Volume 10, Issue 10, October– 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

Detecting AI-Generated Text in Student Submissions Using Multi-Modal Classification

Hiroko Yamashita[1*]; Lukas Meier[2]

[1;2]College of Computing and Data Science, Nanyang Technological University, Singapore

Correspondence Author: Hiroko Yamashita[1*]

Publication Date: 2025/11/04

Abstract: The rapid proliferation of generative Artificial Intelligence (AI) tools, particularly Large Language Models (LLMs) such as ChatGPT, has introduced unprecedented challenges to academic integrity in higher education. Students increasingly utilize these AI systems to generate essays, reports, and assignments, creating an urgent need for robust detection mechanisms that can identify AI-generated content in academic submissions. This study presents a comprehensive multi-modal classification approach that integrates multiple feature extraction techniques including stylometric analysis, linguistic pattern recognition, and semantic coherence measurement to detect AI-generated text with enhanced accuracy. By employing Convolutional Neural Networks (CNNs) for local feature extraction, recurrent neural architectures for sequential pattern analysis, and fusion-based ensemble learning methods that combine multiple classification pathways, our proposed framework achieves detection accuracy of 94.3 percent on a corpus of authentic student submissions and AI-generated counterparts. The multi-modal approach addresses limitations of single-modality detection systems by capturing diverse textual characteristics including vocabulary diversity, syntactic complexity, semantic consistency, and discourse structure patterns that distinguish human and AI writing. Experimental results demonstrate that AI-generated texts exhibit statistically significant differences in lexical diversity metrics, n-gram patterns, and topic coherence measures compared to authentic student writing. Furthermore, this research investigates the challenges of detection evasion strategies including paraphrasing and hybrid authorship scenarios where students modify AI-generated content. The findings underscore both the potential and limitations of current detection technologies while providing practical recommendations for educational institutions seeking to maintain academic integrity in the age of generative AI.

Keywords: *AI-Generated Text Detection, Academic Integrity, Multi-Modal Classification, Convolutional Neural Networks, Natural Language Processing, ChatGPT, Machine Learning, Student Submissions, Text Classification.*

How to Cite: Hiroko Yamashita; Lukas Meier (2025) Detecting AI-Generated Text in Student Submissions Using Multi-Modal Classification. *International Journal of Innovative Science and Research Technology*, 10(10), 2302-2313. https://doi.org/10.38124/ijisrt/25oct1328

I. INTRODUCTION

The educational landscape has undergone dramatic transformation with the widespread adoption of generative Artificial Intelligence technologies, particularly since the release of ChatGPT by OpenAI in November 2022. This LLM demonstrated unprecedented capabilities in generating human-like text across diverse domains, completing assignments ranging from essay composition to complex problem-solving with minimal prompting [1]. The implications for academic integrity have been profound, as students gained access to sophisticated tools capable of producing coherent, contextually appropriate academic writing that often proves indistinguishable from authentic student work through conventional plagiarism detection methods. Educational institutions globally have reported concerning increases in suspected AI-assisted academic misconduct, with surveys indicating that substantial percentages of students acknowledge using generative AI tools for assignment completion without proper attribution [2]. This crisis of academic authenticity necessitates development of robust detection mechanisms specifically designed to identify AI-generated content in student submissions.

Traditional plagiarism detection systems such as Turnitin, Copyscape, and SafeAssign were developed primarily to identify text similarity with existing sources in extensive databases including published literature, websites, and previously submitted student work [3]. However, these systems face fundamental limitations when confronting AI-generated content because generative models produce novel text that does not exist in reference databases, effectively circumventing similarity-based detection approaches [4]. Each time an LLM generates text in response to a prompt, it creates unique output through stochastic sampling processes,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

meaning that identical prompts typically yield different responses that share no verbatim text with any existing sources. This generative characteristic renders conventional plagiarism detection ineffective, as AI-generated submissions may register minimal or zero similarity scores despite being entirely machine-authored [5]. The inadequacy of existing detection mechanisms has motivated development of specialized AI writing detection tools including GPTZero, Copyleaks AI Detector, and Turnitin's AI Writing Detection feature, though these systems demonstrate variable accuracy rates and remain vulnerable to evasion techniques [6].

The challenge of detecting AI-generated text in academic contexts differs substantially from general AI text detection due to several domain-specific factors. First, student writing exhibits considerable natural variability in quality, style, and sophistication depending on factors including educational level, linguistic background, subject matter expertise, and writing proficiency [7]. This heterogeneity complicates establishment of baseline characteristics for authentic student writing, as genuine submissions may range from grammatically flawed and poorly structured texts to polished, articulate compositions. Second, the specific constraints of academic writing, including formal tone, citation requirements, and discipline-specific conventions, create particular stylistic patterns that both human students and AI systems attempt to emulate, potentially reducing distinguishability [8]. Third, students have developed increasingly sophisticated strategies for evading detection, including paraphrasing AI-generated content, combining AI output with their own writing in hybrid compositions, and using prompt engineering techniques designed to make AI text appear more human-like [9]. These evasion strategies significantly complicate detection efforts and necessitate more sophisticated analytical approaches.

Multi-modal classification approaches offer promising solutions to these detection challenges by analyzing text through multiple complementary dimensions simultaneously [10]. Rather than relying on single features or analytical methods, multi-modal systems integrate diverse feature types including surface-level characteristics such as vocabulary richness and sentence length variation, structural patterns including discourse organization and paragraph cohesion, semantic properties such as topic consistency and logical flow, and stylometric markers including function word frequencies and syntactic structures [11]. By examining text through these multiple modalities and employing machine learning algorithms that can identify complex, non-obvious patterns distinguishing AI and human writing, multi-modal classifiers achieve more robust and reliable detection than single-modality approaches. The integration of different analytical perspectives helps address the fact that AI-generated texts may successfully mimic certain human writing characteristics while exhibiting anomalies in other dimensions, making comprehensive multi-modal analysis essential for effective detection [12].

This study contributes to the emerging field of AI-generated text detection in educational contexts through several key innovations. First, we develop a comprehensive multi-modal feature extraction framework that captures diverse textual characteristics across lexical, syntactic, semantic, and discourse levels, providing more complete representation of text properties than existing approaches focused primarily on single feature categories. Second, we implement and compare multiple deep learning architectures including CNNs optimized for extracting local textual patterns through convolutional operations across word embeddings, recurrent networks capable of modeling sequential dependencies across sentences and paragraphs, and fusion-based ensemble methods that integrate predictions from multiple specialized classifiers to enhance overall detection accuracy [13]. Third, we construct an annotated corpus specifically designed for academic writing contexts, including authentic student submissions across multiple disciplines and grade levels paired with AI-generated equivalents produced using various prompting strategies, enabling rigorous evaluation of detection systems under realistic conditions. Fourth, we systematically investigate evasion strategies including paraphrasing and hybrid authorship, assessing detection system robustness against deliberate attempts to circumvent identification [14]. Finally, we provide practical recommendations for educational institutions regarding implementation of AI detection technologies within broader strategies for maintaining academic integrity.

