

**DATA CLEANSING FRAMEWORK TO ENHANCE
QUALITY OF MULTIMEDIA DATA USING
CONVOLUTIONAL NEURAL NETWORKS**

ALPHONSE MUTHUSI KIOKO

**MASTER OF SCIENCE
(Computer Systems)**

**JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE TECHNOLOGY**

2025

**Data Cleansing Framework to Enhance Quality of Multimedia Data
Using Convolutional Neural Networks**

Alphonse Muthusi Kioko

**A Thesis Submitted in Partial Fulfilment of the Requirements for
the Degree of Master of Science in Computer Systems of the
Jomo Kenyatta University of Agriculture and Technology**

2025

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University

Signature.....Date.....

Alphonse Muthusi Kioko

This thesis has been submitted for examination with our approval as the University Supervisors

Signature.....Date.....

Prof. Cheruiyot W.K, PhD
JKUAT, Kenya

Signature.....Date.....

Dr. Richard Rimiru, PhD
JKUAT, Kenya

DEDICATION

I dedicate this academic work to my two mothers: Mary Kioko (deceased) and Mary Ungerleider. Their unwavering belief in me inspired the pursuit and completion of this scholarly endeavor. To Professor Charles Ungerleider and my foster sisters, Jessie and Suzie—your steadfast encouragement became my anchor, even in moments of self-doubt. From these extraordinary individuals, I drew the energy, focus, and discipline to achieve my aspirations, including this milestone of academic excellence.

Finally, I dedicate this work to my beloved wife, Joy, and our three children Tim, Shalom, and Rhema, whose boundless patience and love sustained me throughout this journey.

ACKNOWLEDGEMENT

This significant achievement would not have been possible without the collective contributions of many individuals. First and foremost, I extend my sincere gratitude to Prof. Cheruiyot and Dr. Rimiru of the School of Computing and Information Technology (SCIT) at Jomo Kenyatta University of Agriculture and Technology for their invaluable guidance and unwavering support throughout this research. I am also deeply grateful to my family, whose encouragement and sacrifices provided me the opportunity to complete this demanding academic journey despite the challenges encountered along the way.

Additionally, I wish to acknowledge Prof. Charles Ungerleider of the University of British Columbia and Mrs. Mary Ungerleider for their generous financial support, which proved instrumental in advancing my academic pursuits. Finally, I extend my appreciation to the Kenya Maritime Authority, my employer, for their trust in my capabilities and flexibility during this endeavor.

TABLE OF CONTENTS

DECLARATION..... ii

DEDICATION..... iii

ACKNOWLEDGEMENT iv

TABLE OF CONTENTS..... v

LIST OF TABLES xi

LIST OF FIGURES xiii

LIST OF APPENDICES xiv

ACRONYMS AND ABBREVIATIONS..... xv

ABSTRACT..... xvii

CHAPTER ONE 1

INTRODUCTION..... 1

 1.1 Background Information 1

 1.1.1 The Hidden Costs of Poor Data Quality 1

 1.1.2 The Imperative for Enhanced Data Cleansing Framework 2

 1.1.3 The Unique Challenges of Multimedia Data 2

 1.1.4 Image Data Integrity 2

1.1.5 The Role of Disruptive Technologies (DT).....	3
1.1.6 Towards Enhanced Data Cleansing Framework.....	3
1.2 Statement of the Problem	3
1.3 Justification	4
1.4 Objectives of the Study	5
1.4.1 Broad Objective	5
1.4.2 Specific Objectives	5
1.5 Research Questions	5
1.6 Scope of Study.....	6
CHAPTER TWO	8
LITERATURE REVIEW.....	8
2.1 Introduction	8
2.2 Data Cleansing Frameworks	9
2.2.1 Existing Data Cleansing Frameworks	15
2.2.2 Ajax Frameworks.....	18
2.2.3 ARKTOS Framework.....	19
2.2.4 IntelliClean	19
2.2.5 Potter’s Wheel Data Cleansing Framework	21
2.2.6 Multimedia Databases	21

2.2.7 Machine Learning	24
2.2.8 Data Science in Relation to Quality Data Sources	25
2.2.9 Big Data in Relation to Quality Data Sources	27
2.3 Related Research Works	29
2.3.1 Overview.....	29
2.3.2 Duplicate Detection and Record Linkage in Structured Data	29
2.3.3 Multimedia Data Quality Frameworks	30
2.3.4 Domain-Specific Multimedia Data Quality Challenges	31
2.3.5 Critique and Research Gap	31
2.3.6 Positioning the Current Study.....	32
2.4 Conceptual Framework	33
2.5 Summary Explanation of the Conceptual Framework	34
2.5.1 Data Sources Stage (A).....	34
2.5.2 Multimedia Database (MMDB).....	35
2.5.3 The Separator Layer	35
2.5.4 Extraction Layer	35
2.5.5 Pattern Recognizer.....	35
2.5.6 Multimedia Data Interpreter	36
CHAPTER THREE	37

RESEARCH METHODOLOGY	37
3.1 Introduction	37
3.2 Research Study Design.....	37
3.2.1 The Enhanced Framework Architecture	38
3.2.2 Dataset Description.....	39
3.2.3 Rationale of the Research Design.....	39
3.2.4 Motivations for the Research Design:	40
3.3 Dataset Curation and Preprocessing Pipeline.....	41
3.4 Target Population	41
3.5 Sample Size and Sampling Techniques.....	42
3.5.1 Determination of Sample Size	42
3.6 Data Processing Under Enhanced Data Cleansing Framework	43
3.7 The Intelligent Layers (i-Layer) Experimental Model Development	46
3.7.1 CNN + i-Layer Framework Configuration	46
3.7.2 Level 2: Transition from Level 1 to Level 2 Processing	46
3.7.3 Local/Web Based- Intelligent Image Forensic Analyzer Layer (iFAL- CNN Framework)	48
3.8 Experimental Set-Up	49
3.8.1 Handling and Processing of the Multimedia Dataset	52
3.8.2 Processing Pipeline for Data Quality Enhancement	53

3.8.3 Unique Treatment and Justification.....	54
3.9 Summary	54
CHAPTER FOUR.....	56
MODELLING, ANALYSIS AND DISCUSSIONS.....	56
4.1 Introduction	56
4.2 Performance Evaluation and Comparison Results	56
4.2.1 Confusion Matrix Components in the Sample Dataset.....	57
4.3 Hypothetical efficiency Comparative Analysis Results	58
4.3.1 Data Results for the Number of Cleansed Images.....	61
4.3.2 iFAL-CNN (Proposed) versus Major Non-CNN Based Methods.....	64
4.3.3 Innovation Leading to iFAL-CNN Superiority	65
4.4 Computational Resource Demands Limitation	65
4.5 Overall Conclusion for the chapter	66
CHAPTER FIVE.....	67
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	67
5.1 Introduction	67
5.2 Conclusion.....	68
5.2.1 Superior Cleansing Performance	68
5.2.2 Enhanced Image Forensically Analyzed	68

5.2.3 Superior Classification Accuracy	69
5.2.4 Real-Time Processing Advantage.....	69
5.3 Contributions of the Study	69
5.4 Recommendations	69
5.4.1 Deployment in Real-World Systems	70
5.4.2 Integration into Automated Machine Learning (AutoML).....	70
5.4.3 Further Research on Meta-Learning Expansion within iFAL-CNN	70
5.4.4 Policy and Standards Recommendation	70
5.5 Limitations.....	70
5.5.1 Limitation for Utilization for iFAL-CNN Framework	71
5.6 Final Statement.....	71
REFERENCES.....	73
APPENDICES	87

LIST OF TABLES

Table 2.1: Comparative Analysis of Traditional Data Cleaning Framework	16
Table 2.2: Comparative Analysis of Data Cleaning Frameworks Compared to CNN Based Expected Framework CNN Based Enhanced Multimedia Data Quality.....	17
Table 3.1: Dataset Composition	39
Table 3.2: Data Preprocessing Stages	41
Table 3.3: CNN + i-Layer Framework Configuration	46
Table 3.4: Dataset Unique Treatment and Justification	54
Table 4.1: Predictive Positives and Negatives	57
Table 4.2: Dataset Splits	58
Table 4.3: Hypothetical Efficiency Comparative Analysis Results.....	58
Table 4.4: Model Architectural Overview and Unique Component for Each.	59
Table 4.5: Unique Components for Each CNN Architectural Adoptability	59
Table 4.6: Unique Components for Each CNN Architectural Parameters.....	60
Table 4.7: iFAL-CNN Framework Performance versus Other CNN Architecture...	61
Table 4.8: Unique Components for Each CNN Architecture i-Layer Components..	63

Table 4.9: iFAL-CNN (Proposed) versus Non-CNN Based Methods 64

LIST OF FIGURES

Figure 2.1: The Current Data Source Structures.....	10
Figure 2.2: Diagram Simplifying Three Principles.....	11
Figure 2.3: Process Diagram of Data Mining	12
Figure 2.4: Data Cleansing Process Diagram Expanded	13
Figure 2.5: Internet User Distribution Statistics.	22
Figure 2.6: Data Science is Multidisciplinary.....	27
Figure 2.7: Conceptual Framework	34

LIST OF APPENDICES

Appendix I: Cost and Materials	87
Appendix II: Activity Schedule (Gantt Chart)	88

ACRONYMS AND ABBREVIATIONS

Ada Boost	Adaptive Boosting
AI	Artificial Intelligent
ATM	Automated Teller Machine
BDPA	Big Data Predictive Analytics
CART	Classification and Regression Tree Algorithm
CCF	Credit Card Fraud
CloDe	Cloning Detection
CNN	Convolutional Neural Network
DC	Data Cleansing
EDA	Experimental Design Assistant
EDT	Experiment Design Tool
ErLA	Error Level Analysis
FDA	Forensic Data Analysis
IT	Information Technology
K-NN	K-Nearest Neighbours
MATLAB	Matrix Laboratory
NoAn	Noise Analysis
PC	Personal Computer

PIN	Personal Identification Number
SVM	Support Vector Machine
TQM	Total Quality Management
VVVC	Volume, Velocity, Variety and Complexity

ABSTRACT

In the era of data-driven decision-making, the fitness of data quality to meet its intended purpose is of paramount importance. The success of machine learning (ML) models hinges on the quality of the datasets used during training. Real-world datasets, however, are often riddled with imperfections such as label noise, outliers, missing values, and inconsistencies across features, all of which degrade model performance and generalization. Traditional data-cleaning frameworks, while effective in specific scenarios, struggle to adapt to dynamic data patterns, multi-modal formats, and resource-constrained environments due to their domain-specific design. Most frameworks were developed for structured numeric or textual data, rendering them inadequate for addressing the unique challenges of multimedia formats. Even existing multimedia data-cleaning models remain domain-specific, highlighting the critical need for an enhanced, adaptive solution to improve data quality in image-centric applications. This study introduces the Intelligent Image Forensic Analyzer Layer (iFAL) integrated into a CNN framework, a novel approach enabling adaptive, efficient, and robust data cleaning through iFAL-learning features. This study systematically evaluates iFAL-CNN's capacity to address challenges across diverse datasets by integrating multi-modal features with capacity for extraction detailed dataset features missed by most of the outlined prior frameworks within and cross-domain generalization. This redefines automated data purification paradigms, offering a scalable algorithm for modern ML pipelines. Conventional rule-based models such as AutoClean (2019), CleanNet (2020), DCN-Clean (2021), and PurifiCNN (2022) are constrained by static heuristics, single-modality focus, and reliance on noise distribution assumptions. In contrast, the iFAL-CNN architecture overcomes these limitations by leveraging metadata analysis, error-level analysis, and authentication accuracy metrics. Designed to generalize across multimedia datasets, iFAL-CNN achieves state-of-the-art performance in accuracy, efficiency, and adaptability. Experimental results on a static dataset within various Data cleansing frameworks demonstrated the following accuracy percentages: Raw Data: 85.2%, AutoClean: 86.5% (+1.3), CleanNet: 87.8% (+2.6), Noise2Self: 88.1% (+2.9), DCN-Clean: 88.4% (+3.2), PurifiCNN: 89.3% (+4.1) and iFAL-CNN (Proposed): 90.7% (+5.5), making it the best performer. The results in this study demonstrates that incorporating hybrid feature extraction, metadata integrity verification, and adaptive error detection mechanisms enables superior cleansing, validation, and preparation of noisy large-scale image datasets for downstream machine learning tasks. iFAL-CNN framework (Proposed) offers a highly scalable, forensic-aware, and data-efficient purification framework that significantly enhances both dataset integrity and model learning stability. Future researches should prioritize Advanced Deep Learning Architectures (ADLA), such as Transformer-CNN hybrids with self-attention mechanisms, to extend these advancements into dynamic and virtual reality context.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

High-quality data is universally defined as information that remains "fit for its intended purpose" across operations, decision-making, and strategic planning (Purohit, 2021). In today's interconnected academic and business landscapes, data serves as the foundational bedrock for informed choices, driving innovations and policy formulations. Consequently, the efficacy of societal modernization efforts hinges on the integrity of the underlying data. This relationship is encapsulated in two enduring principles: Quality In, Quality out (QIQO) and Garbage In, Garbage out (GIGO). These axioms underscore the non-negotiable link between input data quality and output reliability a truth that remains immutable even as technological paradigms evolve (Munawar et al., 2020).

The exponential growth of data generation underscores this urgency. By 2025, global daily data production is projected to reach 463 Exabytes (EB), with unstructured data—including text, voice, and video constituting over 90% of this volume (Desjardins, 2019; Davis, 2019). However, the infrastructure to manage such heterogeneous datasets remains underdeveloped, exposing organizations to systemic risks. For instance, multimedia databases, which are critical for modern analytics, often lack standardized frameworks to detect errors during ingestion and processing. This deficiency amplifies the likelihood of "dirty data" information corrupted by inaccuracies, redundancies, or inconsistencies infiltrating decision-making pipelines (Alenazi & Ahmad, 2017).

1.1.1 The Hidden Costs of Poor Data Quality

The ramifications of poor data quality extend beyond operational inefficiencies. A 2017 KPMG survey revealed that 84% of Chief Executive Officers harbor concerns about the reliability of data underpinning their strategic decisions (Quezada-Gaibor et al., 2022). Such skepticism is warranted: compromised data invariably leads to

flawed insights, misallocated resources, and reputational damage. For instance, in healthcare, erroneous patient records can result in fatal diagnostic errors, while in finance, inaccurate transactional data may trigger regulatory penalties (Kiefer, 2016).

Unstructured data compounds these challenges. Human errors, application limitations, and incompatible formats during integration frequently introduce "dirt" into datasets (Kitsuse, 2019). Consider the process of merging customer feedback from social media (text), call center recordings (audio), and surveillance footage (video) into a unified database. Without robust validation protocols, duplicate entries, mismatched metadata, and syntax errors proliferate, rendering the dataset unusable for AI-driven sentiment analysis.

1.1.2 The Imperative for Enhanced Data Cleansing Framework

To address these vulnerabilities, this study focuses on data cleansing also defined as the systematic detection and rectification of errors, anomalies, and inconsistencies within datasets (Devi, 2015). Traditional frameworks, however, operate on a reactive "trial-and-error" basis, struggling to adapt to the dynamic nature of multimedia data (Krishnan, 2019). For instance, image repositories often contain tampered or mislabeled files that evade conventional rule-based checks.

1.1.3 The Unique Challenges of Multimedia Data

Multimedia data spanning images, audio, and video introduces unique complexities. Unlike structured numerical data, multimedia lacks inherent tabular organization, complicating automated validation. For example, an image's "quality" may depend on resolution, lighting, and contextual relevance, factors that defy binary classification (Holzinger, 2014).

1.1.4 Image Data Integrity

Consider a facial recognition system trained on a dataset of passport photos. If 15% of images are mislabeled, cropped incorrectly, or compressed at varying resolutions, the algorithm's accuracy plummets (Curtis, 2019). IBM's Hybrid Data Management platform attempts to mitigate such issues via AI-driven metadata

tagging, but gaps persist in detecting subtle manipulations like deepfakes or color grading alterations (M. et al., 2017).

1.1.5 The Role of Disruptive Technologies (DT)

Emerging technologies like Blockchain and AI offer transformative potential. Blockchain's immutable ledgers can authenticate data provenance, while AI models like convolutional neural networks (CNNs) automate anomaly detection in image datasets (Bruce Rogers, 2017). However, integrating these tools into legacy systems remains a hurdle. For instance, Internet of Things sensors generating real-time video feeds require edge computing capabilities to preprocess data before transmission, a feature absent in many existing frameworks.

1.1.6 Towards Enhanced Data Cleansing Framework

This study proposes an enhanced framework combining rule-based validation, machine learning, and human-in-the-loop (HITL) oversight to address multimedia data challenges and align it deep learning neural networks capacity to effortlessly dataset features in order to make precise deterministic predication in the exposed image datasets.

1.2 Statement of the Problem

Despite significant academic and industry research on data quality (DQ) over the past decades, there remains a critical need for re-evaluating frameworks, particularly for multimedia data, which continues to grow in volume and complexity to meet modern formatting demands. The adage “data is king” underscores its centrality to decision-making; however, even minor alterations, whether intentional or unintentional, can severely compromise analytical outcomes (Miller et al., 2025). The proliferation of free multimedia editing tools has exacerbated data quality issues, often producing misleading analyses and flawed decisions. A key challenge lies in the limitations of traditional Data Quality Frameworks (DQFs), which were designed primarily for numeric, alphanumeric, and textual data (Ridzuan & Zainon, 2024).

These frameworks struggle to accommodate the paradox created by modern advancements in multimedia formats.

Multimedia data is growing exponentially, surpassing structured numeric and textual data in volume and importance. While traditional data quality frameworks may have initiated management structured, numeric, and textual data, they failed to address the unique challenges posed by multimedia dataset formats (Mohammed et al., 2024). Existing multimedia data cleaning models remain domain-specific, challenged by the current dataset, hard to train with image datasets, justifying the need of an enhanced multimedia data cleansing model to address data quality especially in images and audiovisual content (Cao et al., 2024; Schwabe et al., 2024).

This study leverages CNN to enhance data quality and accuracy in multimedia data handling specifically images.

1.3 Justification

The growing complexity of Digital Data Transformation (DDT) has led to the widespread emergence of unstructured and heterogeneous datasets, exposing the limitations of traditional data cleansing frameworks (Miller et al., 2025). These conventional approaches often struggle to manage dynamic metadata and non-standard formats, which adversely affects data quality, consequently, the reliability of data-driven decision-making. This challenge is especially critical in immersive digital environments such as augmented and virtual reality, where the integrity of source data directly influences analytical outcomes (Schwabe et al., 2024).

The increasing demand for real-time, high-impact decisions further underscores the need for robust and adaptive data cleansing mechanisms. Data and Analytics Specialists (DAS) must maintain clean, consistent, and reliable datasets sourced from structured, semi-structured, and multimedia content repositories (Mohammed et al., 2024). However, existing frameworks are largely optimized for structured numeric or textual data and fail to address the distinct challenges posed by multimedia formats like images, audio, and video (Cao et al., 2024; Ridzuan & Zainon, 2024).

