IJWR

From Cover to Story: AI-Driven Genre Classification and Illustrated Narrative Creation for Children's Literature

Maedeh Mosharraf *, Reyhane NaseriMoghadam
Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran; m_mosharraf@sbu.ac.ir, r.naserimoghadam@alumni.sbu.ac.ir

ABSTRACT

Storytelling is a fundamental pillar of childhood development, where visual narratives play a crucial role in enhancing engagement and cognitive processing. While Generative Artificial Intelligence (GAI) has revolutionized content creation, its application for automated story generation from book covers remains largely unexplored. This study presents an innovative pipeline that combines computer vision for genre classification with GAI to create tailored illustrated stories. After evaluating four deep learning architectures widely used in image classification tasks, ConvNeXt-Tiny was selected as the final model, achieving a Weighted F1-score of 0.6898 in categorizing children's books into 13 distinct genres through cover image analysis. To address the lack of benchmark datasets, we compiled and rigorously validated a specialized collection of 4,085 Persian children's book covers. The proposed system leverages both cover design elements and predicted genre features within structured prompts to generate coherent illustrated stories through LLMs and image-synthesis models. A sample of 26 generated stories was qualitatively evaluated by three child psychologists based on narrative coherence, genre alignment, age appropriateness, character continuity, and visual congruence. This research makes significant contributions to both Persian literary analysis and AI-driven creative systems, demonstrating how machine learning can enhance educational storytelling while preserving cultural authenticity.

Keywords*— *Genre Classification, Narrative Creation, Deep Learning, Large Language Model (LLM), Generative Artificial Intelligence (GAI); Children Literature.

1. Introduction

Narratives constitute an integral component of human life, serving a pivotal function in the cognitive, emotional, and social development of individuals. They assist in comprehending complex ideas, enhancing empathy, and facilitating future-oriented thinking [1]. In the realm of childhood, the significance of storytelling is magnified, as it acts as a potent medium for stimulating imagination and curiosity. The visual elements that accompany stories play a vital role in character formation, the advancement of cognitive abilities, and the development of effective social interactions [2-3]. When a storybook is read to a child, the child often internalizes the narrative, revisits it frequently, and reimagines themselves as part of its world [4]. Illustrated books, in particular, have a substantially greater influence on a child's understanding of narrative events. According to [5], children dedicate approximately 90% of their storybook reading time to observing illustrations. Furthermore, research by [6] indicates that large, colorful images can enhance textual comprehension by up to 40% among children under the age of five.

Generative Artificial Intelligence (GAI) has opened new horizons for the creation of imaginative, content-rich, and visually engaging stories tailored to various literary genres. This technology has the potential to enrich the reading experience and help children form deeper emotional and cognitive connections with story content. A critical stage in this process is ensuring the alignment of the generated content with the intended genre [7]. The book cover, as the first point of interaction between the reader and the book, conveys valuable visual cues about the narrative within. By analyzing cover design features, AI systems can identify the visual patterns associated

http://dx.doi.org/10.22133/ijwr.2025.553068.1313
Citation M. Mosharraf, R. NaseriMoghadam, "From Cover to Story: AI-Driven Genre Classification and Illustrated Narrative Creation for Children's Literature", *International Journal of Web Research*, vol.9, no.1,pp.57-73, 2026, doi: http://dx.doi.org/ 10.22133/ijwr.2025.553068.1313.

*Coressponding Author
Article History: Received: 13 October 2025; Revised: 27 December 2025; Accepted: 30 December 2025.

with different genres and leverage this understanding to generate illustrations that align with the story's tone and stylistic framework. The integration of genre classification with generative content creation thus enables the development of cohesive visual narratives that remain faithful to their respective genres.

While various studies have addressed book genre classification using Machine Learning (ML) techniques [8-12], research focused specifically on Persian children's and young adult literature remains scarce. Moreover, the creative potential of GAI to automatically produce children's stories and illustrations based on genre classification has yet to be thoroughly investigated.

This study proposes an intelligent system that not only identifies the genre of Persian children's books based on their cover designs but also generates genre-consistent narratives and illustrations. In doing so, it expands the possibilities of content creation beyond pre-existing storybooks. By using only the book cover as input, the system can produce multiple stories that are stylistically and thematically aligned with the same genre. To accomplish this, deep neural networks are employed to extract and analyze the visual features of children's book covers for genre recognition. Subsequently, leveraging GAI and Large Language Models (LLMs), the system creates cohesive stories and visuals that resonate with the identified genre—offering children a compelling and personalized reading experience.

Building on these gaps, the central claim of this study is that the visual information in children's book covers can be used not only to predict narrative genre but also to guide the creation of coherent, genre-consistent, and developmentally appropriate multimodal stories. This capability is enabled by modern Convolutional Neural Networks (CNN) architectures, which can learn distinctive visual patterns associated with literary genres. The resulting genre labels, when combined with salient visual elements from the cover, can then guide LLMs to generate narratives that maintain genre fidelity, narrative coherence, and developmental suitability.

This work offers three main contributions. First, it introduces a curated dataset of 4,085 Persian children's book covers, consolidated into 13 coherent genres, addressing the absence of publicly documented resources for visual analysis in children's literature—especially in the Persian-language context, where no comparable dataset has been reported. Second, it provides an empirical benchmark for visual genre classification, building on prior image-based literary analysis but extending it to the children's domain; to the best of our knowledge, no previous study has examined genre prediction from children's book covers, and none has done so in Persian. Third, the study presents an integrated and fully documented pipeline that links genre prediction with LLM-based story generation, scene segmentation, and image synthesis, offering a reproducible workflow for generating illustrated stories grounded in cover design features.

The remainder of this paper is organized as follows. In Section 2, related works are reviewed, highlighting key findings and effective methodologies from prior studies. Our proposed methodology, presented in Section 3, consists of two main parts: genre classification and AI-based content generation. Section 4 details the data collection process and the classification of the dataset using AI techniques. Section 5 focuses on generating illustrated stories based on the genre classification outputs and evaluating the results. Finally, Section 6 concludes the study and outlines its limitations along with potential directions for future research.

2. Literature Review

The classification of genres across diverse media formats—such as books, music albums, and films—has persistently constituted a critical area of research within the domains of data classification and processing. A substantial body of literature has sought to enhance the accuracy of genre prediction by leveraging visual, textual, and multimodal analytical methodologies, while concurrently addressing the prevailing challenges inherent to this field. These investigations have primarily emphasized the integration of Natural Language Processing (NLP) with computer vision techniques to reinforce the semantic coherence between visual representations and narrative content.

The first subsection herein provides a comprehensive review of seminal studies pertinent to this topic. Subsequently, the second subsection examines relevant research pertaining to the application of AI, with a particular focus on LLMs, in the generation of illustrated textual content.

2.1. Genre Classification Across Various Media

Several studies have addressed the classification of book genres based solely on cover images and titles, without relying on auxiliary data. The application of transfer learning and the integration of textual and visual embeddings have demonstrated that such approaches can achieve model accuracies competing with human-level performance [8]. Some research efforts have developed multimodal models that process both images and text simultaneously, employing Deep Canonical Correlation Analysis (DCCA) to examine the impact of this fusion. However, findings indicate that simpler models without DCCA often outperform their more complex counterparts, potentially due to genre overlap and misleading cover designs [11].

Other research has explored multimodal architectures employing sparse cross-entropy loss

functions to achieve improved alignment between visual and textual features. Nevertheless, persistent challenges—including ambiguous labeling and the presence of multi-genre data—continue to impede optimal model performance [12]. Alternatively, approaches integrating color distribution-based visual features with Support Vector Machine (SVM) classifiers via late fusion strategies have demonstrated favorable outcomes [10].

Additionally, studies evaluating the efficacy of lightweight CNN reveal that compact architectures frequently surpass more complex models in both accuracy and training efficiency. The incorporation of Sobel and Gabor filters during preprocessing has been shown to significantly enhance the extraction of spatial features, thereby contributing to improved model performance [9].

In the domain of music genre classification, investigations have demonstrated that multimodal models outperform ResNet-based architectures in analyzing cover images within the MuMu dataset. This finding underscores the substantial benefits of integrating visual, textual, and audio data to significantly enhance genre classification accuracy. Nonetheless, key challenges persist, including label imbalance and the necessity for multi-label classification, which have been addressed through matrix factorization and logistic regression techniques [13]. Further studies have highlighted the critical role of album cover design in music genre prediction, illustrating that deep learning methods for extracting visual features relevant to genre can improve classification performance. These insights have practical implications not only for music recommendation systems but also for music content organization platforms [14].

