International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

Neuro-Symbolic Architectures for Explainable Multi-Modal Plagiarism Detection in Academic Assessment

Paul Showemimo

Date of Submission: 05-07-2025 Date of Acceptance: 15-07-2025

ABSTRACT
The pervasive integration of Artificial Intelligence (AI) into educational ecosystems has introduced both unprecedented opportunities and complex challenges, particularly concerning academic integrity. While AI-driven tools have become indispensable for identifying traditional forms of plagiarism, the emergence of sophisticated content generation models and multi-modal assignments necessitates a paradigm shift in detection methodologies. This paper proposes a novel **Multi-Modal Neuro-Symbolic Assessment System (MNSAS)** architecture specifically designed for explainable plagiarism detection. Our approach integrates the robust pattern recognition capabilities of neural networks with the logical reasoning and transparency of symbolic AI to process and analyze diverse academic submissions, including text, code, images, audio, and video. This neuro-symbolic synergy not only enhances the accuracy and robustness against advanced obfuscation techniques but, crucially, provides human-understandable explanations for detected anomalies, fostering trust and promoting genuine learning. We detail the technical components, architectural considerations, mathematical formulations, evaluation metrics, and research questions that underpin this innovative framework, demonstrating its potential to significantly advance the field of academic integrity and contribute to a more transparent and equitable educational landscape.

I. INTRODUCTION

The digital transformation of education has profoundly reshaped learning and assessment practices. Online platforms, digital submissions, and the proliferation of information have made academic integrity both more critical and more challenging to uphold [1]. Traditional plagiarism detection systems, primarily reliant on text-matching algorithms, are increasingly insufficient in an era where students engage with multi-modal content and sophisticated AI tools can generate highly coherent and contextually relevant text, code, or even visual and auditory material [2]. The rise of AI-generated content (AIGC) poses a significant threat, as it can mimic human writing styles, paraphrase effectively, and even synthesize information from disparate sources, making detection by conventional means exceedingly difficult [3].

Furthermore, modern academic assignments often transcend traditional text-based formats. Students submit programming projects, design portfolios, video presentations, audio recordings, and scientific reports that integrate text with complex diagrams, data visualizations, or experimental results. Plagiarism in such multi-modal contexts can manifest in subtle ways, such as the uncredited reuse of a diagram, the conceptual appropriation of an idea presented verbally, or the structural replication of code without proper attribution. Existing detection mechanisms frequently operate in silos, analyzing each modality independently, thereby failing to capture cross-modal plagiarism or the nuanced intent behind content reuse [4].

Beyond detection accuracy, a critical limitation of many contemporary AI systems, particularly deep learning models, is their inherent lack of transparency—the so-called "black box" problem. When a system flags a student for potential academic misconduct, educators and students require clear, interpretable justifications for the decision. Without such explainability, AI-driven integrity systems can erode trust, lead to disputes, and hinder the pedagogical goal of fostering genuine understanding and ethical behavior [5]. A system that merely identifies a problem without explaining why it is a problem offers limited educational value and can be perceived as arbitrary or unfair.

This paper addresses these pressing challenges by introducing a novel **Multi-Modal Neuro-Symbolic Assessment System (MNSAS)**

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

architecture for explainable plagiarism detection. Our proposed framework moves beyond the limitations of unimodal, black-box approaches by:

1. **Integrating Multi-Modal Data:** Processing and fusing information from diverse modalities (text, code, images, audio, video) to enable comprehensive detection of plagiarism across various forms of academic submissions.
2. **Leveraging Neuro-Symbolic AI:** Combining the strengths of neural networks (for robust pattern recognition and feature extraction from complex, unstructured data) with symbolic reasoning (for logical inference, rule-based checking, and generating human-understandable explanations).
3. **Prioritizing Explainability:** Designing the system to not only detect plagiarism but also to articulate why a particular instance is flagged, referencing specific rules, patterns, and contextual cues, thereby fostering transparency and trust.
4. **Enhancing Robustness:** Developing mechanisms to counter sophisticated obfuscation techniques and adversarial attacks, including those employed by AIGC.

The remainder of this paper is structured as follows: Section 2 provides background on academic integrity challenges and reviews related work in AI-driven plagiarism detection, multi-modal AI, and neuro-symbolic systems. Section 3 details the proposed MNSAS architecture for explainable multi-modal plagiarism detection, including descriptions of key diagrams. Section 4 elaborates on the technical components and implementation details. Section 5 presents the mathematical formulations underpinning the system. Section 6 discusses evaluation metrics and experimental design. Section 7 outlines key research questions. Finally, Section 8 concludes the paper and suggests future research directions. All references are consolidated at the end of the document.

II. BACKGROUND AND RELATED WORK

2.1. The Evolving Landscape of Academic Integrity and AI-Generated Content

Academic integrity, traditionally defined by principles of honesty, trust, fairness, respect, and responsibility [1], faces unprecedented challenges in the contemporary digital educational landscape. The proliferation of online learning platforms, readily accessible digital resources, and, most significantly, the rapid advancements in Artificial Intelligence (AI) have fundamentally altered the dynamics of academic misconduct [2]. Students now have access to sophisticated tools capable of generating highly coherent and contextually relevant content across various modalities, ranging from textual essays to complex programming assignments and multimedia presentations. This includes advanced paraphrasing tools, essay mills leveraging AI, and large language models (LLMs) that can produce original-sounding prose, code, or even visual and auditory material [3].

The emergence of AI-Generated Content (AIGC) poses a particularly insidious threat. Unlike traditional plagiarism, which involves direct copying or minor rephrasing of existing human-authored material, AIGC can synthesize novel content that bypasses conventional text-matching algorithms [4]. This necessitates a paradigm shift in detection methodologies, moving beyond surface-level lexical analysis to deeper semantic, structural, and stylistic understanding, and even to the identification of statistical fingerprints indicative of machine generation. The challenge is compounded by the fact that AIGC can mimic human writing styles, effectively paraphrase complex ideas, and integrate information from disparate sources, making it increasingly difficult to distinguish between legitimate student work and AI-assisted or AI-generated submissions [5].