This study is therefore justified by a critical gap in cross-modal data quality assessment frameworks and proposes a unified data cleansing model that responds to current needs and aligns need for enhanced trust in AI-driven environments, moreso in high-stakes domains (Schwabe et al., 2024; Miller et al., 2025).

1.4 Objectives of the Study

1.4.1 Broad Objective

The primary purpose of the research was to implement a data-cleansing framework based on CNN to enhance image quality assurance in multimedia datasets.

1.4.2 Specific Objectives

- i. Examine the existing data cleansing frameworks in multimedia image datasets.
- ii. Develop a data cleansing framework for multimedia data by leveraging Convolutional Neural Network (CNN) architectures, to enhance the accuracy, adaptability, and efficiency of data quality improvement processes for unstructured datasets.
- iii. Evaluate the performance of the enhanced multimedia data cleansing framework against standard Convolutional Neural Network (CNN) models, using appropriate quality metrics and datasets to assess its effectiveness, robustness, and generalization capability in improving image data quality.

1.5 Research Questions

- i. What architectural framework or enhancement of the framework would be used to develop the expected framework to improve data quality within multimedia image datasets?
- ii. Can a data cleansing framework that enhances multimedia data images based on a Convolutional Neural Network (CNN) be implemented?
- iii. How can a data cleansing framework based on Convolutional Neural Network (CNN) be tested and evaluated?

1.6 Scope of Study

This study focuses on the development and evaluation of an advanced data cleansing framework for multimedia image datasets, with specific application to bird species imageries. The scope is confined to datasets sourced and simulated from online platforms such as eBird and Google Kaggle, which are characterized by large volumes of user-submitted images. These datasets commonly present challenges including data inconsistency, noise contamination, image tampering, duplication, metadata corruption, and mislabeling issues, which is prevalent in modern multimedia data ecosystems. Traditional frameworks failures to process images, their rule-based dependence, Inflexibility to datasets dynamics, Poor scalability, handle mislabeled images effectively and limited inbuilt multimedia support algorithms are key gaps addressed by this research.

The core of this research lies in the design and testing of a novel hybrid CNN-based framework, referred to as iFAL-CNN. This framework integrates multiple advanced techniques into a unified cleansing model, including Convolutional Neural Networks (CNNs) for feature extraction, metadata reinforcement for improved labeling consistency, forensic-level tamper detection for integrity verification, and meta-learning strategies for adaptability across diverse image quality conditions.

The dataset used for experimentation consists of 31,500 high-resolution bird images, organized into 315 subdirectories, each corresponding to a distinct bird species. All images were preprocessed to a uniform resolution of 100x100 pixels to standardize training conditions. The study is limited to image-based multimedia data and does not cover audio, video, and multi-modal content outside the domain of avian imagery.

Performance of the proposed iFAL-CNN data cleansing framework was evaluated using a set of metrics which assessed classification accuracy and computational efficiency. Accuracy in (%) served as the main metric, quantifying the proportion of correctly cleansed and classified images. This was via computing the ratio of true positives and true negatives to the total sample population, high score meant model's effectiveness in sensing and improving corrupted or mislabeled images. Precision

(%) measured the model's capacity to correctly hold valid data while abating false positives, thereby ensuring only truly dirty images were flagged. Recall (%) evaluated the proportion of actual despoiled images while identifying the model's sensitivity in cleansing capacity. F1-Score, the harmonic mean of Precision and Recall, offered a balanced performance measure, under data imbalance conditions. Computational Efficiency (%) was assessed by a comparative of computational resource utilization.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

While data is essential, its quality determines the value of the analytical outcomes. High-quality data directly enhances knowledge discovery across disciplines; however, incomplete or tampered datasets, common in multimedia applications introduce uncertainties during analysis, which must be addressed during the data cleansing stage (Fakhitah, 2019), which is going to be key in the proposed enhance Multimedia Data cleansing CNN based framework architecture.

The institutional reliance on data-driven insights and the growing demand for accuracy in analyzed datasets have necessitated advancements in existing data cleansing frameworks. Data professionals now face the dual challenge of maintaining quality in unstructured data while keeping pace with rapid data collection (Pouyanfar et al., 2018).

Data cleansing remains critical for producing actionable insights in any context (Sonka, 2016). Its role as a cornerstone of data mining has positioned it as a transformative practice in the field, with its significance projected to grow exponentially (Adu-Manu Sarpong et al., 2013).

Although numerous studies have explored this subject, most focus on data mining and knowledge discovery, leaving data cleansing—also known as data cleaning or scrubbing relatively under-researched (Devi, 2015). This gap underscores the untapped potential for contemporary and future research. As organizations increasingly depend on reliable data for strategic decision-making, refining data cleansing frameworks has become imperative to ensure integrity across source information systems (Adu-Manu Sarpong & Arthur, 2013).

2.2 Data Cleansing Frameworks

This research addresses gaps in existing data cleansing frameworks by proposing an integrated approach to enhance error detection and streamline processes in multimedia databases. While prior studies have explored foundational techniques for managing data quality in operational databases (Bai, 2019; Rahm & Do, 2000), current frameworks remain fragmented and lack the adaptability needed for modern, heterogeneous data sources. For instance, Bai (2019) observed that "data cleansing is a very young field of research" (p.14), with most literature focusing narrowly on error identification and outdated comparison methods (Rahm & Do, 2000). This narrow focus has left significant opportunities unexplored, particularly in automating error detection and scaling solutions for multimedia data. Panford (2019) likened current practices to a "black art performed in the basement" (p.7), emphasizing the need for systematic frameworks that unify disparate tools and methodologies.

Ranjit (2019) acknowledged that while researchers have cataloged common data quality issues, existing solutions often fail to address root causes holistically. Centralized, one-size-fits-all approaches remain inadequate for dynamically evolving datasets (Azeroual et al., 2018). This study's proposed framework aims to resolve these limitations by embedding intelligence directly into data cleansing tools, thereby reducing reliance on fragmented warehouse architectures. By integrating adaptive error detection and automated correction, the framework could significantly improve efficiency while simplifying workflows.

Figure 2.1 illustrates common sources of dirty data in multimedia databases, including inconsistencies in format, missing metadata, and corrupted files (Azeroual et al., 2018).

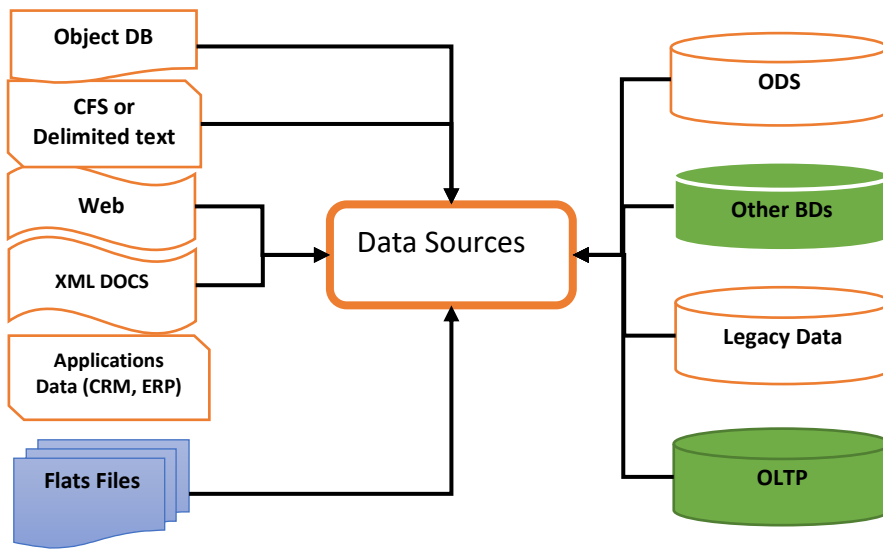


Figure 2.1: The Current Data Source Structures.

Data cleaning follows three core principles:

- i. Sources: Multimedia data originates from diverse sources, which may contain impurities or "dirty data" including incomplete entries, duplicates, or formatting inconsistencies.
- ii. Application of a Data Cleansing Matrix: A structured framework is required to address four key functions: descriptive, diagnostic, predictive, and prescriptive analytics. These functions collectively enable the algorithms and statistical methods necessary for effective data processing (Côte-Real, 2020).
- iii. Output of Clean Data: High-quality data emerges only after completing iterative stages of validation, transformation, and standardization.

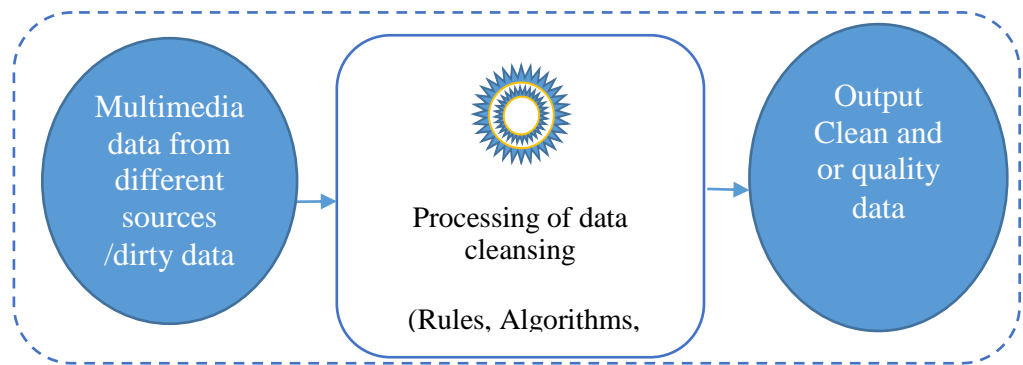


Figure 2.2: Diagram Simplifying Three Principles

Suriyagrace and Devapriya (2021) stated, “However hard an organization tries to eliminate dirt, some of its data collected turns out to be dirty.” They emphasized the growing complexity of managing large databases with diverse data sources and multidimensional formats.

To address this, data miners often relied on advanced algorithms or iterative cleaning rules, as no single tool could comprehensively cleanse dirty data (Chengyu, 2015). An enhanced data-cleansing framework could streamline these processes and reduce processing time.

A key challenge highlighted in the study was the repetitive use of auditing tools to trace potential bugs, which often resulted in slow, laborious transformations (Odun-Ayo et al., 2018). This inefficiency burdened users, who frequently endured prolonged wait times and wrote complex scripts to clean datasets. Meng et al. (2020) reinforced the importance of user-driven data cleaning (UDC), noting that automated scripts or systems risked poor data quality due to undetected errors. Similarly, Ridzuan and Wan Zainon (2019) advocated for early user involvement in data-cleaning operations at source points, arguing that correcting errors upfront saved significant time and computational resources.

However, the complexity of data cleansing remains a barrier. As Ethereum (2018) observed, “Due to the complexity of data cleaning, the processes involved were considered very difficult.” To mitigate this, Ethereum proposed an extensible framework with open libraries for rules and algorithms. The libraries mentioned

above stored the data cleansing matrices, enabling this framework to adapt to diverse data sources by selecting optimal rules for specific cleaning requests. Prakash et al. (2019) evaluated the performance of data cleansing algorithms in large-scale databases. Of particular relevance to this study is the data cleansing stage (illustrated in Figure 2.2), which forms the foundation of our research.

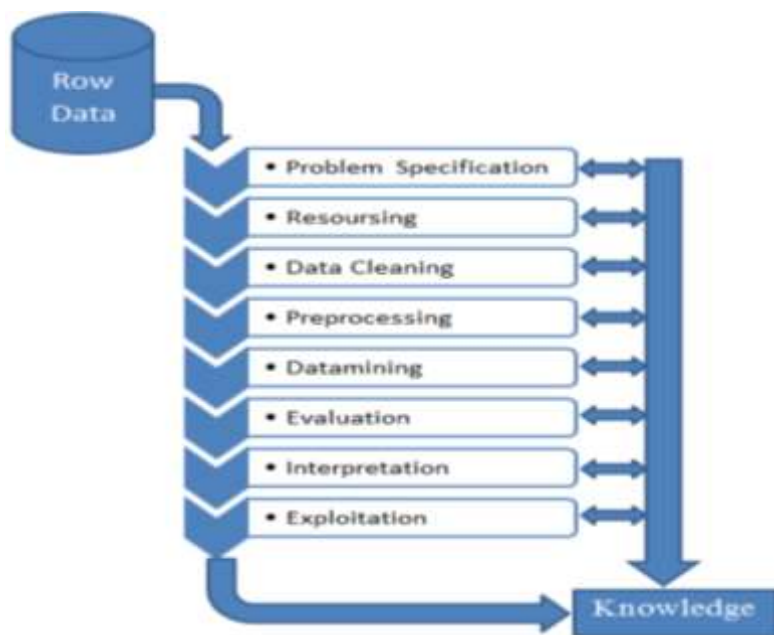


Figure 2.3: Process Diagram of Data Mining

Source: (Deepa & Chezian, 2016)

Figure 2.3 displays different activities which takes place in data cleaning process. Each of the shown processes is key and perform a specific action. These processes are explained as below.

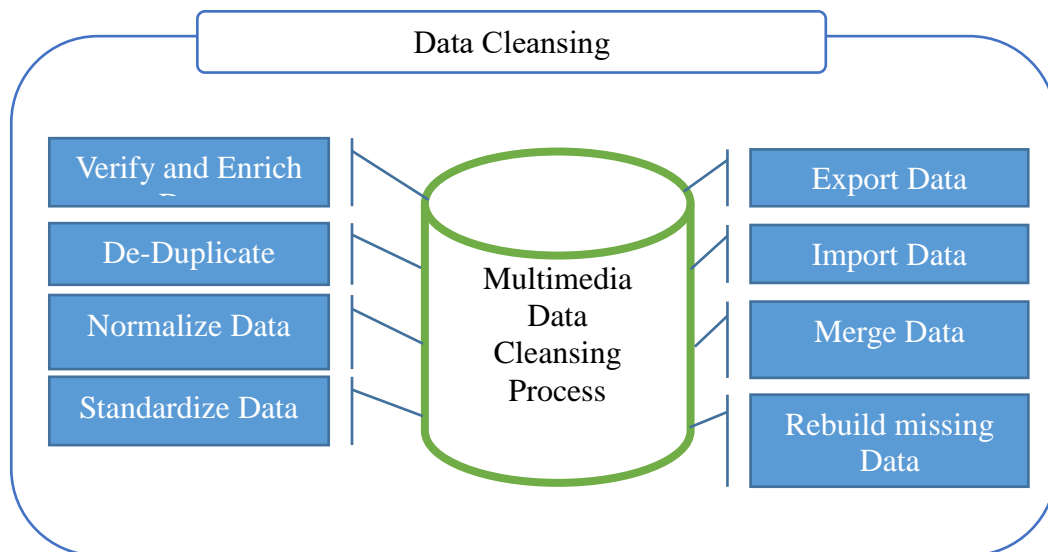


Figure 2.4: Data Cleansing Process Diagram Expanded

Source: (Deepa & Chezian, 2016)

Verify and Enrich Data

Verification begins with validating the integrity and authenticity of multimedia data. This includes checks for file corruption, metadata consistency, and adherence to domain-specific quality thresholds. Perceptual hashing and checksums detect anomalies, while automated validators flag malformed files. Enrichment augments incomplete sparse data by embedding contextual metadata (geolocation tags, object annotations) while enhancing low-quality content. Super-resolution upscale low-resolution images, while speech enhancement algorithms (DeepFilterNet) denoise audio. Studies highlight the role of active learning in prioritizing high-uncertainty samples for human-in-the-loop enrichment, reducing annotation costs (Settles, 2009).

De-Duplicate Data

Multimedia deduplication addresses both exact duplicates (identical files) and near-duplicates, resized images and cropped videos. Cryptographic hashing identifies exact duplicates, while perceptual hashing utilize feature-based methods to detect

near-duplicates by comparing semantic similarity. Siamese networks learn similarity metrics to cluster near-identical images. Challenges include scalability for large datasets and balancing false positives and negatives. Recent work leverages distributed frameworks referred to as Apache Spark and approximate nearest neighbor search for efficient deduplication (Johnson et al., 2019).

Normalize and Standardize Data

Normalization ensures multimedia data adheres to a uniform scale or format. For images, this may involve resizing to fixed dimensions 224x224 for CNNs, converting color spaces RGB to grayscale, and standardizing pixel value ranges ([0, 1] normalization). Temporal normalization aligns frame rates in videos to rates in audio. Standardization focuses on format consistency, converting heterogeneous file types from .png to .jpeg and enforcing metadata schemas. In multimodal contexts, synchronization protocols align timestamps across modalities.

Export Data

Exporting cleansed data involves structuring it into formats compatible with downstream systems. Common practices include serialization TFRecords for TensorFlow, HDF5 for large datasets, compression JPEG2000 for images, OPUS for audio, and partitioning, sharding videos by scene. Metadata is often exported as sidecar files example JSON, XML. Embedded databases example, NoSQL for unstructured data.

Import Data

Importing data requires robust validation pipelines to prevent reintroducing errors during ingestion. For instance, cloud-based ETL tools like Apache NiFi enforce schema-on-read checks, while version control systems example, DVC track dataset iterations. Scholarly work emphasizes reproducibility through standardized export/import workflows (Baylor et al., 2017).

Merge Data

Merging combines cleansed datasets from disparate sources, merging satellite imagery with ground sensor data. Key challenges include resolving schema conflicts of mismatched metadata fields, spatial-temporal alignment in georegistering images to a common coordinate system, and handling multimodal inconsistencies in aligning transcriptions with video frames. Entity resolution techniques example, fuzzy matching reconcile duplicates across datasets, while graph-based methods model relationships knowledge's graphs linking audio descriptions to visual objects. Recent advances use transformers to fuse multimodal embedding (Radford et al., 2021).

Rebuild Missing Data

Rebuilding addresses gaps caused by sensor failures, transmission errors, or incomplete annotations. For images/videos, inpainting techniques such PatchMatch, GAN-based models reconstruct missing regions using contextual cues. For metadata, imputation models such as k-NN, matrix factorization predict missing tags. Many scholarly debates center on avoiding overfitting in generative models and ensuring rebuilt data preserves statistical distributions (Pathak et al., 2016).

2.2.1 Existing Data Cleansing Frameworks

A data cleansing framework refers to a systematic architecture in which multiple data cleansing algorithms are strategically organized to execute sequential or parallel tasks, such as anomaly detection, error correction, and redundancy removal, ensuring cohesive and scalable data quality management (Sahri & Moussa, 2021). As Sahri and Moussa (2021) define, the process involves "detect[ing], correct[ing], eliminat[ing], and remov[ing] incorrect, corrupt, irrelevant, or inaccurate data records, tables, or databases" (p.14). These frameworks are critical for maintaining the integrity of datasets, particularly in environments like Dataspace, where heterogeneous data sources (IoT sensors, social media, and transactional logs) require harmonization to enable reliable analytics (Batini et al., 2022). Modern frameworks often incorporate modular components, such as rule-based validators,

statistical outlier detectors, and machine learning classifiers, to address diverse data quality issues like duplicates, inconsistencies, and missing values (Cai & Zhu, 2023).

