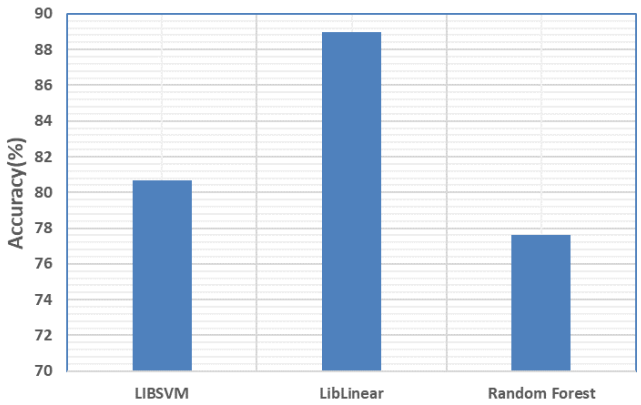


Figure 2.12. Different Classifiers Accuracy Prediction

2.4.1.2 *Deep Learning Based Approach for Image Source Identification*

A variety of machine learning techniques have been used recently to increase image source detection accuracy compared to earlier methods. In which the model is trained to identify the provided input picture while features are taken from the image dataset. Apply several denoising filters to the picture dataset to recover the original image after which classifiers are used to find the original source and boost prediction accuracy. Comparisons of accuracy enhancement following the use of various filters are shown in the Figure 2.13.

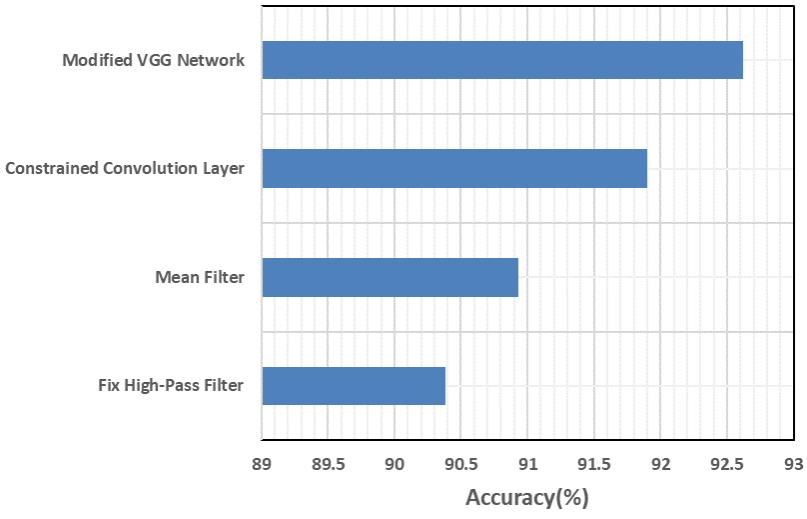


Figure 2.13. Data-driven Approaches Comparison

Deep learning models perform better on computer vision issues as a result of the quick growth of artificial intelligence. The identification of the camera model or brand is a difficult problem in picture forensic science. In order to increase the accuracy of the predictions, the deep learning technique uses a data-oriented paradigm as opposed to the inherent characteristics of the camera or picture. Large datasets are better served by these models, although prediction time is longer. It's essential to choose the features from the dataset before designing a specific deep learning model for a specified job. A basic convolution network with 3 convolution layers and 2 fully connected layers was used by [83] to establish the first deep learning model to identify picture sources, and accuracy was noted at 72.9%. [84] Use the Leaky Rectified Linear Unit (L-ReLU) and enhance the number of layers to attain improved accuracy. [85] presented a CNN architecture with linked layers and additional convolutional layers. In order to improve accuracy, batch normalization is also used.[86] Create a model based on Residual Neural Network (Res Net) to better accurately determine the source. It boosts the model's layer count to 26, which improves the potential of making predictions with an estimated 99% accuracy.[87] To locate small-size picture source cameras, use ResNet to examine saturated photographs, smooth images, and other images. [88] To increase accuracy and rate the degree of similarity between two input images from different sources, two neural networks were used [89]. A denoising convolution neural network (DnCNN) model was proposed to link the device footprint based on maximum likelihood estimate and the picture footprint relying on sensor pattern noise, pick the best attributes for model fitting, and improve accuracy forecasting compared to previously applied models. The camera source identification using convolution neural network (CSI-CNN) method picks the best patch from the picture, calculates SPN for all of the patches, and then adds residual blocks to the network structure to boost accuracy. Recursively extracting camera characteristics from several CNN layers is how [90] suggested camera attribute classifier works. In the Table-2.1, a number of deep learning models created using various datasets to identify the original source are compared.

Accuracy increased above traditional and classifier-based methods after investigation of several deep learning model architectures as ResNet, XceptionNet, and DenseNet using various authentic datasets. Changes in convolutional layers, linked layers, and activation functions in various topologies cause variations in accuracy. The methodologies based on deep learning are data-driven and do not rely on the internal workings of the gadgets. The

accuracy forecast of all recently developed deep learning models is studied and shown in the Table 2.1.

Table 2.1: Deep learning models for image source identification[91].

Architecture	Input Size	Convolution Part			Fully Connected Part		
		Layers	Activation	Pooling	Layers	Activation	Dropout
A1	48×48×3	3	ReLU	Max	1	ReLU	Y
A2	32×32×3	2	L-ReLU	Max	2	L-ReLU	Y
A3	36×36×3	3	ReLU	Avg	1	ReLU	Y
A4	64×64×3	13	ReLU	Max	2	-	Y
A5	256×256×3	1 conv 12 Residual	ReLU	-	-	-	-
A6	64×64×3	4	-	Max	1	ReLU	-
A7	64×64×3	10	-	Max	1	ReLU	-
A8	256×256×2	4	TanH	Max, Avg	2	TanH	-
A9	256×256	3	ReLU	Max	2	ReLU	Y
A10	256×256×3	3	ReLU	Max	2	ReLU	Y
A11	64×64×3	6	ReLU	Avg	-	-	-
A12	64×64×3	1 conv 3 Residual	ReLU	Avg	-	-	-

Table 2.2: Deep learning architectures analysis with accuracy prediction

Architecture	Input Size	Classifiers	Train: Test	Dataset	Model Accuracy (%)
A1	48×48×3	Softmax	7:3	Dresden	72.9
A2	32×32×3	SVM	-	MICHE-I	98.1
A3	36×36×3	Softmax	8:2	Dresden	-
A4	64×64×3	Softmax	3:2	Dresden	93
A5	256×256×3	Softmax	7:3	Dresden	94.7
A6	64×64×3	Softmax	8:2	Dresden	93
A7	64×64×3	Softmax	-	Dresden	94.93
A8	256×256×2	ET	4:1	Dresden	98.58
A9	256×256	Softmax	8:2	Dresden	98.01
A10	256×256×3	Softmax	8:2	Dresden	97.41
A11	64×64×3	Softmax	4:1	Dresden	94.14

A12	64×64×3	Softmax	4:1	Dresden	97.03

2.4.2 *Image Forgery Detection Techniques*

With the proliferation of digital cameras and smart gadgets, image editing has become easily accessible to anyone. While some alterations, like adjusting brightness or converting to black and white, are innocuous, others can be malicious and damaging, especially when aimed at public figures and politicians. The main motivations behind image fabrication are often driven by sinister intentions, such as disseminating distorted information, promoting immorality and fake news, fraudulently obtaining money from unsuspecting audiences, tarnishing the reputation of well-known individuals, and exerting negative political influence on digital platform users.

Consequently, ensuring trustworthy digital information exchange necessitates unequivocal identification of photographs and videos before their utilization [92]. By enforcing robust image forensics and verification techniques, digital media platforms can mitigate the spread of harmful content and safeguard against the negative consequences of manipulated images. Such measures play a vital role in combating the proliferation of misinformation and malicious practices within the digital landscape. As technology advances and image editing tools become increasingly sophisticated, it becomes imperative to continually enhance image forensics and verification mechanisms to maintain the integrity and credibility of digital media content [93]. Among these, the three most frequent modifications are:

- Copy-move is the method entails copying a specific section of one picture into another.
- Image splicing is the process of copying a section of one image and combining it with another.
- Object removal is a process that involves the elimination of a specific region within an image, followed by the restoration of the surrounding area to fill in the gap. This restoration is achieved by painting or reconstructing the remaining portion, ensuring the cohesiveness and visual continuity of the image.

We strive to precisely detect these adjustments in our work. Figure 2.14 displays the fake picture. Forgery detection systems fall into two primary categories: active (non-blind) and passive (blind) approaches. Active methods require prior information about the image, integrated at stages like capture, acquisition, or post-processing.

Figure 2.14. Digital Image Forgery Example

Techniques like digital watermarking and digital signatures exemplify active forgery detection, embedding identifiable information directly into the image to verify its authenticity. Passive methods, however, don't rely on prior data and instead analyze the image's inherent properties. These techniques scrutinize statistical anomalies or inconsistencies in the image, detecting potential forgeries without prior knowledge. The distinction between active and passive approaches lies in their reliance on prior information; while active methods embed specific data, passive methods infer authenticity based on intrinsic image traits. Both approaches serve as crucial tools in detecting image tampering, contributing to the robustness of forensic analysis in various domains. The choice between active and passive methods often depends on the available information and the desired level of intervention or analysis required in different forensic scenarios. In these methods, specific data is embedded into the image, enabling subsequent validation and authentication processes [94]. Figure 2.15 illustrates the principle of utilizing embedded data to verify the image's authenticity.

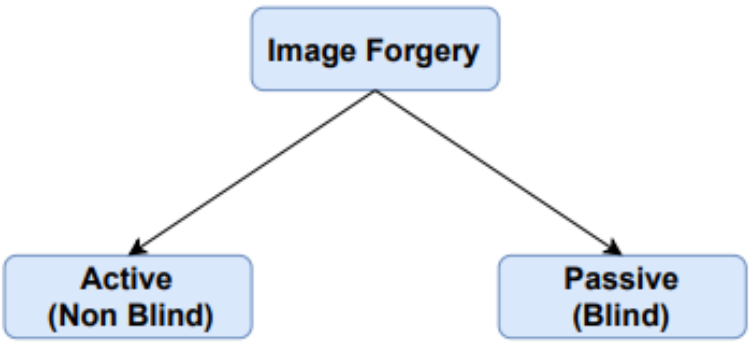


Figure 2.15. Digital Forgery Classification

Active forgery detection has the advantage of higher accuracy and robustness since it operates with prior knowledge of the embedded information. However, its effectiveness relies heavily on maintaining the integrity and security of the embedded data. On the other hand, passive forgery detection systems, also known as blind techniques, do not require any prior knowledge about the image. Instead, they rely solely on analyzing the image's content and statistical characteristics to identify potential manipulations or forgeries. Passive methods are advantageous in scenarios where prior knowledge is unavailable or difficult to obtain, but they may exhibit reduced accuracy and sensitivity compared to active techniques [95]. To enhance forgery detection capabilities, researchers are continuously exploring hybrid approaches that combine the strengths of both active and passive methods. Such advancements aim to provide more comprehensive and reliable solutions for detecting various types of image forgeries and ensuring the integrity of digital media content in diverse applications, including forensic investigations, copyright protection, and digital content authentication.

2.4.2.1 *Non-Blind Image Forgery*

Methods like digital watermarking and digital signatures, as seen in Figure 2.16, are used for active forgery detection. Here's how they work: before an image is sent through an untrustworthy public channel, a specific authentication code is added into the image content. This code acts like a unique tag for the image. Later on, when the image is received, this code can be extracted and compared with the original one that was added before. This comparison helps verify if the image has been tampered with or forged in any way. However, the successful application of this technique requires specialized equipment or software capable of inserting the authentication code into the image before its distribution [92]. Numerous review articles have investigated active forgery detection and have established a hierarchical framework for its classification.

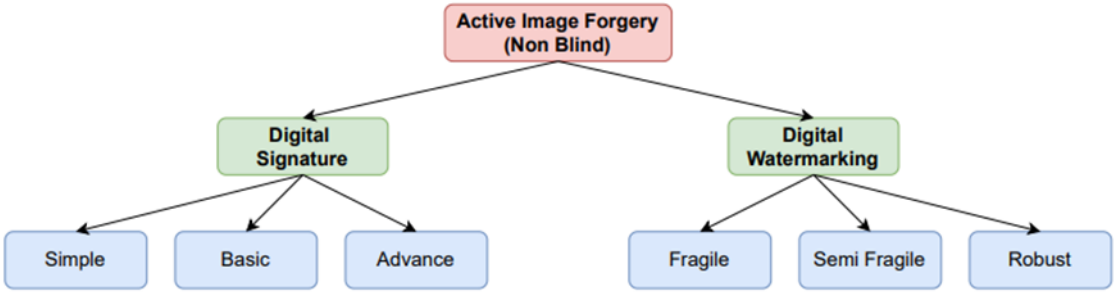


Figure 2.16. Non-Blind Image Forgery Classification.

This hierarchy is structured based on the complexity and effectiveness of different approaches. At the forefront of the hierarchy are techniques that employ robust and imperceptible watermarking, making it difficult for adversaries to remove or alter the embedded information. Within this category, researchers have explored digital signature-based approaches, where cryptographic methods are used to ensure the authenticity and integrity of the image data. In the next tier of the hierarchy, researchers have investigated steganography-based techniques. Steganography involves concealing information within the image data itself, making it less visible to potential forgers. This category includes approaches that leverage various image transformations or frequency domain techniques to embed the authentication code securely. Finally, the hierarchy also encompasses methods that rely on specialized hardware or specific image acquisition processes to insert the authentication code, ensuring tamper-resistant authentication. These techniques offer an additional layer of protection but may impose practical constraints on their widespread adoption due to hardware limitations or complex deployment requirements [95]. As active forgery detection continues to evolve, researchers are actively working on enhancing its robustness, efficiency, and compatibility with various image formats and transmission channels. The ultimate goal is to provide reliable and scalable solutions to counter the rising challenges posed by image forgeries, ensuring the integrity and authenticity of digital media in a wide range of applications, including copyright protection, image forensics, and secure digital content distribution [94].

2.4.2.2 *Blind Forgery Techniques*

Passive or blind forgery detection techniques don't need the sender's signature or watermark to check if received images are genuine. In Figure 2.17, you can see how these methods work. They're built on the idea that even if digital forgeries aren't visible to us, they might still change the statistics or consistency of a natural scene in an image. This could create new differences or strange bits that can be used to spot if something's been tampered with. What's great about passive forgery detection is that it doesn't need any info about the original image beforehand [96]. In real life, current passive forgery detection methods use different ways to find signs of tampering and pinpoint the changed parts in an image. These methods analyze statistical features, noise patterns, or inconsistencies in the image to identify potential alterations. By detecting each sign of tampering separately, these techniques increase their robustness and accuracy in detecting various types of forgeries [92]–[96].

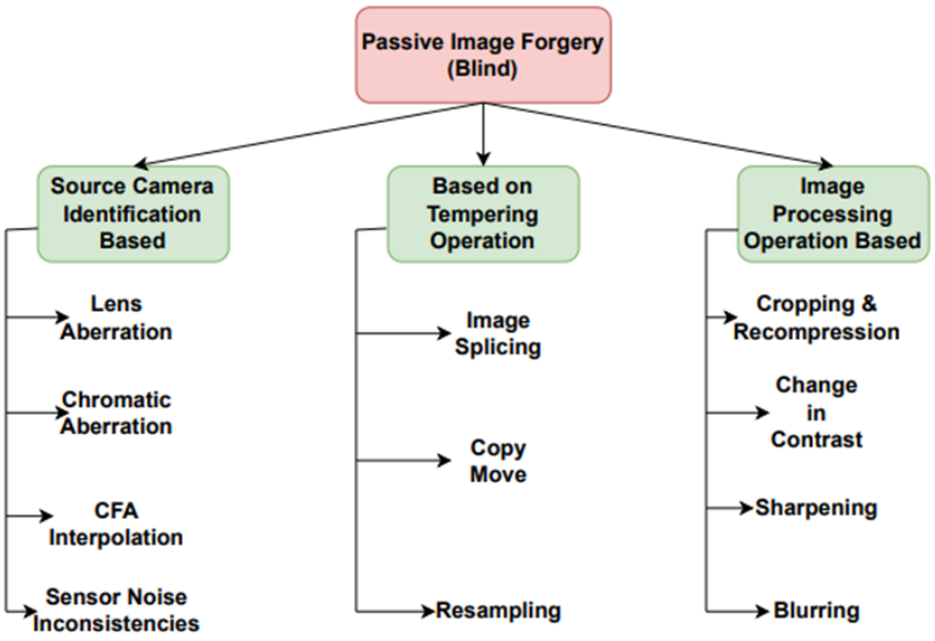


Figure 2.17. Classification Passive Image Forgery

Some common signs of tampering that passive forgery detection techniques may look for include:

1) Inconsistencies in noise patterns or statistical properties, which can indicate regions that have been manipulated.
2) Abnormalities in image compression artifacts, suggesting regions that have undergone editing or alteration.
3) Inconsistent lighting or color inconsistencies that may indicate blending or manipulation of image components.

Passive forgery detection is particularly valuable in scenarios where prior knowledge about the image is unavailable or when images have undergone sophisticated alterations that can bypass active forgery detection methods [97]. By exploiting subtle inconsistencies and artifacts introduced during the forging process, passive forgery detection approaches contribute significantly to the field of image forensics, enabling reliable identification of manipulated images and ensuring the integrity of digital content in various applications, such as law enforcement, digital media forensics, and content verification. Ongoing research in this area continues to advance the effectiveness and versatility of passive forgery detection techniques in combating the ever-evolving landscape of image manipulation and fraudulent activities. In the study of preventing picture fakes, various strategies are available. Some of the older methods rely on particular clues or traces that forged images often leave behind

[97]–[99]. In contrast, modern approaches leverage Convolutional Neural Networks (CNNs) and deep learning methods to address this challenge. Before delving into the deep learning-based strategies, it is pertinent to discuss various conventional methodologies. Conventional forgery detection methods rely on analyzing characteristic artifacts that emerge during the image manipulation process. These artifacts may include inconsistencies in noise patterns, discrepancies in compression artifacts, or irregularities in lighting and color distribution. By identifying and analyzing such artifacts, conventional techniques attempt to detect and locate image forgeries. While these methods have been effective to some extent, they may struggle with complex or well-crafted forgeries that skillfully conceal the typical signs of tampering. In recent years, deep learning techniques, particularly CNNs, have revolutionized image forensics. These approaches are data-driven and can automatically learn intricate patterns and features that indicate image manipulations. Deep learning-based strategies often involve training CNN models on large datasets of authentic and manipulated images to learn to distinguish between genuine and forged content. One of the primary advantages of deep learning-based forgery detection is its ability to adapt and generalize across various types of forgeries. By learning from extensive data, these models can identify subtle and complex alterations that may go unnoticed by conventional methods. However, these deep learning approaches require substantial computational resources for training and can be data-hungry, necessitating access to diverse and well-labeled datasets for effective learning [100]. In conclusion, the battle against picture counterfeiting has seen significant advancements with the emergence of deep learning techniques. While older conventional methodologies continue to be relevant and useful in certain scenarios, deep learning-based strategies offer the potential for enhanced accuracy and versatility in detecting various types of image forgeries. The ongoing development and refinement of deep learning approaches hold promise for further improving image forensics and bolstering the security and reliability of digital media content.

2.4.3 *Video Source Identification Techniques*

In recent times, a proliferation of digital devices has seen the integration of high-quality video cameras, enabling the unhindered and cost-free capture of videos. The surge in digital video usage on various online platforms like YouTube, Facebook, Twitter, and WhatsApp has led to a significant trend. However, this widespread adoption has also brought forth a multitude of security challenges. If unattended, these challenges could have severe consequences, particularly in situations where video content plays a crucial role in critical decisions related

to illegal activities, including issues like movie piracy and child pornography. The growing reliance on digital videos across diverse multimedia platforms amplifies the urgency to address these challenges. Neglecting to tackle these security issues could not only compromise the integrity of online content but also have broader societal implications, underscoring the importance of robust security measures in managing digital video content on these platforms. To bolster the reliability of incorporating digital video into everyday life scenarios, the incorporation of copyright protection and video authentication mechanisms becomes imperative. While the realm of source camera identification rooted in digital images has commanded substantial research attention, the forensic scrutiny of videos has received comparatively less focus. This disparity in attention can be attributed to a spectrum of complexities encompassing compression, stabilization, scaling, cropping, and the inherent dissimilarities within frame types that manifest when storing videos on digital platforms. Resultantly, the availability of comprehensive and sizeable standard digital video databases, augmented by current repositories reflective of novel devices grounded in emergent technologies, remains wanting. The overarching objective of this paper resides in furnishing an all-encompassing survey of advancements witnessed over the preceding decade in the domain of source video identification. This exploration is undertaken through a critical examination of prevailing techniques, notably the likes of photo response nonuniformity (PRNU) and machine learning methodologies, which have contributed to the progress in this arena[116]. Gaining a nuanced comprehension of the fundamental mechanisms underlying the video production process within digital cameras is of paramount importance. This intricate process is elucidated schematically in Figure 2.18. Commencing with the initial stage, the optical lens dutifully captures the incident light from the scene. A pivotal stride in video generation involves the downsizing of the output originating from the full-frame sensor. This deliberate reduction in spatial dimensions serves to curtail the data volume necessitating subsequent processing. The facets of acquisition and color manipulation are tactically orchestrated: color-interpolated image data is subject to down sampling, while pixel readout data undergoes sub-sampling during the acquisition phase. A prevalent technique harnessed for rectifying blurring stemming from inadvertent camera movement is electronic image stabilization. This methodology is judiciously applied during postprocessing in contemporary cameras. Additionally, the postprocessing phase encompasses the prospect of image scaling and cropping, facilitating further diminution in dimensions. To optimize the efficiency of storage and transmission for the postprocessed images, a pivotal operation ensues wherein the sequential imagery is encoded into a standardized video format.

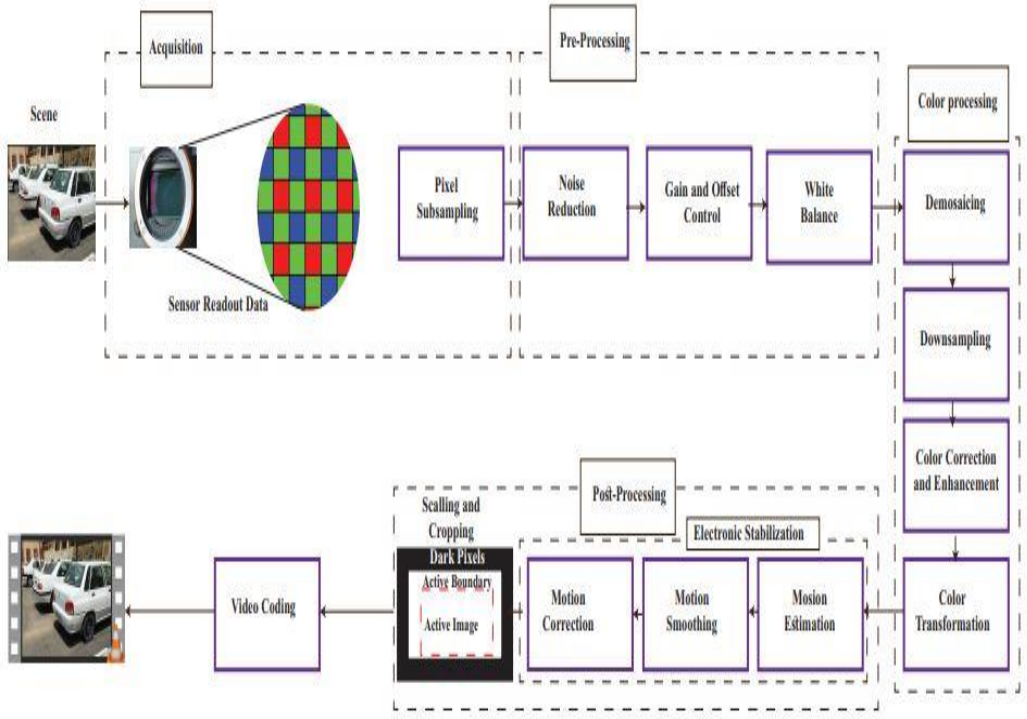


Figure 2.18. Video Acquisition Cycle

Source camera identification is systematically evaluated with respect to video content, delineated into two distinct categories depicted in Figure 2.19: the utilization of Photo Response Nonuniformity (PRNU) analysis and the employment of advanced machine learning methodologies.

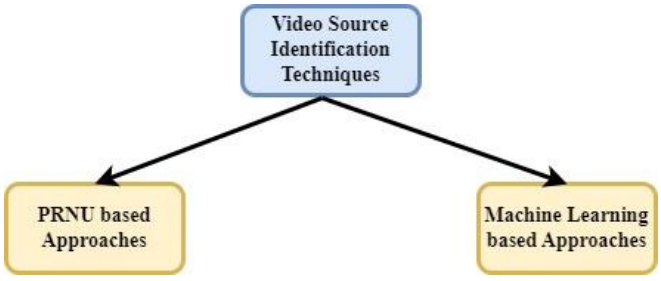


Figure 2.19. Video Source Identification Techniques

2.4.3.1 *PRNU Based Approach*

The research carried out by [117] used a Photo Response Nonuniformity (PRNU) technique involving a minimum average correlation energy (MACE) filter [118]. This method was designed to reduce the effect of noise on Normalized Cross-Correlation (NCC). They extracted PRNU from reference videos and applied the MACE filter to it. Interestingly, this process didn't affect the test (query) videos. The investigation encompassed seven camcorders, revealing a potential accuracy enhancement of up to 10% through the application

of this filter. [119] adopted a classification-oriented paradigm for source camera identification using PRNU. They extracted PRNU-based features from estimated frames and deployed wavelet sub band decomposition. The feature space was classified using a Support Vector Machine (SVM). In a noteworthy departure, [120] introduced a pragmatic yet efficacious approach. PRNU extraction was confined exclusively to the green channel of each frame, chosen for its inherent noisiness within the RGB channels. Frames were then resized to 512x512 pixels, followed by wavelet denoising to derive residual signals. This strategy, evaluated across 256 videos from six devices, convincingly underscored the efficacy of resizing in enhancing source camera identification. [121] conducted a comprehensive case study evaluating existing methods for source camera identification, thereby serving as a valuable resource for nascent researchers in this domain. [122] introduced an innovative method centered on estimating PRNU from I-frame camera video rolls, known for their camera axis rotation. They improved the process using Wiener filter (WF) and zero-mean (ZM) operations in the Fourier domain, and they further enhanced it with rotation normalization. When tested on the VISION database [123], their approach showed significant advancements compared to existing methods. Meanwhile, [124] focused on refining PRNU through an enhancement and clustering strategy. They initially estimated PRNU from a macroblock within frames and then enhanced it based on the method outlined in [125]. This technique effectively boosted high-frequency elements, overpowering noise patterns. To improve the sensor noise fingerprint, they applied smaller weighting factors to strong signal components in the wavelet transform domain. This refined the effectiveness of PRNU-based identification. They concluded by using the unsupervised agglomerative clustering technique, previously described in [126], for the categorical classification of videos obtained from the VISION database, as detailed in [123].

In a recent investigation by [127], they meticulously examined the optimal frame type suitable for identifying the source camera. This assessment, undertaken within the context of compression and stabilization, unveiled noteworthy revelations. Specifically, it was demonstrated that I-frames manifest superior outcomes in instances of stabilization, with the foremost PRNU insights emanating from the initial I-frame. Among the realm of P-frames, the acme of dependable PRNU insights is concentrated within the P-frames constituting the inaugural Group of Pictures (GOP). This revelation is meticulously explored and substantiated utilizing the VISION database. The realm of PRNU-based camera identification within the context of video content is significantly challenged by the intricate process of

image stabilization, whether enacted during capture or post-processing. Essentially, the process of digital stabilization entails a tripartite workflow: motion estimation, smoothing, and the alignment of frames predicated on meticulous motion correction analysis. This endeavor involves the calculation of feature trajectories by tracking key points across successive frames and estimating motion, often accomplished through either parametric models or by harnessing the geometric interrelationships between frames, as highlighted in the works of [128], [129],[130]. The nature of camera motion stabilization, be it two-dimensional (2D) or three-dimensional (3D), holds paramount significance. Contemporary methodologies tend to gravitate toward the utilization of 3D motion models to surmount the inherent constraints associated with 2D modeling. These advanced approaches, while acknowledging the complexities of reconstructing 3D with depth information, streamline the 3D structure and heavily rely on the precision of feature tracking accuracy, as exemplified by the contributions of [131], [132]. The main goal in video stabilization is to precisely line up consecutive frames using geometric registration. This counters any perspective distortion by using Euclidean transformations, like scaling, rotation, and translation, alone or combined. These transformations are applied carefully, often using a varying warping method, to adapt to how the camera moves during recording. An interesting thing happens when these transformations are applied to each frame: they create a lot of variation in camera motion, making it tricky to match pixels between frames. This makes the usual method of aligning or averaging PRNU patterns at the frame level less effective in accurately estimating a reference PRNU pattern. The final step in stabilizing a video is figuring out and undoing these frame-level transformations. This step is crucial for identifying where the video came from. The aggregate outcomes of these algorithms are visually depicted in Figure 2.20.

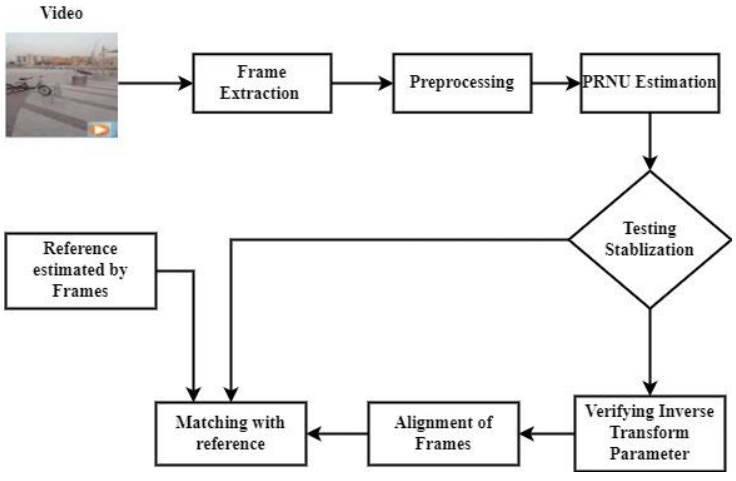


Figure 2.20. Stabilization Process.