2.2. Limitations of Traditional Plagiarism Detection Systems

Traditional plagiarism detection systems, such as Turnitin and SafeAssign, have been instrumental in upholding academic standards for decades. These systems primarily rely on string matching, n-gram analysis, and lexical comparison algorithms to identify identical or highly similar text segments against vast databases of academic papers, web content, and previously submitted assignments [6]. While effective for detecting direct copying, their efficacy diminishes significantly when confronted with more sophisticated forms of academic misconduct:

- **Sophisticated Paraphrasing:** These systems often struggle to detect plagiarism when the original text is extensively rephrased, even if the underlying ideas, arguments, or conceptual structures are appropriated without proper attribution. Semantic similarity measures are often rudimentary, failing to capture deep conceptual equivalence.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

- **Conceptual Plagiarism:** This involves the uncredited appropriation of ideas, theories, methodologies, or structural frameworks, even if the wording is entirely original. Traditional systems, being primarily text-centric, are ill-equipped to identify such abstract forms of intellectual theft.
- **Source Code Plagiarism:** Detecting plagiarism in programming assignments requires specialized analysis beyond natural language processing. Similarities in code structure (e.g., Abstract Syntax Trees - ASTs), algorithms, variable naming conventions, and logical flow are critical indicators that are often missed by text-based tools [7].
- **Cross-Language Plagiarism:** The translation of content from one language to another without proper attribution is another challenge. While some advanced systems incorporate machine translation, the semantic and stylistic nuances across languages can still pose significant detection hurdles.
- **Multi-Modal Content:** Crucially, traditional systems are inherently unimodal, focusing almost exclusively on text. They are blind to plagiarism occurring in images, audio, video, or other non-textual modalities, which are increasingly common in modern academic submissions.

The advent of LLMs has exposed the fundamental limitations of these traditional approaches. The ability of LLMs to generate highly fluent, contextually appropriate, and novel text means that they can often bypass detection by systems reliant on surface-level textual similarities. This necessitates a move towards more intelligent, context-aware, and multi-modal detection mechanisms that can understand the deeper semantic meaning and conceptual nuances of content, irrespective of its modality or generation source.

2.3. Advancements in Multi-Modal AI for Content Understanding

The increasing prevalence of multi-modal assignments (e.g., video presentations, interactive simulations, design portfolios, scientific reports with embedded figures and data) highlights a significant gap in current academic integrity systems. Plagiarism is no longer confined to text; it can involve the uncredited use of images, audio clips, video segments, or even the conceptual structure of a visual argument. Multi-modal AI, which focuses on processing, understanding, and integrating information from multiple distinct modalities (e.g., text, image, audio, video) simultaneously, offers a promising avenue to address this challenge [8].

Recent breakthroughs in multi-modal learning have enabled models to learn joint representations across different data types, capturing complex inter-modal relationships. Key developments include:

- **Vision-Language Models (VLMs):** Models like CLIP [9], DALL-E [10], and Flamingo [11] have demonstrated remarkable capabilities in understanding the semantic relationship between images and text. They can perform tasks such as image captioning, visual question answering, and text-to-image generation, indicating a deep cross-modal comprehension. Applying these to academic integrity could enable detection of uncredited image use or cross-modal conceptual plagiarism where a textual description is derived from an uncredited visual source.
- **Audio-Visual Models:** Architectures like Audio-Visual Transformers [12] and models for synchronized audio and video streams can process and understand spoken content in conjunction with visual cues. This is crucial for analyzing video presentations, where plagiarism might involve uncredited audio segments, re-used video clips, or even the appropriation of a speaker's delivery style.
- **Code-Text Models:** Emerging models are capable of understanding both natural language descriptions and programming code, facilitating tasks like code generation from natural language or code summarization. This is directly relevant for detecting plagiarism in programming assignments, where both the code and accompanying documentation need to be analyzed [13].

The technical challenge in multi-modal plagiarism detection lies in the heterogeneity of data formats, the complexity of feature extraction from each modality, and the difficulty of fusing these features into a coherent, unified representation that preserves semantic meaning and contextual relationships across different forms of expression. Effective multi-modal fusion techniques are required to learn a shared embedding space where content from different modalities can be directly compared for similarity and originality.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

2.4. The Imperative of Explainable AI (XAI) in Academic Contexts

The demand for Explainable AI (XAI) has grown exponentially, particularly in high-stakes applications where AI decisions have significant consequences, such as in education and assessment [14]. When AI systems make determinations that directly impact students' academic standing (e.g., flagging for plagiarism), transparency and interpretability are not merely desirable but paramount. A "black box" AI system, which provides a decision without a clear rationale, can lead to profound distrust, perceived unfairness, and an inability for educators to provide targeted feedback or for students to learn from their mistakes. For plagiarism detection, XAI is crucial for several reasons:

- **Justification and Transparency:** Providing clear, auditable reasons why a submission is flagged. This involves pinpointing specific sections, identifying the suspected source, and articulating the precise type of misconduct (e.g., direct copying, paraphrasing, structural appropriation, conceptual theft). This transparency is essential for due process and maintaining trust in automated systems.
- **Pedagogical Value:** Enabling educators to leverage the AI's insights to guide students towards a deeper understanding of academic integrity principles and best practices. Instead of merely identifying a violation, an explainable system can illustrate how and why a particular action constitutes plagiarism, offering concrete examples of proper citation, paraphrasing, and original thought.
- **Dispute Resolution:** Offering a transparent and evidence-based foundation for discussing and resolving academic integrity cases. When a student disputes a finding, the system's explanation can serve as a neutral arbiter, facilitating a constructive dialogue rather than an adversarial confrontation.
- **System Improvement and Debugging:** Explainable models are inherently easier to debug and improve. If a system makes an erroneous judgment, the explanation can reveal the underlying reasoning flaw, allowing developers to refine the model or update its knowledge base.

Existing XAI techniques, such as LIME (Local Interpretable Model-agnostic Explanations) [15] and SHAP (SHapley Additive exPlanations) [16], provide local explanations for individual predictions by approximating complex models with simpler, interpretable ones. However, applying these post-hoc techniques to complex multi-modal and neuro-symbolic systems, and generating pedagogically meaningful, actionable, and coherent explanations that span multiple modalities, remains a significant and active area of research. The challenge is to move beyond mere feature importance to provide explanations that reflect the system's underlying reasoning process.

2.5. Neuro-Symbolic AI: Bridging the Gap between Perception and Reasoning

Neuro-Symbolic AI represents a rapidly evolving paradigm that seeks to combine the complementary strengths of neural networks (connectionist models) with symbolic AI (logic-based models) [17]. This hybrid approach aims to overcome the limitations of purely connectionist or purely symbolic systems. Neural networks excel at pattern recognition, learning from vast amounts of data, and handling noisy or incomplete information. However, they often operate as "black boxes," lacking explicit mechanisms for reasoning, knowledge representation, or providing transparent explanations. Symbolic AI, on the other hand, excels at logical inference, knowledge representation using formal languages (e.g., ontologies, rules), and providing interpretable reasoning chains. However, symbolic systems often struggle with the ambiguity and complexity of real-world perceptual data and can be brittle when faced with noisy or incomplete information.

