Enhancing Book Recommendation with Automated Genre Mining

MS(Research) Thesis

By

Prolay Mallick

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

June 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 their 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 is essential for improving discovery, search, and recommendation in large-scale digital book platforms, where readers face an ever-growing volume of titles. This thesis investigates automated book genre classification through two independent yet complementary modalities: visual semantics from book cover images and textual semantics from book descriptions and user reviews, enabling flexible deployment under different metadata availability settings. A significant contribution of this research is the construction of a comprehensive custom dataset acquired through large-scale web scraping, which integrates diverse cover imagery with multi-source textual metadata. To overcome the inherent challenges of data scarcity and severe class imbalance in specific genre nodes, we implement a generative data augmentation strategy leveraging the Gemini API to synthesize high-quality, genre-aligned text, thereby enriching underrepresented categories.

For the visual modality, we propose a hierarchical framework based on the Swin Transformer to learn genre-relevant cues from cover designs, such as typography, layout, color patterns, and illustrative elements. The system performs coarse-to-fine classification, first predicting Fiction vs. Nonfiction (Level-1) and then predicting fine-grained subgenres (Level-2). Experiments demonstrate strong accuracy and robustness, with improved generalization compared to CNN-based baselines across diverse cover styles.

For the textual modality, our second work focuses on mining deep semantic signals from book descriptions and user reviews while explicitly mitigating the impact of review noise and off-topic sentiment. We introduce a sophisticated semantic filtering pipeline using BERT embeddings and cosine similarity to ensure that only reviews content-aligned with the core book description are retained. This is followed by a hierarchical multi-label classification approach that evaluates performance across broad and fine-grained categories. Ablation studies across Description-only, Review-only, and combined modalities confirm that this pipeline consistently outperforms models trained on raw, unfiltered reviews by focusing learning on content-relevant signals.

Table of Contents

Contents

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

June 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 “Nonfiction” 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 multi-modal 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 Refined 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 multi-modal 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, qualitative analysis, 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 multi-modal 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 Nonfiction, 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 Literatutre Review

The task of book genre classification has been approached through three primary modalities: **Visual Methods**, **Textual Methods**, and **Multi-modal 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.

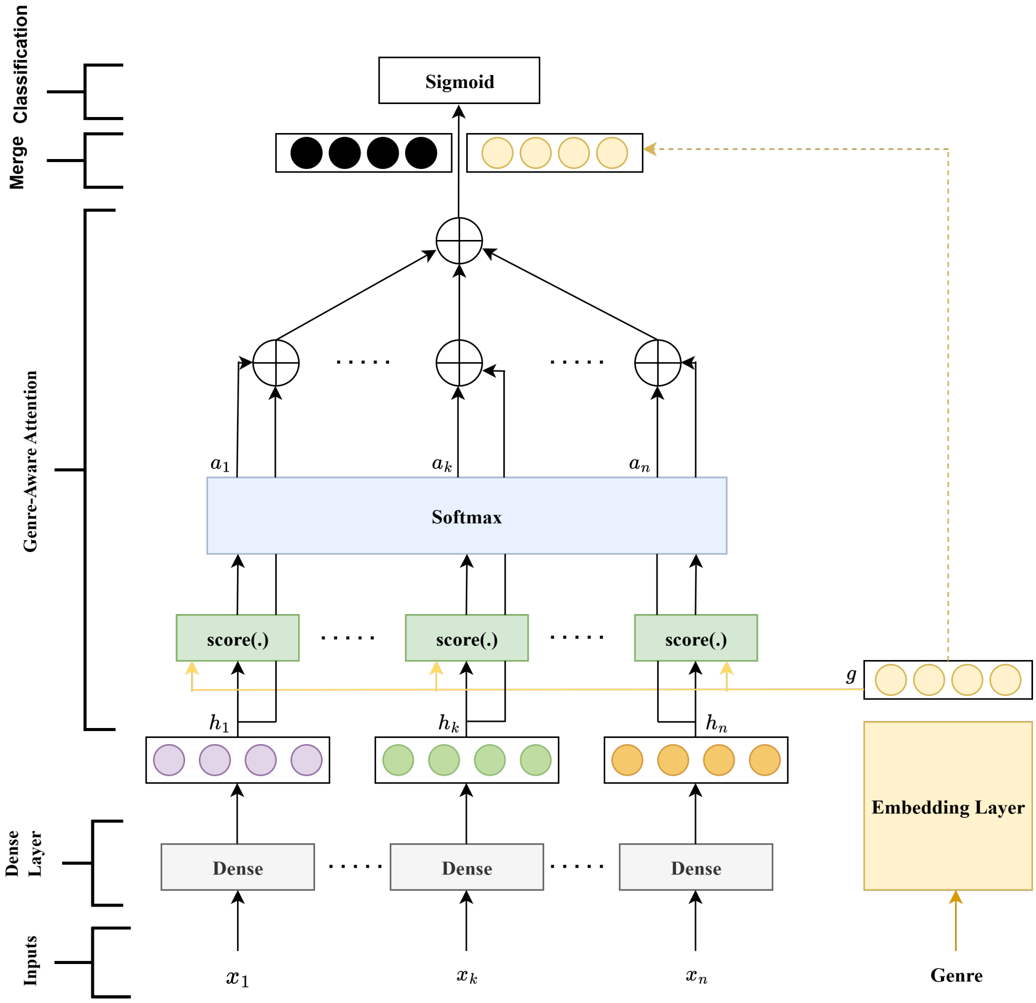


Figure 2.1: Overview of Maharjan et al. genre-aware attention model [3].

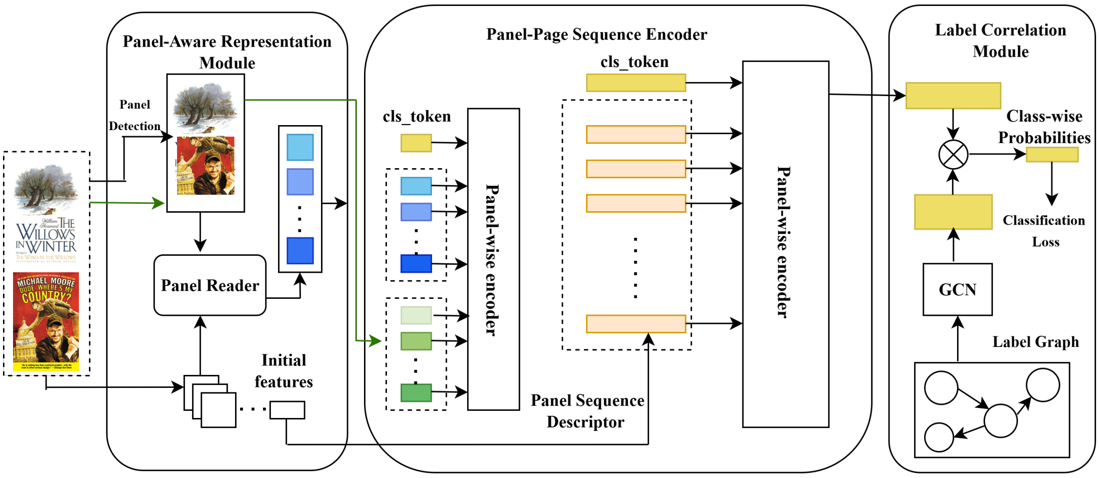


Figure 2.2: Overview of Xu et al. paper panel aware attention model [4].

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.

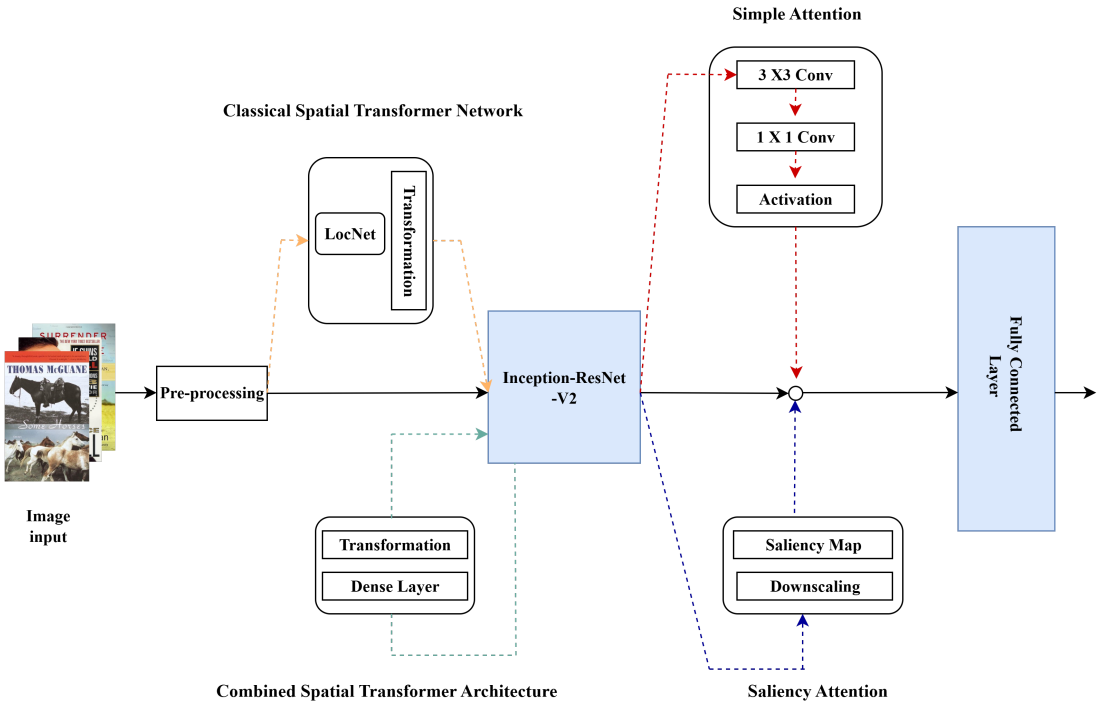


Figure 2.3: Architecture combining Inception-ResNet-V2 with spatial transformer and attention modules for book cover genre classification [2].

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 multi-modal 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.

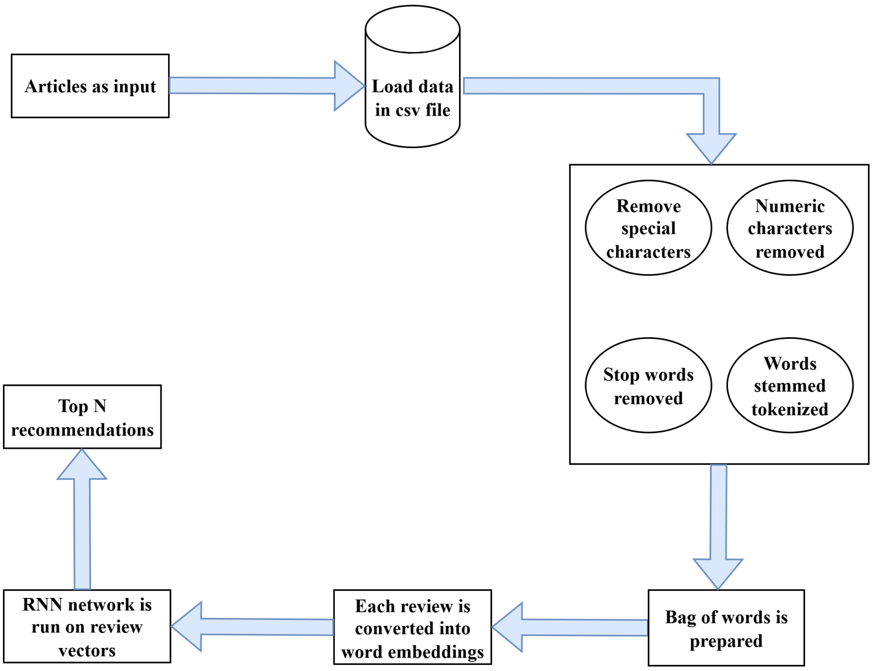

Articles as input
Load data in csv file
Remove special characters
Numeric characters removed
Stop words removed
Words stemmed tokenized
Bag of words is prepared
Each review is converted into word embeddings
RNN network is run on review vectors
Top N recommendations

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

2.2.3 Multi-modal Method

Multi-modal 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 multi-modal neural network architecture for genre classification. Visual features were extracted using a *ResNet-50* backbone pre-trained 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.