The examination of visual attributes in film posters, which serve as critical promotional materials, employs a wide spectrum of computational techniques ranging from CNNs to transformer-based models. Empirical evidence suggests that incorporating Gram layers within CNN architectures, along with features such as color palettes, object spatial arrangements, and compositional balance, significantly enhances the accuracy of multi-label genre classification. Furthermore, the fusion of stylistic and spatial information contributes to improved model generalizability in predicting film genres [15]. In efforts to forecast individual viewer preferences, certain studies have integrated graphical poster elements with XGBoost classifiers to assess attributes including color distribution, luminance, saturation, layout complexity, and design simplicity. The results reveal a meaningful relationship between graphic design characteristics and audience engagement metrics [16].

In the task of multi-genre identification from film posters, the Residual Dense Transformer (RDT) architecture has demonstrated superior performance compared to CNNs, effectively capturing global dependencies within visual data. [12] indicates that genre classification plays a pivotal role in film recommendation systems and advertising, with models such as ResNet, VGG-16, and DenseNet achieving high accuracy in this domain. Moreover, analyses of compositional elements in posters reveal that audiences subconsciously recognize genre-specific patterns, thereby underscoring the significant influence of visual design on genre prediction [17]. Additionally, compositional features—including color, lighting, and element orientation—have been shown to impact genre perception and contribute to the enhancement of genre classification models [18].

2.2. Generation of Illustrated Texts Using GAI

Advancements in GAI models have enabled the automatic creation of illustrated textual narratives. Numerous studies have investigated story generation leveraging LLMs in conjunction with multimodal inputs to synchronize narrative content with corresponding visual elements [19-20]. These investigations have demonstrated significant progress, while also addressing the challenges inherent in maintaining narrative coherence and quality. Furthermore, they have proposed methodologies aimed at enhancing the consistency and overall quality of AI-generated stories.

Building on these methodologies, recent works have introduced multimodal frameworks for narrative generation that integrate language models with vision-guided mechanisms to jointly process textual and visual inputs. These approaches utilize initial images as narrative anchors, enabling models to employ AI-driven guidance, including multimodal attention mechanisms and visual decoding strategies, to dynamically select relevant scenes and characters. This process facilitates the creation of coherent narratives that maintain strong semantic congruence between visual content and textual description. Addressing the complexities of fusing heterogeneous data modalities, these studies have explored sophisticated techniques such as deep correlation learning and interaction modeling between images and sentences to significantly improve the coherence and quality of generated stories [7,21].

Meanwhile, some studies have sought to optimize the narrative generation process by introducing advanced techniques such as Dynamic Beam Search, which enhances text generation through real-time evaluation of emotional impact and narrative coherence. This approach facilitates the integration of affective dimensions into storytelling in a more nuanced and adaptive manner, thereby enabling the generation of narratives with heightened emotional depth and engagement [22].

In parallel, other lines of research have adopted

goal-oriented frameworks to guide LLMs, leveraging reinforcement learning paradigms to align generated content with predefined narrative goals and structural patterns. Notably, the application of algorithms such as Proximal Policy Optimization has demonstrated efficacy in steering model outputs toward coherent and purpose-driven narratives [23]. These advancements not only increase the controllability and interpretability of generative models but also contribute to improved character-event coherence and narrative integrity within the generated texts.

In the context of leveraging pre-trained language models, studies such as [24-25] have demonstrated that combining transfer learning with task-specific fine-tuning for narrative generation significantly enhances the coherence and stylistic quality of the output. These models are capable of producing narratives that exhibit both structural consistency and a well-defined narrative voice. Nevertheless, ongoing challenges remain—particularly in maintaining character consistency, preserving narrative coherence across longer texts, and ensuring stylistic alignment throughout the story. These issues continue to be the focus of active research in the field.

While recent multimodal language models such as Flamingo [26], PaLM-E [27], MiniGPT-4 [28], and LLaVA [29] demonstrate that LLMs can integrate visual information to produce coherent text, these models are designed as broad, general-purpose assistants trained on large-scale heterogeneous datasets. In contrast, our work focuses on a domain-specific and low-resource scenario: using children's book cover visuals as explicit, interpretable control signals for guiding story generation in Persian. Unlike prior multimodal studies that emphasize architectural advances, our contribution lies in applying multimodal conditioning to a culturally grounded task, providing a reproducible pipeline, documenting prompt templates, and conducting expert-based evaluation within the context of children's literature.

3. Methodolgy

To provide a clear overview of the research, Figure 1 presents the complete end-to-end workflow of the proposed system. The diagram summarizes all major components of the study—dataset construction, visual genre classification, narrative generation, scene decomposition, and illustration synthesis—highlighting the order of operations and the information exchanged between modules. This high-level representation clarifies how the two main phases of the research are integrated and how outputs from the classification stage inform the subsequent story-generation process.

As illustrated in Figure 1, the methodology comprises two tightly interconnected components.

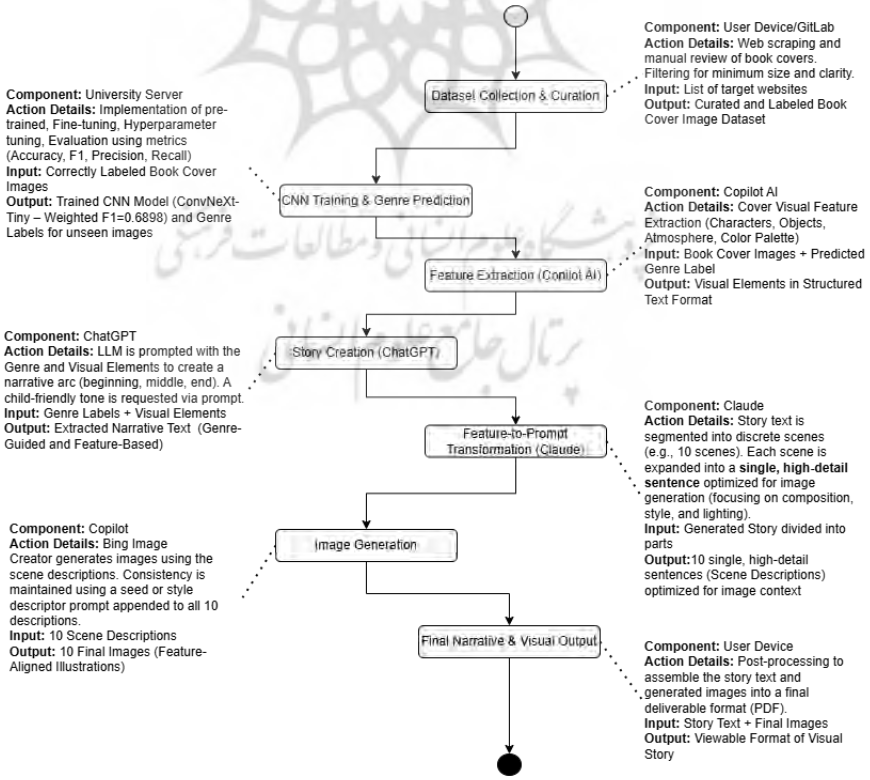


Figure. 1. The complete research workflow, integrating visual genre classification with LLM-based story generation and image synthesis

3.1. Visual Genre Classification

The pipeline begins with dataset compilation, image preprocessing, and training of multiple CNN architectures. The ConvNeXt-Tiny model, selected based on its superior Weighted F1-score, produces both the predicted genre label and a structured set of descriptive visual cues extracted from the book cover (e.g., dominant colors, characters, background objects, emotional tone).

These features constitute the core semantic signals used to bridge the classification and generative stages.

3.2. Story and Illustration Generation

The predicted genre and extracted visual elements then form the structured input prompts for the LLM responsible for story generation. This ensures that the narrative reflects both the thematic expectations of the genre and the specific stylistic cues from the cover image. The generated story is subsequently decomposed into scene-level descriptions, which guide the image-synthesis model in producing coherent illustrations aligned with each narrative segment.

This explanation demonstrates that the pipeline is not composed of two independent systems but rather an integrated multimodal workflow in which visual information captured by the classification model directly influences narrative style, content, and visual aesthetics, thereby addressing reviewer concerns regarding methodological coherence.

4. Classification of Children's Book Covers by Genre

The primary objective of this study is to develop an AI–based system capable of generating illustrated stories for children. For this purpose, CNNs are employed to perform visual analysis of existing book covers and extract salient features—most notably, the genre of the book. Providing the identified genre as input to a text-generation model allows the system to adapt narrative tone, thematic elements, and stylistic conventions according to genre-specific requirements.