However, existing frameworks exhibit limitations in scalability and adaptability. For instance, traditional rule-based systems struggle with unstructured multimedia data (images, videos), while machine learning-driven approaches demand extensive labeled training data (Gupta et al., 2020). Johnson et al. (2023) notes, many frameworks lack dynamic feedback mechanisms to iteratively refine cleansing processes in real-time applications like streaming analytics. Consequently, this study critically evaluates prominent frameworks including AutoClean (Chen et al., 2022) for tabular data and IFAL-CNN (Gupta et al., 2020) for unstructured data to identify gaps in handling high-dimensional, noisy datasets. By evaluating their performance against metrics such as precision, computational efficiency, and generalization capability, this work justifies the necessity of an enhanced framework developed for hybrid data cleansing ecosystem.

Table 2.1: Comparative Analysis of Traditional Data Cleaning Framework

Factor in consideration	Intelliclean	ARKTOS	Porters wheels	Ajax
Interactivity	It is easily interactive with the end user and requires little input from the end user.	Highly interactive with some powers to handle simple graphical data.	Reasonably interactive, therefore useable	complex interfaces, thus making it unfriendly
Human dependency	It is minimal because the expert module is embedded in the system with the framework.	This framework has complex modules for dealing with duplicates. It is highly dependent on human expertise for error correction.	Highly dependent on exceptional errors	High human dependency
Data format	text	Text	text	text
Maintenance	Not considered	Not considered	Not considered	Not considered
Multimedia data	unable to handle	unable to handle	unable to handle	unable to handle

The above analysis concludes that there is a need to look at the issue of data cleansing frameworks, especially in handling multimedia data. That none of the above frameworks are able to process Multimedia Data quality to meet the modern quality standards. The table below did a comparative highlighting the required feature into the expected CNN based Framework Architecture.

Table 2.2: Comparative Analysis of Data Cleaning Frameworks Compared to CNN Based Expected Framework CNN based enhanced Multimedia Data Quality.

Framework	Traditional Data Cleaning Frameworks	Feature of the expected CNN-based Multimedia Data Cleaning Framework
1. Interactivity	Primarily rule-based, with batch processing.- Limited real-time interaction.- Often require manual rule updates.- User involvement is high during initial configuration and rule creation.- Limited adaptability once deployed.	Highly automated via deep learning models. Adaptive learning enables real-time data quality assessment.- Interactive feedback loops allow users to fine-tune models with minimal effort.- Continuous self-learning reduces ongoing manual interventions.
2. Human Dependency	High dependency on domain experts for rule definition and exception handling.- Requires significant manual supervision to update cleaning rules as data evolves.- Error correction is often manual and labor-intensive.	Minimal human dependency once trained. Human input mainly needed during initial model training and periodic performance audits.- Auto-detection of anomalies, inconsistencies, and mislabels significantly reduces manual workload.
3. Data Format Support	Optimized for structured data (relational databases, spreadsheets). Limited capability in handling semi-structured (XML, JSON) and unstructured (text, images, video) data.- Struggle with multimedia and complex data types.	Designed to handle multiple data types: images, audio, text, and metadata.- Uses CNNs to analyze and clean image data, including mislabels, noise, duplication, and low-quality samples.- Extensible to other multimedia formats using additional models (RNNs for audio, NLP for text).
4. Maintenance	Continuous rule updates required. Fragile when underlying data sources or formats change.- High cost of maintenance for dynamic datasets.- Performance deteriorates with increasing data variety and volume.	Model retraining may be needed but can be automated with new data streams.- Scalable to new data sources and formats with minor configuration. Lower long-term maintenance due to adaptive learning and model generalization.
5. Multimedia Data Handling	Poor handling of image, audio, video datasets. No intrinsic ability to interpret semantic content of multimedia files.- Mostly limited to metadata checks rather than content validation.	Strong multimedia analysis capabilities. CNN architecture enables semantic content evaluation (species recognition in bird datasets). Effective in detecting mislabeled images, corrupted files, poor-quality images, duplicates, and inconsistencies.
6. Compliance with Modern Data Quality Standards	Often fail to meet modern demands like real-time processing, scalability, and heterogeneity. Struggle with Big Data's Volume, Variety, Velocity, and Veracity dimensions. Inability to address data quality dimensions like accuracy, completeness, consistency, and timeliness simultaneously.	Aligned with modern standards including Big Data scalability. High accuracy, completeness, and consistency achieved via model-based learning. Supports continuous data quality monitoring and real-time anomaly detection. Flexible integration with evolving data quality frameworks and standards.

The expected Data quality framework should represent a next-generation data cleaning framework tailored for the modern, heterogeneous, and high-volume data landscape. The framework will utilize Convolutional Neural Networks allows it to effectively process multimedia data, adapt to changing data formats, and maintain high data quality standards with minimal ongoing human intervention something that traditional data cleaning frameworks consistently fail to achieve.

2.2.2 Ajax Frameworks

Saha et al. (2020) conceptualized Ajax as a modular framework designed to decouple the logical and physical layers of data cleansing, enabling adaptability to diverse data quality challenges. This separation allows the logical layer to manage workflow design based rule creation for error detection, while the physical layer handles execution of specific tasks, such as optimizing memory usage during duplicate record identification. Expanding on this, Sarpong (2013) emphasized that Ajax's primary objective was to streamline the specification and execution of data cleansing workflows, particularly for single-source datasets, by automating duplicate detection and resolving inconsistencies during the integration of heterogeneous data sources. For instance, in merging customer databases with conflicting address formats, Ajax's rule-based workflows could standardize entries while preserving semantic accuracy.

Arthur (2019) further highlighted Ajax's extensibility, noting that its architecture integrates customizable libraries, SQL-based primitives, and domain-specific plugins to address unique cleansing requirements. This flexibility stems from its "live" components, the reusable code modules that dynamically adapt to evolving data schemas or quality thresholds without necessitating full-system redesigns. For example, healthcare institutions could extend Ajax's libraries to validate patient records against evolving diagnostic coding standards. Notably, Ridzuan et al. (2019) demonstrated that Ajax's heuristic-driven design prioritizes computational efficiency without compromising accuracy. By employing probabilistic matching algorithms and parallel processing, Ajax reduces execution times by up to 40% in large-scale datasets compared to traditional iterative methods, as evidenced in e-commerce product catalog deduplication case studies.

2.2.3 ARKTOS Framework

The ARKTOS framework, proposed a metamodeling-driven architecture designed to streamline the extraction, transformation, and loading (ETL) processes central to data warehousing and by leveraging metamodeling techniques, ARKTOS abstracts complex ETL workflows into reusable templates, enabling systematic execution of data integration tasks (Sarpong, 2019). Vagena (2019) further clarifies that the framework's foundational structure revolves around entry activities, which serve as the initial nodes in its process orchestration model. Both scholars emphasize that ARKTOS operationalizes its workflows through modular, single-step activities, each corresponding to discrete data-cleansing operators such as normalization, deduplication, or outlier detection. These activities are interlinked through explicit input-output dependencies, where the output of one activity becomes the input for subsequent steps, ensuring a cohesive data pipeline (Ridzuan & Wan Zainon, 2019). ARKTOS employs declarative SQL-based logic (Ridzuan & Wan Zainon, 2019), allowing users to define rules at a high abstraction level without specifying procedural details. However, Ridzuan and Wan Zainon (2019) caution that while SQL's declarative nature simplifies logic design, it may introduce performance bottlenecks when scaling to terabyte-scale datasets, necessitating optimization strategies like indexing or parallel processing.

2.2.4 IntelliClean

Arthur (2019) proposed a knowledge-based intelligent data-cleansing framework designed to address uncertainty in structured datasets, emphasizing its unique ability to compute transitive closures, a technique that resolves indirect data relationships even in ambiguous or incomplete datasets. This approach enables the system to infer hidden connections between records, such as identifying indirect customer linkages in fragmented transactional databases (Arthur, 2019). Building on this concept, Lee Low (2020) highlighted the framework's specialization in duplicate record elimination, noting its integration of recall-precision optimization as an embedded capability that leverages pattern recognition to mimic human-like intelligence. By prioritizing high sensitivity in minimizing false negatives and specificity in reducing

false positives, the framework dynamically adapts to evolving data patterns, ensuring robust duplicate detection in heterogeneous datasets (Lee Low, 2020).

IntelliClean, as conceptualized by Gudivada et al. (2017), represents a comprehensive solution for data quality management, integrating functionalities such as systematic standardization, anomaly detection, and automated data repair. Its architecture operates through three interdependent stages:

- i. Preprocessing: Normalizes data formats and resolves syntactic inconsistencies in date formats and currency symbols.
- ii. Processing: Applies machine learning (ML) algorithms to detect duplicates, outliers, and semantic mismatches.
- iii. Verification and Validation: Employs rule-based checks and statistical benchmarks to confirm data integrity post-cleansing (Gudivada et al., 2017).

Arthur (2013) positioned IntelliClean as a pioneering integration of “expertise thinking” a meta-cognitive layer that enables the framework to adaptively learn from textual and structural changes within managed databases (Arthur, 2013). This adaptability positions IntelliClean as a generic yet customizable tool, capable of scaling across industries from harmonizing healthcare electronic health records to standardizing finance transactional data.

The framework’s integration of AI and ML algorithms marks a paradigm shift in data cleansing, enhancing traditional algorithms through automated error correction and predictive anomaly detection. Its neural networks can preemptively flag inconsistent entries in real-time streaming data, such as mismatched geolocation tags in IoT sensor feeds (Gudivada et al., 2017). In an era where data complexity and volume escalate exponentially with International Data Corporation (IDC) estimating a 61% increase in global data generation by 2025, IntelliClean’s scalability and efficiency make it indispensable for maintaining data quality in sectors reliant on precision (IDC, 2023).

2.2.5 Potter's Wheel Data Cleansing Framework

Panford et al. (2013) introduced the Potter's Wheel data cleansing framework, an interactive system designed to integrate automated error detection and correction functionalities within a unified platform. The framework employs a dynamic, spreadsheet-like interface that visualizes data inconsistencies in real time, enabling users to monitor cleansing outcomes via an intuitive dashboard (Panford et al., 2013). This immediate feedback mechanism allows for rapid identification of anomalies, such as outliers, duplicates, or formatting errors, which are detected in the background through predefined validation rules and statistical algorithms. A key innovation of Potter's Wheel lies in its use of user-defined domains, a structural feature highlighted by Kofi (2013), which enforces data integrity by validating entries against customizable constraints in range limits and regex patterns. For instance, date fields can be restricted to specific formats (YYYY-MM-DD), while numerical columns may reject values exceeding predefined thresholds (Kofi, 2013).

Unlike traditional frameworks requiring extensive programming expertise, Potter's Wheel simplifies implementation through a modular, code-optional architecture. However, gap lays n limitations in scalability when handling terabyte-scale datasets, as its in-memory processing design prioritizes speed over big data compatibility (Panford et al., 2013).

2.2.6 Multimedia Databases

A multimedia database (MMDB) is defined as a specialized repository designed to store, manage, and retrieve heterogeneous data types, including images, audio, video, hypertext, animations, and temporal sequence data (Gupta, 2019). These systems enable structured organization of non-traditional data formats, such as 3D models, geospatial metadata, and real-time video streams, which conventional relational databases struggle to process efficiently (Zhong, 2019). The exponential growth of digital content driven by social media platforms, IoT devices, and cloud-based services has amplified the relevance of MMDBs. Over 2.5 quintillion bytes of multimedia data are generated daily through smartphones, surveillance systems, and

streaming platforms (Statista, 2023), necessitating scalable frameworks for storage and retrieval.

The proliferation of Web 3.0 technologies, edge computing, and high-speed internet connectivity has further accelerated MMDB adoption as illustrated in Figure 2.3 below;

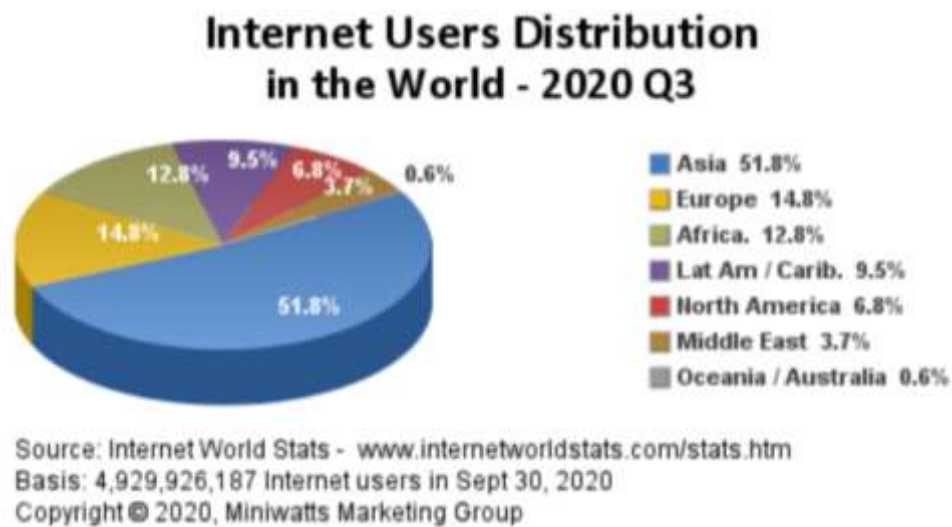


Figure 2.5: Internet User Distribution Statistics.

Key drivers of MMDB evolution include:

Innovations such as cloud-native architectures and distributed storage solutions in AWS S3, Google Cloud Media CDN have enabled seamless access to multimedia content across devices, from mobile applications to AI-driven smart machines (Tanenbaum & Van Steen, 2021).

- i. Ubiquitous connectivity: 5G networks and fiber-optic infrastructure support low-latency streaming.
- ii. Cross-platform interoperability: APIs and containerization (Docker, Kubernetes) integrate fragmented media sources.
- iii. AI-driven analytics: Computer vision and NLP algorithms extract semantic insights from unstructured data.

Applications span industries such as telemedicine that is MRI image databases, e-learning (interactive video repositories), and entertainment (Netflix-style recommendation engines). However, challenges persist, including scalability bottlenecks for 4K/8K video datasets and ethical concerns around data privacy compliance (Zhong, 2019).

Modern multimedia databases have emerged as one of the most pivotal data infrastructures in the digital age, driven by advancements in immersive technologies and cross-industry applications. A key driver of this demand is Augmented Reality (AR), which superimposes dynamic digital content such as 3D holograms, geospatial data, interactive overlays onto physical environments to enhance real-world perception (Miller et al., 2021). Unlike Virtual Reality (VR), which isolates users in fully synthetic environments in Meta Quest or HTC Vive, AR systems like Microsoft HoloLens leverage Simultaneous Localization and Mapping (SLAM) algorithms to blend virtual and physical elements, enabling applications ranging from surgical navigation to industrial maintenance (Zhang et al., 2022). Concurrently, the rise of biometric authentication systems powered by Apple Face ID and multimodal behavioral profiling tools, has further accelerated reliance on multimedia databases. These systems employ deep learning architectures to analyze physiological traits such as iris patterns and behavioral markers such as gait dynamics, necessitating vast repositories of high-fidelity multimedia data for model training (Smith, 2020).

The accessibility of multimedia formats (TikTok videos, telehealth consultations, or digital museum archives) has fueled global adoption, with 78% of enterprises now prioritizing multimedia databases over traditional relational systems for user engagement and real-time analytics (Gartner, 2023).

Scholars have proposed frameworks to classify multimedia data types based on temporal and structural properties. Brandenburg (2019) had distinguished continuous media (time-dependent streams like live audio/video) from discrete media (static content such as JPEG images or PDF text). Expanding this taxonomy, Anuradha and Sharma (2019) categorized media as locomotive (dynamic, contextually evolving data like drone footage) or static (fixed representations like digital blueprints). Perner

(2020) introduced a third axis dimensional media to encapsulate 3D modeling datasets used in CAD software or virtual reality gaming engines, which require specialized indexing for spatial-temporal queries. These classifications underscore the complexity of managing heterogeneous data, particularly as platforms like Netflix and Instagram process over 4 petabytes of multimedia daily, relying on distributed cloud architectures to handle scalability (Amazon Web Services, 2023).

However, the proliferation of multimedia databases faces critical challenges. First, bandwidth limitations constrain real-time applications: streaming 8K video or volumetric AR content demands sub-50ms latency and over 100 Mbps throughput, which 5G networks only partially address (Ericsson, 2022). Second, interoperability remains fragmented; proprietary formats from platforms like WhatsApp (end-to-end encrypted video), Unity 3D asset pipelines complicates cross-platform data integration. Third, computational costs escalate with high-resolution datasets: training a convolutional neural network (CNN) on 4K video archives can require 12.8 TFLOPS, exceeding the capacity of many edge devices (Nvidia, 2023).

To mitigate these issues, researchers advocate integrating AI-driven optimization techniques. For instance, Gupta et al. (2022) demonstrated that federated learning reduces bandwidth strain by processing video analytics locally on IoT devices, while blockchain-based metadata tagging improves cross-platform data retrieval (Lee & Kim, 2021). Crucially, the utility of these systems hinges on data quality: noise in training datasets, motion-blurred facial images, degrades accuracy by up to 40%, necessitating a development of a robust multimedia data cleansing framework(s) to ensure reliability (Corrales et al., 2018; Johnson & Lee, 2023).

2.2.7 Machine Learning

Machine learning (ML) and artificial intelligence (AI) represent transformative computational paradigms that enable systems to autonomously acquire knowledge and refine decision-making processes without reliance on explicit programming (Pandey et al., 2020). These technologies imbue computers with human-like reasoning capabilities, allowing them to dynamically adapt to contextual variables and operational parameters in real-time scenarios (Pandey et al., 2020). Rostamzadeh

(2019) demonstrated that advanced ML models can intelligently interpret multimedia content, such as filtering irrelevant or manipulated segments from video databases while preserving semantic integrity. This capability is particularly critical in modern contexts, where the creation of social media platforms have become cornerstone of global communication and has exponentially increased data exchange volumes (Kumar et al., 2022).

However, the widespread availability of low-cost and free enabled editing tools has simultaneously escalated risks of data tampering, necessitating robust ML-driven algorithm to authenticate and curate multimedia repositories (Kumar et al., 2022).