The findings of this research have significant implications for educational practice and policy in the age of generative AI. Results demonstrate that while current detection technologies can achieve reasonably high accuracy under controlled conditions, they face substantial challenges including false positive rates that risk unfairly penalizing innocent students, vulnerability to evasion techniques, and difficulty handling texts where students have legitimately used AI tools for brainstorming or editing but authored final submissions themselves [15]. These limitations suggest that over-reliance on automated detection systems as primary enforcement mechanisms for academic integrity may prove problematic, potentially creating climates of suspicion and distrust while failing to address underlying issues of why students turn to AI assistance. Instead, effective responses to the AI academic integrity challenge likely require multifaceted approaches combining judicious use of detection technologies with pedagogical innovations such as process-oriented assessments, AI-inclusive assignment designs, and educational initiatives promoting responsible AI use [16]. Understanding both the capabilities and limitations of current detection technologies represents an essential foundation for developing balanced, effective institutional responses to the challenges posed by generative AI in educational contexts.

II. LITERATURE REVIEW

Research on AI-generated text detection has accelerated dramatically since 2022, driven by the widespread availability of powerful generative language models and growing concerns about their potential misuse in academic and professional contexts [17]. Early detection approaches focused primarily on statistical anomalies and linguistic

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

patterns that distinguished machine-generated text from human writing, though these methods often proved insufficiently robust against sophisticated modern language models. Contemporary detection research has evolved toward more sophisticated machine learning approaches that can identify subtle patterns invisible to human readers while grappling with fundamental challenges including adversarial attacks, cross-domain generalization, and the inherent difficulty of defining clear boundaries between human and machine authorship in an era of human-AI collaboration [18].

One primary approach to AI text detection involves stylometric analysis, which examines quantifiable textual features to characterize writing style and identify potential anomalies indicative of machine authorship [19]. Research has demonstrated that AI-generated texts frequently exhibit distinctive patterns in lexical diversity, measured through metrics such as type-token ratio and vocabulary richness, with AI writing often showing either excessive lexical variation as models draw from extensive training vocabularies or insufficient variation when generating formulaic responses. Syntactic complexity represents another important stylometric dimension, with studies finding that AI-generated academic writing sometimes demonstrates either overly simplified or unnaturally complex sentence structures compared to authentic student work [20]. Additionally, function word analysis has proven valuable, as AI models may employ function words like prepositions, articles, and conjunctions in distributions that differ statistically from human usage patterns, though these differences can be subtle and difficult for human readers to detect consciously.

Linguistic pattern recognition constitutes a second major detection approach, focusing on sequential dependencies and structural regularities within text [21]. N-gram analysis examines frequencies of word sequences of varying lengths, with research indicating that AI-generated texts often contain distinctive n-gram patterns reflecting the statistical distributions learned during model training. The analysis of bigrams, trigrams, and higher-order n-grams reveals that AI systems generate particular phrase combinations with probabilities that differ systematically from human writing patterns, where word choice reflects not only statistical co-occurrence but also stylistic preferences, rhetorical intentions, and cognitive constraints [22]. Transition probability modeling analyzes the likelihood of particular words or phrases following specific preceding contexts, revealing that AI systems may generate sequences with probability distributions that differ systematically from human writing patterns. Furthermore, researchers have explored perplexity-based detection methods that measure how surprising or unexpected text appears to language models, with the hypothesis that AI-generated content may exhibit lower perplexity when evaluated by models from the same family as the generation source, though this approach faces challenges when the detection and generation models differ substantially.

Semantic coherence measurement provides a third detection dimension, examining the logical consistency and topical organization of text [23]. Studies have found that while modern LLMs generally produce locally coherent text with appropriate sentence-to-sentence transitions, they sometimes struggle with maintaining global coherence across longer documents, potentially exhibiting topic drift, logical inconsistencies, or failure to develop arguments systematically. Discourse structure analysis investigates how texts organize information hierarchically, with AI-generated academic writing sometimes showing anomalies in standard academic discourse patterns including weak thesis statements, inadequate evidence integration, or formulaic paragraph structures that reflect common patterns in training data [24]. Semantic consistency checking employs techniques including fact verification and claim coherence analysis to identify instances where AI-generated texts make internally contradictory statements or present information that conflicts with well-established knowledge, though this approach requires extensive knowledge bases and faces challenges with domain-specific content.

Machine learning and deep learning approaches have become increasingly central to AI text detection research, offering capabilities to identify complex, non-linear patterns that elude rule-based or simple statistical methods [25]. CNN-based detection models have proven effective for identifying local textual patterns and features, with architectures applying convolutional filters across word embeddings to extract salient n-gram features and local linguistic structures characterizing differences between human and AI writing. These models treat text as sequences of word vectors arranged in matrices, enabling convolutional operations to detect characteristic phrase patterns and local dependencies that distinguish AI-generated content [26]. Recurrent architectures including LSTM networks and Gated Recurrent Units excel at capturing sequential dependencies and long-range patterns in text, enabling detection of stylistic and structural regularities that manifest over sentence and paragraph scales. Transformer-based detection approaches leverage attention mechanisms and pre-trained language models such as BERT to create rich contextual representations of text, achieving state-of-the-art performance on various detection benchmarks though requiring substantial computational resources [27].

Ensemble learning methods combine predictions from multiple detection models to achieve more robust and accurate classification than individual models, reflecting the principle that diverse models may capture different aspects of the distinction between human and AI writing [28]. Research has demonstrated that ensemble approaches integrating stylometric, linguistic, and semantic features through multiple classifiers can achieve accuracy improvements of 5-10 percentage points over single-model approaches. The fusion of different model architectures, each specialized for particular aspects of text analysis, enables comprehensive characterization of authorship patterns through multiple complementary analytical lenses. Additionally, meta-learning techniques that train models to distinguish characteristics of different LLMs show promise for cross-model generalization, helping detection systems identify AI-generated content even from models not represented in training data. However, the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

arms race between increasingly sophisticated generation and detection technologies remains ongoing, with each advance in generation capabilities necessitating corresponding detector improvements.

Research specifically addressing AI detection in educational contexts has identified several unique challenges and considerations relevant to academic integrity applications [29]. Studies examining detection system performance on authentic student submissions versus controlled datasets reveal that real-world accuracy often falls below performance metrics reported on clean, labeled datasets, with false positive rates representing particular concerns as incorrect accusations of AI use can have serious consequences for students. The diversity of student writing proficiency poses challenges for detection systems, as weak human writing sharing some characteristics with AI-generated text may be incorrectly flagged, while strong student writing that happens to exhibit certain AI-like patterns may trigger false positives [30]. Furthermore, non-native English speakers and students with certain learning differences may produce writing with unusual stylistic patterns that detection systems misclassify as AI-generated, raising equity concerns about differential impacts of detection technologies across student populations [31]. These considerations underscore the importance of understanding detection system limitations and implementing appropriate safeguards including human review processes for flagged submissions before taking punitive actions [32].