Accordingly, the AI model is expected to synthesize both visual features and linguistic patterns to generate coherent and imaginative narratives. The storytelling process is structured into multiple segments, each accompanied by a detailed scene description—including character traits, setting, and overall atmosphere. These descriptions not only guide the narrative progression but also serve as inputs to a generative image model tasked with producing corresponding illustrations. As a result, the final output comprises stories that are not only visually aligned with the original book cover but also genre-consistent in their narrative content. Although the classification model operates purely on visual features and is therefore language-independent, the story generation component—Persian-language narrative generation as the primary focus of this study- depends on the linguistic capabilities of the selected LLM. As modern LLMs support multilingual inputs and outputs, the proposed methodology is language-independent and can be generalized to other languages with minimal modification.

4.1. Data Collection and Preprocessing

The absence of a standardized dataset for Persian-language children's book covers prompted the creation of a custom dataset, curated from reputable online platforms specializing in children's books. Sample visualizations are included in the paper for illustrative purposes under fair-use guidelines.[1]

These websites typically categorize their offerings based on age group, theme, or genre. However, the diversity of children's books and the fragmented nature of stakeholders in this domain have resulted in the lack of a unified classification model. Given that genre-based categorization tends to be less influenced by individual preferences compared to age or thematic classification, this study adopted genre as the primary axis for organizing and labeling the dataset. Initial genre assignments were guided by the hierarchical classification system presented in [30].

Due to imprecise labeling in online sources and the inherent difficulty of distinguishing certain sub-genres, ambiguous samples were removed to preserve the coherence and effectiveness of the training data. Ultimately, the selected genres were those that exhibited sufficiently distinctive visual characteristics, allowing for meaningful analysis and comparison. Cover images were sourced from 5 reputable online book retailers and archives[2], and the Image browser extension[3] was utilized to streamline the image download process.

In the image labeling phase, book cover images were annotated using both single-label and multi-label strategies. For single-label annotation, the hierarchical structure illustrated in Table 1 was followed. In the multi-label annotation process—designed to

[1] The book-cover images used in this study are copyright-protected and were accessed solely for non-commercial academic analysis. No images are redistributed or publicly released. Only derived metadata and model outputs are shared to preserve reproducibility while complying with intellectual property requirements.

[2] https://badbaddak.ir/product-age/childbook, https://bazidone.com/product-category/book, https://amooketabi.ir/product-category/kids-book/, https://ketabak.org/, https://hodhod.com/

[3] https://www.imageye.net/

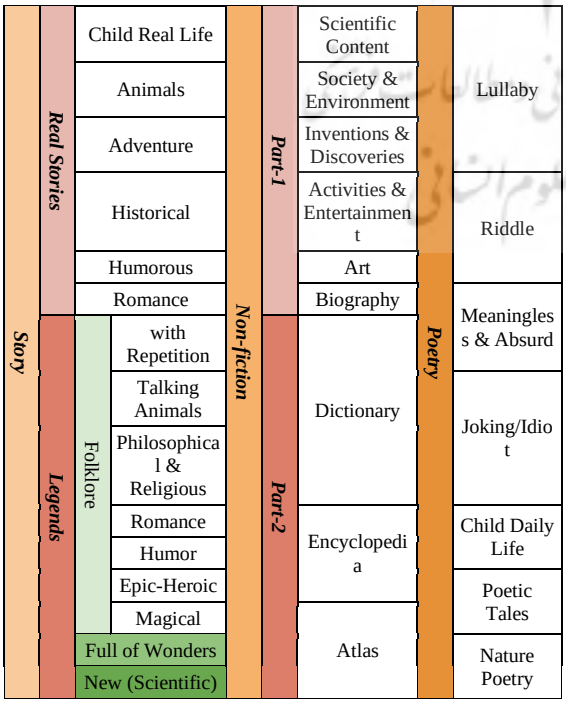
identify duplicate images and account for books belonging to multiple genres—the perceptual hash (pHash) algorithm [31] was employed to generate a unique hash for each book cover image. Given that the dataset was compiled based on genre rather than visual similarity, it was possible for identical or nearly identical images (e.g., those with minor variations in resolution, brightness, or file naming) to exist within the same genre category. This redundancy posed challenges to accurate genre assignment and threatened the overall consistency of the dataset. The pHash algorithm enabled the generation of a compact representation of each image's visual content, facilitating the detection of near-duplicates even in the presence of subtle variations. By comparing the generated hashes, we were able to identify and remove visually redundant or duplicate samples. Moreover, this process allowed us to accurately extract and assign multiple genre labels to relevant images, ensuring comprehensive multi-label representation.

As a result, the precision of multi-genre labeling was significantly improved, and we ensured that the single-label subset remained visually clean, de-duplicated, and well- prepared for model training.

Ultimately, a dataset comprising 4,085 cover images of Persian-language children's books across 15 distinct genres was compiled. To address class imbalance—caused by data scarcity in certain genres (such as lullabies and riddles) and the exclusion of categories outside the target age group (such as inventions and discoveries)—a range of techniques was applied during preprocessing and model evaluation. These included resampling, data augmentation, and the use of class weighting in performance metrics. Tables 2 and 3 respectively display the data distribution across genres for single-label and multi-label classification scenarios.

The dataset was subsequently divided into training (80%), validation (10%), and test (10%) subsets. To ensure stable model performance in the presence of class imbalance, class distributions were maintained uniformly across all subsets.

Table 1. Hierarchical classification of Persian children's and young adult literature genres



4.2. Book Genre Classification

To classify children's book covers into distinct genres, we employed deep learning models based on CNNs. These models receive cover images as input and output the predicted genre as a single classification label. As a first step, various CNN architectures reported in prior research were reviewed and compared. Based on their demonstrated effectiveness and reported accuracy in related visual

Table 2. Genre classification and data distribution in single-label mode

Genre	Genre Subcategory	Images Num	Genre	Genre Subcategory	Images Num	Genre	Genre Subcategory	Images Num
Story	Child Real Life	237	Non-fiction	Scientific Content	480	Poetry	Lullaby	92
	Animals	270		Society & Environment	270			
	Adventure	533		Inventions & Discoveries	76		Riddle	38
	Historical	123		Activities & Entertainment	299			
	Humorous	382		Biography	329		Child Poetry	291
	Legends	359		Reference	306			

Table 3. Genre classification and data distribution in multi-label mode

Genre	Genre Subcategory	Images Num	Genre	Genre Subcategory	Images Num	Genre	Genre Subcategory	Images Num
Story	Child Real Life	315	Non-fiction	Scientific Content	574	Poetry	Lullaby	96
	Animals	314		Society & Environment	282			
	Adventure	559		Inventions & Discoveries	76		Riddle	38
	Historical	123		Activities & Entertainment	307			
	Humorous	464		Biography	333		Child Poetry	309
	Legends	374		Reference	318			

tasks, the ResNet architecture—a widely used and well-established model in computer vision—was selected as the baseline model.

Recognizing that label type and data structure significantly influence classification outcomes, three distinct pipelines were designed to explore alternative data organization and annotation strategies. Each pipeline represents a complete workflow encompassing data loading, preprocessing, model training, evaluation, and result storage.

Approach I: Multi-class (15 classes), Multi-Label Classification

The goal of this approach is to train a model capable of identifying all relevant genre labels—out of a possible 15—for each book cover image. To this end, one-hot encoding was employed for label representation. In this method, each genre is represented as a binary vector in which only the position corresponding to the relevant genre is set to 1, while all other positions are set to 0. This encoding scheme is particularly well-suited for multi-label classification tasks, as it allows a single image to be associated with multiple genre labels simultaneously.

a)Evaluation Metric Selection

One of the primary challenges in multi-label classification lies in the complexity of performance evaluation. Given that each image may be associated with multiple classes, traditional metrics such as accuracy—which require exact label matches across all classes—are not well-suited for this task [32]. Accordingly, as shown in Equation (1), Macro F1-Score was adopted to evaluate model performance. This metric provides a more balanced assessment across all genres and is particularly suitable for imbalanced datasets [33].

$$F1\ Score_i = \frac{2.Precision_{\ i}.Recall_i}{Precision_{\ i} + Recall_i}, Macro\ F1 = \frac{1}{C}\sum_{i=0}^{C} F1\ Score_i \quad (1)$$

$$C = Total\ number\ of\ classes$$

b) Evaluation of Implemented Models

Table 4 presents the various configurations implemented for multi-label classification using the ResNet architecture. Throughout the model design process, we experimented with multiple components including learning rate schedulers, samplers, diverse data augmentation techniques, optimizers, and loss functions. The objective behind testing these combinations was to explore a range of strategies in an unbiased manner and to identify a convergent set of techniques best aligned with the characteristics of the dataset. The selection of methods was guided by project goals, data properties, and the compatibility of each technique with the structure of the problem.