The overarching focal point of these algorithms resides in furnishing a transformation mechanism adept at surmounting alignment quandaries that arise in diverse scenarios. In methodologies that entail the extraction of reference patterns, the transformation is often informed by images captured by the same camera, encompassing both video frames and flat frames, with further integration of initial frames from videos achieved through averaging procedures. In the study conducted by [139],a resolution to the challenge of stabilization was sought through the implementation of a straightforward inverted transformation technique during the process of noise pattern extraction. In contrast, [140] undertook a distinct approach, wherein the identification of stabilization instances within each video was accomplished through a two-pronged strategy. Primarily, the PRNU of the initial and concluding frames was juxtaposed in order to discern the occurrence of stabilization at the outset. In the analysis performed by [141] the identification of the source camera was executed under the presumption that the reference PRNU pattern could have been established from images or an unstabilized video source. Addressing the variance in resolutions between videos and images, the alignment of PRNU patterns was undertaken for a subset of I-frames, typically ranging between 5 to 10 frames. This alignment procedure encompassed the determination of appropriate scaling, shifting, and cropping parameters for each individual frame, leading to the creation of an aligned PRNU pattern through frame combinations exhibiting congruence. [142] devised an innovative approach to PRNU pattern estimation that accommodates weakly stabilized video sources. This methodology engenders the generation of an alignment reference predicated upon a selection of frames. By employing pairwise matching between PRNU estimates from diverse frames, the detection of stabilization instances is facilitated. Subsequently, a reference PRNU pattern is derived through the aggregation of frames displaying a significant match, thus establishing a reference for alignment purposes. In cases where the reference PRNU pattern is already discerned at an alternative resolution, as exemplified in [141], the said reference pattern is harnessed for comparative analysis against other PRNU patterns. Particle swarm optimization (PSO) methods, as elucidated in [155], are adeptly employed to ascertain the transformation parameters in this context, effectively streamlining the comparative process. In scenarios of weakly stabilized videos, a notable observation was made wherein rotation parameters could be disregarded, thus expediently expediting the search process. For the estimation of the PRNU pattern, an initial foundation is established through weak stabilization applied to flat

and stationary content. Subsequently, verification of stabilized videos involves the extraction of five I-frames, which are subsequently juxtaposed against the established reference PRNU pattern. The comprehensive evaluation of outcomes utilizing the VISION database robustly validates the efficacy of this method. In an evolutionary progression of their prior work, [156] further elaborate on their original research, wherein they introduced a comprehensive approach to scrutinizing source camera attribution in videos. This advanced method delves into the nuanced spatial fluctuations inherent to stabilization transformations, postulating an augmented degree of freedom in the exploration of these transformations. Notably, the discernment of transformations is conducted at a subframe granularity, entailing the incorporation of an array of constraints geared towards validating their accuracy. This intricate validation process is underpinned by a judicious computational framework, affording the requisite flexibility in the pursuit of optimal transformation solutions.

The post-decoding phase, following a filtering procedure within the decoder (i.e., the loop filter), involves the translation of the bitstream into individual video frames. Subsequently, each extracted frame undergoes a sequential processing stage aimed at the extraction of the PRNU pattern. Prior to commencing the analytical process, an assessment of the stabilization level inherent within the videos is performed, utilizing the criteria delineated by [141], [142]. This preliminary evaluation serves to eliminate videos characterized by either inadequate or weak stabilization.

To accommodate the spatially variable characteristics of stabilization transformations, the PRNU patterns are subdivided into smaller blocks, a practice found to yield optimal outcomes when employing 500 X 500 blocks. The identification of PRNU transformation parameters is then orchestrated through a block-specific search mechanism, obviating the likelihood of false inversions. A weighting protocol considers the compression levels of transformed blocks before aggregating them. The final alignment evaluation entails a detailed comparison of the estimated PRNU pattern with the reference PRNU pattern. [144] takes a distinct approach, emphasizing the creation of a sturdy reference. This deviates from previous methods that aimed to remove stabilization effects from query frames. The focus shifts toward establishing a reliable reference, departing from earlier methods centered on eliminating stabilization effects from the frames under examination. Their approach involves several essential steps applied to flat I-frames, including cropping, shifting, and using inverse transformations. Additionally, they propose an improved framework to compare PRNUs from motion-stabilized videos. [142] A groundbreaking search technique is introduced to swiftly

determine scaling and rotation parameters in the frequency domain, expediting the discovery of inverse transformations. Leveraging the Fourier-Mellin transform outlined in [157], it estimates scale, rotation, and shift between images, providing straightforward solutions. Experiments conducted on the VISION database affirm its significantly enhanced efficiency compared to existing methods. When sensor resolution exceeds the desired resolution, a combined strategy of cropping and scaling is deployed to downsize images or frames. Cameras utilize bicubic or Lanczos scaling and their variants for downsizing still images. Additionally, pixel binning and line skipping techniques are concurrently employed to alleviate camera processing overheads, primarily in video capture scenarios. These approaches work synergistically to manage resolutions, ensuring efficient downsizing while minimizing computational burdens in diverse imaging contexts. This method's acceleration of inverse transformation discovery, coupled with the amalgamated strategies for resolution management, presents a promising paradigm in handling imaging processes, particularly in scenarios demanding swift transformations or resolution adjustments. Central pixel utilization, entailing the exclusion of peripheral pixels, constitutes a prevalent strategy in video capture, efficaciously reducing camera processing demands. Nonetheless, the exclusive reliance on cropping bears a substantial drawback when applied to still images, manifesting as a narrowed field of view as the cropping region expands. To counterbalance this limitation, cropping is frequently amalgamated with resizing operations. Building upon the foundations laid by [145], [158] embark on a comprehensive extension of their research, concentrating on distinct aspect ratios engendered by resizing and cropping techniques for both images and videos, thereby enriching the arsenal of methodologies for source camera identification. This augmentation further encompasses the introduction of a dedicated database tailored to the intricacies of resizing and cropping concerns, furnishing a versatile resource for empirical investigations in this domain.

2.4.3.2 *Machine Learning Based Video Source Detection*

Table 2.3 concisely encapsulates the array of machine learning methodologies presented in this study. Broadly categorized, the works of [138] [119] are also positioned within the realm of machine learning techniques. [159] undertook an in-depth exploration, deploying machine learning strategies for source camera identification. Their approach involved the extraction of distinctive features from the bitstream, encompassing quantification factors and motion vectors. They harnessed these traits for subsequent classification via a Support Vector Machine (SVM) classifier.

Table:2.3 Machine learning Based Source Identification.

References	Machine learning based Methods	Year	Features
[159]	SVM Classification	2010	Feature extraction involves obtaining characteristics including bitstream data, quantization factors, and motion vectors.
[160]	SVM Classification	2012	A feature extraction technique grounded in Conditional Probability (CP).
[119]	SVM Classification	2016	Attributes derived from wavelet transform
[88]	CNN Framework	2019	Deriving discernible noise signals from a provided frame.
[161]	MISLnet CNN Framework	2020	Introducing a restricted convolutional layer within grayscale mode.
[162]	MISLnet CNN Framework	2020	Incorporating a constrained convolutional layer within the RGB mode.
[163]	MISLnet CNN Framework	2012	The common network functions by receiving two deep feature vectors as input and transforming them into a 2D similarity vector.

Methodically, they extracted motion vectors from each macroblock within P-frames and scrutinized multiple bitstream attributes. These attributes encompassed metrics such as the count of bits, P-frames, and B-frames within a Group of Pictures (GOP). They explored differences between adjacent P-frames and B-frames, delving into granular quantization factor details. This included scrutinizing the maximum consecutive macroblocks sharing identical quantization values within frames categorized as I, P, and B, organized in a specific sequence. Statistical measures such as mean and variance were employed to analyze the number of consecutive macroblocks sharing these quantization values across frames of varying types. Their approach involved a comprehensive assessment of multiple parameters within the video bitstream, extracting nuanced details related to frame types, quantization factors, and statistical attributes. This detailed analysis provided insights into the intricacies of video compression and encoding, aiding in feature extraction and subsequent classification using the SVM classifier. By examining these various attributes within the video stream, they aimed to identify patterns or anomalies contributing to improved classification accuracy in their forensic analysis of video content. Their exploration of quantization factor attributes involved looking at how much the quantization parameters differed among neighbouring macroblocks in different frame types, along with the corresponding average disparity. Motion vector attributes were equally scrutinized, wherein a defined search window facilitated the

estimation of maximum horizontal and vertical dimensions. This methodological synthesis substantiates a meticulous investigation into machine learning-driven source camera identification. [160] introduced an innovative paradigm of feature extraction centered on conditional probability (CP) for the explicit purpose of source camera identification. The efficacy of these features was also demonstrated in the domain of steganalysis applications, as exemplified in the work of [164]. The foundation of conditional probabilities lies in quantifying the likelihood of occurrence of event B, given the prior occurrence of event A. The extraction process encompassed the retrieval of JPEG Discrete Cosine Transform (DCT) coefficient arrays from individual frames, subsequently facilitating the derivation of CP features from these coefficients. The extraction of features was executed at the granularity of blocks, each comprising an 8x8 array of coefficients within a frame. Employing an SVM classifier, the extracted features were subjected to classification to discern the source camera. Worth noting, the method's validation was confined to a selection of videos emanating from four distinct devices. In [88], a Convolutional Neural Network (CNN) underpinned by sensor pattern noise (SPN) was introduced, aptly named SPN-CNN. The conceptual basis rested on CNN's inherent capability to extract signal signatures characterized by noise from an assemblage of images, as theorized by [89]. Consequently, the network was systematically trained to discern noise patterns. Rigorous testing on the VISION database, as curated by,[123] unequivocally showcased the method's superiority over the Wavelet denoiser. Remarkably, the study elucidated a marked enhancement in results when restricting the CNN inputs exclusively to I-frames. [161], [162] contributed to the panorama of deep learning methodologies through the introduction of the MISLnet CNN architecture. Rooted in an extension of the constrained convolutional layer initially introduced by [17], the architecture demonstrated its salience through an innovative majority voting strategy that aggregated decisions at the video level. This was achieved by feeding frames into the network. A distinct feature of this architecture is the incorporation of a foundational layer utilizing three kernels with a size of 5, meticulously designed to elicit inter-pixel relationships independent of the scene's content. Rigorous experimentation, conducted on the VISION database, reaffirmed the potent efficacy of this constrained convolutional layer in augmenting the performance of deep learning architectures, especially when compared to counterparts lacking such a feature. It's pertinent to note that the disparity between these methods lies in the dimensions of images and color modes employed, with [162] utilizing RGB mode and [161] employing grayscale mode. Image patches for the former are sized at 480, while the latter employs patches of dimension 256. [163] harnessed a CNN architecture akin to the prior works of [17],

Expanding its capabilities for extracting features, along with a specially designed similarity identification network to confirm the source camera, sets this approach apart. What makes it unique is how it uses a similarity network to connect pairs of various deep learning feature vectors to a enhanced 2D similarity network vector. Constructed on foundational principles detailed in [165]. For video-level decisions, they introduced a fusion method leveraging the mean of the inactivated output layer from the similarity network. Validating this on the SOCRatES dataset, curated by [166], showcased the method's efficacy and applicability in forensic video analysis. Experimental results unequivocally demonstrated the method's notable superiority over conventional approaches, exemplified by [167].

2.4.3.3 *Deep Learning based Approach*

Deep learning, which falls under machine learning, focuses on using complex hierarchical architectures in artificial neural networks. These structures can learn in different ways: supervised, semi-supervised, or unsupervised. Most developed models are designed on Convolutional Neural Networks (CNNs), but they can also involve propositional formulas or hidden variables organized in layers. These layers, akin to nodes, resemble the architectural patterns seen in enhanced deep learning networks or deep Boltzmann framework machines, as elucidated by [168] of significance, the current landscape of deep learning primarily employs convolutional layers to capture and encapsulate the essence of scene content, shifting the emphasis away from the traditional role of detecting camera-specific attributes such as noise patterns. However, it is worth noting that deep learning methodologies, such as CNNs and Siamese networks, as expounded upon by [169], can be suitably harnessed to fulfil this particular objective. Recognizing the growing importance of identifying camera models is crucial in multimedia forensic investigations. The abundance of digital content—images, videos, audio sequences, and the like—is continuously increasing, a trend expected to continue with ongoing technological progress. This surge is largely due to the internet's rise and the globally use of social media network to share content, which have sped up the proliferation of forged data of digital content. As a result, tracing the origins of this content has become a challenging task [170].

In the realm of forensic investigations, the ability to trace the lineage of digital content assumes paramount importance. This capability is pivotal in unmasking the culprits behind a spectrum of crimes, including untracked rape case, remote areas drug trafficking, and acts of terrorism, by establishing the provenance of digital materials. The unfortunate proliferation of incidents like revenge porn further underscores the potential for private content to go viral on

the internet. Given these multifaceted scenarios, the imperative to retrieve the origin of multimedia content is a foundational tenet [171]. In light of these considerations, the focus of this study is to ascertain the smartphone unique model identification employed for capturing digital video. This objective is pursued through a composite approach involving the extraction and fusion of both visual and auditory information derived from the multimedia content. The domain of forensic literature has seen limited dedicated exploration towards identifying the provenance of video sources, prompting our focused investigation into video source attribution. In juxtaposition, the sphere of digital image analysis has garnered substantial attention within the domain of digital imaging. Evidential traces imprinted onto photographs at the instance of image capture proffer an avenue to discern the specific camera model employed for image acquisition [172]. This pursuit unfolds through two cardinal trajectories: model-based and data-driven methodologies. The former, namely the model-based approach, delves into harnessing the distinctive traces emanating from the digital image capture process to decipher the camera type. These traces, intricately entwined with procedural nuances, serve as conduits to unveil the camera's identity through meticulous tracing. Numerous processing operations and imperfections inherent to the image acquisition pipeline, including residual dust particles and noise patterns [173], have been harnessed as conduits to communicate informative cues, thereby substantiating accurate camera model identification.

In recent years, the advent of digital data and computational prowess has ushered forth data-driven approaches that markedly outshine their model-based counterparts. The data-driven paradigm adeptly captures an array of model-related traces, diverging from the conventional focus on specific traces arising from image acquisition. This expanded scope results from the intricate interplay of system components that facilitate the capture of diverse model traces. One of the key data-driven methods revolves around learned features. These approaches involve feeding digital images into deep learning structures, allowing the models to pick up specific features related to the model and establishing connections between images and where they come from [174]. The CNN models have emerged as the preeminent solutions within this domain, gaining widespread prominence. To the extent of our current understanding, the realm of video sequence-based camera model identification remains relatively uncharted, with only a singular study found in existing literature. In this manuscript, we harness sophisticated deep-learning methodologies to forge efficacious avenues for unique camera model of device for attribution through video sequences. Our approach creates the

segmentation of video frames into discernible patches, from which pertinent features from patch dataset are extracted for better result. These features are subsequently amalgamated to yield a precise classification outcome for each video. Elucidating further, our study centres on harnessing advanced deep-learning paradigms to formulate potent methodologies for source camera identification within video frame sequences. In this endeavor, we propound an approach that entails the automated extraction of pertinent attributes from both visual and auditory components of videos. This is achieved through the adept utilization of CNNs, endowed with the capability to effectuate classification through the amalgamation of the extracted features. Our proposed approach centres on a mixed-modal framework we've termed "multi-modal." This name summarizes how our method works: it simultaneously gathers visual and auditory details from query videos to solve the identification challenge. Our methodology relies on both visual and audio content. To explain it more, in terms of visual content, we carefully isolate specific sections extract patches from the frames. In view of, the audio domain, we judiciously extract content from patches from the LMS of the separated audio-based track enshrined within the video frames, lending credence to our solution for the identification predicament. Pertinently, the proposed technique espoused by the authors aligns with the mono-modal paradigm, characterized by an exclusive reliance on the visual facet of a given video to fuel the classification endeavor. In our quest to fathom multi-modal camera model identification, we proffer two discrete methodologies, each hinged on the pivotal information thus garnered [175]. Both methodologies operate in the realm of Convolutional Neural Networks (CNNs), whereby a dyad of visual and audio patches is adeptly presented to the networks for informational ingestion. The inaugural approach entails a juxtaposition and amalgamation of individual scores furnished by a duo of CNNs. These CNNs, meticulously primed in consonance with a mono-modal tenet, cater to distinct data domains—wherein one CNN is rigorously calibrated to decipher solely visual data, while the other grapples solely with auditory data. In contrast, our second approach embarks upon the tutelage of a solitary multi-input CNN. This multifaceted CNN is astutely cultivated to concurrently process both visual and audio patches, thereby coalescing the potency of dual domains. In the pursuit of methodological robustness, we undertake a comprehensive analysis of each proposed approach, delving into the intricacies of three distinct network configurations and data pre-processing protocols. These configurations are meticulously tailored around established CNN architectures that reign supreme in the realm of cutting-edge video processing, ensuring an optimal amalgamation of efficacy and performance.

Our evaluative endeavors are underpinned by a meticulous examination of the Vision dataset—a comprehensive compendium comprising approximately 650 unaltered video sequences, accompanied by their corresponding renditions across various social media platforms. This corpus constitutes a rich tapestry of nearly 2000 videos, all impeccably captured by 35 contemporary smartphones [176]. The scope of our experimentation extends beyond the confines of the original pristine footage. By incorporating videos subjected to the transformative algorithms of WhatsApp and YouTube, we engage in a multifaceted exploration. This dual-pronged approach not only scrutinizes the ramifications of data recompression but also probes the uncharted waters of disparate training and testing datasets, a scenario often rife with challenges. In our pursuit of comprehensive assessment, we establish a baseline benchmark by unraveling the intricacies of mono-modal attribution quandaries. This strategic comparison serves as a beacon, allowing us to gauge the attained results against a yardstick of reference. It is undeniable that recent strides in the domain of multimedia forensics predominantly gravitate toward video sequences. The contemporary landscape is typified by an array of approaches that either dissect visual or audio constituents in isolation or opt for a symbiotic synergy of both modalities. This entrenched dichotomy serves as a cornerstone, propelling our dedicated exploration into the intricate intricacies of multimedia source identification. The utilization of both visual artifacts and audio content cues into the domain of multimedia forensics has emerged as a relatively nascent endeavor, albeit one that has yet to comprehensively address the intricate task of camera model identification within its purview. Our proposal entails a systematic examination of outcomes garnered from the isolated exploitation of either visual or audio patches, thus yielding a mono-modal avenue for the classification of query video sequences [177].

The bedrock of our investigation lies in a meticulous experimental campaign, orchestrated to unravel the comparative efficacy of mono-modal and multi-modal methodologies. The empirical insights garnered unequivocally affirm the supremacy of the latter, casting a shadow of inefficiency over the former. The demonstrated prowess of our pursued multi-modal strategies aptly surpasses the conventional mono-modal paradigms, thereby offering a streamlined and more potent solution to the task at hand. Furthermore, an intriguing observation surfaces in relation to data compression's impact on classification endeavors. Our empirical analysis reveals a salient pattern wherein data subjected to more robust compression, such as videos transmitted via the WhatsApp application, pose a formidable challenge in the classification realm. In stark contrast, data subjected to milder compression,

exemplified by content uploaded onto YouTube, exhibits a comparably favorable ease of classification [178].

This holistic analysis synthesizes a compelling narrative, underscored by empirical evidence, highlighting the nascent convergence of audio-visual cues in the realm of multimedia forensics. Moreover, it showcases the ascendancy of our multi-modal methodologies in unravelling intricate camera model identification puzzles, all while shedding light on the nuanced interplay between data compression and classification challenges. Despite these challenges, our investigation has revealed a notable trend: even within this intricate context, multi-modal strategies continue to exhibit superior performance compared to their mono-modal counterparts. This observation underscores the resilience and adaptability of multi-modal approaches in addressing the complexities inherent in the task. In pursuit of feature extraction and subsequent categorization within a sequential image dataset, the video classification algorithm leverages advanced feature extractors, notably convolutional neural networks (CNNs). These CNN-based feature extractors parallel their image classification counterparts, facilitating the process of categorizing videos based on extracted descriptors. Harnessing the power of deep learning-driven video categorization, a realm encompassing activities and events within visual data sources like video streams is attainable. The intricacies of such activities are scrutinized, categorized, and tracked, enabling comprehensive analysis within the visual domain. The implications of deep learning-powered video classification transcend mere surveillance, extending into various domains including anomaly detection, gesture recognition, and the discernment of human activities. These multifaceted applications highlight the versatility of video classification as a potent tool in the realm of information processing and understanding.

In essence, our exploration delves into the crux of multi-modal supremacy within intricate contexts, while the utilization of CNN-based feature extractors for video classification underscores the intersection between image and video analysis. This convergence bears testimony to the profound impact of deep learning in extracting meaningful insights from dynamic visual data sources, fostering a plethora of practical applications beyond the scope of traditional video understanding paradigms.

Initiating the process of video classification entails a systematic sequence of steps, each contributing to the comprehensive categorization of video content.

- The initial phase involves the creation and curation of training materials, constituting a pivotal foundation for subsequent classifier development and refinement.
- A judicious selection of a suitable classifier constitutes a pivotal decision in the video classification process, the choice of which should align with the specific objectives and characteristics of the dataset.
- The classifier's proficiency and effectiveness must be honed through a continuous process of education and assessment, wherein its performance is meticulously evaluated and optimized.
- Harnessing the potential of the chosen classifier, the subsequent step involves the adept processing of video data, wherein the classifier's learned knowledge is applied to categorize and label diverse video content.
- Training the classifier extensively on specialized video datasets like the Kinetics-400 Human Action Dataset significantly amplifies its efficacy in activity recognition. These curated datasets serve as pivotal resources specifically designed to facilitate and enhance activity recognition pursuits.

An enhanced classifier can be effectively trained the data by harnessing the vast and high-fidelity reservoir of activity recognition video data present in expansive datasets like the Kinetics-400 Human Action Dataset. This meticulously compiled collection encompasses a plethora of meticulously tagged video clips, each contributing to the enhancement of the classifier's knowledge base [179]. The initial stage of the process entails provisioning the video classifier with annotated footage or video clips, thus initiating the intricate classification process.

Within this framework, a robust and intricate deep learning-based video classifier is constructed, comprising convolutional neural networks meticulously engineered for video analysis. This classifier exhibits the capability to prognosticate and classify videos content based on the inherent characteristics of the video inputs, thereby demonstrating the prowess of deep learning methodologies in the realm of video classification [180]. A crucial component of the procedure involves the rigorous evaluation of the classifier's performance, a step indispensable in ensuring its efficacy and fine-tuning. Furthermore, the versatility of the classifier extends to real-time applications, as it can be adeptly employed to categorize activities depicted in live webcam video streams or collections of dynamically streamed video clips. The profound capabilities of the classifier extend to encompass various training paradigms, encompassing techniques and utilized the developed the Slow path and Fast

improved paths (Slow Fast), ResNet with (2+1) D convolutions, and enhanced two-stream approach to Inflated-3D approaches, as illustrated in Figure 2.21, all facilitated by the expansive resources provided by the Computer Vision Toolbox. Manufacturers of DSLR cameras, including industry giants like Canon, Nikon, and others, routinely employ intricate calibration algorithms as a prelude to capturing scene imagery, a procedure that significantly contributes to the elevated cost of professional-grade DSLR cameras. In light of this, there is a compelling impetus to engineer novel calibration techniques that are not only computationally efficient but also cost-effective, aiming to render them on par with established methodologies employed on a global scale. The overarching objective is to democratize the calibration process, thereby rendering it accessible and economically viable for a wider demographic. The calibration protocols undertaken by leading DSLR camera manufacturers involve intricate procedures aimed at achieving optimal performance and accuracy in image capture. The deployment of such sophisticated algorithms imparts a substantial financial burden, consequently elevating the price point of premium DSLR cameras. To address this challenge, it is imperative to embark on the development of alternative calibration techniques that circumvent excessive computational demands while concurrently maintaining an uncompromised standard of quality. The successful implementation of such techniques would, in turn, lead to a notable reduction in cost, rendering professional-grade image-gathering equipment more affordable and accessible to a broader spectrum of users.

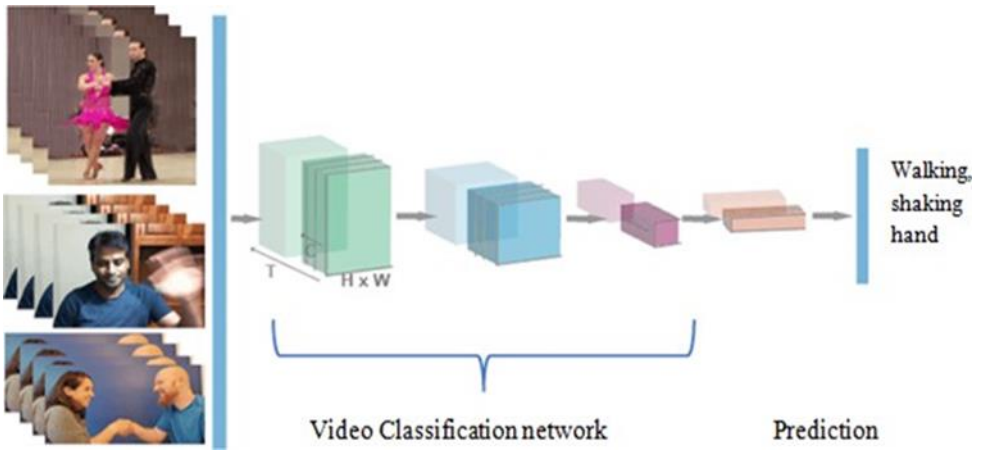


Figure 2.21. 3D Methods for Video Classification Classifier Training

Efforts in this direction are propelled by the aim of democratizing advanced imaging technologies, aligning with the principles of cost-effectiveness and widespread accessibility. By engineering calibration techniques that are both computationally streamlined and financially viable, it becomes feasible to equip a larger populace with the means to engage in

high-quality image acquisition, thus fostering a more inclusive and participatory imaging landscape on a global scale [181]. The process of identifying the specific camera model employed to capture the photographic and video frames showcased within this article hinges upon the meticulous scrutiny of numerous distinctive traces embedded within the images and video frames during the image acquisition process. This endeavor encompasses an exploration of the idiosyncrasies inherent to the acquisition of digital photographs, thereby fostering a comprehensive understanding for readers seeking to delve into this domain. Subsequently, a comprehensive exposition on the Mel scale and its implications for audio content in video sequences will be presented, thereby facilitating a nuanced comprehension of the subsequent analytical framework. The exposition elucidates the merits of the Log-Mel Spectrogram (LMS) as an invaluable tool for scrutinizing temporal variations in audio tracks, as well as their evolving spectral attributes [182]. Over the preceding decades, the endeavor to discern image camera models has engendered an array of methodological avenues, each honing distinctive methodology. Central to these methodologies is the pursuit of discerning noise pattern characteristics germane to individual camera models from the furnished images and videos. These noise patterns, often dubbed traces, are conjectured to stem from manufacturing anomalies that manifest uniquely within each camera model [183]. In seeking to fulfill this aim, diverse approaches have emerged, each striving to ascertain the distinctive noise patterns harbored by varying camera models. This proactive endeavor hinges upon a multifaceted analysis of images or videos, culminating in the extraction of discernible traces, thereby constituting a foundational framework for camera model attribution. The core tenet driving these methodologies is the premise that the distinctiveness of these noise patterns is intimately tied to the intricate manufacturing nuances intrinsic to each camera model, thus propounding a novel facet for model differentiation within the realm of multimedia forensic investigations[184]–[186].

Within the realm of multimedia forensics, an extensive focus has been directed toward the intricate endeavor of blind source device identification. This pursuit entails meticulous analysis of discernible traces, encompassing phenomena such as sensor dust and defective pixels, thereby culminating in the formulation of multifaceted strategies aimed at discerning the originating capturing device. A pivotal turning point in this trajectory was instigated by Lukas et al., who introduced the pioneering concept of harnessing Photo-Response Non-Uniformity (PRNU) noise as an unequivocal marker for defining the distinctive geometry of a camera sensor, thus engendering a notable paradigm shift [187]. An inherent characteristic of

PRNU noise lies in its multiplicative nature, a trait that imbues it with a remarkable resilience against removal even when subjected to sophisticated high-end processing equipment. This intrinsic multiplicative essence renders PRNU noise exceptionally tenacious, resisting effective elimination. Notably, this persistence persists unrelentingly even after subjecting the image to JPEG compression at an average quality level. In the context of exploring the applicability of PRNU-based camera forensics for image recovery from standard Scene Matching Points (SMPs), investigations have unveiled a salient caveat. It has come to light those alterations, whether instigated by users or the SMPs themselves, have the potential to vitiate the efficacy of PRNU-based source identification. Such alterations could erode the fidelity of the PRNU-based inference, thus impeding its reliability as a robust source identification mechanism. This nuanced insight accentuates the need for a comprehensive understanding of the interplay between PRNU-based identification and potential image modifications within the broader domain of multimedia forensic analysis. Emerging advancements in camera software integration encompass novel digital identification technologies aimed at mitigating the destabilizing ramifications stemming from unsteady hands during video capture. This innovative paradigm involves a programmatic evaluation of user-induced movements, discerning their impact on the pixel allocation within the camcorder's image sensor. Within this dynamic framework, the manipulation of specific pixels within the image sensor is orchestrated to counteract the destabilizing influences associated with unsteady hands. For Android-based devices, a degree of user agency is accorded, enabling the activation or deactivation of image stabilization features. In contrast, iOS-based devices do not provide users with the capacity to modulate this setting, indicative of a distinguishing feature between these two ecosystems.

In the pursuit of attributing video sources through the prism of active digital identification hinging upon the PRNU fingerprint, an intricate challenge surfaces in the form of alignment disruption during the identification process. This perturbation-induced misalignment renders the task of source identification elusive, thereby incapacitating the discernment of video sources characterized by active digital identification mechanisms. The underlying implication is that the inherent dynamics of active digital identification methods cast a shadow of uncertainty upon the viability of PRNU-based source attribution within this context [188]. It is essential to recognize and address these intricate interactions to refine the accuracy and reliability of source identification in the realm of digitally enhanced video capture. Notwithstanding the strides made by HSI in formulating a reference-side solution,

specifically pertaining to the estimation of fingerprints from static photographs, the underlying challenge remains unresolved. Within the domain of forensic video analysis, a myriad of evolving techniques hold promise for unearthing evidentiary insights. Yet, prior to their application, a multitude of unanswered inquiries necessitate comprehensive exploration to validate their efficacy within this context. Moreover, the realm of forensic video analysis unveils heightened complexities compared to its image analysis counterpart, posing formidable obstacles in comprehending the intrinsic data of videos. This intricacy is rooted in the intricate compression structures that videos adopt, presenting a stark contrast to the relatively straightforward formats employed by images. While an image frame encapsulates a sequence of discrete images that collectively unfold over time, imbuing the visual narrative with motion and temporal evolution, a video entails a reservoir of information ingeniously encoded and decoded through mathematical methodologies, colloquially known as codecs. These encoded frames, pivotal components of the multimedia tapestry, are encapsulated within a multimedia file format, interwoven with complementary tracks housing audio, metadata, and subtitles. Converging as multimedia files, this amalgamation mirrors the complexity inherent in the video medium, requiring an intricate orchestration of encoding and decoding mechanisms to facilitate seamless interpretation and playback. The nuanced intricacies of multimedia files, brimming with encapsulated visual and auditory dimensions, underscore the multifaceted challenge that forensic video analysis endeavors to surmount [189]. By delving into the depths of video compression paradigms and the intricate interplay of multimedia elements, the quest for refining and advancing forensic video analysis confronts multifarious dimensions demanding systematic exploration and resolution.