By integrating these two paradigms, neuro-symbolic systems aim to achieve the best of both worlds: robust learning and perception capabilities from neural networks, coupled with the rigorous reasoning, knowledge integration, and explainability of symbolic methods [18]. In the context of plagiarism detection, a neuro-symbolic approach offers several advantages:

- **Enhanced Explainability:** Symbolic components can provide a structured, logical framework for interpreting the patterns detected by neural components, leading to more intuitive and human-understandable explanations for plagiarism detection.
- **Knowledge Integration:** Symbolic knowledge bases (e.g., ontologies of academic misconduct, rules for citation, definitions of plagiarism types) can be explicitly incorporated to guide the detection process and enrich the explanations.
- **Robustness to Adversarial Attacks:** The explicit reasoning capabilities of symbolic AI can make the system more robust to

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

adversarial attacks designed to fool purely neural systems.

- **Improved Generalization:** By grounding neural learning in symbolic knowledge, neuro-symbolic systems can potentially generalize better from limited data and adapt more effectively to new forms of plagiarism.

Several neuro-symbolic architectures have been proposed, ranging from loosely coupled systems where neural and symbolic components interact at a high level, to tightly integrated models where neural networks are designed to perform symbolic reasoning or where symbolic knowledge is directly embedded into neural architectures [19]. The choice of architecture depends on the specific task requirements, the nature of the available data, and the desired level of explainability.

2.6. Gaps in Current Research

Despite significant advancements in AI-driven plagiarism detection, multi-modal AI, XAI, and neuro-symbolic systems, several critical research gaps remain, particularly at the intersection of these fields:

1. **Lack of Integrated Multi-Modal Neuro-Symbolic Frameworks for Plagiarism Detection:** While individual components exist, there is a dearth of comprehensive frameworks that seamlessly integrate multi-modal data processing, neuro-symbolic reasoning, and robust explainability specifically tailored for the complexities of academic plagiarism detection.
2. **Limited Explainability in Multi-Modal Contexts:** Generating coherent, human-understandable explanations that span multiple modalities (e.g., explaining how a textual claim is plagiarized from a visual diagram or an audio narration) is a largely unsolved problem.
3. **Scalability and Efficiency:** Processing and analyzing large volumes of diverse multi-modal data in real-time while performing complex neuro-symbolic reasoning presents significant computational challenges.
4. **Robustness against Sophisticated Evasion Techniques:** As AI-generated content and obfuscation methods become more advanced, developing detection systems that are resilient to these evolving threats is crucial.
5. **Standardized Datasets and Evaluation Metrics:** The field lacks standardized, large-scale, multi-modal datasets specifically designed for plagiarism detection research, along with comprehensive evaluation metrics that capture not only detection accuracy but also the quality and utility of the generated explanations.

Our proposed MNSAS architecture aims to address these gaps by providing a foundational framework for building the next generation of explainable, multi-modal, and robust plagiarism detection systems.

III. PROPOSED MNSAS ARCHITECTURE FOR EXPLAINABLE MULTI-MODAL PLAGIARISM DETECTION

We propose a Multi-Modal Neuro-Symbolic Assessment System (MNSAS) architecture designed to address the limitations of existing plagiarism detection methods. The MNSAS framework is conceptualized as a layered, modular system that synergistically combines neural perception with symbolic reasoning to achieve accurate, robust, and, critically, explainable plagiarism detection across diverse academic submission types. Figure 1 presents a high-level overview of the MNSAS architecture, Figure 2 details the Neuro-Symbolic Integration Layer, and Figure 3 illustrates the Explainable Report Generation process.

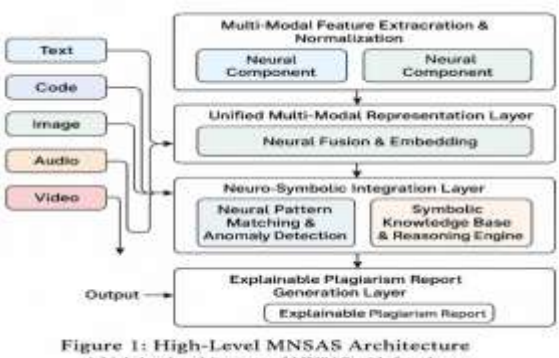


Figure 1: High-Level MNSAS Architecture
A high-level architecture of MNSAS with four layers.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

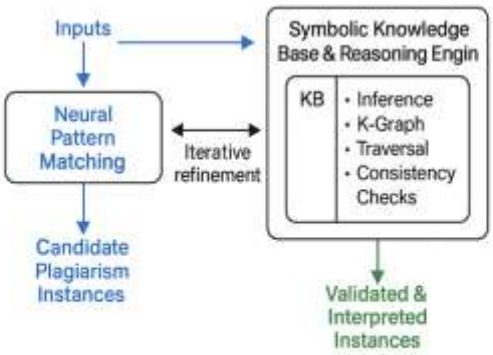


Figure 2: Detailed Neuro-Symbolic Integration Layer
Neural pattern matching identifies candidate instances, which are ylidated and interpreted by the reasoning engine using knowledge base informio-

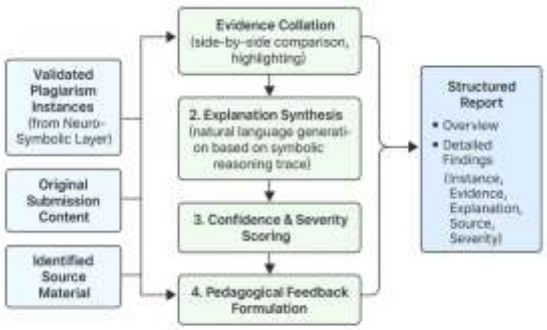


Figure 3: Explainable ReportGeneration Process

IV. TECHNICAL COMPONENTS AND IMPLEMENTATION DETAILS

This section delves into the specific technical components of the proposed MNSAS architecture, outlining the methodologies and technologies envisioned for each layer. The implementation will prioritize modularity, allowing for individual components to be updated or replaced as new research emerges.