Table 4. Comparison of evaluation metrics for the implemented models in the 15-class multi-label classification approach

ResNet			
Multi-label - 15 Classes			
Macro F1-Score			
Test	***Validation***	***Train***	***Pipeline***
0.1566	0.154	0.248	4116
0.1255	0.1315	0.113	4114
0.4648	0.4395	0.2475	4070
0.409	0.4444	0.4355	4061
0.5038	0.5074	0.3143	4053
0.5129	0.5667	0.453	4046
0.5073	0.5365	0.3752	4044
0.5202	0.5029	0.7124	4037
0.4837	0.5023	0.7387	4035

In this context, a deliberate balance between performance and generalizability was pursued, ensuring that the selected combinations demonstrated the highest degree of statistical and empirical consistency with the dataset and the intended task. Model evaluation was carried out using Macro F1-Score, reflecting the dual emphasis on overall accuracy and class-wise balance.

The selected pipeline, as evaluated by the Macro F1-Score (4037), employs the streamlined ResNet-18 architecture, thereby circumventing the added computational complexity associated with larger models. This configuration integrates regularization techniques, including Dropout, alongside a suite of data augmentation strategies such as 30-degree rotation, color jittering, and random horizontal flipping. Collectively, these methods facilitate an optimal balance between model capacity and generalization performance.

This pipeline addresses class imbalance using a "WeightedRandomSampler," a crucial strategy for multi-label datasets with skewed distributions, which significantly improves learning from underrepresented classes. Furthermore, the use of a higher initial learning rate coupled with the "ReduceLROnPlateau" scheduler allows dynamic rate adjustment during training, a key factor in achieving the best F1-score.

Conversely, alternative models assessed within this framework—despite employing more sophisticated architectures and advanced techniques such as Focal Loss, Mixup augmentation, and Adaptive Thresholding—either succumbed to overfitting or failed to yield substantive improvements in model performance.

Collectively, the results underscore that a judicious balance of architectural simplicity, comprehensive data augmentation, and strategically

optimized training protocols is paramount to achieving stable and generalizable model performance. This finding highlights the importance of harmonizing key design elements to align with the specific characteristics of the dataset and task.

Approach II: Multi-class (15 classes), Single-Label Classification

The analysis of the test dataset revealed that the proportion of multi-label images was relatively low, comprising approximately 9% of the total dataset. This imbalance significantly limited the model's ability to effectively learn from multi-label data, resulting in a maximum accuracy of only 52% in multi-label classification. Consequently, a single-label classification (one genre per image) was prioritized for reliable outcomes.

a)Selection of Evaluation Metrics

One of the primary challenges in multi-label classification lies in the complexity of performance evaluation. Given that each image may be associated with multiple classes, traditional metrics such as accuracy—which require exact label matches across all classes—are not well-suited for this task [32]. Accordingly, as shown in Equation (2), Macro F1-Score was adopted to evaluate model performance. This metric provides a more balanced assessment across all genres and is particularly suitable for imbalanced datasets [33].

$$Accuracy = \frac{\sum_{i=1}^{N} 1(y_i = \hat{y}_i)}{N}$$

$$y_i = true\ label\ for\ sample\ i,$$
$$\hat{y}_i = the\ predicted\ label,$$
$$N = total\ number\ of\ samples$$

$$F1_C = \frac{2.Precision_C.Recall_C}{Precision_C + Recall_C},$$

$$Weighted\ F1 = \sum_{c=1}^{C} \left(\frac{n_c}{N} \times F1_c\right) C \quad (2)$$

$$= Number\ of\ classes,$$

$$n_c = Number\ of\ rue\ instances\ in\ class\ C$$
$$N = The\ total\ number\ of\ instances\ across\ all\ classes,$$
$$F1_c = The\ F1\ Score\ for\ class\ c$$

b)Evaluation of Implemented Models

Table 5 presents the performance of the ResNet architecture on the single-label classification task with 15 classes. The results are reported using two evaluation metrics, Weighted F1-Score and Accuracy, across multiple pipelines and segmented by training, validation, and test datasets. This comparative analysis facilitates a more precise examination of model generalizability and the impact of various technical configurations on overall performance.

As observed in the overall trend of pipeline results for this approach—evaluated by the Weighted F1-Score metric (from pipeline 4200 to 4245)—and considering the shift in the target objective, the tuning of component configurations to achieve satisfactory model performance has been time-consuming and required extensive experimentation.

Table 5. Comparison of evaluation metrics for implemented models in the 15-class single-label classification approach

ResNet			
Single-label - 15 Classes			
Weighted F1-Score			
Test	***Validation***	***Train***	***Pipeline***
0.5609	0.5741	0.8347	4245
0.3668	0.3753	0.2375	4238
0.1669	0.1985	0.0956	4237
0.4337	0.4418	0.3731	4228
0.3165	0.3529	0.2814	4224
0.3811	0.4233	0.6521	4200
Accuracy			
Test	Validation	Train	Pipeline
0.5612	0.6097	0.9519	4195
0.4745	0.477	0.6562	4193
0.5434	0.6046	0.8896	4192
0.5842	0.6505	0.9969	4190

Ultimately, pipeline 4245, by striking a balance between model complexity, optimization settings, and data augmentation, has delivered the best performance on the test set based on the Weighted F1-Score criterion. The integration of Dropout and Batch Normalization within the model architecture, combined with a relatively higher learning rate and the use of Cosine Annealing as the learning rate scheduling strategy, has enabled effective training without overfitting.

In other configurations, despite employing more complex architectures and advanced techniques such as Mixup and Adaptive Loss Weighting, the lack of precise tuning and component synchronization prevented satisfactory performance on the test data. In these models, uncontrolled complexities impeded better generalization. Consequently, the results from this phase further underscore the critical importance of balanced model design and meticulous component selection in optimizing final performance.

Regarding efforts evaluated using the Accuracy metric, an examination of pipeline performances alongside their respective components reveals that, under appropriate training conditions—particularly when overfitting is a concern—architectures with lower complexity can outperform deeper models with heavier regularization. While other pipelines incorporated class weighting, perspective augmentations, or more aggressive data augmentations, the restrained augmentation strategy combined with a balanced and straightforward application of "CrossEntropyLoss" enabled the

architecture in pipeline 4190 to learn stable representations without introducing excessive variability or suffering from increased noise or convergence degradation. Moreover, a fixed training schedule utilizing "StepLR" with a relatively high learning rate supported fast yet stable learning. Overall, pipeline 4190 exemplifies the efficacy of architectural simplicity paired with carefully calibrated training procedures and controlled data augmentation in delivering superior performance.

Ultimately, it can be stated that the top-performing models in this approach have succeeded in delivering optimal test set performance by maintaining a proper balance in model complexity, optimization strategies, and data augmentation. They have also leveraged effective architectural simplicity combined with precise training configurations and targeted data augmentation, thereby mitigating issues of overfitting or underfitting observed in other models.

c) Comparison of Selected Models

Considering the performance differences observed between pipelines 4245 and 4190 in Table 5, we proceed to analyze the implementation codes for each pipeline, as detailed in Table 6.

Table 6. Comparison of key implementation components in two selected pipelines for the 15-class single-label classification approach

15 Classes		
Single-Label		
Difference Variable	Pipeline: 4190	Pipeline: 4245
Metric	Accuracy	Weighted F1-Score
Test Score	0.5842	0.5609
Batch Size	64	16
Model	ResNet-18 (Small or simple dataset)	ResNet-50 (Large or complex dataset, Overfitting concerns)
Train Transforms	-	Additional: RandomCrop
Loss Function	Cross Entropy loss	Focal loss
Sampler	Not used (Only shuffle in train loader)	Weighted random sampler (Used for train loader)
Fully Connected Layer	Linear, layer	Multi-layer head
Regularization	Minimal regularization (no dropout, batch normalization in the head)	Better regularization via dropout and batch normalization
Scheduler	StepLR	Cosine Annealing LR

The implementation of pipeline 4190 employs a comparatively simpler architecture, specifically ResNet-18, with core evaluation metrics including Accuracy and the CrossEntropy loss function guiding model assessment. This pipeline utilizes a fixed learning rate governed by a step-wise scheduler (StepLR) and applies random sampling of the dataset without explicitly addressing class imbalance.