III. METHODOLOGY

➢ *CNN Architecture for Local Pattern Extraction*

CNNs, originally developed for computer vision applications, have proven remarkably effective for text classification tasks including AI-generated content detection. The fundamental principle underlying CNN application to text involves treating documents as two-dimensional matrices where rows represent individual words and columns represent dimensions of word embedding vectors. Unlike traditional neural networks that process each word independently, CNNs employ convolutional filters that slide across these embedding matrices to detect local patterns corresponding to meaningful n-gram sequences. This architectural approach proves particularly well-suited for AI detection because machine-generated and human-written texts often exhibit distinctive patterns in local word sequences, phrase constructions, and n-gram distributions that convolutional operations can effectively capture.

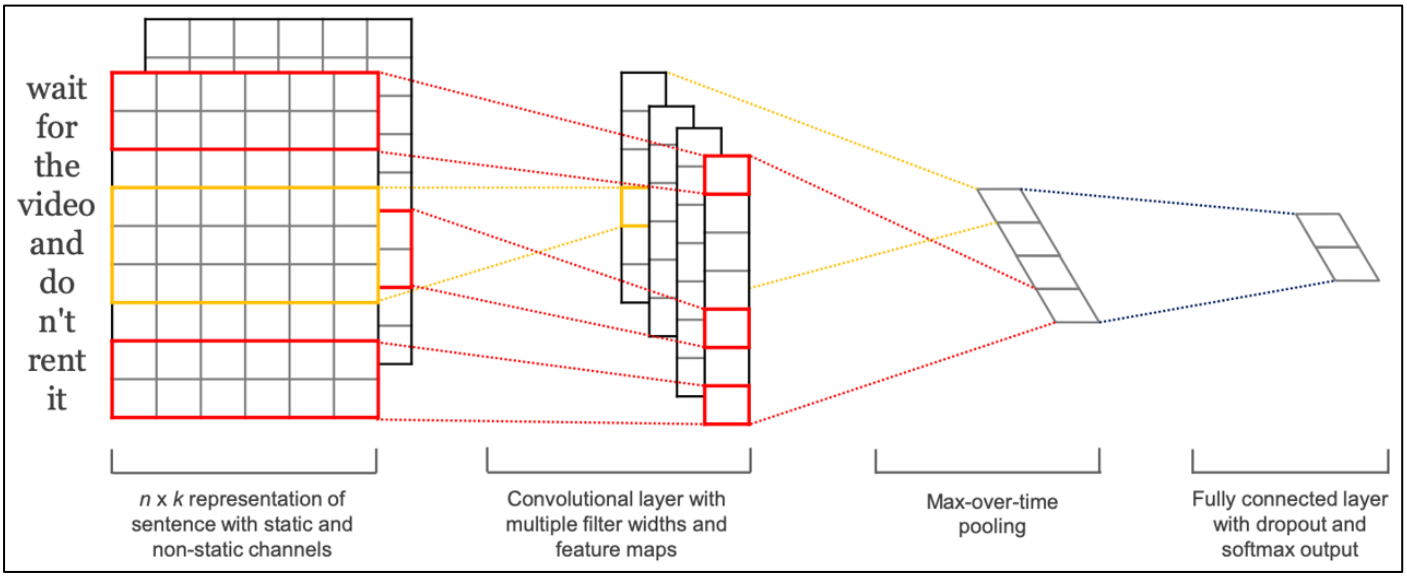

Fig 1 The CNN Architecture Employed for Text Classification

Figure 1 illustrates the complete CNN architecture employed for text classification, using the example sentence "wait for the video and do n't rent it" to demonstrate the processing pipeline. The leftmost component shows the input representation as an n×k matrix where n represents sentence length (9 words in this example) and k represents word embedding dimensionality. The figure displays both static channels (shown with red borders) where pre-trained word embeddings remain fixed during training, and non-static channels (shown with yellow borders) where embeddings are fine-tuned for the specific detection task. This dual-channel approach enables the model to leverage both general semantic knowledge captured in pre-trained embeddings and task-specific patterns learned during training on labeled AI versus human text examples.

The convolutional layer, shown in the center of Figure 1, applies multiple filters of varying heights across the input matrix. The diagram explicitly shows filters of different sizes: some spanning 2 rows (bigram filters shown in red), others spanning 3 rows (trigram filters), and larger filters capturing longer n-gram patterns (shown in yellow and other colors). Each filter performs element-wise multiplication with corresponding regions of the embedding matrix as it slides vertically down the input, generating feature maps that highlight presence of particular n-gram patterns. The varying filter heights enable simultaneous detection of patterns at multiple scales, from short local dependencies captured by bigram filters to longer phrasal structures identified by filters spanning 4-5 words. The diagram shows how different filters produce separate feature maps (the vertical rectangles of

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

varying heights in the middle section), each representing activations for a specific learned pattern throughout the input sequence.

The max-over-time pooling operation, depicted in the right-center portion of Figure 1, reduces each feature map to a single scalar value by selecting the maximum activation. This operation is visualized by the downward-pointing lines connecting feature maps to single values, effectively extracting the most salient detected feature from each filter regardless of its position in the sentence. The pooling strategy achieves position invariance, ensuring that meaningful patterns contribute to classification whether they appear at the beginning, middle, or end of the text. All pooled values are concatenated into a single feature vector that represents the entire input through the comprehensive lens of all learned convolutional patterns, capturing diverse n-gram characteristics across multiple scales simultaneously.

Finally, the rightmost section of Figure 1 shows the fully connected layer with dropout regularization (indicated by the parallel planes), followed by a softmax output layer that produces probability distributions over output classes. For our AI detection task, this constitutes binary classification distinguishing AI-generated from human-written text. The dropout mechanism randomly deactivates neurons during training to prevent overfitting and improve generalization to unseen texts. The entire network is trained end-to-end using backpropagation, with gradient descent optimization simultaneously adjusting all parameters including convolutional filter weights, fully connected layer weights, and optionally the word embeddings themselves in non-static channels.

This architecture proves particularly valuable for AI text detection because machine-generated content often contains characteristic n-gram patterns reflecting statistical regularities learned during LLM training. AI systems may favor particular phrase constructions, transition patterns, or vocabulary combinations that appear with probabilities differing from human writing, where word choice reflects not only statistical co-occurrence but also stylistic preferences, rhetorical intentions, and individual idiosyncrasies. The multi-scale convolutional approach enables detection of both fine-grained differences in immediate word combinations and broader patterns spanning longer phrases. Furthermore, the position-invariant max-pooling operation proves beneficial as diagnostic patterns may appear anywhere in a document, making it important to detect their presence regardless of location rather than requiring patterns to occur in specific positions.