4.1. Multi-Modal Feature Extraction and Normalization Layer

This initial layer is responsible for processing raw input data from various modalities and transforming it into numerical feature representations suitable for subsequent neural processing. Each modality requires specialized techniques:

4.1.1. Text Processing

- **Input:** Plain text from documents (e.g., .txt, .docx, .pdf), textual components of presentations, or transcribed audio.
- **Techniques:**
- **Preprocessing:** Tokenization, stemming/lemmatization, stop-word removal, sentence segmentation.
- **Feature Extraction:**
- Transformer-based Embeddings: Utilize pre-trained language models like BERT [20], RoBERTa [21], or specialized academic domain models (e.g., SciBERT [22]) to generate contextual word and sentence embeddings. These capture deep semantic meaning. For a sentence $S = \{w_1, w_2, ..., w_{*n*}\}$, the embedding E(S) is obtained.
- TF-IDF (Term Frequency-Inverse Document Frequency): For capturing keyword importance and as a supplementary feature.
- Stylometric Features: Analysis of writing style (e.g., sentence length distribution, vocabulary richness, use of specific grammatical structures) can help identify authorship or detect AIGC [23].
- **Normalization:** Embeddings are typically normalized (e.g., L2 normalization) to ensure consistent scales for downstream tasks.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

4.1.2. Source Code Processing

- **Input:** Source code files (e.g., Python, Java, C++, JavaScript).
- **Techniques:**
- **Preprocessing:** Parsing code into Abstract Syntax Trees (ASTs), tokenization of code elements (keywords, identifiers, operators).
- **Feature Extraction:**
- AST Embeddings: Graph embedding techniques (e.g., Graph Convolutional Networks - GCNs [24], Graph Attention Networks - GATs [25]) applied to ASTs to capture structural similarity.
- CodeBERT/GraphCodeBERT: Transformer models pre-trained on large code corpora to learn contextual representations of code tokens and structure [13, 26].
- Halstead Complexity Metrics & Cyclomatic Complexity: Quantitative measures of code complexity and structure.
- Token n-grams: Sequences of code tokens.
- **Normalization:** Similar to text embeddings, code embeddings are normalized.

4.1.3. Image Processing

- **Input:** Images embedded in documents, standalone image files (e.g., .jpg, .png, .svg).
- **Techniques:**
- **Preprocessing:** Resizing, color normalization, data augmentation (if training detectors).
- **Feature Extraction:**
- Convolutional Neural Networks (CNNs): Pre-trained CNNs like ResNet [27], VGG [28], or EfficientNet [29] to extract rich visual features. The output of intermediate or final convolutional layers serves as the image embedding.
- Vision Transformers (ViT): Transformer architectures applied directly to image patches, capturing global context [30].
- Perceptual Hashing: Algorithms like pHash, aHash, dHash to generate compact image fingerprints for fast similarity comparison [31].
- Object Detection & Segmentation: Identifying and localizing objects or regions within an image (e.g., using YOLO [32], Mask R-CNN [33]) can be relevant for detecting partial image plagiarism or understanding image content.
- **Normalization:** Feature vectors are normalized.

4.1.4. Audio Processing

- **Input:** Audio files (e.g., .mp3, .wav, .aac) from presentations, discussions, or embedded media.
- **Techniques:**
- **Preprocessing:** Resampling, noise reduction, silence removal.
- **Feature Extraction:**
- Mel-Frequency Cepstral Coefficients (MFCCs): Standard features for speech and audio recognition.
- Wav2Vec 2.0 / HuBERT: Self-supervised learning models that learn powerful contextual representations from raw audio waveforms [34, 35].
- Audio Fingerprinting: Techniques like Shazam's algorithm to identify known audio clips.
- Speaker Diarization& Recognition: Identifying who spoke when, which can be relevant for detecting uncredited speech.
- **Normalization:** Feature vectors are normalized.

4.1.5. Video Processing

- **Input:** Video files (e.g., .mp4, .mov, .avi).
- **Techniques:** Video processing often combines image and audio processing techniques applied frame-by-frame or to segments, along with temporal modeling.
- **Preprocessing:** Frame extraction, temporal segmentation.
- **Feature Extraction:**
- 3D CNNs (e.g., C3D, I3D): Networks that convolve over both spatial and temporal dimensions to capture spatio-temporal features [36].
- Two-Stream Networks: Separate streams for appearance (spatial features from individual frames using 2D CNNs) and motion (optical flow features), later fused.
- Frame-level Image Embeddings + Audio Embeddings: Extracting features from individual frames and the accompanying audio track, then fusing them.
- Video Transformers (e.g., ViViT): Applying transformer architectures to sequences of frame embeddings [37].
- **Normalization:** Feature vectors are normalized.

4.2. Unified Multi-Modal Representation Layer

This layer is crucial for enabling cross-modal comparison and reasoning. It takes the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

normalized unimodal feature vectors (or sequences of vectors) from the previous layer and projects them into a shared, unified embedding space where semantically similar content from different modalities is located close together. Let F_m be the feature representation for modality $m \in {text, code, image, audio, video}$. The goal is to learn a mapping $g(F_m) \rightarrow U$, where U is the unified representation.

- **Techniques:**
- **Concatenation & Fully Connected Layers:** A simple approach where feature vectors from different modalities are concatenated and then passed through one or more fully connected layers to learn a joint representation. This is often a baseline.
- **Cross-Modal Attention Mechanisms:** Attention mechanisms allow the model to weigh the importance of different modalities or different parts of each modality when constructing the unified representation. For instance, when comparing a text description to an image, attention can highlight relevant regions in the image and corresponding words in the text [38].
- **Multi-Modal Transformers:** Architectures like ViLBERT [39], LXMERT [40], or UNITER [41] use transformer blocks to jointly process sequences of features from different modalities (e.g., image regions and text tokens), learning deep cross-modal interactions.
- **Graph Neural Networks (GNNs):** If the submission can be represented as a multi-modal graph (e.g., nodes are text segments, images, code blocks, with edges representing relationships like "mentions," "depicts," "explains"), GNNs can learn a holistic representation of the entire submission by propagating information across modalities [42].
- **Contrastive Learning:** Training the fusion model using contrastive loss functions (e.g., InfoNCE) to pull representations of semantically similar cross-modal pairs closer and push dissimilar pairs apart. This is particularly effective for learning aligned embedding spaces (e.g., CLIP [9]).
- **Output:** A single vector or a graph structure representing the entire multi-modal submission in a unified semantic space.

4.3. Neuro-Symbolic Integration Layer

This is the core reasoning layer where the perceptual capabilities of neural networks are combined with the logical inference of symbolic systems.

4.3.1. Neural Pattern Matching and Anomaly Detection Module

This module operates on the unified multi-modal representations to identify potential instances of plagiarism or unoriginality. It does not make definitive judgments but rather flags suspicious patterns for further symbolic analysis.

- **Cross-Modal Similarity Network:**
- **Technique:** Siamese networks [43] or triplet networks trained with contrastive or triplet loss. Given two multi-modal inputs (e.g., a segment from the student's submission and a segment from a source document, which could be of different modalities), the network learns to output a similarity score. $Sim(U_{submission}, U_{source})$
- **Application:** Comparing student submissions against a vast database of known source materials (academic papers, web pages, code repositories, image databases) and also against other student submissions (for collusion detection).