Conversely, pipeline 4245 is implemented with a more sophisticated architecture featuring modifications in the terminal layers, such as Dropout and Batch Normalization, to enhance regularization. This pipeline places greater emphasis on mitigating class imbalance through the application of weighted sampling techniques and the incorporation of the Focal Loss function. Furthermore, it adopts an advanced learning rate scheduling approach based on Cosine Annealing, facilitating more dynamic optimization.

Therefore, assuming a prior consideration of these differences before execution, it is anticipated that pipeline 4245 would yield superior results when faced with imbalanced or challenging datasets. However, upon executing both pipelines, it was observed that, due to the deeper model architecture, smaller batch size, and computational resource constraints, the model implemented in pipeline 4245 marginally underperformed relative to pipeline 4190.

Approach III: Multiclass (13 classes), Single-Label Classification

Although the single-label classification approach led to improved model accuracy, a detailed examination of the assigned labels and repeated error analyses from the previous two approaches revealed that the highest error rates—and consequently the lowest Weighted F1-Scores—were concentrated within specific classes. To reduce model complexity and enhance generalizability, two classes, Adventure_Stories and Animal_Stories, which consistently exhibited the greatest ambiguity and classification errors, were excluded from the classification task. This exclusion aimed to maximize the reduction of ambiguity and improve the overall accuracy of the model.

Analysis of the Weighted F1-Score metric indicated that the primary challenge in distinguishing these classes stemmed from limited sample sizes and overlapping visual features across categories. This reflects insufficient distinction in cover designs to effectively differentiate genres, thereby negatively impacting model performance. On the other hand, removing more than two classes risked diminishing genre coverage and reducing the model's generalizability. Therefore, the removal of these two classes was considered a balanced trade-off between reducing complexity and preserving genre diversity. Subsequently, the focus shifted towards enhancing

the model's performance using the Weighted F1-Score metric, with the objective of improving overall accuracy.

a) Evaluation of Fixed Configurations Across Different Neural Networks

In light of the structural similarity in label annotation between this approach and the second one, the previously optimized configurations were systematically re-evaluated within the scope of the third approach. This investigation aimed not only to re-examine the impact of key implementation components on model performance, but also to enhance their effectiveness and pursue performance gains surpassing those of the baseline CNN architecture (ResNet). The outcomes derived from various combinations of implementation components are summarized in Table 7, facilitating a more granular analysis and comparative assessment of pipeline-specific performance dynamics.

A comparative analysis of the results presented in the current table against those in Table 6 demonstrates consistency in the fundamental components, accompanied by an overall enhancement in model performance.

Pairwise evaluation of successive pipelines indicates that pipelines 4248 and 4265 exhibit comparable performance metrics, situating them in a lower tier relative to pipelines 4249 and 4263. Specifically, the models corresponding to pipelines 4248 and 4265 achieved moderate accuracy on the training dataset and acceptable results on the validation set. However, their diminished performance on the independent test dataset reflects suboptimal generalization capability.

In contrast, pipeline 4263 demonstrates the highest accuracy among all evaluated models—surpassing even pipeline 4249 within the same category—and consistently outperforms others across training, validation, and test phases. The combination of high training accuracy alongside strong validation and test results indicates that this model not only excels in learning from the training data but also generalizes more effectively to unseen data compared to the other pipelines. These attributes have established this model as the preferred choice for deployment.

To facilitate a fair and rigorous comparison among various neural network models, the optimized hyperparameters and training configurations developed for the top-performing ResNet-based model were applied unchanged to other widely used architectures in the computer vision domain, including ConvNeXt, DenseNet, and EfficientNet. This approach allowed isolation of the pure model impact by minimizing confounding effects introduced by differing training setups. As detailed in Table 8, different complexity variants of each model were selected to enable a comprehensive evaluation.

This selection strategy not only permitted an in-depth comparison of structural differences among CNN models but also facilitated assessment of the internal component complexities within each network.

Analyzing the employed models' performance enables the identification of the optimal one within each model family, based on a rigorous set of evaluation metrics:

- **EfficientNet CNNs:** An assessment of various complexity levels across EfficientNet architectures reveals that each model consistently attained high training accuracy

Table 7. Comparison of evaluation metrics for ResNet in the 13-class single-label classification approach

ResNet			
Single-label - 13 Classes			
Weighted F1-Score			
Test	Validation	Train	Pipeline
0.5478	0.5814	0.7818	4248
0.6239	0.6517	0.8692	4249
0.6517	0.6884	0.94	4263
0.5544	0.607	0.8733	4265

Table 8. Comparison of evaluation metrics for implemented models in the 13-class single-label classification approach

Single-label - 13 Classes											
Weighted F1-Score											
ConvNeXt				DenseNet				EfficientNet			
Tiny Version				DenseNet-121				EfficientNet-B3			
Test	Validation	Train	Pipeline	Test	Validation	Train	Pipeline	Test	Validation	Train	Pipeline
0.6898	0.7324	0.9973	4280	0.671	0.7133	0.9858	4272	0.6484	0.7242	0.9916	4267
Base Version				DenseNet-169				EfficientNet-B4			
				0.6694	0.7437	0.9885	4293	0.6537	0.7015	0.9831	4269
				DenseNet-201				EfficientNet-B7			
0.6572	0.736	0.9981	4291	0.6722	0.7027	0.9904	4294	0.675	0.7046	0.9678	4271

accompanied by strong validation and test results, indicating a well-calibrated balance between learning capacity and generalization ability. Given the marginal differences observed across evaluation metrics on different datasets, the test set performance was designated as the decisive criterion for model selection. In this context, EfficientNet-B3 (pipeline 4267) demonstrated superior performance relative to the more complex B5 and B7 variants.

- **DenseNet CNNs**: Analysis of the performance across various DenseNet architectures indicates that the fewer complex variants were excluded from final consideration due to their suboptimal validation and test accuracies. Conversely, the most complex architecture, DenseNet-201 (pipeline 4294), achieved the highest test accuracy among all models evaluated, demonstrating superior generalization capability on unseen data and thereby establishing itself as the leading model within this network family.
- **ConvNeXt CNNs**: Comparative evaluation of ConvNeXt architectures across different data splits reveals marginal differences in test accuracy between the examined pipelines. Specifically, the base version exhibited a slightly larger gap between training and test accuracies, suggestive of potential overfitting. While the base model's performance remains robust, it is considered inferior relative to the Tiny version (pipeline 4280), which presents a more balanced accuracy profile and is thus the preferable choice for final deployment.

These results indicate that under equivalent training conditions, there are no significant differences in the performance of the models on the test data. Rather, certain models are less suitable for final selection due to inferior generalization and stability.

This outcome suggests that the chosen hyperparameter settings possess sufficient flexibility to be effectively applied across diverse models, and that the conducted evaluation provides an appropriate criterion for final selection.

b) Comparison of Selected Models within the Approach

All pipelines evaluated in this approach utilized identical configurations regarding data preprocessing, evaluation metrics, and training protocols. The sole variable affecting their performance was the type of employed neural network model. Table 9 presents the results across the training, validation, and test datasets for all pipelines.

Table 9. Comparative metrics of top performing approaches in the 13-class single-label classification approach

Single-label - 13 Classes							
Weighted F1-Score							
Pipeline	Train	Validation	Test	Pipeline	Train	Validation	Test
ResNet				ConvNeXt - Tiny Version			
4263	0.94	0.6884	0.6517	4280	0.9973	0.7324	0.6898
EfficientNet – B3				DenseNet – 201			
4267	0.9916	0.7242	0.6484	4294	0.9904	0.7027	0.6722

As presented in the table, pipeline number 4280, based on the Tiny ConvNeXt model, achieved the highest F1-Score on both the validation and test datasets. This result indicates the superior generalization capability of this model to unseen data and its robustness when encountering samples outside the training distribution. Accordingly, relying on the principle of stable performance on real-world data (test set), the ConvNeXt-based pipeline (4280) can be considered the optimal choice within this approach. This selection is justified both technically and statistically, underscoring the efficacy of the ConvNeXt model in the single-label classification task of children's book genres across 13 classes.

Optimal Model Selection

The final model selection was undertaken through a rigorous analysis of the performance metrics attained across the training, validation, and test datasets, as delineated in Table 10.

The principal criterion guiding this selection was the identification of a model that not only demonstrated robustness and stability throughout the training and evaluation phases but also exhibited superior generalization capacity when exposed to previously unseen, real-world data.

Based on the evaluation metrics presented in Table 10, Pipeline 4280—implementing the ConvNeXt-Tiny architecture—demonstrates exceptional performance in handling class imbalance and extracting meaningful patterns from the dataset. The model consistently achieves the highest scores across training, validation, and test sets, indicating both robust learning capabilities and strong generalization. This consistent performance across all evaluation phases highlights the architectural superiority and adaptability of ConvNeXt, particularly in managing datasets characterized by high variability and complexity.

