Enhancing Book Recommendation with Automated Genre Mining

MS(Research) Thesis

By

Prolay Mallick

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Enhancing Book Recommendation with Automated Genre Mining** in the partial fulfillment of the requirements for the award of the degree of **MS(Research)** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore,** is an authentic record of my work carried out during the period from August 2023 to May 2025. The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Prolay Mallick 22/05/2025

Signature of the Student with Date

(Prolay Mallick)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Soumi 22/5/25

Signature of Thesis Supervisor with Date

(Dr. Soumi Chattopadhyay)

Prolay Mallick has successfully given his MS(Research) Oral Examination held on

---

Signature of Head of Discipline

Date: 23/5/25

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my heartfelt gratitude to a number of persons who in one or the other way contributed by making this time as learnable, enjoyable, and bearable. At first, I would like to thank my supervisor **Dr. Soumi Chattopadhyay's** constant source of inspiration during my work. Without her constant guidance and research directions, this research work could not have been completed. Her continuous support and encouragement have motivated me to remain streamlined in my research work.

I would also like to acknowledge **Dr. Chandranath Adak**, Assistant Professor at IIT Patna, for guiding me to do this dissertation. Without his precious support, it would not be possible to conduct this project.

I would also like to acknowledge **Mr. Suraj Kumar**, PhD Scholar at IIT Indore and **Mr. Utsav Kumar Nareti**, PhD Scholar at IIT Patna, for guiding and developing my technical and soft skills, I am gratefully indebted to his very valuable contribution to this project.

I am also grateful to **Dr. Ranveer Singh**, HOD of Computer Science and Engineering, for his help and support.

My sincere acknowledgement and respect to **Prof. Suhas S. Joshi**, Director, Indian Institute of Technology Indore for providing me the opportunity to explore my research capabilities at Indian Institute of Technology Indore.

I would like to express my heartfelt respect to my parents for their love, care and support they have provided to me throughout my life.

Finally, I am thankful to all who directly or indirectly contributed, helped and supported me.

Prolay Mallick

Abstract

Genre classification of books plays a pivotal role in enhancing the overall user experience in the rapidly evolving digital landscape of literature consumption. As the number of e-books and digital titles continues to grow exponentially, readers often face an overwhelming amount of content, making it increasingly difficult to discover books that match their interests. Accurate genre classification not only simplifies this discovery process but also serves as a foundational element for personalized recommendation systems, intelligent content organization, and improved search relevance. These systems rely on genre tags to guide readers through curated suggestions, thereby increasing engagement, satisfaction, and the likelihood of continued platform usage. Furthermore, for publishers and platform designers, effective genre classification provides insights into emerging trends, audience preferences, and market dynamics, supporting decisions related to marketing, acquisition, and catalogue management.

This thesis investigates the task of automated book genre classification through two complementary yet independent modalities: visual semantics derived from book cover images and textual semantics mined from book blurbs and user-generated reviews. Each modality addresses specific challenges in genre inference and serves distinct practical use cases. The visual approach is particularly beneficial when textual metadata is unavailable, unreliable, or intentionally misleading, while the textual analysis is better suited to capture deeper thematic and narrative content.

In the visual modality, we propose a robust deep learning-based framework that leverages the Swin Transformer architecture to extract and interpret genre-relevant visual cues from cover images. Book covers, often designed to encapsulate the book's tone, target audience, and thematic elements, serve as a rich source of visual information. The Swin Transformer, with its hierarchical structure and window-based attention mechanism, enables both local and global feature learning, making it highly suitable for modeling the spatial complexity and design variation present in book covers. The extracted features are processed through a hierarchical classification scheme comprising two stages: a Level-1 classifier that distinguishes between Fiction and

Non-Fiction, and a Level-2 classifier that further refines the prediction into specific subgenres. This method ensures coarse-to-fine granularity in genre labeling, while maintaining scalability for large genre taxonomies. Experimental evaluations show that this model significantly outperforms traditional CNN-based approaches, demonstrating strong generalization even across visually diverse cover designs.

In the textual modality, we focus on mining semantic insights from book descriptions and reader reviews. While descriptions provide a structured summary authored by publishers, user reviews offer subjective, experience-driven commentary that often reveals latent genre indicators. However, reviews are also noisy, sentiment-driven, and vary in relevance. To overcome this, we introduce a two-phase architecture that begins with a semantic filtering stage. Using cosine similarity between contextual embeddings of the book description and individual reviews (generated via BERT), the system selects only those reviews that are contextually aligned with the book's core content. This is followed by hierarchical multi-label classification, where the encoded representations of the filtered reviews are used to predict both broad categories and fine-grained genres. This approach enables the system to learn both lexical and semantic genre patterns from crowdsourced content, improving performance over models that use raw, unfiltered reviews.

Experimental results across both modalities demonstrate the effectiveness of the proposed methods in accurately identifying book genres. Importantly, each modality operates independently, allowing for flexible deployment depending on data availability. For example, platforms lacking access to user reviews can rely solely on visual analysis, while review-rich environments can utilize textual classification. Together, these contributions offer a scalable and adaptable framework for genre mining, with implications for digital cataloging, metadata enrichment, personalized recommendations, and automated literary curation.

This work contributes to the advancement of content-aware systems in the digital publishing ecosystem. By leveraging both visual and textual modalities, it sets a strong foundation for future research in multimodal genre classification and intelligent book discovery tools tailored to reader preferences in diverse contexts.

Table of Content

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Challenges in Book Genre Identification 2

1.3 Contribution 4

1.3.1 Literature Review Contribution 4

1.3.2 Book Genre Classification from Reliable Sources 5

1.3.3 Book Genre Classification from Unreliable Reviews Refines with Blurbs 5

1.4 Organizing the Thesis 6

2 Literature Review **9**

2.1 Background 9

2.2 Review 11

2.2.1 Visual Method 12

2.2.2 Textual Method 14

2.2.3 Multimodal Method 16

2.3 Limitations of Existing Work 18

3 Dataset Creation **21**

3.1 Dataset details 21

3.1.1 Dataset Metadata 21

3.2 Dataset Creation Process 23

3.2.1 Cover Page Image Collection 23

3.2.2 Blurb Collection . . . 23
3.2.3 User Reviews Collection . . . 24
3.2.4 Data Annotation and Validation . . . 24
3.2.5 Final Dataset Statistics . . . 25
3.3 Challenges in Dataset . . . 25
3.3.1 Mitigation of Dataset Issues . . . 27
3.4 Dataset Analysis . . . 30
3.4.1 Hierarchical Genre-wise Count of this Dataset: . . . 31
3.4.2 Statistical Analysis of Cover Page Images: . . . 31
3.4.3 Characteristics of Visual Dataset: . . . 32

4 Book Genre Classification from Reliable Sources 35
4.1 Problem Formulation . . . 35
4.1.1 Problem Definition . . . 35
4.1.2 Mathematical Representation . . . 36
4.1.3 Objective . . . 37
4.2 Proposed Method . . . 37
4.2.1 Feature Extractor: Swin Transformer . . . 37
4.2.2 Level 1 Classifier: Binary Genre Classifier . . . 40
4.2.3 Level 2 Classifiers: Multi-label Subgenre Prediction . . . 41
4.3 Experimental Analysis . . . 42
4.3.1 Evaluation Metrics . . . 42
4.3.2 Experiment . . . 45
4.3.3 Experiment Result . . . 45
4.3.4 Genrewise Analysis . . . 46
4.3.5 Genre Co-occurrence Heatmaps . . . 48
4.4 Summary . . . 49

5 Book Genre Identification from Unreliable Reviews Refines with Blurbs 53
5.1 Problem Formulation . . . 53

5.2 Proposed Method 54
5.2.1 Review Filtering and Vocabulary Creation 54
5.2.2 Feature Extractor: BERT 57
5.2.3 Level-1 Binary Classifier 61
5.2.4 Level-2 Multi-Label Classifier 62
5.3 Experimental Analysis 63
5.3.1 Evaluation Metrics 63
5.3.2 Experiment 66
5.3.3 Experiment Result 66
5.4 Summary 67

6 Conclusion and Future Work **69**
6.1 Conclusion 69
6.2 Future Work 70

Dissemination of the Thesis **72**

Enhancing Book Recommendation with Automated Genre Mining

A THESIS

submitted to the

INDIAN INSTITUTE OF TECHNOLOGY INDORE

in partial fulfillment of the requirements for

the award of the degree

of

MS(Research)

By

Prolay Mallick

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2025

Chapter 1

Introduction

1.1 Motivation

The digital transformation of literary consumption has led to a rapid expansion of online book repositories and digital libraries, fundamentally changing the way readers discover, categorize, and engage with books. As the volume of accessible literary content continues to grow exponentially, effective genre classification has become increasingly critical for enhancing searchability, improving recommendation systems, and supporting large-scale digital libraries. Accurate genre tagging enables readers to easily navigate vast collections, discover books aligned with their preferences, and explore new genres with confidence.

Traditional genre classification methods primarily rely on structured metadata, such as book titles, blurbs, author information, and publisher-provided descriptions. While these metadata-based approaches have been effective to some extent, they are not without limitations. Metadata is often noisy, inconsistent, or even completely unavailable, particularly for newly released or independently published books. Moreover, it is largely text-based and fails to capture the multi-layered and nuanced nature of genre definitions. Many books span multiple genres—like a novel that could simultaneously belong to "Science Fiction," "Romance," and "Adventure"—which traditional flat classification methods struggle to represent accurately.

To address these challenges, this research is driven by two key motivations:

- **Leveraging Visual Semantics from Book Covers:** Book covers are more than decorative elements; they are carefully designed visual representations intended to convey genre, mood, and thematic essence. Elements such as color schemes, typography, imagery, and composition are often curated to signal the book's genre, target audience, and stylistic tone. Despite their ubiquity and visual richness, cover images remain underutilized in genre classification tasks. Recent advancements in deep learning and computer vision provide an opportunity to tap into the visual semantics of book covers, enabling models to decode genre-specific visual cues. This modality, being universally available and independent of language barriers, can significantly enhance genre prediction without relying solely on textual metadata.

- **Mining Semantic Insights from User-Generated Reviews:** User-generated reviews are a promising yet underexplored resource for understanding genre. Unlike structured metadata, reviews are rich in narrative descriptions, emotional tone, and thematic analysis written by readers who have engaged deeply with the content. These reviews offer unique insights into genre-specific elements that are often absent from traditional metadata. However, the unstructured nature of reviews introduces challenges such as noise, sentiment bias, and inconsistent terminology. To address this, a semantic filtering mechanism is proposed to filter out irrelevant or sentimentally skewed reviews, retaining only those that are semantically aligned with the book's thematic content. This step enables cleaner and more precise genre mining from crowdsourced text.

1.2 Challenges in Book Genre Identification

Book genre identification plays a pivotal role in enhancing digital library experiences and powering recommendation systems. However, the task is fraught with multiple challenges stemming from the complexity of literary content, variability in expression, and limitations of existing metadata. Below, we outline the primary challenges faced in genre identification:

- **Genre Overlap and Ambiguity:** Many books do not fit neatly into a single genre but span multiple categories such as "Science Fiction," "Romance," and "Adventure." This overlap introduces ambiguity, making it difficult for classification models to disambiguate genres based solely on textual or visual cues. Furthermore, subgenres often inherit characteristics from parent genres, complicating the boundary definitions.

- **Limited and Noisy Metadata:** Traditional genre classification systems primarily rely on structured metadata such as blurbs, author names, and publisher descriptions. However, this metadata is often incomplete, inconsistent, or missing for newly published or independently released books. Furthermore, metadata can be noisy or misleading, affecting the reliability of genre predictions.

- **Subjectivity in User Reviews:** User-generated reviews provide rich semantic information but are inherently subjective and sentiment-driven. Personal biases, emotional reactions, and varied interpretations contribute to noisy data, complicating the task of extracting consistent genre-specific signals. Filtering out non-informative or sentiment-biased content is a crucial yet challenging step.

- **Cross-Language and Cultural Differences:** Books are published and reviewed in multiple languages, with cultural interpretations affecting genre perception. For instance, what is considered "Thriller" in one cultural context may overlap with "Mystery" or "Drama" in another. This cross-linguistic and cross-cultural variability demands robust language-agnostic models for effective classification.

- **Visual Abstraction and Symbolism in Covers:** Book covers are designed to capture symbolic elements and thematic moods rather than explicit genre labels. For instance, dark color palettes may suggest "Horror" or "Mystery," while vibrant, colorful designs may indicate "Children's Literature" or "Fantasy." Decoding these abstract visual signals into specific genres requires advanced computer vision models capable of high-level semantic understanding.

- **Hierarchical Genre Structuring:** Literary genres are inherently hierarchical, with broad categories like "Fiction" and "Non-Fiction" containing numerous subgenres. Traditional classification methods often treat genres as flat labels, ignoring the parent-child relationships within the genre taxonomy. This limitation reduces the granularity of predictions and affects multi-label classification.

- **Data Imbalance Across Genres:** Genre distribution in book datasets is often imbalanced, with certain popular genres being significantly overrepresented, while more niche categories remain underrepresented. This imbalance biases learning algorithms, making it challenging for models to generalize well across less-populated categories and affecting the accuracy of genre classification in underrepresented segments.

Addressing these challenges requires a robust framework that can effectively integrate visual and textual cues while handling noise, cross-language variability, and hierarchical genre relationships. Our proposed approach aims to bridge these gaps through a structured genre classification mechanism that leverages both book covers and user-generated content for richer, more accurate genre predictions.

1.3 Contribution

This thesis presents significant contributions across three dimensions: critical review of existing literature, development of a visual-based genre classification framework, and a complementary textual analysis-based classification system. These contributions are structured as follows:

1.3.1 Literature Review Contribution

A thorough literature survey was undertaken to understand the current landscape of book genre classification techniques. Key contributions in this area include:

- Categorization of prior work into visual, textual, and multimodal methodologies.

- Identification of critical gaps such as lack of hierarchical genre modeling, outdated cover image datasets, and insufficient handling of genre ambiguity.

- Evaluation of the limitations of unimodal systems and establishment of the motivation for a dual-modality approach combining visual and textual cues.

1.3.2 Book Genre Classification from Reliable Sources

The first major technical contribution of this thesis involves developing a deep learning-based framework to classify book genres solely from cover images. The primary elements include:

- Design and implementation of a hierarchical two-level classification pipeline using the Swin Transformer architecture.

- Level-1 binary classifier distinguishes between *Fiction* and *Nonfiction*; Level-2 classifier performs fine-grained multi-label genre prediction across 30 sub-genres.

- Integration of advanced data augmentation techniques using *Stable Diffusion* to synthetically enhance visual diversity and address class imbalance.

- Robust evaluation across multiple vision models (e.g., ResNet, ViT, BLIP), establishing Swin Transformer as the best-performing backbone in terms of accuracy and generalization.

1.3.3 Book Genre Classification from Unreliable Reviews Refines with Blurbs

The second technical contribution extends the genre classification framework by incorporating semantic information from book descriptions and user reviews. This component consists of:

- A novel review filtering mechanism using cosine similarity to retain only those user reviews that align closely with the book's description.

- Use of BERT-based embeddings to extract deep semantic representations from both descriptions and filtered reviews.
- Implementation of a hierarchical multi-label classifier, aligned with the structure of the visual model, enabling consistent genre taxonomies across modalities.
- Application of large language model APIs (e.g., Gemini) for augmenting incomplete or non-English descriptions to improve text quality and classification performance.

1.4 Organizing the Thesis

This thesis is organized as follows:

- **Chapter 1** introduces the research problem, outlines the motivation behind automated genre mining, identifies key challenges in genre classification, and summarizes the main contributions of the thesis.
- **Chapter 2** provides a comprehensive literature review of previous work in visual, textual, and multimodal genre classification. It also identifies critical limitations in existing methods and highlights the need for a dual-modality approach.
- **Chapter 3** describes the dataset creation process, including metadata curation, web scraping of book covers, blurbs, and user reviews, data preprocessing, annotation strategies, and augmentation techniques. It concludes with a statistical and qualitative analysis of the dataset.
- **Chapter 4** presents the first major technical contribution: a hierarchical visual genre classification framework using Swin Transformer. It includes the problem formulation, architecture, training setup, experimental evaluation, and results analysis.
- **Chapter 5** introduces the second technical contribution: a textual genre classification model leveraging BERT-based review filtering and description integration.

It details the proposed method, hierarchical classifiers, experiment setup, and result discussion.