CHAPTER 3

IMAGE SOURCE IDENTIFICATION USING TWIN CNN ARCHITECTURE

3.1 *Proposed Framework*

This section introduces a Twin CNN Architecture for image source identification. A novel model is devised to preprocess the dataset and accurately classify images with appropriate device class labels, as outlined shown in Figure 3.1. Within this proposed framework, the input consists of an image dataset, and the output involves classifying the source device along with class-level information. The effectiveness of prediction is quantified through high-accuracy measurements. The procedure is detailed as follows: The initial step involves generating patches of size 256×256 from the original dataset, considering the varying resolutions of device models for captured images.

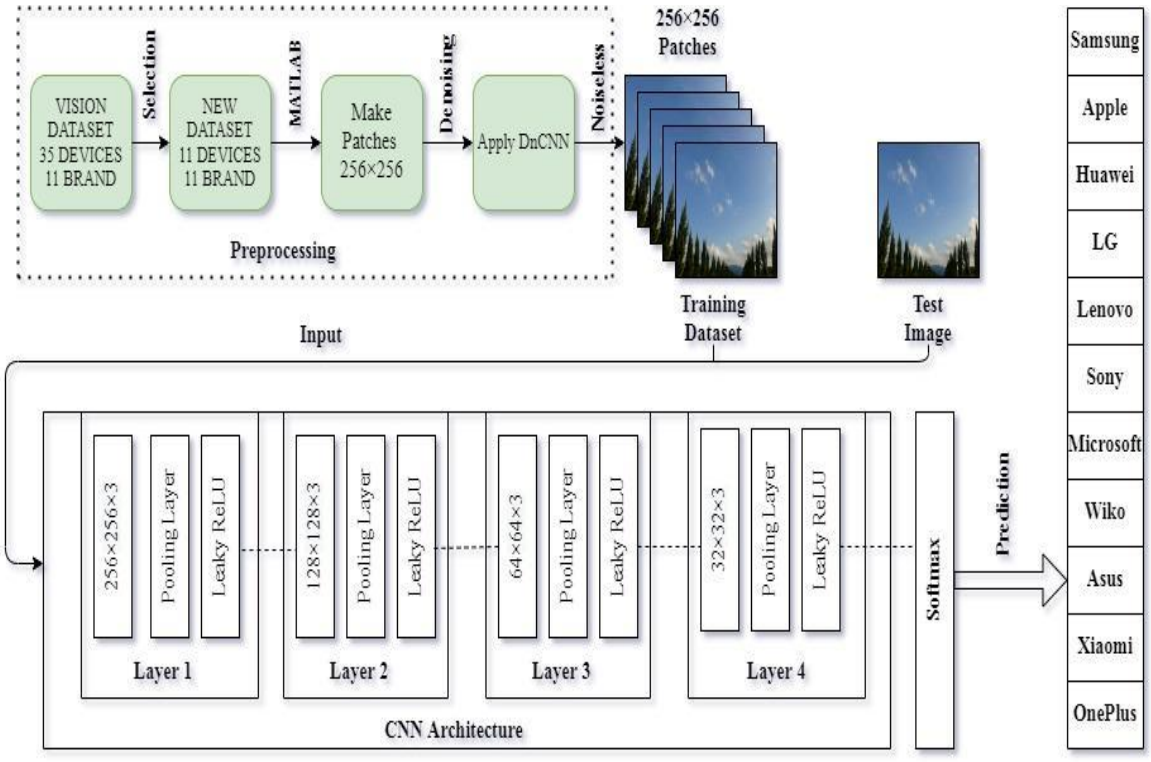


Figure 3.1. Proposed Framework using Twin CNN Architecture for Image Source Identification

3.2 *Conversion of Data Sets into Patches*

This conversion of images into patches is executed using MATLAB to ensure quality preservation. The optimal patch is selected from the images to proceed with the analysis depicted in Figure 3.2.

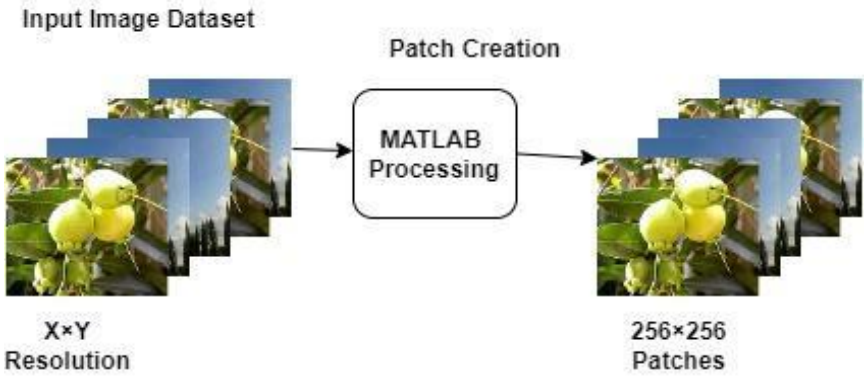


Figure 3.2. Conversion Image into Patches

Upon establishment of a meticulous dataset with distinct device IDs, the subsequent phase involves employing the 'MATLAB' software to generate patches of dimensions 256×256 from the initial dataset. It is noteworthy that these patches inherently harbor noise attributed to material imperfections. To mitigate this noise, the DnCNN model is harnessed for denoising the image dataset.

3.3 *Denoising of Patches*

The selection of DnCNN Architecture for denoising is substantiated by its competence in handling Gaussian noise instances with indeterminate noise labels within random images. DnCNN architecture leverages residual learning techniques and integrates batch normalization to optimize denoising outcomes for the image dataset. Employing a 3×3 kernel, the patch is subjected to convolutional processing, while the omission of a pooling layer maintains the patch's dimensions post denoising. The DnCNN model is configured with dimensions of 35×35 and is structured with an expanded depth ranging from 17 to 20 layers, categorized into three distinctive layers. The initial layer employs convolutional operations combined with the ReLU activation function, where a 3×3×3 filter yields the generation of 64 feature maps. This strategic arrangement facilitates the initial feature extraction process. Within the second layer, the convolutional filter is invoked, accompanied by batch normalization, which synergistically heightens the learning rate to optimize prediction accuracy. The incorporation of the ReLU activation function within this layer further

augments learning efficiency. In the ensuing third layer, a single convolution filter, structured with dimensions of 3×3×64, is employed to effectuate the reconstruction of the native picture dimension within the DnCNN framework. The holistic architectural progression of DnCNN is concisely outlined as follows shown in Figure 3.3.

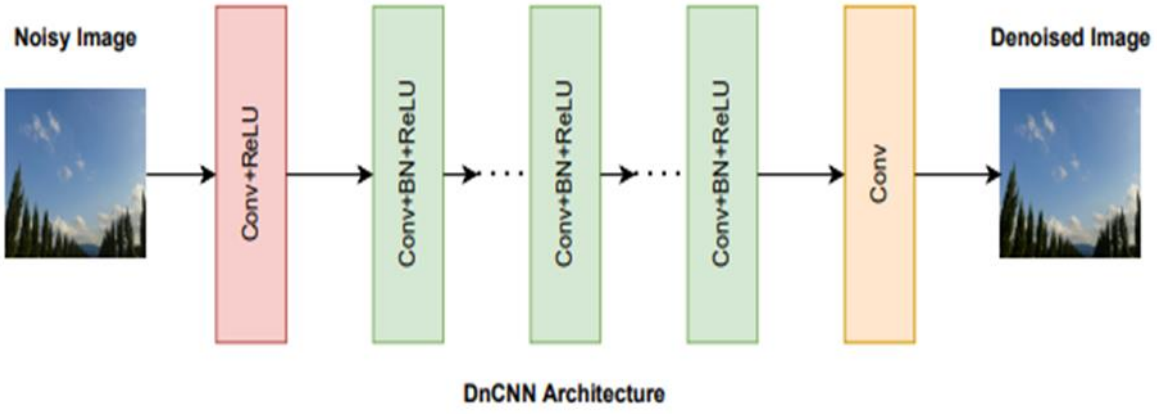


Figure 3.3. DnCNN Architecture for Denoising the Image

Following dataset pre-processing, we introduce a CNN architecture comprising four distinct layers, aimed at source identification by virtue of feature extraction via convolution kernels. The classification CNN architecture takes a 256×256 resolution image as input. To effectuate feature extraction from the input image, we deploy a convolution filter characterized by dimensions of 3×3×3. It is imperative to note that the trailing dimension varies depending on the image type. For grayscale images, '1' is utilized, whereas color images entail a '3' due to their trichromatic nature. Subsequent to convolutional filtering, we employ a max pooling layer, serving to extract maximal values post-convolution while concurrently reducing the feature dimensions to 64×64. The leaky rectified linear unit (ReLU) activation function is harnessed to generate an output, which subsequently serves as the input for the successive layer within the CNN Architecture. This approach enables robust feature extraction and encapsulates the inherent multi-channel nature of color images while ensuring dimensionality reduction via pooling, thereby facilitating subsequent processing and classification.

3.4 *CNN Architecture Operations*

The architecture's design encompasses tailored strategies for varying image characteristics, thereby enhancing adaptability and accuracy in the source identification process. In our approach, we incorporate padding and a stride value of 2 within the convolutional layers. To enhance accuracy, we iteratively execute this convolutional process for a total of four layers, culminating in the presentation of the resulting output to the SoftMax layer. This latter layer

serves to produce probabilistic values ranging from '0' to '1' for each device category, effectively discerning distinctive feature characteristics. In the context of source identification, the procedure entails inputting a random image from the test set into the model, which subsequently predicts the original source of said image.

The convolutional network's functionality can be mathematically encapsulated through the following equation:

$$P_{i,j} = f\left(\sum_{x=0}^{x} \sum_{y=0}^{y} W_{x,y} \cdot I_{i+x,j+y} + W_b\right) \quad (3.1)$$

In this context, f represents the activation function employed within the convolutional framework. The notation $P_{i,j}$ pertains to the estimated pixel value at the specific coordinate (i,j) within a given image. $W_{x,y}$ signifies the shared weight value utilized in the convolutional layer, while W_b denotes the bias element integrated within the filter structure to achieve balance. Notably, these parameters are iteratively learned from the available data and are optimized throughout the training process to attain their most effective values. The max pooling layer serves to extract features by identifying the maximum value following the convolutional operation depicted in Figure 3.4, subsequently diminishing the dimensionality of the input image.

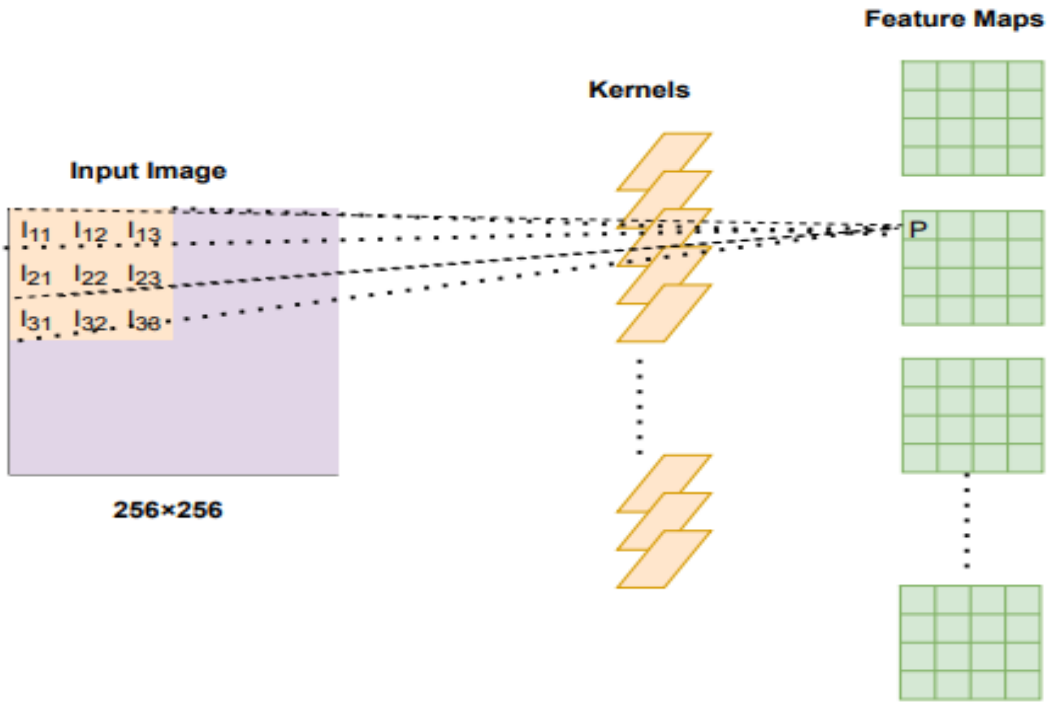


Figure 3.4. Kernels Operation in Convolution Layer

The designated kernel is applied to the input image (I), engaging convolutional operations to transform the original pixel values, and subsequently discerning the maximum value among them. The conversion process adheres to the ensuing function:

$$f(I) = max(0, I) = \{I, I \geq 0\ 0, I \leq 0 \tag{3.2}$$

The max pooling layer functions to filter out low-intensity feature values within the image, thereby enhancing both the quality and predictive accuracy. Within this network, a 2×2 filter is implemented within the max pooling layer, coupled with a stride of '2' and padding. The conversion process is visually represented in the accompanying figure:

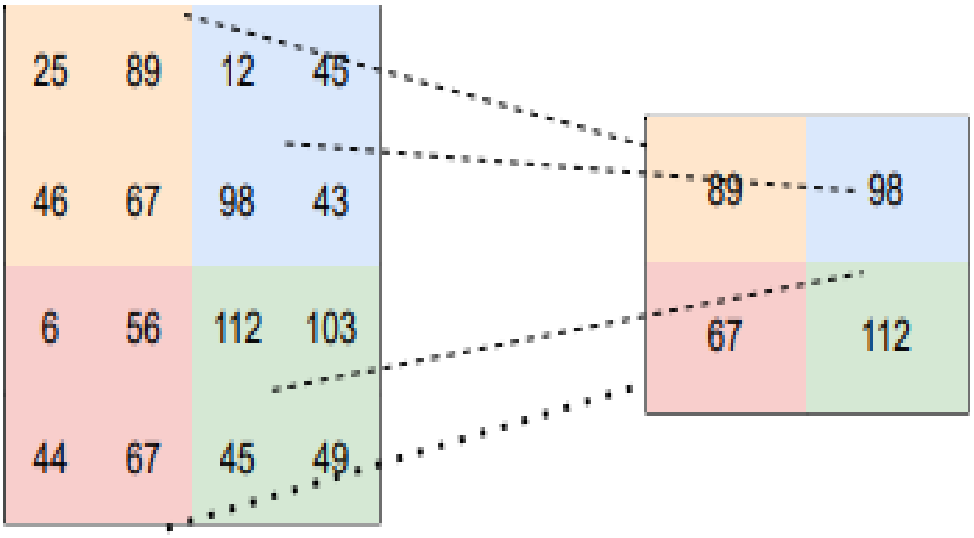


Figure 3.5. Max Pooling Layer Conversion Operation

Upon extracting features from the patches, a precaution against overfitting is taken by employing a parameter of 0.5 between the fully connected layers. The process of feature extraction is carried out across all convolutional layers, after which a softmax classifier is applied. This classifier predicts values ranging from 0 to 1 through the utilization of the subsequent expression:

$$\psi_n(I) = \frac{exp(C_n(I))}{\sum_{n=1}^{m} (C_n(I))} \tag{3.3}$$

Here "I" is the input image patch, "n" denotes the number of clusters based on a number of devices $n\epsilon(1, m)$. Where $Cn(I) = ln(P(I/Cn)P(Cn))$, $P(I/Cn)$ denotes the conditional probability, and "ψ" is the final output of softmax function which belongs to $0 \leq \psi \leq 1$. The flow of architecture is shown in the below Figure 3.6.

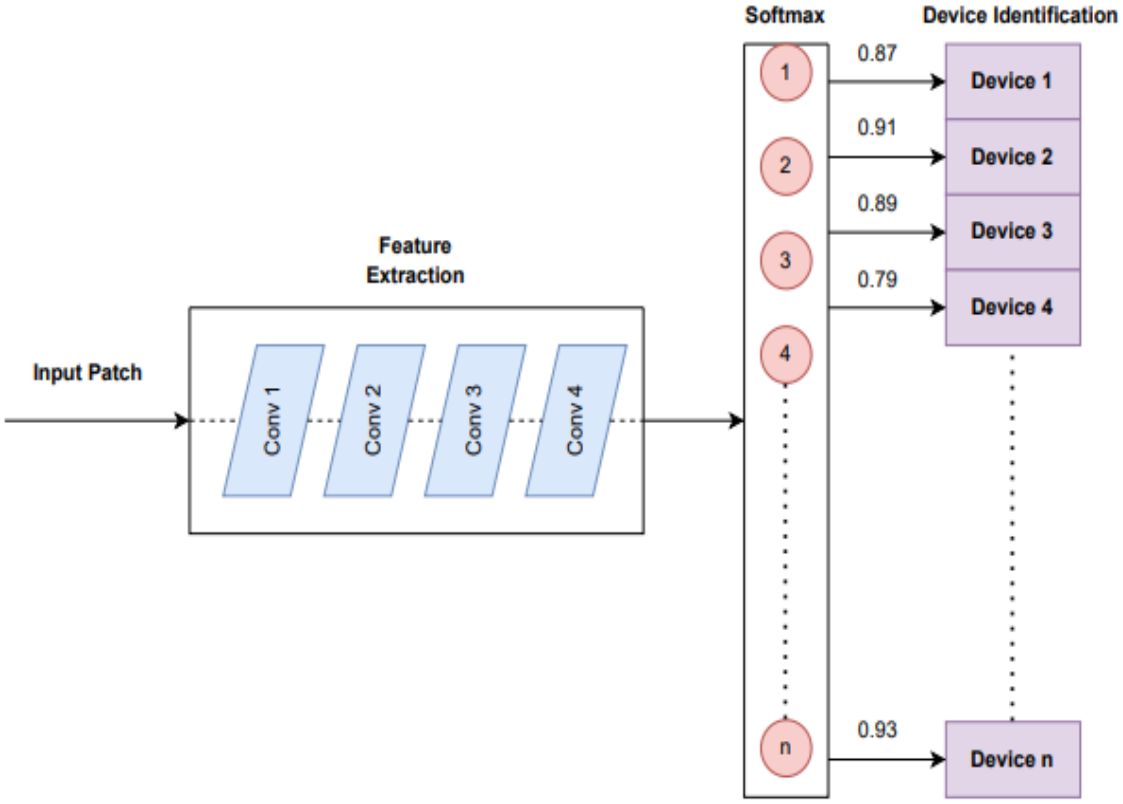


Figure 3.6 . CNN Classification using Softmax Function.

Table 3.1: Proposed CNN Architecture for device classification.

Input	Output	Operator Dimension	Activation Function
Image	256×256	Conv, 3×3, Stride=2	Leaky ReLU
Layer 1	128×128	Conv, 3×3, Stride=2	Leaky ReLU
Layer 2	64×64	Conv, 3×3, Stride=2	Leaky ReLU
Layer 3	32×32	Conv, 3×3, Stride=2	Leaky ReLU
Layer 4	16×16	Conv, 3×3, Stride=2	Leaky ReLU

3.5 *Result Analysis*

The datasets employed in this study originate from the 'VISION' dataset developed by the Communications Signal Processing Laboratory (https://lesc.dinfo.unifi.it/VISION/dataset/). This dataset has been meticulously curated to cater to the digital forensic community's requirements, encompassing high dynamic range images and videos. The compilation comprises media content captured by 35 contemporary devices from 11 distinct brands, namely Samsung, Sony, Wiko, Xiaomi, Microsoft, LG Electronics, Lenovo, Huawei, Asus,

Apple, and OnePlus. Within this dataset, there are a total of 11,732 native images, among which 7,565 are shared images distributed via platforms like WhatsApp and Facebook, encompassing both low and high quality variants. The images are categorized into various classes depicted in Figure 3.7 within the dataset.

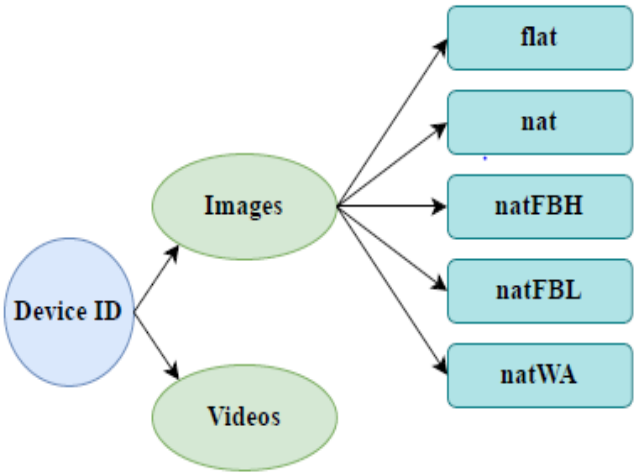


Figure 3.7 . Vision Dataset Organization

In this research endeavor, we have conscientiously opted to focus on a solitary device image emanating from a specific brand. This strategic decision aims to expedite and enhance the precision of our image source identification process. To achieve this, we have meticulously culled 'FLAT' and 'NATIVE' images at random from the comprehensive dataset, thereby crafting a distinct dataset exclusively comprised of these flat or native images, encompassing all 11 different brands. This methodological approach ensures that we have a streamlined and consistent set of images for rigorous analysis. The selected images correspond to a diverse range of brands, each bringing its unique characteristics to the fore. This meticulous curation and subsequent analysis of individual device images provide us with a rich and comprehensive dataset that is essential for our investigation into image source identification.

For elucidation, the specific details pertaining to the chosen device images, along with their respective brands, are furnished in the Table 3.2. Within our devised framework, the initial input to our model comprises patches of dimensions 256×256, a well-suited scale for robust feature extraction. Following the comprehensive assembly of the dataset, a pivotal pre-processing step ensues: the entire dataset is systematically transformed into corresponding patches. An indispensable augmentation to our methodology lies in the application of denoising procedures across the entire dataset, effectuated through the sophisticated DnCNN

architecture. This denoising process culminates in the attainment of patches characterized by uniform dimensions, thereby facilitating subsequent classification.

Table 3.2: Device characteristics with image type.

Device	Brand	Resolution	Image Category	Images
D1	Samsung	2560×1920	Flat, Native	30
D2	Apple	3264×2448	Flat, Native	30
D3	Huawei	3968×2976	Flat, Native	30
D4	LG	3264×2448	Flat, Native	30
D5	Lenovo	4784×2704	Flat, Native	30
D6	Sony	5248×3936	Flat, Native	30
D7	Mocrosoft	3264×1840	Flat, Native	30
D8	Wiko	3264×2448	Flat, Native	30
D9	Asus	3264×1836	Flat, Native	30
D10	Xiaomi	4608×2592	Flat, Native	30
D11	OnePlus	4640×3480	Flat, Native	30

Our proposed approach hinges upon the utilization of a meticulously curated high-resolution image dataset, a strategic maneuverer that underpins the discerning identification of image source. Pertinent to the experimental phase, the model is rigorously trained utilizing the designated dataset. For the subsequent testing regimen, a subset of 10 images per distinct device serves as the basis for evaluation. It is noteworthy that the training and testing accuracies, although commendably hovering around the 90% mark, are somewhat constrained due to the relatively limited number of available images, an inherent constraint that calls for strategic consideration in the overall assessment. To quantitatively evaluate the efficacy of our classification framework, we rely on a confusion matrix, a quintessential tool for gauging the accuracy of our classification endeavours.

The evaluation parameters are computed using the provided equations:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3.4}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3.5}$$

$$F1 = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{3.6}$$

$$Accuracy = \frac{\text{TP} + \text{TN}}{TP + FN + TN + FP} \qquad (3.7)$$

The specifics of this confusion matrix are succinctly provided in Figure 3.8, thereby affording a comprehensive overview of the classification performance across distinct categories.

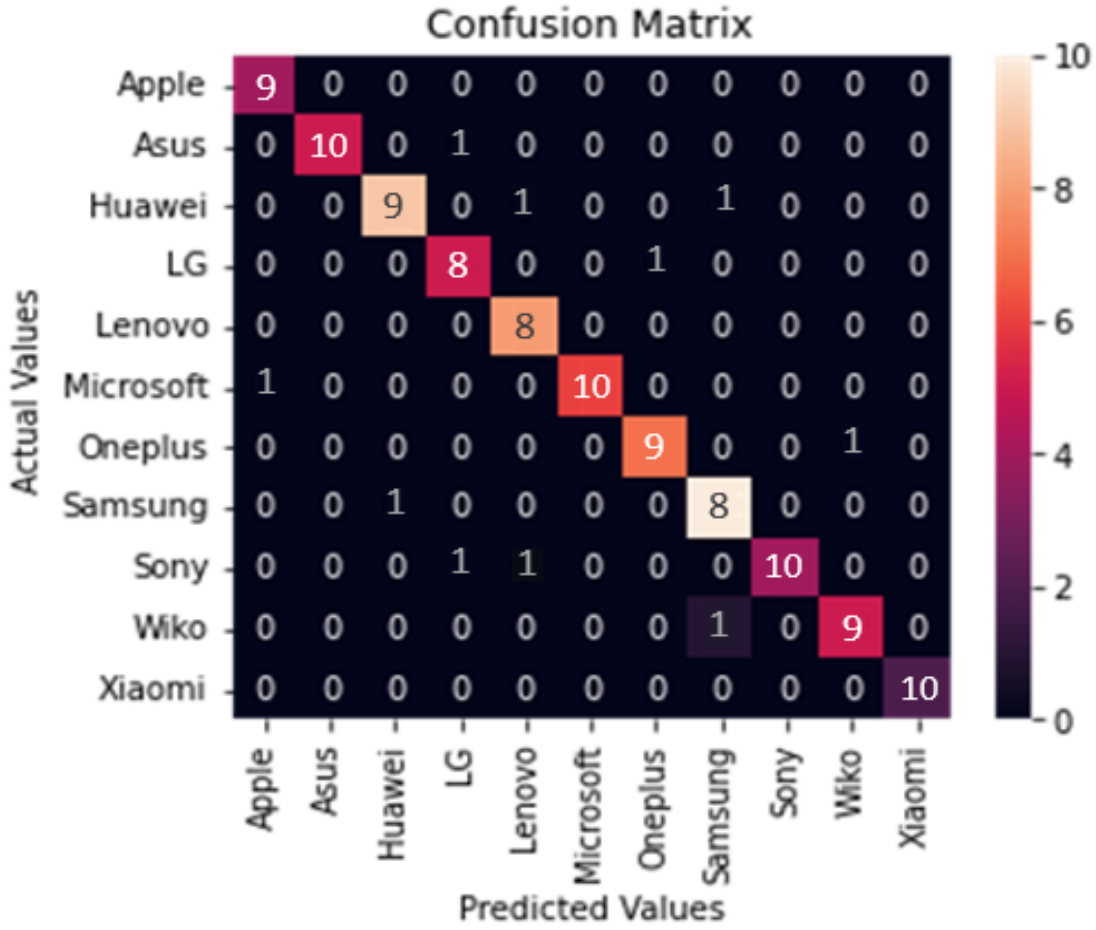


Figure 3.8. Accuracy Confusion Matrix for Fewer Data input Patches

Upon augmentation of the image dataset by increasing the count of images per individual device, an intensive training regimen spanning 400 epochs was undertaken. This strategic augmentation, along with the extended training interval, facilitated a noteworthy enhancement in both training and validation accuracy, culminating in an impressive achievement of 93.6%.

For a comprehensive understanding of the dynamic evolution of accuracy and loss throughout the training process, we present the accuracy and loss curve in the subsequent visualization in

Figure 3.9, providing a concise depiction of the progressive convergence of our model's performance metrics.