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 multi-modal 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.

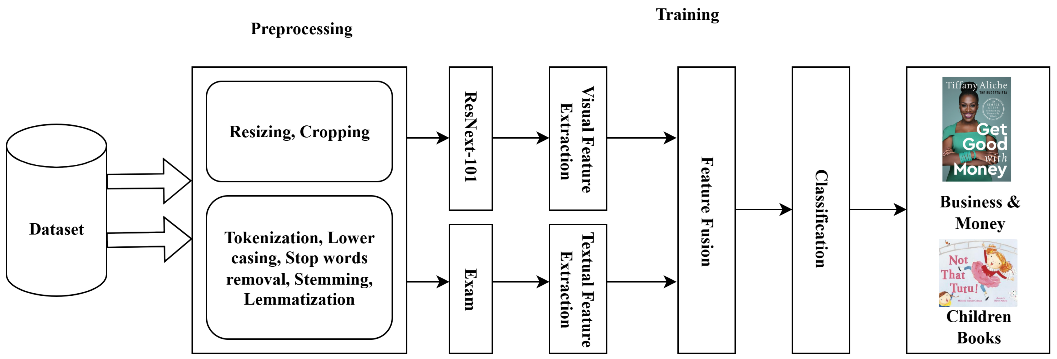


Figure 2.5: Rasheed et al.'s multi-modal framework combining SE-ResNeXt-101 and EXAN with attention-based fusion [17].

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

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 Multi-modal 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 **multi-modal 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 Details

3.1 Dataset Creation

To the best of our knowledge, there is no publicly available dataset for hierarchical multi-label book genre prediction with multi-modal information. To address this gap, we constructed a dataset from scratch. Initially, we web-scraped book cover images, blurb texts, metadata (such as authors and publishers), and associated genres from Goodreads. However, Goodreads lists approximately 2055 user-defined genres, many of which are redundant, anomalous, or highly subjective. Hence, genre labeling required standardization.

To address the complexity of multi-label genre annotation, we simplified the task of managing numerous interrelated genres by grouping them into broader, hierarchically structured parent categories. This approach enhanced consistency and interpretability in the labeling process. At Level-1, books were classified into *fiction* and *nonfiction* categories. At Level-2, each book was annotated with 1 to 6 sub-genre labels, with 29 labels defined for each top-level category. Eight annotators, in consultation with three linguistic experts, performed the genre assignments. To ensure labeling consistency, we measured inter-annotator agreement using Krippendorff's alpha (α) [19], which achieved a score of $\alpha = 0.83$, indicating a high level of agreement and annotation reliability. The final dataset contains 11,302 books, comprising 6,704 fiction and 4,598 nonfiction samples.

In the multi-modal setting, each book sample comprised a cover image, blurb text, and metadata. Additionally, we engaged Gemini Pro Vision [20] to perform optical character recognition (OCR) on the cover images to extract embedded textual content or cover texts.

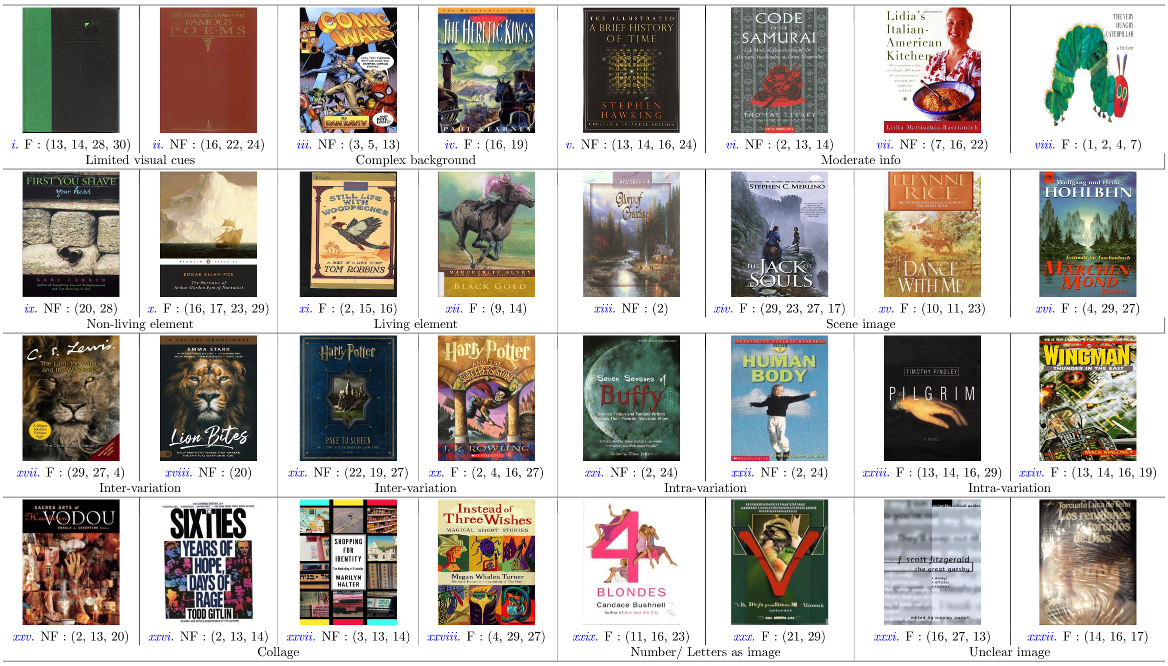
3.2 Dataset Challenges

The book cover page image presents complexity in accurately identifying book genres. We summarize these challenges as follows:

- ***Challenges in Cover Images:***

  The complexity of book cover images poses challenges in accurately identifying book genres.

  - *Background:* Book cover image background plays a crucial role in setting the tone or mood, sparking interest, and assisting readers in determining whether the book aligns with their tastes and preferences. For instance, sci-fi & fantasy book covers frequently showcase detailed background elements, featuring enchanting landscapes and vivid colors, designed to captivate readers with a sense of wonder and adventure (Figs. 3.1: *viii*). Based on the visual cues available in the background of the cover image, we can group it as follows:
  - *Limited Visual Cues:* Sometimes, the cover page provides minimal or no visual information, containing only text in standard fonts, making genre identification challenging (Figs. 3.1: *i–ii*).
  - *Moderate Information:* The cover page often includes moderate visual features to represent some of its genres but may fail to capture others due to the multi-label nature of genres (Figs. 3.1: *v–viii*).
  - *Complex Background:* In some instances, the background of a book cover image becomes convoluted due to the presence of extensive or composite visual effects or elements (Figs. 3.1: *iii-iv*).



i. F : (13, 14, 28, 30) *ii.* NF : (16, 22, 24)
Limited visual cues
iii. NF : (3, 5, 13) *iv.* F : (16, 19)
Complex background
v. NF : (13, 14, 16, 24) *vi.* NF : (2, 13, 14) *vii.* NF : (7, 16, 22) *viii.* F : (1, 2, 4, 7)
Moderate info
ix. NF : (20, 28) *x.* F : (16, 17, 23, 29)
Non-living element
xi. F : (2, 15, 16) *xii.* F : (9, 14)
Living element
xiii. NF : (2) *xiv.* F : (29, 23, 27, 17) *xv.* F : (10, 11, 23) *xvi.* F : (4, 29, 27)
Scene image
xvii. F : (29, 27, 4) *xviii.* NF : (20)
Inter-variation
xix. NF : (22, 19, 27) *xx.* F : (2, 4, 16, 27)
Inter-variation
xxi. NF : (2, 24) *xxii.* NF : (2, 24)
Intra-variation
xxiii. F : (13, 14, 16, 29) *xxiv.* F : (13, 14, 16, 19)
Intra-variation
xxv. NF : (2, 13, 20) *xxvi.* NF : (2, 13, 14) *xxvii.* NF : (3, 13, 14) *xxviii.* F : (4, 29, 27)
Collage
xxix. F : (11, 16, 23) *xxx.* F : (21, 29)
Number/ Letters as image
xxxi. F : (16, 27, 13) *xxxii.* F : (14, 16, 17)
Unclear image

Figure 3.1: Examples of some challenging cover page images with mentioned issues

- *Foreground:* The foreground of a book cover image plays a vital role in capturing the reader's attention. However, foreground images on book covers pose challenges for genre identification due to ambiguity, cross-genre similarities, abstract designs, stylistic variations, cultural influences, evolving trends, and marketing-driven misrepresentations.

- *Living/ Non-living Element:* The foreground image can effectively communicate the book's theme or atmosphere by strategically placing engaging foreground elements. Some non-living elements (e.g., mountains, sea, etc.) are typically associated with genres like travel. However, in some cases (Figs. 3.1: *x*), such elements are used to represent genres like literature, mystery & thriller & suspense & horror & adventure, romance, and sci-fi & fantasy, which creates a disconnect and makes genre identification challenging. Similarly, living objects contribute to this ambiguity. For example, while an animal on the cover often implies a connection to genres like animals & wildlife & pets, there are instances (Figs. 3.1: *xi-xii*) where this association does not hold, making genre identification more challenging.

- *Scene Image:* Scene-based images on book covers further complicate genre prediction due to their intricate visual details, clutter, and lack of a cohesive composition (Figs. 3.1: *xiii-xvi*). Unlike single-object covers, scene-based images often contain multiple elements—characters, landscapes, objects, or action sequences—that may belong to different genres. This complexity increases ambiguity, as the dominant visual theme may not be immediately apparent. Additionally, certain scenes may be common across multiple genres, making classification more challenging. Variations in artistic styles, lighting, and color palettes further contribute to genre ambiguity.

- *Inter-variation:* Sometimes, books from different genres may contain similar visual information (Figs. 3.1: *xvii-xviii*, and Figs. 3.1: *xix-xx*). This may be intentional to challenge reader expectations, create intrigue, or highlight genre blends. For example, cover image of Figs. 3.1: *xvii* comprises genres children's book, comics & graphics, sci-fi & fantasy, teen & young adult, but another book's cover image (Figs. 3.1: *xviii*) with similar visual cues belongs to arts & photography.

- *Intra-Variation:* The visual styles and design elements used in cover images within a specific genre can vary widely. This variation is often influenced by the author/artist's unique vision and the intended mood of the book, resulting in diverse interpretations of the same genre. Additionally, publishers frequently adhere to specific house styles or branding guidelines, incorporating consistent visual elements or color schemes across their catalog. While this ensures some level of uniformity within a publisher's collection, the broader diversity in design choices makes genre identification more challenging (Figs. 3.1: *xxi-xxii*, and Figs. 3.1: *xxiii-xxiv*).

- *Collage:* In some cases, book covers integrate multiple aforementioned background and foreground elements, forming a collage of numerous smaller images. This visual complexity often leads to information overload, making it challenging to accurately determine the book's genre (Figs. 3.1:

xxv-xxviii).

  - *Number/ Letter as Image:* Occasionally, book covers feature numbers or letters stylized as visual elements. This artistic choice is often intended to emphasize a theme, establish a distinctive style, or highlight key information related to the book's content. However, such designs pose a unique challenge for genre identification, as the graphical representation of text or numbers can obscure their intended meaning, making interpretation more difficult (Figs. 3.1: *xxix-xxx*).