- **Chapter 6** summarizes the key findings and contributions of the thesis and discusses possible directions for future research, such as multimodal integration and multilingual genre classification.

Chapter 2

Literature Review

2.1 Background

The rapid expansion of digital content in the modern era has revolutionized how books are discovered, categorized, and recommended. With the continuous rise of e-books and online reading platforms, the sheer volume of available literary works has grown exponentially. In this vast digital landscape, effective genre classification has emerged as a cornerstone for enhancing user experience, powering intelligent search engines, and enabling personalized content delivery in online bookstores and digital libraries. Accurate genre classification allows readers to seamlessly discover books that align with their preferences, while also enabling efficient organization and retrieval of content in digital ecosystems.

Traditionally, genre identification has relied heavily on metadata-based methods that utilize structured information such as blurbs, author details, and publisher information. These sources provide foundational context for genre categorization, yet they often fall short of capturing the nuanced thematic elements of literary works, particularly in cases of multi-genre and cross-genre books. Furthermore, metadata is sometimes noisy, inconsistent, or altogether unavailable—especially for newly released or lesser-known titles. This lack of consistency introduces ambiguity in genre prediction and can degrade the performance of recommendation systems.

In addition to structured metadata, **book covers** have emerged as a strategic yet

underexplored source of genre information. Book covers are not merely decorative elements; they are designed with visual cues that reflect tone, target audience, and thematic essence, offering a visually intuitive hint of the book's content. The color palette, typography, imagery, and layout are often carefully curated to resonate with specific genres. For instance, dark and moody aesthetics are typical of horror and thriller genres, while bright and playful designs are often associated with children's literature. These visual indicators make book covers a compelling modality for genre classification. With advancements in deep learning and computer vision, it is now feasible to extract meaningful visual representations from book covers. Techniques such as Convolutional Neural Networks (CNNs) and Transformer-based architectures enable models to capture hierarchical and abstract visual features, making the visual modality a viable input for automated genre classification. Visual-based genre mining has the potential to operate independently of noisy or incomplete textual metadata, thus contributing to more robust and scalable recommendation systems.

Parallel to visual analysis, **user-generated book reviews** present another promising avenue for genre mining. Unlike metadata, which is static and often limited in scope, reviews are dynamic, continually generated, and rich with semantic insights. These reviews, written by readers who have engaged deeply with the book's content, offer valuable perspectives on narrative themes, emotional tone, character development, and genre-specific characteristics that are often not captured by traditional metadata alone. They encapsulate subjective perceptions and thematic impressions, adding a layer of contextual understanding that can enhance traditional classification methods. However, mining genres from user reviews introduces challenges, including noise, reader subjectivity, and linguistic ambiguity, which complicate straightforward genre identification. Sentiment biases, off-topic discussions, and language inconsistencies often obscure the true thematic essence of the book.

To address these challenges, this research introduces a **dual-modality framework for book genre classification** that leverages both **visual cues from book covers** and **semantic insights from user-generated reviews**. The proposed framework is designed to seamlessly integrate these two rich modalities:

- **Visual Genre Mining:** Advanced deep learning architectures are utilized to extract hierarchical and abstract visual features from book covers. By bridging the gap between visual design elements and genre semantics, this approach aims to enhance visually driven book recommendation systems. The model is designed to capture fine-grained visual patterns that are indicative of genre categories, moving beyond simple pattern recognition to include thematic and stylistic representations.

- **Textual Genre Mining:** User-generated reviews are harnessed to mine semantic insights for genre classification. This process includes

  - **Semantic Filtering:** It employs zero-shot learning to filter out noisy and sentiment-biased reviews, ensuring that only semantically consistent reviews are retained, and

  - **Hierarchical Classification:** It maps the reviews to a genre taxonomy using graph-based methods. This classification mechanism first distinguishes between Fiction and Non-Fiction, followed by further categorization into subgenres, allowing for fine-grained and multi-label genre predictions.

This dual exploration of visual and textual data—while handled independently—offers two distinct perspectives for genre classification, enhancing the reliability and depth of literary categorization in digital platforms.

2.2 Review

The task of book genre classification has been approached through three primary modalities: **Visual Methods**, **Textual Methods**, and **Multimodal Methods**. Each of these strategies leverages distinct information sources—cover images, textual descriptions, and a combination of both—to predict genre labels. This section reviews key contributions in each of these modalities.

2.2.1 Visual Method

The visual aesthetics of a book cover often provide significant cues about its genre, tone, and target audience. Leveraging these visual elements for automated genre classification has been an area of increasing interest.

Iwana et al. [1] pioneered visual-based genre classification by applying classical CNN architectures such as *AlexNet* and *LeNet*. Their study highlighted the importance of cover layout, typography, and color composition in genre prediction, demonstrating that even shallow networks could extract meaningful genre-specific patterns.

Lucieri et al. [2] advanced this concept by introducing an architecture that fuses *Inception-ResNet-V2* with attention mechanisms and a spatial transformer network. Their model preprocesses cover images, highlights salient regions, and employs dual attention branches to capture genre-defining features. This modular design mitigates high intra-class variance and low inter-class variance—challenges typical in visual genre classification. A detailed architecture is illustrated in Figure 2.3.

Maharjan et al. [3], in their work A Genre-Aware Attention Model to Improve the Likability Prediction of Books, introduced a Genre-Aware Attention Model (GA) that integrates genre-specific attention to improve book likability prediction by weighting textual and visual features with genre supervision. The input features include both textual metadata and visual elements, and the architecture supports multi-label classification to reflect multiple genre associations. Their model was evaluated across 18 genres for likability prediction. A schematic overview of their framework is illustrated in Figure 2.1.

Xu et al. [4], in their paper Panel-Page-Aware Comic Genre Understanding, proposed P2Comic, a genre classification model specifically designed for comic books. The model leverages panel-page representations as its primary input feature, processed through attention mechanisms and label correlation modules. The architecture is multi-label, capable of capturing complex genre overlaps in comic books, and introduces the first multi-genre comic book dataset for robust evaluation, covering 15 genres. Detailed architecture is illustrated in Figure 2.2.

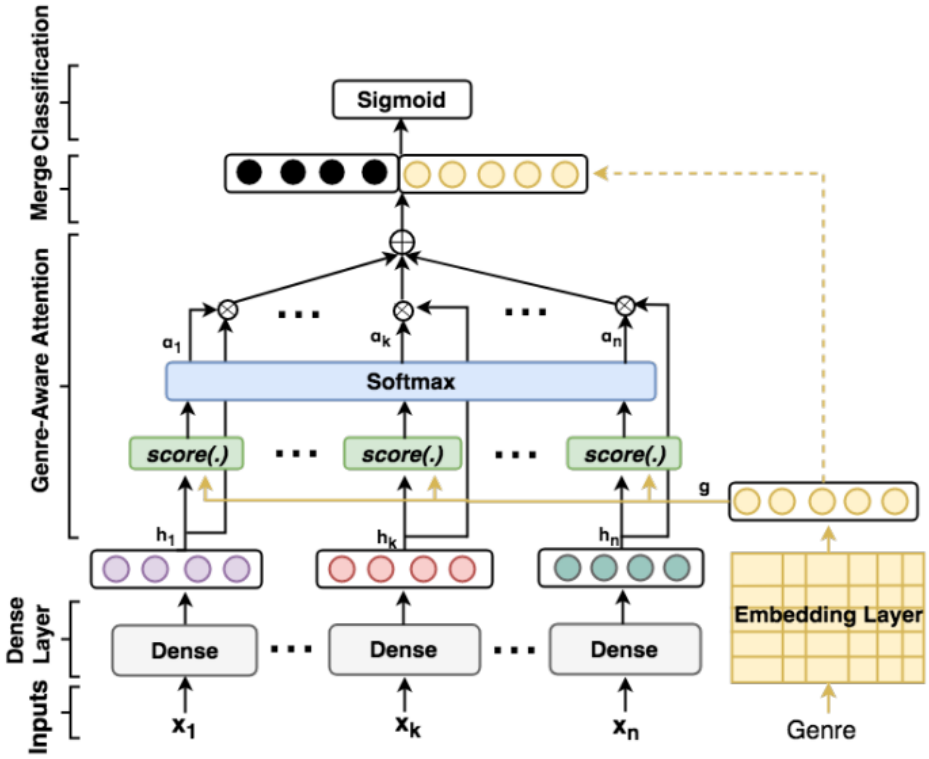


Figure 2.1: Overview of Maharjan et al. genre-aware attention model [3].

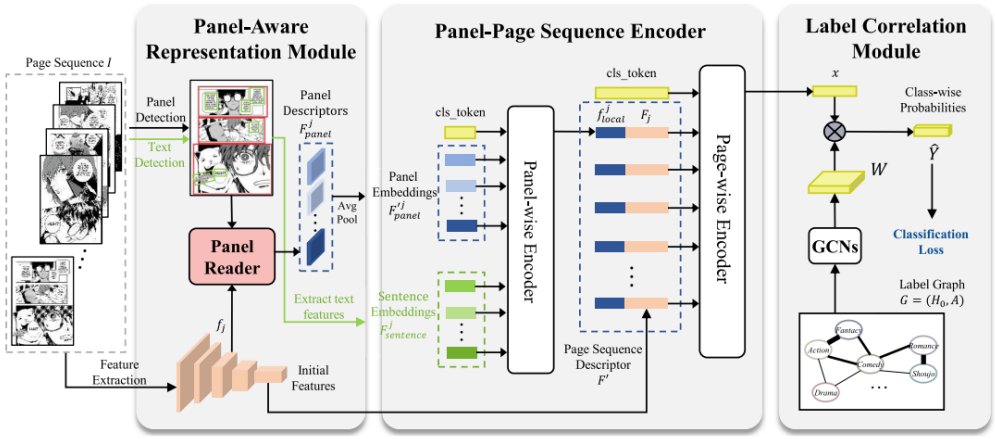


Figure 2.2: Overview of Xu et al. paper panel aware attention model [4].

Buczkowski et al. [5] explored the Goodreads dataset with VGG-inspired CNN models optimized for the structure of book covers. Their tailored network depth and convolutional filter sizes were specifically designed for the visual patterns typical of book genres, achieving significant differentiation among 14 genre categories.

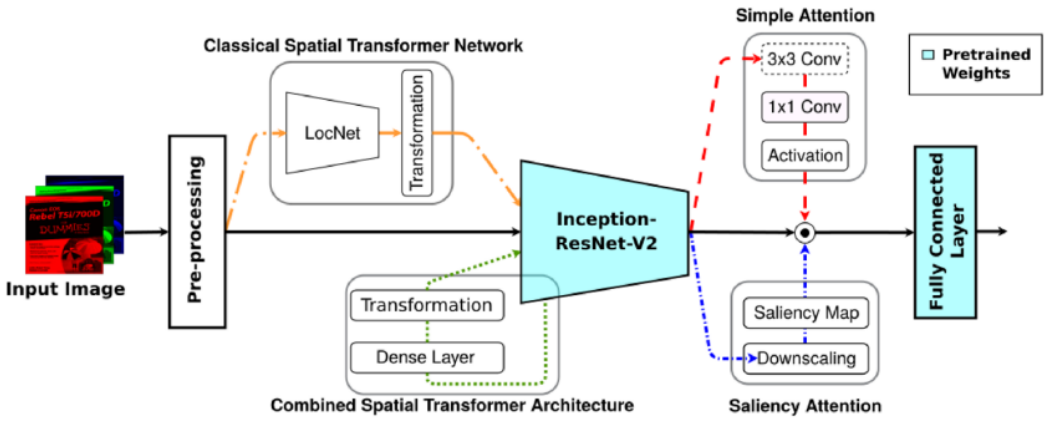


Figure 2.3: Architecture combining Inception-ResNet-V2 with spatial transformer and attention modules for book cover genre classification [2].

2.2.2 Textual Method

Textual information, such as book descriptions, blurbs, and user-generated reviews, has been widely explored for genre classification. This modality captures narrative themes, stylistic elements, and semantic features that are indicative of genre.

Ullah et al. [6], in their work Classifying Bangla Book's Context. A Multi-Label Approach, presented a novel model for classifying the contextual genres of Bangla books. Their approach leverages both textual metadata and contextual analysis to identify multiple genre labels simultaneously, reflecting the complexity and overlap inherent in literary categorization. The model employs multi-label classification techniques tailored for the Bangla language, addressing challenges specific to linguistic structure and semantic interpretation. Evaluated on a curated dataset of Bangla literature, the study underscores the effectiveness of combining contextual understanding with machine learning for genre classification in non-English corpora.

Sobkowicz et al. [7] applied Naive Bayes and Doc2Vec to book description texts from the Goodreads dataset, achieving genre classification for 14 categories. Despite its simplicity, the model demonstrated the effectiveness of unsupervised embeddings for capturing genre-related semantics.

Khalifa et al. [8] introduced a CNN-based architecture that integrates *Universal*

Sentence Encoder (USE) embeddings and readability scores. By processing fixed-length chunks of sentences through convolutional layers, the model effectively captured semantic and readability cues, boosting genre classification performance across eight genres.

Nolazco-Flores et al. [9], in their work Genre Classification of Books on Spanish, proposed a genre classification model specifically designed for Spanish-language books. The study introduced a multimodal approach that integrates both textual metadata and visual elements from book covers to enhance genre prediction accuracy. Their model was evaluated across a comprehensive dataset of Spanish literature, demonstrating robust performance in multi-label genre classification. The research highlighted the importance of language-specific adaptations in genre detection, addressing linguistic nuances and cultural context in Spanish literary works.

Worsham and Kalita [10], in their work Genre Identification and the Compositional Effect of Genre in Literature, explored the influence of compositional genre elements on literary classification. Their study introduced a model that not only classifies books by genre but also analyzes the compositional impact of multiple genre influences within a single work. Utilizing a combination of deep learning techniques and linguistic analysis, the model accounts for genre blending and hierarchical genre relationships. The architecture capturing overlapping genre associations, and primarily relies on textual features extracted from literary content for genre prediction. Evaluated on a diverse literary dataset, their findings emphasize the complexity of genre classification when literary works exhibit multi-genre characteristics.

Scofield et al. [11], in their work Book Genre Classification Based on Reviews of Portuguese-Language Literature, proposed a genre classification model leveraging textual features extracted from book reviews written in Portuguese. Their approach utilizes Long Short-Term Memory (LSTM) networks for capturing contextual information from reviews, alongside word embeddings for semantic representation. Its evaluated on a dataset of Portuguese-language literary works, the study highlights the effectiveness of reader-generated content as a rich source of genre-indicative information.

Alzetta et al. [12], in their work Tell me how you write and I'll tell you what you read: a study on the writing style of book reviews, explored genre classification through the lens of writing style analysis in book reviews. Their model leverages textual features, specifically stylistic elements such as syntax, sentiment, and lexical choices, to infer the genres of books being reviewed. The study employs Transformer-based architectures, utilizing contextual embeddings to capture nuanced writing styles that correlate with specific genres. The model is designed for multi-label classification, allowing it to associate multiple genres with a single book based on stylistic patterns observed in reviews. Evaluated on a dataset of literary critiques, the research demonstrates the link between review writing styles and book genres.

Ng et al. [13] employed a hybrid RNN-GRU model to classify books into 31 genres using a combination of descriptions and metadata. This approach modeled temporal dependencies in text, outperforming traditional static embeddings.

Saraswat et al. [14] proposed a review-driven pipeline that utilizes LSTM-based RNNs for genre classification and book recommendation. Raw reviews are preprocessed and transformed into word embeddings, which are fed into the RNN for sequential sentiment and thematic analysis. This method proved effective for capturing reader perceptions and latent themes, although it struggled with sparsity for books with limited reviews.

2.2.3 Multimodal Method

Multimodal learning seeks to combine visual and textual information to enhance the robustness and accuracy of genre classification. These methods capitalize on complementary signals from both book covers and textual descriptions.

Kundu et al. [15] introduced a deep multimodal neural network architecture for genre classification. Visual features were extracted using a *ResNet-50* backbone pretrained on ImageNet, while textual features were derived using the *Universal Sentence Encoder (USE)*. These features were concatenated and passed to a softmax classifier for final genre prediction, demonstrating substantial improvements over unimodal baselines.