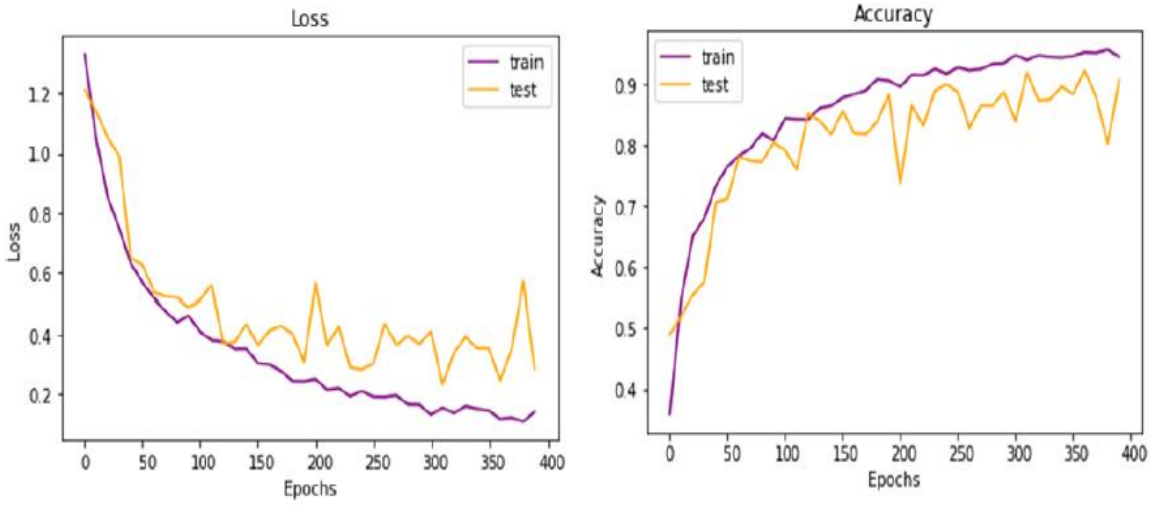


Figure 3.9. Train and Testing Loss and Accuracy Comparison

Upon completion of the model's rigorous training and subsequent meticulous testing, we present the achieved classification accuracy, meticulously organized into a confusion matrix with increasing the number of images to 30 per device, as follows shown in Figure 3.10

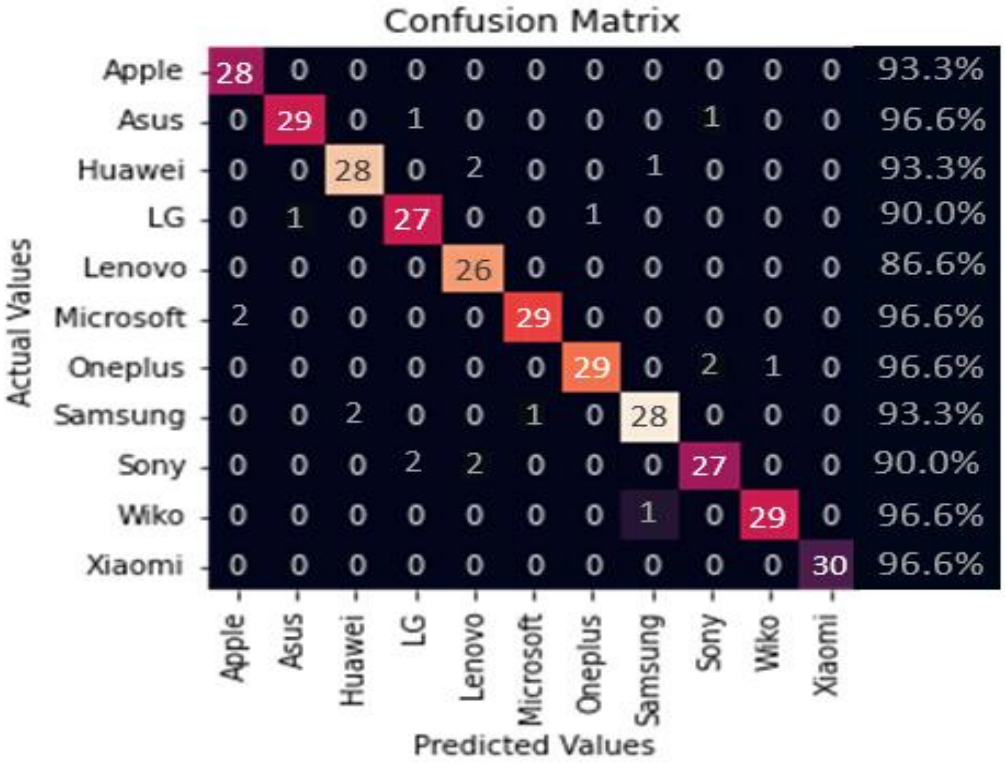


Figure 3.10. Actual and Predicted Class Accuracy Comparison

During the training model on dataset the training accuracy achieve 96.7% and validation accuracy is approximate 93.6%.

Choosing the architecture topology and fine-tuning hyperparameters are not straightforward tasks; they necessitate in-depth analysis and a comprehensive understanding of both theoretical principles and practical considerations. This complexity arises from the fact that the configuration of a CNN model depends on various factors, including the nature of the data under consideration. For example, the data may differ in terms of size, image complexity, or the specific task at hand. In this section, we discuss the various factors that inform the selection of the CNN architecture we propose.

The Figure 3.11 provided displays a comparison of accuracy with other techniques.

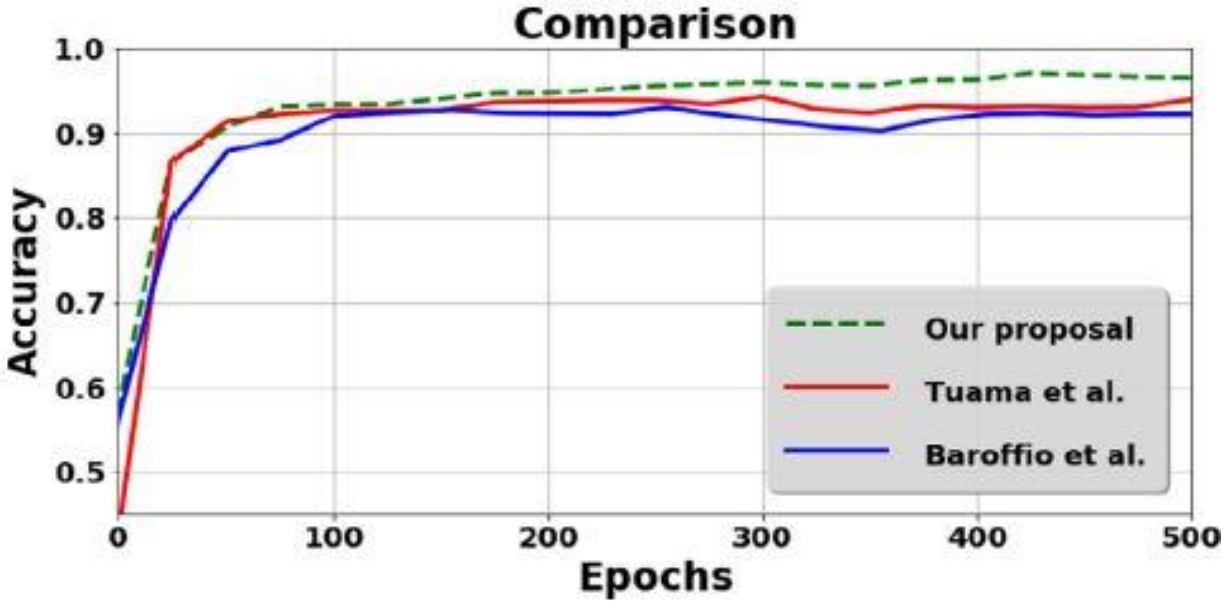


Figure 3.11. Comparison of other Techniques

In the final experiment, we evaluate the impact of altering the hyper parameters of the regularization algorithm. When we introduced the dropout technique to mitigate overfitting, a new hyperparameter was introduced: the probability 'p' that dictates the retention rate for each node in a layer. During training, dropout functions by either alive a node with a probability 'p' or deactivating it (setting its value to zero) otherwise. In this proposed model, conducted an experiment in which we tested various values for this hyperparameter 'p' depicted in Figure 3.12.

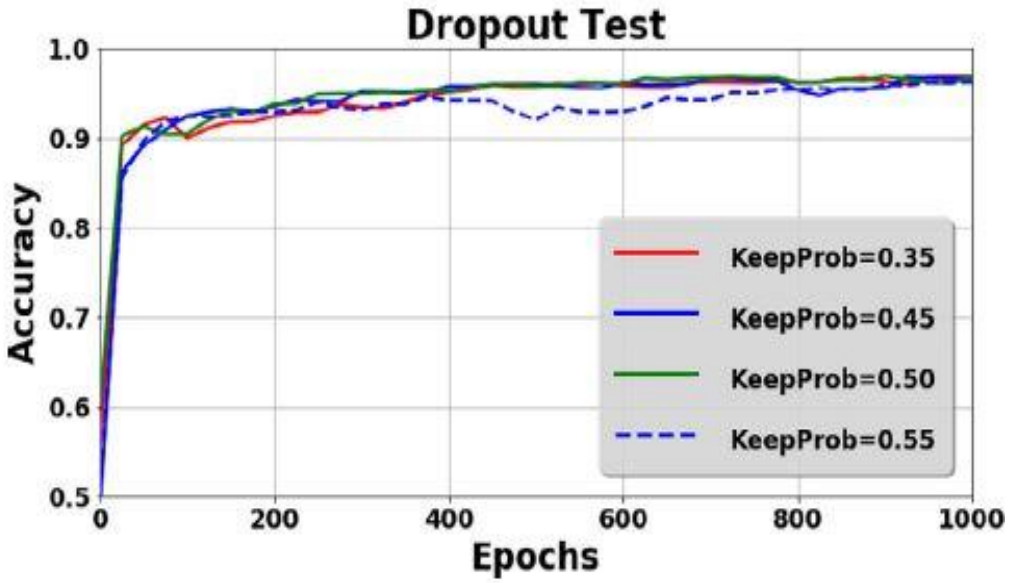


Figure 3.12. The dropout hyperparameter analysis, we examine various node retention probabilities, specifically, 0.35%, 0.45%, 0.5%, and 0.55%, respectively

This study introduced a novel deep convolutional neural network architecture, characterized by four convolutional layers, one fully connected layer, and a SoftMax classifier, aimed at discerning the original camera source of images. Through a systematic exploration of various learning configurations, an optimal balance between testing and learning performance was achieved. The key advantages of the proposed approach for camera source identification are:

1. Mitigation of the challenge posed by limited forensic image samples through image patch cropping, yielding ample training data. This is achieved by enabling the input layer to process image patches of dimensions 256x256x3.
2. Empirical validation substantiates the claim of enhanced efficacy, reinforcing the viability of the proposed method.

To gain deeper insights and further refine the technique, future research endeavours will extend the application of the proposed strategy to a more extensive dataset, thereby the proposed method aims to be applicable and effective in real-world situations beyond controlled experiments, with the goal of combating fake news, deepfake videos, and the malicious use of forged images to defame individuals in today's social media landscape. The outcomes of the proposed method could serve as valuable evidence in forensic investigations of images and videos from specific devices.

3.6 *Conclusions*

In conclusion, this study introduces a groundbreaking deep convolutional neural network as a robust solution for identifying the source of images. The model's commendable testing and learning performance are the result of a combination of architectural innovation, input preprocessing, and learning paradigms working synergistically. As the field of image forensics evolves, the approach revealed in this study marks a significant step forward in improving accuracy and efficiency in the crucial task of identifying the primary camera of images.

Looking ahead, we anticipate a sustained trajectory of research and refinement that will ultimately contribute to the progress of image forensics and its practical applications. This future-oriented perspective underscores our commitment to advancing the field and addressing the challenges of image source identification.

Some potential limitations and challenges associated with the proposed Twin CNN architecture for image source identification include computational complexity, scalability issues, and performance evaluation under varying conditions. The computational complexity of the architecture may pose challenges in terms of processing time and resource requirements, particularly for large-scale datasets or real-time applications. Scalability issues may arise when scaling the architecture to handle increasingly large datasets or when deploying it in different environments. Additionally, accurately evaluating the performance of the architecture under diverse conditions, such as varying lighting conditions or image qualities, may be challenging and require comprehensive testing and validation procedures.

CHAPTER 4
IMAGE FORGERY DETECTION USING CNN ARCHITECTURE WITH SVM CLASSIFIER

4.1 *Introduction and Motivation:*

CNNs are designed with a structure that mirrors the intricate workings of the human visual system, with interconnected nonlinear neurons. This design has already exhibited remarkable potential across a spectrum of computer vision applications, prominently in tasks like object recognition and image detection. However, the versatility of CNNs extends beyond traditional visual tasks and into realms such as image forensics, presenting intriguing prospects for detecting image manipulations. In contemporary digital landscapes, image forgery has become alarmingly accessible due to the proliferation of sophisticated editing tools. Consequently, the ability to identify manipulated images is of paramount importance, as the consequences can be far-reaching. CNNs, equipped with their ability to discern intricate patterns and deviations, emerge as potential tools for tackling this challenge. The process of image forgery often involves transplanting components from one image onto another. This manipulation gives rise to a host of artifacts that might evade casual human observation. However, CNNs possess the computational prowess to perceive these subtle deformities, even when imperceptible to the naked eye. By analysing the finer details and discrepancies in pixel-level distributions, CNNs can act as vigilant detectors of tampered imagery. The convolutional layers of CNNs play a pivotal role in this discernment process. Through a hierarchical analysis of features, these layers break down the image into smaller, more manageable components. This dissection enables the network to identify irregularities that might indicate image manipulation. Furthermore, the non-linear activation functions integrated within the architecture empower CNNs to model complex relationships, making them adept at identifying patterns that might signify tampering. To harness the potential of CNNs for image forensics, it is imperative to develop tailored training methodologies. This involves exposing the network to a diverse array of manipulated and authentic images, enabling it to learn the intricacies of image tampering. The training process involves fine-

tuning the network's parameters to enhance its discriminatory capabilities. Techniques such as transfer learning approach, which enhanced pre-trained models on large HDR image datasets, can expedite this process by providing a solid foundation for image analysis. In practice, CNNs' proficiency in image forensics could manifest in various applications. These networks could serve as integral components of digital platforms that scrutinize media content for authenticity. By systematically analysing images and flagging potential anomalies, CNNs can offer an additional layer of security and trust in an increasingly visual digital landscape. The convergence of CNNs and image forensics holds significant promise. By capitalizing on the networks' inherent ability to decipher intricate patterns and detect deviations, the field of image manipulation detection stands to benefit greatly. CNNs, through their non-linear interconnected architecture and hierarchical analysis, offer a compelling avenue for bolstering the integrity of digital imagery and upholding the authenticity of visual content. The utilization of such images introduces a distinct enhancement in the manipulated content, primarily attributed to the dissimilarities in compression algorithms between the manipulated region's source and the background images.

4.2 *Proposed CNN Architecture with SVM Classifier*

Leveraging this insight, we proceed to train a CNN model to improve accuracy with a focus on picture authenticity. The practical implementation of this novel approach is depicted through the procedural depiction in Figure 4.1. Our approach involves harnessing the divergent compression artifacts present within the manipulated area and the background. The CNN model is meticulously trained on a diverse dataset encompassing both manipulated and genuine images. The network learns to distinguish the nuanced patterns arising from compression-induced inconsistencies. The proposed method showcases the culmination of this theoretical framework, culminating in a practical and applicable solution. Figure 4.1 visually encapsulates the sequential stages of our method. Initiated by the input image, the process encompasses intricate feature extraction, enabled by the CNN's hierarchical layers. Subsequently, the model's learned discriminative capabilities come into play, facilitating the classification of the image's authenticity. This comprehensive workflow epitomizes the operationalization of our pioneering approach, underscored by the fusion of compression artifact analysis and CNN-based image evaluation. In this context, a novel technique has been devised for the automated detection of counterfeit images, leveraging CNNs.

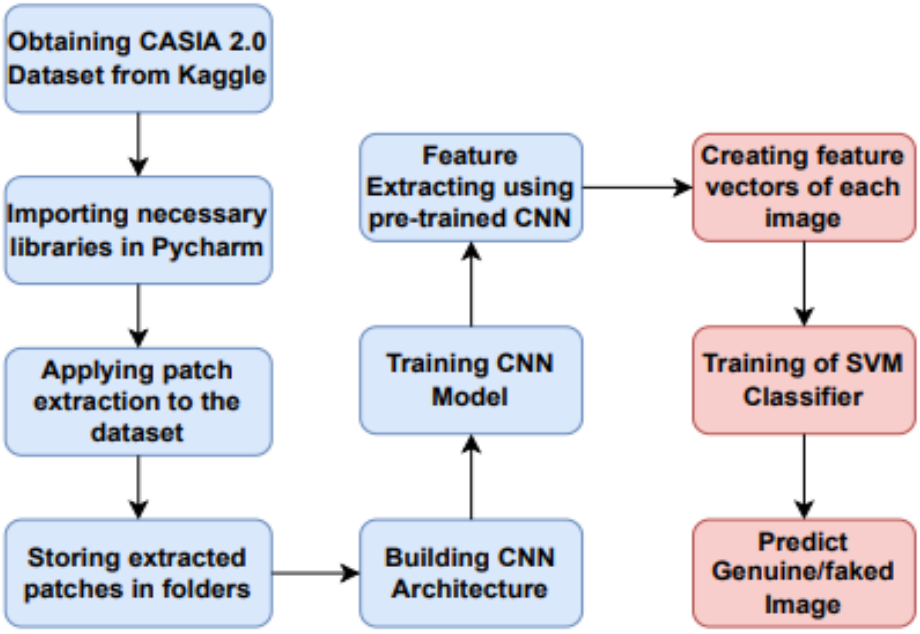


Figure 4.1: Proposed Classification Model

Our approach involves the automatic construction of hierarchical representations through convolutional operations, applied to input-coloured images or patches. Specifically, the CNN architecture proves to be highly advantageous for addressing challenges within copy-move detection and image splicing scenarios, wherein tampered regions are manipulated or duplicated. A distinctive feature of our proposed technique lies in the initialization of the first layer's weights within the CNN architecture. These weights are strategically set to correspond with fundamental high-pass filters, as utilized in the creation of residual maps within a spatial rich model (SRM). The incorporation of SRM-based filters functions as an innovative regularization strategy, yielding multiple advantages. This incorporation of SRM-based initialization serves a dual purpose. Firstly, it effectively mitigates the influence of the underlying image contents during the detection process. Traditional detection mechanisms often struggle with capturing subtle inconsistencies arising from tampering due to the dominance of content-related features. By integrating SRM-derived filters, our CNN-based approach acquires an enhanced capability to discern these intricate artifacts, thereby reinforcing its sensitivity to tampering cues.

Secondly, the utilization of SRM-driven initialization aligns with the requirement to address diverse tampering procedures. The complex nature of image manipulation techniques necessitates an adaptable approach that can accommodate a spectrum of potential alterations. By utilizing filters rooted in the SRM framework, our CNN model gains a robustness that

extends across a wide array of tampering scenarios, enhancing its versatility and applicability. The hierarchical construction of representations through CNN layers acts as a pivotal element in our approach. These layers progressively extract abstract features, allowing the network to progressively learn and represent increasingly complex patterns inherent in the images. Through these hierarchical transformations, the CNN adapts to the intricacies of both authentic and manipulated visual content, thereby enabling accurate differentiation between the two. In essence, our technique embodies a synergistic fusion of CNNs and the foundational principles of a spatial rich model. By initiating the CNN architecture with SRM-derived filters, we harness the advantages of regularization and robustness, culminating in a discernment mechanism that excels in detecting counterfeit images. The hierarchical extraction of features within the CNN further augments its proficiency, ensuring a comprehensive grasp of the underlying visual elements. Collectively, these components coalesce to establish an innovative approach that exhibits notable potential within the realm of counterfeit image detection. Upon extracting dense features from test images utilizing a pre-trained CNN as a descriptor, the culmination of our approach involves the generation of final features for support vector machine (SVM) classification through a feature fusion process. This step amalgamates the intricate visual characteristics captured by the CNN, paving the way for robust classification. To rigorously evaluate the efficacy of our method, a comprehensive assessment was conducted. A publicly accessible dataset renowned for its utilization in picture forgery identification was employed for comparative analysis. This enabled a meticulous appraisal of our approach's accuracy in contrast to a previously employed image forgery detection system, fostering a data-driven validation of our technique's proficiency. The cornerstone of our approach lies in the strategic architecture of the CNN employed for feature extraction. This architecture, depicted in Figure 4.2, is purposefully designed to encapsulate the unique requirements of our forgery detection framework. The CNN's hierarchical layers enable the progressive extraction of intricate visual elements, allowing the network to comprehend both subtle and prominent cues indicative of image tampering. By utilizing a pre-trained CNN as a descriptor, our approach taps into the wealth of knowledge encoded within the network's learned weights. This enables an efficient and effective portrayal of images in a feature-rich manner, forming the basis for subsequent SVM classification. The fusion of features accentuates the discriminative capabilities, ensuring a comprehensive representation that is well-suited for accurate classification. Our methodology culminates in the creation of robust final features for SVM classification through a feature fusion process, leveraging dense features extracted from test

images using a pre-trained CNN. The extensive evaluation against a benchmark dataset reinforces the effectiveness of our approach in image forgery identification. This validation is substantiated by the tailored CNN architecture designed to encapsulate the intricacies of the forgery detection task. The holistic framework, as depicted in Figure 4.2, illustrates the orchestrated integration of these elements, ultimately contributing to a potent image forgery detection methodology.

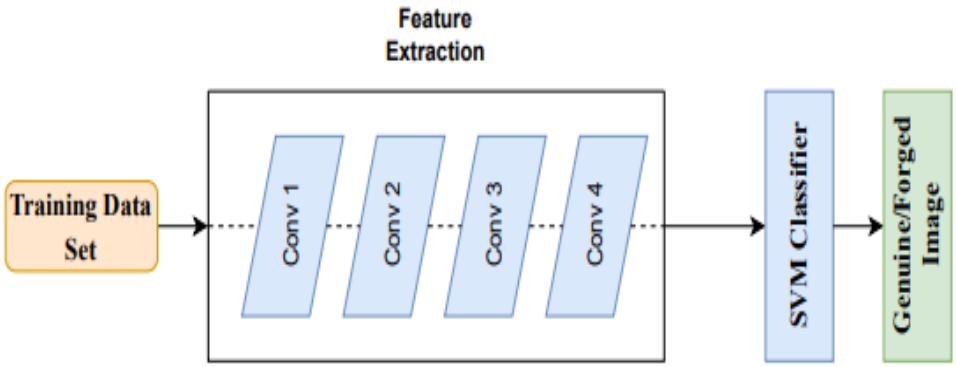


Figure 4.2. CNN Architecture for Classification of Forged Images

4.3 *CNN Layers Operations Involved in Proposed Framework*

A CNN model is developed with key components: it starts with convolution layers, followed by fully connected layers, and ends with a SoftMax classifier. Each convolution layer is composed of three main components: convolution operations, non-linear activation functions, and pooling operations. Notably, feature maps serve as the input to these convolutional layers, facilitating the extraction of hierarchically structured features.

The intricate orchestration of these components is encapsulated within the CNN Architecture execution sequence, as detailed below:

- **Convolution Operation:** In this initial phase, convolutional filters are applied to the input feature maps. This process involves sliding these filters across the input, detecting localized patterns and generating feature maps that emphasize distinct visual cues.
- **Non-Linear Activation:** After the convolution step, non-linear activation functions like ReLU are applied to the resulting feature maps. This brings in non-linearity, allowing the network to understand intricate relationships within the data.
- **Pooling Operation:** After applying the non-linear activation, the network performs pooling operations. Pooling helps shrink the feature maps, reducing their size while

keeping vital information. Max pooling and average pooling are popular methods used for this purpose.

- **Fully Connected Layers:** Following several convolutional and pooling layers, the network incorporates fully connected layers. These layers create broad connections throughout the neural network, aiding in a comprehensive understanding of features gathered from earlier steps.
- **SoftMax Classifier:** The final phase involves the SoftMax classifier, responsible for assigning probabilities to classes. It computes the likelihood of each class being the correct classification for a given input, aiding in the ultimate categorization.

This intricate sequence reflects the CNN's capacity to progressively extract intricate features through successive convolutional and pooling layers. The utilization of non-linear activation functions enriches its ability to capture nuanced relationships, while pooling operations and fully connected layers contribute to global feature integration. The SoftMax classifier provides a probabilistic framework for making categorical predictions based on the network's learned representations. In essence, the CNN Architecture execution sequence underscores the interplay of convolution, non-linearity, and pooling in a hierarchical manner, culminating in a powerful framework for feature extraction, abstraction, and classification.

After extracting extensive features from dataset images using a pre-trained CNN, we generate N sets, each containing 400 features. These sets are amalgamated into a unified representation per image to facilitate subsequent support vector machine (SVM) classification. The SVM's role is crucial in detecting potential image alterations. It operates by leveraging these combined feature vectors as input, enabling the system to discern and classify alterations within the images based on the consolidated representations. This methodological approach, utilizing pre-extracted features and SVM classification, forms a fundamental stage in identifying and categorizing potential modifications or anomalies present within the images sourced from the dataset. The SVM leverages these high-dimensional feature vectors to learn patterns and relationships indicative of image tampering, subsequently providing predictions on, determining if any modifications have been made to an image. The integration of these steps underscores the holistic approach in which the pre-trained CNN, feature extraction, and SVM classification collectively contribute to the task of image authenticity assessment. By strategically merging the N feature representations and harnessing the discriminative power of the SVM, our methodology encapsulates both local and global visual cues, thus enhancing its accuracy in detecting image manipulation.

4.4 *Dataset Discussion and Result Analysis*

The CASIA v2.0 dataset encompasses a fusion of three distinct categories of image collections, each contributing to the dataset's comprehensive composition:

- Authentic Images: This subset encompasses unaltered images, devoid of any editing interventions. Constituting a significant portion, this category comprises 7,491 images, each in its original state, serving as a benchmark for authentic imagery.
- Tampered Images: Within this category, images have undergone diverse forms of manipulation, resulting in a total of 5,123 photographs. Primarily characterized by operations like copying, pasting, and combining, these images illustrate various manifestations of tampering and represent a critical aspect of the dataset.
- Masks: The technique of pixel masking is employed to accentuate regions of alteration. Specifically, the pixel values of tampered images are strategically set to 0, delineating the altered region. Concurrently, all non-altered background pixels are also assigned a value of 0. This meticulous masking procedure enables the precise extraction of altered regions while emphasizing their contrast with the unchanged background.

The amalgamation of these three distinct image categories within the CASIA v2.0 dataset culminates in a robust and comprehensive resource for image forensics and tampering detection research. The juxtaposition of authentic images, manipulated counterparts, and the innovative masking approach collectively equips researchers and practitioners with a diverse range of data, fostering a deeper understanding of image manipulation and enhancing the efficacy of tampering detection algorithms. The subsequent table outlines the description of attributes present within the dataset:

Table4.1: Image dataset detail.

Authentic Images	**Tampered Images**	**Masks**
7,491	5,123	5,123

The devised methodology encompasses the development of a pivotal function, aptly named "Patch Extractor," tailored to the precise extraction of patches from images. This function is invoked by furnishing it with a set of essential arguments, elucidating the intricate parameters and specifications governing the patch extraction process. Upon invoking the Patch Extractor

function, a series of critical arguments are supplied to orchestrate the extraction procedure. These encompass the dataset's path, serving as the source repository of images, and the designated path of the output, which designates the location where the extracted patches will be deposited. Additionally, the function accommodates the specification of a stride value, dictating the spatial shift between consecutive patches shown in Figure 4.3. Moreover, a parameter controlling rotation is integrated, enabling the generation of rotated patches, thereby augmenting the diversity of the extracted dataset. A fundamental component of the Patch Extractor function is the determination of the desired number of patches to be extracted from each image. This parameter facilitates the fine-tuning of the granularity of patch extraction, catering to specific research objectives and experimental requirements. By encapsulating these diverse arguments and parameters, the Patch Extractor function demonstrates a versatile and adaptive framework for efficient patch extraction. This foundational function not only streamlines the extraction process but also offers a customizable avenue to tailor the extraction process in alignment with distinct research needs, ultimately enhancing the precision and versatility of patch-based analysis.

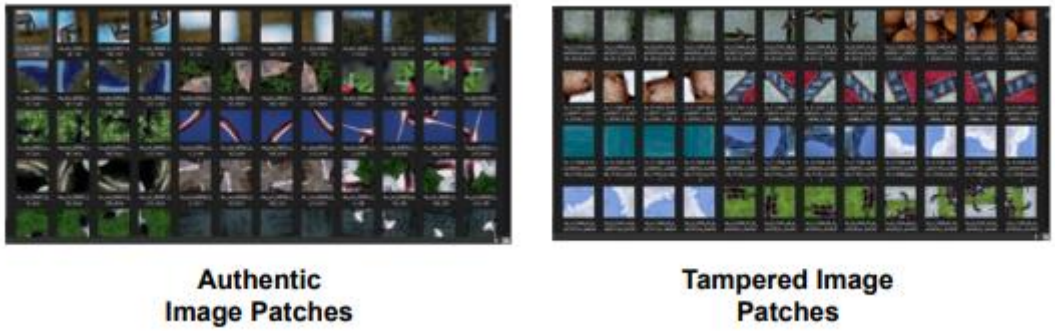


Figure 4.3. Extraction of Patches from Genuine and Altered Images

Our algorithm effectively identified instances of image tampering, achieving a notable training accuracy of approximately 96%. Notably, the accuracy curve showcases a discernible pattern of improvement throughout the training process. Commencing at an initial level of around 50% accuracy during the early epochs, the algorithm's performance progressively advances. The trajectory of accuracy demonstrates a consistent upward trend, exhibiting a steady and incremental rise. This upward momentum continues until a point of stability is reached at approximately 86-87%, a level consistently maintained during the final epochs. This plateau in accuracy implies the algorithm's proficiency in reliably discerning tampered images, as indicated by its sustained performance over multiple iterations. Simultaneously, the loss function, a pivotal metric in optimization, exhibits an inverse pattern. Initiated with

the first epoch, the loss value gradually decreases over subsequent epochs. This gradual reduction underscores the algorithm's ability to align its predictions with actual outcomes, optimizing its parameter settings to minimize discrepancies. The loss function gradually converges, culminating in a stable pattern during the concluding epochs. This convergence further corroborates the algorithm's convergence towards an optimal state, signifying an effective adaptation to the underlying data distribution. Our algorithm showcases commendable performance in detecting image tampering, yielding an impressive training accuracy of approximately 96% depicted in Figure 4.4. The accuracy's consistent ascent and subsequent stabilization, along with the gradual decline in loss, collectively underscore the algorithm's robustness and efficacy in addressing the complex task of image tampering detection.

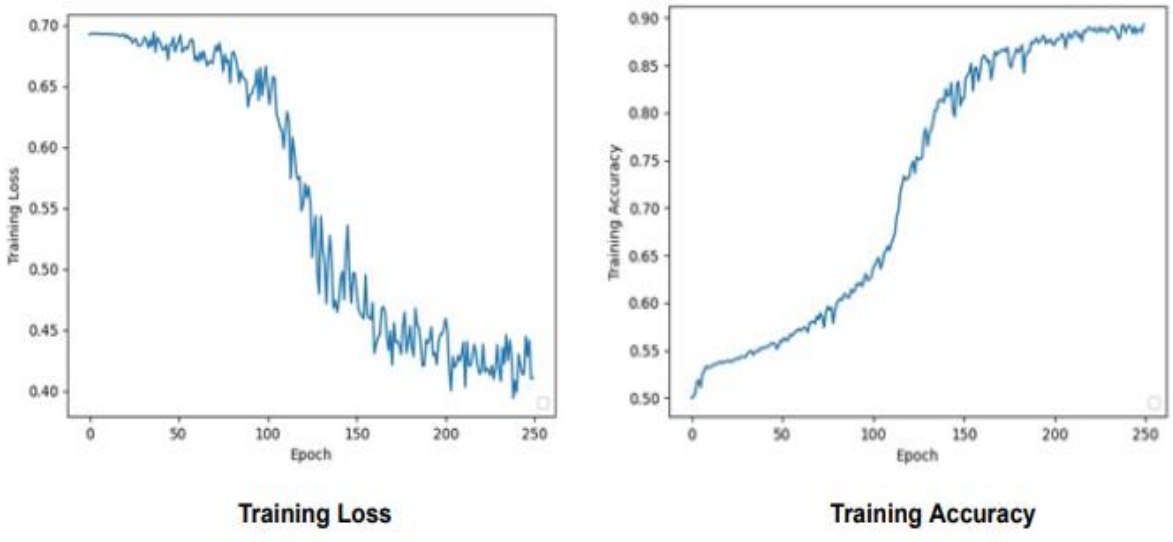


Figure 4.4. Training Loss and Accuracy of the CNN Model

For the comprehensive assessment of the model's performance, the test dataset serves as the foundation for the construction of a confusion matrix. In this evaluative endeavor, the support vector machine (SVM) demonstrated its adeptness in discerning between altered and original images. Specifically, among the pool of 1,008 tampered images, the SVM proficiently distinguished them from 1,426 unaltered counterparts, yielding a commendable accuracy. While the SVM's classification prowess is evident, a marginal subset of misclassifications occurred. Specifically, a total of 72 altered images were erroneously categorized, falsely resembling original images. Furthermore, a limited count of 17 genuine images encountered misclassification, being inaccurately associated with the tampered category. The resulting confusion matrix encapsulates this interplay of correct and misclassifications, offering a quantitative depiction of the SVM's performance across the dataset. This analytical

framework furnishes an insight into the model's precision, recall, and overall discriminative capabilities, thus providing a comprehensive profile of its efficacy in image tampering identification. The SVM's evaluation via the confusion matrix showcases its prowess in accurately discerning between altered and original images, exemplified by the substantial accuracy achieved. The slight occurrence of misclassifications, though limited, offers an avenue for potential refinements, thereby contributing to the continuous enhancement of the algorithm's detection accuracy.