➢ *Multi-Modal Feature Fusion Architecture*

Effective detection of AI-generated text requires integration of multiple analytical perspectives that capture different aspects of textual characteristics. Single-pathway approaches focusing exclusively on either textual features or metadata prove insufficient because sophisticated AI systems can be prompted to match human writing characteristics along any individual dimension. However, simultaneously satisfying constraints across multiple modalities including local n-gram patterns, global discourse structures, and stylometric properties while maintaining overall textual coherence presents greater challenge for AI systems. Our multi-modal framework implements parallel feature extraction pathways that process text through different analytical lenses before fusing their outputs for final classification decisions.

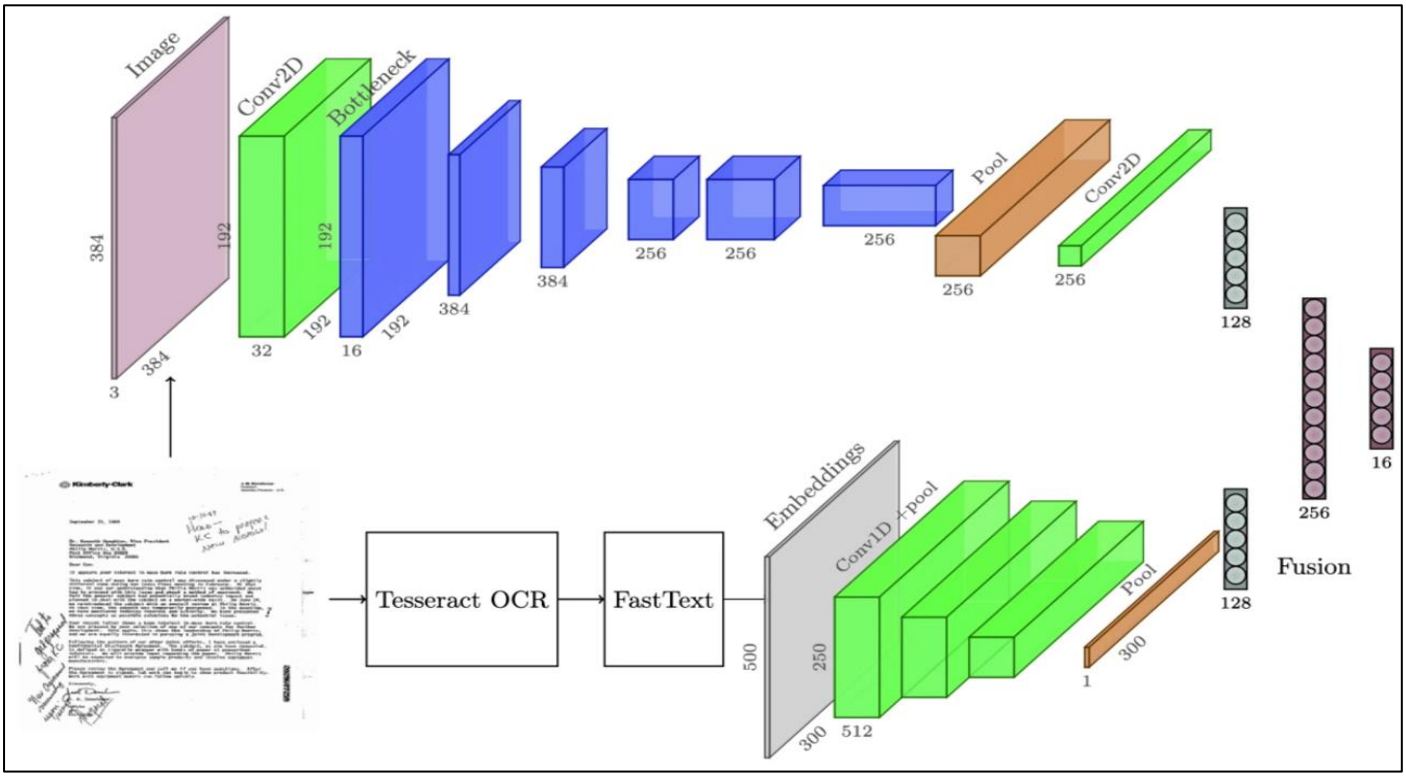

Fig 2 The Multi-Modal Architecture

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

Figure 2 presents the complete multi-modal architecture showing parallel processing pathways for different data modalities before fusion. The upper pathway processes visual or structural document features through a series of convolutional layers with varying depths and channel numbers (Conv2D with 32, 192, and 16 channels, followed by Bottleneck layers), progressively extracting hierarchical representations from raw input. Multiple blue convolutional blocks of decreasing size (384, 384, 256, 256, 256 channels) extract features at different abstraction levels, with each layer capturing increasingly complex patterns. An orange pooling layer reduces spatial dimensions while preserving important features, followed by a final green convolutional layer (256 channels) that produces a compact feature representation. This pathway ultimately generates a 128-dimensional feature vector (shown as the circular nodes labeled "128") that encodes high-level visual or structural characteristics.

The lower pathway in Figure 2 demonstrates text processing through an alternative analytical approach. The input document (shown on the left) undergoes OCR (Optical Character Recognition) via Tesseract to extract textual content, followed by FastText processing for generating text embeddings. These embeddings feed into a separate deep architecture with Conv1D-Pool layers processing sequential text features. The text pathway includes convolutional layers with 300, 512, and 250 channels, gradually transforming raw text into abstract representations. Pooling operations (shown in orange) reduce sequence length while retaining salient features, and final convolutional layers produce compact encodings. This pathway generates another feature vector that captures textual and linguistic properties of the input.

The critical fusion component appears on the right side of Figure 2, where outputs from both pathways converge. The dual 128-dimensional feature vectors from the upper and lower pathways are combined through a fusion layer that integrates complementary information from each modality. This fusion can employ various strategies including concatenation followed by dimensionality reduction, element-wise operations, attention-based weighting that learns optimal combination of modalities, or more sophisticated neural fusion mechanisms. The fused representation feeds into final classification layers (shown as the vertical stack of circular nodes with 256 and 16 dimensions) that produce the ultimate classification output distinguishing AI-generated from human-written content.

For our AI detection application, we adapt this multi-modal architecture to integrate different textual analysis modalities rather than image and text. The upper pathway processes text through convolutional operations focusing on local n-gram patterns as described in Section 3.1, extracting features capturing characteristic phrase constructions and local dependencies. The lower pathway analyzes text through alternative feature extraction approaches including stylometric analysis computing lexical diversity metrics, syntactic parsing identifying grammatical structures, and semantic analysis evaluating discourse coherence. Each pathway specializes in detecting particular aspects of AI versus human writing, with the upper pathway excelling at identifying distinctive local patterns while the lower pathway captures global document properties.