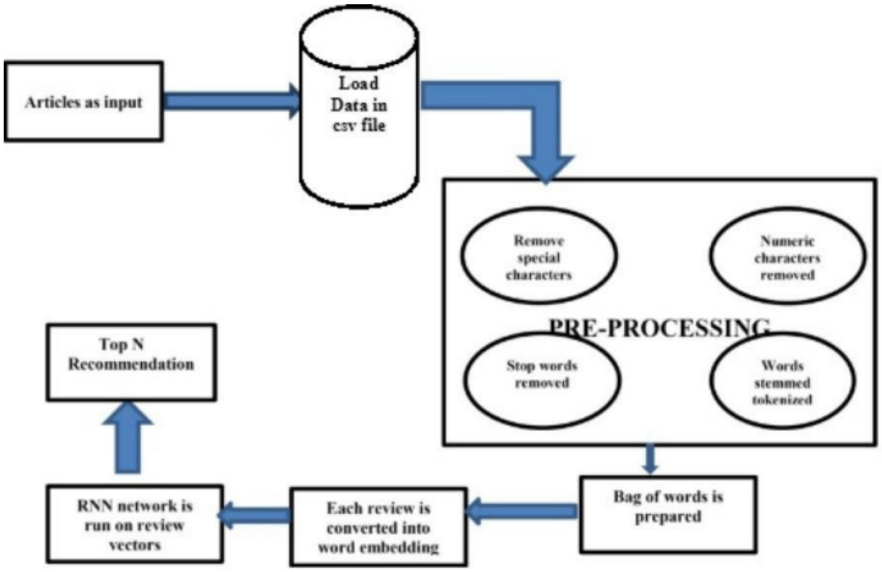


Figure 2.4: Review-driven genre classification and recommendation framework of Saraswat et al. [14].

Biradar et al. [16] adopted a late-fusion strategy that combined visual features from *XceptionNet* and textual features from *GloVe* embeddings. Their model utilized multinomial logistic regression to classify books into five genres, highlighting the efficacy of late fusion for genre separation.

Rasheed et al. [17] proposed an attention-based multimodal framework that integrated a modified *SE-ResNeXt-101* for cover images with a novel textual model called *EXAN*. Their framework dynamically balanced the contributions of visual and textual modalities using attention mechanisms, achieving state-of-the-art results on the BookCover28 dataset and a custom Arabic dataset.

[18] further explored multimodal fusion by combining features from *Inception-v3* and Naive Bayes-based text classifiers. Both early and late fusion strategies were evaluated, revealing that multimodal integration led to improved consistency across 30 genres.

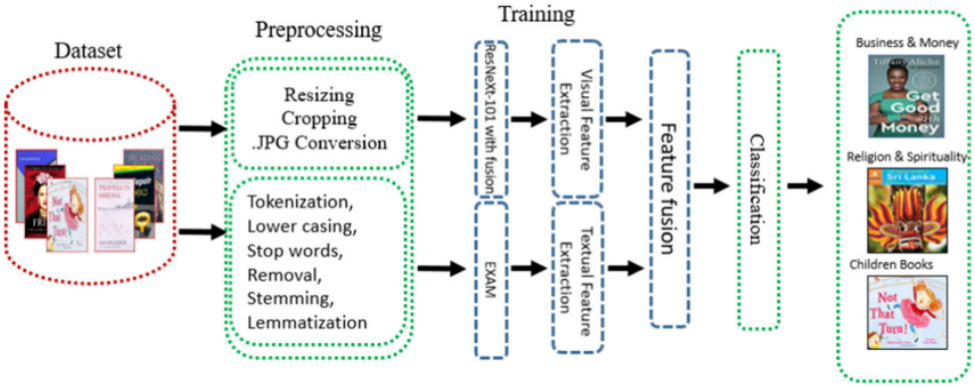


Figure 2.5: Rasheed et al.'s multimodal framework combining SE-ResNeXt-101 and EXAN with attention-based fusion [17].

2.3 Limitations of Existing Work

- **Limited Research on Book Genre Identification:** Existing studies on book recommendation systems focus mainly on user behavior or text-based metadata, with minimal attention given to accurately identifying genres based on book features like cover images.

- **Outdated Datasets:** Many commonly used datasets in this domain contain outdated cover images, which may not reflect current design trends, genre conventions, or reader expectations. This limits the relevance and applicability of models trained on such data to modern publishing scenarios.

- **Single-Label Constraints:** Several studies, including **Iwana et al.** [1] and **Lucieri et al.** [2], employ **single-label classification**, where each book is assigned only one primary genre. This oversimplifies the genre representation, as many books naturally belong to multiple genres (e.g., *Science Fiction Thriller* or *Historical Romance*), limiting the model's expressiveness.

- **Lack of Multimodal Integration:** Some works, such as **Sobkowicz et al.** [7] and **Worsham and Kalita** [10], rely purely on **textual metadata** or **linguistic analysis**, overlooking the rich semantic information that visual features (like book covers) can provide. Conversely, models like **Iwana et al.** [1] and

Lucieri et al. [2] focus heavily on visual features, ignoring textual content. This compartmentalized approach misses out on the enhanced performance that **multimodal fusion** could achieve.

- **Lack of Hierarchical Genre Classification:** Most current approaches do not consider a hierarchical structure for genres. A hierarchical classification system (e.g., Fiction/Nonfiction and sub-genres) is missing, which could provide more detailed and organized genre categorization.

- **Language and Regional Limitations** Many studies focus on English-language books (**Worsham and Kalita** [10], **Iwana et al.** [1]), with limited exploration of non-English literature. While **Ullah et al.** [6] and **Scofield et al.** [11] address Bangla and Portuguese literature, other major languages remain underrepresented, resulting in limited generalizability across global literary contexts.

- **Inadequate Results from Existing Methods:** Many existing methods fall short in achieving high accuracy and relevance, indicating a need for improved models and approaches.

Chapter 3

Dataset Creation

3.1 Dataset details

This study leverages a publicly available book dataset from Kaggle as the foundational resource for constructing our custom dataset. The Kaggle dataset contains detailed records of user interactions with books, which include metadata about the books, user-generated ratings, and user profiles.

- **User Information:** Consists of 2.78 lakh unique users.
- **Book Information:** Contains metadata for 2.71 lakh unique books.
- **Valid Ratings:** Includes approximately 11.49 lakh valid user-generated ratings.

Table 3.1: Dataset Summary

Attributes	**Kaggle Book Dataset**
Number of Users	2.78 Lakhs
Number of Books	2.71 Lakhs
Number of Valid Ratings	11.49 Lakhs

3.1.1 Dataset Metadata

The Kaggle dataset is organized into two primary metadata files:

- **Book Metadata:** Contains key information about each book, including:

  - ISBN: Unique identifier for each book.
  - Book-title: The title of the book.
  - Book-author: The author's name.
  - YOP (Year of Publication): The year the book was published.
  - Publisher: The name of the publishing house.

- **Rating Metadata:** Includes user-generated ratings, structured as follows:
  - User-ID: Unique identifier for each user.
  - ISBN: Corresponding ISBN for the rated book.
  - Book Rating: The rating given by the user.

- **User Metadata:** Includes user-generated ratings, structured as follows:
  - User-ID: Unique identifier for each user.
  - Location: Location of the reader.
  - Age: Age of the reader.

Table 3.2: Book Metadata (books.csv)

ISBN	**Book-title**	**Book-author**	**YOP**	**Publisher**
195153448	Classical Mythology	Ann Beattie	2002	Scribner
0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo
0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial

Table 3.3: Rating Metadata (ratings.csv)

User-ID	**ISBN**	**Book Rating**
276725	034545104X	0
276726	0155061224	5
276727	0446520802	0

Table 3.4: User Metadata (users.csv)

User-ID	Location	Age
1	nyc, new york, usa	NULL
2	stockton, california, usa	18
3	Moscow, Yukon Territory, Russia	NULL

3.2 Dataset Creation Process

The dataset creation process is structured to ensure that each book entry is enriched with critical components, including cover images, blurbs, and multiple user reviews. To achieve this, we utilized **web scraping techniques** to extract data from **Goodreads.com**, a well-known literary database. This approach enabled us to gather comprehensive and diverse book-related information directly from the source. To maintain high data quality and reliability, additional steps for annotation and validation were also performed.

3.2.1 Cover Page Image Collection

Since the Kaggle dataset does not inherently provide book cover images, we extended our dataset by scraping high-resolution cover images for each book from Goodreads.com. The ISBN was used as the primary key for querying and fetching the images. Images were collected, resized to a standard resolution of 224×224 pixels, and stored in a structured format. This step ensured uniformity in image input for visual-based genre classification models.

3.2.2 Blurb Collection

Blurbs, which serve as brief descriptions or summaries of the book, were collected for each sample using web scraping methods from Goodreads.com. These blurbs were extracted in a clean and structured format. To ensure data consistency, all text was preprocessed by:

- Converting to lowercase for uniformity.

- Removing special characters and excessive whitespace.
- Normalizing punctuation and encoding formats.

The collected blurbs act as primary textual signals for genre inference in our experiments.

3.2.3 User Reviews Collection

To capture diverse reader perceptions, we collected up to **10 user reviews** per book through web scraping from Goodreads.com. These reviews were selected based on their relevance and completeness. If fewer than 10 reviews were available, all existing reviews were used. Text preprocessing included:

- Tokenization and removal of stop words.
- Lemmatization to standardize word forms.
- Elimination of URLs, emojis, and non-alphanumeric characters.

This structured collection of reviews enriched our dataset, enabling fine-grained sentiment and genre analysis.

3.2.4 Data Annotation and Validation

To ensure the reliability of genre labels, a multi-step annotation and validation process was performed:

- **Manual Genre Labeling:** Ambiguous or missing genre tags were manually inspected and annotated based on blurbs and user reviews.
- **Cross-Validation with External Sources:** Collected genre information was cross-checked with reliable literary databases to ensure accuracy.
- **Multi-Label Annotation:** For books spanning multiple genres, multi-label annotations were applied to capture the thematic diversity.

This rigorous annotation process minimized labeling noise and prepared the dataset for effective model training and evaluation.

3.2.5 Final Dataset Statistics

The finalized dataset consisted of:

- **27,953** books with corresponding cover images.
- **Blurbs** for descriptive context.
- Up to **10 user reviews** per book for sentiment and genre analysis.
- **Validated genre labels** for high-quality, structured input.

The entire dataset creation was conducted using automated web scraping techniques from Goodreads.com, ensuring comprehensive coverage of metadata and user-generated content.

3.3 Challenges in Dataset

- **Non-English Blurbs and Reviews:** During the web scraping process, a portion of book blurbs and reviews were found to be written in non-English languages. Since the intended model is trained primarily on English text, these multilingual entries introduce noise and inconsistencies in the textual data, leading to reduced effectiveness in semantic understanding and classification accuracy. The distribution of non-English blurbs and reviews is shown in Table 3.5.
- **Absence of Blurb Text:** Blurbs provide structured and curated descriptions of books, offering valuable insights into their themes, narrative style, and intended audience. They serve as a condensed summary that is crucial for genre classification. The lack of blurbs forces the model to rely solely on unstructured text, increasing the difficulty of capturing genre-specific signals accurately.

Table 3.5: Non-English Language Distribution in Descriptions and Reviews

Descriptions (Language)	**Counts**	**Reviews (Language)**	**Counts**
German	109	Dutch	2
French	108	Spanish	44
Spanish	128	Italian	6
Greek	1	Portuguese	7
Portuguese	8	Arabic	17
Maltese	1	Indonesian	5
Italian	14	French	4
Japanese	2	Turkish	4
Russian	1	Romanian	2
Ukrainian	2	German	3
Indonesian	1	Greek	1
Sinhala	1	Persian	5
Korean	1	Russian	3
Turkish	3	Persian (Farsi)	1
Serbian	1	-	-
Japanese	1	-	-
Others	5	-	-

- **Genre Imbalance:** The dataset exhibits a significant imbalance in genre distribution after hierarchical labeling. This skew can bias the learning model, causing it to underperform on less-represented genres. The genre imbalance is illustrated in Figure 3.1 and Figure 3.2.

- **Lack of User-Generated Reviews:** User reviews capture readers' perceptions, thematic observations, and emotional responses, adding semantic layers that are not evident from metadata alone. The absence of such reviews reduces the contextual understanding of the book's content, making genre inference less precise.

- **Absence of Hierarchical Labeling:** Genre classification is inherently hierarchical, with broader categories like *Fiction* and *Non-Fiction* containing multiple subgenres. Without hierarchical labeling, the model is limited to flat classification, ignoring the multi-layered genre structure. This oversimplification results

in less granular and less meaningful genre predictions.

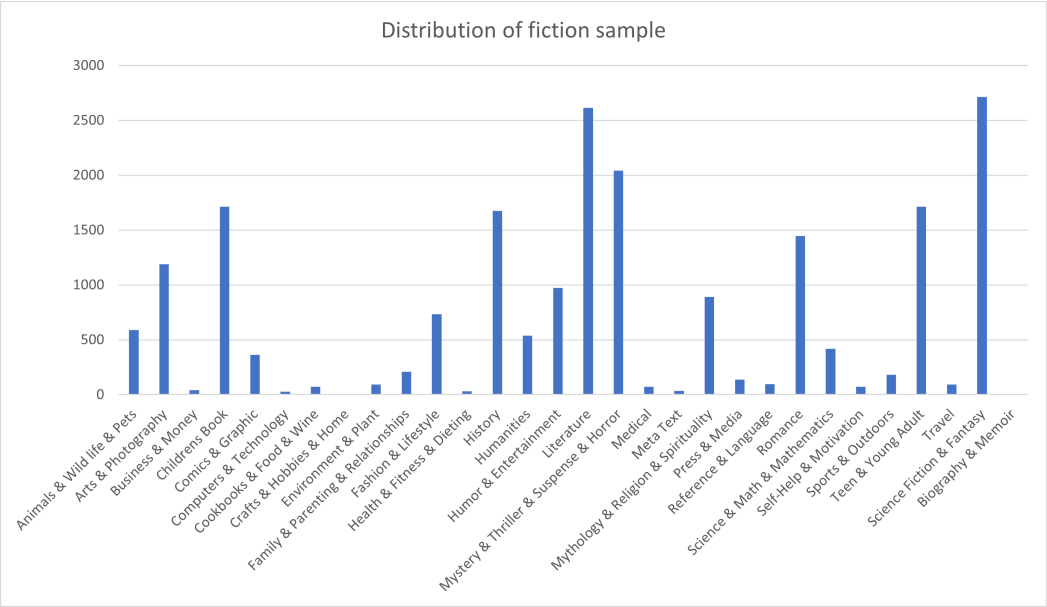


Figure 3.1: Genre Imbalance in Fiction Class

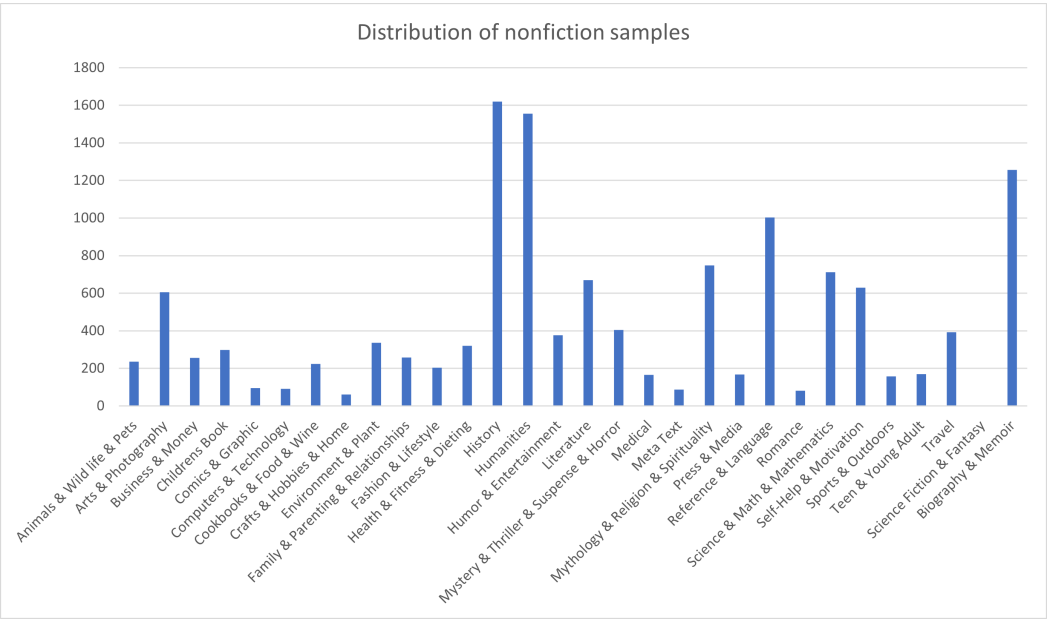


Figure 3.2: Genre Imbalance in Nonfiction Class

3.3.1 Mitigation of Dataset Issues

To address the issues identified in the curated dataset, we employed a combination of data augmentation techniques and hierarchical labeling strategies. These steps

enhanced the dataset's quality, diversity, and usability for genre classification tasks.