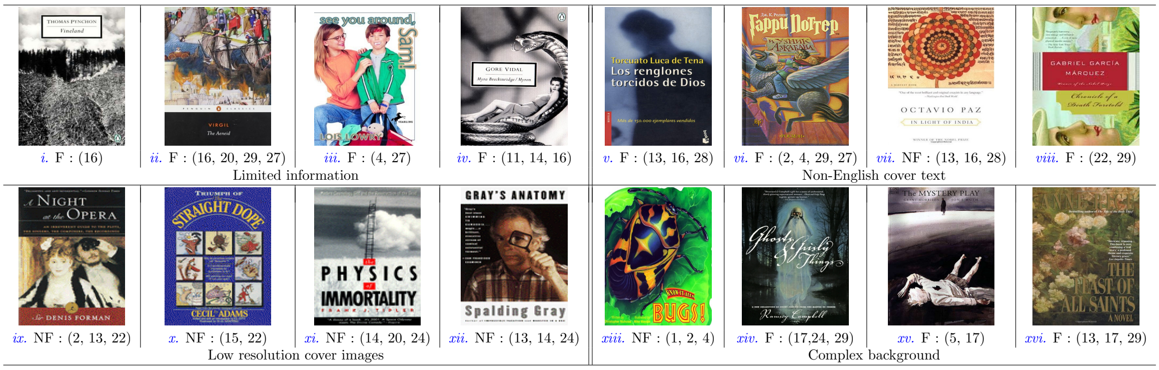
  - *Unclear Image:* Sometimes, book covers feature images that are blurry, hazy, or of low resolution. Extracting meaningful features from such images becomes difficult, making accurate genre identification challenging (Figs. 3.1: *xxxi-xxxii*).

- ***Challenges in Cover Text:***

  We extracted text from the book cover images using an OCR (Optical Character Recognition) engine. However, these extracted texts present additional challenges, which we summarize below.

  - *Limited Text:* Some book covers feature only minimal text, such as the author name, or the book title, without supplementary elements like subtitles, or descriptive taglines. These additional details often play a crucial role in genre identification. The absence of such details forces the OCR system to rely solely on sparse information, significantly increasing the risk of misclassification (Figs. 3.2: *i-iv*).

  - *Linguistic Issues:* OCR systems are often designed with a primary focus on specific languages, e.g., English. When book covers feature text in non-English languages, these systems may face difficulties in accurately recognizing and processing the characters. Errors in language detection or character recognition can lead to misinterpretations, further complicating text-based genre classification (Figs. 3.2: *v-viii*).



i. F : (16) | *ii.* F : (16, 20, 29, 27) | *iii.* F : (4, 27) | *iv.* F : (11, 14, 16)

Limited information

v. F : (13, 16, 28) | *vi.* F : (2, 4, 29, 27) | *vii.* NF : (13, 16, 28) | *viii.* F : (22, 29)

Non-English cover text

ix. NF : (2, 13, 22) | *x.* NF : (15, 22) | *xi.* NF : (14, 20, 24) | *xii.* NF : (13, 14, 24)

Low resolution cover images

xiii. NF : (1, 2, 4) | *xiv.* F : (17,24, 29) | *xv.* F : (5, 17) | *xvi.* F : (13, 17, 29)

Complex background

Figure 3.2: Some challenges of extracting OCR from book cover page

– *Low Resolution Cover Page:* Low-quality and glossy book covers present significant challenges for accurate text extraction using OCR. Glossy surfaces can create glare and reflections when photographed or scanned, interfering with character recognition. Additionally, low-resolution images often result in blurred text, making it difficult to distinguish individual characters. These factors contribute to errors in text extraction and misinterpretation (Figs. 3.2: *ix-xii*).

– *Complex Background:* Colorful or busy backgrounds on book covers present a significant challenge for OCR systems. These complex backgrounds can interfere with the OCR processes in several ways:

– *Visual Clutter:* A busy background with multiple colors, patterns, or images can create visual clutter. This clutter makes it difficult for the OCR system to distinguish the text from the background. The presence of various elements can cause the OCR algorithm to misidentify parts of the background as text or fail to recognize the text altogether (Figs. 3.2: *xiii-xvi*).

– *Color Contrast:* Text on colorful backgrounds might not have sufficient contrast. When the text color closely matches the background colors, the OCR system struggles to differentiate between them. High contrast between text and background is crucial for accurate OCR, and colorful backgrounds often fail to provide this (Figs. 3.2: *xiii-xvi*).

- ***Challenges in Blurbs:***

  The blurb often provides genre-related cues, but accurately identifying genres from it poses several challenges, as outlined below.

  – *Insufficient Information:* Several books include blurbs that lack sufficient detail, making it difficult to accurately determine their genres.

  – *Irrelevant Information:* We encounter many books with blurbs that contain irrelevant information. These blurbs often include vague or generic statements, offering little insight into the genre and making accurate genre identification challenging.

  – *Multilinguality:* Blurb texts may appear in multiple languages, each with distinct scripts and sentence structures, creating challenges in designing a framework that can uniformly interpret and process multilingual content.

- ***Challenges of Metadata:***

  There are several challenges for identifying book genres using book metadata information. We summarized these challenges as follows.

  – *Surface-Level Information:* Metadata doesn't reveal a user's deeper interests, reading level, or genre preferences. One author writes books on different genres. So, someone who enjoys a particular author might not like all of their books.

- ***Challenges of Ground-Truthing:***

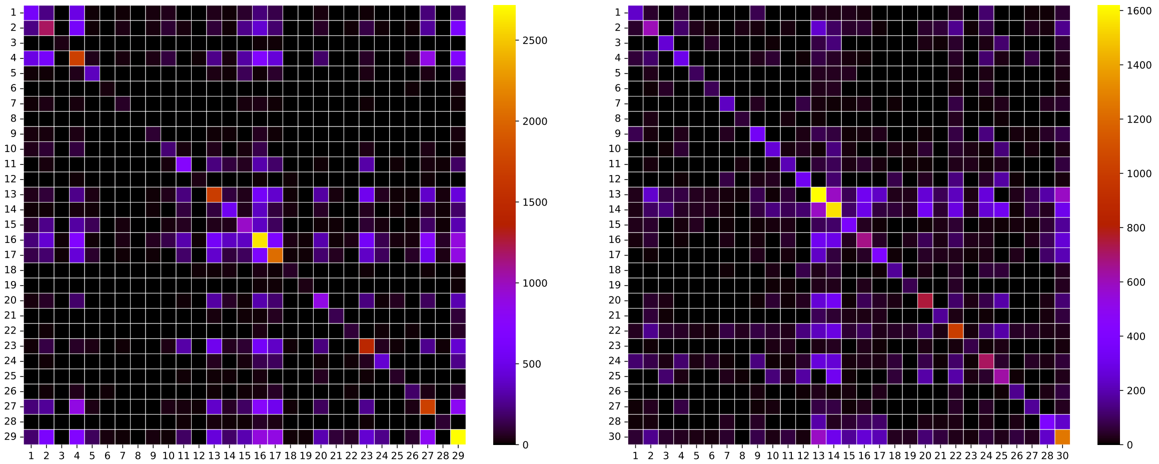
  Establishing reliable ground-truth genre labels presented significant challenges due to the inherent subjectivity and complexity of the task.

  – *Partial Understanding by Human Annotators:* Since it is often impractical for annotators to read an entire book, the ground-truthing process primarily relied on linguistic experts who assigned genre labels based on limited content—such as publisher blurbs, user reviews (e.g., from Goodreads), and selected excerpts. This partial exposure may lead to a superficial or incomplete understanding of the book's thematic nuances, thereby impacting the accuracy of genre labeling.

  - *Subjectivity and Inconsistency in Genre Interpretation:* Genre classification is often influenced by subjective interpretation. Annotators might interpret the same book differently based on the emphasis placed on particular elements. For instance, a book that blends psychological drama and crime might be labeled as a thriller by one expert, and as a mystery or drama by another, depending on their reading perspective.
  - *Fiction vs. Nonfiction Ambiguities:* Differentiating between fiction and nonfiction genres was particularly challenging for works related to well-known fictional universes. For example, while a "Harry Potter" novel is clearly fiction, a companion book like "Harry Potter – Page to Screen: The Complete Filmmaking Journey" is nonfiction, despite sharing the same universe and characters. Such genre-conflicting edge cases complicated the labeling process, requiring careful contextual analysis.
  - *Multi-Label Overlaps and Unclear Boundaries:* Many books span multiple genres, making it difficult to determine which labels are most appropriate and at what level of the hierarchy.
  - *Lack of Standardized Taxonomy:* Genre taxonomies vary significantly across publishers, retailers, and literary databases, resulting in inconsistent ground-truth references. Aligning expert annotations with a unified genre hierarchy was non-trivial and often required manual reconciliation.

- ***Challenges of Overall Dataset:*** In preparing a high quality dataset for hierarchical multi-label book genre classification, several overarching challenges emerged, particularly due to the multi-label nature of the task and the complexity of real-world genre distributions.

  - *Class Imbalance in Multi-Label Context:* The dataset exhibits significant



(a)

(b)

Figure 3.3: Co-occurrence matrices for book genres: (a) *fiction*, (b) *nonfiction*

class imbalance, with certain genres being highly represented, while others appear far less frequently. However, unlike single-label classification, this imbalance cannot be easily corrected through traditional resampling techniques because books often belong to multiple genres simultaneously. Balancing one under-represented genre could unintentionally disrupt the co-occurrence patterns with more frequent genre.

- *Constraints in Data Augmentation:* Augmentation techniques commonly used to balance datasets pose unique risks in the multi-label setting. When augmenting samples from a genre with few examples, the co-occurring genres in the original sample, many of which may already be overrepresented, also get replicated. This reinforces existing imbalances and biases, making it difficult to selectively boost specific genre classes without inadvertently inflating others. (Figs. 3.3 *(a), (b)*) Visually depict the relationships and co-occurrence patterns among book fiction and nonfiction genres.

- *Sparse Label Combinations:* The multi-label setting also leads to a large number of unique label combinations, many of which appear very infrequently. This sparsity in the label space challenges the model's ability to generalize well to unseen or rare combinations and limits the effectiveness of frequency-based heuristics.

- *Dependency between Genre Labels:* Usually, genre labels are not mutually independent; the presence of one genre may often influence the likelihood of another. Modeling these dependencies becomes increasingly difficult as the number of labels grows, especially under imbalance and sparsity, and it further complicates synthetic balancing or sampling strategies.

3.3 Data Augmentation

We observed that some specific genres within both fiction and nonfiction categories were underrepresented in our dataset. To address this imbalance, we applied data augmentation techniques. Each book sample in our dataset consists of a cover image, cover text, blurb, metadata, and genre labels. For augmentation purposes, we focused on the cover image, cover text, and blurb, while keeping metadata and genre labels unchanged. We now discuss our data augmentation strategies.

- ***Visual Data Augmentation:*** We augmented the cover images using SDEdit [21], a diffusion-based generative model. SDEdit introduces noise to an input image and then refines it to produce high-quality synthetic variations, preserving the semantic structure while diversifying the visual representation.

- ***Textual Data Augmentation:*** For cover text and blurb augmentation, we leveraged the Gemini large language model. By applying prompt engineering, we guided the model to generate alternative versions of the text that retained the contextual integrity and relevance of the original content.

  All augmented samples were manually reviewed and validated by human annotators. As a matter of fact, these augmented data instances were used solely during the model training phase to enhance generalization and robustness.

3.4 Statistical Information

Table 3.1 presents a statistical overview of each input modality, both before and after augmentation. The reported metrics include the minimum, maximum, mean, median, and standard deviation (SD). The measured attributes are area of the cover page image (in pixels^2), word counts from the blurb and cover text, and the number of books authored or published as derived from the metadata.

Table 3.1: Statistical analysis

Input		**Minimum**	**Maximum**	**Mean**	**Median**	**SD**
Cover page images	Image area before augmentation	20460	16301712	435282.41	146775	1125224.73
	Image area after augmentation	20460	16301712	325205.84	135850	980565.34
Blurb text	No. of words before augmentation	0	1786	120.52	98	102.37
	No. of words after augmentation	0	1953	110.01	96	80.89
Cover text	No. of words before and after augmentation	1	944	20.67	15	27.69
Metadata	No. of books written by an author	1	43	1.57	1	1.91
	No. of books published by a publisher	1	268	5.80	1	16.64

SD: Standard Deviation

Chapter 4

Book Genre Classification from Reliable Sources

4.1 Problem Formulation

We are given:

- A set of n books $\mathcal{B} = \{\mathcal{B}_1, \ldots, \mathcal{B}_n\}$, each with a cover image I_i.
- A two-level genre hierarchy. Level-1 labels are $L^1 = \{0, 1\}$, indicating *fiction* (0) vs. *nonfiction* (1). Level-2 labels are partitioned into

$$L^2_F = \{\gamma_1, \ldots, \gamma_{m_F}\} \quad \text{(fiction subgenres)}, \qquad L^2_N = \{\lambda_1, \ldots, \lambda_{m_N}\} \quad \text{(nonfiction subgenres)}$$

Each book $\mathcal{B}_i$ has a hierarchical label $L_i = (\ell^1_i, L^2_i)$ with $\ell^1_i \in L^1$ and

$$L^2_i \subseteq \begin{cases} L^2_F, & \text{if } \ell^1_i = 0, \\ L^2_N, & \text{if } \ell^1_i = 1. \end{cases}$$

The goal is: given $(\mathcal{B}_i, I_i)$, predict (ℓ^1_i, L^2_i). Level-1 is a *binary classification* task, and Level-2 is a *multi-label classification* within the branch selected by Level-1.