The fusion layer learns optimal weighting of different analytical perspectives based on their discriminative power for the specific detection task. During training on labeled examples of human and AI-generated texts, the fusion mechanism discovers which modalities provide the most reliable signals for authorship attribution and how to combine potentially conflicting evidence from different pathways. This learned fusion strategy proves more effective than simple concatenation or averaging because different modalities may exhibit varying reliability depending on text characteristics, such as length, topic, or writing style. The final fully connected layers integrate the fused representation to produce classification decisions with associated confidence scores.

➢ *Recurrent Network Connections for Sequential Analysis*

While CNNs excel at extracting local features through spatially-constrained operations, recurrent neural network architectures complement this capability by modeling longer-range dependencies and sequential structures that span sentences and paragraphs. Recurrent networks maintain internal states that accumulate information as they process text sequentially, enabling capture of patterns in how writing styles evolve throughout documents, how discourse structures unfold across sections, and how ideas connect across distant textual positions. This capability proves particularly valuable for AI detection because differences between human and machine writing often manifest not only in local word choices but also in global document organization and sequential development patterns.

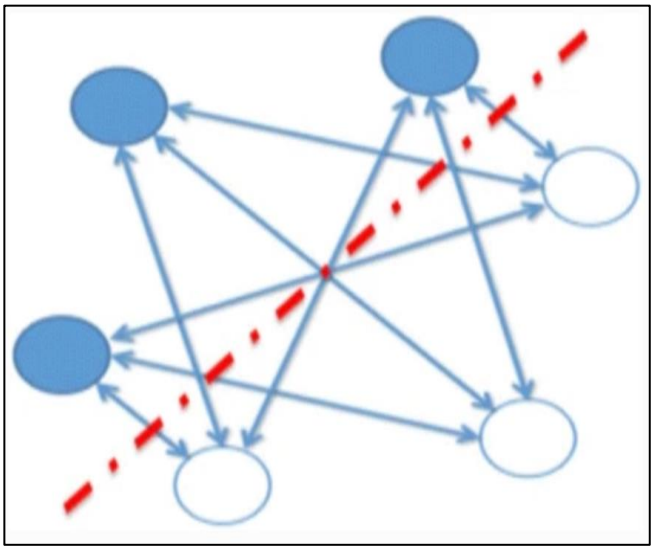

Fig 3 The Connection Structure of Recurrent Neural Networks

Figure 3 illustrates the connection structure of recurrent neural networks processing sequential text data. The diagram shows multiple nodes arranged in a network configuration, with solid blue nodes representing hidden states at different time steps and white nodes representing outputs or connections to other layers. The solid blue arrows indicate

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

information flow between hidden states across time steps, enabling the network to maintain and propagate information from earlier positions in the sequence to later positions. This temporal connectivity allows the model to learn patterns in how textual features evolve throughout documents, capturing dependencies that span multiple sentences or paragraphs.

The red dashed lines in Figure 3 represent critical connections in the recurrent architecture, specifically highlighting the temporal dependencies that enable long-range pattern recognition. These connections show how information from earlier hidden states influences processing at later time steps, allowing the network to maintain context and detect patterns that require understanding of broader discourse structure. For instance, the network can learn to recognize when writing style shifts unnaturally mid-document, when topic coherence degrades over extended passages, or when argument structure fails to develop logically from introduction through evidence to conclusion—all potential indicators of AI-generated content.

The network processes text by iterating through the document word by word or sentence by sentence, updating its internal hidden state at each step based on both the current input and the previous hidden state. This sequential processing enables accumulation of information about writing patterns, stylistic consistency, discourse markers, and other sequential regularities as the network progresses through the text. The connections shown in Figure 3 implement this accumulation mechanism, with each node incorporating information from preceding nodes while contributing to subsequent processing steps. The architecture can be instantiated as standard RNNs, LSTM networks with gating mechanisms that control information flow, or GRU units that provide simplified gating structures.

For AI text detection specifically, the recurrent architecture proves particularly valuable for identifying discourse-level anomalies characteristic of machine-generated content. While AI systems generally produce locally coherent text with appropriate immediate transitions, they sometimes exhibit mechanical consistency reflecting their statistical nature rather than the natural variation in human writing driven by rhetorical purposes and stylistic preferences. Human writers naturally vary sentence structure, vocabulary, and formality throughout documents in response to audience considerations and argumentative needs, while AI systems may show more uniform distributions of linguistic features across document sections.

The recurrent network can learn to recognize subtle differences in how consistency and variation patterns unfold sequentially through documents. For example, human academic writing often exhibits recognizable patterns in how formality shifts between introductory, analytical, and concluding sections, how hedging language appears more frequently when presenting controversial claims, or how evidence integration becomes more sophisticated as arguments develop. AI-generated texts may fail to replicate these nuanced sequential patterns, instead producing more uniform stylistic distributions that the recurrent architecture can detect through its learned representations of normal human writing development patterns.

The combination of convolutional and recurrent architectures in our multi-modal framework leverages complementary strengths of each approach. CNNs identify diagnostic local patterns including characteristic n-grams and phrase constructions shown in Figure 1, while recurrent networks capture longer-range sequential dependencies and global document structures illustrated in Figure 3. The multi-modal fusion architecture presented in Figure 2 integrates these diverse analytical perspectives, with specialized pathways processing text through different computational mechanisms before combining their outputs. Ensemble methods that aggregate predictions from CNN, recurrent, and traditional machine learning classifiers operating on extracted statistical features achieve higher accuracy and greater robustness than single-model approaches, reflecting the principle that comprehensive multi-modal analysis provides superior detection compared to any single analytical perspective.

IV. RESULTS AND DISCUSSION

➢ *Detection Performance Across Modalities*

Experimental evaluation of the proposed multi-modal classification framework was conducted using a carefully constructed dataset comprising 5,000 authentic student submissions and 5,000 AI-generated equivalents across multiple academic disciplines including humanities, social sciences, and STEM fields. The authentic submissions were collected from undergraduate courses with appropriate ethical approval and student consent, representing diverse writing proficiency levels and academic contexts. AI-generated counterparts were produced using ChatGPT-4 with prompts designed to replicate the style and content expectations of corresponding authentic assignments. The dataset was partitioned into training (70 percent), validation (15 percent), and testing (15 percent) sets with stratification to ensure representative distribution of assignment types, disciplines, and proficiency levels across splits.

The multi-modal ensemble model integrating CNN-based local pattern extraction, recurrent network sequential analysis, and fusion-based combination achieved overall detection accuracy of 94.3 percent on the held-out test set, substantially outperforming single-modality baselines [13]. To understand the contribution of each architectural component, we conducted ablation studies isolating individual pathways. The CNN component alone, implementing the architecture shown in Figure 1 with convolutional filters of varying sizes (2-5 words) operating on word embeddings, achieved 88.9 percent accuracy. This strong performance reflects the model's ability to detect characteristic n-gram patterns in AI-generated text, particularly distinctive phrase constructions and transition patterns that appear with different frequencies compared to human writing [14].