- **Unoriginality/Anomaly Detection:**
- **Technique:** Autoencoders (AEs) or Variational Autoencoders (VAEs) [44] trained on a large corpus of known original student work. Submissions that cannot be reconstructed accurately (i.e., have a high reconstruction error) are flagged as potentially unoriginal or anomalous. Generative Adversarial Networks (GANs) [45] can also be used, where a discriminator tries to distinguish between real student work and plagiarized/generated content.
- **Application:** Identifying content that deviates significantly from typical student patterns, which might indicate sophisticated paraphrasing, contract cheating, or AIGC.

- **AIGC Fingerprint Detection:**
- **Technique:** Specialized classifiers (e.g., SVMs, Transformers) trained to detect statistical artifacts, stylistic inconsistencies, or known watermarks/fingerprints associated with AI content generators [23, 46]. This might

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

involve analyzing perplexity scores, burstiness, or specific linguistic patterns.

- **Application:** Specifically targeting content generated by LLMs or other AI tools.
- **Output:** A set of candidate plagiarism instances, each with associated features: location within the submission (across modalities), potential source(s), type of anomaly (e.g., high similarity, high unoriginality score, AIGC suspected), and a confidence score from the neural model.

4.3.2. Symbolic Knowledge Base and Reasoning Engine Module

This module takes the candidate instances from the neural module and applies explicit knowledge and logical reasoning to validate, interpret, and explain them.

- **Knowledge Base (KB):**
- **Ontology of Academic Misconduct:** A formal, hierarchical representation (e.g., using OWL [47] or RDF-S) defining different types of plagiarism (e.g., DirectCopying, Paraphrasing, MosaicPlagiarism, StructuralPlagiarism, ConceptualPlagiarism, SelfPlagiarism), their properties (e.g., requiresCitation, involvesText, involvesCode), and relationships (e.g., Paraphrasing is a subTypeOfPlagiarism). This ontology also includes concepts related to AIGC, collusion, and contract cheating.
- **Citation and Referencing Rules:** Formalized rules for major citation styles (e.g., APA, MLA, IEEE) represented in a machine-interpretable format (e.g., SWRL [48], Prolog rules). These rules define correct citation practices, quotation marks, bibliography formats, etc.
- **Academic Integrity Policies:** Institution-specific policies encoded as rules.
- **Source Material Database Interface:** An interface to query indexed source materials, including metadata (author, date, publication venue) and semantic fingerprints.
- **AIGC Characteristics:** A set of rules or patterns describing known characteristics or limitations of common AI generation tools (e.g., typical phrasing, lack of deep domain knowledge, specific error types).

- **Reasoning Engine:**
- **Technique:** A hybrid reasoning engine combining probabilistic reasoning (e.g., Bayesian networks to handle uncertainty from neural outputs) with deductive reasoning (e.g., Prolog, Description Logic reasoners like Pellet or HermiT) to apply rules from the KB.
- **Process:**

1. **Validation:** The reasoning engine examines each candidate plagiarism instance. It queries the KB for relevant rules (e.g., if similarity > 0.9 and no citation, it's a potential direct copy).
2. **Contextualization:** Considers assignment guidelines (e.g., is collaboration allowed? are specific sources permitted/forbidden?).
3. **Refinement:** May query the neural module for more focused analysis if ambiguity exists.
4. **Classification:** Classifies the plagiarism type based on the ontology (e.g., StructuralCodePlagiarism, ParaphrasedTextPlagiarismFromWebSource).
5. **Explanation Generation:** Constructs a logical chain of reasoning for each confirmed instance, linking the evidence (neural scores, matched patterns) to the violated rules in the KB and the classification from the ontology. This forms the basis for the human-readable explanation.

- **Output:** Validated and interpreted plagiarism instances, each with a refined classification, identified rule violations, links to source material, and a structured explanation trace.

4.4. Explainable Plagiarism Report Generation Layer

This layer synthesizes the findings from the Neuro-Symbolic Integration Layer into a comprehensive, human-readable report for educators and students.

- **Components:**
- **Evidence Collation:** Gathers all relevant evidence, such as side-by-side comparisons of the submitted content and source material (highlighting similar text, code, or image regions), links to original sources, and visualizations of structural similarities.
- **Explanation Synthesis:** Converts the structured explanation trace from the symbolic reasoner into natural language. For example, if the symbolic reasoner determined "This text segment (neural match confidence 0.95) is a direct copy of Source A, violating Rule R1 (Direct Copying without Citation)," the explanation synthesis component would generate: "The system detected a high degree of textual similarity (95% confidence) between this paragraph and content from Source A.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

This constitutes direct copying without proper citation, violating academic integrity guidelines."

- **Confidence and Severity Scoring:** Assigns an overall confidence score to the plagiarism detection and a severity level (e.g., minor, moderate, severe) based on the type of plagiarism, extent, and intent (if inferable).
- **Pedagogical Feedback Formulation:** Provides actionable advice to the student based on the detected plagiarism, guiding them on how to improve their academic integrity practices (e.g., "Review guidelines on proper paraphrasing and citation for Source A.")
- **Output:** A structured, interactive plagiarism report.

V. MATHEMATICAL FORMULATIONS

This section outlines the key mathematical formulations underpinning the MNSAS architecture, particularly focusing on similarity measures, anomaly detection, and the integration of neural and symbolic components.

5.1. Multi-Modal Feature Embeddings

Let X_m be the raw input data for modality $m \in {text, code, image, audio, video}$. A feature extractor f_m transforms X_m into a high-dimensional embedding $E_m \in \mathbb{R}^{d_m}$.
$E_m = f_m(X_m)$

For text, using a Transformer-based model, the embedding for a sentence $S = (w_1, w_2, \dots, w_n)$ can be represented as the concatenation or pooling of token embeddings:

$E_{text} = \text{POOL}(\text{Transformer}(w_1, \dots, w_n))$

For images, a CNN extracts features:

$E_{image} = \text{CNN}(X_{image})$

5.2. Unified Multi-Modal Representation

To create a unified representation U from individual modality embeddings E_m, we employ a multi-modal fusion function g that maps all E_m into a shared embedding space $\mathbb{R}^D$.