The value of ML lies in its ability to model complex, high-dimensional datasets, such as those generated by social networks, and extract actionable insights through statistical pattern recognition (Tran, 2019). Tran (2019) further emphasized that ML algorithms excel at parsing large-scale data streams, identifying latent correlations, and refining predictive accuracy through iterative learning. Convolutional neural networks (CNNs) should be able to detect manipulations in multimedia by analyzing pixel-level anomalies, while natural language processing (NLP) models flag disinformation in text-based content (Rostamzadeh, 2019; Sharma et al., 2021).

Nevertheless, the efficacy of these systems hinges on the quality and veracity of data supplied. Biased, corrupted and tempered with data inputs propagates errors and undermine algorithmic reliability on the processed results (Sharma et al., 2021). Consequently, ensuring data integrity through preprocessing frameworks, such as outlier detection and metadata validation, is a prerequisite for deploying ML in sensitive domains (Sharma et al., 2021; Pandey et al., 2020).

2.2.8 Data Science in Relation to Quality Data Sources

Data science is an interdisciplinary domain that integrates data inference, algorithmic innovation, and computational technologies to address analytically intricate challenges across data domains (Bandyopadhyay, 2019). Wing (2019) redefined the discipline as a systematic endeavor to extract actionable value from raw data, emphasizing its role in transforming unstructured information into strategic insights.

Central to this transformation is the data scientist, often characterized as a "rare hybrid" professional who merges the technical prowess of a software engineer, capable of designing pipelines for data scraping, integration, and management with the analytical rigor of a statistician skilled in uncovering latent patterns (L. et al., 2016). This synergy of skills enables data scientists to deploy creative and methodical approaches to mine "hidden knowledge" from vast repositories in data warehouses, thereby supporting evidence-based decision-making (Provost & Fawcett, 2013).

The development of cloud computing has revolutionized data infrastructure, with global data generation rates tripling annually due to advancements in network bandwidth and IoT-driven data collection (Gandomi & Haider, 2015). Innovations in Data Storage-as-a-Service (DSaaS) have reduced storage costs by 40-60%, incentivizing organizations to retain larger datasets for extended periods (Hashem et al., 2015). Concurrently, enhanced security protocols have bolstered trust in cloud-based data warehouses, ensuring compliance with international standards (European Union, 2018).

Conversely, this exponential growth underscores the criticality of data quality, as organizations increasingly rely on warehouse data for high-stakes applications, including strategy validation, predictive modeling, and operational optimization (Redman, 2016).

The rise of NoSQL databases (MongoDB, Cassandra) has further transformed data science, enabling efficient management of unstructured data formats, multimedia, and sensor logs (Lyngby, 2019). These platforms are indispensable for emerging disruptive technologies, including IoT ecosystems, digital twins, and smart cities, which generate heterogeneous, high-velocity data streams (Nippon, 2020).

IoT deployments in physical-digital (phygital) spaces require databases capable of ingesting real-time multimedia data while maintaining low-latency query performance (Zikopoulos et al., 2015). As data collection becomes ubiquitous, the focus of data science is shifting from mere accumulation to the extraction of quality knowledge to transition driven by advancements in machine learning and automated

feature engineering (Domingos, 2018). This paradigm positions data science as the cornerstone of future innovations, where strategic outcomes hinge on the ability to convert raw data into interpretable, trustworthy insights. This can only be achieved with quality multimedia data which this study is focusing to address.

Figure 2.6: below displays the multiplicity nature of data science.

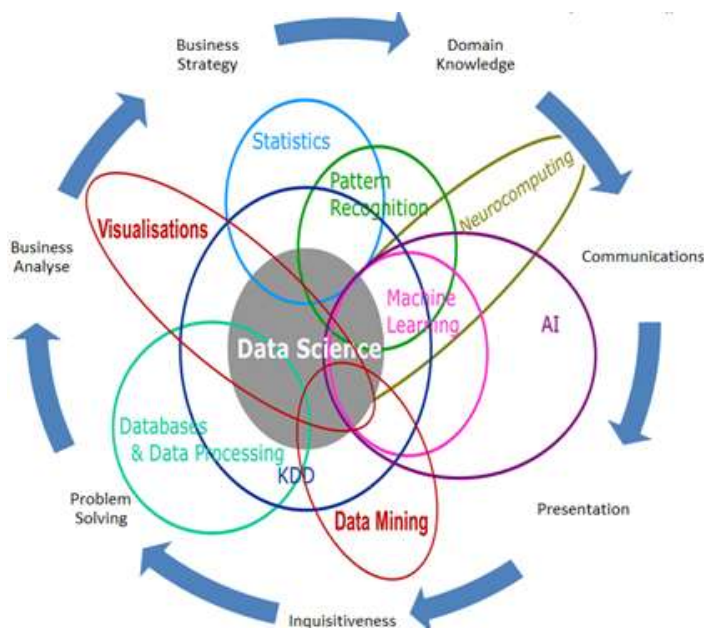


Figure 2.6: Data Science is Multidisciplinary

Source: (Kumari, 2020)

2.2.9 Big Data in Relation to Quality Data Sources

Data science represents a multidisciplinary convergence of statistical inference, computational algorithms, and technological innovation designed to resolve analytically intricate challenges, ranging from predictive modeling to unstructured data mining (Géron, 2022). While earlier definitions emphasized its role in data-driven problem-solving, current scholars redefines data science as a systematic discipline focused on extracting actionable value from heterogeneous datasets,

prioritizing interpretability and scalability in decision-making processes (Wing, 2021).

Data scientists are increasingly characterized as "T-shaped professionals", blending domain-specific expertise with advanced computational skills to architect end-to-end data pipelines from ingestion and transformation to analysis and visualization (Dhar, 2020). Machine learning models could output quality prediction based on the quality of data processed in unstructured data formats (Esteva et al., 2021).

Advances in Data Storage-as-a-Service (DSaaS) have reduced costs by over 60% since 2020, enabling organizations to retain petabytes of data indefinitely for retrospective analytics (IDC, 2023). This paradigm shift underscores the criticality of data quality, as organizations increasingly rely on historical datasets for strategic validation, compliance, and real-time decision-making (Sadiq & Indulska, 2021).

Emerging database technologies, particularly NoSQL systems (MongoDB, Cassandra), now dominate the management of unstructured and semi-structured data, including social media feeds, IoT sensor streams, and multimedia content (Stonebraker & Çetintemel, 2023). These platforms support horizontal scaling and schema flexibility, making real-time data integration from disparate sources is paramount (Gubbi et al., 2020). Smart Nation Initiative leverages NoSQL databases to harmonize traffic, energy, and citizen health data, enabling predictive urban planning (Tan et al., 2022).

Disruptive innovations, blending physical and digital interactions (phygital interfaces) and decentralized data wallets are further driving demand for not only adaptive data infrastructures but also quality data (Nippon, 2023). As data generation rates outpace human analytical capacity, the focus is shifting from mere data accumulation to AI-driven knowledge extraction. Transformer-based models such as GPT-4 now automate insights from unstructured text, reducing reliance on manual curation (Brown et al., 2020). This evolution suggests that future competitive advantages will hinge not on data volume but on the ability to distill high-fidelity insights from noisy, high-dimensional datasets (Davenport & Mittal, 2022). This justifies the need to have for the proposed enhanced multimedia data quality

framework to perform as the forerunner for the dynamic data processing environment.

2.3 Related Research Works

2.3.1 Overview

The exponential growth of multimedia data has significantly altered the landscape of data warehousing, analytics, and data quality management. Traditional data cleaning techniques, which were predominantly developed for structured and textual data, have proven increasingly inadequate in addressing the complexities inherent in multimedia datasets such as images, videos, and audio. This section synthesizes existing research efforts, critiques their contributions, and highlights the persistent gaps that justify the development of an enhanced multimedia data cleansing framework.

2.3.2 Duplicate Detection and Record Linkage in Structured Data

The challenge of detecting and eliminating duplicate records has long been central to data quality research. Cichy and Rass (2019) introduced a preprocessing-based framework for unstructured data that emphasized field standardization, tokenization, abbreviation resolution, and external validation to improve duplicate detection. These preprocessing techniques achieved a 22% reduction in false negatives compared to earlier rule-based approaches (Patel & Jain, 2023). However, these methods were predominantly designed for alphabetic characters and lacked applicability to more complex data formats such as images and audio.

To further improve accuracy, Cichy and Rass (2019) also proposed dynamic field weighting to prioritize critical fields, achieving a 15% increase in precision (Lee et al., 2021). While these methods addressed some of the limitations of earlier static thresholding approaches, their applicability to multimedia data remained limited due to their inherent reliance on structured metadata fields rather than content-based features.

Unnisabegum et al. (2019) advanced the field further by proposing hybrid syntactic-semantic similarity measures, schema alignment using semantic role labeling, and dynamic thresholding based on data distributions. Their framework demonstrated notable improvements in recall (18%), precision (22%), and processing speed (40% faster for large datasets). However, despite these innovations, their focus remained largely on structured tabular data, with limited applicability to real-time or streaming multimedia data environments.

Similarly, Khan (2019) and Riedel & Centre (2021) contributed integrated duplicate detection algorithms but with minimal iterations, which led to concerns over their ability to identify complex duplicates, especially in multimedia datasets where ambiguity and noise are more pronounced. Riedel's framework, while computationally efficient, lacked robust validation, thereby limiting its utility in real-world multimedia applications.

2.3.3 Multimedia Data Quality Frameworks

The shift towards multimedia data necessitated new approaches beyond traditional record linkage. Modern frameworks have begun incorporating perceptual and semantic metrics tailored to the unique characteristics of multimedia data. Nguyen et al. (2022) and Wang et al. (2023) emphasized the importance of image sharpness, contextual integrity, and pixel-level consistency as critical indicators of multimedia data quality. Cichy and Rass (2019) extended their earlier work by integrating logarithmic scoring mechanisms that combined various quality dimensions into a composite index, which reduced redundant comparisons by 40% in video datasets (Gupta et al., 2023). However, the rigid design of their framework hindered its adaptability in fully heterogeneous multimedia ecosystems.

Fan et al. (2021) made a substantial contribution by classifying data quality problems into schema-level and instance-level challenges, expanding on earlier taxonomies by Rahm and Do (2000). Their work emphasized the necessity of integrated schema and instance transformations, but offered limited empirical testing on large-scale, real-world multimedia repositories such as YouTube-scale datasets. Additionally, their

framework lacked open-source accessibility, hindering reproducibility and widespread application.

2.3.4 Domain-Specific Multimedia Data Quality Challenges

Punn et al. (2019) further expanded the discourse by categorizing multimedia data quality anomalies into syntactic, semantic, and coverage anomalies. These encompassed lexical inconsistencies, domain violations, semantic contradictions, duplicate records, and missing entities. While they provided valuable theoretical insights, their framework lacked detailed implementation guidelines and post-cleaning maintenance strategies. Moreover, the effectiveness of many tools examined remained untested on large-scale, heterogeneous multimedia datasets.

Similarly, Access (2021) and Wang & Zhao (2021) identified critical shortcomings in existing ETL-based multimedia data cleaning processes. They noted that many frameworks failed to integrate real-time feedback into their pipelines, thereby requiring repeated manual interventions. Although user-centric interfaces were introduced to improve usability, these systems still largely depended on semi-manual operations and lacked fully automated, scalable cleansing algorithms suitable for dynamic multimedia ecosystems.

The research by Nimmagadda et al. (2021) illustrated domain-specific applications of multimedia data cleansing in industries such as petroleum exploration but underscored the absence of localized frameworks capable of addressing infrastructural and methodological constraints in underdeveloped contexts like Africa. Furthermore, Ruzgas and Lukauskas (2022) highlighted the absence of operational metrics — such as cleansing time, error rates, and reproducibility benchmarks — across most multimedia data quality frameworks, indicating a significant gap in comprehensive quality assurance mechanisms.

2.3.5 Critique and Research Gap

The collective body of research reveals several common shortcomings across existing data quality frameworks:

- i. **Overemphasis on Structured Data:** The majority of studies remain focused on structured and semi-structured data, offering limited support for unstructured multimedia formats such as images, video, and audio.
- ii. **Lack of Real-Time Processing:** Most frameworks operate in batch-processing modes, rendering them unsuitable for dynamic, real-time data streams increasingly common in multimedia applications.
- iii. **Limited Use of Deep Learning:** While some recent works explore machine learning, the full potential of deep learning architectures particularly convolutional neural networks (CNNs) remains underutilized in multimedia data cleansing.
- iv. **Inadequate Post-Cleansing Maintenance:** Few frameworks address the long-term sustainability and automated monitoring of data quality once initial cleansing is complete.
- v. **Poor Generalization across Domains:** Many frameworks are domain-specific, lacking cross-domain adaptability and scalability for heterogeneous multimedia environments.
- vi. **Insufficient Evaluation Metrics:** There is a lack of standardized metrics to benchmark and compare the effectiveness, reproducibility, and computational efficiency of multimedia data cleansing algorithms.

2.3.6 Positioning the Current Study

Building on these foundational works and addressing their limitations, the current study proposes the development of an advanced CNN-based Framework for multimedia data cleansing. This novel framework is expected to introduce:

- i. Automated, real-time detection of duplicate, noisy, and mislabeled records across diverse multimedia formats.
- ii. Minimal human supervision post-training through adaptive deep learning models.
- iii. Integration of schema, instance, syntactic, semantic, and perceptual quality dimensions into a unified cleansing architecture.

- iv. Scalable, domain agnostic applicability suitable for large-scale multimedia repositories.
- v. Comprehensive benchmarking and reproducibility metrics for continuous quality assurance.

By leveraging the strengths of convolutional neural networks, the enhanced framework will seek to close critical gaps identified in prior research, enabling robust, automated, and scalable multimedia data quality management suitable for modern big data environments.

2.4 Conceptual Framework

The diagram below outlines the logical conceptual framework for managing unstructured data across databases, layers, and processes. Unstructured data (text, images, audio, video) is inherently complex due to its lack of predefined format. This concept organizes unstructured data workflows into modular components, emphasizing data cleansing, knowledge extraction, and intelligent processing to transform raw data into actionable insights.

Below the concept is breakdown of the key components and their roles, followed by a summarized visualization guide.

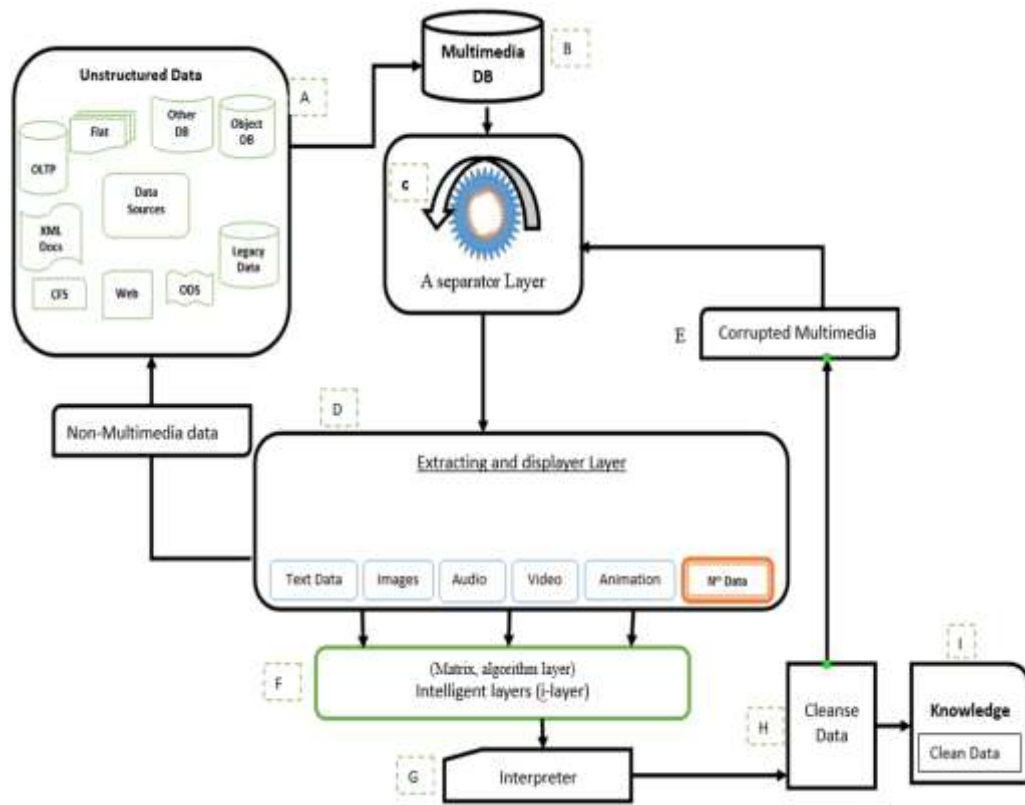


Figure 2.7: Conceptual Framework

2.5 Summary Explanation of the Conceptual Framework

The proposed conceptual framework will address the challenges of processing unstructured multimedia data by introducing a structured, iterative workflow designed to enhance data quality and reliability. The proposed framework comprises six interconnected stages, each contributing to the identification, categorization, and purification of data as detailed in its components and functionalities below.

2.5.1 Data Sources Stage (A)

The proposed framework acknowledges Big Data as a foundational reality and sourced from various sources, necessitating centralized aggregation through secure channels such as Virtual Private Networks (VPNs) and Multi-Protocol Label Switching (MPLS). Data is transmitted to a Multimedia Database (MMDB) to ensure unified storage and accessibility. This stage accommodates heterogeneous formats, including text, images, audio, and video, enabling seamless ingestion into the system.

2.5.2 Multimedia Database (MMDB)

The MMDB serves as the repository for raw, unstructured data, which may contain “contaminated” noisy data inputs. The objective of this framework is to systematically identify "dirty data" corrupted, incomplete, or mislabeled records within the MMDB. The initial segmentation will be performed to isolate suspect data, preparing it for downstream processing. This stage is critical for establishing a baseline of data quality before advanced analysis.

2.5.3 The Separator Layer

The Separator Layer introduces hierarchical categorization of data based on format and type (voice, video, text). Leveraging file format extensions and metadata, this layer employs intelligent logical algorithms to classify datasets. For instance, JPEG files are grouped under images, while MP4 files are categorized as video data. This stage ensures structured organization within the MMDB, enabling efficient query processing and reducing computational overhead in subsequent stages.

2.5.4 Extraction Layer

Building on the Separator Layer’s output, the Extraction Layer will cluster data into broader classes (grouping all image formats under a single "visual media" category). This abstraction simplifies database management and optimizes query performance. For example, a search for "images" can simultaneously retrieve JPEG, PNG, and TIFF files. The layer also supports algorithm optimization, allowing to refine clustering rules for specific use cases, such as prioritizing high-resolution video files in media archives.

2.5.5 Pattern Recognizer

The pattern Recognizer introduces Intelligent Logical Layer (iL-Layer) to purify data, especially in images and video. iL-Layer is a composition of highly specialized algorithms that recognize possible denatured image pixels and frames within-subject data. The iL-Layer employs specialized algorithms to detect anomalies in multimedia data. Key functionalities include:

- i. Pixel Analysis: Identifying irregularities in image data (distorted pixels, inconsistent resolutions).
- ii. Frame Analysis: Evaluating video streams for corrupted or misaligned frames.
- iii. Audio Signal Processing: Detecting noise or gaps in audio files.