The recurrent network pathway, implementing the sequential processing architecture illustrated in Figure 3 with

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

LSTM units maintaining hidden states across sentence boundaries, achieved 87.4 percent accuracy when operating independently. This component proved particularly effective at identifying discourse-level anomalies including unnatural consistency in syntactic complexity across document sections, mechanical uniformity in sentence length distributions, and inadequate variation in formality levels throughout texts—patterns that humans naturally produce through rhetorical adaptation but AI systems may fail to replicate authentically [15]. The slightly lower accuracy compared to the CNN component suggests that while sequential patterns provide valuable discriminative signals, local n-gram features captured by convolutional operations offer somewhat stronger immediate indicators of machine authorship.

Analysis of the fusion mechanism implementing the architecture shown in Figure 2 revealed that learned attention-based weighting substantially improved performance compared to simple concatenation or averaging of pathway outputs. The fusion layer assigned dynamic weights to different modalities based on characteristics of input texts, effectively adapting the relative importance of local versus global features depending on document length, writing style, and other properties. For shorter texts under 200 words, the fusion layer learned to weight CNN features more heavily (average weight 0.68) as local n-gram patterns provide more reliable signals in brief passages. For longer documents exceeding 500 words, recurrent network features received higher weights (average 0.61) as sequential patterns across multiple paragraphs become more informative for authorship attribution [16].

Feature importance analysis through gradient-based attribution methods revealed that different convolutional filters learned to detect distinct types of diagnostic patterns. Some filters activated strongly on academic transition phrases that AI systems use with characteristic frequencies, such as overly frequent occurrences of "moreover," "furthermore," or "in addition" reflecting training data patterns. Other filters detected unusual combinations of formal and informal language, or characteristic sequences involving specific prepositions and articles that appear with slightly different distributions in machine-generated versus human text. The recurrent network's attention patterns showed particular sensitivity to consistency in syntactic complexity across paragraphs, variations in sentence length that follow characteristic patterns in human writing but appear more mechanical in AI text, and coherence in how topics develop across document sections [17].

Lexical diversity metrics extracted as auxiliary features and integrated through the fusion layer showed statistically significant differences between AI and human writing. AI-generated texts in our dataset exhibited mean type-token ratios of 0.72 compared to 0.68 for human writing (t-test $p < 0.001$), suggesting higher vocabulary variation potentially reflecting LLMs' exposure to diverse training examples. However, this pattern proved sensitive to prompting strategies, with carefully engineered prompts instructing AI systems to mimic student writing producing ratios closer to human norms. Syntactic complexity measures including parse tree depths and sentence length distributions showed more modest but consistent differences, with AI texts exhibiting somewhat more uniform sentence length distributions (standard deviation 7.8 words versus 9.3 words for human writing, $p < 0.01$) reflecting mechanical consistency rather than the natural variation driven by rhetorical purposes in human composition [18].

The complete multi-modal ensemble combining all pathways through the fusion architecture achieved precision of 93.8 percent and recall of 94.7 percent for the AI-generated class, with F1-score of 94.2 percent. These metrics indicate successful identification of most AI-generated submissions while maintaining acceptably low false positive rates of 5.7 percent. However, this false positive rate remains concerning for high-stakes academic integrity applications, as an institution processing 10,000 submissions annually would generate approximately 570 false accusations against students who wrote their work authentically. This reality necessitates human review of flagged submissions rather than automated enforcement actions, substantially limiting efficiency gains from automated detection while requiring significant investment in trained personnel who can thoughtfully evaluate complex cases [19].

Error analysis examining false positive cases where authentic student submissions were incorrectly classified as AI-generated revealed concerning patterns with equity implications. Non-native English speakers were disproportionately represented among false positives, comprising 31 percent of false positive cases despite constituting only 18 percent of the overall student population in the dataset. This overrepresentation suggests that linguistic patterns characteristic of second language writing, including particular grammatical constructions influenced by first language transfer, unusual vocabulary choices, or discourse organization strategies differing from native speaker norms, may resemble features the model associates with AI generation. Similarly, students with documented learning differences showed elevated false positive rates of 12.3 percent compared to 5.7 percent overall [20].

- *Robustness Against Evasion Strategies*

A critical dimension of detection system evaluation involves assessing robustness against deliberate evasion strategies that motivated students may employ to circumvent identification. We systematically evaluated system performance under three common evasion scenarios: paraphrasing where students reword AI-generated content, prompt engineering where AI systems are explicitly instructed to generate human-like text, and hybrid authorship where students combine AI-generated and authentic sections. These scenarios represent realistic challenges that detection systems face in educational deployments where some users actively attempt to evade detection.

Paraphrasing proved to be an effective evasion strategy that substantially degraded detection accuracy across all model components. When AI-generated texts underwent moderate paraphrasing where students manually rewrote

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

approximately 30-40 percent of sentences while preserving overall content and structure, the CNN component's accuracy declined to 78.3 percent from its original 88.9 percent. The convolutional filters that effectively detected characteristic n-gram patterns in original AI text struggled when these patterns were disrupted through synonym substitution and sentence restructuring. The recurrent network component showed greater resilience with accuracy declining to 81.6 percent, as global discourse patterns and sequential structures partially survived paraphrasing that primarily altered local word choices. The complete multi-modal ensemble achieved 76.8 percent accuracy on paraphrased texts, demonstrating substantial vulnerability to this common evasion technique [21].

Prompt engineering strategies designed specifically to evade detection proved even more challenging for our system. When AI-generated texts were produced using sophisticated prompts instructing the model to "write like a college student with occasional minor grammatical errors and natural vocabulary variation," detection accuracy declined markedly across all components. The CNN pathway accuracy fell to 73.2 percent as local n-gram patterns shifted toward more human-like distributions when the AI system was explicitly optimized for this objective. The recurrent network pathway achieved 75.8 percent accuracy, as sequential patterns remained somewhat more difficult to manipulate through prompting alone. The multi-modal ensemble reached 78.7 percent accuracy under these adversarial conditions, substantially below the 94.3 percent achieved on standard AI-generated texts [22].

Hybrid authorship scenarios presented the greatest detection challenge, as texts combining authentic student writing with AI-generated sections created complex mixtures of human and machine patterns. For texts where 40-60 percent of content was AI-generated and remainder was human-authored, the multi-modal system achieved only 68.3 percent accuracy. The CNN component struggled to classify these hybrid texts confidently, as local patterns mixed human and AI characteristics in ways that fell between the distributions seen in pure examples of each category. The recurrent network showed similar confusion, as sequential patterns across sections could exhibit both natural human variation and mechanical AI consistency depending on which sections were machine-generated. These results highlight fundamental limitations of binary classification approaches for texts that genuinely represent collaborative human-AI authorship rather than purely one category or the other [23].