4.2 Solution Architecture

Figure 4.1 illustrates our unimodal (vision-only) hierarchy. A Swin Transformer extracts a visual feature $\mathbf{f}_i^v$ from I_i. A Level-1 classifier Φ_B predicts $\mathcal{Y}_1$ (fiction vs. nonfiction) and yields an intermediate embedding $\mathbf{h}_B$. A gating step then routes the concatenated vector $[\mathbf{f}_i^v; \mathbf{h}_B]$ to a Level-2 classifier Φ_V with two heads: Φ_V^F for fiction and Φ_V^N for nonfiction. The active head outputs multi-label scores $\mathcal{Y}_2^F \in [0,1]^{m_F}$ or $\mathcal{Y}_2^N \in [0,1]^{m_N}$.

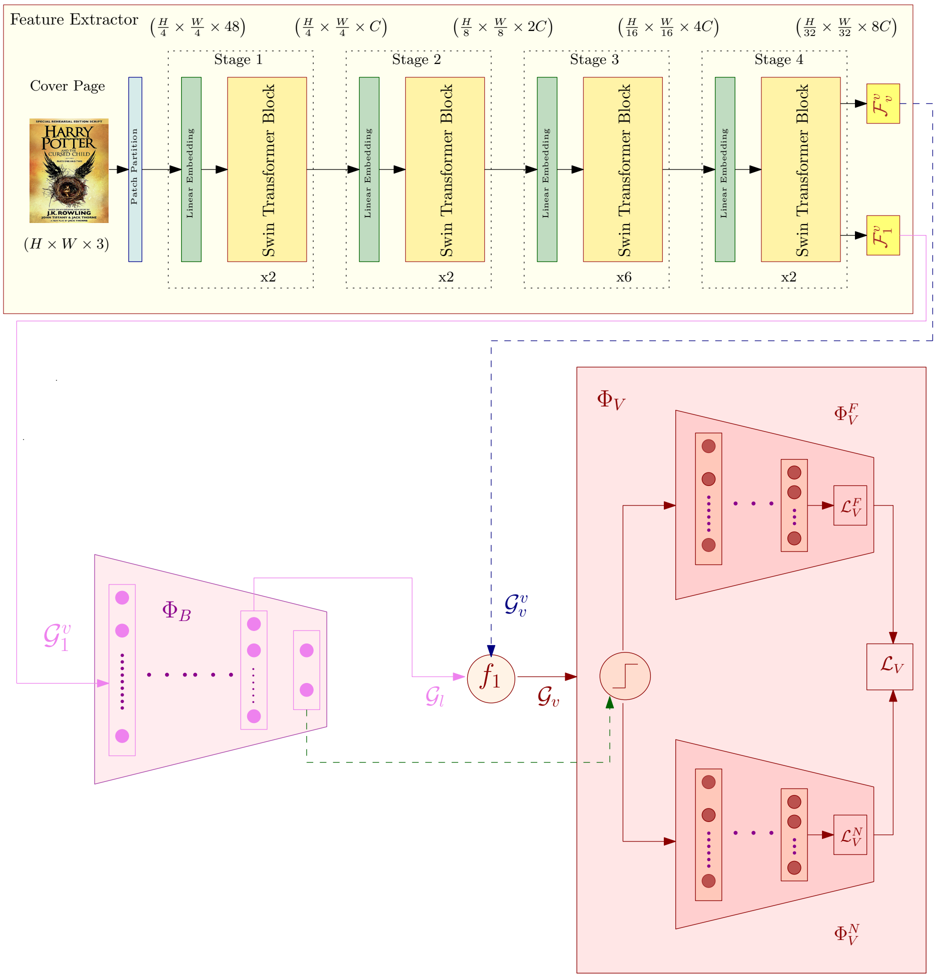


Figure 4.1: Architecture of the visual inference pipeline ψ_V: Swin $\rightarrow \mathbf{f}_i^v$; Level-1 Φ_B produces $\mathcal{Y}_1$ and $\mathbf{h}_B$; the gate selects Φ_V^F or Φ_V^N; Level-2 outputs $\mathcal{Y}_2^F$ or $\mathcal{Y}_2^N$.

4.3 VIS: Visual Inference Sub-Architecture ψ_V

The pipeline ψ_V comprises two stages linked by a gate:

1. **Level-1 (Φ_B):** Input $\mathbf{f}_i^v$; output $\mathcal{Y}_1 \in [0,1]^2$ and embedding $\mathbf{h}_B$.

2. **Level-2 (Φ_V):** Input $[\mathbf{f}_i^v; \mathbf{h}_B]$; two heads Φ_V^F and Φ_V^N. The branch is selected using $\mathcal{Y}_1$. Outputs are $\mathcal{Y}_2^F$ or $\mathcal{Y}_2^N$ (sigmoid activations for multi-label prediction).

4.4 Loss Function

Level-1 uses standard binary cross-entropy on $\mathcal{Y}_1$. For Level-2, we adopt an *asymmetric multi-label loss* computed only on the active branch $b \in \{F, N\}$ (fiction or nonfiction).

Let m_b be the number of labels in branch b, and for book i, let $y_{i,j}^{(2,b)} \in \{0,1\}$ and $\hat{y}_{i,j}^{(2,b)} \in (0,1)$ be the ground-truth and predicted scores for label j, respectively. Define

$$P_\epsilon(\hat{y}) = \max(\hat{y} - \epsilon, 0),$$

where $\gamma^+, \gamma^- > 0$ are focusing exponents and $\epsilon, \epsilon_0 > 0$ are small stability constants. The per-branch loss is

$$\mathcal{L}_M^{(b,i)} = \frac{1}{m_b} \sum_{j=1}^{m_b} \Big(- y_{i,j}^{(2,b)} \big(1 - \hat{y}_{i,j}^{(2,b)}\big)^{\gamma^+} \log\big(\hat{y}_{i,j}^{(2,b)} + \epsilon_0\big) - \big(1 - y_{i,j}^{(2,b)}\big) \big(P_\epsilon(\hat{y}_{i,j}^{(2,b)})\big)^{\gamma^-} \log\big(1 - P_\epsilon(\hat{y}_{i,j}^{(2,b)}) + \epsilon_0\big) \Big).$$

The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}^{(1)} + \mathcal{L}_M^{(b,i)},$$

where b denotes the branch determined by the Level-1 target (or by its prediction, depending on the training variant).

4.5 Feature Extractor

We extract visual semantics from book cover images using a hierarchical transformer backbone. This module provides the sole modality in our unimodal framework and produces the feature vector consumed by the downstream classifiers.

4.5.1 Visual Feature Extractor from Cover Image

We adopt the *Swin Transformer* [22] as our visual backbone owing to its window-based self-attention and shifted-window mechanism, which enable efficient long-range interactions with lower computational cost than global self-attention. The input cover image is split into non-overlapping patches and linearly embedded into a d_v-dimensional space (where d_v is the visual embedding dimension). Stacked Swin stages process these embeddings and yield the final visual representation

$$\mathbf{f}_i^v \in \mathbb{R}^{d_v}.$$

This representation $\mathbf{f}_i^v$ is used by all downstream components: the Level-1 classifier Φ_B and the Level-2 classifier Φ_V (with two heads, Φ_V^F for fiction and Φ_V^N for nonfiction). In later sections, we concatenate $\mathbf{f}_i^v$ with the intermediate embedding $\mathbf{h}_B$ from Φ_B to form the input to Φ_V.

4.6 Training Strategy

Our unimodal visual architecture is trained in two sequential phases to ensure effective hierarchical supervision and robust specialization across the classification branches.

In the first phase, we construct a reliable dataset D. This dataset includes samples that contain book cover page images and accurate genre labels. Initially, the Level-1 binary classifier Φ_B is trained using D to predict whether a book is fiction or nonfiction. Once Φ_B is trained, we proceed to train the Level-2 classifiers: Φ_V^F for fiction and Φ_V^N

for nonfiction genres. Both classifiers are trained using D, but only on the respective branches based on the ground truth fiction/nonfiction label. During this phase, the output from the penultimate layer of Φ_B is concatenated with the visual representation $\mathbf{f}_i^v$ from the Swin Transformer to form the input to the Level-2 classifiers. The classifier Φ_B is frozen during Level-2 training to retain its learned coarse-level decision boundary.

In the second phase, we jointly fine-tune all components of the model—Φ_B, Φ_V^F, and Φ_V^N—using the complete dataset D. During this stage, the model alternates between updating the Level-1 and Level-2 classifiers, enabling consistent supervision across hierarchical levels. This alternating training schedule ensures that the binary classifier remains aligned with the evolving genre classifiers and that the branch-specific genre predictors continue to specialize based on refined binary guidance.

This two-phase training process enhances the model's ability to perform coarse-to-fine hierarchical classification using only visual input, while leveraging both low-level and high-level features to improve genre prediction accuracy.

4.7 Experimental Setup

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 4.1 provides detailed experimental settings.

4.8 Database Employed

The primary goal of this study is to analyze visual data from book covers and identify associated multi-label genres through a hierarchical classification framework.

Since no publicly available hierarchical book genre dataset exists, to the best of our knowledge, we created a comprehensive dataset containing 11,302 book samples, categorized into 6,704 fiction and 4,598 nonfiction books, each annotated with 1 to 6

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

genre labels. The dataset includes the book cover images and associated genre labels.

Table 4.2 presents the distribution of genre labels across both fiction and nonfiction categories. Most samples are annotated with multiple genres, leading to overlapping counts and contributing to significant class imbalance. For example, genres like *sci-fi & fantasy* and *history* are highly frequent, while genres such as *computer & technology* and *crafts & hobbies & home* are relatively rare.

To address this imbalance, we applied data augmentation techniques to the cover images, which helped in improving the distribution across genres. However, due to the inherently multi-label nature of the dataset and co-occurrence of majority and minority genres, a perfectly balanced dataset was not achievable. After augmentation, *literature and humanities* emerged as the most frequent genre, whereas *press & media* and *sports & outdoors* remained among the least frequent.

For model training and evaluation, the dataset was divided into three non-overlapping subsets: D_{train} (80%), D_{valid} (10%), and D_{test} (10%). This stratified split preserved the overall genre label distribution, allowing for effective supervised learning. Only the cover images and genre annotations were used throughout all stages of training and evaluation, aligning with our unimodal visual learning approach.

4.9 Evaluation Metrics

To comprehensively evaluate model performance, we report the following metrics for Level-1 (binary) classification: F1-score ($\mathcal{F}$), and accuracy ($\mathcal{A}$), all in percentage.

For Level-2 (multi-label) classification, we report F1-score and balanced accuracy in micro ($\mathcal{F}_\mu$, $\mathcal{BA}_\mu$), macro ($\mathcal{F}_m$, $\mathcal{BA}_m$), weighted ($\mathcal{F}_w$, $\mathcal{BA}_w$), and samples-based ($\mathcal{F}_s$, $\mathcal{BA}_s$) forms. Additionally, Hamming loss ($\mathcal{HL}$) is used to capture label-wise mismatches. These different variants offer complementary perspectives: micro averages emphasize frequent classes, macro treats all classes equally, weighted accounts for class frequency, and sample-based metrics reflect per-instance performance—crucial for multi-label tasks with class imbalance.