$U = g(E_{text}, E_{code}, E_{image}, E_{audio}, E_{video})$

If using a cross-modal attention mechanism, the unified representation could be a weighted sum of modality-specific representations, where weights are learned via attention scores:

$U = \sum_{m \in \text{modalities}} \alpha_mE_m$, where $\alpha_m = \text{softmax}(\text{AttentionScore}(E_m, \text{Query}))$

For contrastive learning, a common objective is InfoNCE loss, which maximizes agreement between positive pairs (e.g., different modalities describing the same concept) and minimizes agreement for negative pairs:

$\mathcal{L}{NCE} = -\mathbb{E}{x,y} \left[\log \frac{\exp(\text{sim}(g(x), g(y)) / \tau)}{\sum_{y' \in Y} \exp(\text{sim}(g(x), g(y')) / \tau)} \right]$

where sim is a similarity function (e.g., cosine similarity), τ is a temperature parameter, and Y includes one positive pair y and multiple negative pairs y'.

5.3. Plagiarism Similarity Score

Given a unified representation of a student's submission U_{sub} and a unified representation of a potential source U_{src}, the plagiarism similarity score S_P can be calculated using cosine similarity:

$S_P(U_{sub}, U_{src}) = \frac{U_{sub} \cdot U_{src}}{|U_{sub}| |U_{src}|}$

This score ranges from -1 to 1, with 1 indicating perfect similarity. A threshold θ_S is applied to identify candidate instances:

$\text{Candidate}(U_{sub}, U_{src}) \text{ if } S_P(U_{sub}, U_{src}) \ge \theta_S$

5.4. Anomaly Detection (Unoriginality Score)

For unoriginality detection using an Autoencoder, the reconstruction error R_E for a submission U_{sub} is calculated as:

$R_E(U_{sub}) = |U_{sub} - \text{Decoder}(\text{Encoder}(U_{sub}))|^2$

Submissions with $R_E > \theta_R$ (a predefined threshold) are flagged as anomalous.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

5.5. AIGC Detection

For AIGC detection, a classifier C_{AIGC} (e.g., a neural network) is trained on features F_{AIGC} extracted from the submission (e.g., perplexity, burstiness, specific n-gram frequencies):

$P(\text{AIGC} | F_{AIGC}) = C_{AIGC}(F_{AIGC})$

If $P(\text{AIGC} | F_{AIGC}) > \theta_{AIGC}$, the submission is flagged as potentially AI-generated.

5.6. Symbolic Reasoning and Rule Application

The symbolic reasoning engine applies logical rules to the outputs of the neural modules. For example, a rule for direct copying might be:

$\text{Rule}{DirectCopy}: \text{IF } S_P(U{sub}, U_{src}) \ge \theta_S \text{ AND } \neg \text{HasCitation}(U_{sub}, U_{src}) \text{ THEN } \text{Flag}(\text{DirectCopying})$

More complex rules can involve multiple conditions and modalities. For instance, detecting structural code plagiarism might involve comparing AST similarity and variable naming conventions:

$\text{Rule}{CodeStruct}: \text{IF } S{AST}(U_{sub}, U_{src}) \ge \theta_{AST} \text{ AND } S_{VarNames}(U_{sub}, U_{src}) \ge \theta_{Var} \text{ AND } \neg \text{HasAttribution}(U_{sub}, U_{src}) \text{ THEN } \text{Flag}(\text{StructuralCodePlagiarism})$

These rules are typically expressed in a logic programming language (e.g., Prolog) or a rule-based system, allowing for transparent and auditable decision-making.

VI. EVALUATION METRICS AND EXPERIMENTAL DESIGN

Evaluating a multi-modal neuro-symbolic plagiarism detection system requires a comprehensive approach that considers not only detection accuracy but also the quality of explanations and robustness against adversarial attacks. We propose the following evaluation metrics and experimental design considerations:

6.1. Datasets

Developing a robust evaluation requires diverse and realistic datasets. Since no large-scale, publicly available multi-modal plagiarism dataset exists, a significant effort will be dedicated to its creation. This dataset will include:

- **Original Submissions:** A corpus of genuine student work across various disciplines and modalities (text essays, code projects, image-based assignments, audio/video presentations).
- **Plagiarized Submissions:** Artificially generated or human-crafted plagiarized versions of original submissions, covering different types of plagiarism (direct copy, paraphrasing, mosaic, structural, conceptual) and across modalities. This will include examples of cross-modal plagiarism (e.g., text plagiarized from an image, audio from a video).
- **AI-Generated Content (AIGC):** Submissions generated by various LLMs and generative AI models, designed to mimic student work.
- **Source Materials:** A corresponding database of potential source documents (web pages, academic papers, books, code repositories, image/audio/video databases) from which plagiarism instances are derived.

Each instance in the dataset will be meticulously annotated with ground truth labels indicating the presence, type, and source of plagiarism, as well as the specific segments involved across modalities.

6.2. Detection Accuracy Metrics

Standard classification metrics will be adapted for multi-modal plagiarism detection:

- **Precision, Recall, F1-score:** Calculated at different granularities (e.g., document-level, segment-level, modality-specific). For a given plagiarism type T, Precision = $\frac{\text{TP}_T}{\text{TP}_T + \text{FP}_T}$, Recall = $\frac{\text{TP}_T}{\text{TP}_T + \text{FN}_T}$, F1-score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$.
- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** Useful for evaluating the system's ability to distinguish between plagiarized and original content across various thresholds.
- **Mean Average Precision (mAP):** Particularly relevant for scenarios where multiple plagiarism instances might exist within a single submission.

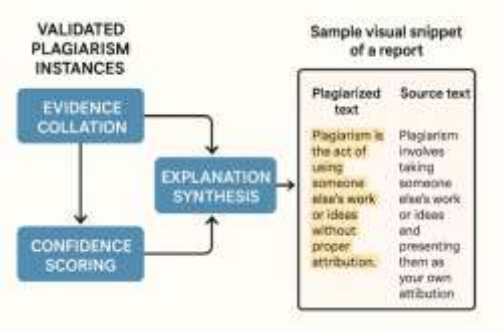
International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 483-499 www.ijaem.net ISSN: 2395-5252

- **Modality-Specific Metrics:** For example, for code plagiarism, metrics like AST similarity, token overlap, and control flow graph similarity will be used in addition to general text-based metrics.

6.3. Explainability Metrics

Evaluating the quality of explanations is challenging but crucial. We propose a combination of quantitative and qualitative metrics:

- **Fidelity:** How well the explanation reflects the true reasoning of the model. This can be measured by perturbing input features and observing changes in both the model's prediction and the explanation.
- **Sufficiency and Comprehensiveness:** Does the explanation provide enough information to understand the decision? This can be assessed via human evaluation (e.g., Likert scale ratings by educators).
- **Actionability:** Does the explanation provide clear guidance for students to correct their mistakes? Also assessed via human evaluation.
- **Consistency:** Do similar plagiarism instances receive similar explanations?
- **Compactness/Conciseness:** Is the explanation easy to understand without being overly verbose?