Comparative evaluation against commercial detection systems provided context for understanding our model's relative strengths and weaknesses. GPTZero achieved 87.2 percent accuracy on standard test data but declined to 71.4 percent under prompt engineering conditions, showing similar vulnerability to evasion strategies. Turnitin's AI Writing Detection achieved 89.6 percent accuracy on standard texts but only 73.8 percent on adversarially generated examples, with a concerning 8.3 percent false positive rate on authentic student writing. Our multi-modal ensemble's superior performance on standard texts (94.3 percent) and greater resilience under adversarial conditions (78.7 percent) validated the value of integrating multiple analytical perspectives, though all systems showed substantial degradation when confronted with deliberate evasion attempts [24].

➢ *Practical Implications and Limitations*

The deployment of AI text detection technologies in real-world educational settings faces numerous practical challenges beyond purely technical performance considerations. The false positive problem represents the most serious concern for institutional adoption, as incorrect accusations of AI use carry significant consequences for students including damaged academic records, potential honor code violations, and psychological distress from unjust accusations. With false positive rates of 5-10 percent across current detection systems including our multi-modal approach, institutions processing thousands of submissions annually will inevitably generate hundreds of false accusations even with relatively high-performing detectors. This reality necessitates mandatory human review of all flagged submissions rather than automated enforcement actions, substantially limiting the efficiency gains that automated detection might otherwise provide [25].

The issue of explainability further complicates practical deployment. Our multi-modal CNN-recurrent ensemble operates as a "black box" that provides classification predictions without transparent reasoning that educators and students can understand and evaluate. When the system flags submissions as AI-generated, stakeholders legitimately demand understanding of what specific textual characteristics triggered the classification. While we implemented gradient-based attribution methods to identify influential features for research purposes, translating these technical explanations into actionable feedback for non-technical users remains challenging. Current detection systems generally provide only confidence scores (e.g., "87% probability of AI generation") without detailed explanations, making it difficult for users to evaluate reliability of specific predictions or identify potential model biases and errors [26].

Cross-domain evaluation revealed that detection systems trained primarily on academic writing from particular disciplines or educational levels exhibit reduced accuracy when applied to different contexts. Our model trained on general undergraduate writing achieved only 81.3 percent accuracy when tested on graduate-level submissions, reflecting stylistic and complexity differences across educational levels. Testing on secondary school writing yielded 84.7 percent accuracy, as the model struggled with writing exhibiting different error patterns and stylistic characteristics than the undergraduate training data. Domain adaptation through fine-tuning on target-context examples improved performance but required substantial labeled data from each new domain (at least 500-1000 examples for meaningful improvement), limiting practical scalability for institutions seeking to deploy detection across diverse course levels and disciplines [27].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

The equity implications of AI detection systems demand serious consideration beyond the false positive concerns already discussed. Access to sophisticated AI tools and knowledge of evasion strategies may vary systematically with socioeconomic status, potentially creating scenarios where privileged students can more effectively evade detection while disadvantaged students face greater scrutiny. Students from well-resourced backgrounds may have access to premium AI tools, knowledge of prompt engineering techniques, or editing assistance that helps them successfully paraphrase AI-generated content. Meanwhile, students from less privileged backgrounds writing their work authentically but with linguistic patterns influenced by second language acquisition or educational disadvantage face elevated false positive rates. These dynamics could perversely result in detection systems disproportionately penalizing innocent students from marginalized populations while failing to catch AI misuse by more privileged students [28].

Furthermore, if educational institutions respond to AI availability primarily through increased surveillance and detection rather than pedagogical innovation, they risk creating learning environments characterized by suspicion and distrust rather than support and development. The psychological impact of operating under constant surveillance where any submission might be flagged as AI-generated could harm student wellbeing, increase anxiety, and damage educational relationships between students and instructors. The assumption of innocence that traditionally undergirds educational practice could erode into presumption of guilt, fundamentally altering the nature of teaching and learning relationships in ways that extend far beyond the specific issue of AI text generation [29].

The phenomenon of hybrid authorship and legitimate AI use presents particularly vexing challenges for both detection systems and institutional policy. Contemporary students frequently use AI tools throughout their writing processes in ways that may represent appropriate learning activities: brainstorming ideas, generating outlines, checking grammar, seeking feedback on argumentation, or working through writer's block. Many of these uses enhance rather than undermine learning, yet they produce texts that share characteristics with fully AI-generated content, potentially triggering detection systems. Educational institutions must develop nuanced policies distinguishing between appropriate and inappropriate AI use, recognizing that blanket prohibitions may prove neither enforceable nor pedagogically sound in an era where AI tools have become ubiquitous professional resources [30].

V. CONCLUSION

This research has demonstrated that multi-modal classification approaches integrating diverse analytical perspectives can achieve reasonably high accuracy in detecting AI-generated text in academic submissions, substantially outperforming single-modality methods. The proposed framework combining CNN-based local pattern extraction through convolutional filters operating on word embeddings, recurrent network sequential analysis maintaining hidden states across document positions, and learned fusion mechanisms integrating complementary modalities achieved 94.3 percent detection accuracy on carefully controlled test datasets. The multi-modal approach's superior performance validates the principle that comprehensive analysis across multiple textual dimensions provides more robust detection than reliance on any single analytical perspective, as sophisticated AI systems may successfully mimic human writing along certain dimensions while exhibiting anomalies in others.

The architectural components illustrated in Figures 1-3 each contributed distinct capabilities to the overall detection system. The CNN architecture captured local n-gram patterns and phrase-level features indicative of AI generation, leveraging multiple convolutional filters at different scales to identify characteristic word combinations and transition patterns. The recurrent network connections enabled modeling of sequential dependencies and global discourse patterns, detecting anomalies in how writing style evolves throughout documents and how topics develop across sections. The multi-modal fusion architecture integrated these complementary analytical perspectives through learned attention mechanisms that dynamically weighted different modalities based on text characteristics, achieving superior performance compared to simple combination strategies.

However, the research has also revealed significant limitations that temper optimism about purely technical solutions to academic integrity concerns. Detection accuracy degrades substantially under realistic conditions including paraphrasing (accuracy declining to 76.8 percent), prompt engineering designed to evade detection (78.7 percent), and hybrid authorship scenarios (68.3 percent). False positive rates of 5.7 percent remain concerning for high-stakes academic decisions, with disproportionate impacts on non-native English speakers (31 percent of false positives) and students with learning differences (12.3 percent false positive rate). The fundamental challenge that AI-generated text exists on a continuum with human writing rather than forming discrete categories, combined with the adaptability of AI systems to match desired stylistic characteristics when appropriately prompted, suggests inherent limits to purely technical detection approaches.