Table 4.2: Hierarchical genre-wise count of the dataset

Class ID	**Genre label**	**Fiction**		**Nonfiction**	
		Before augmentation	After augmentation	Before augmentation	After augmentation
1	Animals & Wildlife & Pets	590	1260	235	924
2	Arts & Photography	1188	2566	606	1782
3	Business & Money	42	624	256	760
4	Children's 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	40	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 & Adventure	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	Sci-Fi & Fantasy	2715	5034	–	–
30	Biographies & Memoir	–	–	1256	3466

4.10 Experimental 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 [22], 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 4.3, Transformer-based models, particularly Swin Transformer [22], consistently outperform earlier convolutional architectures across both levels of the hierarchy, demonstrating their strength in capturing the visual semantics of book cover images.

Table 4.3: Result of hierarchical classification for book cover page image (complete dataset)

Model	**Level-1**		**Level-2**							
			Fiction				**NonFiction**			
	$\mathcal{FM}$	$\mathcal{A}$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$
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
SwinT V2 [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 [22]	87.32	**86.22**	**68.31**	**79.88**	**68.08**	**79.46**	**60.83**	**75.94**	58.71	74.13

4.11 Ablation Studies

4.11.1 Performance Analysis on Non-Augmented Data

To evaluate the effectiveness of the proposed data augmentation techniques, we conducted a detailed ablation study comparing the performance of various models trained without augmented data. This comparison is critical to understanding how the framework handles the inherent challenges of the dataset.

The results from this study reveals that without data augmentation, models suffer from a substantial drop in performance across all hierarchical levels.

Table 4.4: Result on non-augmented dataset

Model	**Level 1**		**Level 2**							
			Fiction				**NonFiction**			
	$\mathcal{FM}$	$\mathcal{A}$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$
Efficientformer [27]	58.68	59.36	52.23	59.24	49.32	57.02	49.13	55.32	**48.89**	54.12
ViT [28]	61.23	60.91	53.14	59.26	**50.69**	58.45	50.17	58.06	48.36	54.28
EfficientNetB3 [23]	59.17	60.45	48.51	61.28	44.32	59.27	46.13	60.25	42.06	58.35
RegNet [24]	62.77	63.95	50.12	64.79	48.86	61.68	48.22	61.39	47.58	57.01
Dinov2 [29]	60.53	61.12	49.52	62.36	45.36	59.17	44.92	60.08	44.26	57.63
ResNext-50 [30]	62.25	64.19	53.14	59.13	50.23	58.93	50.43	55.69	47.06	55.31
MobileNet-v2 [31]	58.74	61.26	47.41	59.55	46.32	56.12	45.23	57.14	45.96	54.25
VGG-19 [32]	51.03	53.14	45.23	50.74	42.09	49.22	42.03	48.59	40.29	47.15
Swiftformer [33]	58.55	59.13	49.63	60.56	46.39	58.93	43.78	59.24	45.33	55.46
ResNet50 [34]	59.78	58.07	48.21	57.06	46.29	55.46	45.16	55.64	46.02	53.39
VGG-16 [35]	60.98	61.25	50.23	54.33	48.75	51.89	48.46	51.02	45.96	49.28
DenseNet-121 [36]	63.87	64.19	52.67	60.26	49.76	56.78	49.27	56.98	46.02	54.08
SwinT V2 [37]	67.12	66.58	53.16	66.89	49.37	63.19	**50.33**	62.01	44.51	61.11
Xception [38]	58.01	58.49	48.32	55.69	45.06	53.26	47.36	56.93	42.97	50.49
BLIP [26]	61.23	60.47	49.07	60.11	45.19	58.34	46.87	58.02	43.09	56.21
ResNet152 [34]	60.08	61.11	48.34	58.15	43.93	55.29	47.55	55.12	42.28	52.15
CLIP [25]	60.69	59.72	46.53	61.06	44.65	59.77	44.26	58.79	44.69	52.13
SwinT [22]	**70.89**	**69.33**	**53.49**	**68.82**	50.16	**67.48**	50.19	**64.19**	48.01	**64.33**

4.11.2 Performance on Flat Multi-label Classification

To validate the necessity of the proposed two-stage framework, we conducted an experiment using a flat classification approach. In this setup, the models were trained to predict all 56 genres simultaneously, treating them as independent labels without the “Fiction vs. Nonfiction” categorical guidance.

Table 4.5: Performance Measure on Flat Multi-label Classification

Model	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$
Efficientformer [27]	43.26	55.12	40.89	53.09
ViT [28]	42.04	50.06	41.12	48.27
EfficientNetB3 [23]	40.23	56.89	38.65	55.22
RegNet [24]	43.26	58.45	41.29	55.79
Dinov2 [29]	41.02	57.99	38.02	56.29
ResNext-50 [30]	44.29	55.16	40.69	53.62
MobileNet-v2 [31]	40.15	54.04	36.13	53.02
VGG-19 [32]	39.26	44.78	38.14	43.22
Swiftformer [33]	44.27	55.43	40.11	52.61
ResNet50 [34]	40.66	56.13	37.89	54.08
VGG-16 [35]	39.01	52.01	46.13	50.66
DenseNet-121 [36]	42.99	56.23	39.93	52.99
SwinT V2 [37]	44.25	54.25	41.06	52.13
Xception [38]	41.34	55.15	37.12	50.26
BLIP [26]	42.08	57.92	38.59	54.01
ResNet152 [34]	40.29	55.13	38.55	51.28
CLIP [25]	37.16	57.11	34.33	**56.33**
SwinT [22]	**45.16**	**59.19**	**43.97**	56.29

4.11.3 Inference Time

To evaluate the practical deployability of the proposed framework, we conducted a rigorous analysis of inference latency and model complexity. This evaluation is critical for understanding the trade-off between the high-fidelity genre mining achieved by transformer-based architectures and the real-time processing requirements of large-scale digital libraries.

Table 4.6: Inference Latency Analysis for Hierarchical Genre Classification

Model	**Parameters (Millions)**	**Latency per Level (ms)**		**Total Inference Time (ms)**
		Level-1 (Binary)	**Level-2 (Multi-label)**	
Efficientformer [27]	12.3	51.5	77.2	128.7
ViT [28]	86.0	93.3	140.0	233.3
EfficientNetB3 [23]	12.0	61.3	92.0	153.3
RegNet [24]	3.9	25.1	37.6	62.7
Dinov2 [29]	21.0	68.1	102.1	170.2
ResNext-50 [30]	25.0	77.8	116.7	194.5
MobileNet-v2 [31]	3.5	20.3	30.5	50.8
VGG-19 [32]	143.6	176.4	264.6	441.0
Swiftformer [33]	6.3	38.6	57.9	96.5
ResNet50 [34]	25.6	62.7	94.0	156.7
VGG-16 [35]	138.0	157.1	235.6	392.7
DenseNet-121 [36]	8.0	45.7	68.6	114.3
SwinT V2 [37]	28.0	72.2	108.3	180.5
Xception [38]	22.9	70.9	106.3	177.2
BLIP [26]	224.0	182.3	263.2	445.5
ResNet152 [34]	60.0	115.9	173.8	289.7
CLIP [25]	151.0	163.7	245.6	409.3
SwinT [22]	197	170.2	203.5	373.7

4.12 Qualitative Analysis

4.12.1 Heatmap Encoding

Table 4.7 presents 12 selected data samples, and provides a qualitative evaluation of the performance of the proposed model through heatmap encodings for them. The grayscale heatmaps show the ground-truth and model-predicted genre labels, with white representing positive and black indicating negative genres.

4.12.2 Performance Stagnation Analysis

In this section, we first explore the quantitative analysis for genre indistinguishability, and then the qualitative analysis for misprediction by our model.

- ***Qualitative and Quantitative Analysis for Genre Indistinguishability:*** In multi-label classification, especially with imbalanced datasets, label co-occurrence can significantly affect prediction accuracy. This becomes particularly problematic when a dominant genre frequently overlaps with a less represented one, leading the model to underrepresented the minority due to skewed training distributions. Such patterns can result in false positives or false negatives, and obscure the actual genre structure. The interaction between sci-fi & fantasy and press & media in the fiction category in our dataset exemplifies this issue, as illustrated below.

The training set contains 396 samples labeled as press & media and 3980 as sci-fi & fantasy, with only 264 samples overlapping between the two. This means that 93% of the scifi & fantasy training samples do not belong to press & media, while only 7% are shared. Conversely, 67% of the press & media samples also belong to sci-fi & fantasy, leaving just 33% exclusive to press & media. This co-occurrence imbalance leads to two key observations:

- **Case-(i):** Because the vast majority (93%) of sci-fi & fantasy samples are not associated with press & media, the model tends to predict only sci-fi & fantasy for test instances that actually belong to both categories, leading to false negatives for press & media (refer to Table 4.10: (a),(b)).
- **Case-(ii):** Since only one-third of press & media training samples are exclusive, the model often overgeneralizes, predicting sci-fi & fantasy for test samples that are only labeled press & media, resulting in false positives for sci-fi & fantasy (refer to Table 4.10: (c)).

These patterns highlight the challenges of genre prediction in multi-label settings with skewed co-occurrence distributions.

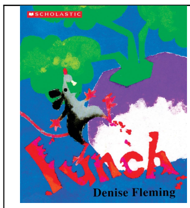
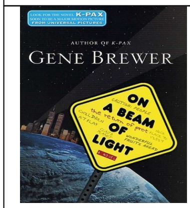
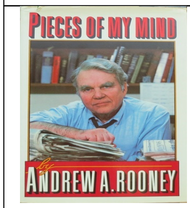
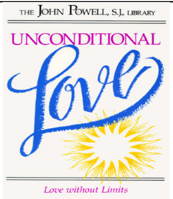
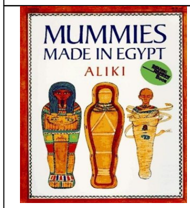
Tables 4.8: (a) and 4.9: (a) present pair-wise genre sample count for fiction and nonfiction genre in $\mathcal{D}_{\text{train}}$. Each cell in the table represents the number of samples belonging to both the row and column genre class IDs. Tables 4.8: (b) and 4.9: (b) illustrates the co-occurrence ratio between row class ID i in association with column class ID j for fiction and nonfiction genres, respectively. Each cell contains the ratio of samples belonging to both class IDs i and j to the total number of training samples associated with class ID j. Tables 4.8: (c) and

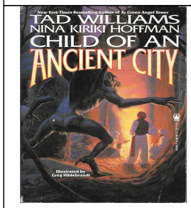
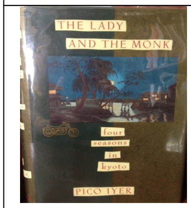
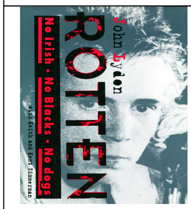
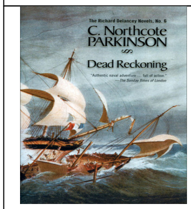
4.9: (c) reports the misclassification rate by our proposed model for fiction and nonfiction genres, respectively. Here, each cell corresponds to row class ID i and column class ID j represents the ratio between the number of testing samples associated with class ID j, but wrongly identified as class ID i and the total number of testing samples associated with class ID j that has been misclassified.

Our analysis of Tables 4.8 and 4.9 suggests that the high co-occurrence of the two genres, coupled with a significant imbalance in the number of training samples associated with one genre without the other, potentially causes the high misclassification rate. For instance, in Table 4.8, 264 out of 396 press & media (class ID 21) books are associated with sci-fi & fantasy (class ID 29) genre (refer to Table 4.8: (a)). This leaves only 132 samples that are not categorized as sci-fi & fantasy. Conversely, there are 3716 sci-fi & fantasy books that do not belong to the press & media genre, creating an imbalance between non-sci-fi & fantasy press & media and non-press & media sci-fi & fantasy samples. This leads to press & media genre being misclassified as sci-fi & fantasy with a misclassification rate of 0.39. Standard data augmentation methods are limited in their ability to address this imbalance issue, as increasing the number of samples for press & media genre, also increases the sci-fi & fantasy samples due to intergenre relation. We observe similar issues between humanities (class ID 14) and literature (class ID 16), animals & wildlife & pets (class ID 1) and children's book (class ID 4) (refer to Table 4.8), and nonfiction genres teen & young adult (class ID 27) and history (class ID 13) (refer to Table 4.9). Table 4.10 (a) showcases the example of misclassified data samples due to genre indistinguishability.