3.3.1.1 Cover Page Image Augmentation:

One of the core challenges in genre classification using book cover images is the limited visual diversity in certain genre categories, especially those with fewer samples. To mitigate this, we applied data augmentation using **('Stable-diffusion-xl-refiner-1.0')**. [19]

These models were used to generate novel cover page images that preserved the visual semantics of their respective genres while introducing variability in style, layout, and background. This not only improved class balance but also increased the robustness of the vision model in learning genre-specific visual patterns. Image augmentation using the diffusion model is illustrated in Figure 3.3.

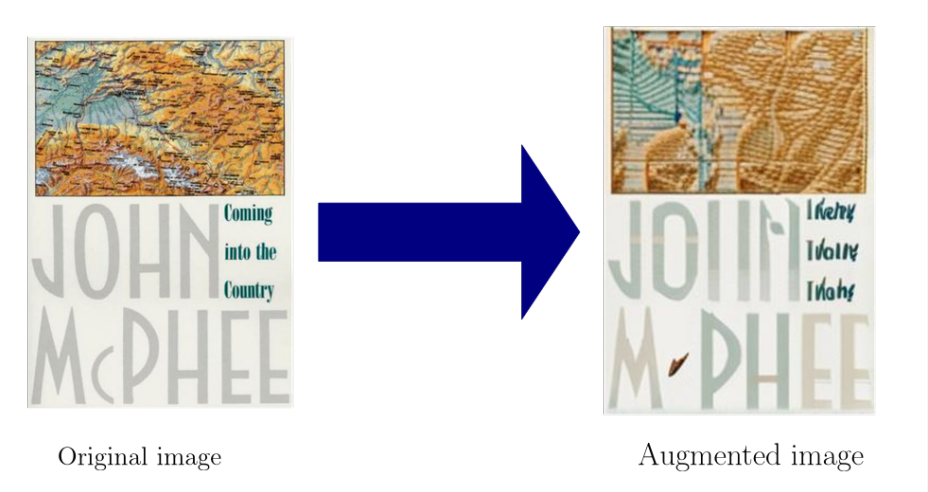


Figure 3.3: Image Augmentation using Diffusion Model

3.3.1.2 Description Augmentation:

To enhance the textual modality, we addressed the limitations posed by non-English and short or incomplete descriptions. For this, we utilized the **Gemini** API [20], a large language model API, to perform text augmentation.

The Gemini API was used to paraphrase and extend existing English descriptions while generating meaningful and coherent content for missing or low-quality descrip-

tions. The augmented descriptions retained the core semantic meaning of the original content while improving linguistic consistency and diversity. Description augmentation using the **Gemini** API [20] is illustrated in Figure 3.4.

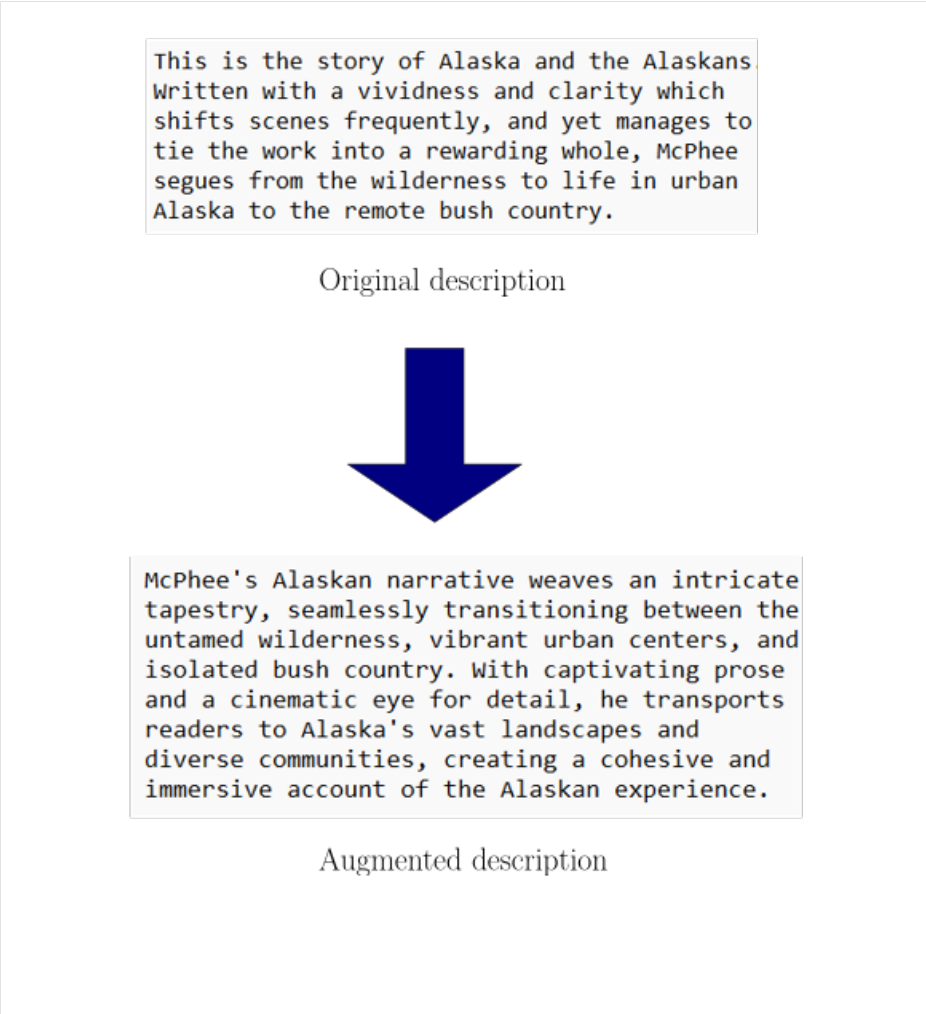


Figure 3.4: Description Augmentation using **Gemini** [20]

3.3.1.3 Hierarchical Labeling of Genres:

Given the complexity and overlapping nature of book genres, a **hierarchical labeling approach** was employed to structure the genre classification in a scalable and semantically meaningful manner.

- **Identifying Unique Genres**: From the scraped dataset, over **2,055 unique genre tags** were identified. These genres included a wide range of descriptors,

many of which were redundant, overly specific, or inconsistently labeled. The presence of niche categories and overlapping terminologies made it challenging to directly utilize these tags for model training or evaluation.

- **Genre Consolidation and Selection**: To simplify and standardize the classification process, the long tail of infrequent or ambiguous genres was pruned. A curated list of **30 primary sub-genres** was finalized by analyzing genre frequency, semantic clarity, and practical relevance. This step ensured better balance, cleaner annotations, and more meaningful categorization for downstream tasks.

- **Two-Level Genre Structure**: The final hierarchy was designed with **two levels**:

  - **Level 1**: A binary classification into **Fiction** and **Nonfiction**, offering a coarse-grained division aligned with standard literary taxonomy.

  - **Level 2**: Each of these main categories was further classified into **30** refined sub-genres. This structure enables both high-level abstraction and granular analysis, thereby enhancing classification accuracy and interpretability in both training and inference stages.

Despite applying various augmentation techniques such as diffusion-based image generation and language model-driven description synthesis, the dataset continues to exhibit noticeable imbalance. This residual skew is primarily due to overlapping genre associations—where books naturally span multiple thematic categories. For instance, titles tagged with both *Romance* and *Historical Fiction*, or *Science Fiction* and *Biography*, introduce classification ambiguity. Such genre co-occurrences lead to an uneven sample distribution across mutually exclusive genre labels, which in turn complicates the training of a robust and unbiased genre prediction model.

3.4 Dataset Analysis

3.4.1 Hierarchical Genre-wise Count of this Dataset:

The hierarchical genre-wise distribution of our curated dataset is summarized in Table 3.6. The dataset was categorized under two primary classes—Fiction and Nonfiction—and further subdivided into 30 fine-grained genres. Each genre is associated with four distinct counts: the number of samples before and after augmentation for both Fiction and Nonfiction categories.

Table 3.6: Genrewise count of samples before and after augmentation

Class Id	Genre Label	$\mathcal{F}_{BA}$	$\mathcal{F}_{AA}$	$\mathcal{NF}_{BA}$	$\mathcal{NF}_{AA}$
1	Animals & Wildlife & Pets	590	1260	235	924
2	Arts & Photography	1188	2566	606	1782
3	Business & Money	42	624	256	760
4	Childrens Book	1714	3281	298	1080
5	Comics & Graphic	364	791	96	1002
6	Computers & Technology	27	596	92	595
7	Cookbooks & Food & Wine	72	808	223	608
8	Crafts & Hobbies & Home	4	520	61	541
9	Environment & Plant	92	905	337	1028
10	Family & Parenting & Relationships	208	800	257	1089
11	Fashion & Lifestyle	732	1426	204	1100
12	Health & Fitness & Dieting	32	710	320	1171
13	History	1677	3123	1619	4095
14	Humanities	537	1481	1555	4223
15	Humor & Entertainment	972	2009	376	1360
16	Literature	2615	6088	670	1786
17	Mystery & Thriller & Suspense & Horror	2043	4289	404	920
18	Medical	70	950	165	1080
19	Meta Text	35	774	88	954
20	Mythology & Religion & Spirituality	890	1909	748	1376
21	Press & Media	138	486	167	903
22	Reference & Language	97	694	1003	3077
23	Romance	1445	2275	80	1065
24	Science & Math	419	718	712	2221
25	Self-help & Motivation	72	993	630	1865
26	Sports & Outdoors	183	825	157	522
27	Teen & Young Adult	1712	2524	170	817
28	Travel	92	743	393	1090
29	Science Fiction & Fantasy	2715	5034	0	0
30	Biographies & Memoir	0	0	1256	3466

3.4.2 Statistical Analysis of Cover Page Images:

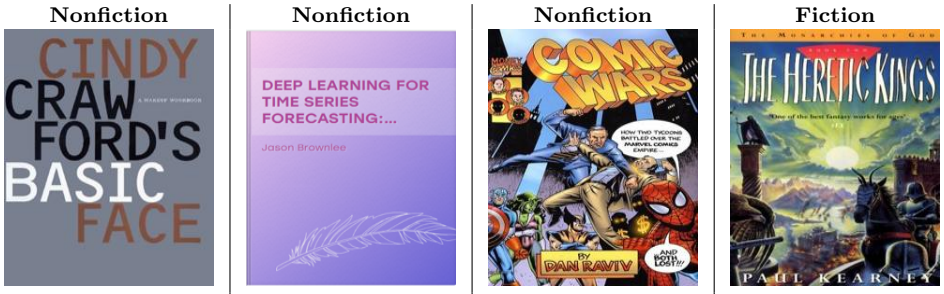
We analyze the distribution of cover page images based on their pixel area. This analysis helps in understanding the variation in image resolutions across the dataset and offer insights into how augmentation enhances diversity and uniformity in visual input.

Table 3.7: Image area-wise distribution before and after augmentation

Image area range (10000 pixel^2)	Image count before augmentation	Image count after augmentation
< 5	20	50
$5 - 10$	264	9847
$10 - 15$	7057	11124
$15 - 20$	2205	3906
$20 - 25$	278	584
$25 - 30$	125	269
$30 - 35$	52	87
$35 - 40$	41	63
> 40	1298	2091

3.4.3 Characteristics of Visual Dataset:

The visual diversity of book cover images in our dataset reflects a wide range of design styles, thematic elements, and genre-specific cues. These variations are significant in influencing the performance of visual classification models. Some covers are minimalistic with sparse textual or graphical content, while others are busy. Additionally, there exists substantial variation both across and within genres—some genres share overlapping visual traits, and books within the same genre can differ greatly in visual presentation. The following figures 3.5 showcase representative samples that illustrate these nuanced visual properties, helping to better contextualize the complexity of image-based genre classification.



(a) Computers & Technology, — — —

(b) Fashion & Lifestyle, – – —

(c) Business & Money, Comics & Graphic, History

(d) Literature, Science Fiction & Fantasy – –

The illustrative samples presented in Figure 3.5 reveal several critical challenges associated with genre classification from book cover images. Subfigures (a) and (b)

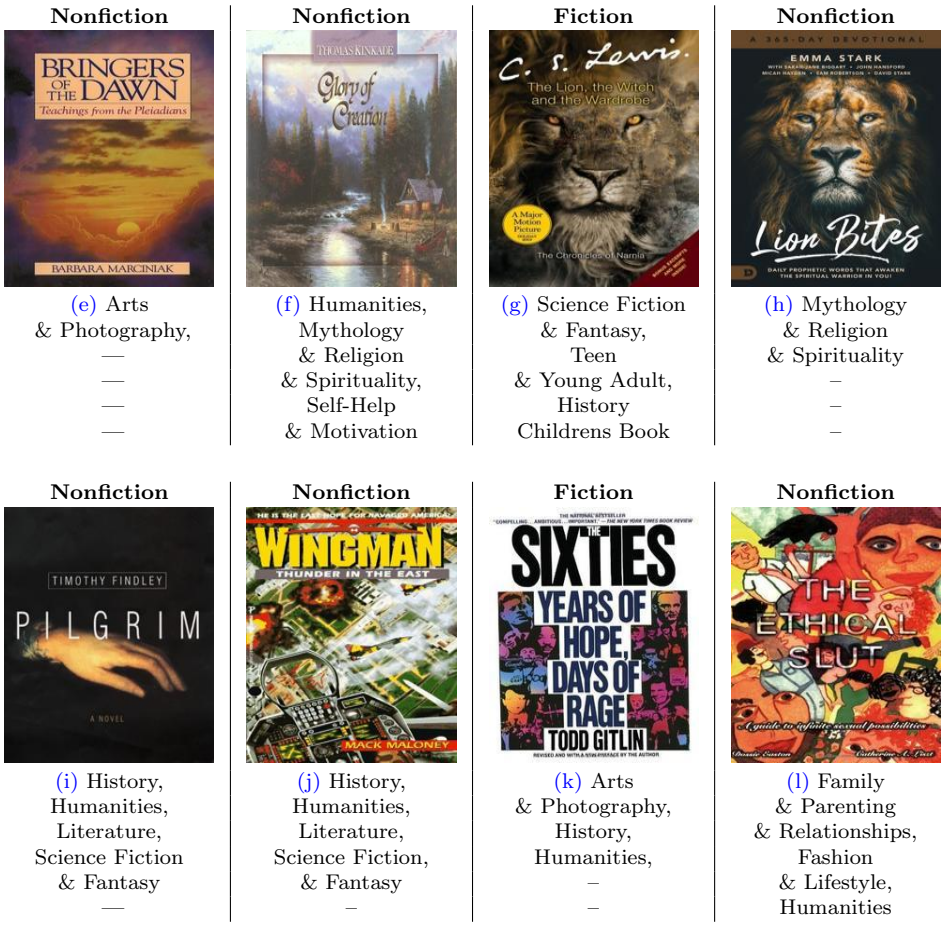


Figure 3.5: Examples of coverpage challenges: (a, b) less information, (c, d) complex background, (e, f) scene complexity, (g, h) inter-genre variation, (i, j) intra-variation, (k, l) collage

depict minimalist covers containing very limited visual information. The absence of explicit graphical or textual cues in such designs makes it particularly difficult for a model to extract genre-relevant features. In contrast, subfigures (c) and (d) exhibit visually complex compositions, which, while richer in information, may introduce noise and ambiguity due to cluttered design elements.

Subfigures (e) and (f) represent scene-based illustrations, where the cover portrays a visual narrative or setting. While such images may offer semantic cues, the interpretation of scenes can be highly subjective and genre overlap is common. Subfigures (g) and (h) illustrate instances of inter-genre similarity, where visually similar covers belong to entirely different genres. This visual convergence across genres introduces confusion in model predictions. Conversely, subfigures (i) and (j) demonstrate intra-

genre variability, where books of the same genre exhibit vastly different visual styles, reducing the consistency of genre-specific patterns.