For selecting the optimal model, it is essential not only to consider quantitative metrics such as accuracy and weighted F1-score, but also to carefully evaluate how well each model aligns with the specific requirements of the project. Table 11 outlines the strengths and weaknesses of each pipeline, providing a comprehensive basis for informed model selection.

Table 10. Evaluation metrics of the top-performing models across the three approaches in training, validation, and test data classification

15 Classes								***13 Classes***			
Multi-Label				***Single-Label***				***Single-Label***			
Macro F1-Score				***Accuracy***				***Weighted F1-Score***			
ResNet-18								***ConvNeXt-Tiny***			
Test	Validation	Train	Pipeline	Test	Validation	Train	Pipeline	Test	Validation	Train	Pipeline
0.5202	0.5029	0.7124	4037	0.5842	0.6505	0.9969	4190	0.6898	0.7324	0.9973	4280

Table 11. Analysis of strengths and weaknesses in the implementation of selected pipelines

Pipeline 4280 Single-label 13 Classes ConvNeXt	***Strengths***	Use of ConvNeXt with modern architecture (Vision Transformer-inspired)
		Advanced augmentations (e.g., vertical flip, Gaussian blur, random erasing)
		Smooth learning rate scheduling via CosineAnnealingLR
		Improved class balance handling using CBLoss
	Weaknesses	High training complexity and resource requirements
		Potential loss of long-term patterns due to overfitting prevention focus
Pipeline 4190 Single-label 15 Classes ResNet-18	***Strengths***	Lightweight and proven ResNet-18 architecture
		Simple training setup with basic augmentations and StepLR
		Suitable for quick prototyping and experiments
	Weaknesses	Basic augmentations insufficient for complex data variability
		No strategies to address class imbalance
Pipeline 4037 Multi-label 15 Classes ResNet-18	***Strengths***	Supports multi-label tasks with flexible architecture
		Diverse input data through random cropping and normalization
		Easily switchable between classification strategies
	Weaknesses	Unnecessary complexity for single-label scenarios
		No adaptive learning rate or class balancing

Based on a thorough analysis of the pipeline's performance—including both evaluation metrics and technical implementation—we can now more precisely highlight Pipeline 4280's technical strengths and justify why it stands out as the best choice.

Although no prior work exists for Persian children's book cover genre classification, A synthesis of prior studies on visual genre prediction—including book-cover classification [8]-[11], album-cover genre analysis [13]-[14], and movie-poster genre identification [12], [15], [17] —shows that purely image-based CNN models typically achieve Weighted F1-scores in the range of approximately 0.60–0.75. This range reflects common performance levels reported across datasets of comparable size and visual complexity, where genre distinctions rely heavily on stylistic and compositional cues rather than textual metadata.

Our ConvNeXt-Tiny model, achieving a weighted F1-score of 0.6898, falls within this expected performance range, suggesting that despite linguistic and cultural differences, visual cues in children's literature exhibit similarly learnable patterns. While direct numerical benchmarking is not feasible due to dataset incompatibility, this qualitative comparison situates our findings within existing research.

5. Generating Illustrated Children's Stories Using AI Tools

Following the classification process which identified 13 distinct genres of children's literature, our implemented system utilizes these categorized book covers and their genre labels to generate illustrated content through a multi-stage AI-powered process. By integrating various GAI engine APIs, the system produces children's stories that maintain strong thematic alignment with their respective genres while incorporating visual and contextual elements derived from the original book covers. This methodology enables the automated production of age-appropriate, genre-specific illustrated narratives for young readers, demonstrating an effective approach to AI-assisted children's story generation that remains faithful to established literary classifications.

5.1. Illustrated story generation process based on book covers

The process of creating illustrated stories was conducted in four sequential stages.

Extraction of visual elements

The identification of visual elements embedded in children's book covers—such as objects, color schemes, and artistic styles—plays a critical role in guiding the generation of coherent narratives and

their corresponding illustrations. To facilitate this process, Microsoft's Copilot AI, integrated within the Bing search engine, was utilized for visual analysis. By uploading the cover images into the system, the tool automatically extracted salient visual features, which then served as structured inputs for the subsequent story generation phase.

Story generation

The genre and visual elements extracted from the book covers were provided to the ChatGPT model to generate children's narratives. Leveraging the model's advanced NLP capabilities, the objective was to produce stories characterized by three essential attributes: narrative engagement, linguistic fluency, and age-appropriate comprehensibility.

Furthermore, the generated texts were designed to preserve thematic alignment with both the extracted visual elements and the genre assigned to each book cover, thereby ensuring cohesion, simplicity, and relevance for a young audience.

Scene Description Generation

Each story was segmented into a series of discrete "windows," with each window encapsulating a single scene from the narrative. Given the target demographic of children aged 3 to 10 years, the number of sentences per window was limited to a maximum of three [34], and the total number of windows per story was kept below 32 [35]. This segmentation not only facilitated easier narrative comprehension but also supported the generation of scene-specific illustrations, promoting better cognitive and emotional engagement among young readers.

To generate detailed scene descriptions, the Claude model was employed. This model, integrating advanced language processing and visual analysis technologies, produced semantically aligned and context-aware descriptions suitable for use in image generation tools. Each scene description encompassed three principal components:

- **Overall scene setting**: A contextual overview of the location and spatial dynamics of the scene.
- **Visual details**: Specific depictions of characters, objects, colors, and patterns present in the scene.
- **Mood and atmosphere**: An articulation of the emotional tone conveyed, such as excitement, joy, or tranquility.

Image generation

The scene descriptions for each narrative window were subsequently provided to the Copilot AI tool, which utilizes AI-based image synthesis techniques to generate illustrations. These images were required to reflect both the narrative content and the visual style implied by the original book cover, ensuring age-appropriate visual engagement. Ultimately, the generated illustrations were integrated with their corresponding textual segments to construct complete, visually enriched stories tailored for young readers.

5.2. Evaluation of AI-generated Stories

To assess the stories produced by the AI engine, we adopted a qualitative evaluation framework guided by multiple narrative and developmental criteria, and three child psychologists independently reviewed and evaluated the selected stories.

- **Age Appropriateness of the Story**: The story must align with the developmental stage of the child in terms of language, content, and structure. The use of simple and comprehensible concepts, while avoiding complex or inappropriate themes, is a critical factor in designing a children's story. Furthermore, the narratives should address the emotional needs of children, assisting them in coping with fears, challenges, and emotions. When placed in relatable scenarios, children are more likely to respond effectively and benefit significantly from the story's content [36].
- **Positive Moral Lessons**: Children's books offer a valuable opportunity to impart moral values in an indirect and engaging manner. Stories that convey virtues such as kindness, honesty, empathy, teamwork, problem-solving, and acceptance of differences—while maintaining an entertaining and appealing tone—tend to be more effective than direct instruction of these concepts. The objective is to enable children to enjoy the narrative while developing a deeper understanding of human interactions [37].
- **Engaging and Relatable Characters**: Story characters should be crafted in a way that allows children to emotionally connect with them and reinforce learning through imitation of their traits. When a child can identify with a character and place themselves in the character's situation, they become more immersed in the story and demonstrate greater interest. Characters that face realistic challenges—such as success, failure, joy, or fear—foster a sense of empowerment in children and motivate them to confront their own difficulties [38-39].
- **Cultural Sensitivity and Inclusivity**: Stories that embrace cultural diversity and demonstrate respect for various values can broaden a child's perspective on the world. Such narratives help foster an appreciation for

differences and teach children that every culture, language, and background hold intrinsic value [40].

- **Emotional Impact**: Stories should be crafted to evoke emotional engagement in children. The tone of the narrative should avoid inducing excessive stress or sadness and instead promote a sense of security and hope [41-42]. Narratives that portray emotional challenges but ultimately conclude with a reassuring and uplifting ending can significantly contribute to the emotional development of children and enhance their ability to understand both their own emotions and those of others.
- **Encouragement of Imagination and engagement**: Stories that stimulate a child's imagination can enhance creative thinking and problem-solving abilities. Narratives involving fantastical worlds, magical characters, or unusual scenarios encourage exploration and learning. Additionally, stories that pose questions or present challenges solved through creativity and teamwork can ignite curiosity and motivate children toward active thinking and participation.

The evaluation phase was conducted manually by child psychologists. Given that a comprehensive review of all generated stories would be prohibitively costly and labor-intensive, we adopted a sampling approach for efficient assessment. From each genre, two stories were randomly selected, resulting in a final sample of 26 stories for evaluation.