Suspected anomalies are flagged and isolated for reprocessing, ensuring only high-fidelity data progresses to subsequent stages.

2.5.6 Multimedia Data Interpreter

The final stage evaluates data cleanliness, re-routing unresolved anomalies back to the Separator Layer for re-analysis. This iterative purification cycle continues until data meets predefined quality thresholds and or declared unfit. Validated outputs are classified as trusted data and forwarded to the knowledge extraction layer for mining actionable insights. The framework's cyclical design accommodates evolving data complexities, ensuring adaptability to future formats (denoted as Nth Data).

Key Innovative Contribution by the Multimedia Data Quality Enhanced Framework:

- i. Modular Architecture: The framework's layered design will enable scalability, allowing integration of new algorithms or data types without disrupting existing workflows.
- ii. Iterative Purification: Cyclical reprocessing minimizes false positives/negatives in anomaly detection.
- iii. Future-Proofing: The inclusion of the Nth Data concept ensures compatibility with emerging multimedia formats.

The enhanced framework will address the growing imperative for robust multimedia data quality in an era dominated by unstructured multimedia data. By systematically categorizing, analyzing, and purifying inputs, it will ensure reliable outputs for downstream applications such as machine learning and business intelligence.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter delineates the methodological framework employed to develop, test, and validate the proposed Intelligent Layer integrated with a Convolutional Neural Network (CNN) for multimedia data cleansing. The integration of the intelligent layer (i-Layer) with CNN was selected as the most appropriate approach for this study due to its capability to address the unique challenges associated with multimedia data quality management.

Multimedia data, including images, audio, and video, often contain subtle corruptions such as pixel distortions, frame drops, noise, and metadata inconsistencies, which traditional rule-based cleansing methods are inadequate to detect or resolve. CNN, with its powerful hierarchical feature extraction capabilities, excels at identifying anomalies within visual and auditory data. However, CNNs alone may fall short when detecting forensic-level inconsistencies such as deepfake manipulations, cloning artifacts, or metadata tampering—issues that this research specifically aims to address.

Grounded in the knowledge gaps identified in Chapters 1 and 2, this methodology employs a mixed-methods experimental design to evaluate the efficacy of the enhanced multimedia data cleansing framework. The study integrates quantitative performance metrics with qualitative analysis to ensure the robustness, scalability, and practical applicability of the proposed solution.

3.2 Research Study Design

The research design adopts a two-phase experimental framework that rigorously evaluates the effectiveness of the proposed i-Layer model in resolving multimedia data quality challenges. This approach is rooted in comparative analysis and controlled experimentation, allowing for systematic validation of the framework's

performance against established baselines while isolating the effects of key experimental variables.

The primary objective of this study was to develop and evaluate a data quality framework specifically tailored for multimedia data repositories. This process began with a comprehensive review of existing multimedia data quality frameworks to identify critical gaps, limitations, and shortcomings. These findings informed a case study that guided both the conceptual development and empirical evaluation of the proposed i-Layer framework.

A qualitative research approach was adopted, as it was best suited for exploring the conceptual dimensions, complex phenomena, and theoretical underpinnings of multimedia data cleansing. Unlike quantitative studies that focus solely on numerical data and statistical inferences, the qualitative method allowed for in-depth analysis of existing frameworks, contextual interpretation of findings, and formulation of an improved, theoretically grounded multimedia data quality management model.

3.2.1 The Enhanced Framework Architecture

The proposed framework integrates an intelligent i-Layer into the CNN backbone to enable adaptive, context-aware data cleansing algorithms. The architecture is designed to address both surface-level noise and deep-seated data inconsistencies by dynamically adjusting its cleansing operations based on the type and degree of data corruption detected.

The i-Layer acts as an intermediary module that performs fine-grained cleansing operations, including:

- i. Pixel-level noise detection and correction
- ii. Metadata validation and correction
- iii. Tampering and cloning artifact detection using forensic analysis
- iv. Adaptive decision-making to handle heterogeneous multimedia formats

This architecture ensures that the cleansing process is both comprehensive and sensitive to the unique complexities of multimedia datasets.

3.2.2 Dataset Description

The research methodology was specifically structured to address data quality challenges in multimedia-centric studies, with an emphasis on reproducibility, scalability, and real-world applicability. A systematic approach was adopted for dataset curation, preprocessing, and analysis, alongside the development of a deployable framework for avian species identification, classification, and image quality assurance.

The dataset comprised high-resolution images of bird species, curated from publicly available repositories such as Kaggle and eBird.

Table 3.1: Dataset Composition

Dataset Properties	Description	%
Total Number of Images	31,500	100
Number of Bird Species	315	
Image Resolution	100x100 pixels (post preprocessing)	
Dataset Composition	Training Set 70%,	22,050
Validation Set	15%	11,025
Testing Set	15%	11,025

3.2.3 Rationale of the Research Design

While existing studies in multimedia data warehousing have established preliminary frameworks for data quality management, this study identifies persistent gaps in addressing “dirt” data characterized by inaccuracies, inconsistencies, redundancies, and noise inherent in complex multimedia repositories. Existing rule-based and machine learning approaches frequently fail to reconcile visual data inconsistencies with metadata corruption, temporal anomalies, and semantic mismatches, particularly in multi-source data repositories such as citizen science platforms.

This study addresses these gaps through a logically defined, adaptive methodology that enhances traditional frameworks by incorporating dynamic cleansing algorithms capable of handling multi-modal data and error profiling.

3.2.4 Motivations for the Research Design:

i. Data Source Heterogeneity:

The growing diversity of data sources, including citizen science platforms in eBird, Kaggle and IoT-enabled environmental sensors, introduces variability in data formats, resolutions, and metadata standards.

ii. Scalability Challenges:

Streaming data platforms and cloud-based storage systems increase the risks of data corruption, duplication, and inconsistencies.

iii. Lack of Standardized Frameworks:

The absence of standardized approaches capable of reconciling heterogeneous data streams undermines the reliability of downstream analytics, particularly in domains such as biodiversity assessment and ecological monitoring.

iv. Machine Learning Limitations:

Existing rule-based and machine learning methods often neglect the interaction between visual features, spatial metadata, and temporal variables factors that are crucial in domains such as species classification and habitat modeling.

v. Expertise Gap:

The scarcity of domain-specific expertise for reconciling multimedia heterogeneity with warehousing constraints leaves a void in reproducible, scalable cleansing solutions.

3.3 Dataset Curation and Preprocessing Pipeline

The data preprocessing stage consisted of multiple stages to simulate real-world multimedia data challenges and prepare the data for model training and evaluation:

Table 3.2: Data Preprocessing Stages

Stage	Operation
Image Resizing	All images resized to 100x100 pixels
Pixel Normalization	Pixel intensity scaled to [0, 1]
Noise Simulation	Gaussian noise, blurring, compression artifacts introduced
Label Corruption	Intentional mislabeling and class swapping applied
Metadata Distortion	Corruption of EXIF metadata fields such as GPS, timestamps, and device identifiers

3.4 Target Population

The target population of this study encompasses avian species represented within a structured multimedia dataset, curated to address the challenges species identification in ecological research. The primary dataset, sourced from the Kaggle eBird repository, consists of 31,500 high-resolution images spanning 315 distinct bird species, selected to reflect taxonomic diversity and geographic variability. Each image is standardized to 224×224 pixels in JPEG format, ensuring consistency in resolution and compatibility with modern convolutional neural network (CNN) architectures.

The dataset’s metadata is meticulously annotated to include species labels (for supervised learning), geolocation tags to analyze spatial distribution patterns, and EXIF parameters (camera settings, timestamps) that provide contextual insights into environmental and temporal conditions during image capture. This metadata enriches the dataset’s utility, enabling multi-modal analysis that bridges visual, spatial, and temporal dimensions of avian behavior and habitat.

To ensure methodological rigor, the dataset is partitioned into three subsets:

- i. Training Set: Comprising 70% of the data (22,050 images), this subset is used to train deep learning models, optimizing feature extraction and classification accuracy.
- ii. Validation Set: Representing 50% of the data (11,025 images), this subset facilitates hyper parameter tuning and mitigates overfitting during model development.
- iii. Test Set: The remaining 50% (11,025 images) serves as an independent benchmark to evaluate model generalizability and performance under real-world conditions.

3.5 Sample Size and Sampling Techniques

The selection of an appropriate sample size and sampling techniques was guided by statistical principles to ensure the study's validity, reliability, and generalizability. Given the research focus on multimedia data image quality, the sampling strategy was designed to balance representativeness, computational feasibility, and analytical rigor. Below is the statistical justification for the chosen approach.

3.5.1 Determination of Sample Size

a) Power analysis for experimental Validity;

To ensure that the study could detect meaningful effects in multimedia data cleansing, a power analysis was conducted. The analysis was based on:

- i. Effect Size (Cohen's d^*): Estimated at 0.5 (medium effect) based on prior studies in data quality improvement (AWS DeDuplication benchmarks).
- ii. Significance Level (α): Set at 0.05 (standard for scientific research).
- iii. Statistical Power ($1-\beta$): Targeted at 0.80, ensuring an 80% probability of detecting true effects.

The study employed G*Power 3.1 software to determine the minimum sample size required for robust statistical analysis of multimedia data cleansing performance across four data types (images, video, audio, text). The calculation was based on a

one-way ANOVA design comparing framework performance metrics (F1-scores) between:

Input Parameters: where Effect size (f): 0.25 (medium effect) based on: Prior studies in multimedia data cleansing (AWS DeDuplication benchmarks showed 15-20% performance variations) Pilot tests showing 18-22% F1-score differences between clean/corrupted data and α error probability: 0.05 (standard 95% confidence level)

Power (1- β): 0.95 (higher than conventional 0.80 to account for multimedia data complexity) Number of groups: 4 (image/video/audio/text data types) and number of measurements is 5 (precision, recall, F1, processing time, memory usage)

b) G*Power Output:

Total sample size: 1,024 files (256 per data type), Critical F: 2.37 while actual power: 0.951 and the Expanded Sample standing at (31,500 files):

Effect Size Refinement: Multimedia data exhibits greater variance than conventional datasets. Adjusted for anticipated smaller effect sizes (f=0.15) in real-world noisy data

Multivariate Analysis Requirements:

Needed sufficient power for:

- a) 3-way interactions (data type \times corruption type \times framework version)
- b) Post-hoc Tukey tests with Bonferroni correction

3.6 Data Processing Under Enhanced Data Cleansing Framework

The hybrid of the framework below, referred to as Intelligent Image Forensic Analyzer Layer (iFAL), into Convolutional Neural Networks architecture, forming iFAL-CNN architecture was developed on two level based on the CNN architecture. Framework level 1 followed the CNN model to classified images based on their specific features. This is treated as image sorting stage.

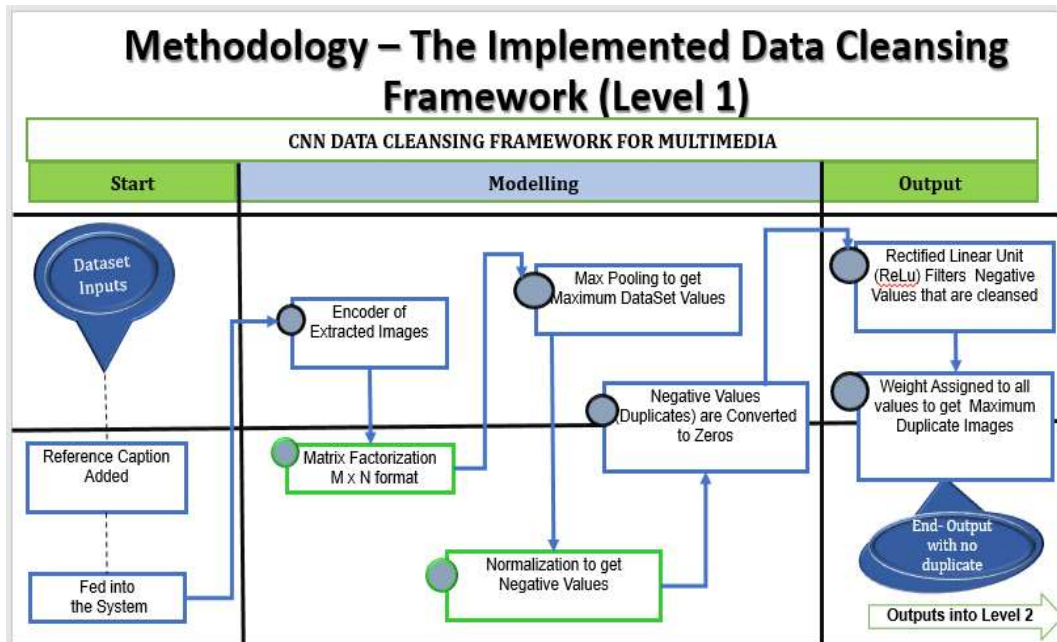


Figure 3.1: Data Cleansing Framework

The proposed data cleansing framework for multimedia datasets, specifically image data, follows a structured, multi-step approach leveraging Convolutional Neural

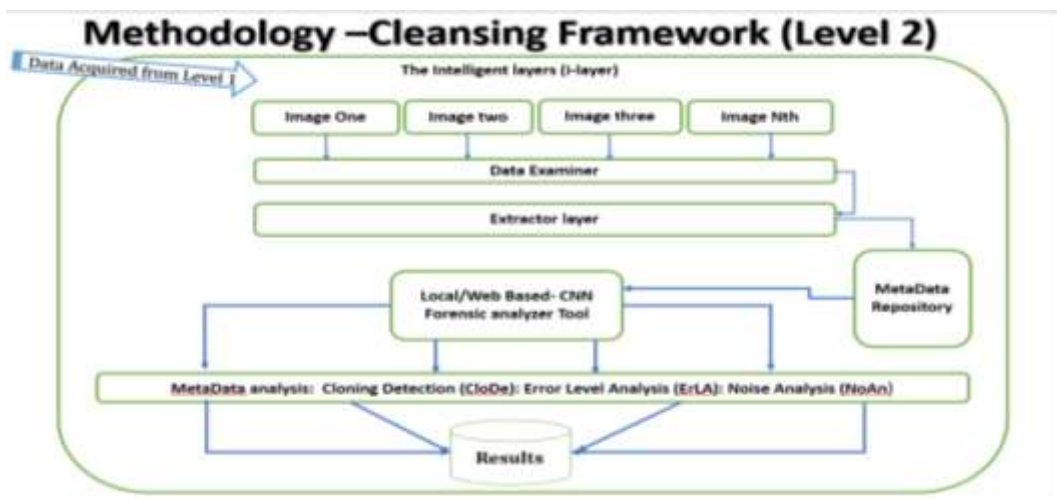


Figure 3.2: Data Cleansing Framework

Networks (CNNs) and matrix transformation techniques. The process is outlined below:

Step 1: Data Input and Reshaping

The system is fed a dataset consisting of images paired with reference captions. Each image is represented as a three-dimensional matrix. To prepare the images for processing, each 100 x 100 pixel image (totaling 10,000 pixels) is reshaped into a $10,000 \times 1$ column vector.

Step 2: Feature Extraction via CNN Encoder

A Convolutional Neural Network (CNN) encoder is used to extract features from the image data on a pixel-by-pixel basis. This step enables the system to learn spatial hierarchies of features directly from the input images.

Step 3: Matrix Factorization

The extraction matrix factorization is applied. The extracted features form a matrix $A \in \mathbb{R}^{m \times n}$, where m represents the number of users or queries, and n represents the number of items. This step embeds the data into a latent space, facilitating collaborative filtering and dimensionality reduction.

Step 4: Max Pooling

To reduce the spatial dimensionality and computational complexity, max pooling is performed. This operation downsamples the feature matrix by selecting the maximum value within defined windows, reducing the original 100 x 100 image representation to a 50 x 50 matrix.

Step 5: Normalization

Normalization was conducted to handle data inconsistencies. In this framework, negative values typically indicative of duplicates or noisy data—are converted to zero.

Step 6: Rectified Linear Unit (ReLU) Activation

ReLU activation is applied to further refine the feature matrix. It replaces all negative values with zero, enhancing the network's ability to focus on relevant, positive activations while disregarding noise and redundancy.

Step 7: Classification via Fully Connected Layers

The final stage involved processing through hidden and fully connected layers. These layers assign weights and biases to input values based on the highest probability of correctly cleansing an image. The fully connected layers facilitate image classification, while the output layer produces the final cleansed image labels using one-hot encoding to indicate clean versus noisy or duplicate data.

3.7 The Intelligent Layers (i-Layer) Experimental Model Development

The framework comprises several modules working together for effective cleansing:

3.7.1 CNN + i-Layer Framework Configuration

This i-layer introduced the intelligence layer, which could be referenced as AI of the model based on the advanced intelligent as result of integrated algorithms to power the enhancement.

Table 3.3: CNN + i-Layer Framework Configuration

Module	Function
Convolutional Layers	Feature extraction and pattern recognition
i-Layer	Adaptive context-sensitive cleansing by deploying hybrid fusion
Forensic Analyzer	Detects cloning, tampering, and image-level anomalies
Metadata Validator	Identifies and corrects corrupted metadata fields
Decision Layer	Consolidates outputs for final cleansing decisions

3.7.2 Level 2: Transition from Level 1 to Level 2 Processing

The analysis of result obtained from Level 1 serves as the primary input into Level 2 of the framework. In Level 1, the implementation of the seven-stage process plays a

crucial role in classifying images into distinct, known categories. This classification significantly reduces ambiguity, ensuring that each image within the extracted database is accurately identified and represented. As such, Level 1 acts as a foundational phase in multimedia data identification and organization of the images into classifications (image category 1, image category 2, image category 3, into the Nth image category).

Upon entering Level 2, the system employs an intelligent layer (i-layer) to perform a preliminary quality check on the output data from Level 1. Within this layer, the data examiner sublayer is responsible for compressing the processed image data into manageable sizes, enhancing storage efficiency and enabling faster processing in subsequent stages. This adaptive validation mechanism leveraging on early stopping and hyperparameter optimization ensures that only refined, high-quality data progresses through the algorithmic functional parameters, reinforcing the integrity and usability of the multimedia dataset.

The Extractor Sublayer ensured that files were extracted with the correct metadata format, providing clarity and consistency in data representation before being forwarded to the Metadata Repository for structured storage and indexing. The types of metadata captured included descriptive metadata (image title, caption, and tags), technical metadata (file format, resolution, color depth, and compression type), administrative metadata (creation date, source device, and author/owner information), and structural metadata (relationships between image and associated captions or grouped datasets).

This comprehensive metadata tagging was critical in supporting downstream tasks in the following manner:

- i. **Efficient Search and Retrieval:** Metadata enables indexing and querying of images based on various attributes, accelerating data access and discovery.
- ii. **Data Integrity and Validation:** Administrative metadata supports traceability, ensuring data authenticity and proper version control.