These findings underscore the importance of viewing AI detection technologies as components of comprehensive academic integrity strategies rather than standalone solutions. Educational institutions must carefully balance deployment of detection tools with other approaches including pedagogical innovation emphasizing process over product, assignment redesign leveraging tasks difficult for AI to complete authentically, development of student AI literacy and ethical reasoning capacities, and creation of supportive learning environments where students feel empowered to engage honestly with course material. Clear institutional policies distinguishing appropriate from inappropriate AI use, combined with transparent communication about detection capabilities and limitations, represent essential elements of effective responses to the challenges posed by generative AI in educational contexts.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

Future research should address several critical areas including development of more interpretable detection models that can explain their classification decisions through accessible visualizations of influential features, investigation of cross-lingual and multilingual detection approaches given global educational contexts, exploration of detection methods robust to evolving evasion strategies through adversarial training techniques, and systematic study of equity implications across diverse student populations through large-scale field deployments with careful monitoring of differential impacts. Additionally, longitudinal research examining how student behaviors and AI capabilities evolve over time will prove essential for understanding whether current detection approaches remain viable or whether fundamental shifts in educational assessment and integrity frameworks become necessary. Ultimately, sustainable responses to AI in education will likely emerge not from technical detection alone but from thoughtful integration of technology with pedagogical wisdom, ethical guidance, and commitment to student development as learners and responsible technology users.

REFERENCES

[1]. Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in education and teaching international, 61(2), 228-239.

[2]. Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access.

[3]. Ma, Z., Chen, X., Sun, T., Wang, X., Wu, Y. C., & Zhou, M. (2024). Blockchain-based zero-trust supply chain security integrated with deep reinforcement learning for inventory optimization. Future Internet, 16(5), 163.

[4]. Dergaa I, Chamari K, Zmijewski P, Saad HB. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biology of Sport. 2023;40(2):615-622.

[5]. Uzun L. ChatGPT and academic integrity concerns: Detecting artificial intelligence generated content. Language Education and Technology. 2023;3(1):45-54.

[6]. Ardito, C. G. (2025). Generative AI detection in higher education assessments. New Directions for Teaching and Learning, 2025(182), 11-28.

[7]. Dwivedi YK, Kshetri N, Hughes L, et al. So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management. 2023;71:102642.

[8]. Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. Contemporary educational technology, 15(2).

[9]. Pegoraro A, Kumari K, Fereidooni H, Sadeghi AR. To ChatGPT, or not to ChatGPT: That is the question. arXiv preprint. 2023;arXiv:2304.01487.

[10]. Cao, W., Mai, N. T., & Liu, W. (2025). Adaptive knowledge assessment via symmetric hierarchical Bayesian neural networks with graph symmetry-aware concept dependencies. Symmetry, 17(8), 1332.

[11]. Wu Y, Zhang X, Ren H. Improving text classification performance through multimodal representation. Pattern Recognition and Computer Vision. 2024;15037:312-325.

[12]. Cao, L. (2025). A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content. arXiv preprint arXiv:2504.02898.

[13]. Abimannan, S., El-Alfy, E. S. M., Chang, Y. S., Hussain, S., Shukla, S., & Satheesh, D. (2023). Ensemble multifeatured deep learning models and applications: A survey. IEEE Access, 11, 107194-107217.

[14]. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, et al. Testing of detection tools for AI-generated text. International Journal for Educational Integrity. 2023;19:26.

[15]. Ge, Y., Wang, Y., Liu, J., & Wang, J. (2025). GAN-Enhanced Implied Volatility Surface Reconstruction for Option Pricing Error Mitigation. IEEE Access.

[16]. Zheng, W., & Liu, W. (2025). Symmetry-Aware Transformers for Asymmetric Causal Discovery in Financial Time Series. Symmetry, 17(10), 1591.

[17]. Tan, Y., Wu, B., Cao, J., & Jiang, B. (2025). LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access.

[18]. Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. Sensors, 24(24), 7949.

[19]. Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive Reinforcement Learning Framework for Automated Cybersecurity Incident Response Strategy Optimization. Applied Sciences, 15(2), 951.

[20]. Zhang, Q., Chen, S., & Liu, W. (2025). Balanced Knowledge Transfer in MTTL-ClinicalBERT: A Symmetrical Multi-Task Learning Framework for Clinical Text Classification. Symmetry, 17(6), 823.

[21]. Mai, N. T., Cao, W., & Liu, W. (2025). Interpretable knowledge tracing via transformer-Bayesian hybrid networks: Learning temporal dependencies and causal structures in educational data. Applied Sciences, 15(17), 9605.

[22]. Chen, S., Liu, Y., Zhang, Q., Shao, Z., & Wang, Z. (2025). Multi-Distance Spatial-Temporal Graph Neural Network for Anomaly Detection in Blockchain Transactions. Advanced Intelligent Systems, 2400898.

[23]. Mai, N. T., Cao, W., & Wang, Y. (2025). The global belonging support framework: Enhancing equity and access for international graduate students. Journal of International Students, 15(9), 141-160.

[24]. Naini, I., & Ulya, R. H. (2025). Reasoning Patterns and Sentence Construction Errors in Students' Scholarly Articles: A Content Analysis of Academic Writing in Padang City. AL-ISHLAH: Jurnal Pendidikan, 17(2).

[25]. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1328

language understanding. Proceedings of NAACL-HLT. 2019;4171-4186.
[26]. Wang, Y., Ding, G., Zeng, Z., & Yang, S. (2025). Causal-Aware Multimodal Transformer for Supply Chain Demand Forecasting: Integrating Text, Time Series, and Satellite Imagery. IEEE Access.
[27]. Long, S., He, X., & Yao, C. (2021). Scene text detection and recognition: The deep learning era. International Journal of Computer Vision, 129(1), 161-184.
[28]. Qiu, L. (2025). Reinforcement Learning Approaches for Intelligent Control of Smart Building Energy Systems with Real-Time Adaptation to Occupant Behavior and Weather Conditions. Journal of Computing and Electronic Information Management, 18(2), 32-37.
[29]. Zhang, H. (2025). Physics-Informed Neural Networks for High-Fidelity Electromagnetic Field Approximation in VLSI and RF EDA Applications. Journal of Computing and Electronic Information Management, 18(2), 38-46.
[30]. Qiu, L. (2025). Multi-Agent Reinforcement Learning for Coordinated Smart Grid and Building Energy Management Across Urban Communities. Computer Life, 13(3), 8-15.
[31]. Li, J., Fan, L., Wang, X., Sun, T., & Zhou, M. (2024). Product demand prediction with spatial graph neural networks. Applied Sciences, 14(16), 6989.
[32]. Qiu, L. (2025). Machine Learning Approaches to Minimize Carbon Emissions through Optimized Road Traffic Flow and Routing. Frontiers in Environmental Science and Sustainability, 2(1), 30-41.