During the evaluation phase, the generated stories were carefully reviewed and assessed against the aforementioned criteria. Table 12 presents the evaluation results for three of the 26 generated stories.

As part of the comparative evaluation between AI-generated stories and authentic children's books, a targeted selection of excerpts from original publications was undertaken to serve as representative samples of overall narrative structures. This selection included the official synopsis as well as excerpts from the beginning, middle, and ending of each story to ensure a concise yet comprehensive review of semantic, linguistic, and narrative elements. This approach maintained the analytical rigor while enabling effective assessment of components such as meaning, narrative flow, and resolution.

In the comparison process between real stories and AI-generated ones, four key components were analyzed:

- **Lexical Similarity** (linguistic matching, word frequency): This criterion aims to measure the superficial proximity of texts based on word count and the frequency of shared vocabulary. It is considered one of the simplest yet effective comparative methods to assess the degree of surface-level similarity between two texts.
- **Concept and Meaning** (theme, message, audience comprehension): This component evaluates the depth of the story, including its moral and emotional values, as well as the ultimate message conveyed. It is a critical element in children's content creation and reflects the difference between human and machine perspectives on conceptualization.
- **Narrative Progression** (story structure, sequence of events, coherence): This dimension analyzes the story's coherence, logical flow of events, and the presence of conflict and resolution, indicating narrative maturity and the ability of AI systems to produce story-driven content.
- **Ending and Conclusion** (final message, emotional impact, story closure): This assesses how the story concludes and whether it delivers a tangible, instructive, or emotionally resonant ending—an essential factor for impacting the child reader.

Table 13 compares these criteria in the original text versus the AI-generated story (whose cover design is shown in Figure 2).

The comparison between original children's stories and AI-generated versions reveals that despite stylistic, linguistic, and structural differences, there are notable overlaps in terms of content, message, and narrative elements. The original stories often feature richer language, multilayered narration, and thought-provoking themes; whereas the AI-generated versions, employing simpler structures, fluent language, and an emphasis on visual or adventurous elements, manage to provide an accessible and engaging experience for the child audience.

Figure. 2. The cover design of the book: when I was Da Vinci's housemate

Table 12. Analysis of stories based on children's literature evaluation criteria

Story Title	Genre	Age Appropriateness	Moral Lessons	Characters	Cultural Sensitivity	Emotional Impact	Imagination & Creativity
The Lost World	Adventure	Creative and entertaining (Ages 6–10)	Cooperation, gratitude, problem-solving through creativity	Clover (a hen), Shelly (a snail), the Guardian Fox	Imaginary world enabling intercultural empathy among children	Clear excitement, humorous and satisfying ending	Unique concept: Sock Forest and surreal portals
Madame Charlotte's Strange Journey	Humorous	Joyful and interactive learning experience (Ages 6–10)	Teamwork, perseverance, joy of discovery	Madame Charlotte and her students (Pippin, Leila, Omar, Samir, Hana)	Multicultural classroom with diverse names promoting coexistence	Playful, inspiring, and engaging teamwork dynamics	Magic hat, school adventures, bright and enchanted classrooms
Luca – Leonardo da Vinci's Curious Housemate	Biographical	Simplification of genius-related concepts, inspiring curiosity	Curiosity, creativity, experiential learning	Luca and Leonardo da Vinci	Renaissance Italy enriched with historical and cultural depth	Sense of admiration, enthusiasm, and creative wonder	Flying machines, paintings, inventions – the imaginative world of da Vinci

Table 13. Comparative analysis of a real book and its AI-generated story based on key narrative components

Parameter	When I Was Da Vinci's Housemate (Real Story)	Luca: The Curious Housemate of Leonardo da Vinci (AI-Generated)
Word Count	*More than 440 words (summary + selected paragraphs from Chapters 1, 5, 6, and 10)*	*655 words (entire story)*
Shared Vocabulary (Objects, People, Concepts)	*Leonardo/Da Vinci, painting, invention, flying machine, secret, Mona Lisa, notebook, master, curiosity, king, flight, art, plants, nature* ***Note:*** *Estimated –15 overlapping key concepts or words, indicating high thematic similarity (creativity, discovery, and admiration for Da Vinci) despite differences in language and style.*	
Theme and Meaning	***Core theme:*** *Mysterious, reflective, humorous, driven by childlike curiosity* ***Interpretation:*** *Da Vinci appears as an enigmatic figure—brilliant but somewhat obscure* ***Narrative tone:*** *Imaginative, lightly mysterious, and humorous* ***Conclusion:*** *AI story is more idealized and didactic; real story blends humor and mystery. Both, however, center on curiosity, mentorship, and admiration.*	***Core theme:*** *Educational, entertaining, and creatively inspiring* ***Interpretation:*** *Da Vinci is portrayed as a kind and brilliant mentor who encourages learning and creativity* ***Narrative tone:*** *Positive, inspirational, and calm* ***Conclusion:*** *More structured in the tradition of children's literature—challenge → lesson → resolution.*
Narrative Structure	***Beginning:*** *The narrator becomes Da Vinci's housemate; adventures unfold gradually* ***Plot development:*** *Includes questions, challenges, and new characters; narrator experiences mental and emotional growth* ***Climax:*** *Flight experiment, public reactions, awe at Da Vinci's genius, and internal dilemmas* ***Resolution:*** *Emotionally and intellectually layered ending*	***Beginning:*** *Luca becomes Da Vinci's housemate; instantly immersed in inventions and excitement* ***Plot development:*** *Encounters with machines, art lessons, puzzles, and nature* ***Climax:*** *Discovering secrets, solving puzzle boxes, viewing inventions, drawing lessons from Da Vinci* ***Resolution:*** *Follows a more linear, conventional child-story format*
Ending and Impact	***Closure:*** *Philosophical reflection—everyone has a role in life; the narrator was a "good companion" to Da Vinci* ***Story message:*** *Deep respect for genius, introspection, and a mysterious bond of friendship*	***Closure:*** *Luca learns from Da Vinci, continues on his path, feels inspired, and shares his story* ***Story message:*** *Growth through learning, curiosity, honoring a mentor*
Conclusion	*Both narratives conclude with a sense of admiration for Da Vinci, albeit with different emotional tones.* ***Real Story:*** *Thought-provoking and emotionally resonant* ***AI Story:*** *Uplifting, clear, and optimistic*	

These AI-generated stories successfully preserve core aspects such as inspiring characters, human relationships, adventure, and cognitive activities like exploration or problem-solving. They reconstruct reflections of the original narratives while maintaining machine-specific storytelling features such as clarity, organization, and linear progression.

Such a comparison suggests that GAI can effectively reproduce key conceptual elements of storytelling, although it may face limitations in conveying the nuanced human subtleties, implicit humor, or psychological complexities embedded in authentic children's literature.

6. Conclusions

This research presents the development of an ML-based framework for the classification of children's book genres utilizing cover image analysis. The proposed model effectively categorizes books into thirteen distinct genres and underpins an automated story generation system, which generates creative

narratives based on cover imagery, genre, and thematic elements.

6.1. Limitations

This study encountered several limitations related to both data quality and computational resources.

Data-Related Limitations

A major challenge was the absence of any standardized dataset of Persian children's book covers. The dataset had to be constructed manually from online bookstores, relying on the genre labels provided by retailers, which may contain inconsistencies. Some images were of low resolution or visually ambiguous, reducing classification accuracy. Genre boundaries in children's literature are inherently fuzzy, and certain book covers share visual similarities across genres, limiting model discriminability. The dataset was also imbalanced, with some genres being underrepresented.

Resource-Related Limitations

Training modern deep learning models required computing resources beyond the capacity of personal devices or free cloud platforms. Additionally, the reliance on proprietary LLMs and image-generation tools limits full reproducibility. These limitations highlight the need for larger, publicly available datasets, open-source multimodal generation tools, and more robust evaluation frameworks for future research.

Evaluation-Related Limitations

The evaluation relied on qualitative expert reviews, which are time-consuming, costly, and inherently subjective. Differences in expert judgment may influence the results, and the small evaluation group limits generalizability. Future work should involve a larger pool of reviewers and develop clearer, standardized criteria—or quantitative metrics—for assessing the quality and genre alignment of AI-generated children's stories.

6.2. Future Work

This study represents a pivotal step toward the digital transformation of children's publishing by advancing the development of automated digital libraries equipped with genre-based content filtering capabilities. Building on this foundation, several promising research directions emerge. First, we aim to extend the system by integrating real-time feedback mechanisms from children, enabling dynamic story generation that adapts to instantaneous user input. Second, we plan to broaden the scope beyond literary narratives to explore applications in educational and therapeutic contexts. These advancements would not only enhance the system's versatility but also amplify its broader societal impact.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

MM: proposed the idea and study design, supervised the project, revised, and approved the final version.