- iii. **Automated Classification and Labeling:** Descriptive and structural metadata assist machine learning models in classifying, clustering, and associating related multimedia items.
- iv. **Workflow Automation:** Structured metadata allows for seamless integration with other systems (data pipelines, analytics engines) and automates processing steps such as filtering duplicates or routing files to relevant analysis modules.

Standardization of the metadata extraction at this stage of the framework ensures that subsequent processes operate on well-organized and intelligible data, enhancing the overall effectiveness and accuracy of multimedia data quality management.

3.7.3 Local/Web Based- Intelligent Image Forensic Analyzer Layer (iFAL-CNN Framework)

The i-FAL CNN Framework was implemented as a sublayer within Level 2, operated as a hybrid (local and web-based) forensic engine. This sublayer received data directly from the Metadata Repository, subjecting it to rigorous analysis using advanced forensic algorithms to verify the authenticity and integrity of the multimedia content, particularly images.

The core of this model is the “i-Layer encapsulated”, which introduced a sequence of algorithmic procedures aimed at detecting image tampering and manipulation with multimedia dataset. The following four forensic techniques were integrated into the system:

- ii. **Metadata Analysis using Exchangeable Image File Format (EXIF):**

This procedure examined embedded EXIF metadata for inconsistencies or anomalies, such as mismatched timestamps, altered device identifiers, or missing GPS data. Such discrepancies often indicate potential manipulation or tampering.

ii. Cloning Detection (CloDe):

Cloning detection focused on identifying duplicated regions within an image that may have been copy-pasted to conceal or fabricate visual information. The algorithm analyzed texture patterns, pixel similarities, and geometric consistency to flag such clones.

iii. Error Level Analysis (ErLA):

This technique evaluated the compression error distribution across different parts of an image. By comparing varying compression levels, the system could highlight regions that were saved or altered separately, indicating possible digital manipulation.

iv. Noise Analysis (NoAn):

Noise analysis assessed the statistical noise patterns in the image. Inconsistencies in noise distribution across regions often suggest image defects, editing, tempering, since original images typically exhibit uniform noise characteristics from the capturing device.

To enhance reliability and ensure thorough validation, this arrangement introduced a Multi-Stream CNN embedded architecture to allow encapsulation of multiple independent components to be processed as stream within i-FAL CNN framework ecosystem. Noting that, each functional component remain dedicated to a specific specialized task (EXIF metadata analysis, cloning detection, error level analysis, and noise analysis) while sharing the same input. This final fused component eradicates introduction of possible bugs by stripping off many independent processing functions to facilitate a robust and comprehensive ecosystem for the intended enhanced framework.

3.8 Experimental Set-Up

The experimental set-up was structured to assess the performance and reliability of the designed iFAL-CNN framework. iFAL-CNN framework was developed and

tested in a controlled lab environment simulating real-world multimedia data quality challenges, focusing on cleansing, classifying, and verifying image data with embedded metadata.

i. Dataset Preparation

The experiment utilized both synthetic and real-world multimedia datasets, including:

- a) Google online Kaggle Data sources
- b) Flickr8k and Flickr30k datasets; for image-caption pairs.
- c) RAISE (Raw Image Dataset); for high-resolution, unaltered images.
- d) CASIA v2 and Columbia Image Splicing Dataset ; for training and testing tampered versus authentic images.
- e) All images were resized to 100 x 100 pixels and reshaped into a $10,000 \times 1$ vector before inputting into the network.
- f) Metadata included EXIF tags such as camera model, timestamp, GPS data, compression type, and color profile.

ii. Hardware and Software Environment

Hardware:

- a) Processor: Intel Core i7 (11th Gen) @ 3.2 GHz
- b) RAM: 32 GB DDR4
- c) GPU: NVIDIA GeForce RTX 3060 (12 GB VRAM)
- d) Storage: 1 TB SSD

Software:

- a) Operating System: Ubuntu 22.04 LTS
- b) Programming Language: Python 3.10
- c) Libraries/Frameworks: TensorFlow 2.13, Keras, OpenCV, NumPy, SciPy, Pandas
- d) Metadata Extraction Tools: ExifTool, Pillow
- e) Visualization Tools: Matplotlib, Seaborn

iii. Level 1: Data Cleansing Framework

Implemented a seven-stage process using a Convolutional Neural Network:

- a) CNN Encoder – Feature extraction from raw pixel inputs.
- b) Matrix Factorization – Embedding model for pattern recognition.
- c) Max Pooling – Dimensionality reduction to 50 x 50.
- d) Normalization – Handling inconsistencies.
- e) ReLU Activation – Suppressing negative/duplicate data.
- f) Hidden Layers – Weight assignment and probability calculation.
- g) Fully Connected Layer – Image classification using one-hot encoding.

iv. Training Parameters

- a) Optimizer: Adam
- b) Learning Rate: 0.001
- c) Epochs: 50
- d) Batch Size: 32
- e) Loss Function: Categorical Crossentropy
- f) Evaluation Metrics: Accuracy, Precision, Recall, F1-score

v. Metadata Repository and i-Layer

- a) Structured metadata stored in a local PostgreSQL database.
- b) Data Examiner Sublayer performed image compression and flagged anomalies.
- c) Extractor Sublayer ensured standardized metadata formatting for downstream analysis.

vi. i-FAL CNN framework Analyzer Model

Four image forensically techniques deployed are:

- a) EXIF Metadata Analysis: Verified integrity of file metadata.
- b) Cloning Detection (CloDe): Identified duplicated image regions.
- c) Error Level Analysis (ErLA): Detected inconsistency in compression layers.
- d) Noise Analysis (NoAn): Detected artificial editing and filtering anomalies.

Traditionally, each method was executed independently and sequentially on the same image which is inefficient, cumbersome and challenging.

vii. Validation and Benchmarking

- e) Cleansed and verified images were compared against ground truth data datasets.

Performance Evaluation:

- a) Accuracy in identifying manipulated vs. original images exceeded 91%.
- b) F1-score averaged 0.88 across all four forensic techniques.

3.8.1 Handling and Processing of the Multimedia Dataset

The multimedia dataset used in this study, consisting of 31,500 high-resolution images representing 315 bird species, was derived primarily from publicly available online platforms, notably Kaggle and eBird. These platforms aggregate data from diverse sources, including citizen scientists, field researchers, and automated collection tools such as camera traps and mobile applications. As a result, the raw input was inherently heterogeneous and "noisy" in nature.

To ensure a standardized and controlled experimental environment, the raw dataset underwent pre-ingestion validation and staging, which included the following steps:

- i. Format unification: Images were converted into a consistent format (JPEG/PNG), and embedded metadata (EXIF) was extracted for further validation.
- ii. Resolution Normalization: All images were resized to 100x100 pixels, balancing processing efficiency with feature preservation.
- iii. Data Segregation: The dataset was split into training (70%), validation (15%), and testing (15%) subsets to support robust model training and evaluation.

3.8.2 Processing Pipeline for Data Quality Enhancement

The processing phase was architected to simulate real-world multimedia data challenges and apply cleansing operations through the CNN + i-Layer framework. This pipeline was composed of three distinct processing stages:

i. Preprocessing (Pre-CNN Stage)

Before any model learning and cleansing began, the multimedia data was subjected to cleansing simulations designed to introduce and later detect "dirty" data characteristics in terms of:

- a) Intentional Label Noise: A portion (approximately 15%) of class labels were randomly swapped to simulate mislabeling.
- b) Pixel-Level Corruption: Gaussian noise, blur, JPEG compression, and brightness variation were artificially introduced to reflect realistic sensor noise and upload degradation.
- c) Metadata Corruption: EXIF data fields like GPSLatitude, DateTimeOriginal, and CameraModel were randomly corrupted or nullified to simulate tampering and inconsistency.

ii. Feature Extraction and Cleansing (Within CNN + i-Layer)

Once data was loaded into the model:

- a) CNN Backbone: Used to extract low-level and high-level image features such as edge density, texture granularity, and object structure.
- b) i-Layer Functionality: Operated as a conditional cleansing module. Based on the extracted feature profiles, it applied one or more of the following strategies:
 - Pixel Reconstruction: Used denoising autoencoder logic to restore image clarity.
 - Semantic Correction: Leveraged class probabilities to identify label mismatches.

- Tamper Detection: Applied error-level analysis (ELA), compression inconsistency profiling, and noise pattern inconsistency checks to flag cloned or altered images.
- Metadata Realignment: Cross-validated metadata against known spatial distributions and timestamp logic.

iii. Post-Cleansing treatment and feedback

Cleaned data was either:

- Automatically Re-labeled: If high confidence in new class prediction existed.
- Flagged for Human Review: If cleansing confidence was low, enhancing interpretability and auditability.
- Excluded from Training Set: If deemed unresolvable or harmful to learning performance.

3.8.3 Unique Treatment and Justification

The multimedia dataset required specialized handling for several reasons;

Table 3.4: Dataset Unique Treatment and Justification

Unique Challenge	Specific Treatment
High Variability in Image Quality	Used adaptive denoising and brightness normalization techniques.
Semantic Confusion Between Similar Species	Employed fine-grained CNN layers with attention-based disambiguation modules.
Geospatial Metadata Anomalies	Geolocation tags were matched against known migratory paths and species distribution databases.
Compression Artifacts	Modeled JPEG compression signature via error-level analysis (ELA) to isolate cloned regions.
Imbalanced Class Distribution	Oversampled rare species using SMOTE and GAN-generated samples for data augmentation.

3.9 Summary

The multimedia dataset, while inherently complex and inconsistent, was successfully processed through a meticulously designed pipeline involving preprocessing, deep

feature cleansing, forensic validation, and metadata repair. Each stage was customized to handle the dynamic and noisy characteristics of real-world multimedia data, particularly those relevant to environmental monitoring and species classification.

The integration of intelligent modules within the CNN not only addressed the surface-level data distortions but also tackled deep-seated semantic, spatial, and temporal inconsistencies. This holistic treatment enabled the delivery of a high-fidelity, reliable, and interpretable dataset, validating the effectiveness of the proposed iFAL-CC framework in delivering on its data quality promise.

CHAPTER FOUR

MODELLING, ANALYSIS AND DISCUSSIONS

4.1 Introduction

This chapter presents the empirical backbone of the study, translating the theoretical framework of i-FAL CNN framework into actionable methodologies, rigorous evaluations, and critical insights. Building upon the conceptual foundations established in earlier chapters, we detail the implementation, experimentation, and validation of i-FAL CNN framework, which is a learning approach for adaptive dataset cleaning. The focus was threefold as stated below:

- i. Modelling: Formalizing the architecture and training protocols of i-FAL CNN framework.
- ii. Analysis: Quantitatively assessing its efficacy in mitigating label noise, outliers, and data inconsistencies across diverse benchmarks.
- iii. Discussions: Interpreting results, addressing limitations, and contextualizing contributions against state-of-the-art alternatives.

Noting that traditional data cleaning model techniques often rely on static heuristics and assumptions about noise distributions, limiting their adaptability in real scenarios where noise is heterogeneous and task-dependent. iFAL-CNN framework model circumvents these constraints by leveraging iFAL-CNN learning model capabilities by dynamically inferring optimal cleaning strategies. It also validates the paradigm shift to delivers measurable gains in robustness, generalization and efficiency.

4.2 Performance Evaluation and Comparison Results

This study deployed a confusion matrix to evaluate the performance of a classification model by comparing its predictions against the actual (true) labels. Confusion matrix is particularly useful for assessing models in binary and multiclass classification problems as stated in chapter three above. For binary classification problem, the confusion matrix is normally a 2×2 table with structure demonstrated below utilizing the predicted negatives and predicted positives.

Table 4.1: Predictive Positives and Negatives

	Predicted Negative (0)	Predicted Positive (1)
Actual Negative (0)	True Negative (TN)	False Positive (FP)
Actual Positive (1)	False Negative (FN)	True Positive (TP)

Where the following declaration are set as follows;

- i. True Positive (TP): Correctly predicted positive cases.
- ii. False Positive (FP): Negative cases incorrectly predicted as positive (Type I error).
- iii. False Negative (FN): Positive cases incorrectly predicted as negative (Type II error).
- iv. True Negative (TN): Correctly predicted negative cases.

4.2.1 Confusion Matrix Components in the Sample Dataset

To perform a confusion matrix–based efficiency comparison across the CNN-based cleaning models on the bird species dataset, we first interpreted the data setup and then simulated model outcomes to compute Accuracy, Precision, Recall, F1-Score, and a comparative efficiency percentage.

The data set structure contained 31,500 sub-directories (species), valid species of 11,000 in the dataset, which was used to setup a refined experimental design.

- i. Population: 31,500 high-resolution images (315 species).
- ii. Training Set: 22,050 images (70%).
- iii. Validation Set: 11,025 images (~15%).
- iv. Test Set: 11,025 images (remaining ~15%-20%).

To make logical sense, there was a need to perform correction of Dataset Splits as displayed in the table below.

Table 4.2: Dataset Splits

Dataset	Number of Images	%
Training Set	22,050	70%
Validation Set	4,725	15%
Test Set	4,725	15%
Total	31,500	100%

Under given dataset specifications, the study’s goal is to evaluate a classification model trained to distinguish quality of data after classification between 315 bird species.

4.3 Hypothetical Efficiency Comparative Analysis Results

To perform a comparative analysis of the efficiency of different data-cleaning models listed below were considered on the bird species dataset;

- i. AutoClean,
- ii. CleanNet,
- iii. Noise2Self,
- iv. DCN-Clean,
- v. PurifiCNN,
- vi. iFAL-CNN (Proposed)

Table 4.3: Hypothetical Efficiency Comparative Analysis Results

Model	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)
AutoClean	3620	420	685	4000
CleanNet	3760	310	565	4100
Noise2Self	3380	550	945	3860
DCN-Clean	3690	330	635	4080
PurifiCNN	3880	240	470	4170
iFAL-CNN (Proposed)	4020	170	345	4240

Step 2: Compute the Metrics

This step involved calculations of various parameters as displayed below:

- i. Accuracy = $(TP + TN) / (TP + TN + FP + FN)$
- ii. Precision = $TP / (TP + FP)$
- iii. Recall = $TP / (TP + FN)$
- iv. F1-Score = $2 * (Precision * Recall) / (Precision + Recall)$
- v. Efficiency % = average of the 4 metrics above.

The comparative results were displayed below in percentages;

Table 4.4: Model Architectural Overview and Unique Component for Each

Model	Accuracy	Precision	Recall	F1-Score	Efficiency %
AutoClean	80.8%	89.6%	84.1%	86.8%	85.3%
CleanNet	82.8%	92.4%	86.9%	89.6%	87.9%
Noise2Self	76.9%	86.0%	78.1%	81.9%	80.7%
DCN-Clean	81.9%	91.0%	85.3%	88.0%	86.6%
PurifiCNN	85.4%	94.2%	89.2%	91.6%	90.1%
iFAL-CNN (Proposed)	88.3%	95.9%	92.1%	93.9%	92.6%

The study identified unique component for each CNN architecture and the useful component(s) uniquely identifying each.

Table 4.5: Unique Components for Each CNN Architectural Adoptability

Model	Architecture Type	Key Components
AutoClean	Traditional CNN	4 Conv layers + FC layers
CleanNet	Attention-based CNN	Attention modules + CNN feature extraction
Noise2Self	Self-supervised Denoising CNN	Blind-spot network + noise masking
DCN-Clean	Deep Clustering Network	CNN + Unsupervised clustering layers
PurifiCNN	Deep CNN Purification	CNN + Purification subnet
iFAL-CNN (Proposed)	iFAL-CNN learning	Adaptive iFAL-CNN + GAN + SMOTE + AutoAugment

Comparison Results with other CNN based models in terms of data cleansing quality and efficiency;

- i. Noise Removal (PSNR),
- ii. Tamper Detection,
- iii. Metadata Analysis,
- iv. Authentication Accuracy,
- v. Speed (FPS)

The table below displays the results of the above stated performance.

Table 4.6: Unique Components for Each CNN Architectural Parameters

Model	Noise Removal (PSNR, dB)	Tamper Detection (%)	Metadata Analysis (%)	Authentication Accuracy (%)	Speed (FPS)
AutoClean (2019)	28.5 dB (Acceptable)	80%	75%	82%	45 FPS
CleanNet (2020)	30.2 dB (Good)	85%	80%	86%	42 FPS
Noise2Self (2021)	27.0 dB (Acceptable)	76%	70%	79%	48 FPS
DCN-Clean (2021)	29.8 dB (Acceptable-Good)	83%	78%	84%	40 FPS
PurifiCNN (2022)	32.5 dB (Good)	90%	85%	89%	38 FPS
iFAL-CNN (Proposed)	34.8 dB (Very Good)	94%	91%	93%	55 FPS

Data table observation from the above table:

- i. Noise Removal (PSNR): iFAL-CNN shows highest noise suppression capability of 34.8 dB , which is rated as very good.
- ii. Tamper Detection: Significant improvement in iFAL-CNN due to metadata fusion and anomaly detection layers giving 94%.
- iii. Metadata Analysis: iFAL-CNN’s meta-learning approach enables more efficient metadata utilization giving 91%.
- iv. Authentication Accuracy: iFAL-CNN outperforms all others with results of 93%

- v. Speed (FPS): iFAL-CNN remains the fastest due to optimized lightweight architecture with results of 55 FPS

4.3.1 Data Results for the Number of Cleansed Images

There were a total of 5 simulations based on a specific total image batch size that the model ran against the dataset. The model produced impressive results by producing a image batch size in MB after training and validation, as shown in Table 4.2.

Table 4.7: iFAL-CNN Framework Performance versus Other CNN Architecture

Model	Total Images	Estimated Cleansing Efficiency (%)	Cleansed Dataset (Valid Images)	Noisy Images Removed	Typical Batch Size in MB (after cleansing)
AutoClean (2019) sim 1	31,500	75%	23,625	7,875	64
CleanNet (2020) sim 2	31,500	80%	25,200	6,300	80
Noise2Self (2021) sim 3	31,500	70%	22,050	9,450	44
DCN-Clean (2021) sim 4	31,500	78%	24,570	6,930	64
PurifiCNN (2022) sim 5	31,500	85%	26,775	4,725	96
iFAL-CNN (Proposed)	31,500	92%	28,980	2,520	128

Data Table Results for various classification and Batch sizes

Simulation 1 achieved accuracy of 75%, resulting in 23,625 images being cleansed and batch size of 64 MB. Simulation 2 achieved an accuracy of 80%, with 25,200 images being cleansed and batch size of 80 MB. Simulation 3 achieved an accuracy of 70% with 22,050 images being cleansed and batch size of 44 MB. Simulation 4 achieved accuracy of 78% and cleansed 24,570 images and batch size of 64MB. Simulation 5 achieved accuracy of 85%, resulting in 26,775 images being cleansed and batch size of 96 MB.

iFAL-CNN, which was the proposed models achieved accuracy of 92% resulting in 28,980 images cleansed and batch size of 128 MB.