Table 4.2: Outcome of Classification Prediction.

Image Category	Actual Image	Predicted	Misclassified
Native	1443	1426	17
Forged	1152	1008	72

The application of SVM classification yielded a commendable accuracy of 96.8% within our model. To gauge the efficacy of various recommended strategies for detecting image forgeries, a comprehensive comparative analysis was undertaken is shown in Figure 4.5.

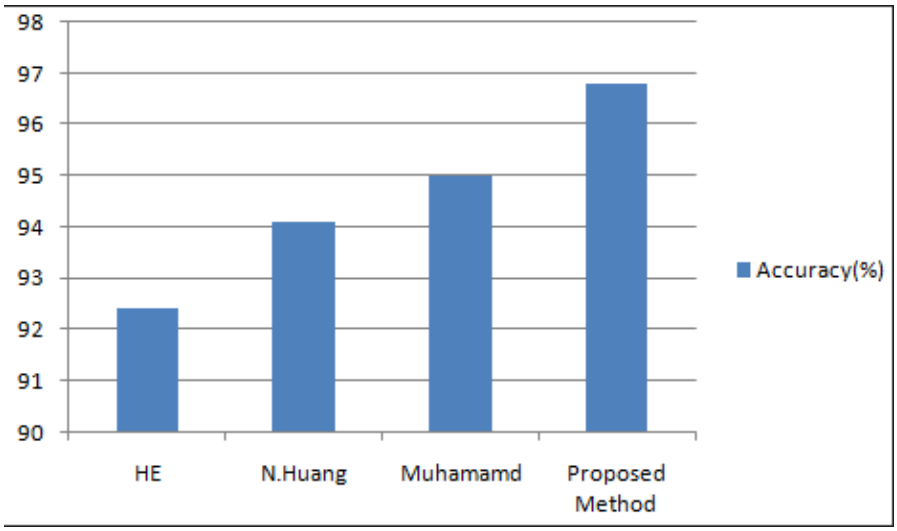


Figure 4.5. Comparative Accuracy Analysis

Despite the notable progress achieved in image forgery detection, the field remains ripe for further advancements to enhance its efficacy in the future. Notably, neural networks have demonstrated remarkable capabilities even in the face of challenges, exhibiting a high degree of performance and showing promise in their potential to discern altered images. The augmentation of the CNN layer holds the prospect of refining the model's detection prowess. Multiple adjustments can be explored to fine-tune its architecture, thereby potentially elevating the detection rate and fostering even more precise results. However, a critical

consideration arises from the dataset composition, particularly in the context of real-world manipulation scenarios. The CASIA dataset, while comprehensive, may lag behind in accurately reflecting the diverse array of alterations encountered in practical situations. To address this limitation, a more expansive and representative dataset encompassing a wider spectrum of tampering techniques is imperative. Upon delving into the details of the confusion matrix, intriguing observations emerge. Instances of misclassification highlight the complex nuances inherent in image analysis. Certain images were incorrectly categorized, often sharing common characteristics such as blurring, fog, sun flare, or reflection. These intricacies underscore the intricacies of tampering detection, where the visual interplay of authentic and manipulated elements can lead to subtle misinterpretations. Of particular note is the instance where genuine images, albeit featuring blurred areas, were mistakenly labelled as manipulated. This scenario emphasizes the sensitivity of the detection process to nuanced visual traits, where certain genuine images might inadvertently exhibit features characteristic of tampering.

4.5 *Conclusions*

In conclusion, the domain of image forgery detection continues to hold untapped potential for future advancements. Neural networks, such as CNNs, exhibit prowess in spite of challenges, hinting at their evolving capacity to identify tampered images. The refinement of model architectures and the integration of diverse datasets are key pathways towards augmenting accuracy. The intricacies revealed in the confusion matrix emphasize the intricacies of image analysis, warranting a nuanced approach to ensure precision in discerning authentic and manipulated content.

CHAPTER 5
MULTI-MODAL CAMERA MODEL IDENTIFICATION IN VIDEOS USING DEEP LEARNING-BASED CNNS

5.1 *Video Forensic Process*

In comparison to the well-established use of traditional photography-based evidence in legal cases, the areas of source forensic query video analysis and processing multimedia evidence are relatively new and continually developing fields. The latter segment, referred to as "enhanced forensic video analysis" [190], is the focal point of our endeavors. It entails the meticulous scrutiny of videos and associated data through advanced analytical tools, aiming to uncover intricate details and insights. Our work is specifically oriented towards the domain of enhanced forensic video analysis. This involves the orchestration of a comprehensive architectural framework, as illustrated in Figure 5.1. This architecture encompasses three pivotal components, each contributing to the holistic process:

- Crime scene analysis: This foundational phase involves a meticulous assessment of the crime scene, establishing a context for subsequent analysis. Factors such as lighting, camera angles, and environmental conditions are scrutinized to inform the subsequent data collection and analytical phases.
- Data collection, video enhancement and analysis: Central to the process is the systematic collection of relevant data, which forms the basis for subsequent analysis. Advanced video enhancement techniques are employed to optimize the visual quality of the content. The resultant data is then subjected to rigorous analysis, aiming to unveil hidden details, patterns, and anomalies.
- Presentation and findings enlargement: The culmination of the analysis phase involves the synthesis of findings into a coherent and compelling narrative. Employing effective presentation techniques, the enhanced insights are showcased to stakeholders, enhancing their understanding and contributing to informed decision-making.

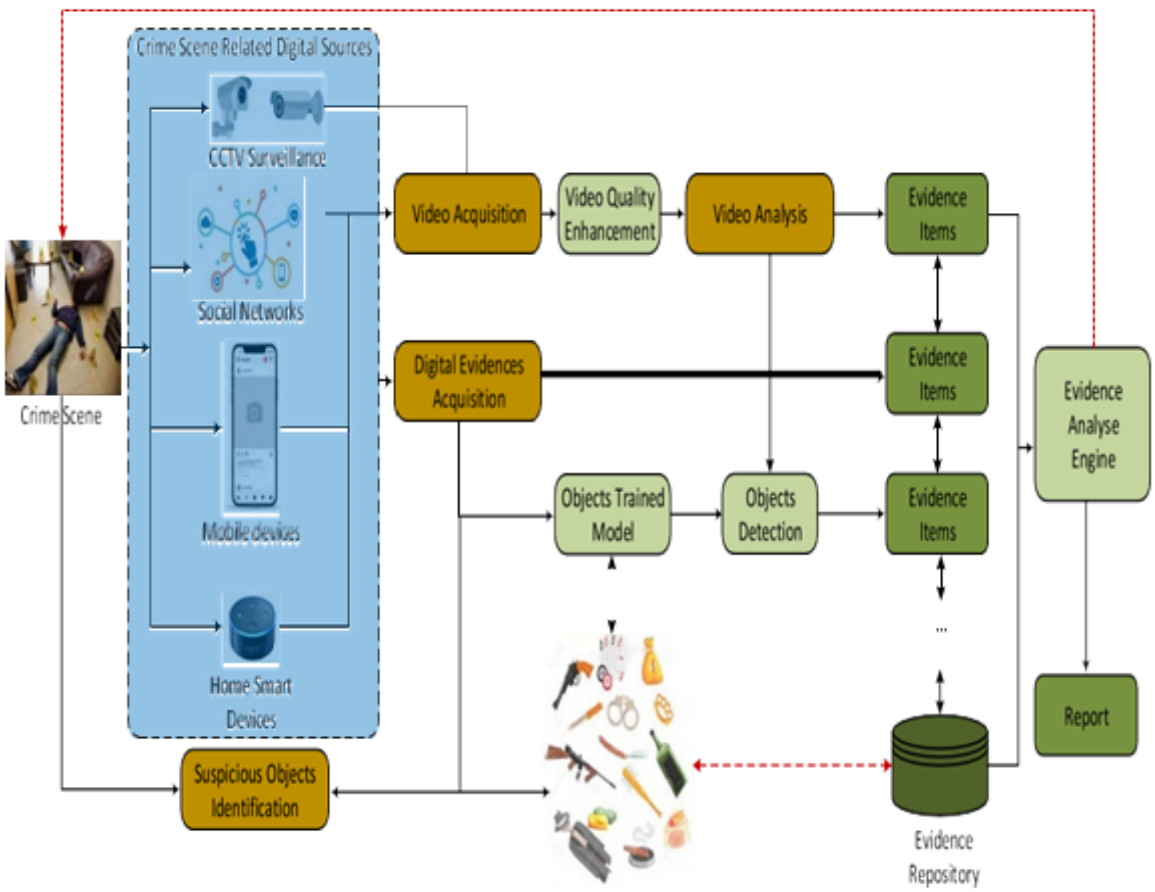


Figure 5.1. Advanced Framework for Forensic Video Analysis

The essence of our work lies in harnessing cutting-edge video analysis tools to decipher complex forensic evidence. Through the intricate interplay of crime scene analysis, data collection, video enhancement, analysis, and presentation, our enhanced forensic video analysis architecture embodies a comprehensive approach to elucidate intricate details that might otherwise elude conventional scrutiny. This systematic approach aligns with the evolving landscape of forensic video analysis, contributing to the advancement of this nascent yet crucial domain.

5.2 *Approach for the Analysis of Query Videos for Source Identification Forensic*

The preceding framework underscores the clear delineation of two fundamental categories within forensic video analysis. These categories are notably characterized by a systematic classification based on the scrutiny of video content and its inherent type.

5.2.1 *An Investigation into Types of Forensic Video and Analytical Approaches*

A conspicuous aim of forensic video analysis is the discernment of unauthorized reproduction or tampering within video files. Furthermore, an imperative task is to ascertain potential alterations inflicted upon the video content. Additionally, the study is positioned to unveil concealed information through the identification of the video source and a meticulous examination of video steganography for the detection of covert data. Notably, the identification of the video source serves as a pivotal evidentiary foundation, as evidenced by research [161]. This identification process holds significance in determining whether the video source emanates from a camera or a tokenized device, as depicted in Figure 5.2. Forensic audio analysis, forensic video analysis, image analysis, and computer forensics have all been formally established as distinct fields of inquiry by the American society of crime laboratory directors laboratory accreditation board (ASCLD/LAB). Notably, a surge in the formation of digital and multimedia divisions is witnessed across a spectrum of private, public, and state/local law enforcement entities. These specialized units often encompass various or all of the aforementioned disciplines. Notably, certain scenarios witness the same examiner undertaking examinations for multiple agencies, showcasing the interdisciplinary nature of their role. Specialization frequently ensues for examiners within larger federal and state agencies, as well as across various fields, culminating from extensive training and evolving into subject matter expertise over years. The realm of video evidence enhancement offers a range of techniques, as exemplified by studies. Of paramount importance is the initial submission of high-quality video recordings, a crucial prerequisite for yielding optimal outcomes through the enhancement process. It is imperative to refrain from submitting digitally compressed or analog copies that have undergone additional compression. Such files are rendered unsuitable for enhancement due to the cumulative effect of compression, which diminishes their capacity to undergo further improvement.

5.2.2 *Enhancement of Videos Techniques*

To accomplish this objective, a diverse spectrum of methodologies has been employed in the past decade to enhance video quality. These approaches have found applications in video monitoring systems, intelligent highway systems, safety-monitoring systems, and various other contexts. For instance, introduced an innovative technique that incorporates color

information into low-quality video footage to facilitate luggage identification. A distinctive strategy involves the construction of human-like temporal templates to discern an object's motion direction. By accurately aligning these templates with pertinent parameters, the object's trajectory can be effectively ascertained. Numerous researchers advocate for the establishment of luggage detection systems. Chuang et al., for instance, conducted a study aimed at identifying missing colors via a ratio histogram. This endeavor exemplifies the breadth of techniques employed to improve video quality and underscores the multifaceted nature of video enhancement in diverse application domains. The variable under consideration corresponds to the ratio derived from color histograms [174]. To identify absent colors, the integration of a tracking model becomes imperative. In the context of low-quality videos, the foremost objective of forensics is to extract maximal information from them, thereby bolstering the investigative process. This section endeavors to outline strategies aimed at augmenting video quality to amplify information extraction capabilities. Specifically, when dealing with low-quality videos or images, employing histogram equalization (HE)-based methodologies exhibits heightened potential for detecting supplementary information in contrast to conventional techniques. A pertinent illustration involves the utilization of a webcam to discern objects, employing the recommended technique as depicted in Figure 5.2. This exemplifies the efficacy of the proposed approach in enhancing information retrieval from low-quality visual data.

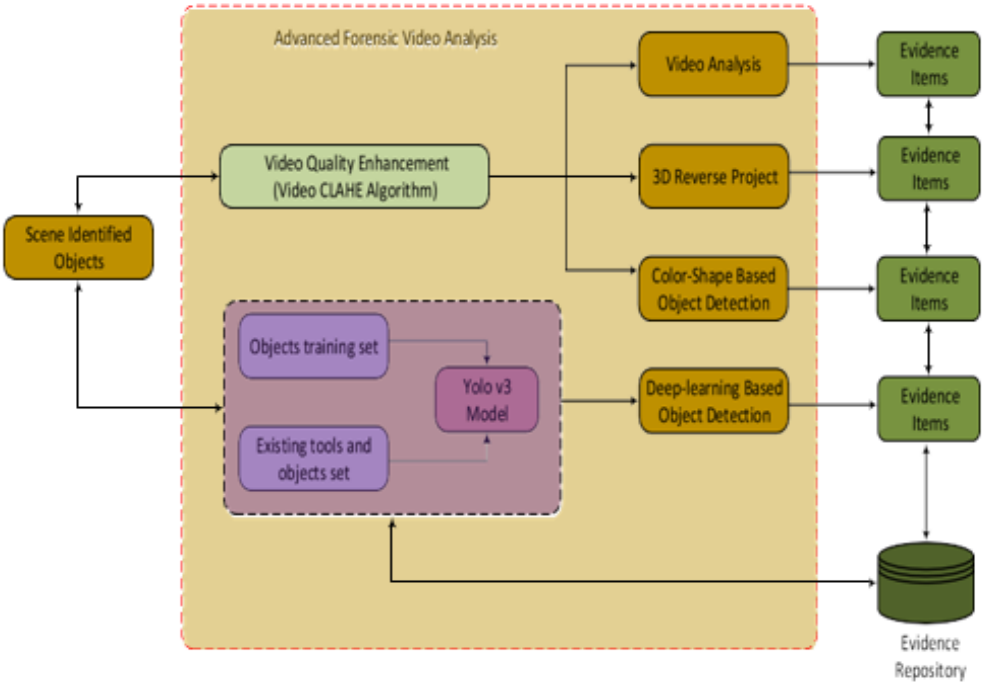


Figure 5.2. Advanced Forensic Video Analysis Techniques

Our research focal point centers on the intricate task of camera model identification within video sequences, predicated on the intrinsic content of these visual data streams. The principal thrust of our investigation resides in the discernment of the originating camera model from digital video sequences, as delineated by the work of [191]. The impetus for this inquiry stems from the extensive exploration of digital image analysis within the forensic domain, resulting in remarkable achievements. Within the scope of this study, our attention is honed on video sequences hailing from a diverse array of smartphone models. Our innovative approach converges informational and auditory components within these videos, fostering a comprehensive analytical framework, an approach articulated by [192]. The research trajectory commences with an exploration of the classic mono-modal paradigm. This facet delves into the endeavor of source camera model identification rooted exclusively in either visual or auditory attributes. The ensuing sections of this discourse expound upon this mono-modal issue in detail. Subsequently, the research landscape transcends into the realm of the multi-modal quandary, a core tenet of our investigation. This intricate facet amalgamates both visual and auditory cues, culminating in an evolved problem wherein the confluence of these modalities contributes to the determination of the sound source's origin. our research endeavors to unravel the complexities of camera model identification within video sequences.

5.3 *Camera Model Identification Approaches*

We navigate through the monolithic mono-modal pursuit before delving into the nuanced intricacies of the multi-modal inquiry, encompassing both visual and auditory dimensions. This comprehensive exploration promises to unveil novel insights and methodologies in the realm of advanced forensic video analysis.

5.3.1 *Mono-Modal Camera Model Identification*

Consequently, the underlying quandary materializes in the form of discerning the device model employed for capturing specific media within a singular modality. For instance, consider the scenario where an image is captured; in such instances, it becomes indispensable to ascertain the precise camera model responsible for the image's acquisition. This attribution serves a pivotal purpose: it enables the retracing of the image's lineage to its point of origin. Furthermore, this attribution extends to audio recordings, necessitating the inclusion of the recorder's model alongside the recorded audio, as expounded by [193]. Within the mono-modal attribution paradigm, the crux lies in associating a video, the focus of our inquiry, with

the device category responsible for its capture. This attribution rests solely upon the visual or auditory cues intrinsic to the video's content. In essence, the essence of the attribution process hinges on the distinct characteristics exhibited by the visual and auditory components within the video. In practical terms, this implies that the model of the camera or recorder wielded for capturing the media leaves an indelible imprint on the ensuing visual or auditory data. The mono-modal framework operates within the confines of each modality, discerning the device model with a singular focus—be it the camera that captured a compelling image or the recorder that preserved a captivating audio snippet. As we delve into the intricacies of this mono-modal model attribution, we unravel the interplay between media, modality, and device model. This exploration ventures beyond the surface and delves into the nuanced dynamics that underlie the mono-modal attribution process, enhancing our comprehension of the intricate tapestry that defines the origins of visual and auditory data within the realm of video analysis. Our study navigates the intricacies of device model attribution within a solitary modality, unveiling the essence of tracing media origins through the identification of capturing devices. The mono-modal attribution, dissected within the context of video analysis, underscores the significance of visual and auditory cues as the linchpin for unveiling the device model that shapes the recorded content. This framework is integral to unravelling the intricate tapestry of multimedia attribution and embodies a pivotal stepping stone within advanced forensic analysis.

5.3.2 *Multi-Modal Camera Model Identification*

In the context of a video, identifying the camera model through multiple modes becomes a complex challenge: accurately determining the device used to record the video. This task involves gathering both visual and auditory data from the video. In the following example, we'll explore a closed-set identification process where the main goal is to figure out the exact camera model that recorded the video sequence. This determination is grounded in a predetermined roster of known devices previously employed for recording purposes, as elucidated by [194]. In this analytical pursuit, a foundational assumption is established: the video under scrutiny emanates from a device within a designated family, one acquainted with the investigator. This familiarity guides the investigator's inclination to associate the video with a device belonging to this familiar device family. The assumption presupposes that the recorded video aligns with the characteristic traits and idiosyncrasies intrinsic to the devices encompassed within the designated family. However, an inherent susceptibility prevails within this attribution process. The investigator could potentially err in ascribing the video to

a specific device within the familiar family, inadvertently assigning it to a device that it did not originate from. This misattribution can materialize if the video has been captured by a device external to the pre-established familial pool. To mitigate this potential misclassification, a comprehensive evaluation of both visual and aural attributes within the video sequence becomes paramount. The intricate interplay between these modalities contributes to a more accurate device model identification, reducing the likelihood of erroneous attributions. By scrutinizing the amalgamation of visual and aural cues, the investigator gains a holistic perspective, discerning nuances that aid in the precise identification of the originating camera model. The multi-modal camera model identification endeavor is encapsulated within the intricate domain of video sequence analysis. This challenge hinges upon the discernment of recording device models, enlisting both visual and aural cues as pivotal discriminators. The closed-set identification process, characterized by a roster of familiar device models, underscores the investigatory nature of the pursuit. However, the propensity for misattribution necessitates a comprehensive and nuanced evaluation, ensuring the alignment of the video's inherent traits with the designated familial attributes for accurate device model assignment.

5.4 *Proposed Methodology*

This research introduces a robust approach to the identification of closed-set multi-modal camera models within video sequences, warranting further exploration and investigation. The schematic representation of this pioneering approach is depicted in Figure 5.3, encapsulating the essence of the proposed methodology. Central to our method is the utilization of both visual and aural attributes intrinsic to the video content to ascertain the specific smartphone model employed for recording. Through the integration of these dual modalities, the inherent disparities among diverse camera models utilized within the source video cameras can be effectively discerned, thereby enabling precise model identification, as elucidated by [195].

The proposed strategy can be succinctly outlined through two pivotal stages:

- **Preprocessing and Content Extraction:** This phase encompasses the extraction of salient visual and auditory information embedded within the scrutinized videos. Prior to the input into the CNNs, the data undergoes meticulous manipulation and enhancement, an operation referred to as preprocessing and content extraction. This preparatory stage is instrumental in facilitating effective data representation within the CNNs, enhancing their analytical efficacy.

- **CNN Processing Block:** The crux of the method resides within this block, which is divided into two integral components: the Extraction Block and the Classification Block, each contributing distinct functions. The Extraction Block processes the raw data, effectively parsing it into discriminative features that encapsulate the essential attributes of the visual and auditory content. Subsequently, the Classification Block, constituted by a CNN, undertakes the intricate task of model classification based on the feature-rich input. This involves the CNN's capacity to identify and differentiate the unique characteristics that demarcate various camera models, thereby elucidating the specific smartphone model utilized for recording.

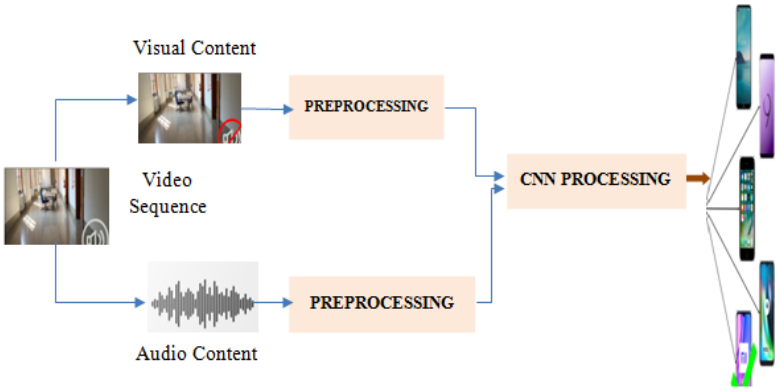


Figure 5.3. Flowchart Illustrating the Proposed Methodology

The synergy between these steps culminates in a comprehensive and sophisticated approach to multi-modal camera model identification within video sequences. By harnessing the power of visual and auditory cues, coupled with the analytical prowess of CNNs, the methodology transcends the boundaries of conventional unimodal techniques. This innovative method equips researchers and practitioners with a robust tool for the precise determination of camera models, thereby augmenting the realm of advanced forensic video analysis. As future investigations delve deeper into this method, its potential for refinement and enhancement remains an exciting avenue for exploration.

5.4.1 *Content Extraction and Pre-Processing*

Our methodology starts with extracting and preparing visual and audio content, focusing on thorough data standardization. This first phase involves a three-part process (shown in Figure 5.4) that's highly detailed in extracting and preparing the visual content found in the examined video. These phases encompass:

- Temporal Frame Selection: The extraction of color frames from the video stream (Nv) is executed strategically, selecting frames that are uniformly distributed over an extended temporal interval [196]. This approach ensures temporal diversity and robustness in frame representation. The video frames are categorized into two dimensions, Hv and Wv, which correspond to their height and width, dictated by the resolution of the video under examination.
- Random Patch Extraction: In this phase, a random sampling process is employed to extract NPv color patches, each characterized by dimensions HPV and WPV. These patches serve as the data input for the CNNs. Subsequently, the extracted patches undergo normalization to facilitate optimal training conditions within the CNNs. This normalization step aims to achieve a zero mean and unit variance across the data, contributing to enhanced convergence and performance during the subsequent analysis.
- Audio Content Extraction and Pre-processing: Beyond visual content, this phase encompasses the extraction and pre-processing of audio data embedded within the video. Techniques such as spectrogram analysis may be applied to convert audio signals into a visual representation, facilitating subsequent analysis within the multi-modal framework.

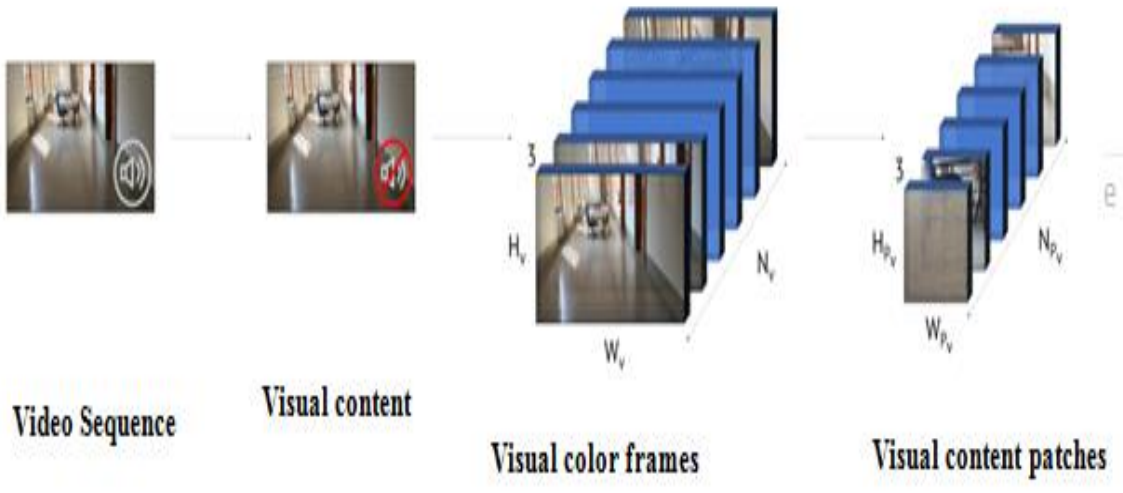


Figure 5.4. color frames from Nv, extracted as HV and WV sizes. As a result of this analysis, randomly extraction of NPv visual patches of size HPvWPv from these frames

The intricate orchestration of these phases within the extraction and pre-processing stage lays the foundation for comprehensive data representation, poised for insightful analysis. This preparatory groundwork is instrumental in fostering the accuracy and efficacy of the ensuing CNN-based analysis. The extraction and pre-processing phase of our methodology constitute a pivotal preliminary step in the multi-modal camera model identification process. By harmoniously integrating the temporal frame selection, random patch extraction, and audio content pre-processing, the method primes the data for ingestion into the Convolutional Neural Networks. This synergistic approach ensures data uniformity, diversity, and optimal normalization, culminating in a robust data representation that forms the bedrock for subsequent advanced analyses within the proposed framework.

The extraction and preparation of audio content from the scrutinized movie encompass a structured process outlined in three distinct phases, as delineated in Figure 5.5:

- **Audio Content Extraction**: The initial phase entails extracting audio content from the linking matrix set (LMS L) associated with the video sequence. The significance of the LMS L as a robust tool for audio data is underscored by its extensive application in various audio and speech classification studies. Through exploratory experimentation, several audio attributes were derived from the short-time fourier transform (STFT) signal's magnitude and phase. Notably, the LMS (predicated on the STFT signal's magnitude) emerged as the optimal choice, rendering superior results. It is noteworthy that LMS outperformed phase-based methods [197], yielding an accuracy rate exceeding 80%. The LMS L, visualized as a matrix of dimensions Ha Wa, is characterized by rows representing temporal nuances (varying in alignment with video length) and columns delineating frequency content in Mel units.
- **Random Patch Extraction**: Subsequent to the audio content extraction, NPa patches, each of dimensions HPaWPa, are randomly extracted from the LMS L. This step enhances the diversity and comprehensiveness of the audio data that will be subject to further analysis.
- **Patch Normalization**: The third phase of this process revolves around patch normalization, akin to the normalization applied to the visual patches. This normalization is instrumental in ensuring that the extracted audio patches exhibit a zero mean and unitary variance. This enhancement facilitates consistent and optimized data representation, fostering precision and efficacy during subsequent analytical operations.

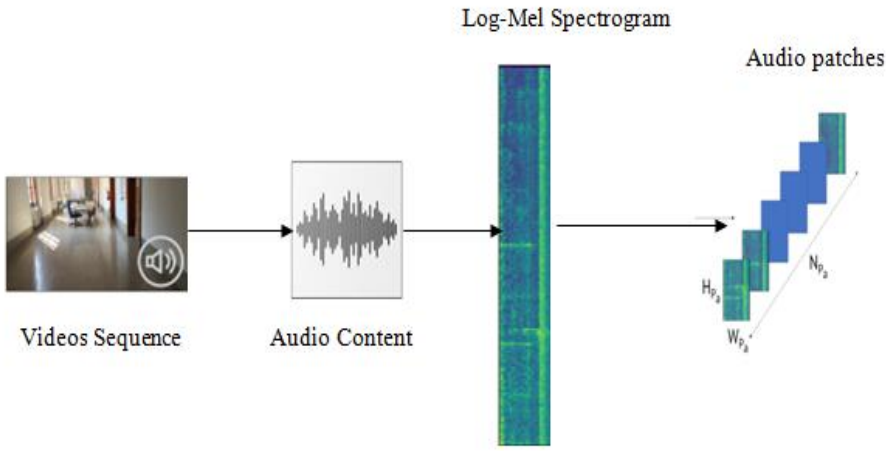


Figure 5.5. Computation of LMS after the audio content, which has the size Ha Wa, has been selected. Random extraction of NPa audio patches with sizes HPaWPa from the NPa audio patches

In summation, the comprehensive treatment of audio content extraction and preparation within the proposed methodology encompasses a tri-fold process. Commencing with the extraction of audio content from the LMS L, and proceeding to the extraction of randomized audio patches, culminating in patch normalization, this systematic approach ensures the meticulous handling and optimization of audio data. These preparatory steps collectively lay the groundwork for subsequent analysis within the multi-modal camera model identification framework.