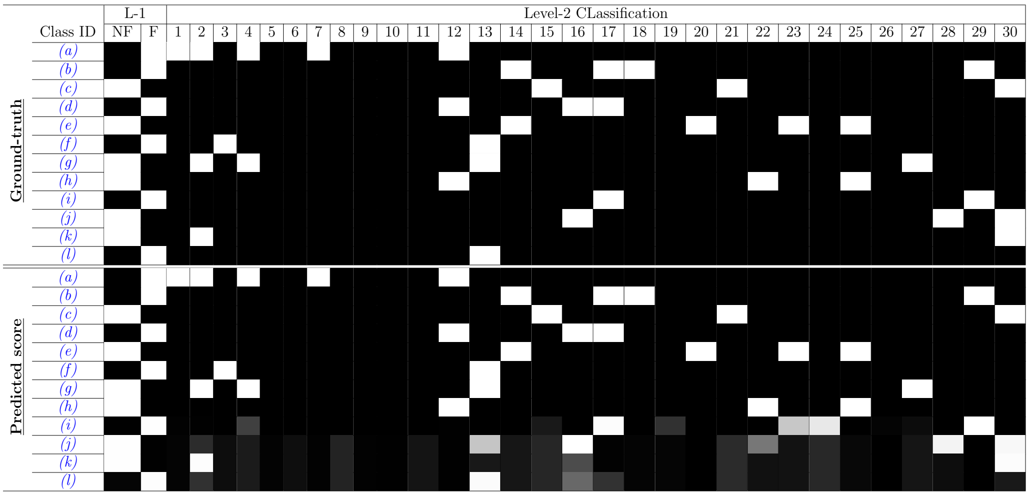
- ***Qualitative Analysis of Model Limitation:*** Another significant reason for mispredictions is the inherent limitations of the model. From Table 4.10 (b), we observe genre misprediction in Level-2 classification due to misclassification in Level-1 classification for some samples.

Table 4.7: Qualitative samples of book entries with image, metadata, and ISBN

 	***(a) 9780590467650:*** Lunch; **Author:** Denise Fleming; **Publisher:** Scholastic Inc
 K-PAX AUTHOR OF K-PAX GENE BREWER ON A BEAM OF LIGHT	***(b) 9780312269265:*** On a Beam of Light; **Author:** Gene Brewer; **Publisher:** St. Martin's Press
 	***(c) 9780689114922:*** Pieces of My Mind; **Author:** Andy Rooney; **Publisher:** Scribner
 	(d) 9780399149368: Prince of Lost Places; **Author:** Kathy Hepinstall; **Publisher:** Putnam Publishing Group
 THE JOHN POWELL, S.J. LIBRARY UNCONDITIONAL Love Love without Limits	***(e) 9780883473122:*** Unconditional Love: Love Without Limits; **Author:** John Joseph Powell; **Publisher:** Thomas More Pr
  "HIGH DRAMA." "FASCINATING." WALTER J. BOYNE Bestselling coauthor of THE WILD BLUE TROPHY FOR EAGLES Determined to be the best, two rival pilots compete in the air and on the ground, in a battle that only one will survive.	***(f) 9780804105293:*** Trophy for Eagles; **Author:** Walter J. Boyne; **Publisher:** Ivy Books
 	***(g) 9780064460118:*** Mummies Made in Egypt; **Author:** Aliki; **Publisher:** HarperCollins

	(h) 9780451208620: Simple Steps: 10 Weeks to Getting Control of Your Life; **Author:** Lisa Lelas, Linda McClintock; **Publisher:** New American Library
	***(i) 9780812533910:*** Child of an Ancient City; **Author:** Tad Williams; **Publisher:** Tor Books
	***(j) 9780679403081:*** The Lady and the Monk: Four Seasons in Kyoto; **Author:** Pico Iyer; **Publisher:** Knopf
	***(k) 9780312099039:*** Rotten: No Irish, No Blacks, No Dogs; **Author:** John Lydon, Keith Zimmerman; **Publisher:** St Martins Pr
	***(l) 9781590130384:*** Dead Reckoning; **Author:** C. Northcote Parkinson; **Publisher:** McBooks Press

(b) Ground-truth and model-predicted confidence score heatmap encodings in gray color code