Finally, subfigures (k) and (l) present collage-style covers that combine multiple graphical elements or themes. While these may contain rich content, the lack of focus or the heterogeneous nature of elements complicates the extraction of coherent genre indicators.

These diverse visual representations underscore the inherent challenges in cover-based genre classification and highlight the need for sophisticated feature extractors capable of capturing both subtle patterns and abstract semantics.

Chapter 4

Book Genre Classification from Reliable Sources

4.1 Problem Formulation

The task of book genre classification from cover page images is a challenging problem that aims to predict the literary genre of a book solely based on its visual representation. Book covers are meticulously designed to convey essential information about the book's genre, narrative style, and intended audience through visual cues such as color schemes, typography, imagery, and layout. These design elements are often aligned with genre conventions—thrillers may feature dark tones and bold fonts, while romance novels might employ softer color palettes and delicate typography. The primary challenge is to effectively leverage these cues to infer genre labels without relying on textual metadata like blurbs or author descriptions.

4.1.1 Problem Definition

We define the problem as a multi-label hierarchical classification task, where each book is categorized into a genre taxonomy structured in two levels:

- **Level 1:** Broad genre categories such as *Fiction* and *Non-Fiction*.
- **Level 2:** Fine-grained subgenres that are nested under the main categories. For example, *Fiction* may include subgenres like *Mystery, Fantasy, Romance*, and

Science Fiction, while *Non-Fiction* may contain *Biography, History, Self-Help*, and *Philosophy*.

4.1.2 Mathematical Representation

Given:

- A dataset of n books represented as:

$$B = \{B_1, B_2, B_3, \ldots, B_n\}$$

  where each book B_i is uniquely identified and characterized by its cover image I_i.

- Each book cover image I_i is represented as a three-dimensional tensor:

$$I_i \in \mathbb{R}^{H \times W \times 3}$$

  where H and W denote the height and width of the image, respectively, and 3 represents the RGB color channels.

- A hierarchical genre label structure L, composed of two primary branches:

$$L = (\{0\} \times L_f) \cup (\{1\} \times L_{nf})$$

  Here:

  - $L_f = \{l_1, l_2, \ldots, l_k\}$ represents the set of subgenres under the **Fiction** category.
  - $L_{nf} = \{l_1, l_2, \ldots, l_m\}$ represents the set of subgenres under the **Non-Fiction** category.
  - $\{0\}$ and $\{1\}$ are binary indicators representing Fiction and Non-Fiction, respectively.

4.1.3 Objective

The main objective is to learn a mapping function f:

$$f : I_i \rightarrow L$$

such that each cover image I_i is accurately mapped to its corresponding hierarchical genre label in L. This mapping is performed in two sequential stages:

1. **Level-1 Classification:** Predict whether the book is Fiction (0) or Non-Fiction (1).

2. **Level-2 Classification:** Given the output of Level-1, classify the book into its specific subgenre within the identified category.

4.2 Proposed Method

In this section, we will discuss about feature extractor, followed by the level 1 binary classifier, and the level 2 multi-label classifier. An overview of the complete architecture is illustrated in Figure 4.1.

4.2.1 Feature Extractor: Swin Transformer

To extract rich and hierarchical visual features from book cover images, we employ the **Swin Transformer** [21] as the backbone of our network. Introduced by Liu et al. (2021), the Swin Transformer is a hierarchical vision transformer that leverages a novel *shifted window mechanism* for efficient and scalable feature representation.

4.2.1.1 Key Characteristics

- **Hierarchical Architecture:** The Swin Transformer processes the input image in four stages, where the spatial resolution is progressively reduced while the feature dimension increases. This design, analogous to conventional CNNs, facilitates multi-scale representation learning.

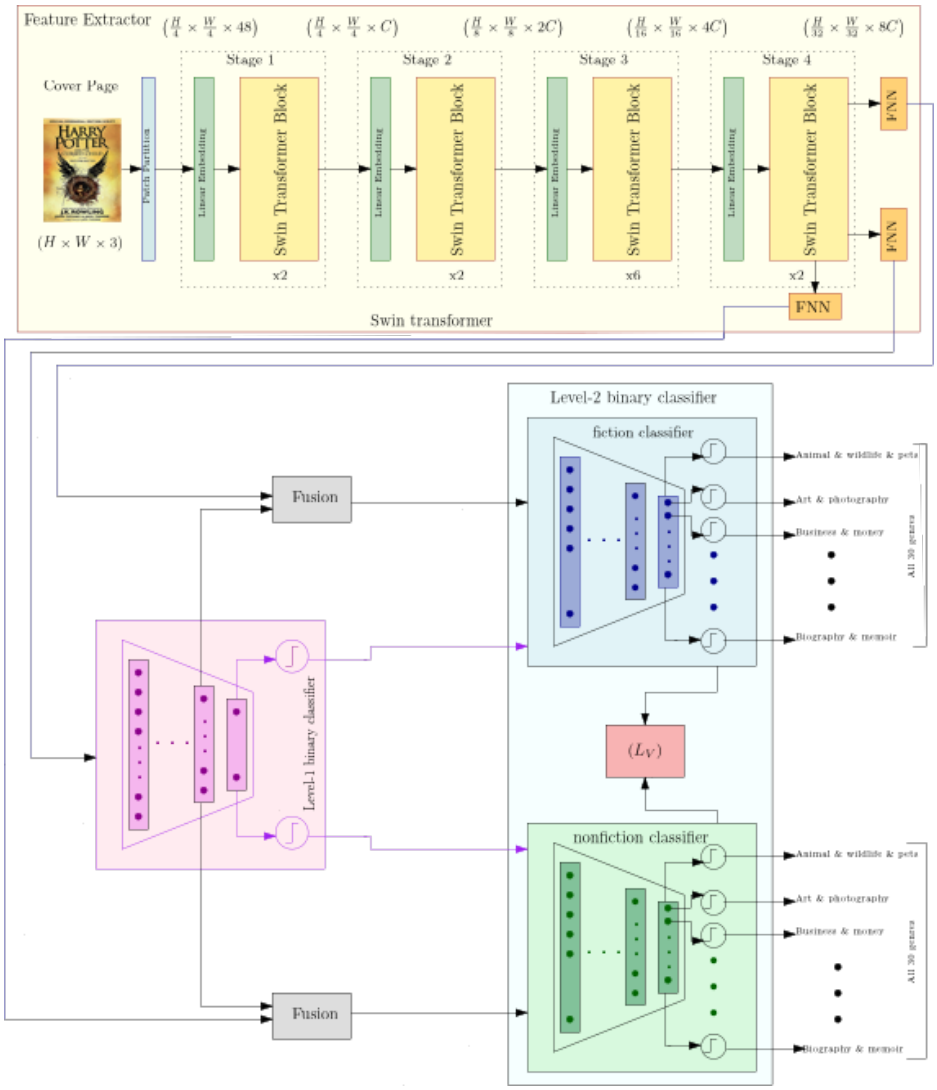


Figure 4.1: Proposed Vision Architecture

- **Shifted Window Self-Attention:** Unlike vanilla Vision Transformers that apply global self-attention, the Swin Transformer restricts self-attention computation within non-overlapping local windows. In alternating layers, windows are shifted by a predefined offset to enable cross-window connections. This approach maintains linear computational complexity with respect to image size while enhancing local-global context integration.

4.2.1.2 Patch Partitioning and Embedding

The input image of size $H \times W \times 3$ is first partitioned into non-overlapping patches of size 4×4. These patches are flattened and linearly projected into an embedding space of dimension 96, resulting in an initial feature map of size $\frac{H}{4} \times \frac{W}{4} \times C$, where $C = 96$.

4.2.1.3 Stage-wise Breakdown

Each stage of the Swin Transformer consists of several Swin Transformer blocks, followed by a patch merging layer that reduces spatial dimensions and increases channel depth:

- **Stage 1:**
  - Input resolution: $\frac{H}{4} \times \frac{W}{4} \times C$
  - Contains 2 Swin Transformer blocks.
- **Stage 2:**
  - Resolution: $\frac{H}{8} \times \frac{W}{8} \times 2C$
  - Contains 2 blocks.
- **Stage 3:**
  - Resolution: $\frac{H}{16} \times \frac{W}{16} \times 4C$
  - Contains 6 blocks, offering deeper contextual modeling.
- **Stage 4:**
  - Resolution: $\frac{H}{32} \times \frac{W}{32} \times 8C$
  - Contains 2 blocks.

Each block incorporates:

- Layer Normalization (LN)

- Multi-Head Self-Attention (MSA) within windows
- A Feed Forward Network (FFN) with GELU activation
- Residual connections for stable and efficient learning

The final feature representation output from Stage 4 encapsulates high-level semantic information and serves as the input to the subsequent classification modules. An overview of the detailed Swin Transformer architecture is illustrated in Figure 4.2 below.

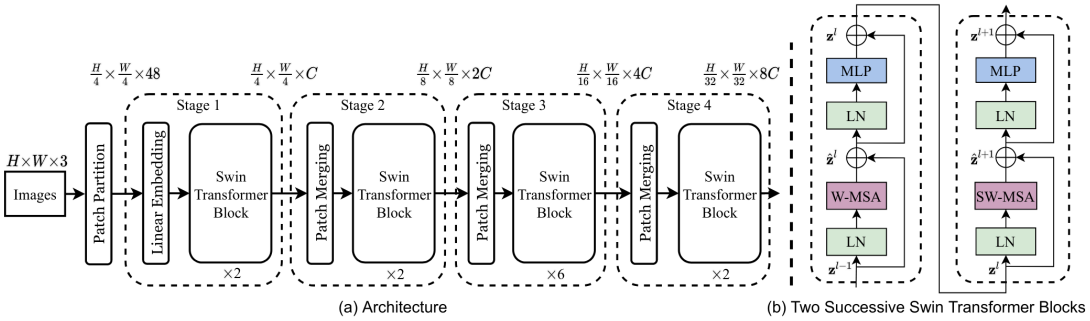


Figure 4.2: Swin Transformer Architecture [21]

4.2.2 Level 1 Classifier: Binary Genre Classifier

This level determines whether the input is **Fiction (1)** or **Nonfiction (0)**, and uses the Swin Transformer [21] extracted feature as an input feature.

4.2.2.1 Architecture:

- A stack of **four fully connected layers**:

$$\mathbf{h}^{(1)} = \text{ReLU}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$$
$$\mathbf{h}^{(2)} = \text{ReLU}(\mathbf{W}_2\mathbf{h}^{(1)} + \mathbf{b}_2)$$
$$\mathbf{h}^{(3)} = \text{ReLU}(\mathbf{W}_3\mathbf{h}^{(2)} + \mathbf{b}_3)$$
$$\hat{y} = \sigma(\mathbf{W}_4\mathbf{h}^{(3)} + \mathbf{b}_4)$$

Where:

- $\mathbf{W}_i$, $\mathbf{b}_i$ are weights and biases of each layer.
- σ is the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

4.2.2.2 Loss Function: Binary Cross-Entropy (BCE)

$$\mathcal{L}_{\text{BCE}} = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]$$

Where:

- $y \in \{0, 1\}$ is the true class label.
- $\hat{y} \in (0, 1)$ is the predicted probability.

4.2.3 Level 2 Classifiers: Multi-label Subgenre Prediction

Based on the output of Level 1:

- If $\hat{y} > 0.5$: pass to **Fiction Classifier**
- Else: pass to **Nonfiction Classifier**

Each Level 2 classifier is a **multi-label** prediction head for 30 genre tags.

4.2.3.1 Architecture:

Same 4-layer FCNN as above, ending in a **30-dimensional sigmoid output**:

$$\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{30}] = \sigma(\mathbf{W}_4 \mathbf{h}^{(3)} + \mathbf{b}_4)$$

Where each $\hat{y}_i \in (0, 1)$ represents the predicted probability of the i^{th} genre.

4.2.3.2 Loss Function: Asymmetric Loss for Multi-label Classification

To address **label imbalance** and **label sparsity**, we use the **Asymmetric Loss** introduced by Ridnik et al. (ASL) [22]:

$$\mathcal{L}_{ASL} = -\frac{1}{C}\sum_{i=1}^{C}[y_i \cdot (1-\hat{y}_i)^{\gamma_+} \cdot \log(\hat{y}_i) + (1-y_i) \cdot \hat{y}_i^{\gamma_-} \cdot \log(1-\hat{y}_i)]$$

Where:

- $C = 30$ is the number of genres,
- $y_i \in \{0, 1\}$ is the true label for the i^{th} class,
- $\hat{y}_i \in (0, 1)$ is the predicted probability,
- $\gamma_+ > \gamma_-$ (commonly $\gamma_+ = 2$, $\gamma_- = 1$) control suppression of easy negatives and enhancement of hard positives.

This asymmetric formulation penalizes **false negatives more** than false positives — helping mitigate the class imbalance issue typical in multi-label setups.

4.3 Experimental Analysis

Our experiments were executed using Pytorch 2.1.0, having Python 3.10.14 on an Ubuntu 20.04.4 LTS. The hardware setup included Intel(R) Xeon(R) W-1270 clocked at 3.40GHz, with 16 CPU cores and 128 GB of RAM. Additionally, the machine was equipped with a 24 GB NVIDIA RTX A5000 GPU. Table 5.1 provide detailed experimental settings.

4.3.1 Evaluation Metrics

To evaluate the performance of our hierarchical multi-label classification model, we employ a comprehensive set of metrics that effectively capture both binary and multi-label classification performance. The following metrics are used:

Table 4.1: Experimental setup details

Level-1 and Level-2 classification settings	
No. of epochs (Level-1 classification)	50
No. of epochs (Level-2 classification)	100
Batch size	16
Learning rate	10^{-5}
Weight decay	10^{-2}
Exponential decay rates	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Zero-denominator avoidance parameter	10^{-8}
Optimizer	AdamW
Patience	5

4.3.1.1 Accuracy

Accuracy measures the proportion of correct predictions over the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

4.3.1.2 Precision

Precision is the proportion of correctly predicted positive observations to the total predicted positives. It is computed in micro, macro, and weighted forms:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

4.3.1.3 Recall

Recall is the proportion of correctly predicted positive observations to all actual positives:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

4.3.1.4 Specificity

Specificity, also known as the True Negative Rate, measures the proportion of actual negatives correctly identified as such. It complements recall:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.4}$$

4.3.1.5 F1-Score

The F1-score is the harmonic mean of precision and recall. We report micro, macro, weighted, and sample-based F1-scores:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.5}$$

4.3.1.6 Balanced Accuracy

Balanced accuracy is the average of recall obtained on each class. It is especially useful when dealing with imbalanced datasets:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{4.6}$$

4.3.1.7 Hamming Loss

Hamming Loss computes the fraction of incorrect labels to the total number of labels, and is particularly suited for multi-label classification:

$$\text{Hamming Loss} = \frac{1}{n \times L} \sum_{i=1}^{n} \sum_{j=1}^{L} \mathbb{1}[y_{i,j} \neq \hat{y}_{i,j}] \tag{4.7}$$

where n is the number of samples, L is the number of labels, $y_{i,j}$ is the ground truth, and $\hat{y}_{i,j}$ is the predicted label.

4.3.1.8 Variants

The following variants are computed for metrics like precision, recall, and F1-score:

- **Micro**: Calculates metrics globally by counting total true positives, false negatives, and false positives.

- **Macro**: Calculates metrics independently for each label and then takes the average.

- **Weighted**: Similar to macro, but each label's metric is weighted by the number of true instances for that label.

- **Samples**: Computes metrics for each instance and then averages across all samples.

4.3.2 Experiment

We constructed our dataset with a total of 27,953 book cover image samples. To facilitate robust model development and evaluation, the dataset was partitioned into three distinct subsets: training, validation, and testing. Specifically, we employed an 8:1:1 split ratio, allocating 80% of the data (22,362 samples) for training, 10% (2,796 samples) for validation, and the remaining 10% (2,795 samples) for testing. This splitting strategy provides enough data for training while keeping separate and fair sets for testing and validating the model.

4.3.3 Experiment Result

To evaluate the effectiveness of different visual feature extractors in our hierarchical classification pipeline, we conducted experiments using a diverse set of state-of-the-art CNN and transformer-based models. These include architectures such as EfficientNet [23], RegNet [24], Swin Transformer [21], CLIP [25], and BLIP [26], among others. Each model was integrated into our two-stage classification framework to assess performance at both Level 1 (binary classification: Fiction vs. Nonfiction) and Level 2 (multi-label classification within each category). As shown in Table 4.2, Transformer-based models, particularly Swin Transformer [21], consistently outperform earlier

convolutional architectures across both levels of the hierarchy, demonstrating their strength in capturing the visual semantics of book cover images.