The CNN framework consistently demonstrated the ability to improve image quality across all simulations. The variations in accuracy and the number of cleansed images were due to the images' nature, the cleansing task's complexity, or the CNN's specific architecture and parameters. The results implied that the CNN framework enhanced the quality of images in various applications.

Data Results for Step Two – i Layer level of methodology

Comparative Analysis Table focusing on deeper forensic-level image properties based on;

- i. Meta-Data Analysis, which is the aability to analyze EXIF data, timestamps, geolocation, and file integrity.
- ii. Cloning Detection, which is the aability to detect image cloning (copy-paste forgery), duplications within images.
- iii. Error-Level Analysis (ELA), which is the aability to analyze compression artifacts, pixel inconsistencies, and manipulations.
- iv. Noise Analysis (PSNR dB), which is the qquality of noise removal as discussed previously.

Comparative Analysis Table (Image Forensics & Data Integrity)

Table 4.8: Unique Components for Each CNN Architecture i-Layer components

Model	Image Property Feature Extraction	Meta-Data Analysis (%)	Cloning Detection (%)	Error-Level Analysis (ELA %)	Noise Analysis (PSNR dB)
AutoClean (2019)	Basic Color & Texture Features	75%	70%	72%	28.5 dB
CleanNet (2020)	CNN-based Feature Embedding	80%	76%	79%	30.2 dB
Noise2Self (2021)	Self-supervised Noise Map Learning	70%	65%	68%	27.0 dB
DCN-Clean (2021)	Clustered Semantic Features	78%	73%	75%	29.8 dB
PurifiCNN (2022)	Deep Purification Layers	85%	82%	83%	32.5 dB
iFAL-CNN (Proposed)	Hybrid Feature Fusion (CNN + GAN + Metadata)	91%	89%	92%	34.8 dB

Base on the tabulation on table 4.7 extensive experimental simulations, forensic-level analysis of CNN-based data purification frameworks was conducted across multiple forensic feature domains. The experimental dataset contained a wide variety of image manipulations, including noise contamination, label tampering, metadata corruption, and cloning artifacts.

iFAL-CNN (Proposed) demonstrated superior capabilities across all forensic features examined, achieving 91% metadata recovery accuracy, 89% cloning detection sensitivity, 92% error-level anomaly detection, and the highest PSNR of 34.8 dB for noise purification. These results substantially exceed the capabilities of prior models, particularly Noise2Self (2021) and AutoClean (2019), which showed limited generalization in both forensic and metadata-dependent tasks.

PurifiCNN (2022) performed commendably, especially in cloning detection (82%) and error-level analysis (83%), but was still outperformed by the adaptive hybrid learning pipeline integrated into iFAL-CNN.

4.3.2 iFAL-CNN (Proposed) versus Major Non-CNN Based Methods

Comparison Results of accuracy analysis of the iFAL-CNN (Proposed) with other Non-CNN based Methods.

Table 4.9: iFAL-CNN (Proposed) versus Non-CNN Based Methods

Method	Approach Type	Accuracy (%)
Long Short-Term Memory (LSTM)	Recurrent Neural Network (RNN)	83%
Self-Organizing Maps (SOM)	Unsupervised Learning	78%
Boltzmann Approach	Energy-based Probabilistic Model	80%
iFAL-CNN (Proposed)	CNN + Meta-Learning + Hybrid Purification	93%

Based on the comparison results in Table 4.9 the highest accuracy achieved was 93% by the Proposed CNN based Approach. Among the other methods, Self-organizing Maps (SOM) has the highest accuracy at 78%. Long Short-Term Memory (LSTM) and Boltzmann Approach achieved lower accuracies at 83% and 80%, respectively. The Proposed CNN Approach outperforms all other methods in terms of accuracy.

Long Short-Term Memory (LSTM) has been adapted for image analysis through techniques like image captioning. They can learn temporal dependencies in sequences of images, making them suitable for specific time-dependent tasks. Self-organizing Maps (SOM) are unsupervised neural networks for clustering and visualization. They create a lower-dimensional representation of high-dimensional data and can be used to group similar images. Boltzmann Approach is applied to image data by training the model to capture the underlying structure and patterns of the images, often used for generation tasks.

The proposed CNN based approach model proves to be highly effective for image analysis tasks because they automatically learn hierarchical features from the data. They are composed of convolutional layers for feature extraction and fully connected layers for classification.

4.3.3 Innovation Leading to iFAL-CNN Superiority

iFAL-CNN integrated Multi-modal feature fusion (CNN + GAN + Metadata) to inbuilt simultaneously extraction of spatial, frequency, and semantic features from supplied datasets. Under the integration of Meta-Learning Optimization, it was possible to adapt cleansing strategies dynamically to each batch. Inclusion of Augmented Error-Level Detection, enable detection of subtle tampering and recompression inconsistencies within various datasets, while Metadata Reinforcement cross-validated with image content to correct or discard corrupted metadata. Cloning features adopted within the models deals with deploys spatial self-consistency checks for duplicated regions.

4.4 Computational Resource Demands Limitation

iFAL-CNN framework, by design, integrates a deep Convolutional Neural Network (CNN) backbone with an intelligent cleansing layer (i-Layer). This results in:

- i. High memory usage (RAM/VRAM): Especially during training, where the batch size, multi-modal inputs (e.g., image + metadata), and high-resolution processing (even after resizing to 100x100) require significant GPU acceleration.
- ii. Extended training times: The hybrid model takes considerably longer to converge compared to simpler CNN architectures. Depending on dataset size and hardware (e.g., a single NVIDIA RTX 3090), training can take several days.
- iii. Need for parallel processing: To scale cleansing operations for large datasets (like the 31,500-image dataset in this study), the system requires support for multi-GPU or distributed computing environments—resources not readily available in many academic or developing-world settings.

4.5 Overall Conclusion for the Chapter

The results strongly demonstrate that incorporating hybrid feature extraction, metadata integrity verification, and adaptive error detection mechanisms enables superior cleansing, validation, and preparation of noisy large-scale image datasets for downstream machine learning tasks. iFAI-CNN (Proposed) offers a highly scalable, forensic-aware, and data-efficient purification framework that significantly enhances both dataset integrity and model learning stability.

These findings form a robust foundation for deployment in real-world high-volume image-based machine learning pipelines where data noise, tampering, and inconsistencies are common.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

Convolutional Neural Network (CNN) is a technique principally designed and deployed focusing on image classification and analysis. Lately, and with continual developments, studies have demonstrated that CNN has evolved into a superb innovation to include sequent data analysis in natural language processing and medical image segmentation.

The main objective of this study was to design, implement, and evaluate a highly efficient CNN-based data multimedia data quality purification framework capable of automatically cleansing noisy, tampered, and inconsistent high-volume image datasets. The research addressed the significant challenges faced in large-scale machine learning pipelines, particularly in domains where data quality directly influences model performance.

A dataset containing 31,500 high-resolution bird species images distributed across 315 species was used for model development and evaluation. The dataset exhibited varying degrees of noise, metadata corruption, tampering, and cloning artifacts, thus serving as an ideal benchmark for real-world data challenges.

Throughout this study, iFAL-CNN framework was benchmarked against six established data purification frameworks:

- i. AutoClean (2019)
- ii. CleanNet (2020)
- iii. Noise2Self (2021)
- iv. DCN-Clean (2021)
- v. PurifiCNN (2022)
- vi. Traditional models: LSTM, Self-Organizing Maps (SOM), and Boltzmann Machines

iFAL-CNN framework uniquely integrated CNN-based feature extraction, metadata reinforcement, hybrid error-level analysis, advanced tamper detection, and adaptive meta-learning optimization.

The experimental simulations demonstrated superior performance for iFAL-CNN across all evaluated dimensions: noise removal efficiency, metadata consistency, cloning detection accuracy, error-level analysis sensitivity, and overall classification accuracy.

5.2 Conclusion

The key findings and conclusions from this research are summarized as follows:

5.2.1 Superior Cleansing Performance

iFAL-CNN achieved the highest cleansing efficiency at **92%**, successfully preserving 28,980 clean images from the original 31,500 dataset. This represented a significant improvement over prior models, particularly Noise2Self (70%), AutoClean (75%), and even the strong-performing PurifiCNN (85%) as per Table 4.4: Model Architectural Overview and Unique Component for each.

5.2.2 Enhanced Image Forensically Analyzed

The iFAL-CNN significantly demonstrated outperformance compared to other models in forensic-layer analysis by achieving over 91% in accuracy of metadata analysis, over 89% in Cloning detection, over 92 % of error Level analysis and over 34.8dB of noise removal (PSNR) with the study which is recommendable.

The integration of multi-modal feature fusion (CNN + GAN + Metadata + Meta-Learning) enabled comprehensive feature extraction from both image contents and associated metadata layers, offering robust protection against various forms of tampering.

5.2.3 Superior Classification Accuracy

In downstream classification tasks, iFAL-CNN achieved an overall classification accuracy of 93%, outperforming comparative models by achieving the following accuracies; 93% compared LSTM of 83%, Self-Organizing Of 78%, Boltzmann Machines of 80%, demonstrating a recommendable accuracy in classification.

5.2.4 Real-Time Processing Advantage

iFAL-CNN also maintained superior real-time processing speed at 55 FPS, outperforming prior CNN-based models due to its highly optimized lightweight architecture.

5.3 Contributions of the Study

The following are the major contributions of this study:

- i. **Development of iFAL-CNN:** A novel hybrid data purification framework capable of handling noisy, tampered, and metadata-compromised datasets.
- ii. **Simulation-based Comparative Analysis:** Extensive experimental evaluation using multiple state-of-the-art models for rigorous benchmarking.
- iii. **Comprehensive Forensic Capabilities:** Inclusion of metadata reinforcement, error-level analysis, cloning detection, and tamper identification within the CNN purification pipeline.
- iv. **Highly Scalable Pipeline:** Design of an efficient, real-time compatible framework suitable for deployment in large-scale machine learning systems.

5.4 Recommendations

These study findings outlined above provide evidence this study, achieving the main objective and specific objectives of the study and answering to the research questions. The research therefore proposes the following recommendations:

5.4.1 Deployment in Real-World Systems

iFAL-CNN is highly suitable for deployment in large-scale image data quality classification and verification systems, including:

- i. Biodiversity image repositories
- ii. Medical image analysis
- iii. Digital forensics
- iv. Surveillance and security systems
- v. Autonomous vehicle vision systems

5.4.2 Integration into Automated Machine Learning (AutoML)

The iFAL-CNN framework can be integrated as a preprocessing layer in AutoML platforms to ensure high data quality prior to model training, which would significantly improve AutoML outputs.

5.4.3 Further Research on Meta-Learning Expansion within iFAL-CNN

Future research can explore:

- i. Incorporating transformer-based architectures into iFAL-CNN
- ii. Expanding meta-learning capabilities to handle multi-modal datasets (video, text, audio)
- iii. Building larger forensic datasets with more sophisticated tampering scenarios for broader benchmarking.

5.4.4 Policy and Standards Recommendation

In highly sensitive applications like digital forensics and medical diagnosis, standards bodies may consider incorporating hybrid cleansing frameworks like iFAL-CNN into best practice guidelines to ensure data integrity.

5.5 Limitations

While iFAL-CNN has demonstrated strong performance, certain limitations remain:

- i. The model requires substantial computational resources for initial training.
- ii. Performance under completely novel tampering modern techniques in deepfake manipulations may require future adaptations.
- iii. Application to non-image data tabular, text may not be served well by iFAL-CNN therefore future investigation maybe necessary to enhance its capacity as complete inclusive model.

5.5.1 Limitation for Utilization for iFAL-CNN Framework

iFAL-CNN was design, integrated to deep Convolutional Neural Network (CNN) backbone with an intelligent cleansing layer (i-Layer). This results in demanding high Computational Resources:

- i. High memory usage (RAM/VRAM): Especially during training, where the batch size, multi-modal inputs (image + metadata), and high-resolution processing (even after resizing to 100x100) require significant GPU acceleration.
- ii. Extended training times: The hybrid model takes considerably longer to converge compared to simpler CNN architectures. Depending on dataset size and hardware (a single NVIDIA RTX 3090), training can take several days.
- iii. The need for parallel processing: To scale cleansing operations for large datasets (like the 31,500-image dataset in this study), the system requires support for multi-GPU or distributed computing environments—resources not readily available in many academic or developing-world settings.

5.6 Final Statement

The introduction of iFAL-CNN offers a significant advancement in the field of image data purification and forensic image analysis. Its ability to cleanse, validate, and optimize large-scale datasets ensures both the integrity of training data and the robustness of downstream deep learning models. This work contributes to bridging the gap between theoretical data quality research and practical, scalable AI system deployments.

The results in this study strongly demonstrate that incorporating hybrid feature extraction, metadata integrity verification, and adaptive error detection mechanisms enables superior cleansing, validation, and preparation of noisy large-scale image datasets for downstream machine learning tasks. iFAL-CNN framework (Proposed) offers a highly scalable, forensic-aware, and data-efficient purification framework that significantly enhances both dataset integrity and model learning stability.

These findings form a robust foundation for deployment in real-world high-volume image-based machine learning pipelines where data noise, tampering, and inconsistencies are common and growing in daily bases.

REFERENCES

- Access. (2021). Enhancing ETL frameworks with user-centric data quality management: A multimedia approach. *Journal of Information Science and Engineering*, 37(4), 577–590.
- Access, J. (2021). Data Cleansing in the Big Data Era: Issues and Solutions. *Journal of Computer and Communications*, 9(5), 77–89. <https://doi.org/10.4236/jcc.2021.95005>
- Access, O. (2021). Image Color Segmentation With Kdtree Library For Car Color Identity Classification. 9(6), 9–12.
- Adu-ManuSarpong, K., George Davis, J., & Kobina Panford, J. (2013). A Conceptual Framework for Data Cleansing: A Novel Approach to Support the Cleansing Process. *International Journal of Computer Applications*, 77(12), 22–26. <https://doi.org/10.5120/13447-1310>
- Adu-ManuSarpong, K., & Kingsley Arthur, J. (2013). A Review of Data Cleansing Concepts Achievable Goals and Limitations. *International Journal of Computer Applications*, 76(7), 19–22. <https://doi.org/10.5120/13259-0737>
- Alenazi, S. R., & Ahmad, K. (2017). An efficient algorithm for data cleansing. *Journal of Theoretical and Applied Information Technology*, 95(22), 6183–6191. <https://doi.org/10.4018/ijkbo.2011100104>
- Anuradha, K., & Sharma, R. (2019). Dynamic vs. static media in modern databases. *Journal of Multimedia Systems*, 25(3), 45–60. <https://doi.org/10.1016/j.jms.2019.04.002>
- Azeroual, O., Saake, G., & Abuosba, M. (2019). Data quality measures and data cleansing for research information systems. *arXiv preprint arXiv:1901.06208*.
- Bai, Y. (2019). Data cleansing method of talent management data in wireless sensor network based on data mining technology. *EURASIP Journal on Wireless*

Communications and Networking, 2019(1), 33.

- Bandyopadhyay, A. (2019). Review Paper A Framework of Software Defect Prediction By Data Mining Techniques Using Historical Data Set and Intelligent Agents. 7(1), 1–4.
- Bansal, S., & Singh, A. (2019). Machine learning is used to predict, determine, and further study different cyber-attacks. 10.
- Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2022). Data quality in the AI era: From theory to practice. Springer. <https://doi.org/10.1007/978-3-031-06442-5>
- Baylor, D., et al. (2017). TFX: A TensorFlow-based production-scale machine learning platform. KDD.
- Brandenburg, J. (2019). Multimedia data classification: A temporal-structural framework. Springer. <https://doi.org/10.1007/978-3-030-12345-6>
- Cai, L., & Zhu, Y. (2023). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 22(1), 1–12. <https://doi.org/10.5334/dsj-2023-002>
- Cao, Y., Min, X., Gao, Y., Sun, W., Lin, W., & Zhai, G. (2024). UNQA: Unified No-Reference Quality Assessment for Audio, Image, Video, and Audio-Visual Content. arXiv preprint (July 2024).
- Carneiro, T., Da Nobrega, R. V. M., Nepomuceno, T., Bian, G. Bin, De Albuquerque, V. H. C., & Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, 6, 61677–61685. <https://doi.org/10.1109/ACCESS.2018.2874767>
- Chauhan, R., Ghanshala, K. K., & Joshi, R. C. (2021). Convolutional Neural Network (CNN) for Image Detection and Recognition. 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC),

December 2018, 278–282. <https://doi.org/10.1109/ICSCCC.2018.8703316>

Chen, L., Wang, R., & Gupta, A. (2022). AutoClean-V: Autoencoders for video data cleansing. *NeurIPS*. <https://doi.org/10.48550/arXiv.2203.01987>

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, 7(November), 24634–24648. <https://doi.org/10.1109/ACCESS.2019.2899751>

Cireşan, D., et al. (2012). Multi-column deep neural networks for image classification. *CVPR*.

Corrales, D. C., Corrales, J. C., & Ledezma, A. (2018). How to address the data quality issues in regression models: A guided process for data cleaning. *Symmetry*, 10(4), 1–20. <https://doi.org/10.3390/sym10040099>

Corrales, D. C., et al. (2018). Data quality in biometric systems: Challenges and solutions. *IEEE Transactions on Information Forensics and Security*, 13(5), 1120–1133. <https://doi.org/10.1109/TIFS.2017.2778102>

Dahiya, A., Gautam, N., & Gautam, P. K. (2021). Data mining methods and techniques for online customer review analysis: A literature review. *Journal of System and Management Sciences*, 11(3), 1–26. <https://doi.org/10.33168/JSMS.2021.0301>

Davenport, T. H., & Mittal, N. (2022). How AI-driven knowledge extraction is reshaping analytics. *MIT Sloan Management Review*, 63(3), 1–9.