L-1: Level-1 classification, NF: *nonfiction*, F: *fiction*

Table 4.8: Fiction genre indistinguishability analysis

Class ID ↓→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
(a) **Sample Count**																													
1	1002	539	0	817	23	0	153	0	206	143	0	102	41	38	190	362	38	45	158	52	0	0	16	33	27	0	248	0	304
2	539	2024	12	1279	32	68	308	44	206	319	24	141	129	54	435	716	46	81	204	137	23	102	154	44	71	53	327	39	882
3	0	12	510	0	28	18	0	0	0	0	0	0	154	109	44	160	0	0	213	30	14	0	0	0	106	0	0	14	69
4	817	1279	0	2575	104	69	291	44	262	387	0	174	273	72	449	997	70	109	484	242	14	51	74	25	83	189	979	19	876
5	23	32	28	104	643	19	9	0	0	0	0	0	62	32	270	32	0	0	137	0	9	19	69	17	0	19	56	0	299
6	0	68	18	69	19	473	0	0	0	0	0	0	39	18	54	137	0	0	154	0	0	19	16	18	0	179	16	36	419
7	153	308	0	291	9	0	646	0	16	62	38	102	24	40	191	306	16	0	133	16	8	0	92	17	0	0	52	32	103
8	0	44	0	44	0	0	0	408	0	0	77	0	103	0	0	191	0	0	268	44	0	52	0	52	0	0	52	0	0
9	206	206	0	262	0	0	16	0	713	40	26	0	141	104	40	441	0	0	158	84	0	0	36	48	0	50	76	25	184
10	143	319	0	387	0	0	62	0	40	638	80	48	130	0	103	371	31	0	20	48	0	0	97	0	36	0	94	8	28
11	0	24	0	0	0	0	38	77	26	80	1128	96	254	139	99	614	89	54	256	151	0	9	376	0	153	0	18	66	159
12	102	141	0	174	0	0	102	0	0	48	96	566	30	200	79	365	257	0	84	21	0	0	9	0	0	0	142	0	76
13	41	129	154	273	62	39	24	103	141	130	254	30	2473	138	70	985	108	155	650	382	73	20	574	77	187	65	400	174	506
14	38	54	109	72	32	18	40	0	104	0	139	200	138	1169	85	693	226	40	256	287	0	34	97	8	229	26	146	86	339
15	190	435	44	449	270	54	191	0	40	103	99	79	70	85	1595	604	25	172	271	38	40	42	100	19	34	40	159	83	454
16	362	716	160	997	32	137	306	191	441	371	614	365	985	693	604	4730	378	218	1044	496	67	216	803	219	277	88	899	425	1175
17	158	204	213	484	137	154	133	268	158	20	256	84	650	256	271	1044	200	214	3371	258	168	0	379	158	90	236	471	242	1184
18	38	46	0	70	0	0	16	0	0	31	89	257	108	226	25	378	752	0	200	29	46	0	117	24	0	8	114	0	144
19	45	81	0	109	0	0	0	0	0	0	54	0	155	40	172	218	0	598	214	29	0	64	74	16	0	0	85	0	182
20	52	137	30	242	0	0	16	44	84	48	151	21	382	287	38	496	29	29	258	1517	0	73	352	18	615	0	171	56	427
21	0	23	14	14	9	0	8	0	0	0	0	0	73	0	40	67	46	0	168	0	**396**	23	0	16	0	0	0	14	**264**
22	0	102	0	51	19	19	0	52	0	0	9	0	20	34	42	216	0	64	0	73	23	560	0	118	0	63	76	11	379
23	16	154	0	74	69	16	92	0	36	97	376	9	574	97	100	803	117	74	379	352	0	0	1778	49	193	12	247	67	423
24	33	44	0	25	17	18	17	52	48	0	0	0	77	8	19	219	24	16	158	18	16	118	49	580	0	0	92	0	281
25	27	71	106	83	0	0	0	0	0	36	153	0	187	229	34	277	0	0	90	615	0	0	193	0	799	0	51	45	81
26	0	53	0	189	19	179	0	0	50	0	0	0	65	26	40	88	8	0	236	0	0	63	12	0	0	657	114	0	366
27	248	327	0	979	56	16	52	52	76	94	18	142	400	146	159	899	114	85	471	171	0	76	247	92	51	114	1973	0	766
28	0	39	14	19	0	36	32	0	25	8	66	0	174	86	83	425	0	0	242	56	14	11	67	0	45	0	0	603	94
29	304	882	69	876	299	419	103	0	184	28	159	76	506	339	454	1175	144	182	1184	427	**264**	379	423	281	81	366	766	94	**3980**
(b) **Cooccurrence**																													
1	1	0.27	0	0.32	0.04	0	0.24	0	0.29	0.22	0	0.18	0.02	0.03	0.12	0.08	0.05	0.08	0.05	0.03	0	0	0.01	0.06	0.03	0	0.13	0	0.08
2	0.54	1	0.02	0.50	0.05	0.14	0.48	0.11	0.29	0.50	0.02	0.25	0.05	0.05	0.27	0.15	0.06	0.14	0.06	0.09	0.06	0.18	0.09	0.08	0.09	0.08	0.17	0.06	0.22
3	0	0.01	1	0	0.04	0.04	0	0	0	0	0	0	0.06	0.09	0.03	0.03	0	0	0.06	0.02	0.04	0	0	0	0.13	0	0	0.02	0.02
4	0.82	0.63	0	1	0.16	0.15	0.45	0.11	0.37	0.61	0	0.31	0.11	0.06	0.28	0.21	0.09	0.18	0.14	0.16	0.04	0.09	0.04	0.04	0.10	0.29	0.50	0.03	0.22
5	0.02	0.02	0.05	0.04	1	0.04	0.01	0	0	0	0	0	0.03	0.03	0.17	0.01	0	0	0.04	0	0.02	0.03	0.04	0.03	0	0.03	0.03	0	0.08
6	0	0.03	0.04	0.03	0.03	1	0	0	0	0	0	0	0.02	0.02	0.03	0.03	0	0	0.05	0	0	0.03	0.01	0.03	0	0.27	0.01	0.06	0.11
7	0.15	0.15	0	0.11	0.01	0	1	0	0.02	0.10	0.03	0.18	0.01	0.03	0.12	0.06	0.02	0	0.04	0.01	0.02	0	0.05	0.03	0	0	0.03	0.05	0.03
8	0	0.02	0	0.02	0	0	0	1	0	0	0.07	0	0.04	0	0	0.04	0	0	0.08	0.03	0	0.09	0	0.09	0	0	0.03	0	0
9	0.21	0.10	0	0.10	0	0	0.02	0	1	0.06	0.02	0	0.06	0.09	0.03	0.09	0	0	0.05	0.06	0	0	0.02	0.08	0	0.08	0.04	0.04	0.05
10	0.14	0.16	0	0.15	0	0	0.10	0	0.06	1	0.07	0.08	0.05	0	0.06	0.08	0.04	0	0.01	0.03	0	0	0.05	0	0.05	0	0.05	0.01	0.01
11	0	0.01	0	0	0	0	0.06	0.19	0.04	0.13	1	0.17	0.10	0.12	0.06	0.13	0.12	0.09	0.08	0.10	0	0.02	0.21	0	0.19	0	0.01	0.11	0.04
12	0.10	0.07	0	0.07	0	0	0.16	0	0	0.08	0.09	1	0.01	0.17	0.05	0.08	0.34	0	0.02	0.01	0	0	0.01	0	0	0	0.07	0	0.02
13	0.04	0.06	0.30	0.11	0.10	0.08	0.04	0.25	0.20	0.20	0.23	0.05	1	0.12	0.04	0.21	0.14	0.26	0.19	0.25	0.18	0.04	0.32	0.13	0.23	0.10	0.20	0.29	0.13
14	0.04	0.03	0.21	0.03	0.05	0.04	0.06	0	0.15	0	0.12	0.35	0.06	1	0.05	0.15	0.30	0.07	0.08	0.19	0	0.06	0.05	0.01	0.29	0.04	0.07	0.14	0.09
15	0.19	0.21	0.09	0.17	0.42	0.11	0.30	0	0.06	0.16	0.09	0.14	0.03	0.07	1	0.13	0.03	0.29	0.08	0.03	0.10	0.08	0.06	0.03	0.04	0.06	0.08	0.14	0.11
16	0.36	0.35	0.31	0.39	0.05	0.29	0.47	0.47	0.62	0.58	0.54	0.64	0.40	0.59	0.38	1	0.50	0.36	0.31	0.33	0.17	0.39	0.45	0.38	0.35	0.13	0.46	0.70	0.30
17	0.16	0.10	0.42	0.19	0.21	0.33	0.21	0.66	0.22	0.03	0.23	0.15	0.26	0.22	0.17	0.22	0.27	0.36	1	0.17	0.42	0	0.21	0.27	0.11	0.36	0.24	0.40	0.30
18	0.04	0.02	0	0.03	0	0	0.02	0	0	0.05	0.08	0.45	0.04	0.19	0.02	0.08	1	0	0.06	0.02	0.12	0	0.07	0.04	0	0.01	0.06	0	0.04
19	0.04	0.04	0	0.04	0	0	0	0	0	0	0.05	0	0.06	0.03	0.11	0.05	0	1	0.06	0.02	0	0.11	0.04	0.03	0	0	0.04	0	0.05
20	0.05	0.07	0.06	0.09	0	0	0.02	0.11	0.12	0.08	0.13	0.04	0.15	0.25	0.02	0.10	0.04	0.05	0.08	1	0	0.13	0.20	0.03	0.77	0	0.09	0.09	0.11
21	0	0.01	0.03	0.01	0.01	0	0.01	0	0	0	0	0	0.03	0	0.03	0.01	0.06	0	0.05	0	1	0.04	0	0.03	0	0	0	0.02	0.07
22	0	0.05	0	0.02	0.03	0.04	0	0.13	0	0	0.01	0	0.01	0.03	0.03	0.05	0	0.11	0	0.05	0.06	1	0	0.20	0	0.10	0.04	0.02	0.10
23	0.02	0.08	0	0.03	0.11	0.03	0.14	0	0.05	0.15	0.33	0.02	0.23	0.08	0.06	0.17	0.16	0.12	0.11	0.23	0	0	1	0.08	0.24	0.02	0.13	0.11	0.11
24	0.03	0.02	0	0.01	0.03	0.04	0.03	0.13	0.07	0	0	0	0.03	0.01	0.01	0.05	0.03	0.03	0.05	0.01	0.04	0.21	0.03	1	0	0	0.05	0	0.07
25	0.03	0.04	0.21	0.03	0	0	0	0	0	0.06	0.14	0	0.08	0.20	0.02	0.06	0	0	0.03	0.41	0	0	0.11	0	1	0	0.03	0.07	0.02
26	0	0.03	0	0.07	0.03	0.38	0	0	0.07	0	0	0	0.03	0.02	0.03	0.02	0.01	0	0.07	0	0	0.11	0.01	0	0	1	0.06	0	0.09
27	0.25	0.16	0	0.38	0.09	0.03	0.08	0.13	0.11	0.15	0.02	0.25	0.16	0.12	0.10	0.19	0.15	0.14	0.14	0.11	0	0.14	0.14	0.16	0.06	0.17	1	0	0.19
28	0	0.02	0.03	0.01	0	0.08	0.05	0	0.04	0.01	0.06	0	0.07	0.07	0.05	0.09	0	0	0.07	0.04	0.04	0.02	0.04	0	0.06	0	0	1	0.02
29	0.30	0.44	0.14	0.34	0.47	0.89	0.16	0	0.26	0.04	0.14	0.13	0.20	0.29	0.28	0.25	0.19	0.30	0.35	0.28	**0.67**	0.68	0.24	0.48	0.10	0.56	0.39	0.16	1
(c) **Misprediction**																													
1	0	0.07	0	0.09	0	0	0.09	0	0.09	0.06	0	0.06	0.01	0.01	0.04	0.03	0.02	0.01	0.02	0.01	0.01	0	0	0	0.01	0.01	0.03	0.01	0.03
2	0.14	0	0	0.12	0	0.01	0.15	0.05	0.09	0.12	0	0.07	0.02	0.01	0.07	0.03	0.01	0.02	0.05	0.03	0.01	0.01	0.03	0	0.03	0.01	0.02	0.02	0.03
3	0	0	0	0	0.03	0.01	0	0	0	0	0.01	0	0.02	0.03	0.01	0.01	0.02	0.01	0	0.01	0.01	0	0	0	0.04	0	0	0.01	0.01
4	0.25	0.2	0	0	0.09	0.06	0.17	0.05	0.12	0.16	0.01	0.09	0.04	0.02	0.15	0.08	0.06	0.03	0.07	0.05	0.01	0.03	0.03	0	0.03	0.13	0.15	0.01	0.1
5	0.02	0.02	0.03	0.03	0	0.04	0.01	0	0	0	0	0	0.01	0.01	0.06	0.01	0.01	0	0	0	0	0.01	0.02	0	0	0.03	0.02	0.01	0.02
6	0	0.01	0.03	0.01	0.02	0	0	0	0	0	0	0	0.01	0	0	0.01	0.02	0	0	0	0.01	0.02	0.01	0	0	0.12	0	0.02	0.05
7	0.03	0.04	0	0.03	0	0	0	0	0.01	0.02	0.02	0.04	0	0.01	0.04	0.02	0.01	0.01	0	0	0.02	0	0.02	0	0	0	0.01	0.02	0.01
8	0	0.01	0	0.01	0	0	0	0	0	0	0.03	0	0.02	0	0	0.02	0.04	0	0	0.01	0	0.05	0	0	0	0	0.01	0	0
9	0.06	0.03	0.01	0.03	0	0	0	0	0	0.03	0.02	0.01	0.03	0.02	0.01	0.04	0.02	0	0	0.02	0	0.01	0.01	0	0	0.03	0.02	0.02	0.02
10	0.02	0.03	0	0.03	0	0	0.03	0	0.02	0	0.01	0.01	0	0	0.01	0.01	0	0.01	0	0.01	0	0	0.01	0	0.01	0	0.01	0	0
11	0	0.01	0	0	0	0.01	0.02	0.08	0.02	0.04	0	0.04	0.02	0.01	0.01	0.03	0.02	0.04	0.03	0.03	0	0	0.05	0	0.05	0	0.01	0.05	0.01
12	0.02	0.01	0	0.01	0	0	0.05	0	0	0.02	0.02	0	0.01	0.04	0.01	0.02	0.01	0.1	0	0.01	0	0	0	0	0	0	0.02	0	0.01
13	0.03	0.04	0.18	0.05	0.02	0.04	0	0.11	0.09	0.06	0.09	0.02	0	0.07	0.02	0.1	0.1	0.05	0.12	0.1	0.09	0.03	0.17	0	0.09	0.04	0.08	0.11	0.08
14	0	0	0.09	0	0	0.01	0.01	0	0.03	0	0.01	0.08	0.01	0	0.01	0.03	0.02	0.1	0.01	0.06	0	0.01	0.02	0	0.1	0	0.02	0.04	0.02
15	0.08	0.08	0.04	0.08	0.22	0.02	0.1	0	0.01	0.05	0.03	0.05	0.02	0.02	0	0.04	0.03	0.01	0.11	0.02	0.02	0.01	0.01	0	0.01	0.02	0.03	0.05	0.03
16	0.12	0.11	0.16	0.13	0.07	0.08	0.14	0.2	0.24	0.17	0.21	0.19	0.18	0.19	0.15	0	0.14	0.18	0.16	0.13	0.07	0.18	0.17	0	0.12	0.07	0.16	0.26	0.12
17	0.04	0.05	0.24	0.06	0.15	0.17	0.08	0.29	0.06	0.03	0.1	0.06	0.14	0.1	0.1	0.1	0	0.12	0.15	0.07	0.21	0.05	0.11	0	0.04	0.17	0.08	0.14	0.17
18	0.01	0	0	0	0	0	0.01	0	0	0.02	0.03	0.12	0.02	0.06	0.01	0.03	0.02	0	0	0.01	0.06	0	0.03	0	0.01	0	0.02	0	0.02
19	0.01	0.02	0.01	0.02	0	0	0	0	0	0	0.02	0	0.03	0.01	0.04	0.02	0.03	0.01	0	0	0	0.03	0.02	0	0	0	0.01	0	0.02
20	0.01	0.01	0.02	0.01	0	0	0	0.05	0.03	0.03	0.06	0.01	0.04	0.06	0	0.03	0.02	0.01	0	0	0	0.02	0.06	0	0.26	0	0.02	0.04	0.02
21	0	0	0.01	0	0	0	0	0	0	0	0	0	0.01	0	0.01	0	0.01	0.02	0	0	0	0.01	0	0	0	0	0	0.01	0.02
22	0	0.01	0	0.01	0.03	0.03	0	0.05	0	0	0	0	0	0.01	0	0.01	0	0	0.02	0.01	0.01	0	0	0	0	0.04	0.02	0.02	0.02
23	0.01	0.03	0.01	0.03	0.02	0.02	0.03	0	0.02	0.07	0.12	0.01	0.12	0.03	0.01	0.06	0.05	0.06	0.05	0.1	0.01	0.01	0	0	0.1	0.02	0.05	0.04	0.05
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0.01	0.01	0.1	0.01	0	0	0	0	0	0.02	0.05	0	0.03	0.06	0.01	0.02	0.01	0	0	0.14	0	0.01	0.05	0	0	0	0	0.03	0.01
26	0	0.01	0.01	0.03	0.03	0.15	0.01	0	0.02	0.01	0	0.01	0.01	0.01	0.01	0	0.03	0	0	0	0.02	0.04	0	0	0	0	0.03	0	0.04
27	0.06	0.03	0	0.09	0.02	0	0.02	0.05	0.02	0.03	0.02	0.07	0.05	0.04	0.04	0.06	0.03	0.06	0.06	0.04	0	0.08	0.04	0	0.02	0.03	0	0	0.05
28	0	0	0.02	0	0	0	0.02	0	0.01	0.01	0.02	0	0.02	0.02	0.01	0.02	0.02	0	0	0.01	0.01	0.01	0.01	0	0.02	0	0	0	0
29	0.08	0.14	0.06	0.12	0.29	0.34	0.03	0.01	0.1	0.05	0.09	0.04	0.11	0.13	0.12	0.12	0.21	0.11	0.11	0.1	0.36	0.28	0.13	0	0.04	0.26	0.15	0.08	0
30	0.12	0.2	0.3	0.18	0.22	0.31	0.11	0.28	0.17	0.09	0.23	0.11	0.27	0.26	0.24	0.25	0	0.26	0.28	0.18	0.32	0.42	0.26	0	0.09	0.34	0.26	0.23	0

Predicted Class ID ↓

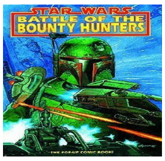
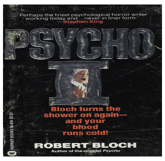
Table 4.9: Nonfiction genre indistinguishability analysis