5.4.2 *CNN Processing*

Upon retrieval of the pre-processed data, it is subsequently channeled into one or multiple CNNs within the CNN processing stage. Here, the objective is to elicit distinctive features corresponding to diverse source camera models and effectuate their subsequent classification. A tangible manifestation of this approach lies in addressing the mono-modal camera identification problem by feeding either the retrieved visual or auditory data into a CNN, as illustrated by [198].

While any CNN architecture capable of data classification can, in principle, be harnessed during this stage, our rationale behind the chosen architecture is expounded upon in subsequent sections. The final layer of the classification network consists of a fully connected layer containing nodes equal to the total number of camera models (M). Each node represents a distinct camera model incorporated into the network. In application, the result is an M-element vector called "y". Within this vector, the element "ym" encapsulates the probability or likelihood that the model affiliated with the respective node accurately captured and processed the input data. This vector consequently facilitates the extraction of valuable

insights from the classification process, thereby furnishing a means to identify the anticipated model "m" that engendered the input data.

5.4.3 *Early Fusion Methodology*

Similar to the first method, the second technique, known as "Early Fusion," involves merging two separate CNNs to create a single CNN with multiple inputs. This fusion is achieved by combining the final fully-connected layers of both networks and then adding three more fully-connected layers to generate the final predictive outcome. This process imparts the camera type classification with dimensionality cues, as delineated in Figure 5.6. Leveraging paired visual and audio patches, each instance of Early Fusion prognosticates the projected camera model predicated upon its respective estimation. The ultimate determination is rendered through the final fully connected layer, culminating in the computation of "yEF," which represents the score emanating from this terminal layer [199]. During the training phase, our methodology employs pairs of visual and audio patches as a cohesive mechanism to train the complete network. It is imperative to underscore that this differs from the Late Fusion approach, wherein there exists no discrete training procedure for the visual and audio branches. In a parallel vein, both the training and testing stages mirror those employed in the monomodal technique. However, a notable departure lies in the allocation of visual and audio patch pairs throughout the entire network, unlike before, where single patches were primarily used (mostly focusing on visual or audio content), the process shown in Figure 5.6 explains the workflow of the Early Fusion technique, summarized in a flowchart. Understanding the dimensions of input and output features related to the fully-connected layers is essential for designing this methodology. This architectural understanding helps structure the Early Fusion scheme systematically [200]. It's important to note that the output feature from the final network layer matches the size of M, representing the number of assessed camera models. This pivotal characteristic enhances the predictive capabilities of the network by encapsulating the potential camera models within the final output layer. As the training phase unfolds, the integrated utilization of visual and audio patch pairs instils a holistic perspective, fostering a comprehensive understanding of the multi-modal intricacies inherent in the data. The absence of disjoint training sequences, as witnessed in Late Fusion, underscores the seamless synergy between the visual and audio domains. In this cohesive framework, the amalgamation of information from both modalities imbues the network with a heightened capacity for discernment and prediction.

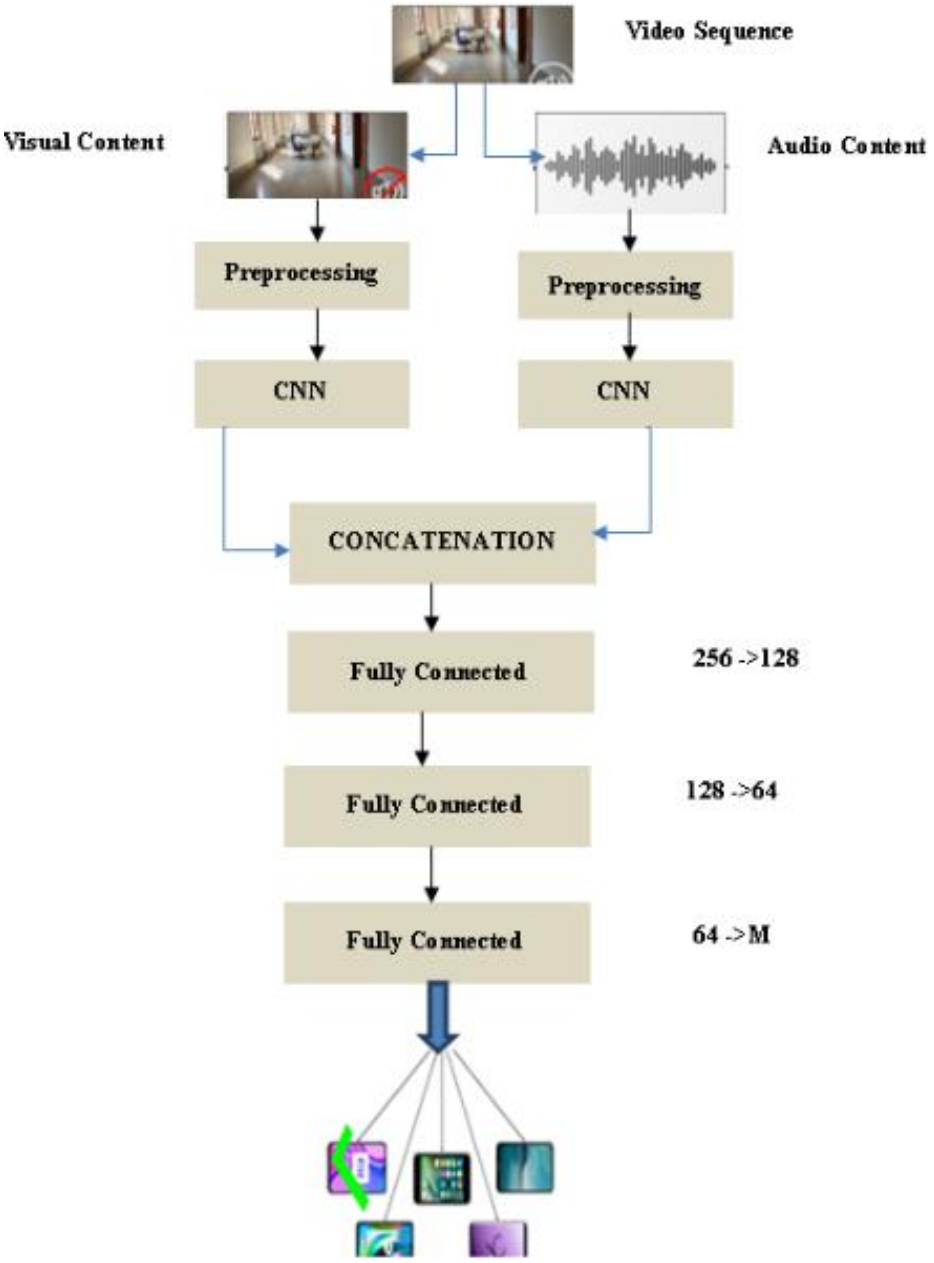


Figure 5.6. Pipeline of the Early Fusion Methodology

In the context of testing, the congruity with the monomodal technique remains palpable, thereby ensuring a cohesive and coherent evaluation process. While parallel to its monomodal counterpart, the Early Fusion methodology draws its potency from the integrated presentation of visual and audio patch pairs, ultimately augmenting its predictive prowess and enabling nuanced camera model identification.

5.4.4 *CNN Architectures*

Our approach centers on employing two distinct CNNs, namely EfficientNetB0 and VGGish, pivotal in addressing the specified issue. EfficientNetB0 occupies a pivotal position within

the avant-garde EfficientNet family of CNN models. Renowned for its exemplary performance in multimedia forensics tasks, it stands out as a highly promising candidate within this lineage. Our strategic selection of the EfficientNetB0 stems from its foundational nature, rendering it an ideal choice to serve as our cornerstone model. Its inherent simplicity allows for an extensive range of experimentation across various evaluation configurations. Notably, the expedited training phases facilitated by EfficientNetB0 offer an invaluable advantage, fostering iterative experimentation across diverse parameter settings. Crucially, preliminary investigations affirm the model's stability, corroborated by minimal deviations when scrutinizing parameters, thus underpinning its efficacy in comparison to computationally more intricate counterparts [126]. Furthermore, our methodology incorporates the VGGish CNN, an exemplar derived from the esteemed VGG network lineage, renowned for its efficacy in image classification. This strategic inclusion attests to the inherent adaptability of CNN architectures, adeptly extending image-centric principles to the realm of audio classification. By leveraging the strengths of VGGish, our framework augments its multi-modal capabilities, collectively harnessing visual and auditory cues for enhanced camera model identification. The dual utilization of EfficientNetB0 and VGGish engenders a potent synergy, orchestrating a comprehensive analysis that transcends single-modal paradigms. This tandem approach harnesses the prowess of each CNN to decipher the intricate nuances embedded within distinct video sequences, accentuating the proficiency of camera model identification. Through the orchestrated orchestration of EfficientNetB0 and VGGish, we orchestrate a nuanced understanding of the inherent attributes within diverse video streams, effectively bridging the domains of visual and auditory data. This integrated methodology fortifies our capacity to discern and distinguish camera models, propelling the boundaries of multi-modal analysis and deepening our comprehension of camera model identification. The realm of audio classification is enriched by the utilization of several CNNs, with the VGGish CNN standing as an eminent exemplar. This network draws inspiration from the well-established VGG networks renowned for their prowess in image classification. To address our specific challenge, we adopt a dual-CNN strategy, incorporating two distinct CNNs: EfficientNetB0 and VGGish. EfficientNetB0 holds a significant position within the newly introduced Efficient Net family of CNN models. As a member of this family, it is hailed for its outstanding performance within multimedia forensics tasks. Of noteworthy distinction, EfficientNetB0 emerges as a frontrunner among its counterparts, demonstrating remarkable capabilities in the realm of camera model identification. Situated within the cutting-edge Efficient Net model family, EfficientNetB0

showcases exceptional promise, particularly within the context of multimedia forensics, substantiated by its performance in diverse scenarios (Pandeya & Lee, 2021). Our comprehensive approach embraces both EfficientNetB0 and VGGish, synergizing their strengths to tackle the intricacies of camera model identification. By intertwining the capacities of these two CNNs, we forge a robust foundation for multi-modal analysis, advancing our understanding of camera model discernment through a confluence of visual and auditory cues. Our selection of the Efficient Net model stems from its fundamental nature, rendering it an ideal candidate for our research objectives. Its foundational simplicity grants us ample leeway to explore a spectrum of evaluation configurations, enabling a thorough investigation of our objectives. A notable advantage lies in the expeditiousness of its training phase, affording us a considerable temporal allowance for comprehensive experimentation. This temporal abundance facilitates the meticulous tuning of our model's performance, contributing to a heightened understanding of its capabilities. Notably, our preliminary experiments have already unveiled a pivotal insight: the absence of any substantial discernible divergence when employing varying parameters. Specifically, when evaluating EfficientNetB0's performance vis-à-vis computationally more intricate models characterized by an augmented parameter count, no significant discrepancies have emerged. This assertion, substantiated by empirical evidence, reinforces the efficiency and potency of our chosen Efficient Net model, debunking concerns of parameter-heavy models that entail greater computational demands [201]. Within the realm of audio classification, a diverse array of CNNs is harnessed for discerning auditory features. Among these, the VGGish CNN stands out, drawing its architectural inspiration from the renowned VGG network, initially tailored for image classification. Notably, the design of VGGish capitalizes on the proven efficacy of CNNs in audio classification tasks.

Upon traversing the dataset, the subsequent procedural imperative involves partitioning it into distinct training and validation sets. The training set serves as the foundation upon which the model's learning is cultivated, while the validation set serves as the crucible for assessing the trained model's performance. An effective strategy encompasses the extraction of frames from each video constituting both the training and validation sets. These frames, culled from the videos, lay the groundwork for subsequent analysis. The trajectory proceeds with preprocessing the extracted frames, culminating in their transformation into refined data representations. Subsequently, the training set of pre-processed frames is marshalled to train a specialized model, calibrated to glean nuanced insights from the audiovisual content. This strategic training phase imparts a model with the requisite cognitive machinery for discerning

intricate patterns and characteristics inherent within the audiovisual input. An intrinsic facet of this process resides in the pivotal role of the validation set. During the evaluation phase, the model's predictive capacity is scrutinized, with the frames extracted from the validation set serving as the input for these assessments. The model's efficacy and aptitude are gauged through the lens of these validation set frames, encapsulating its prowess in interpreting and classifying the audiovisual content. Upon achieving satisfactory performance benchmarks on the validation set, the trained model is poised to undertake the task of categorizing additional videos, thereby extending its utility. This pivotal step leverages the model's acquired cognitive capacity to discern and classify intricate audiovisual nuances within new video inputs. Figure 5.7 offers a visual representation of the processing flow within the spatial stream. The upper segment of the figure intricately delineates this trajectory. The classification CNN is meticulously designed to categorize visual content, resembling the structure of typical deep CNNs used for image classification. Each frame extracted from a video serves as input to this network. The architectural enhancements involve a sequence of convolutional layers, pooling layers, and fully connected (FC) layers. In this construct, frames undergo a sequential process. Initially, they pass through convolutional layers, instrumental in extracting complex features and patterns from the visual data. Subsequently, pooling layers down sample the information, retaining crucial features while reducing spatial dimensions. The network then integrates fully connected (FC) layers, facilitating comprehensive connections across the processed visual information. These FC layers allow for a holistic comprehension of features extracted from earlier stages, aiding in higher-level decision-making. This architecture aims to systematically process individual frames, extracting nuanced visual information through convolutional layers, distilling essential features via pooling layers, and finally, comprehensively understanding these features through fully connected layers. The network's design enables it to grasp intricate details and patterns within each frame, contributing to a comprehensive assessment of visual content in the classification process. These components collectively synergize to facilitate the intricate analysis and interpretation of the visual data ingrained within each frame. The convolutional layers meticulously detect salient features, the pooling layers condense and abstract these features, and the fully connected layers amalgamate these insights to make informed classification determinations. In essence, the framework orchestrates an intricate cascade of computational operations, which, guided by the model's training and prior learned insights, culminate in the accurate categorization of visual content encompassed within each video

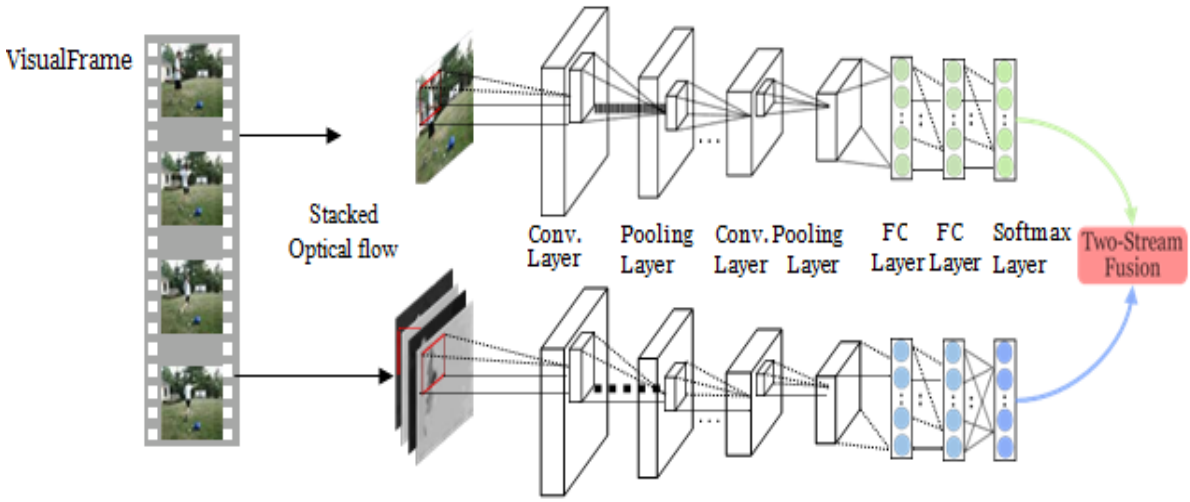


Figure 5.7. Processing Pipeline for the Extraction of Two-Stream Features from CNNs

frame. This procedural framework ultimately empowers the model to systematically categorize videos, affording a mechanism to extrapolate its acquired knowledge and proficiency to novel audiovisual inputs.

5.5 *Result Analysis*

Within this section, our focus initially lies on introducing the dataset and outlining the experimental framework that will underpin our investigations. This encompasses network training parameters and configurations essential for network training. Subsequently, we detail the chosen evaluation metrics and provide an insightful commentary on the achieved results.

Dataset: In this research, we utilize captured video frame patches obtained from the Vision dataset, a newly introduced collection of images and videos specifically designed for investigations in multimedia forensics. This dataset contains roughly 650 original videos from different time variant in HDR captured by 35 latest smartphones, DSLR, and their respective social media variants. The collection comprises approximately 2050 video with different length, each clearly identifying the device used to capture it. For our experimentation, we deliberately selected non-flat videos, depicting genuine scenarios with various objects, from both the unprocessed source (i.e., videos captured directly via smartphone camera without post-processing) and those subjected to compression through WhatsApp and YouTube platforms. In our pursuit of the desired analytical granularity, we aggregated videos from diverse devices belonging to the same model, facilitating comprehensive model-level analysis. Notably, videos sourced from devices D04, D12, and D17 are considered for evaluation, with the exclusion of D21 and D22 due to frame or audio

track extraction difficulties, as per the Vision dataset nomenclature. Furthermore, original videos not available in compressed form on WhatsApp or YouTube are omitted from our analysis. Diverging from prevalent video analysis services, our approach extends beyond the realm of high-resolution videos, encompassing a spectrum ranging from resolutions equal to or exceeding 720p down to 640x480. Our dataset comprises 1110 videos, each approximately one minute in duration, sourced from 25 distinct cameras. In order to assess how well our method classifies videos, we use the given details about the original camera model as the reference for each video sequence. To analyze the visual content, we extract 50 frames from every video sequence, evenly spread out across the entire duration. Each frame is divided into 10 randomly positioned patches, resulting in a total of NPv = 500 color patches per video. These patches are of dimension 256x256 pixels, a choice that yields favourable outcomes in our study. The collected dataset exhibits considerable diversity, encompassing varying camera models and resolutions, thereby enabling a comprehensive evaluation of the proposed technique's performance.

This inclusive dataset approach enhances the robustness of our methodology, as it is designed to handle a wide array of video scenarios encountered in real-world multimedia forensics tasks. By encompassing a broad range of resolutions and camera models, our approach accounts for real-world variations and challenges, bolstering the practical applicability and generalizability of our findings. It allows us to validate the effectiveness of the suggested technique under diverse conditions, ensuring its relevance across a spectrum of multimedia forensics scenarios. our dataset comprises 1110 videos with resolutions spanning from 720p to 640x480, recorded by 25 different cameras. Ground truth information about the source camera model is employed for classification assessment. We systematically extract visual content through a well-defined process of frame and patch selection. The resultant dataset demonstrates an appropriate diversity of scenarios, models, and resolutions, serving as a comprehensive foundation for evaluating and validating the proposed technique's performance and effectiveness. This thorough and inclusive approach enhances the practical utility and applicability of our methodology, enabling its potential deployment in real-world multimedia forensics applications. To ensure robustness and address potential concerns of overfitting and reduced prediction accuracy, we strategically devised a custom dataset for our feature extraction process, departing from the Kaggle dataset that comprises ten classes and 275 instances. Recognizing the limitations associated with the Kaggle dataset, we meticulously curated a new dataset with enhanced parameters, consisting of 1300 instances distributed across three distinct classes: iPhone 6s, Xiaomi Note 4x, and Samsung Galaxy J7.

Our decision to construct this new dataset was driven by the need to mitigate the aforementioned challenges and cultivate a dataset that aligns more closely with the objectives of our study. This strategic approach empowers us to curtail potential overfitting concerns and elevate the precision of predictive outcomes. Furthermore, our dataset enhancement strategy encompassed the introduction of two novel classes into the analytical framework. Inclusive of 275 samples each, these classes encompassed Samsung Galaxy Note 3 and HTC One M7 camera instances, thereby augmenting the dataset's comprehensiveness and diversity. The crux of our methodology hinges on feature extraction, a pivotal process that underpins the classification task. This involved the meticulous provision of our curated dataset to the model, enabling the extraction of salient features intrinsic to each camera model. These features, culled from the intricate nuances of the dataset, hold the key to discerning and characterizing camera models. By employing a feature-centric approach, we transcend the limitations of mere pixel-level analysis, delving deeper into the distinctive traits embedded within the data. The aggregation and analysis of these features facilitate a more profound understanding of the unique characteristics exhibited by each camera model, enabling accurate classification based on a comprehensive array of discriminating attributes. Upon extracting and analyzing the distinctive features of the camera models, our classification process ensued. These features, which encapsulate an array of intricate details and patterns, are harnessed to systematically categorize the camera models. The classification process harnesses the amalgamation of features to make informed and precise determinations, underpinning the core predictive capability of our proposed model. we judiciously transitioned from the Kaggle dataset, steering clear of its potential limitations, and painstakingly curated a novel dataset with heightened attributes and strategic class composition. This dataset refinement was instrumental in surmounting issues such as overfitting and accuracy diminishment. The subsequent feature extraction process facilitated a profound analysis of camera model characteristics, furnishing the groundwork for a robust classification mechanism. By anchoring our classification on these distinctive features, we establish a principled and technologically sophisticated methodology for camera model categorization, offering enhanced accuracy and predictive prowess.

Table 2 presents the average error rate and the standard deviation of the confidence score related to the patch dataset's test split. It also illustrates different values of a crucial variable, which we refer to as. These parameter values have been systematically explored to gauge their impact on the susceptibility to adversarial instances, particularly instances where FGSM perturbations yield negligible visual alterations in the context of untargeted attacks. The

Table 5.1: Details of the dataset.

Device Model Name	Number of Instances	Captured From
IPhone 6s	**1500**	**self**
Xiaomi Note 4x	**1560**	**self**
Samsung Galaxy j7	**1600**	**self**
Samsung Galaxy Note 3	**1000**	**Kaggle**
HTC One M7	**550**	**Kaggle**

investigation unfolds within the realm of the patch test split, yielding noteworthy insights. Among the array of evaluated values, it becomes evident that a value of = 0.005 engenders an optimal trade-off between the error rate and perceptible alterations in the image domain. This strategic selection stems from a meticulous balance achieved between the discriminative prowess of the trained DenseNet model detector and the observable changes in the manipulated image.

The performance metrics substantiate this strategic choice. When exposed to examples created with the determined ideal value of, the improved DenseNet model achieves an average accuracy rate of 93.1 percent. Simultaneously, the model's maximum trust rating is at a remarkable 95.3 percent. This robust performance manifests the detector's aptitude in effectively categorizing instances while maintaining a high degree of certainty in its predictions. An interesting observation emerges as the value of is modulated. It becomes apparent that the perceptibility of manipulations is intrinsically linked to the magnitude of. As this parameter escalates, the visual impact of manipulations becomes progressively pronounced. This nuanced relationship underscores the intricate interplay between model sensitivity and the detectability of adversarial interventions.

In summation, the exploration of diverse values for within the framework of the patch dataset test split furnishes crucial insights into the trade-off between error rates and perceptual alterations. The judicious selection of = 0.005 emerges as a pivotal decision point, offering an optimal equilibrium between classification accuracy and the visual fidelity of manipulated instances. The ensuing performance of the trained DenseNet model is marked by a commendable average error rate of 93.1 percent, complemented by a robust average confidence level of 95.3 percent. Moreover, the delineation of the relationship between and the perceptibility of manipulations accentuates the dynamic nature of the detector's responsiveness and its consequential impact on adversarial instance detection.

Table 5.2: The error rate and confidence score of the DenseNet model.

ϱ Value	Error Rate (%)	Confidence Score (%)
0.01	97.3	97.8
0.02	94.8	91.0
0.03	92.6	93.9
0.04	93.7	92.8
0.05	98.4	94.8
0.06	96.7	98.6
0.07	91.5	99.4
0.08	90.6	97.1
0.09	92.0	92.0
0.11	91.4	91.2

The table demonstrates the error rate and confidence score resulting from an untargeted FGSM attack on the test partition, evaluating the performance of our trained DenseNet model. The provided chart, which is displayed in Figure 5.8, compares the suggested methodology with alternative approaches.

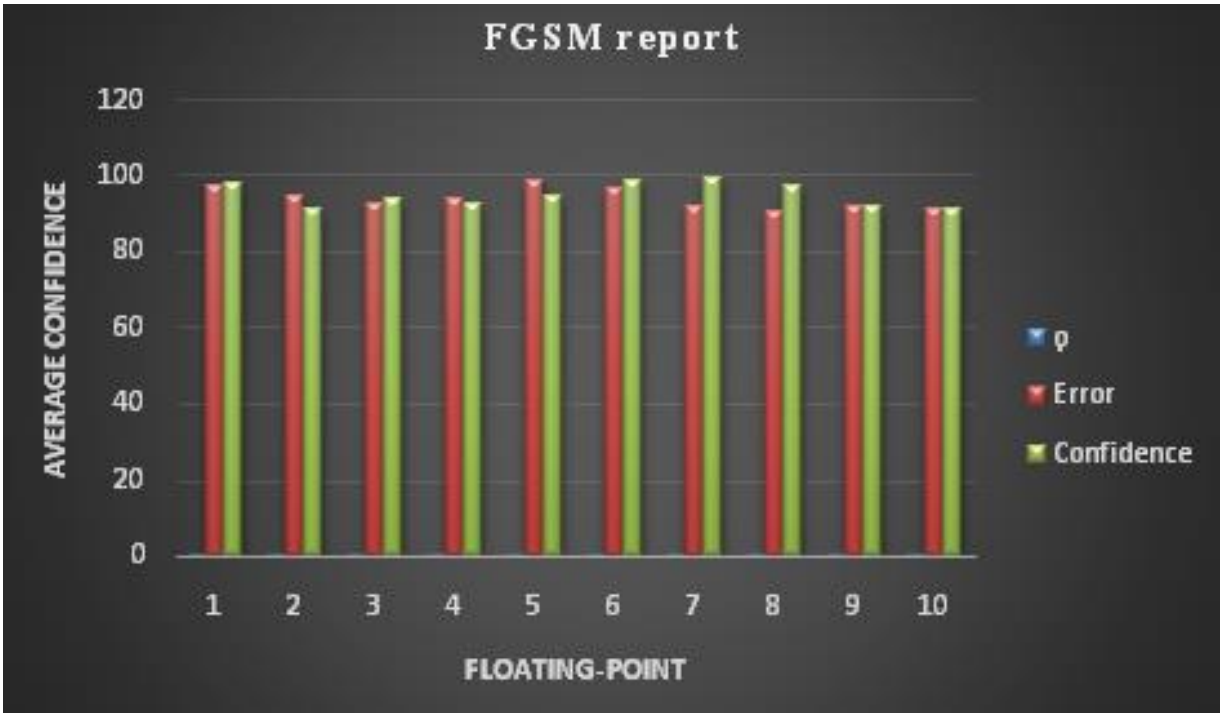


Figure 5.8. Comparing the Suggested Approach with Alternative Approaches

The additional experiment was limited to using the second set of attributes and involved CFA interpolation. The evaluation yielded a precision of 86.93%. Although this outcome is considered acceptable, it falls short of the results achieved in the first experiment that relied solely on co-occurrences. To further enhance accuracy and achieve an average of 97.81%, a fusion approach was employed by combining both feature sets and applying them in tandem. This amalgamated approach resulted in an impressive average accuracy of 98.75% across all three feature sets. Table 5.3 comprehensively illustrates the outcomes of the aforementioned experiments, providing a detailed overview of their respective accuracy rates.

Table 5.3: illustrates classification accuracy derived from the VISION dataset.

Model	***N***	**Constraint Type**	**Overall**	**Flat**	**Indoor**	**Outdoor**	**WA**	**YT**	**NA**
ResNet50	60	Conv	55.20	64.81	50.74	41.71	55.10	51.60	62.80
ResNet50	60	Conv	55.20	64.81	50.74	41.71	55.10	51.60	62.80
MobileNet	60	None	71.57	**85.32**	62.87	**75.45**	**78.66**	67.96	71.66
MobileNet	60	Conv	56.18	64.74	47.21	56.51	53.60	46.20	53.00
MobileNet	60	PRNU	62.70	63.96	53.11	61.12	58.80	63.50	67.30
MobileNet	60	None	**75.87**	76.92	**64.62**	75.02	74.84	**77.68**	**75.90**
MobileNet	60	PRNU	61.74	65.96	54.14	67.14	57.81	65.54	68.31

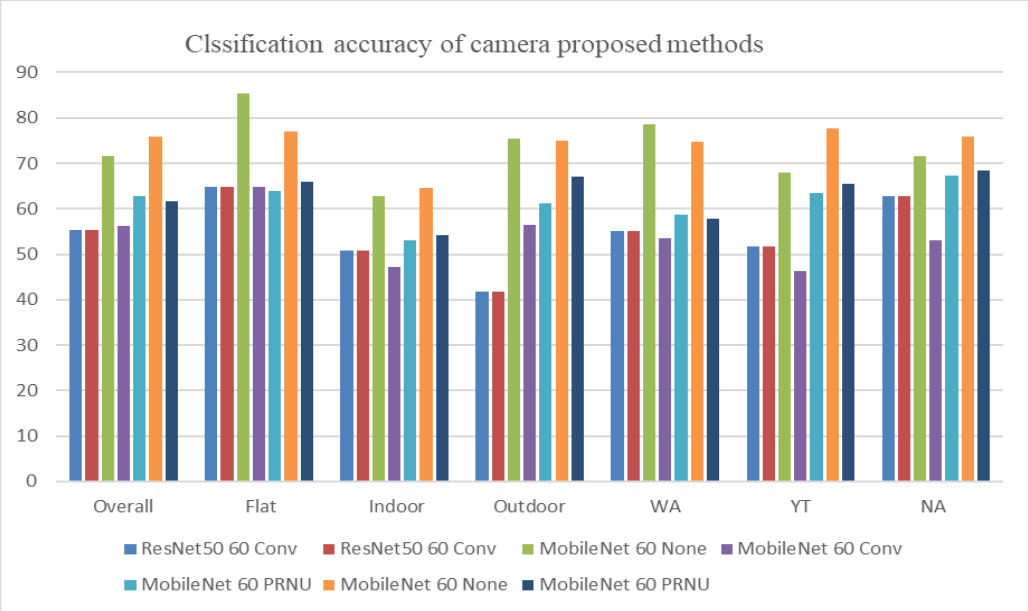


Figure 5.9. Classification Accuracy of Camera Proposed Methods

The table that follows gives a thorough summary of each Convolutional Neural Network's (ConvNet) particular test accuracy as well as the total test efficiency for each of the three different settings—flat, indoor, and outdoor. Furthermore, three types of compression are included in the evaluation: native (NA), WhatsApp (WA), and YouTube (YT). N I-frames per movie were used in the testing as well as training stages of a standard procedure that produced consistent outcomes. It is noteworthy that the obtained results align with those of parallel tests conducted, confirming the reliability of the PRNU-based methodology. Across all tested scenarios and compression types, the accuracy achieved through PRNU analysis significantly outperforms the accuracy observed with limited counterparts. The accuracy measurements are underpinned by a rigorous evaluation process conducted on the VISION dataset. This dataset provides a solid basis for evaluating the effectiveness and efficacy of the ConvNets in different compression circumstances and parameters.