Table 4.2: Result of hierarchical classification for book cover page image

Model	**Level 1**		**Level 2**							
			Fiction				**Non-Fiction**			
	$\mathcal{FM}$	$\mathcal{A}$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$
Efficientformer [27]	72.25	76.77	64.93	78.43	61.45	75.90	56.97	73.62	53.49	71.46
ViT [28]	79.02	75.53	68.36	80.14	67.36	78.59	57.05	73.87	54.27	71.86
EfficientNetB3 [23]	78.98	79.05	61.02	74.78	59.85	73.92	56.08	74.16	52.32	71.26
RegNet [24]	83.79	81.91	49.50	68.43	41.57	64.97	46.72	67.85	42.82	65.80
Dinov2 [29]	83.87	82.71	40.58	63.74	32.77	64.91	12.35	53.31	8.89	51.51
ResNext-50 [30]	84.26	83.05	45.34	66.14	35.65	62.77	37.47	62.86	30.50	60.30
MobileNet-v2 [31]	84.71	83.27	40.58	63.74	32.77	60.96	13.02	53.31	9.19	52.51
VGG-19 [32]	85.25	83.35	44.05	65.47	34.59	61.83	37.41	62.58	31.88	60.52
Swiftformer [33]	84.92	84.31	66.06	78.84	61.13	75.07	63.82	77.75	**61.29**	**75.72**
ResNet50 [34]	85.94	84.40	56.94	72.96	52.46	70.18	50.63	69.81	47.40	68.21
VGG-16 [35]	85.90	84.49	62.76	77.07	60.80	74.60	59.23	75.68	55.65	73.07
DenseNet-121 [36]	85.90	84.80	18.99	55.16	11.75	53.16	30.76	59.37	20.95	56.55
SwinTv2 [37]	86.58	86.13	51.84	69.66	44.97	66.63	29.56	59.08	23.47	57.15
Xception [38]	86.69	84.85	40.01	63.54	28.98	59.49	34.30	61.29	28.71	59.20
BLIP [26]	86.82	85.49	44.61	65.90	32.46	60.94	34.79	61.54	25.87	58.31
ResNet152 [34]	87.14	85.79	55.22	72.46	51.69	70.81	47.40	68.24	41.42	65.18
CLIP [25]	**87.44**	85.65	39.90	63.41	26.01	58.40	24.64	57.13	16.31	54.93
SwinT [21]	87.32	**86.22**	**68.31**	**79.88**	**68.08**	**79.46**	**60.83**	**75.94**	58.71	74.13

4.3.4 Genrewise Analysis

To gain deeper insights into the hierarchical classification performance at a granular level, a genre-wise analysis of fiction books was conducted. This breakdown allows us to assess how well individual genres within the fiction category are being identified by the model. Metrics such as Precision (P), Recall (R), F1-score, Balanced Accuracy (BA), and Specificity (SP) were computed for each genre. The analysis reveals performance disparities across genres, highlighting areas where the model excels and where it struggles. Such fine-grained evaluation is instrumental in understanding model be-

havior and guiding future improvements in dataset balancing and architectural tuning.

Table 4.3 presents the detailed performance for fiction genres, highlighting strong classification results in genres such as "Craft & Hobbies & Home" and "Meta Text", while identifying challenges in genres like "Science & Math" and "Family & Parenting & Relationships".

Similarly, Table 4.4 provides a comprehensive performance breakdown for nonfiction genres, showcasing high classification accuracy in categories such as "Comics & Graphics" and "Romance", while also revealing difficulties in accurately predicting genres like "Sports & Outdoors".

Table 4.3: Genre-wise performance of the proposed method on fiction books

Genre	**Precision ($\mathcal{P}$)**	**Recall ($\mathcal{R}$)**	**F1-score ($\mathcal{F}$)**	**Balanced Accuracy ($\mathcal{BA}$)**	**Specificity ($\mathcal{S}p$)**
Animals & Wildlife & Pets	74.80	71.32	73.02	84.56	97.81
Arts & Photography	77.00	60.29	67.63	78.22	96.15
Business & Money	95.35	71.93	82.00	85.90	99.87
Children's Book	80.86	74.01	77.29	84.40	94.79
Comics & Graphic	66.20	63.51	64.83	80.94	98.37
Computers & Technology	90.74	76.56	83.05	88.11	99.66
Cookbooks & Food & Wine	86.30	77.78	81.82	88.55	99.32
Crafts & Hobbies & Home	**100.00**	89.29	**94.34**	**94.64**	**100.00**
Environment & Plant	80.77	65.62	72.41	82.29	98.96
Family & Parenting & Relationships	61.90	32.10	42.28	65.50	98.91
Fashion & Lifestyle	65.66	43.62	52.42	70.59	97.56
Health & Fitness & Dieting	93.22	76.39	83.97	88.06	99.73
History	77.45	48.62	59.74	72.42	96.23
Humanities	88.10	47.44	61.67	73.36	99.28
Humor & Entertainment	67.46	55.07	60.64	75.48	95.89
Literature	77.36	69.85	73.42	76.89	83.93
Mystery & Thriller & Suspense & Horror	69.79	56.62	62.51	73.10	89.58
Medical	87.36	76.77	81.72	88.00	99.24
Meta Text	94.81	82.95	88.48	91.34	99.73
Mythology & Religion & Spirituality	77.52	51.02	61.54	74.44	97.85
Press & Media	71.43	44.44	54.79	71.96	99.47
Reference & Language	58.90	64.18	61.43	81.07	97.97
Romance	73.96	57.03	64.40	76.59	96.14
Science & Math	30.43	30.43	30.43	63.59	96.75
Self-help & Motivation	91.58	**89.69**	90.62	94.57	99.45
Sports & Outdoors	69.84	51.76	59.46	75.23	98.70
Teen & Young Adult	65.93	43.48	52.40	69.30	95.11
Travel	80.95	48.57	60.71	74.01	99.46
Sci-Fi & Fantasy	71.99	78.76	75.22	81.33	83.91

Table 4.4: Genre-wise performance of the proposed method on nonfiction books

Genre	**Precision** ($\mathcal{P}$)	**Recall** ($\mathcal{R}$)	**F1-score** ($\mathcal{F}$)	**Balanced Accuracy** ($\mathcal{BA}$)	**Specificity** ($\mathcal{S}p$)
Animals & Wildlife & Pets	73.77	51.14	60.40	74.89	98.64
Arts & Photography	74.79	52.66	61.81	74.96	97.26
Business & Money	59.46	30.14	40.00	64.44	98.74
Children's Book	65.43	52.48	58.24	75.03	97.59
Comics & Graphic	84.69	**84.69**	84.69	**91.70**	98.71
Computers & Technology	70.00	37.50	48.84	68.38	99.25
Cookbooks & Food & Wine	50.00	33.33	40.00	65.88	98.43
Crafts & Hobbies & Home	84.62	64.71	73.33	82.11	99.51
Environment & Plant	51.92	29.67	37.76	63.77	97.87
Family & Parenting & Relationships	75.95	60.00	67.04	79.18	98.37
Fashion & Lifestyle	70.27	50.98	59.09	74.54	98.11
Health & Fitness & Dieting	72.37	49.55	58.82	73.86	98.18
History	62.62	65.91	64.22	73.88	81.85
Humanities	59.90	60.05	59.98	70.62	81.18
Humor & Entertainment	68.69	54.84	60.99	76.06	97.28
Literature	66.67	41.42	51.09	69.11	96.80
Mystery & Thriller & Suspense & Horror	72.22	44.32	54.93	71.52	98.72
Medical	85.96	50.52	63.64	74.91	99.31
Meta Text	90.67	70.83	79.53	85.12	**99.40**
Mythology & Religion & Spirituality	55.56	47.24	51.06	71.51	95.78
Press & Media	72.41	50.00	59.15	74.32	98.64
Reference & Language	70.35	54.64	61.51	73.88	93.11
Romance	**92.50**	78.72	**85.06**	89.11	99.49
Science & Math	69.50	47.57	56.48	71.75	95.94
Self-help & Motivation	64.13	68.21	66.11	81.08	93.95
Sports & Outdoors	44.00	22.45	29.73	60.65	98.85
Teen & Young Adult	72.34	43.04	53.97	70.97	98.90
Travel	59.72	42.57	49.71	70.04	97.51
Biographies & Memoir	67.63	63.55	65.53	76.36	89.16

4.3.5 Genre Co-occurrence Heatmaps

To analyze the correlation and distribution patterns among various genres, we employ **co-occurrence heatmaps**. These heatmaps visually represent the frequency of genres appearing together within the dataset, offering insights into genre clustering and thematic overlaps. The co-occurrence matrix M is defined as:

$$M_{i,j} = \frac{C(G_i, G_j)}{N}$$

where:

- $C(G_i, G_j)$ represents the number of times genre G_i and genre G_j co-occur in the

same book.

- N is the total number of books in the corresponding category (Fiction or Non-Fiction).

The heatmaps allow us to:

- Identify **strong genre associations** where certain genres frequently co-exist.
- Detect **underrepresented combinations**, useful for data augmentation strategies.
- Enhance hierarchical classification by leveraging cross-genre relationships.

The next sections present individual heatmaps for **Fiction** and **Non-Fiction**, highlighting their unique co-occurrence patterns and distribution dynamics.

4.3.5.1 Fiction Heatmap

The **Fiction Heatmap** visualizes the co-occurrence patterns of different fiction genres. By identifying these correlations, the heatmap helps in understanding thematic overlaps, which can be leveraged for enhanced hierarchical classification and genre prediction.

4.3.5.2 Non-Fiction Heatmap

The **Non-Fiction Heatmap** illustrates the co-occurrence relationships among non-fiction genres, which often appear together. These insights enable more refined classification by capturing the nuanced associations between non-fiction topics.

4.4 Summary

In this section, we presented a detailed exploration of book genre classification using cover page images. We formulated the problem as a hierarchical multi-label classification task, where the objective is to predict both broad categories (*Fiction*

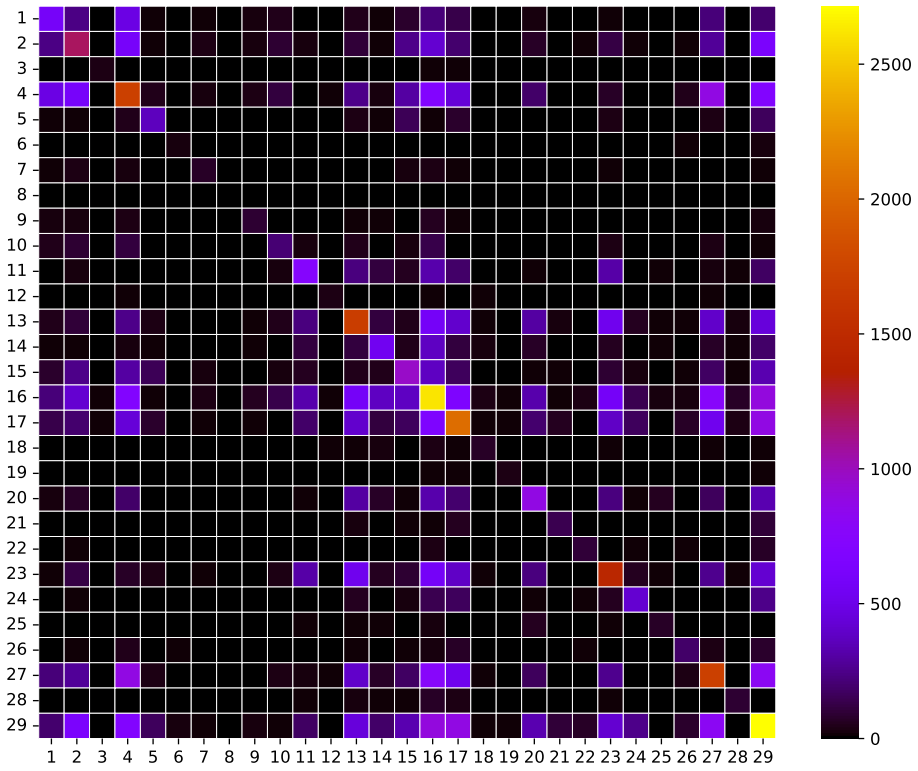


Figure 4.3: Fiction Genre Co-occurrence Heatmap

and *Non-Fiction*) and their respective subgenres based solely on visual features. The hierarchical structure of the genre taxonomy allows for refined categorization, capturing both high-level and granular genre distinctions.

We introduced a robust deep learning framework leveraging the Swin Transformer architecture for feature extraction. The Swin Transformer, with its multi-stage hierarchical processing, enables effective learning of local and global patterns from cover images, enhancing genre classification accuracy. The extracted features are further processed through a two-level hierarchical classification mechanism that first segregates books into major categories before identifying specific subgenres.

The proposed approach addresses common challenges in visual genre prediction, including visual diversity, intra-class variability, and inter-class overlap. Through the integration of hierarchical classifiers and a multi-label loss function, the method demonstrates resilience to noise and design inconsistencies inherent in book covers.

Overall, the findings highlight the untapped potential of book cover images as

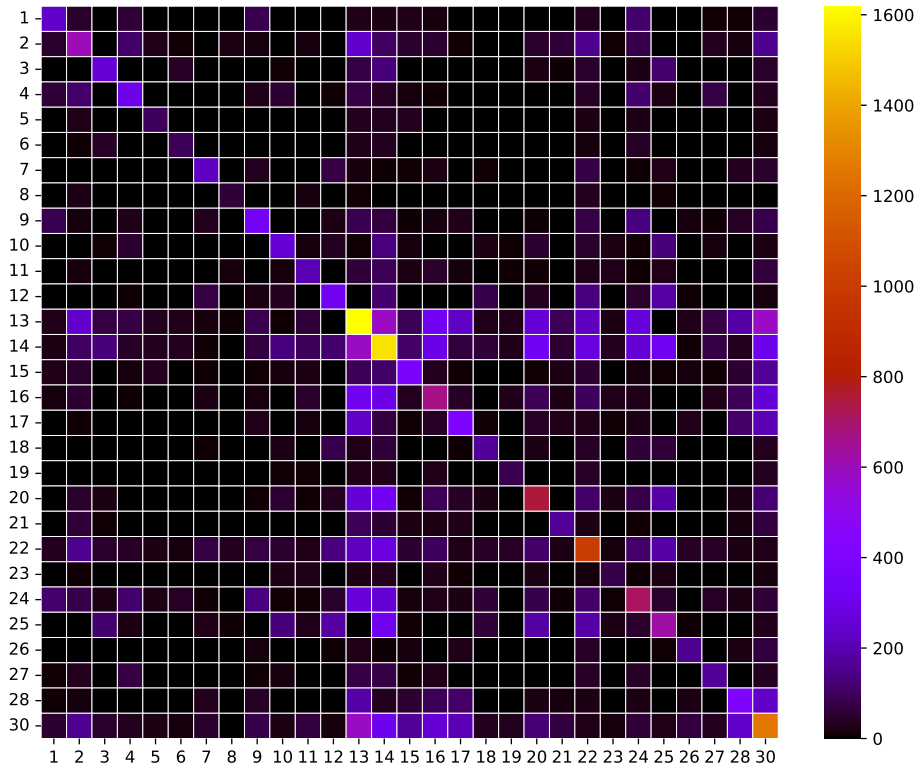


Figure 4.4: Fiction Genre Co-occurrence Heatmap

a reliable modality for genre classification, providing a path forward for metadata-independent categorization in large-scale digital libraries and recommendation systems. This visual-centric methodology not only enriches the genre identification process but also sets the foundation for further research in multimodal classification by seamlessly integrating visual and textual elements.

Chapter 5

Book Genre Identification from Unreliable Reviews Refines with Blurbs

5.1 Problem Formulation

The objective of book genre classification from **user reviews and blurbs** is to predict the literary genre of a book based on rich semantic cues extracted from its textual descriptions. Unlike cover images, user reviews and blurbs provide narrative-driven insights, emotional tone, and thematic depth that are critical for genre identification. However, these text-based sources are often noisy, biased, and sentiment-driven, posing challenges for effective genre mining.