RN: developed the model, performed simulations, collected the data, and drafted the manuscript.

Conflict of interest

The authors declare that no conflicts of interest exist.

References

[1] M. Sunderland, Using Story Telling as a Therapeutic Tool with Children, London: Routledge, 2017.

[2] G. Trionfi and E. Reese, "A Good Story: Children With Imaginary Companions Create Richer Narratives," Child Development, vol. 80, pp. 1301-1313, 2009.

[3] A. Nicolopoulou, "Children and Narratives," in Narrative Development, New York, Routledge, 1997, p. 37.

[4] B. Seuling, How to write a children's book and get it published, New York City: Wiley, 2004.

[5] M. Evans and J. Saint-Aubin, "What children are looking at during shared storybook reading," Psychol Sci., vol. 16, pp. 913-920, 2005.

[6] R. E. Mayer, Multimedia Learning, 3rd ed., Cambridge: Cambridge University Press, 2020.

[7] Y. Li, X. Zhiding, H. Wenxin and Z. Xian, "Enhancing Visual Storytelling with Multi-Modal Large Language Models," in 31st International Conference on Computational Linguistics, Abu Dhabi, 2025.

[8] G. R. Biradar, R. JM, A. Varier and M. Sudhir, "Classification of Book Genres Using Book Cover and Title," in IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), Visakhapatnam, 2019.

[9] P. Buczkowski, A. Sobkowicz and M. Kozlowski, "Deep Learning Approaches towards Book Covers Classification," in Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods, Funchal, Madeira, 2018.

[10] R. Jayaram, M. Harshitha, S. Pavithra, B. Munshira Noor and K. J. Bhanushree, "Classifying Books by Genre Based on Cover," International Journal of Engineering and Advanced Technology, vol. 9, no. 5, pp. 530-535, 30 June 2020.

[11] C. S. Kundu, "Book Genre Classification By Its Cover Using A Multi-View Learning Approach," Masters Theses & Specialist Projects, Kentucky, 2020.

[12] S. Sung and R. Chokshi, "Classification of movie posters to movie genres," California, 2018.

[13] S. Oramas, O. Nieto, F. Barbieri and X. Serra, "Multi-Label Music Genre Classification from Audio, Text and Images Using Deep Features," in 18th International Society for Music Information Retrieval Conference, Suzhou, 2017.

[14] J. Li, D. Sun and T. Cai, "Genre Classification via Album Cover," Stanford University, California, 2019.

[15] J. A. Wi, S. Jang and Y. Kim, "Poster-Based Multiple Movie Genre Classification Using Inter-Channel Features," IEEE Access, vol. 8, pp. 66615-66624, 2020.

[16] J. Kim and H.-J. Suk, "Prediction of the Emotion Responses to Poster Designs based on Graphical Features: A Machine Learning-Driven Approach," Archives of Design Research, vol. 33, no. 2, pp. 39-55, 31 May 2020.

[17] U. K. Nareti, C. Adak and S. Chattopadhyay, "Demystifying Visual Features of Movie Posters for Multi-Label Genre Identification," IEEE Transactions on Computational Social Systems, Doi: 10.1109/TCSS.2024.3481157, 2024.

[18] S. Pooranalingam, "Film Poster Design: Understanding Film Poster Designs and the Compositional Similarities within Specific Genres," Spectrum, no. 12, 8 January 2024.

[19] L. Xiaochuan and C. Xiangyong, "Improving Visual Storytelling with Multimodal Large Language Models," arXiv:2407.02586, 2024.

[20] C. Zang, J. Tang, R. Zhang, Z. Zhao, T. Lv, M. Pei and W. Liang, "Let Storytelling Tell Vivid Stories: An Expressive and Fluent Multimodal Storyteller," arXiv:2403.07301, 2024.

[21] S. Yang, Y. Ge, Y. LI, Y. Chen, Y. Ge, Y. Shan and Y.-C. Chen, "SEED-Story: Multimodal Long Story Generation with Large Language Model," CoRR abs/2407.08683, 2024.

[22] T. Huang, E. Qasemi, B. Li, H. Wang, F. Brahman, M. Chen and S. Chaturvedi, "Affective and Dynamic Beam Search for Story Generation," in Findings of the Association for Computational Linguistics: EMNLP, Singapore, 2023.

[23] A. Alabdulkarim, W. Li, L. J. Martin and M. O. Riedl, "Goal-Directed Story Generation: Augmenting Generative Language Models with Reinforcement Learning," 10.48550/arXiv.2112.08593, 2021.

[24] J. Canary, "Transfer Learning: Leveraging Pretrained Models - Jim Canary - Medium," Medium, 28 1 2025. [Online]. Available: https://medium.com/%40jimcanary/transfer-learning-leveraging-pretrained-models-153ab99b9b00.

[25] K. Juntae, H. Yoonseok, Y. Hogeon and N. Jongho, "A Multi-Modal Story Generation Framework with AI-Driven Storyline Guidance," Electronics, vol. 12, no. 6, p. 1289, 2023.

[26] J.-B. Alayrac et al., "Flamingo: a visual language model for few-shot learning," Advances in Neural Information Processing Systems (NeurIPS), 2022.

[27] D. Driess et al., "PaLM-E: An embodied multimodal language model," arXiv preprint arXiv:2303.03378, 2023.

[28] D. Zhu et al., "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," arXiv preprint arXiv:2304.10592, 2023.

[29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," Advances in Neural Information Processing Systems, vol. 36, pp. 34892–34916, 2023.

[30] B. Hejazi, Children's and Adolescents' Literature: Features and Aspects (In Persian), Tehran: Roshangaran Publications, 2023.

[31] C. Zauner, "pHash – Perceptual Hash Library," [Online]. Available: https://phash.org/docs/design.html.

[32] M. F. Uddin, "Addressing Accuracy Paradox Using Enhanced Weighted Performance Metric in Machine Learning," in Sixth HCT Information Technology Trends (ITT), United Arab Emirates, 2019.

[33] X.-Z. Wu and Z.-H. Zhou, "A Unified View of Multi-Label Performance Measures," in 34th International Conference on Machine Learning, Sydney, 2017.

[34] L. M. Justice and P. C. Pullen, "Promising Interventions for Promoting Emergent Literacy Skills: Three Evidence-Based Approaches," Intervention in Scholl and Clinic, vol. 39, p. 87–98, 2003.

[35] L. R. Buccieri and P. Economy, Writing Children's Books for Dummies, Wiley, 2011.

[36] S. Earnshaw, The Handbook of Creative Writing, Edinburgh: Edinburgh University Press Ltd, 2014.

[37] H. Rahim and M. D. H. Rahiem, "The Use of Stories as Moral Education for Young Children," International Journal of Social Science and Humanity, vol. 2, pp. 454-458, 2012.

[38] Booka, "What Makes a Good Children's Book: 10 Important Characteristics," [Online]. Available: https://appbooka.com/blog/what-makes-a-good-childrens-book.

[39] Brett, "How to Tell Awesome Stories to Your Kids," Brett, 22 October 2020. [Online]. Available: https://www.artofmanliness.com/people/family/how-to-tell-awesome-stories-to-your-kids/. [Accessed 28 December 2024].

[40] A. McCabe, "Developmental and Cross-Cultural Aspects of Children's Narration," in Narrative Development, New York, Routledge, 1997, p. 38.

[41] J. R. Brown and J. Dunn, "Continuities in Emotion Understanding from Three to Six Years," Child Development, vol. 67, pp. 789-802, 1996.

[42] M. N. Sala, F. Pons and P. Molina, "Emotion regulation strategies in preschool children," British Journal of Developmental Psychology, vol. 32, p. 10.1111/bjdp.12055, 2014.

Maedeh Mosharraf is an Assistant Professor of Computer Engineering Software and Information Systems at Shahid Beheshti University (SBU) where she has served since 2021. She received her M.Sc. (2013) and Ph.D. (2019) in Computer Engineering from University of Tehran (UT). Her current research involves "Blockchain technologies and cryptocurrency", "Digital transformation", "Electronic commerce", and "Technology enhanced learning".

Reyhaneh Naseri Moghadam is a Computer Engineering graduate (SBU) with research interests in artificial intelligence, machine learning, and data analytics. Her work includes deep learning–based image classification and predictive modeling applied to real-world datasets. She has also been involved in the development of data-driven enterprise applications. Her academic interests focus on applying modern AI techniques to practical and interdisciplinary research problems.