6.4. Robustness Metrics

Evaluating robustness against evasion techniques and AIGC:

- **Adversarial Attack Success Rate:** Measure how easily the system can be fooled by intentionally obfuscated or AI-generated content. This involves generating adversarial examples (e.g., using paraphrasing tools, synonym replacement, minor structural changes) and measuring the drop in detection accuracy.
- **AIGC Detection Rate:** The system's ability to correctly identify content generated by various LLMs and generative AI models.

6.5. Experimental Design

Experiments will be conducted in phases:

1. **Baseline Comparison:** Compare the MNSAS performance against state-of-the-art unimodal plagiarism detection systems and multi-modal content similarity detection systems.
2. **Ablation Studies:** Evaluate the contribution of each MNSAS component (e.g., impact of neuro-symbolic integration vs. purely neural, impact of specific modality feature extractors) by removing or simplifying them.
3. **Explainability User Studies:** Conduct human-in-the-loop studies with educators and students to assess the utility, clarity, and trustworthiness of the generated explanations.
4. **Robustness Testing:** Systematically test the MNSAS against various obfuscation techniques and AIGC models.

VII. RESEARCH QUESTIONS

This research aims to answer the following fundamental questions:

1. How can neuro-symbolic architectures effectively integrate multi-modal feature representations (text, code, image, audio, video) to achieve superior accuracy and robustness in plagiarism detection compared to unimodal or purely neural approaches?
2. What formal knowledge representations (ontologies, rules) are most effective for capturing academic integrity policies and plagiarism types, and how can they be seamlessly integrated with neural pattern recognition for explainable decision-making?
3. How can the MNSAS generate human-understandable and pedagogically valuable explanations for detected plagiarism instances across diverse modalities, and how do these

explanations impact user trust and learning outcomes?

4. To what extent can the neuro-symbolic approach enhance the system's resilience against sophisticated obfuscation techniques and the detection of AI-generated content in academic submissions?
5. What are the computational trade-offs and scalability considerations for deploying such a comprehensive multi-modal neuro-symbolic system in real-world academic environments?

VIII. CONCLUSION AND FUTURE WORK

This paper has presented the conceptual framework and detailed architecture for a Multi-Modal Neuro-Symbolic Assessment System (MNSAS) for explainable plagiarism detection. By synergistically combining the perceptual strengths of neural networks with the logical reasoning capabilities of symbolic AI, the proposed MNSAS aims to overcome the limitations of current systems, offering enhanced accuracy, robustness, and, crucially, transparency in academic integrity enforcement. The ability to process diverse modalities and provide human-understandable explanations for detected anomalies is paramount in fostering trust and promoting genuine learning in the evolving digital educational landscape.

Future work will focus on several key areas:

- **Dataset Construction:** Developing a large-scale, meticulously annotated multi-modal plagiarism dataset to facilitate robust training and evaluation of the MNSAS.
- **Prototype Implementation and Empirical Validation:** Building a functional prototype of the MNSAS and conducting extensive empirical studies to validate its performance against established baselines across various plagiarism types and modalities.
- **Refinement of Neuro-Symbolic Integration:** Exploring more sophisticated methods for seamless interaction between neural and symbolic components, including iterative feedback loops and dynamic knowledge graph construction.
- **User Experience and Interface Design:** Developing intuitive interfaces for educators and students to interact with the MNSAS, visualize explanations, and provide feedback.
- **Ethical AI Considerations:** Further research into the ethical implications of AI-driven academic integrity systems, including bias detection and mitigation, data privacy, and the psychological impact on students.
- **Real-world Deployment and Longitudinal Studies:** Collaborating with educational institutions to deploy the MNSAS in a controlled environment and conduct longitudinal studies to assess its long-term impact on academic integrity practices and student learning.

REFERENCES

[1]. International Center for Academic Integrity. (n.d.). The Fundamental Values of Academic Integrity. Retrieved from https://academicintegrity.org/fundamental-values

[2]. Dawson, P., & Sutherland-Smith, W. (2018). Plagiarism in the Digital Age: Preventing, Detecting and Combating Academic Dishonesty. Routledge.

[3]. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[4]. Sadasivam, R. S., & Suresh, R. M. (2019). A Survey on Multi-Modal Plagiarism Detection. International Journal of Recent Technology and Engineering (IJRTE), 8(2S11), 244-248.

[5]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.

[6]. Turnitin. (n.d.). How Turnitin Works. Retrieved from https://www.turnitin.com/how-it-works

[7]. Roy, C. K., & Cordy, J. R. (2007). A Survey on Software Plagiarism Detection. ACM Computing Surveys (CSUR), 39(4), 10.

[8]. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 421-443.

[9]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... &Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.

[10]. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2021). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2112.10752.

[11]. Alayrac, J. B., Donahue, J., Luc, P., Miech, K., Seitzer, I., Sharma, A., ... &

Zisserman, A. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. arXiv preprint arXiv:2204.14198.

[12]. Arandjelović, R., & Zisserman, A. (2017). Look, Listen and Learn. In Proceedings of the IEEE International Conference on Computer Vision (pp. 609-617).

[13]. Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., ... & Zhou, M. (2020). CodeBERT: A Pre-Trained Model for Programming Language Understanding and Generation. arXiv preprint arXiv:2002.09425.

[14]. Gunning, D., & Aha, D. W. (2019). DARPA Explainable Artificial Intelligence (XAI). AI Magazine, 40(2), 56-66.

[15]. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).

[16]. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

[17]. Garcez, A. S., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. AI Magazine, 41(1), 13-25.

[18]. Besold, T. R., & Poggio, T. (2020). Neuro-Symbolic AI: A New Frontier. AI Magazine, 41(1), 5-12.

[19]. Hitzler, P., & van Harmelen, F. (2020). Neuro-Symbolic AI: A Brief Introduction. KI - KünstlicheIntelligenz, 34(2), 1-5.

[20]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[21]. Liu, Y., Ott, M., Goyal, N., Du, J., Li, M., Ko, J., ... &Feichtenhofer, C. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[22]. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: Pretrained Contextualized Embeddings for Scientific Text. arXiv preprint arXiv:1903.10676.

[23]. Jakesch, M., Koller, A., & Gartner, M. (2023). AIGC-Detector: Towards Universal Detection of AI-Generated Content. arXiv preprint arXiv:2305.15016.

[24]. Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv:1609.02907.

[25]. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. arXiv preprint arXiv:1710.10903.