Given:

- A dataset of n books represented as $B = \{B_1, B_2, \ldots, B_n\}$
- Each book B_i is associated with:
  - A blurb $b_i \in \mathbb{R}^{l_b}$, where l_b is the length of the blurb text
  - A collection of user reviews $R_i = \{r_{i1}, r_{i2}, \ldots, r_{im}\}$, where m is the number of reviews for book B_i and $r_{ij} \in \mathbb{R}^{l_r}$ is the j^{th} review with length l_r
- A hierarchical genre structure:

$$L = (\{0\} \times L_f) \cup (\{1\} \times L_{nf})$$

where L_f and L_{nf} are subgenres under Fiction and Non-Fiction, respectively.

Objective: The main goal is to predict the genre label(s) in L for each book B_i by leveraging:

1. **Blurb Alignment:** Using blurbs as the primary contextual anchor to filter out noisy or sentiment-biased user reviews.

2. **Review Aggregation:** Aggregating semantic cues from the filtered review set R_i to enhance genre prediction.

3. **Hierarchical Classification:** First, classify the book as Fiction or Non-Fiction, and subsequently map it to its subgenre using the enriched textual features.

Mathematical Representation: We define the mapping as follows:

$$f : (b_i, R_i) \rightarrow L$$

where the function f takes the blurb and filtered user reviews as input and maps it to the hierarchical genre structure L. The mapping is learned through deep semantic alignment and hierarchical classification mechanisms.

5.2 Proposed Method

The proposed architecture is designed to enhance **multi-label genre classification** by leveraging both **book blurbs** and **filtered user reviews**. The architecture consists of four main components: *Review Filtering*, *Feature Extraction*, *Level-1 Binary Classifier*, and *Level-2 Multi-Label Classifier*. An overview of the complete architecture is illustrated in Figure 5.1.

5.2.1 Review Filtering and Vocabulary Creation

User reviews provide a valuable source of user-driven perspectives for genre classification. However, these reviews often contain noise, subjective biases, and off-topic

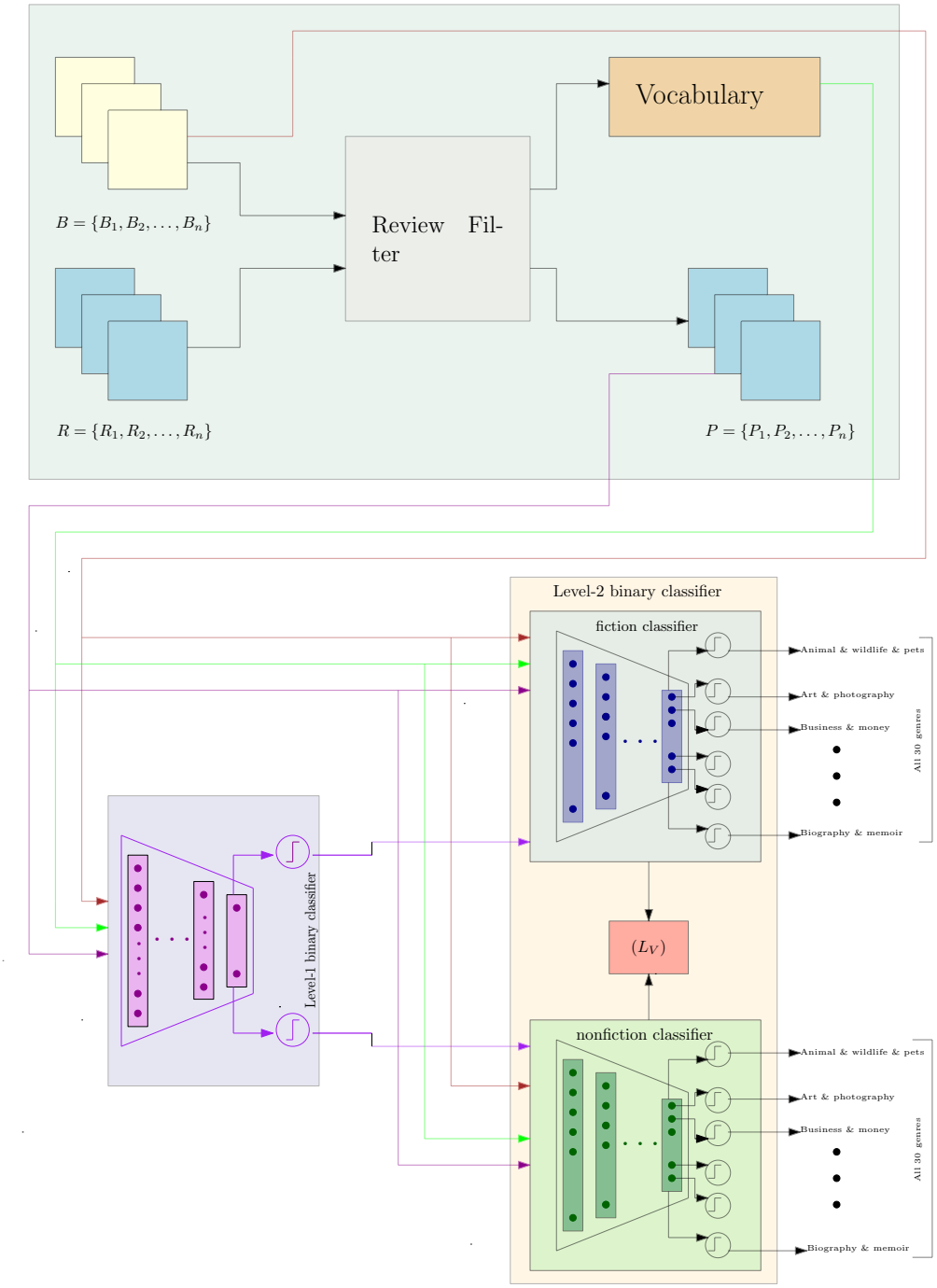


Figure 5.1: Proposed Architecture

discussions, which can degrade model performance if used directly. To address this, we propose a **blurb-guided review filtering mechanism**, where the book's blurb is utilized as an anchor for filtering out irrelevant reviews.

5.2.1.1 Semantic Alignment of Reviews

To ensure that only contextually relevant reviews are included, we employ a **semantic alignment strategy** based on cosine similarity. First, both the *blurb* (B_i) and its associated *reviews* (R_j^i) are embedded into dense vector representations using a pre-trained BERT model $\mathcal{E}_\mathcal{R}$ [39]. The embeddings for the blurb and each review are represented as:

$$b_i = \mathcal{E}_\mathcal{R}(B_i), \quad \delta_j^i = \mathcal{E}_\mathcal{R}(R_j^i)$$

where $b_i \in \mathbb{R}^d$ and $\delta_j^i \in \mathbb{R}^d$ are the embedding vectors for the blurb and the j^{th} review of book S_i, respectively.

To measure semantic alignment, we compute the **cosine similarity** between the blurb embedding and each review embedding:

$$d_j^i = \frac{b_i \cdot \delta_j^i}{\|b_i\| \|\delta_j^i\|}$$

where d_j^i denotes the similarity score for the j^{th} review of book S_i. This score ranges from -1 (completely dissimilar) to 1 (perfectly similar).

5.2.1.2 Threshold-Based Filtering

To retain only the most relevant reviews, we introduce a dynamic threshold Ψ for filtering:

$$\Psi = \min(0.5, Q_{0.75}(d^i))$$

where $Q_{0.75}(d^i)$ represents the 75^{th} percentile of the similarity scores for all reviews of book S_i. A review R_j^i is retained if:

$$d_j^i \geq \Psi$$

This ensures that only the top 25% of semantically aligned reviews are included for further processing, reducing noise and improving the contextual quality of the input data.

5.2.1.3 Consolidated Review Representation

The filtered set of reviews for each book is then concatenated to form a consolidated representation $\mathcal{R}_i$:

$$\mathcal{R}_i = \bigcup_{j \in \mathcal{J}} R_j^i, \quad \text{where} \quad \mathcal{J} = \{j | d_j^i \geq \Psi\}$$

This consolidated representation, which now only contains semantically aligned information, is used for further feature extraction and genre classification.

5.2.1.4 Vocabulary Creation

In addition to review filtering, we perform **vocabulary extraction** from the retained reviews and blurbs. The consolidated text is tokenized, and a vocabulary $\mathcal{V} = \{T_1, T_2, \ldots, T_m\}$ is constructed, where T_i represents a unique term. This vocabulary serves as the foundation for understanding genre-specific terminology and improves the interpretability of the learned model.

The combination of filtered reviews and curated vocabulary not only enhances the semantic quality of the input but also strengthens the model's capacity to capture genre-specific textual patterns.

5.2.2 Feature Extractor: BERT

To extract deep semantic features from both **book blurbs** and **filtered user reviews**, we employ the **BERT (Bidirectional Encoder Representations from Transformers)** model [39] as our primary feature extractor. Introduced by Devlin et al. (2018), BERT revolutionized natural language processing by utilizing a deep bidirectional architecture for learning language representations from both left and right contexts simultaneously.

5.2.2.1 Key Characteristics

- **Bidirectional Contextual Understanding:** Unlike traditional language models that process text either left-to-right or right-to-left, BERT performs deep

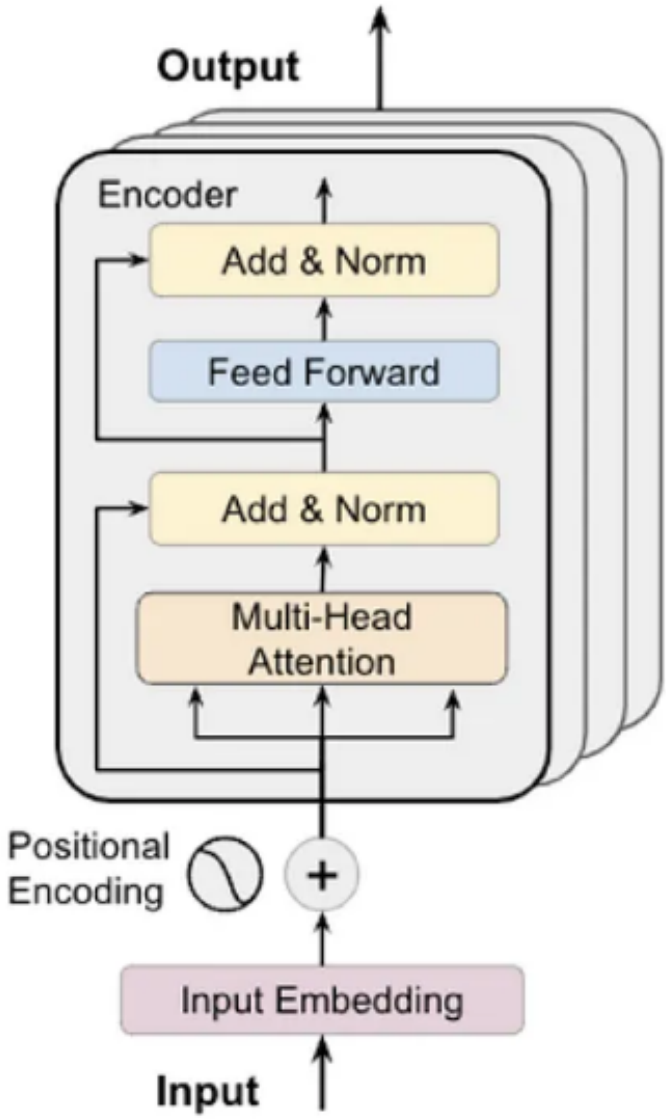


Figure 5.2: BERT encoder architecture

bidirectional learning by considering both previous and next words in all layers simultaneously. This allows BERT to understand the full context of a word based on its surroundings, enhancing semantic understanding. The representation for each word w_i in a sentence is computed as:

$$H_i = \text{BERT}(w_1, w_2, \ldots, w_i, \ldots, w_n)$$

where H_i is the contextualized embedding for the word w_i.

- **Transformer-based Architecture:** BERT is entirely built on the Transformer encoder mechanism, which leverages **self-attention** to compute relationships between words in a sequence. Given an input sequence $X = [x_1, x_2, \ldots, x_n]$, the attention scores are computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

  where Q, K, and V are the query, key, and value matrices, and d_k is the dimensionality of the key vectors.

- **Pre-trained on Massive Corpora:** BERT is pre-trained on two extensive datasets: **BooksCorpus** and **English Wikipedia**. During pre-training, two tasks are optimized:

  - **Masked Language Modeling (MLM):** Randomly masks 15% of the input tokens, and the model attempts to predict them:

$$p(w_i|X_{\text{masked}}) = \text{softmax}(WH_i + b)$$

  - **Next Sentence Prediction (NSP):** Predicts if two sentences are sequentially connected:

$$p(\text{IsNext}|S_1, S_2) = \sigma(W[H_{\texttt{[CLS]}}; H_{\texttt{[SEP]}}] + b)$$

5.2.2.2 Tokenization and Embedding

The input text, which includes both **book descriptions** and **filtered user reviews**, is first processed through BERT's tokenizer:

- Each word or subword is converted into token IDs.

- Special tokens such as `[CLS]` (classification token) and `[SEP]` (separator token) are appended to mark the beginning and end of sequences.

For example, a sample input of "The Great Gatsby" is tokenized as:

```
[CLS]  The  Great  Gatsby  [SEP]
```

This tokenized input is then embedded into dense vector representations, capturing both syntactic and semantic properties of the text. The final embedding for a token w_i is represented as:

$$E_i = E_{\text{token}} + E_{\text{segment}} + E_{\text{position}}$$

where E_{token}, E_{segment}, and E_{position} are the token, segment, and position embeddings, respectively.

5.2.2.3 Transformer Encoder Layers

BERT consists of multiple Transformer encoder layers (12 for $\text{BERT}_{\text{BASE}}$ and 24 for $\text{BERT}_{\text{LARGE}}$). Each encoder layer comprises:

- **Multi-Head Self-Attention (MHSA):** Computes attention weights across all tokens, allowing the model to focus on relevant parts of the input:

  $$\text{MHSA}(H) = [\text{head}_1; \text{head}_2; \ldots; \text{head}_h]W^O$$

  where each head is computed as:

  $$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

- **Layer Normalization (LN):** Ensures stability during training by normalizing activations:

  $$H' = \text{LayerNorm}(H + \text{MHSA}(H))$$

- **Feed Forward Neural Networks (FFN):** Applies two fully connected layers with GELU activation:

  $$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

- **Residual Connections:** Shortcut connections are added to facilitate gradient flow and improve convergence:

$$H'' = \text{LayerNorm}(H' + \text{FFN}(H'))$$

5.2.2.4 Final Representation Extraction

After passing through all transformer layers, the final hidden state corresponding to the `[CLS]` token is extracted as the aggregate representation of the entire input sequence:

$$F_T = H''_{\texttt{[CLS]}}$$

This vector F_T represents the semantic encoding of the description or review and is concatenated with features from other modules for further classification.

5.2.3 Level-1 Binary Classifier

The concatenated feature vector is then passed to the **Level-1 Binary Classifier**, which performs a high-level distinction between:

- **Fiction**
- **Non-Fiction**

The classifier consists of fully connected neural network layers followed by a **sigmoid activation function**. We utilize **Binary Cross-Entropy Loss (BCE)** during training to optimize this binary separation.

$$P_F, P_{NF} = \sigma(W_1 F + b_1)$$

where P_F and P_{NF} represent the probabilities of Fiction and Non-Fiction, and W_1 and b_1 are the learnable weights and biases.

5.2.4 Level-2 Multi-Label Classifier

Once the text is classified as Fiction or Non-Fiction by the Level-1 module, it is sent to the appropriate **Level-2 Multi-Label Classifier**:

- If classified as **Fiction**, it is passed to the *Fiction Multi-Label Classifier*.
- If classified as **Non-Fiction**, it is processed by the *Non-Fiction Multi-Label Classifier*.

Both classifiers are multi-label in nature and consist of fully connected layers with **sigmoid activation** to produce probability scores for each sub-genre. We employ an **Asymmetric Loss Function** to handle genre imbalance effectively, ensuring that less-represented genres are given appropriate learning focus.

The final multi-label classification is represented as:

$$G_F = \sigma(W_2 F + b_2), \quad G_{NF} = \sigma(W_3 F + b_3)$$

where G_F and G_{NF} are the predicted genre probabilities for Fiction and Non-Fiction, respectively, and W_2, W_3, b_2, and b_3 are the learnable parameters of the model.