Table 5.4: compares the accuracy of MobileNet when it is compared to different counts of I-frames per video (I-fpv).

I-fpv	Overall	Flat	Indoor	Outdoor
1	69.12	71.1	57.5	76.5
5	72.31	79.8	59.6	75.4
30	74.10	82.1	62.3	76.0
50	73.51	81.5	61.6	75.4
100	73.71	82.1	61.6	75.4
All	73.71	82.1	61.6	75.4
1	69.12	71.1	57.5	76.5

In order to establish a comprehensive comparative analysis, an analogous experiment was undertaken utilizing the I-frames methodology. The outcomes of this experiment have been meticulously documented in Table 4. The results of this study demonstrate that the model may get a remarkably high degree of accuracy even with a restricted number of I-frames used in testing.

It is worth noting that the VISION dataset, central to this study, comprises movies of relatively short duration. Consequently, the pool of available I-frames is inherently constrained. Attempts to augment the number of extracted I-frames do not yield a commensurate increase in accuracy, owing to the inherent limitations posed by the dataset's compressed temporal scope.

The implications of these results are twofold. Firstly, they underscore the model's inherent resilience and capacity to deliver consistent accuracy levels, even in scenarios characterized by limited data points such as I-frames. Secondly, the study highlights the influence of dataset characteristics, with the temporal brevity of the VISION dataset movies contributing to the observed phenomenon of accuracy stabilization despite efforts to expand the I-frame extraction process.In essence, the study underscores the interplay between the model's robustness and the data constraints presented by the dataset's temporal attributes. This insight is of paramount importance in both refining the model's application and in comprehending the intricacies of video-based analysis within the context of limited temporal information.

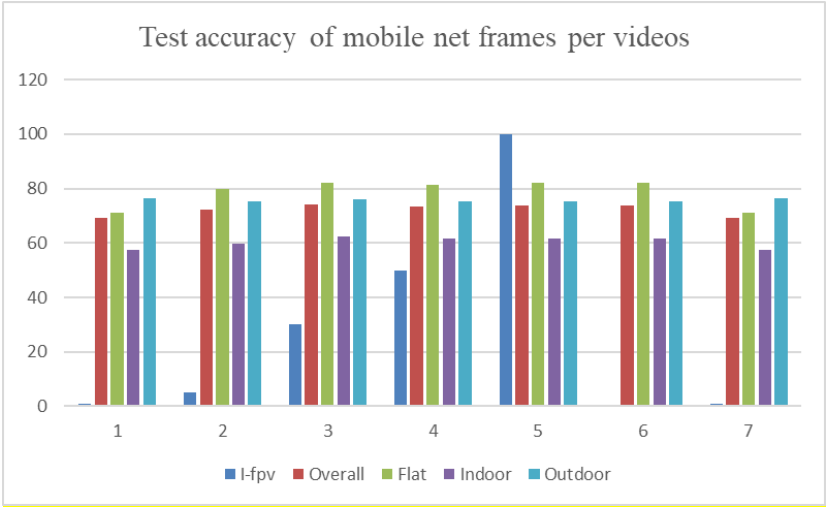


Figure 5.10. Fig: Test Accuracy of Mobile,net Frames per Videos

Drawing from our accumulated expertise, we firmly assert that the most efficacious overarching approach involves the integration of the Late Fusion technique, coupled with a judicious configuration of the EE192 model based on our experiential insights. In scrutinizing both native and YouTube video sequences, this composite methodology consistently delivers the most accurate outcomes across various testing paradigms, regardless of the presence of cross-tests or the distinction between non-cross and cross-tests. Interestingly, the cross-test results, encompassing WhatsApp data, align closely with the performance of the alternative configurations, if not exhibiting a slight diminution. This phenomenon can be attributed to the remarkable adaptability of the trained CNNs within this setup to the specific training data they encounter. These CNNs exhibit a nuanced sensitivity, rendering them less versatile and more susceptible to the pronounced data compression characteristic of platforms like WhatsApp, which, in turn, elucidates the comparatively suboptimal performance observed.

In essence, this observation underscores the intricate interplay between the network's adaptability to diverse training data and its susceptibility to compression-induced variations.

By aligning Late Fusion methodology with EE192 configuration based on these insights, our approach demonstrates superior versatility and robustness, while also shedding light on the nuanced dynamics that govern performance fluctuations across distinct compression contexts.

5.6 *Conclusions and Future Works*

This study introduces an innovative multi-modal approach aimed at discerning closed-set camera models within digital video sequences. This research uses both auditory and visual data that is taken from the video itself in order to pinpoint the exact smartphone model that was used to film a particular video. The suggested approach uses Convolutional Neural Networks (CNNs) to categorize the video according to its audio and visual characteristics. Patches from the video frames are used to create the visual portion, and patches from the audio track's Log-Mel Spectrogram are used to create the sound portion. The main objective is to use the Late Fusion technique to classify the query video by integrating the results of two distinct networks that are specialized in auditory and visual analysis. One person handles aural patches, while the other processes visual patches. After that, these combined scores are sent into a multi-input network that uses paired aural and visual patches fetched from the query video. This holistic approach integrates diverse modalities of information within a singular framework, fostering a comprehensive analysis of video sequences. The combination of visual and auditory cues, harnessed through state-of-the-art CNN architecture, empowers the system to effectively delineate nuances intrinsic to various smartphone models' characteristic capture patterns. By employing the Late Fusion technique, the system capitalizes on the strengths of individual mono-modal networks and orchestrates their outputs to achieve enhanced accuracy and discriminative power. This sophisticated strategy facilitates the identification process, resulting in a more refined and accurate determination of the source camera model for the given query video. This work aims to transform the field of electronic video clips camera model identification. Using the creative combination of visual and auditory data, as well as the calculated application of the Late Fusion technique, the suggested methodology enhances the precision as well as the precision of camera model categorization, hence expanding the potential applications of investigative multimedia processing.

A single multi-input system is used by the Early Fusion technique to process input from both visual and auditory patch pairs retrieved from the query video. The fact that both of these tactics are multi-modal techniques for camera model identification must be emphasized.

Using a range of designs and data pre-processing methods, our study aims to investigate three different topologies for each of these approaches. We use video clips from the Vision dataset to assess the effectiveness of our experimental efforts. The scope of our assessment extends beyond solely original native videos captured directly by smartphone cameras. We intentionally include diverse video sources to explore an array of training and testing configurations, aiming to simulate real-world scenarios necessitating the categorization of data subjected to internet-based compression services. In pursuit of these objectives, we also incorporate videos that have undergone compression through algorithms employed by platforms such as WhatsApp and YouTube (commonly used for social media and uploading purposes). This multifaceted approach enables us to gauge the performance and robustness of our proposed multi-modal attribution strategy under varying conditions, thereby reflecting the intricacies encountered in practical contexts. Additionally, our investigation involves a comparative analysis between the multi-modal attribution strategy we introduce and conventional mono-modal attribution methods, along with other suggested techniques. By subjecting our proposed strategy to rigorous evaluation against established benchmarks and alternative methodologies, we strive to ascertain its distinct advantages and contributions within the realm of camera model identification. Our study employs a systematic approach, utilizing both Early Fusion and Late Fusion methods as multi-modal avenues for camera model identification. Through comprehensive experimentation involving different topologies, architectural variations, and real-world data scenarios, we seek to validate the efficacy and versatility of our proposed strategy, thereby enriching the discourse on multimedia forensics and advancing the field's analytical capabilities. On a comprehensive analysis, the Late Fusion methodology emerges as the frontrunner among diverse multi-modal approaches, notably surpassing the conventional mono-modal counterparts. Empirical evidence consistently substantiates the superiority of multi-modal methodologies over mono-modal ones. Notably, the Late Fusion technique achieves outstanding performance levels, exceeding a 99 percent accuracy threshold in distinguishing original video sequences from YouTube counterparts.

However, it's noteworthy that a fraction of videos, albeit small, poses challenges for accurate modeling, primarily attributed to the pronounced compression characteristic of WhatsApp. This intriguing observation suggests the potential emergence of novel challenges and avenues for advancement, particularly in the domain of identifying originating camera models for videos that find widespread dissemination through social media platforms. The prevalence of extreme compression in WhatsApp raises intriguing questions that warrant further

exploration to comprehend the underlying factors contributing to this challenge. Furthermore, a noteworthy aspect is the adaptability of the proposed multi-modal solutions to broader scenarios encompassing more than two data modalities. The framework's flexibility is readily amenable to extension, accommodating additional data sources seamlessly. This adaptability envisions future scenarios where the complexity of data may demand integration across multiple modalities, reinforcing the robustness and versatility of the Late Fusion approach. It's also pertinent to acknowledge that the adoption of the Late Fusion approach offers a streamlined training paradigm. By training the Convolutional Neural Networks (CNNs) independently for each target within the multi-modal architecture, the complexity of the learning process is effectively compartmentalized. This modular training scheme facilitates efficient handling of varying data modalities, enhancing the agility and scalability of the approach in scenarios requiring expansive data integration.

In summation, the Late Fusion technique emerges as a potent contender in the realm of multi-modal methodologies, showcasing remarkable performance advantages over traditional mono-modal strategies. Its remarkable accuracy in distinguishing video origins, even in challenging scenarios, highlights its potential relevance in contemporary multimedia forensics. Moreover, its inherent adaptability and modularity position it favorably for future explorations into more intricate multi-modal scenarios, underscoring its capacity to serve as a foundational framework for advancing camera model identification and related endeavors.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 *Conclusion*

In this study, we delve into the realm of image source identification, presenting a pioneering deep convolutional neural network (CNN) as a robust solution to achieve commendable testing and learning performance. The overarching goal is to enhance the accuracy of image forgery detection, addressing the growing challenges posed by manipulated and altered images in the contemporary real-world scenario. The proposed CNN model undergoes several modifications in its architecture to improve the detection rate, ensuring its adaptability and efficacy across various types of altered images. This iterative process of refinement is crucial in developing a model that can reliably discern the source of images in a dynamic and ever-evolving digital landscape. Our approach is specifically tailored to function efficiently in the current real-world scenario, where the prevalence of altered images presents a significant obstacle to conventional image forensics techniques. By subjecting the proposed CNN model to rigorous modifications, we aim to fortify its ability to identify the primary source of images amidst the complexities introduced by different types of image alterations.

The outcomes of our study reveal that the suggested multi-modal methods significantly outperform traditional mono-modal methods in the context of image forgery detection. The integration of multiple modes of information and features proves to be a pivotal strategy in enhancing the model's overall performance. This underscores the importance of leveraging diverse sources of data and signals to improve the robustness and reliability of image source identification systems. A noteworthy aspect of our research is the exploration of fusion techniques within the realm of multi-modal methodologies. The fusion technique emerges as a potent contender, showcasing remarkable performance advantages over traditional mono-modal strategies. By combining information from various sources and modalities, the fusion technique demonstrates an enhanced ability to accurately identify the source of altered images, even in scenarios where traditional methods may fall short.

The significance of our findings extends beyond the realm of image forensics. It underscores the broader potential of multi-modal approaches and fusion techniques in addressing complex challenges across various domains. The success of our proposed CNN model and fusion technique not only contributes to the advancement of image source identification but also opens avenues for innovative solutions in related fields, such as computer vision and artificial intelligence.

6.2 *Future Scope*

Looking ahead, our study paves the way for continued research and refinement in the field of image forgery detection. The dynamic nature of digital manipulation calls for ongoing efforts to adapt and improve detection methodologies. We anticipate that our multi-modal approach and fusion technique will inspire further innovations, leading to more robust and reliable solutions for identifying the source of images in an increasingly complex digital landscape. This commitment to advancement reflects our dedication to staying at the forefront of technological developments and addressing the evolving challenges in the realm of image forensics.

References

[1] Wei Lu, Wei Sun, Ji-Wu Huang and Hong-Tao Lu, "Digital image forensics using statistical features and neural network classifier," *2008 International Conference on Machine Learning and Cybernetics, Kunming*, 2008, pp. 2831-2834, doi: 10.1109/ICMLC.2008.4620890.

[2] J. A. Redi, W. Taktak, and J. L. Dugelay, "Digital image forensics: A booklet for beginners," *Multimed Tools Appl*, vol. 51, no. 1, pp. 133–162, Jan. 2011, doi: 10.1007/s11042-010-0620-1.

[3] A. Rocha, W. Scheirer, T. Boult, and S. Goldenstein, "Vision of the unseen: Current trends and challenges in digital image and video forensics," *ACM Comput Surv*, vol. 43, no. 4, Oct. 2011, doi: 10.1145/1978802.1978805.

[4] B. Hosler et al., "A Video Camera Model Identification System Using Deep Learning and Fusion," ICASSP 2019 - *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 8271-8275, doi: 10.1109/ICASSP.2019.8682608.

[5] H. Farid, "Image forgery detection," *IEEE Signal Process Mag*, vol. 26, no. 2, pp. 16–25, Mar. 2009, doi: 10.1109/MSP.2008.931079.

[6] G. K. Birajdar and V. H. Mankar, "Digital image forgery detection using passive techniques: A survey," *Digit Investig*, vol. 10, no. 3, pp. 226–245, Oct. 2013, doi: 10.1016/j.diin.2013.04.007.

[7] A. C. Kot and H. Cao, "Image and Video Source Class Identification," in *Digital Image Forensics*, New York, NY: Springer New York, 2013, pp. 157–178. doi: 10.1007/978-1-4614-0757-7_5.

[8] A. C. Popescu and H. Farid, "Exposing digital forgeries in color filter array interpolated images," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3948–3959, Oct. 2005, doi: 10.1109/TSP.2005.855406.

[9] C. Meij and Z. Geradts, "Source camera identification using Photo Response Non-Uniformity on WhatsApp," *Digit Investigation*, vol. 24, pp. 142–154, Mar. 2018, doi: 10.1016/j.diin.2018.02.005.

[10] S. S. Ali, I. I. Ganapathi, N.-S. Vu, S. D. Ali, N. Saxena, and N. Werghi, "Image Forgery Detection Using Deep Learning by Recompressing Images," *Electronics (Basel)*, vol. 11, no. 3, p. 403, Jan. 2022, doi: 10.3390/electronics11030403.

[11] E. A. Armas Vega, E. Gonzalez Fernandez, A. L. Sandoval Orozco, and L. J. Garcia Villalba, "Passive Image Forgery Detection Based on the Demosaicing Algorithm and JPEG Compression," *IEEE Access*, vol. 8, pp. 11815–11823, 2020, doi: 10.1109/ACCESS.2020.2964516.

[12] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "A Full-Image Full-Resolution End-to-End-Trainable CNN Framework for Image Forgery Detection," *IEEE Access*, vol. 8, pp.

133488–133502, 2020, doi: 10.1109/ACCESS.2020.3009877.

[13] K. B. Meena and V. Tyagi, "Image Forgery Detection: Survey and Future Directions," in *Data, Engineering and Applications*, Singapore: Springer Singapore, 2019, pp. 163–194. doi: 10.1007/978-981-13-6351-1_14.

[14] O. Ismael Al-Sanjary, G. Sulong, and O. Ismael Al-sanjary, "Detection of video forgery: A review of literature HIGH Al-Sanjary, O.I., & Sulong, G. (2015). Detection Of Video Forgery: A Review of Literature. Journal of theoretical and applied information technology, 74, 207-220.

[15] M. Barni *et al.*, "Aligned and Non-Aligned Double JPEG Detection Using Convolutional Neural Networks," Aug. 2017, doi: 10.1016/j.jvcir.2017.09.003.

[16] C. Chen, X. Zhao, and M. C. Stamm, "Generative Adversarial Attacks Against Deep-Learning-Based Camera Model Identification," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2022, doi: 10.1109/TIFS.2019.2945198.

[17] B. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, New York, NY, USA: ACM, Jun. 2016, pp. 5–10. doi: 10.1145/2909827.2930786.

[18] A. Rocha and W. Scheirer, "Vision of the Unseen: Current Trends and Challenges in Digital Image and Video Forensics." [Online]. Available: http://www.bradley.edu/exhibit96/about/twoways.html

[19] R. L. Easton, "Fundamentals of Digital Image Processing," 2010.

[20] R. Rodrigo, "Introduction to Digital Image Processing," 2011.

[21] T. Van Lanh, K. -S. Chong, S. Emmanuel and M. S. Kankanhalli, "A Survey on Digital Camera Image Forensic Methods," *2007 IEEE International Conference on Multimedia and Expo, Beijing*, China, 2007, pp. 16-19, doi: 10.1109/ICME.2007.4284575.

[22] McKay, C., Swaminathan, A., Gou, H., & Wu, M. (2008). Image acquisition forensics: Forensic analysis to identify imaging source. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1657-1660.

[23] A. T. Ho and S. Li, "Handbook of Digital Forensics of Multimedia Data and Devices," 2015. [Online]. Available: http://forensics.inf.tu-dresden.de/ddimgdb/locations

[24] Physioc, L. W. (1932). Photographic Emulsions. *Journal of the Society of Motion Picture Engineers*, 19(1), 913-920.

[25] W. Lefèvre, "Exposing the seventeenth-century optical camera obscura," *Endeavour*, vol. 31, no. 2. pp. 54–58, Jun. 2007. doi: 10.1016/j.endeavour.2007.05.004.

[26] M. Guarnieri, "An Historical Survey on Light Technologies," *IEEE Access*, vol. 6. Institute of Electrical and Electronics Engineers Inc., pp. 25881–25897, May 08, 2018. doi: 10.1109/ACCESS.2018.2834432.

[27] Wang, D. C., Vagnucci, A. H., & Li, C. C. (1983). Digital image enhancement: a survey. Computer vision, graphics, and image processing, 24(3), 363-381.

[28] D. H. Rao and P. P. Panduranga, "A Survey on Image Enhancement Techniques: Classical Spatial Filter, Neural Network, Cellular Neural Network, and Fuzzy Filter," *2006 IEEE International Conference on Industrial Technology*, Mumbai, India, 2006, pp. 2821-2826, doi: 10.1109/ICIT.2006.372671.

[29] P. Suganya, S. Gayathri, and N. Mohanapriya, "Survey on Image Enhancement Techniques," *International Journal of Computer Applications Technology and Research*, pp. 623–627, Sep. 2013, doi: 10.7753/ijcatr0205.1019.

[30] K. Sharumathi and R. Priyadharsini, "A survey on various image enhancement techniques for underwater acoustic images," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, Mar. 2016, pp. 2930–2933. doi: 10.1109/ICEEOT.2016.7755235.

[31] A. N. Aimi Salihah, M. Y. Mashor, N. H. Harun, A. A. Abdullah, and H. Rosline, "Improving colour image segmentation on acute myelogenous leukaemia images using contrast enhancement techniques," in *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, IEEE, Nov. 2010, pp. 246–251. doi: 10.1109/IECBES.2010.5742237.

[32] M. M. Ahmed, J. M. Zain, and M. M. Ahmed, "A Study on the Validation of Histogram Equalization as a Contrast Enhancement Technique," in *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, IEEE, Nov. 2012, pp. 253–256. doi: 10.1109/ACSAT.2012.82.

[33] G. Cao, Y. Zhao, R. Ni, and X. Li, "Contrast Enhancement-Based Forensics in Digital Images," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 515–525, Mar. 2014, doi: 10.1109/TIFS.2014.2300937.

[34] T. Celik, "Spatial Entropy-Based Global and Local Image Contrast Enhancement," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5298–5308, Dec. 2014, doi: 10.1109/TIP.2014.2364537.

[35] H. D. Cheng and Y. Zhang, "Detecting of contrast over-enhancement," in *2012 19th IEEE International Conference on Image Processing*, IEEE, Sep. 2012, pp. 961–964. doi: 10.1109/ICIP.2012.6467021.

[36] X. Chen and L. Lv, "A Compositive Contrast Enhancement Algorithm of IR Image," in *2013*

International Conference on Information Technology and Applications, IEEE, Nov. 2013, pp. 58–62. doi: 10.1109/ITA.2013.20.

[37] C. E. Chang *et al.*, "A hardware-oriented contrast enhancement algorithm for real-time applications," in *Proceedings - 2018 1st International Cognitive Cities Conference, IC3 2018*, Institute of Electrical and Electronics Engineers Inc., Dec. 2018, pp. 183–185. doi: 10.1109/IC3.2018.00-32.

[38] R. K. Jha, R. Chouhan, K. Aizawa, and P. K. Biswas, "Dark and low-contrast image enhancement using dynamic stochastic resonance in discrete cosine transform domain," *APSIPA Trans Signal Inf Process*, vol. 2, 2013, doi: 10.1017/ATSIP.2013.7.

[39] T. H. Kil, S. H. Lee, and N. I. Cho, "A dehazing algorithm using dark channel prior and contrast enhancement," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, May 2013, pp. 2484–2487. doi: 10.1109/ICASSP.2013.6638102.

[40] G. Raju and M. S. Nair, "A fast and efficient color image enhancement method based on fuzzy-logic and histogram," *AEU - International Journal of Electronics and Communications*, vol. 68, no. 3, pp. 237–243, Mar. 2014, doi: 10.1016/j.aeue.2013.08.015.

[41] G. Maragatham and † S Md Mansoor Roomi, "An Automatic Contrast Enhancement method based on Stochastic Resonance."

[42] S. C. Nercessian, K. A. Panetta, and S. S. Agaian, "Non-linear direct multi-scale image enhancement based on the luminance and contrast masking characteristics of the human visual system," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3549–3561, Sep. 2013, doi: 10.1109/TIP.2013.2262287.

[43] C. Reshmalakshmi and M. Sasikumar, "Image contrast enhancement using fuzzy technique," in *2013 International Conference on Circuits, Power and Computing Technologies (ICCPCT)*, IEEE, Mar. 2013, pp. 861–865. doi: 10.1109/ICCPCT.2013.6528836.

[44] K. N. Shukla, "A Review on Image Enhancement Techniques," *International Journal of Engineering and Applied Computer Science*, vol. 02, no. 07, pp. 232–235, Aug. 2017, doi: 10.24032/ijeacs/0207/05.

[45] M. Liang and X. Hu, "Feature selection in supervised saliency prediction," *IEEE Trans Cybern*, vol. 45, no. 5, pp. 900–912, May 2015, doi: 10.1109/TCYB.2014.2338893.

[46] C. Amiot, C. Girard, J. Chanussot, J. Pescatore, and M. Desvignes, "Curvelet based contrast enhancement in fluoroscopic sequences," *IEEE Trans Med Imaging*, vol. 34, no. 1, pp. 137–147, Jan. 2015, doi: 10.1109/TMI.2014.2349034.

[47] X. Guo, Z. Yu, S. B. Kang, H. Lin, and J. Yu, "Enhancing light fields through ray-space stitching," *IEEE Trans Vis Comput Graph*, vol. 22, no. 7, pp. 1852–1861, Jul. 2016, doi:

10.1109/TVCG.2015.2476805.

[48] C. Jung and T. Sun, "Optimized Perceptual Tone Mapping for Contrast Enhancement of Images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1161–1170, Jun. 2017, doi: 10.1109/TCSVT.2016.2527339.

[49] L. Huang, W. Zhao, B. R. Abidi, and M. A. Abidi, "A Constrained Optimization Approach for Image Gradient Enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1707–1718, Aug. 2018, doi: 10.1109/TCSVT.2017.2696971.

[50] R. Liu, L. Ma, Y. Wang, and L. Zhang, "Learning converged propagations with deep prior ensemble for image enhancement," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1528–1543, Mar. 2019, doi: 10.1109/TIP.2018.2875568.

[51] M. Liu, Z. Zhou, P. Shang, and D. Xu, "Fuzzified Image Enhancement for Deep Learning in Iris Recognition," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 1, pp. 92–99, Jan. 2020, doi: 10.1109/TFUZZ.2019.2912576.

[52] S. B. Yoo and E. J. Lee, "Deep Super-Resolution Imaging Technology: Toward Optical Super-Vision," *IEEE Consumer Electronics Magazine*, vol. 10, no. 1, pp. 24–31, Jan. 2021, doi: 10.1109/MCE.2020.2988442.

[53] T. C. Tung and C. S. Fuh, "ICEBIN: Image Contrast Enhancement Based on Induced Norm and Local Patch Approaches," *IEEE Access*, vol. 9, pp. 23737–23750, 2021, doi: 10.1109/ACCESS.2021.3056244.

[54] Y. Zhang, X. Di, B. Zhang, R. Ji, and C. Wang, "Better Than Reference in Low-Light Image Enhancement: Conditional Re-Enhancement Network," *IEEE Transactions on Image Processing*, vol. 31, pp. 759–772, 2022, doi: 10.1109/TIP.2021.3135473.

[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.01497

[56] N. Dalal, B. T. Histograms, and B. Triggs, "Histograms of Oriented Gradients for Human Detection," pp. 886–893, 2005, doi: 10.1109/CVPR.2005.177i.

[57] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-Shot Refinement Neural Network for Object Detection," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1711.06897

[58] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[59] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.02640

[60] R. A. Ciora and C. M. Simion, "Industrial Applications of Image Processing," *ACTA*

Universitatis Cibiniensis, vol. 64, no. 1, pp. 17–21, Nov. 2014, doi: 10.2478/aucts-2014-0004.

[61] M. Kirchner and T. Gloe, "Forensic Camera Model Identification," in *Handbook of Digital Forensics of Multimedia Data and Devices*, Chichester, UK: John Wiley & Sons, Ltd, 2015, pp. 329–374. doi: 10.1002/9781118705773.ch9.

[62] J. A. Redi, W. Taktak, and J. L. Dugelay, "Digital image forensics: A booklet for beginners," *Multimed Tools Appl*, vol. 51, no. 1, pp. 133–162, Jan. 2011, doi: 10.1007/s11042-010-0620-1.

[63] J. Luka, J. Fridrich, and M. Goljan, "Digital Camera Identification From Sensor Pattern Noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, Jun. 2006, doi: 10.1109/TIFS.2006.873602.

[64] B. Wang, J. Yin, S. Tan, Y. Li, and M. Li, "Source camera model identification based on convolutional neural networks with local binary patterns coding," *Signal Process Image Commun*, vol. 68, pp. 162–168, Oct. 2018, doi: 10.1016/j.image.2018.08.001.

[65] T. Van Lanh, K.-S. Chong, S. Emmanuel, and M. S. Kankanhalli, "A Survey on Digital Camera Image Forensic Methods," in *Multimedia and Expo, 2007 IEEE International Conference on*, IEEE, Jul. 2007, pp. 16–19. doi: 10.1109/ICME.2007.4284575.

[66] M. R. Bai and J.-H. Lin, "Source identification system based on the time-domain nearfield equivalence source imaging: Fundamental theory and implementation," *J Sound Vib*, vol. 307, no. 1–2, pp. 202–225, Oct. 2007, doi: 10.1016/j.jsv.2007.06.025.

[67] B. Gupta and M. Tiwari, "Improving source camera identification performance using DCT based image frequency components dependent sensor pattern noise extraction method," *Digit Investig*, vol. 24, pp. 121–127, Mar. 2018, doi: 10.1016/j.diin.2018.02.003.

[68] B. Xu, X. Wang, X. Zhou, J. Xi, and S. Wang, "Source camera identification from image texture features," *Neurocomputing*, vol. 207, pp. 131–140, Sep. 2016, doi: 10.1016/j.neucom.2016.05.012.

[69] J. Luka, J. Fridrich, and M. Goljan, "Digital Camera Identification From Sensor Pattern Noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, Jun. 2006, doi: 10.1109/TIFS.2006.873602.

[70] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining Image Origin and Integrity Using Sensor Noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, 2008, doi: 10.1109/TIFS.2007.916285.

[71] F. Gharibi, F. Akhlaghian, J. RavanJamjah, and B. ZahirAzami, "Using the local information of image to identify the source camera," in *The 10th IEEE International Symposium on Signal Processing and Information Technology*, IEEE, Dec. 2010, pp. 515–519. doi: 10.1109/ISSPIT.2010.5711808.

[72] Y. Tomioka, Y. Ito, and H. Kitazawa, "Robust Digital Camera Identification Based on Pairwise

Magnitude Relations of Clustered Sensor Pattern Noise," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 1986–1995, Dec. 2013, doi: 10.1109/TIFS.2013.2284761.

[73] A. E. Dirik, H. T. Sencar, and N. Memon, "Digital Single Lens Reflex Camera Identification From Traces of Sensor Dust," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 3, pp. 539–552, Sep. 2008, doi: 10.1109/TIFS.2008.926987.

[74] R. Li, C.-T. Li, and Y. Guan, "A compact representation of sensor fingerprint for camera identification and fingerprint matching," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2015, pp. 1777–1781. doi: 10.1109/ICASSP.2015.7178276.

[75] S. Liu, J. Chen, Y. Xun, X. Zhao, and C.-H. Chang, "A New Polarization Image Demosaicking Algorithm by Exploiting Inter-Channel Correlations With Guided Filtering," *IEEE Transactions on Image Processing*, vol. 29, pp. 7076–7089, 2020, doi: 10.1109/TIP.2020.2998281.

[76] G. Wu, X. Kang, and K. J. R. Liu, "A context adaptive predictor of sensor pattern noise for camera source identification," in *2012 19th IEEE International Conference on Image Processing*, IEEE, Sep. 2012, pp. 237–240. doi: 10.1109/ICIP.2012.6466839.

[77] H. Zeng and X. Kang, "Fast Source Camera Identification Using Content Adaptive Guided Image Filter," *J Forensic Sci*, vol. 61, no. 2, pp. 520–526, Mar. 2016, doi: 10.1111/1556-4029.13017.