[26]. Guo, D., Feng, Z., Tang, D., Li, S., Gong, M., & Duan, N. (2020). GraphCodeBERT: Pre-training Code Representations with Structural Information. arXiv preprint arXiv:2009.08179.

[27]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

[28]. Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

[29]. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint arXiv:1905.11946.

[30]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... &Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[31]. Zauner, C. (2010). Implementation and Benchmarking of Perceptual Hashing. Master's thesis, University of Applied Sciences Upper Austria.

[32]. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788).

[33]. He, K., Gkioxari, G., Dollár, P., &Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961-2969).

[34]. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning

of Speech Representations. Advances in Neural Information Processing Systems, 33, 12449-12460.

[35]. Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451-3461.

[36]. Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

[37]. Arnab, A., Dehghani, F., Heigold, G., Sun, C., Lu, T., Schmid, C., & Lucic, M. (2021). ViViT: A Video Vision Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6836-6846).

[38]. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems, 32.

[39]. Li, L., Yatskar, M., Yin, K., Gan, Z., Yu, P., & Chang, B. (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv preprint arXiv:1908.03554.

[40]. Tan, H., & Bansal, M. (2019). LXMERT: Learning Cross-Modality Encoders for Language-Vision Tasks. arXiv preprint arXiv:1908.07490.

[41]. Chen, Y., Li, L., Yu, L., Sayama, K., Gan, Z., Cheng, Y., ... & Wang, L. (2020). UNITER: UNiversal Image-TExt Representation Learning. European Conference on Computer Vision (pp. 104-120).

[42]. Wu, Z., Pan, S., Chen, F., Long, G., Jiang, J., & Zhang, C. (2020). A Comprehensive Survey on Graph Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24.

[43]. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature Verification using a Siamese Time Delay Neural Network. In Advances in Neural Information Processing Systems (pp. 737-744).

[44]. Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114.

[45]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27.

[46]. Sadasivam, R. S., & Suresh, R. M. (2019). A Survey on Multi-Modal Plagiarism Detection. International Journal of Recent Technology and Engineering (IJRTE), 8(2S11), 244-248.

[47]. W3C. (2004). OWL Web Ontology Language Overview. Retrieved from https://www.w3.org/TR/owl-overview/

[48]. Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission.

[49]. Alayrac, J. B., Donahue, J., Luc, P., Miech, K., Seitzer, I., Sharma, A., ... & Zisserman, A. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. arXiv preprint arXiv:2204.14198.

[50]. Arandjelović, R., & Zisserman, A. (2017). Look, Listen and Learn. In Proceedings of the IEEE International Conference on Computer Vision (pp. 609-617).

[51]. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems, 33, 12449-12460.

[52]. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 421-443.

[53]. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: Pretrained Contextualized Embeddings for Scientific Text. arXiv preprint arXiv:1903.10676.

[54]. Besold, T. R., & Poggio, T. (2020). Neuro-Symbolic AI: A New Frontier. AI Magazine, 41(1), 5-12.

[55]. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature Verification using a Siamese Time Delay Neural Network. In Advances in Neural Information Processing Systems (pp. 737-744).

[56]. Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

[57]. Chen, Y., Li, L., Yu, L., Sayama, K., Gan, Z., Cheng, Y., ... & Wang, L. (2020). UNITER: UNiversal Image-TExt Representation Learning. European Conference on Computer Vision (pp. 104-120).

[58]. Dawson, P., & Sutherland-Smith, W. (2018). Plagiarism in the Digital Age: Preventing, Detecting and Combating Academic Dishonesty. Routledge.

[59]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[60]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... &Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[61]. Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., ... & Zhou, M. (2020). CodeBERT: A Pre-Trained Model for Programming Language Understanding and Generation. arXiv preprint arXiv:2002.09425.

[62]. Garcez, A. S., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. AI Magazine, 41(1), 13-25.

[63]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27.

[64]. Guo, D., Feng, Z., Tang, D., Li, S., Gong, M., & Duan, N. (2020). GraphCodeBERT: Pre-training Code Representations with Structural Information. arXiv preprint arXiv:2009.08179.

[65]. Gunning, D., & Aha, D. W. (2019). DARPA Explainable Artificial Intelligence (XAI). AI Magazine, 40(2), 56-66.

[66]. He, K., Gkioxari, G., Dollár, P., &Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961-2969).

[67]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

[68]. Hitzler, P., & van Harmelen, F. (2020). Neuro-Symbolic AI: A Brief Introduction. KI - KünstlicheIntelligenz, 34(2), 1-5.

[69]. Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission.

[70]. Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451-3461.

[71]. International Center for Academic Integrity. (n.d.). The Fundamental Values of Academic Integrity. Retrieved from https://academicintegrity.org/fundamental-values

[72]. Jakesch, M., Koller, A., & Gartner, M. (2023). AIGC-Detector: Towards Universal Detection of AI-Generated Content. arXiv preprint arXiv:2305.15016.

[73]. Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114.

[74]. Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv:1609.02907.

[75]. Li, L., Yatskar, M., Yin, K., Gan, Z., Yu, P., & Chang, B. (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv preprint arXiv:1908.03554.

[76]. Liu, Y., Ott, M., Goyal, N., Du, J., Li, M., Ko, J., ... &Feichtenhofer, C. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[77]. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems, 32.

[78]. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

[79]. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
[80]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... &Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.
[81]. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2021). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2112.10752.
[82]. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788).
[83]. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).
[84]. Roy, C. K., & Cordy, J. R. (2007). A Survey on Software Plagiarism Detection. ACM Computing Surveys (CSUR), 39(4), 10.
[85]. Sadasivam, R. S., & Suresh, R. M. (2019). A Survey on Multi-Modal Plagiarism Detection. International Journal of Recent Technology and Engineering (IJRTE), 8(2S11), 244-248.
[86]. Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.
[87]. Tan, H., & Bansal, M. (2019). LXMERT: Learning Cross-Modality Encoders for Language-Vision Tasks. arXiv preprint arXiv:1908.07490.
[88]. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint arXiv:1905.11946.
[89]. Turnitin. (n.d.). How Turnitin Works. Retrieved from https://www.turnitin.com/how-it-works
[90]. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. arXiv preprint arXiv:1710.10903.
[91]. W3C. (2004). OWL Web Ontology Language Overview. Retrieved from https://www.w3.org/TR/owl-overview/
[92]. Wu, Z., Pan, S., Chen, F., Long, G., Jiang, J., & Zhang, C. (2020). A Comprehensive Survey on Graph Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24.
[93]. Zauner, C. (2010). Implementation and Benchmarking of Perceptual Hashing. Master's thesis, University of Applied Sciences Upper Austria.