5.2.4.1 Loss Function: Asymmetric Loss for Multi-label Classification

To address **label imbalance** and **label sparsity**, we use the **Asymmetric Loss** introduced by Ridnik et al. (ASL) [22]:

$$\mathcal{L}_{ASL} = -\frac{1}{C} \sum_{i=1}^{C} [y_i \cdot (1 - \hat{y}_i)^{\gamma+} \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \hat{y}_i^{\gamma-} \cdot \log(1 - \hat{y}_i)]$$

Where:

- $C = 30$ is the number of genres,
- $y_i \in \{0, 1\}$ is the true label for the i^{th} class,
- $\hat{y}_i \in (0, 1)$ is the predicted probability,

- $\gamma_+ > \gamma_-$ (commonly $\gamma_+ = 2$, $\gamma_- = 1$) control suppression of easy negatives and enhancement of hard positives.

This asymmetric formulation penalizes **false negatives more** than false positives — helping mitigate the class imbalance issue typical in multi-label setups.

5.3 Experimental Analysis

Our experiments were executed using Pytorch 2.1.0, having Python 3.10.14 on an Ubuntu 20.04.4 LTS. The hardware setup included Intel(R) Xeon(R) W-1270 clocked at 3.40GHz, with 16 CPU cores and 128 GB of RAM. Additionally, the machine was equipped with a 24 GB NVIDIA RTX A5000 GPU. Table 5.1 provide detailed experimental settings.

Table 5.1: Experimental setup details

Level-1 and Level-2 classification settings	
No. of epochs (Level-1 classification)	50
No. of epochs (Level-2 classification)	100
Batch size	16
Learning rate	10^{-5}
Weight decay	10^{-2}
Exponential decay rates	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Zero-denominator avoidance parameter	10^{-8}
Optimizer	AdamW
Patience	5

5.3.1 Evaluation Metrics

To evaluate the performance of our hierarchical multi-label classification model, we employ a comprehensive set of metrics that effectively capture both binary and multi-label classification performance. The following metrics are used:

5.3.1.1 Accuracy

Accuracy measures the proportion of correct predictions over the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

5.3.1.2 Precision

Precision is the proportion of correctly predicted positive observations to the total predicted positives. It is computed in micro, macro, and weighted forms:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.2}$$

5.3.1.3 Recall

Recall is the proportion of correctly predicted positive observations to all actual positives:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

5.3.1.4 Specificity

Specificity, also known as the True Negative Rate, measures the proportion of actual negatives correctly identified as such. It complements recall:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5.4}$$

5.3.1.5 F1-Score

The F1-score is the harmonic mean of precision and recall. We report micro, macro, weighted, and sample-based F1-scores:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.5}$$

5.3.1.6 Balanced Accuracy

Balanced accuracy is the average of recall obtained on each class. It is especially useful when dealing with imbalanced datasets:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{5.6}$$

5.3.1.7 Hamming Loss

Hamming Loss computes the fraction of incorrect labels to the total number of labels, and is particularly suited for multi-label classification:

$$\text{Hamming Loss} = \frac{1}{n \times L}\sum_{i=1}^{n}\sum_{j=1}^{L} \mathbb{1}[y_{i,j} \neq \hat{y}_{i,j}] \tag{5.7}$$

where n is the number of samples, L is the number of labels, $y_{i,j}$ is the ground truth, and $\hat{y}_{i,j}$ is the predicted label.

5.3.1.8 Variants

The following variants are computed for metrics like precision, recall, and F1-score:

- **Micro**: Calculates metrics globally by counting total true positives, false negatives, and false positives.
- **Macro**: Calculates metrics independently for each label and then takes the average.
- **Weighted**: Similar to macro, but each label's metric is weighted by the number of true instances for that label.
- **Samples**: Computes metrics for each instance and then averages across all samples.

5.3.2 Experiment

Our curated dataset comprises a total of 23,888 book cover image samples. To ensure effective model training and evaluation, the dataset is systematically divided into three distinct subsets: training, validation, and testing. Following a 7:2:1 split ratio, 70% of the samples (16,722) are designated for training, 20% samples (4,778) for validation, and the remaining 10% samples (2378) for testing. This structured partitioning guarantees sufficient data for learning while preserving unbiased samples for model assessment and fine-tuning.

5.3.3 Experiment Result

The experimental results for hierarchical classification of book genres using crowd-sourced reviews are presented in Table 5.3.3. The table compares various transformer-based models across different modalities—Blurbs, Review, and a combined modality (Blurbs + Review). The performance is evaluated at two hierarchical levels: Level 1 (Fiction vs. Non-Fiction) and Level 2 (sub-genre classification within Fiction and Non-Fiction). Metrics used for evaluation include the Macro F1-Score ($\mathcal{F}_M$), Accuracy ($\mathcal{A}$), and Balanced Accuracy ($\mathcal{BA}$) for both Fiction and Non-Fiction categories.

For the Blurbs modality, transformer models like RoBERTa and XLNet achieve competitive performance in Level 1 classification, with RoBERTa obtaining an F1-Score of 90.90% and XLNet achieving 91.57%. In contrast, the Review modality shows a notable improvement in capturing genre-specific features, with ALBERT and RoBERTa performing strongly at both Level 1 and Level 2, reflecting the richness of user-generated content for contextual understanding.

The combined modality (Blurbs + Review) using BERT significantly outperforms individual modalities, achieving the highest scores across all metrics: an F1-Score of 93.22% at Level 1, 53.44% for Fiction, and an impressive 54.51% for Non-Fiction at Level 2. This demonstrates the complementary nature of descriptive metadata and user reviews in enhancing hierarchical genre classification. The fusion of both modalities not only improves semantic understanding but also provides richer contextual

cues, leading to better discrimination of nuanced genres.

Table 5.2: Result of hierarchical classification for book crowd source reviews and blurbs

Modality	**Model**	**Level 1**		**Level 2**							
				Fiction				**Non-Fiction**			
		$\mathcal{FM}$	$\mathcal{A}$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$
Description	BERT [39]	92.95	91.80	50.76	74.33	41.30	67.94	48.19	74.27	50.29	73.36
	BLIP [26]	86.63	83.60	31.75	61.54	24.69	59.34	37.55	67.92	33.73	67.81
	DistilBERT [40]	92.36	90.85	51.08	72.71	45.46	70.12	48.68	72.60	44.48	72.72
	RoBERTa [41]	90.90	89.42	53.80	75.01	39.69	66.04	53.61	75.84	50.27	73.97
	XLNet [42]	91.57	89.95	47.44	71.23	37.36	66.22	50.54	73.05	49.94	75.32
	ALBERT [39]	91.95	90.11	53.23	75.82	40.08	68.17	50.18	75.18	48.28	74.49
Review	BERT [39]	93.22	91.86	53.44	73.82	44.59	68.65	54.51	76.26	48.52	71.95
	BLIP [26]	88.28	85.78	40.34	68.89	34.11	64.37	38.71	69.06	35.37	66.82
	DistilBERT [40]	92.25	90.91	52.45	73.26	45.48	68.81	57.31	76.48	39.47	68.14
	RoBERTa [41]	92.99	91.65	51.47	73.59	41.29	67.57	52.84	75.30	50.19	74.72
	XLNet [42]	91.32	89.75	58.04	78.32	44.71	70.54	52.68	75.36	51.70	73.78
Description + Review	BERT [39]	**93.22**	**91.86**	53.44	73.82	44.59	68.65	**54.51**	**76.26**	48.52	71.95

5.4 Summary

In this work, we proposed a robust framework for book genre classification leveraging textual information from user reviews and blurbs. Unlike traditional metadata-based classification, this approach harnesses the semantic richness of crowd-sourced reviews, capturing nuanced thematic and emotional elements that are often reflective of a book's genre. Our method incorporates a two-phase process: first, it aligns user-generated reviews with the blurb to filter out noise and retain only the most semantically consistent reviews. This is achieved through a zero-shot semantic alignment mechanism that improves the reliability of genre-specific cues. In the second phase, the filtered reviews and blurbs are utilized within a hierarchical classification architecture to distinguish between Fiction and Non-Fiction genres, followed by finer subgenre classification.

The experimental results demonstrate that integrating user perceptions and narrative-driven insights from reviews significantly enhances genre prediction accu-

racy. This work sets a foundation for multimodal genre mining by validating the effectiveness of textual signals in genre classification, paving the way for more enriched recommendation systems and intelligent digital cataloging. Our framework, by capitalizing on the depth of user engagement in reviews, provides a scalable and effective solution for genre classification in large-scale digital libraries.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis explored two complementary modalities for automated book genre classification: visual analysis of cover page images and textual understanding of user-generated content. In the first part of the work, we demonstrated that book covers—carefully designed to convey genre-specific visual themes—can be effectively used as a stand-alone modality for genre prediction. Leveraging a Swin Transformer-based feature extractor followed by a hierarchical classification scheme, our proposed vision framework showed strong performance in distinguishing both broad categories (Fiction vs. Non-Fiction) and fine-grained subgenres.

In the second part, we focused on semantic insights derived from book reviews and blurbs. We introduced a zero-shot filtering strategy based on BERT embeddings to align reviews with their associated blurbs, effectively reducing noise and retaining only semantically relevant content. This filtered textual information was then used to construct a vocabulary and train hierarchical classifiers, enabling accurate and context-aware genre identification.

Together, these approaches validate the potential of using both visual and textual signals in isolation to predict literary genres, addressing the limitations of traditional metadata-driven methods. Our custom dataset, built through web scraping from Goodreads, contributed significantly to this work by combining cover images, blurbs,

and multiple user reviews for each book.

6.2 Future Work

While the proposed frameworks demonstrate strong potential, several promising research directions can further advance the task of automated genre classification:

- **Scalability to Multilingual and Multicultural Data:** Extending the framework to handle reviews and metadata in multiple languages and from diverse cultural contexts would broaden its applicability across global literary datasets.
- **Temporal Genre Dynamics:** Investigating how genre representations evolve over time by analyzing historical cover design trends and review language can provide insights into shifting reader preferences and publishing patterns.
- **Explainability and Interpretability:** Integrating explainable AI methods to highlight which visual or textual cues led to a particular genre prediction would enhance trust and usability for publishers, authors, and librarians.
- **Reader-Aware Personalization:** Future models can incorporate user review patterns and genre affinity clusters to deliver personalized genre tagging or book recommendations tailored to individual reader profiles.
- **Genre Discovery and Emergence Detection:** Beyond classification, models could be designed to detect emerging subgenres or hybrid genres by clustering books with similar features that do not belong to existing labels.
- **Low-Resource Genre Adaptation:** Applying few-shot or zero-shot learning techniques could improve genre classification performance for rare or underrepresented genres where labeled data is limited.
- **Real-Time and Scalable Deployment:** Optimizing models for faster inference and integrating them into real-world platforms (e.g., bookstores, library

catalogs, or e-readers) could make genre classification systems more practical for large-scale usage.

- **Bias Mitigation in Genre Prediction:** As models may learn biases from skewed datasets, future work should consider fairness-aware learning techniques to ensure genre classification remains inclusive and equitable.

By addressing these directions, future research can build more intelligent, inclusive, and scalable genre classification systems that support richer reader engagement and discovery in digital literary ecosystems.

Dissemination of the Thesis

1. Utsav Kumar Nareti, Soumi Chattopadhyay, **Prolay Mallick**, Ayush Vikas Daga, Chandranath Adak, Suraj Kumar, Adarsh Wase, Arjab Roy. *"An Adaptive Data-Resilient Multi-Modal Framework for Hierarchical Multi-Label Book Genre Identification"*, IEEE Transactions on Knowledge and Data Engineering (Communicated), [Online]. Available: https://arxiv.org/abs/2505.03839

2. Suraj Kumar, Utsav Kumar Nareti, Soumi Chattopadhyay, Chandranath Adak, **Prolay Mallick** have submitted a research paper to an A* ranked international conference. Due to the double-blind review policy of the conference, specific details such as the paper title and conference name cannot be disclosed at this stage.

Bibliography

[1] B. K. Iwana, S. T. R. Rizvi, S. Ahmed, A. Dengel, and S. Uchida, "Judging a book by its cover," *arXiv preprint arXiv:1610.09204*, 2016.

[2] A. Lucieri, H. Sabir, S. A. Siddiqui, S. T. R. Rizvi, B. K. Iwana, S. Uchida, A. Dengel, and S. Ahmed, "Benchmarking deep learning models for classification of book covers," *SN computer science*, vol. 1, pp. 1–16, 2020.

[3] S. Maharjan, M. Montes, F. A. González, and T. Solorio, "A genre-aware attention model to improve the likability prediction of books," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3381–3391.

[4] C. Xu, X. Xu, N. Zhao, W. Cai, H. Zhang, C. Li, and X. Liu, "Panel-page-aware comic genre understanding," *IEEE Transactions on Image Processing*, vol. 32, pp. 2636–2648, 2023.

[5] P. Buczkowski, A. Sobkowicz, and M. Kozlowski, "Deep learning approaches towards book covers classification." in *ICPRAM*, 2018, pp. 309–316.

[6] M. S. Ullah, M. A. Al-Reza, and M. S. Rahman, "Classifying bangla book's context: A multi-label approach," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2023, pp. 1–5.

[7] A. Sobkowicz, M. Kozłowski, and P. Buczkowski, "Reading book by the cover—book genre detection using short descriptions," in *Man-Machine Interactions 5: 5th International Conference on Man-Machine Interactions, ICMMI 2017 Held at Kraków, Poland, October 3-6, 2017*. Springer, 2018, pp. 439–448.

[8] M. Khalifa and A. Islam, "Book success prediction with pretrained sentence embeddings and readability scores," *arXiv preprint arXiv:2007.11073*, 2020.

[9] J. A. Nolazco-Flores, A. V. Guerrero-Galván, C. Del-Valle-Soto, and L. P. Garcia-Perera, "Genre classification of books on spanish," *IEEE Access*, vol. 11, pp. 132 878–132 892, 2023.

[10] J. Worsham and J. Kalita, "Genre identification and the compositional effect of genre in literature," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1963–1973.

[11] C. Scofield, M. O. Silva, L. de Melo-Gomes, and M. M. Moro, "Book genre classification based on reviews of portuguese-language literature," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2022, pp. 188–197.

[12] C. Alzetta, F. Dell'Orletta, A. Miaschi, E. Prat, and G. Venturi, "Tell me how you write and i'll tell you what you read: a study on the writing style of book reviews," *Journal of Documentation*, vol. 80, no. 1, pp. 180–202, 2024.

[13] Y.-K. Ng and U. Jung, "Personalized book recommendation based on a deep learning model and metadata," in *Web Information Systems Engineering–WISE 2019: 20th International Conference, Hong Kong, China, January 19–22, 2020, Proceedings 20*. Springer, 2019, pp. 162–178.

[14] M. Saraswat and Srishti, "Leveraging genre classification with rnn for book recommendation," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3751–3756, 2022.

[15] C. Kundu and L. Zheng, "Deep multi-modal networks for book genre classification based on its cover," *arXiv preprint arXiv:2011.07658*, 2020.

[16] G. R. Biradar, J. Raagini, A. Varier, and M. Sudhir, "Classification of book genres using book cover and title," in *2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT)*. IEEE, 2019, pp. 72–723.

[17] A. Rasheed, A. I. Umar, S. H. Shirazi, Z. Khan, and M. Shahzad, "Cover-based multiple book genre recognition using an improved multimodal network," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 26, no. 1, pp. 65–88, 2023.

[18] R. Jayaram *et al.*, "Classifying books by genre based on cover," *IJEAT*, vol. 9, pp. 530–535, 06 2020.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[20] G. Team, Anil *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[21] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[22] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," *arXiv preprint arXiv:2009.14119*, 2020.

[23] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, vol. 97, 2019, pp. 6105–6114.

[24] J. Xu *et al.*, "RegNet: Self-regulated network for image classification," *IEEE TNNLS*, vol. 34, no. 11, pp. 9562–9567, 2023.

[25] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[26] J. Li *et al.*, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022, pp. 12 888–12 900.

[27] Y. Li *et al.*, "Efficientformer: Vision transformers at mobilenet speed," *NeurIPS*, vol. 35, pp. 12 934–12 949, 2022.

[28] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[29] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *TMLR*, 2024.

[30] S. Xie *et al.*, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.

[31] M. Sandler *et al.*, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[33] A. Shaker *et al.*, "Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications," in *ICCV*, 2023, pp. 17 425–17 436.

[34] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[35] Simonyan *et al.*, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2014.

[36] G. Huang *et al.*, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[37] Liu *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, 2017, pp. 1251–1258.

[39] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[40] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[42] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.