[78] O. Çeliktutan, I. Avcibas, and B. Sankur, "Blind identification of cellular phone cameras," Feb. 2007, p. 65051H. doi: 10.1117/12.703920.

[79] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans Pattern Anal Mach Intell*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.

[80] N. Zandi and F. Razzazi, "Source Camera Identification Using WLBP Descriptor," in *2020 International Conference on Machine Vision and Image Processing (MVIP)*, IEEE, Feb. 2020, pp. 1–6. doi: 10.1109/MVIP49855.2020.9187484.

[81] Hongmei Gou, A. Swaminathan, and Min Wu, "Intrinsic Sensor Noise Features for Forensic Analysis on Scanners and Scanned Images," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 476–491, Sep. 2009, doi: 10.1109/TIFS.2009.2026458.

[82] B. Wang, X. Kong, and X. You, "Source Camera Identification Using Support Vector Machines," 2009, pp. 107–118. doi: 10.1007/978-3-642-04155-6_8.

[83] L. Bondi, L. Baroffio, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro, "First Steps Toward Camera Model Identification with Convolutional Neural Networks," Mar. 2016, doi: 10.1109/LSP.2016.2641006.

[84] D. Freire-Obregón, F. Narducci, S. Barra, and M. Castrillón-Santana, "Deep learning for source camera identification on mobile devices," *Pattern Recognit Lett*, vol. 126, pp. 86–91, Sep. 2019, doi: 10.1016/j.patrec.2018.01.005.

[85] N. Huang, J. He, N. Zhu, X. Xuan, G. Liu, and C. Chang, "Identification of the source camera of images based on convolutional neural network," *Digit Investig*, vol. 26, pp. 72–80, Sep. 2018, doi: 10.1016/j.diin.2018.08.001.

[86] Y. Chen, Y. Huang, and X. Ding, "Camera model identification with residual neural network," in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, Sep. 2017, pp. 4337–4341. doi: 10.1109/ICIP.2017.8297101.

[87] Y. Wang, Q. Sun, D. Rong, S. Li, and L. Da Xu, "Image Source Identification Using Convolutional Neural Networks in IoT Environment," *Wirel Commun Mob Comput*, vol. 2021, pp. 1–12, Sep. 2021, doi: 10.1155/2021/5804665.

[88] M. Kirchner and C. Johnson, "SPN-CNN: Boosting Sensor-Based Source Camera Attribution With Deep Learning," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Dec. 2019, pp. 1–6. doi: 10.1109/WIFS47025.2019.9035103.

[89] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017, doi: 10.1109/TIP.2017.2662206.

[90] M. Zhao, B. Wang, F. Wei, M. Zhu, and X. Sui, "Source Camera Identification Based on Coupling Coding and Adaptive Filter," *IEEE Access*, vol. 8, pp. 54431–54440, 2020, doi: 10.1109/ACCESS.2019.2959627.

[91] P. Yang, D. Baracchi, R. Ni, Y. Zhao, F. Argenti, and A. Piva, "A Survey of Deep Learning-Based Source Image Forensics," *J Imaging*, vol. 6, no. 3, p. 9, Mar. 2020, doi: 10.3390/jimaging6030009.

[92] H. Farid, "Image forgery detection," *IEEE Signal Process Mag*, vol. 26, no. 2, pp. 16–25, Mar. 2009, doi: 10.1109/MSP.2008.931079.

[93] S. Scholarworks@cwu and L. Alzamil, "Image Forgery Detection with Machine Learning." [Online]. Available: https://digitalcommons.cwu.edu/etd/1361

[94] G. K. Birajdar and V. H. Mankar, "Digital image forgery detection using passive techniques: A survey," *Digit Investig*, vol. 10, no. 3, pp. 226–245, Oct. 2013, doi: 10.1016/j.diin.2013.04.007.

[95] J. Ouyang, Y. Liu, and M. Liao, "Copy-move forgery detection based on deep learning," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, Oct. 2017, pp. 1–5. doi: 10.1109/CISP-BMEI.2017.8301940.

[96] S. S. Ali, I. I. Ganapathi, N.-S. Vu, S. D. Ali, N. Saxena, and N. Werghi, "Image Forgery

Detection Using Deep Learning by Recompressing Images," *Electronics (Basel)*, vol. 11, no. 3, p. 403, Jan. 2022, doi: 10.3390/electronics11030403.

[97] K. B. Meena and V. Tyagi, "Image Forgery Detection: Survey and Future Directions," in *Data, Engineering and Applications*, Singapore: Springer Singapore, 2019, pp. 163–194. doi: 10.1007/978-981-13-6351-1_14.

[98] Y. Rao, J. Ni, and H. Zhao, "Deep Learning Local Descriptor for Image Splicing Detection and Localization," *IEEE Access*, vol. 8, pp. 25611–25625, 2020, doi: 10.1109/ACCESS.2020.2970735.

[99] E. A. Armas Vega, E. Gonzalez Fernandez, A. L. Sandoval Orozco, and L. J. Garcia Villalba, "Passive Image Forgery Detection Based on the Demosaicing Algorithm and JPEG Compression," *IEEE Access*, vol. 8, pp. 11815–11823, 2020, doi: 10.1109/ACCESS.2020.2964516.

[100] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "A Full-Image Full-Resolution End-to-End-Trainable CNN Framework for Image Forgery Detection," *IEEE Access*, vol. 8, pp. 133488–133502, 2020, doi: 10.1109/ACCESS.2020.3009877.

[101] T. Mahmood, T. Nawaz, Z. Mehmood, Z. Khan, M. Shah, and R. Ashraf, "Forensic analysis of copy-move forgery in digital images using the stationary wavelets," in *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, IEEE, Aug. 2016, pp. 578–583. doi: 10.1109/INTECH.2016.7845040.

[102] W. Chen, Y. Q. Shi, and W. Su, "Image splicing detection using 2-D phase congruency and statistical moments of characteristic function," Feb. 2007, p. 65050R. doi: 10.1117/12.704321.

[103] Wei Wang, J. Dong, and T. Tan, "Effective image splicing detection based on image chroma," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, IEEE, Nov. 2009, pp. 1257–1260. doi: 10.1109/ICIP.2009.5413549.

[104] Z. Fang, S. Wang, and X. Zhang, "Image Splicing Detection Using Color Edge Inconsistency," in *2010 International Conference on Multimedia Information Networking and Security*, IEEE, 2010, pp. 923–926. doi: 10.1109/MINES.2010.196.

[105] Y. Huang, W. Lu, W. Sun, and D. Long, "Improved DCT-based detection of copy-move forgery in images," *Forensic Sci Int*, vol. 206, no. 1–3, pp. 178–184, Mar. 2011, doi: 10.1016/j.forsciint.2010.08.001.

[106] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "A SIFT-Based Forensic Method for Copy–Move Attack Detection and Transformation Recovery," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1099–1110, Sep. 2011, doi: 10.1109/TIFS.2011.2129512.

[107] G. Liu, J. Wang, S. Lian, and Z. Wang, "A passive image authentication scheme for detecting region-duplication forgery with rotation," *Journal of Network and Computer Applications*, vol.

34, no. 5, pp. 1557–1565, Sep. 2011, doi: 10.1016/j.jnca.2010.09.001.

[108] G. Muhammad, M. Hussain, and G. Bebis, "Passive copy move image forgery detection using undecimated dyadic wavelet transform," *Digit Investig*, vol. 9, no. 1, pp. 49–57, Jun. 2012, doi: 10.1016/j.diin.2012.04.004.

[109] J. Zhao and J. Guo, "Passive forensics for copy-move image forgery using a method based on DCT and SVD," *Forensic Sci Int*, vol. 233, no. 1–3, pp. 158–166, Dec. 2013, doi: 10.1016/j.forsciint.2013.09.013.

[110] L. Li, S. Li, H. Zhu, S.-C. Chu, J. F. Roddick, and J.-S. Pan, "Journal of Information Hiding and Multimedia Signal Processing©2013 ISSN," 2013.

[111] A. A. Alahmadi, M. Hussain, H. Aboalsamh, G. Muhammad, and G. Bebis, "Splicing image forgery detection based on DCT and Local Binary Pattern," in *2013 IEEE Global Conference on Signal and Information Processing*, IEEE, Dec. 2013, pp. 253–256. doi: 10.1109/GlobalSIP.2013.6736863.

[112] E.-S. M. El-Alfy and M. A. Qureshi, "Combining spatial and DCT based Markov features for enhanced blind detection of image splicing," *Pattern Analysis and Applications*, vol. 18, no. 3, pp. 713–723, Aug. 2015, doi: 10.1007/s10044-014-0396-4.

[113] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," in *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Dec. 2016, pp. 1–6. doi: 10.1109/WIFS.2016.7823911.

[114] J. Ouyang, Y. Liu, and M. Liao, "Copy-move forgery detection based on deep learning," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, Oct. 2017, pp. 1–5. doi: 10.1109/CISP-BMEI.2017.8301940.

[115] N. Huang, J. He, and N. Zhu, "A Novel Method for Detecting Image Forgery Based on Convolutional Neural Network," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, IEEE, Aug. 2018, pp. 1702–1705. doi: 10.1109/TrustCom/BigDataSE.2018.00255.

[116] Y. Akbari, S. Al-maadeed, O. Elharrouss, F. Khelifi, A. Lawgaly, and A. Bouridane, "Digital forensic analysis for source video identification: A survey," *Forensic Science International: Digital Investigation*, vol. 41, p. 301390, Jun. 2022, doi: 10.1016/j.fsidi.2022.301390.

[117] D.-K. Hyun, C.-H. Choi, and H.-K. Lee, "Camcorder Identification for Heavily Compressed Low Resolution Videos," 2012, pp. 695–701. doi: 10.1007/978-94-007-2792-2_68.

[118] A. Mahalanobis, B. V. K. Vijaya Kumar, and D. Casasent, "Minimum average correlation energy filters," *Appl Opt*, vol. 26, no. 17, p. 3633, Sep. 1987, doi: 10.1364/AO.26.003633.

[119] L. J. García Villalba, A. L. Sandoval Orozco, R. Ramos López, and J. Hernandez Castro, "Identification of smartphone brand and model via forensic video analysis," *Expert Syst Appl*, vol. 55, pp. 59–69, Aug. 2016, doi: 10.1016/j.eswa.2016.01.025.

[120] A.-A. M, K. F, C. D, and F. M, "Digital Video Source Identification Based on Green-Channel Photo Response Non-Uniformity (G-PRNU)," in *Computer Science & Information Technology (CS & IT)*, Academy & Industry Research Collaboration Center (AIRCC), Sep. 2016, pp. 47–57. doi: 10.5121/csit.2016.61105.

[121] Wales, G. S. (2019). Proposed framework for digital video authentication. *University of Colorado at Denver*.

[122] W.-C. Yang, J. Jiang, and C.-H. Chen, "A fast source camera identification and verification method based on PRNU analysis for use in video forensic investigations," *Multimed Tools Appl*, vol. 80, no. 5, pp. 6617–6638, Feb. 2021, doi: 10.1007/s11042-020-09763-z.

[123] D. Shullani, M. Fontani, M. Iuliani, O. Al Shaya, and A. Piva, "VISION: a video and image dataset for source identification," *EURASIP J Inf Secur*, vol. 2017, no. 1, p. 15, Dec. 2017, doi: 10.1186/s13635-017-0067-2.

[124] R. Ramos Lopez, E. Almaraz Luengo, A. L. Sandoval Orozco, and L. J. G. Villalba, "Digital Video Source Identification Based on Container's Structure Analysis," *IEEE Access*, vol. 8, pp. 36363–36375, 2020, doi: 10.1109/ACCESS.2020.2971785.

[125] Chang-Tsun Li, "Source Camera Identification Using Enhanced Sensor Pattern Noise," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 280–287, Jun. 2010, doi: 10.1109/TIFS.2010.2046268.

[126] R. Caldelli, I. Amerini, F. Picchioni, and M. Innocenti, "Fast image clustering of unknown source images," in *2010 IEEE International Workshop on Information Forensics and Security*, IEEE, Dec. 2010, pp. 1–5. doi: 10.1109/WIFS.2010.5711454.

[127] P. Ferrara, M. Iuliani, and A. Piva, "PRNU-Based Video Source Attribution: Which Frames Are You Using?," *J Imaging*, vol. 8, no. 3, p. 57, Feb. 2022, doi: 10.3390/jimaging8030057.

[128] J. Xu, H. Chang, S. Yang, and M. Wang, "Fast feature-based video stabilization without accumulative global motion estimation," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 993–999, Aug. 2012, doi: 10.1109/TCE.2012.6311347.

[129] M. Grundmann, V. Kwatra, and I. Essa, "Auto-directed video stabilization with robust L1 optimal camera paths," in *CVPR 2011*, IEEE, Jun. 2011, pp. 225–232. doi: 10.1109/CVPR.2011.5995525.

[130] Thivent, D. J., Williams, G. E., Zhou, J., Baer, R. L., Toft, R., & Beysserie, S. X. (2017). U.S. Patent No. 9,596,411. Washington, DC: U.S. Patent and Trademark Office.

[131] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, "Content-preserving warps for 3D video stabilization," *ACM Trans Graph*, vol. 28, no. 3, pp. 1–9, Jul. 2009, doi: 10.1145/1531326.1531350.

[132] Z. Wang, L. Zhang, and H. Huang, "High-Quality Real-Time Video Stabilization Using Trajectory Smoothing and Mesh-Based Warping," *IEEE Access*, vol. 6, pp. 25157–25166, 2018, doi: 10.1109/ACCESS.2018.2828653.

[133] A. Karaküçük, A. E. Dirik, H. T. Sencar, and N. D. Memon, "Recent advances in counter PRNU based source attribution and beyond," A. M. Alattar, N. D. Memon, and C. D. Heitzenrater, Eds., Mar. 2015, p. 94090N. doi: 10.1117/12.2182458.

[134] M. Goljan and J. Fridrich, "Camera identification from cropped and scaled images," E. J. Delp III, P. W. Wong, J. Dittmann, and N. D. Memon, Eds., Feb. 2008, p. 68190E. doi: 10.1117/12.766732.

[135] S. McCloskey, "Confidence weighting for sensor fingerprinting," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Jun. 2008, pp. 1–6. doi: 10.1109/CVPRW.2008.4562986.

[136] W.-H. Chuang, H. Su, and M. Wu, "Exploring compression effects for improved source camera identification using strongly compressed video," in *2011 18th IEEE International Conference on Image Processing*, IEEE, Sep. 2011, pp. 1953–1956. doi: 10.1109/ICIP.2011.6115855.

[137] M. Goljan, M. Chen, P. Comesaña, and J. Fridrich, "Effect of Compression on Sensor-Fingerprint Based Camera Identification." [Online]. Available: http://en.wikipedia.

[138] R. Ramos López, A. L. Sandoval Orozco, and L. J. García Villalba, "Compression effects and scene details on the source camera identification of digital videos," *Expert Syst Appl*, vol. 170, p. 114515, May 2021, doi: 10.1016/j.eswa.2020.114515.

[139] T. Höglund, P. Brolund, and K. Norell, "Identifying camcorders using noise patterns from video clips recorded with image stabilisation," in *2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2011, pp. 668–671.

[140] S. Taspinar, M. Mohanty, and N. Memon, "Source camera attribution using stabilized video," in *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Dec. 2016, pp. 1–6. doi: 10.1109/WIFS.2016.7823918.

[141] M. Iuliani, M. Fontani, D. Shullani, and A. Piva, "Hybrid reference-based Video Source Identification," *Sensors*, vol. 19, no. 3, p. 649, Feb. 2019, doi: 10.3390/s19030649.

[142] S. Mandelli, P. Bestagini, L. Verdoliva, and S. Tubaro, "Facing Device Attribution Problem for Stabilized Video Sequences," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 14–27, 2020, doi: 10.1109/TIFS.2019.2918644.

[143] E. Altinisik and H. T. Sencar, "Source Camera Verification for Strongly Stabilized Videos," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 643–657, 2021, doi: 10.1109/TIFS.2020.3016830.

[144] P. Ferrara and L. Beslay, "Robust video source recognition in presence of motion stabilization," in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, Apr. 2020, pp. 1–6. doi: 10.1109/IWBF49977.2020.9107957.

[145] S. Taspinar, M. Mohanty, and N. Memon, "Camera identification of multi-format devices," *Pattern Recognit Lett*, vol. 140, pp. 288–294, Dec. 2020, doi: 10.1016/j.patrec.2020.10.010.

[146] W. van Houten and Z. Geradts, "Source video camera identification for multiply compressed videos originating from YouTube," *Digit Investig*, vol. 6, no. 1–2, pp. 48–60, Sep. 2009, doi: 10.1016/j.diin.2009.05.003.

[147] Scheelen, Y., & van der Lelie, J. (2012). Camera identification on YouTube. *Chin. J. Forensic Sci*, 64, 19-39.

[148] Brouwers, M., & Mousa, R. (2017). Automatic comparison of photo response non uniformity (PRNU) on Youtube. *System and Network Engineering*", styczeń.

[149] I. Amerini, R. Caldelli, A. Del Mastio, A. Di Fuccia, C. Molinari, and A. P. Rizzo, "Dealing with video source identification in social networks," *Signal Process Image Commun*, vol. 57, pp. 1–7, Sep. 2017, doi: 10.1016/j.image.2017.04.009.

[150] C. Meij and Z. Geradts, "Source camera identification using Photo Response Non-Uniformity on WhatsApp," *Digit Investig*, vol. 24, pp. 142–154, Mar. 2018, doi: 10.1016/j.diin.2018.02.005.

[151] E. K. Kouokam and A. E. Dirik, "PRNU-based source device attribution for YouTube videos," *Digit Investig*, vol. 29, pp. 91–100, Jun. 2019, doi: 10.1016/j.diin.2019.03.005.

[152] A. Pande, S. Chen, P. Mohapatra, and J. Zambreno, "Hardware Architecture for Video Authentication Using Sensor Pattern Noise," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 157–167, Jan. 2014, doi: 10.1109/TCSVT.2013.2276869.

[153] Shaxun Chen, A. Pande, Kai Zeng, and P. Mohapatra, "Live Video Forensics: Source Identification in Lossy Wireless Networks," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 1, pp. 28–39, Jan. 2015, doi: 10.1109/TIFS.2014.2362848.

[154] J. Kaur and D. K. K. Randhawa, "Source Identification of Videos Transmitted in Lossy Wireless Networks," *IJIREEICE*, vol. 5, no. 5, pp. 331–339, May 2017, doi: 10.17148/ijireeice.2017.5552.

[155] C. Sammut and G. I. Webb, Eds., *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2010. doi: 10.1007/978-0-387-30164-8.

[156] E. Altinisik and H. T. Sencar, "Source Camera Verification for Strongly Stabilized Videos,"

IEEE Transactions on Information Forensics and Security, vol. 16, pp. 643–657, 2021, doi: 10.1109/TIFS.2020.3016830.

[157] B. S. Reddy and B. N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Transactions on Image Processing*, vol. 5, no. 8, pp. 1266–1271, Aug. 1996, doi: 10.1109/83.506761.

[158] S. Taspinar, M. Mohanty, and N. Memon, "Effect of Video Pixel-Binning on Source Attribution of Mixed Media," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021, pp. 2545–2549. doi: 10.1109/ICASSP39728.2021.9415094.

[159] Su, Y., Xu, J., Dong, B., Zhang, J., & Liu, Q. (2010). A novel source mpeg-2 video identification algorithm. *International Journal of Pattern Recognition and Artificial Intelligence,* 24(08), 1311-1328.

[160] S. Yahaya, A. T. S. Ho, and A. A. Wahab, "Advanced video camera identification using conditional probability features," in *IET Conference on Image Processing (IPR 2012)*, IET, 2012, pp. 132–132. doi: 10.1049/cp.2012.0426.

[161] B. Hosler *et al.*, "A Video Camera Model Identification System Using Deep Learning and Fusion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2019, pp. 8271–8275. doi: 10.1109/ICASSP.2019.8682608.

[162] D. Timmerman, S. Bennabhaktula, E. Alegre, and G. Azzopardi, "Video Camera Identification from Sensor Pattern Noise with a Constrained ConvNet," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.06277

[163] O. Mayer, B. Hosler, and M. C. Stamm, "Open Set Video Camera Model Verification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020, pp. 2962–2966. doi: 10.1109/ICASSP40776.2020.9054261.

[164] A. W. Wahab, J. A. Briffa, H. G. Schaathun, and A. T. S. Ho, "Conditional Probability Based Steganalysis for JPEG Steganography," in *2009 International Conference on Signal Processing Systems*, IEEE, 2009, pp. 205–209. doi: 10.1109/ICSPS.2009.71.

[165] O. Mayer and M. C. Stamm, "Forensic Similarity for Digital Images," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1331–1346, 2020, doi: 10.1109/TIFS.2019.2924552.

[166] C. Galdi, F. Hartung, and J.-L. Dugelay, "SOCRatES: A Database of Realistic Data for SOurce Camera REcognition on Smartphones." [Online]. Available: http://socrates.eurecom.fr/.

[167] Goljan, M., Fridrich, J., & Filler, T. (2009, February). Large scale test of sensor fingerprint camera identification. *In Media forensics and security* (Vol. 7254, pp. 170-181). SPIE.

[168] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444,

May 2015, doi: 10.1038/nature14539.

[169] D. Cozzolino and L. Verdoliva, "Noiseprint: a CNN-based camera model fingerprint," Aug. 2018, [Online]. Available: http://arxiv.org/abs/1808.08396

[170] T.-C. Phan, A.-C. Phan, H.-P. Cao, and T.-N. Trieu, "Content-Based Video Big Data Retrieval with Extensive Features and Deep Learning," *Applied Sciences*, vol. 12, no. 13, p. 6753, Jul. 2022, doi: 10.3390/app12136753.

[171] G. S. Bennabhaktula, D. Timmerman, E. Alegre, and G. Azzopardi, "Source Camera Device Identification from Videos," *SN Comput Sci*, vol. 3, no. 4, p. 316, Jul. 2022, doi: 10.1007/s42979-022-01202-0.

[172] Y. Akbari *et al.*, "A New Forensic Video Database for Source Smartphone Identification: Description and Analysis," *IEEE Access*, vol. 10, pp. 20080–20091, 2022, doi: 10.1109/ACCESS.2022.3151406.

[173] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Weapon Detection in Real-Time CCTV Videos Using Deep Learning," *IEEE Access*, vol. 9, pp. 34366–34382, 2021, doi: 10.1109/ACCESS.2021.3059170.

[174] Y. Shi, D. Feng, Y. Cheng, and S. Biswas, "A natural language-inspired multilabel video streaming source identification method based on deep neural networks," *Signal Image Video Process*, vol. 15, no. 6, pp. 1161–1168, Sep. 2021, doi: 10.1007/s11760-020-01844-8.

[175] L. Maiano, I. Amerini, L. Ricciardi Celsi, and A. Anagnostopoulos, "Identification of Social-Media Platform of Videos through the Use of Shared Features," *J Imaging*, vol. 7, no. 8, p. 140, Aug. 2021, doi: 10.3390/jimaging7080140.

[176] D. Dal Cortivo, S. Mandelli, P. Bestagini, and S. Tubaro, "CNN-Based Multi-Modal Camera Model Identification on Video Sequences," *J Imaging*, vol. 7, no. 8, p. 135, Aug. 2021, doi: 10.3390/jimaging7080135.

[177] J. Salido, V. Lomas, J. Ruiz-Santaquiteria, and O. Deniz, "Automatic Handgun Detection with Deep Learning in Video Surveillance Images," *Applied Sciences*, vol. 11, no. 13, p. 6085, Jun. 2021, doi: 10.3390/app11136085.

[178] V.-N. Huynh and H.-H. Nguyen, "Fast pornographic video detection using Deep Learning," in *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, IEEE, Aug. 2021, pp. 1–6. doi: 10.1109/RIVF51545.2021.9642154.

[179] Y. Wang, Q. Sun, D. Rong, S. Li, and L. Da Xu, "Image Source Identification Using Convolutional Neural Networks in IoT Environment," *Wirel Commun Mob Comput*, vol. 2021, pp. 1–12, Sep. 2021, doi: 10.1155/2021/5804665.

[180] H. Fan *et al.*, "PyTorchVideo: A Deep Learning Library for Video Understanding," in *MM*

2021 - Proceedings of the 29th ACM International Conference on Multimedia, Association for Computing Machinery, Inc, Oct. 2021, pp. 3783–3786. doi: 10.1145/3474085.3478329.

[181] D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.11126

[182] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, Mar. 2020, doi: 10.1016/j.neucom.2019.11.023.

[183] Lukas, J., Fridrich, J., & Goljan, M. (2006). Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security, 1(2), 205-214.

[184] S. Bayram, H. Sencar, N. Memon, and I. Avcibas, "Source camera identification based on CFA interpolation," in *IEEE International Conference on Image Processing 2005*, IEEE, 2005, pp. III–69. doi: 10.1109/ICIP.2005.1530330.

[185] L.-J. Li, H. Su, E. Xing, and F.-F. Li, "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification.," Aug. 2010, pp. 1378–1386.

[186] L. Bondi, L. Baroffio, D. Guera, P. Bestagini, E. J. Delp, and S. Tubaro, "First Steps Toward Camera Model Identification With Convolutional Neural Networks," *IEEE Signal Process Lett*, vol. 24, no. 3, pp. 259–263, Mar. 2017, doi: 10.1109/LSP.2016.2641006.

[187] A. Ashraf, T. Surya Gunawan, B. Subhan Riza, E. V. Haryanto, and Z. Janin, "On the review of image and video-based depression detection using machine learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 3, p. 1677, Sep. 2020, doi: 10.11591/ijeecs.v19.i3.pp1677-1684.

[188] D. Schofield *et al.*, "Chimpanzee face recognition from videos in the wild using deep learning," *Sci Adv*, vol. 5, no. 9, Sep. 2019, doi: 10.1126/sciadv.aaw0736.

[189] G. Sreenu and M. A. Saleem Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," *J Big Data*, vol. 6, no. 1, p. 48, Dec. 2019, doi: 10.1186/s40537-019-0212-5.

[190] T. Akilan, Q. M. Jonathan Wu, W. Jiang, A. Safaei, and J. Huo, "New Trend in Video Foreground Detection Using Deep Learning," in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, IEEE, Aug. 2018, pp. 889–892. doi: 10.1109/MWSCAS.2018.8623825.

[191] A. Shojaei-Hashemi, P. Nasiopoulos, J. J. Little, and M. T. Pourazad, "Video-based Human Fall Detection in Smart Homes Using Deep Learning," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, May 2018, pp. 1–5. doi: 10.1109/ISCAS.2018.8351648.

[192] E. Athanasiadou, Z. Geradts, and E. Van Eijk, "Camera recognition with deep learning,"

Forensic Sci Res, vol. 3, no. 3, pp. 210–218, Jul. 2018, doi: 10.1080/20961790.2018.1485198.

[193] J. Ott, A. Atchison, P. Harnack, A. Bergh, and E. Linstead, "A deep learning approach to identifying source code in images and video," in *Proceedings of the 15th International Conference on Mining Software Repositories*, New York, NY, USA: ACM, May 2018, pp. 376–386. doi: 10.1145/3196398.3196402.

[194] W. Wang *et al.*, "Learning Two-Stream CNN for Multi-Modal Age-Related Macular Degeneration Categorization," *IEEE J Biomed Health Inform*, vol. 26, no. 8, pp. 4111–4122, Aug. 2022, doi: 10.1109/JBHI.2022.3171523.

[195] F. Abdullakutty, P. Johnston, and E. Elyan, "Fusion Methods for Face Presentation Attack Detection," *Sensors*, vol. 22, no. 14, p. 5196, Jul. 2022, doi: 10.3390/s22145196.

[196] E. Blasch, Z. Liu, and Y. Zheng, "Advances in deep learning for infrared image processing and exploitation," in *Infrared Technology and Applications XLVIII*, G. F. Fulop, M. Kimata, L. Zheng, B. F. Andresen, J. L. Miller, and Y.-H. Kim, Eds., SPIE, May 2022, p. 56. doi: 10.1117/12.2619140.

[197] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, "Semantic Similarity Metrics for Evaluating Source Code Summarization," in *IEEE International Conference on Program Comprehension*, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/.

[198] A. Gona and M. Subramoniam, "Convolutional neural network with improved feature ranking for robust multi-modal biometric system," *Computers and Electrical Engineering*, vol. 101, p. 108096, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108096.

[199] M. A. Uddin, J. B. Joolee, and K.-A. Sohn, "Deep Multi-Modal Network Based Automated Depression Severity Estimation," *IEEE Trans Affect Comput*, pp. 1–1, 2022, doi: 10.1109/TAFFC.2022.3179478.

[200] J. Ott, A. Atchison, P. Harnack, A. Bergh, and E. Linstead, "A deep learning approach to identifying source code in images and video," in *Proceedings of the 15th International Conference on Mining Software Repositories*, New York, NY, USA: ACM, May 2018, pp. 376–386. doi: 10.1145/3196398.3196402.

[201] I. Amerini, A. Anagnostopoulos, L. Maiano, and L. R. Celsi, *Deep Learning for Multimedia Forensics*. Now Publishers, 2021. doi: 10.1561/9781680838558.

List of Publication:

Published:

Journal Publication

1. Singh, S., & Sehgal, V. K. (2023). Deep Learning-Based CNN Multi-Modal Camera Model Identification for Video Source Identification. *Informatica*, *47*(3), doi: https://doi.org/10.31449/inf.v47i3.4392. [Scopus]
2. Singh, S., & Sehgal, V. K. (2023). Exploring Biomedical Video Source Identification: Transitioning from Fuzzy-Based Systems to Machine Learning Models. *Fuzzy Information and Engineering*. [Scopus (Published), ESCI,IF=1.2] https://doi.org/10.26599/FIE.2023.9270030.
3. Singh, S., & Sehgal, V. K. (2023). A Hybrid Data Fusion Approach with Twin CNN Architecture for Enhancing Image Source Identification in IoT Environment. *Computational Intelligence*. [SCIE(Published), IF=2.8] https://doi.org/10.1111/coin.12631

Conference Publication

1. S. Singh and V. K. Sehgal, "A Comprehensive Study: Image Forensic Analysis Traditional to Cognitive Image Processing," *2022 8th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2022, pp. 236-243, doi: 10.1109/ICSC56524.2022.10009322.
2. S. Singh and V. K. Sehgal, "Image Forgery Detection Model using CNN Architecture with SVM Classifier," *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Solan, Himachal Pradesh, India, 2022, pp. 263-268, doi: 10.1109/PDGC56933.2022.10053298.