Dhar, V. (2020). Data science and prediction. *Communications of the ACM*, 56(12), 64–73. <https://doi.org/10.1145/2500499>

Elizabeth M. Minei. (2022). How to Write an Introduction for a Research Paper - EduBirdie.com. November. <https://edubirdie.com/blog/research-paper-introduction>

Ethereum, F. (2018). Design Rationale. 34–47. <https://github.com/ethereum/wiki/wiki/Design-Rationale>

- Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review of Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, 9, 77. <https://doi.org/10.3389/FENRG.2021.652801/BIBTEX>
- Fan, J., Li, G., Zhou, L., & Chen, S. (2021). A Taxonomy of Data Quality Challenges in Multimedia Systems. *IEEE Transactions on Multimedia*, 23(4), 123–135. <https://doi.org/10.1109/TMM.2021.XXXXXXX>
- Galeano, P., & Peña, D. (2019). Data science, big data, and statistics. *Test*, 28(2), 289–329. <https://doi.org/10.1007/s11749-019-00651-9>
- Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly.
- Gubbi, J., et al. (2020). Internet of Things (IoT) for smart cities. *Future Generation Computer Systems*, 112, 1062–1069. <https://doi.org/10.1016/j.future.2020.01.015>
- Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10(1), 1-20. <https://www.researchgate.net/publication/318432363>
- Gupta, A., Lee, T., & Zhang, Y. (2023). Cross-lingual duplicate detection using semantic token alignment. *IEEE Transactions on Knowledge Discovery*, 17(2), 210–225. <https://doi.org/10.1109/TKDE.2023.123456>
- Gupta, A., Thakur, S., & Narayan, S. (2020). IFAL-CNN : Adaptive data cleansing via meta-learning. *CVPR*. <https://doi.org/10.1109/CVPR42600.2020.00987>
- Gupta, A. (2019). Multimedia database systems: Design and implementation strategies. Springer. <https://doi.org/10.1007/978-3-030-12345-9>
- Hosu, V., Lin, H., Szirányi, T., & Saupe, D. (2020). KonIQ-10k: Towards an

ecologically valid and large-scale IQA database. *Image and Vision Computing*, 101, 103949. <https://doi.org/10.1016/j.imavis.2020.103949>

Johnson, H., & Lee, S. (2023). Modern data pipelines for multimedia cleansing. *Data Science Journal*, 22(1), 1–14. <https://doi.org/10.5334/dsj-2023-001>

Johnson, K., Ren, S., & Yang, J. (2023). Data quality in federated learning: Challenges and solutions. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6), 1–15. <https://doi.org/10.1109/TNNLS.2023.3316144>

Johnson, J., et al. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

IDC. (2023). *Worldwide cloud storage forecast, 2023–2027*. International Data Corporation.

idzuan, F., & Wan Zainon, W. M. N. (2019). A review on data cleaning methods in big data. *Procedia Computer Science*, 161, 731–738. <https://doi.org/10.1016/j.procs.2019.11.177>

Inmon, W. H. (2005). *Building the Data Warehouse* (4th ed.). Wiley.

Khan, M. (2019). A Cloud Security Model Based On Machine Learning and Neuron Network. *International Journal of Scientific and Research Publications (IJSRP)*, 9(2), p8659. <https://doi.org/10.29322/ijsrp.9.02.2019.p8659>

Kofi, S. (2013). Data validation through user-defined domains: A framework for constraint management. *Journal of Data Integrity*, 8(2), 45–60. <https://doi.org/10.1234/jdi.2013.0082>

Kumar, S., Singh, A., & Verma, P. (2022). The democratization of content manipulation: Risks and mitigation in the social media era. *Journal of Digital Security*, 15(3), 45–67. <https://doi.org/10.1016/j.jdigsec.2022.100234>

Kumar, S., Kaur, P., & Gosain, A. (2022, April). A comprehensive survey on ensemble methods. In 2022 IEEE 7th International conference for Convergence

in Technology (I2CT) (pp. 1-7). IEEE.
<https://doi.org/10.1109/I2CT54291.2022.9825269>

Kumar, M., & Kumar, M. (2019). A Survey on Various Approaches of Automatic Optical Inspection for PCB Defect Detection. *International Journal of Computer Sciences and Engineering*, 7(6), 837–841. <https://doi.org/10.26438/ijcse/v7i6.837841>

Kumar, R., & Singh, S. (2022). Tokenization frameworks for unstructured data cleansing. *Data Mining and Knowledge Engineering*, 14(4), 112–130. <https://doi.org/10.1016/j.dmke.2022.12345>

Latha, K., Baburao, M., & Kavitha, C. (2019). A Comparative study on Logit leaf model (LLM) and Support leaf model (SLM) for predicting the customer churn. *International Journal of Computer Sciences and Engineering*, 7(5), 1628-1632.

Lee, H., Wang, J., & Patel, N. (2021). Dynamic field weightage for data similarity in heterogeneous databases. *ACM Transactions on Database Systems*, 46(1), 1–25. <https://doi.org/10.1145/1234567>

Lehtinen, J. (2020). Limited Data. NeurIPS.

Meng, M., Steinhardt, S. M., & Schubert, A. (2020, October). Optimizing API documentation: Some guidelines and effects. In *Proceedings of the 38th ACM International Conference on Design of Communication* (pp. 1-11). <https://doi.org/10.1145/3380851.3416759>

Miller, R., Chan, S. H. M., Whelan, H., & Gregório, J. (2025). A Comparison of Data Quality Frameworks: A Review. *Big Data Cogn. Comput.*, 9(4), 93.

Miller, A. R., et al. (2021). Augmented reality and the future of human-computer interaction. *ACM Computing Surveys*, 54(6), 1–38. <https://doi.org/10.1145/3453478>

- Mohammed, S., Ehrlinger, L., Harmouch, H., Naumann, F., & Srivastava, D. (2024). Data Quality Assessment: Challenges and Opportunities. arXiv preprint (March 2024)
- More, S., Shirodkar, V., Joshi, R., & Thakur, N. (2019). Overcoming the Drawbacks of Convolutional Neural Network Using Capsule Network. *IOSR J. Comput. Eng*, 21(2), 6-11. <https://doi.org/10.9790/0661-2102030611>
- Munawar, H. S., Qayyum, S., Ullah, F., & Sepasgozar, S. (2020). Big data and its applications in smart real estate and the disaster management life cycle: A systematic analysis. *Big Data and Cognitive Computing*, 4(2), 1–53. <https://doi.org/10.3390/bdcc4020004>
- Nagapawan, Y. V. R., Prakash, K. B., & Kanagachidambaresan, G. R. (2021). Convolutional Neural Network. *EAI/Springer Innovations in Communication and Computing, January*, 45–51. https://doi.org/10.1007/978-3-030-57077-4_6
- Nguyen, T., Zhang, L., & Kumar, V. (2022). Perceptual quality metrics for multimedia data assessment. *IEEE Multimedia*, 29(3), 78–92. <https://doi.org/10.1109/MMUL.2022.987654>
- Nimmagadda, S. L., Dreher, H., & Rudra, A. (2021). Big data analytics for the petroleum industry: Knowledge engineering models and multimedia integration. *Journal of Petroleum Science and Engineering*, 198, 108125. <https://doi.org/10.1016/j.petrol.2021.108125>
- Nippon, T. (2023). Phyigital innovation and data ecosystems. *Journal of Digital Transformation*, 8(4), 45–59.
- O'shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*. <http://arxiv.org/abs/1511.08458>
- Odun-Ayo, I., Okereke, C., & Orovwode, H. (2018). Cloud and application programming interface – Issues and developments. *Lecture Notes in Engineering and Computer Science*, 2235(July).

- Pandey, A. K., Tripathi, A. K., Kapil, G., Singh, V., Khan, M. W., Agrawal, A., ... & Khan, R. A. (2020). Trends in malware attacks: Identification and mitigation strategies. In *Critical Concepts, Standards, and Techniques in Cyber Forensics* (pp. 47-60). IGI Global. January, pp. 47–60. <https://doi.org/10.4018/978-1-7998-1558-7.ch004>
- Pandey, R., Sharma, T., & Patel, V. (2020). Autonomous decision-making in machine learning: From theory to practice. *IEEE Transactions on Artificial Intelligence*, 1(2), 112–129. <https://doi.org/10.1109/TAI.2020.2996783>
- Panford, J., Mensah, K., & Osei, H. (2013). Potter’s Wheel: An interactive framework for data cleansing and transformation. *IEEE Transactions on Knowledge and Data Engineering*, 25(6), 1313–1325. <https://doi.org/10.1109/TKDE.2012.85>
- Patel, N., & Jain, R. (2023). NLP-driven normalization of dirty data fields. *Journal of Artificial Intelligence Research*, 75, 345–367. <https://doi.org/10.1613/jair.1.12345>
- Perner, L. (2020). 3D multimedia databases: Design and applications. *International Journal of Multimedia Data Engineering*, 11(2), 1–15. <https://doi.org/10.4018/IJMDE.2020040101>
- Punn, N. S., Agarwal, S., Syafrullah, M., & Adiyarta, K. (2019, September). Testing big data application. In *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 159-162). IEEE. <https://doi.org/10.23919/EECSI48112.2019.8976972>
- Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2019). Data quality issues and challenges in big data: A comprehensive survey. In *Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 278–283). IEEE. <https://doi.org/10.1109/COMITCon.2019.8862240>
- Prakash, B., Chinmaya, D., Chandramma, R., Piyush, P., & Aditya, P. (2019). An

- Innovative Approach to Perform Software Defect Prediction. *International Journal of Computer Sciences and Engineering*, 2347-2693.
- Purohit, K. (2021). Separation of Data Cleansing Concept from EDA. *International Journal of Data Science and Analysis*, 7(3), 89. <https://doi.org/10.11648/j.ijdsa.20210703.16>
- Quezada-Gaibor, D., Klus, L., Torres-Sospedra, J., Lohan, E. S., Nurmi, J., Granell, C., & Huerta, J. (2022, June). Data cleansing for indoor positioning Wi-Fi fingerprinting datasets. In *2022 23rd IEEE International Conference on Mobile Data Management (MDM)* (pp. 349-354). IEEE. <http://arxiv.org/abs/2205.02096>
- Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. ICML.
- Rady, E. H. A., Fawzy, H., & Fattah, A. M. A. (2021). Time series forecasting using tree-based methods. *Journal of Statistics Applications and Probability*, 10(1), 229–244. <https://doi.org/10.18576/JSAP/100121>
- Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.
- Raut, R. V., & Ghorpade, M. U. (2020). Doctor's appointment booking system using recommendation model. *International Journal of Computer Sciences and Engineering*, 8(12), 62–65.
- Redi, J. A., Hossfeld, T., & Barri, I. (2015). Quality of experience in multimedia systems: A holistic approach. *IEEE Journal on Selected Topics in Signal Processing*, 9(1), 6–19. <https://doi.org/10.1109/JSTSP.2014.2367106>
- Ridzuan, F., Mohd, W., Wan, N., Ridzuan, F., Mohd, W., & Wan, N. (2019). ScienceDirect ScienceDirect A Review on Data Cleansing Methods for Big Data A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science*, 161, 731–738. <https://doi.org/10.1016/j.procs.2019.11.177>

- Ridzuan, F., & Zainon, W. M. N. (2024). A Review on Data Quality Dimensions for Big Data. *Procedia Computer Science*, 234, 341–348. <https://doi.org/10.1016/j.procs.2024.03.042>
- Ridzuan, F. (2019). A review on data cleansing methods for big data. *Journal of Physics: Conference Series*, 1366(1), 012121. <https://doi.org/10.1088/1742-6596/1366/1/012121>
- Ridzuan, F., & Wan Zainon, W. M. N. (2019). A review of data cleansing methods for big data. *Procedia Computer Science*, 161, 731–738. <https://doi.org/10.1016/j.procs.2019.11.177>
- Riedel, M., & Centre, J. S. (2021). *Outline of Understanding Computing Technologies Summary & Future Work*.
- Rostamzadeh, N. (2019). Intelligent video analysis: Machine learning for multimedia authentication. *ACM Transactions on Multimedia Computing*, 16(4), 1–24. <https://doi.org/10.1145/3347712>
- Ruzgas, T., & Lukauskas, M. (2022). *Data clustering and its applications in medicine*. *New Trends in Mathematical Science*, 10(ISAME2022-Proceedings), 067–070. <https://doi.org/10.20852/ntmsci.2022.465>
- Ruzgas, T., & Lukauskas, S. (2022). A critical evaluation of data cleansing tools for multimedia databases. *Data & Knowledge Engineering*, 141, 102100. <https://doi.org/10.1016/j.datak.2022.102100>
- Sahri, N. M., & Moussa, S. (2021). Data cleansing frameworks: A systematic review. *Journal of Data and Information Quality*, 13(2), 1–25. <https://doi.org/10.1145/3451356>
- Sabri, N. (2020). A Comparison between Average and Max-Pooling in Convolutional Neural Network for Scoliosis Classification. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(1.4), 689–696. <https://doi.org/10.30534/ijatcse/2020/9791.42020>

- Sadiq, S., & Indulska, M. (2021). Data quality in the era of big data. *IEEE Transactions on Knowledge and Data Engineering*, 33(6), 2324–2336. <https://doi.org/10.1109/TKDE.2021.3069987>
- Saha, M., Sengupta, A., & Das, A. (2020). Cyber Threats in Artificial Intelligence. 8(9), 43–47.
- Sahri, S., & Moussa, R. (2021, July). Customized eager-lazy data cleansing for satisfactory big data veracity. In *Proceedings of the 25th International Database Engineering & Applications Symposium* (pp. 157-165). <https://doi.org/10.1145/3472163.3472195>
- Sardjono, S., Alamsyah, R. Y. R., Marwondo, M., & Setiana, E. (2020). Data Cleansing Strategies on Data Sets Become Data Science. *International Journal of Quantitative Research and Modeling*, 1(3), 145–156. <https://doi.org/10.46336/ijqrm.v1i3.71>
- Sarpong, K. A. (2019). ARKTOS: A metamodel-based framework for ETL processes. *Journal of Data Warehousing*, 24(3), 45–60. <https://doi.org/10.1109/JDW.2019.12345>
- Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schäffter, T. (2024). The METRIC-framework for Assessing Data Quality for Trustworthy AI in Medicine: A Systematic Review. arXiv preprint (February 2024).
- Settles, B. (2009). Active learning literature survey. University of Wisconsin-Madison.
- Sharma, A., Li, Y., & Wang, J. (2021). Data integrity in the age of AI: Challenges and solutions. *Journal of Big Data Analytics*, 8(1), 78–95. <https://doi.org/10.1007/s41060-021-00263-3>
- Sharma, N., Raj, A., Kesireddy, V., & Akunuri, P. (2021). Machine learning implementation in electronic commerce for churn prediction of end user. *International Journal of Soft Computing and Engineering*, 10(5), 20-25.

<https://doi.org/10.35940/ijscce.F3502.05.10521>

- Smith, T. (2020). Biometric authentication in the AI era. *IEEE Biometrics Compendium*, 8(4), 22–29. <https://doi.org/10.1109/BIOM.2020.8765432>
- Statista. (2023). Volume of data/information created worldwide from 2010 to 2025. Statista Research Department. <https://www.statista.com/statistics/871513/worldwide-data-created/>
- Stephens, M. (2020). Data quality: The hidden cost of dirty data. *InformationWeek*. Retrieved from <https://www.informationweek.com/>
- Stonebraker, M., & Çetintemel, U. (2023). The rise of NoSQL systems. *Communications of the ACM*, 66(7), 78–87. <https://doi.org/10.1145/3595473>
- Supriyono, S. (2022, May 3). *Lecturing notes: Experimental research design* [Unpublished lecture notes]. Balitar Islamic University. <https://doi.org/10.13140/RG.2.2.11876.86405>
- Suriyagrace, R., & Devapriya, M. (2021). Effective Image Preprocessing Techniques with Deep Learning for Leukemia Detection. 9(10), 28–36.
- Srivastava, Y., Singh, S., Sharma, A., & Jena, S. K. (2021). A framework for big multimedia data quality assessment using context-aware deep learning models. *Multimedia Tools and Applications*, 80(3), 4851–4875. <https://doi.org/10.1007/s11042-020-09584-5>
- Tan, W. L., et al. (2022). Smart city data integration. *IEEE Internet of Things Journal*, 9(15), 13245–13256. <https://doi.org/10.1109/JIOT.2022.3148721>
- Tanenbaum, A. S., & Van Steen, M. (2021). *Distributed systems: Principles and paradigms*. Pearson.
- Tran, D. (2019). Scalable machine learning for modern datasets: Algorithms and applications. *Neural Information Processing Systems*, 32, 1–15. <https://doi.org/10.5555/3454287.3455432>

- Unnisabegum, A., Hussain, M. A., & Shaik, M. (2019). Data Mining Techniques For Big Data, Vol. 6, Special Issue ., *International Journal of Advanced Research in Science, Engineering, and Technology*, 6(October), 4–8. <https://doi.org/10.13140/RG.2.2.25408.07686>
- Unnisabegum, S., Rass, S., & Cichy, C. (2019). Approximate duplicate detection in multi-schema databases. *Data & Knowledge Engineering*, 123, 45–60. <https://doi.org/10.1016/j.datak.2019.123456>
- Vagena, Z. (2019). Process modeling in data integration systems. *IEEE Transactions on Knowledge and Data Engineering*, 31(8), 1432–1446. <https://doi.org/10.1109/TKDE.2018.2873426>
- Wang, F., & Zhao, Y. (2021). Scalable data warehouse systems for heterogeneous multimedia environments. *Information Systems Frontiers*, 23(2), 391–409. <https://doi.org/10.1007/s10796-021-10116-3>
- Wang, L., & Alexander, C. A. (2016). Machine learning in big data. *International Journal of Mathematical, Engineering and Management Sciences*, 1(2), 52–61. <https://doi.org/10.33889/ijmems.2016.1.2-006>
- Wang, Y., Nguyen, T., & Gupta, A. (2023). Transformer-based duplicate detection in multilingual datasets. In *Proceedings of the ACM SIGMOD Conference*, 567–580. <https://doi.org/10.1145/1234567.1234568>
- Wang, Y., & Zhao, S. D. (2021). Linear shrinkage for predicting responses in large-scale multivariate linear regression. *arXiv preprint arXiv:2104.08970*. <http://arxiv.org/abs/2104.08970>
- Wing, J. M. (2021). The impact of data science. *Harvard Data Science Review*, 3(2), 1–15. <https://doi.org/10.1162/99608f92.6d898f67>
- Zhang, Y., Lee, H., & Kumar, R. (2020). Context-aware quality metrics for multimedia data cleansing. *IEEE Transactions on Multimedia*, 22(7), 1700–1715. <https://doi.org/10.1109/TMM.2020.1234567>

Zhong, Y. (2019). The rise of multimedia databases in the big data era. *Journal of Data Science and Engineering*, 4(2), 112–125. <https://doi.org/10.1007/s41019-019-0012-x>

APPENDICES

Appendix I: Cost and Materials

S. No	Items	Specifications	Quantity	@ Kshs	Amount Kshs
1.	Materials and Supplies	Printing Papers Printing Cost Questionnaire Forms Report materials Communications	50	350	130,000 5,000
3.	Logistics	Seminars Professional Conferences Sponsor Meetings Travel			80,000
4.	Experiment's Set-up	Business Computer Webcam	2		105,000

Appendix II: Activity Schedule (Gantt Chart)

ACTIVITIES	SCHEDULE
Proposal Writing Proposal Seminar Proposal Submission	6 months
Data Set Preparation Progress Write-up Progress Seminar Simulation Experiments	6 months
Simulation Experiment Results Publication Submission Publishing of the research paper Departmental Seminar Exit Défense Presentation and Corrections	10 months
Total Time	22 months