Class ID ↓→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	30
(a) **Sample Count**																													
1	748	169	0	222	41	0	12	0	306	14	0	12	89	52	97	44	14	10	0	16	0	95	0	369	14	0	65	37	157
2	169	1444	9	331	203	42	0	143	53	34	103	12	512	237	165	109	26	57	58	58	270	408	125	236	14	8	162	28	259
3	0	9	614	0	46	197	0	0	0	24	0	0	197	303	38	9	0	0	0	34	44	84	0	58	227	0	8	0	107
4	222	331	0	878	41	0	8	28	84	170	0	31	214	140	82	52	37	46	0	9	0	152	0	340	61	9	315	0	109
5	41	203	46	41	806	0	0	17	19	10	47	17	308	274	298	50	0	25	24	16	8	216	11	176	19	0	19	16	187
6	0	42	197	0	0	483	0	0	11	0	0	0	165	183	11	0	8	12	0	0	0	69	0	201	28	9	0	0	73
7	12	0	0	8	0	0	494	0	72	0	9	183	38	26	26	38	41	9	0	11	0	185	12	29	89	0	8	104	111
8	0	143	0	28	17	0	0	439	45	0	110	8	52	44	8	19	0	34	0	32	0	290	0	0	47	8	0	0	12
9	306	53	0	84	19	11	72	45	846	0	0	66	197	132	37	39	11	0	76	17	9	174	0	361	9	55	41	122	196
10	14	34	24	170	10	0	0	0	0	889	119	142	16	521	33	25	149	84	0	193	0	222	225	54	506	9	80	0	57
11	0	103	0	0	47	0	9	110	0	119	896	0	210	398	85	216	12	68	47	71	33	223	282	39	133	12	0	9	204
12	12	12	0	31	17	0	183	8	66	142	0	949	17	381	0	11	456	0	0	92	0	400	11	221	506	16	0	0	84
13	89	512	197	214	308	165	38	52	197	16	210	17	**3297**	1016	250	556	128	258	383	249	412	478	198	590	9	84	**266**	369	1198
14	52	237	303	140	274	183	26	44	132	521	398	381	1016	3417	295	594	353	254	103	532	241	632	369	597	783	24	255	92	581
15	97	165	38	82	298	11	26	8	37	33	85	0	250	295	1112	97	0	37	39	12	104	149	11	77	33	52	40	147	405
16	44	109	9	52	50	0	38	19	39	25	216	11	556	594	97	1448	34	236	93	126	102	207	265	66	17	0	96	204	575
17	0	58	0	0	24	0	0	0	76	0	47	0	383	103	39	93	44	41	744	55	126	103	87	58	9	81	0	221	383
18	14	26	0	37	0	8	41	0	11	149	12	456	128	353	0	34	886	0	44	143	9	241	20	318	361	0	0	0	179
19	10	57	0	46	25	12	9	34	0	84	68	0	258	254	37	236	0	762	41	52	0	315	0	32	27	0	25	37	289
20	16	58	34	9	16	0	11	32	17	193	71	92	249	532	12	126	143	52	55	1122	0	135	205	93	416	12	0	36	147
21	0	270	44	0	8	0	0	0	9	0	33	0	412	241	104	102	9	0	126	0	735	104	42	42	0	11	8	67	278
22	95	408	84	152	216	69	185	290	174	222	223	400	478	632	149	207	241	315	103	135	104	2495	171	327	462	129	149	72	107
23	0	125	0	0	11	0	12	0	0	225	282	11	198	369	11	265	20	0	87	205	42	171	877	70	273	12	0	43	162
24	369	236	58	340	176	201	29	0	361	54	39	221	590	597	77	66	318	32	58	93	42	327	70	1809	148	0	188	41	207
25	14	14	227	61	19	28	89	47	9	506	133	506	9	783	33	17	361	27	9	416	0	462	273	148	1519	25	16	9	63
26	0	8	0	9	0	9	0	8	55	9	12	16	84	24	52	0	0	0	81	12	11	129	12	0	25	424	0	69	185
27	65	162	8	315	19	0	8	0	41	80	0	0	**266**	255	40	96	0	25	0	0	8	149	0	188	16	0	**659**	0	151
28	37	28	0	0	16	0	104	0	122	0	9	0	369	92	147	204	0	37	221	36	67	72	43	41	9	69	0	888	543
30	157	259	107	109	187	73	111	12	196	57	204	84	1198	581	405	575	179	289	383	147	278	107	162	207	63	185	151	543	2802
(b) **Cooccurrence**																													
1	1	0.12	0	0.25	0.05	0	0.02	0	0.36	0.02	0	0.01	0.03	0.02	0.09	0.03	0.02	0.01	0	0.01	0	0.04	0	0.20	0.01	0	0.10	0.04	0.06
2	0.23	1	0.01	0.38	0.25	0.09	0	0.33	0.06	0.04	0.11	0.01	0.16	0.07	0.15	0.08	0.03	0.07	0.08	0.05	0.37	0.16	0.14	0.13	0.01	0.02	0.25	0.03	0.09
3	0	0.01	1	0	0.06	0.41	0	0	0	0.03	0	0	0.06	0.09	0.03	0.01	0	0	0	0.03	0.06	0.03	0	0.03	0.15	0	0.01	0	0.04
4	0.30	0.23	0	1	0.05	0	0.02	0.06	0.10	0.19	0	0.03	0.06	0.04	0.07	0.04	0.04	0.06	0	0.01	0	0.06	0	0.19	0.04	0.02	0.48	0	0.04
5	0.05	0.14	0.07	0.05	1	0	0	0.04	0.02	0.01	0.05	0.02	0.09	0.08	0.27	0.03	0	0.03	0.03	0.01	0.01	0.09	0.01	0.10	0.01	0	0.03	0.02	0.07
6	0	0.03	0.32	0	0	1	0	0	0.01	0	0	0	0.05	0.05	0.01	0	0.01	0.02	0	0	0	0.03	0	0.11	0.02	0.02	0	0	0.03
7	0.02	0	0	0.01	0	0	1	0	0.09	0	0.01	0.19	0.01	0.01	0.02	0.03	0.05	0.01	0	0.01	0	0.07	0.01	0.02	0.06	0	0.01	0.12	0.04
8	0	0.10	0	0.03	0.02	0	0	1	0.05	0	0.12	0.01	0.02	0.01	0.01	0.01	0	0.04	0	0.03	0	0.12	0	0	0.03	0.02	0	0	0
9	0.41	0.04	0	0.10	0.02	0.02	0.15	0.10	1	0	0	0.07	0.06	0.04	0.03	0.03	0.01	0	0.10	0.02	0.01	0.07	0	0.20	0.01	0.13	0.06	0.14	0.07
10	0.02	0.02	0.04	0.19	0.01	0	0	0	0	1	0.13	0.15	0	0.15	0.03	0.02	0.17	0.11	0	0.17	0	0.09	0.26	0.03	0.33	0.02	0.12	0	0.02
11	0	0.07	0	0	0.06	0	0.02	0.25	0	0.13	1	0	0.06	0.12	0.08	0.15	0.01	0.09	0.06	0.06	0.04	0.09	0.32	0.02	0.09	0.03	0	0.01	0.07
12	0.02	0.01	0	0.04	0.02	0	0.37	0.02	0.08	0.16	0	1	0.01	0.11	0	0.01	0.51	0	0	0.08	0	0.16	0.01	0.12	0.33	0.04	0	0	0.03
13	0.12	0.35	0.32	0.24	0.38	0.34	0.08	0.12	0.23	0.02	0.23	0.02	1	0.30	0.22	0.38	0.14	0.34	0.51	0.22	0.56	0.19	0.23	0.33	0.01	0.20	**0.40**	0.42	0.43
14	0.07	0.16	0.49	0.16	0.34	0.38	0.05	0.10	0.16	0.59	0.44	0.40	0.31	1	0.27	0.41	0.40	0.33	0.14	0.47	0.33	0.25	0.42	0.33	0.52	0.06	0.39	0.10	0.21
15	0.13	0.11	0.06	0.09	0.37	0.02	0.05	0.02	0.04	0.04	0.09	0	0.08	0.09	1	0.07	0	0.05	0.05	0.01	0.14	0.06	0.01	0.04	0.02	0.12	0.06	0.17	0.14
16	0.06	0.08	0.01	0.06	0.06	0	0.08	0.04	0.05	0.03	0.24	0.01	0.17	0.17	0.09	1	0.04	0.31	0.13	0.11	0.14	0.08	0.30	0.04	0.01	0	0.15	0.23	0.21
17	0	0.04	0	0	0.03	0	0	0	0.09	0	0.05	0	0.12	0.03	0.04	0.06	0.05	0.05	1	0.05	0.17	0.04	0.10	0.03	0.01	0.19	0	0.25	0.14
18	0.02	0.02	0	0.04	0	0.02	0.08	0	0.01	0.17	0.01	0.48	0.04	0.10	0	0.02	1	0	0.06	0.13	0.01	0.10	0.02	0.18	0.24	0	0	0	0.06
19	0.01	0.04	0	0.05	0.03	0.02	0.02	0.08	0	0.09	0.08	0	0.08	0.07	0.03	0.16	0	1	0.06	0.05	0	0.13	0	0.02	0.02	0	0.04	0.04	0.10
20	0.02	0.04	0.06	0.01	0.02	0	0.02	0.07	0.02	0.22	0.08	0.10	0.08	0.16	0.01	0.09	0.16	0.07	0.07	1	0	0.05	0.23	0.05	0.27	0.03	0	0.04	0.05
21	0	0.19	0.07	0	0.01	0	0	0	0.01	0	0.04	0	0.12	0.07	0.09	0.07	0.01	0	0.17	0	1	0.04	0.05	0.02	0	0.03	0.01	0.08	0.10
22	0.13	0.28	0.14	0.17	0.27	0.14	0.37	0.66	0.21	0.25	0.25	0.42	0.14	0.18	0.13	0.14	0.27	0.41	0.14	0.12	0.14	1	0.19	0.18	0.30	0.30	0.23	0.08	0.04
23	0	0.09	0	0	0.01	0	0.02	0	0	0.25	0.31	0.01	0.06	0.11	0.01	0.18	0.02	0	0.12	0.18	0.06	0.07	1	0.04	0.18	0.03	0	0.05	0.06
24	0.49	0.16	0.09	0.39	0.22	0.42	0.06	0	0.43	0.06	0.04	0.23	0.18	0.17	0.07	0.05	0.36	0.04	0.08	0.08	0.06	0.13	0.08	1	0.10	0	0.29	0.05	0.07
25	0.02	0.01	0.37	0.07	0.02	0.06	0.18	0.11	0.01	0.57	0.15	0.53	0	0.23	0.03	0.01	0.41	0.04	0.01	0.37	0	0.19	0.31	0.08	1	0.06	0.02	0.01	0.02
26	0	0.01	0	0.01	0	0.02	0	0.02	0.07	0.01	0.01	0.02	0.03	0.01	0.05	0	0	0	0.11	0.01	0.01	0.05	0.01	0	0.02	1	0	0.08	0.07
27	0.09	0.11	0.01	0.36	0.02	0	0.02	0	0.05	0.09	0	0	0.08	0.07	0.04	0.07	0	0.03	0	0	0.01	0.06	0	0.10	0.01	0	1	0	0.05
28	0.05	0.02	0	0	0.02	0	0.21	0	0.14	0	0.01	0	0.11	0.03	0.13	0.14	0	0.05	0.30	0.03	0.09	0.03	0.05	0.02	0.01	0.16	0	1	0.19
30	0.21	0.18	0.17	0.12	0.23	0.15	0.22	0.03	0.23	0.06	0.23	0.09	0.36	0.17	0.36	0.40	0.20	0.38	0.51	0.13	0.38	0.04	0.18	0.11	0.04	0.44	0.23	0.61	1
(c) **Misprediction**																													
1	0	0.03	0	0.07	0.02	0	0.03	0	0.11	0	0	0	0.01	0	0.04	0.01	0.01	0	0	0	0	0.01	0	0	0	0.01	0.02	0.02	0.02
2	0.08	0	0.01	0.1	0.08	0	0	0.14	0.02	0	0.03	0	0.04	0.01	0.05	0.02	0.03	0	0.02	0.02	0.1	0.05	0.04	0	0	0.02	0.06	0.02	0.02
3	0	0	0	0	0.01	0.08	0.01	0	0	0.01	0	0.02	0.01	0.02	0.01	0	0.01	0	0	0	0.01	0	0	0	0.04	0	0	0	0
4	0.14	0.07	0	0	0.04	0.01	0.02	0.02	0.05	0.04	0	0	0.03	0.02	0.04	0.01	0.01	0.01	0.02	0.01	0	0.03	0	0	0.01	0.02	0.13	0.02	0.02
5	0.02	0.06	0.03	0.02	0	0.01	0	0.04	0.01	0.01	0.02	0.01	0.04	0.03	0.11	0.02	0.01	0	0.02	0	0	0.04	0	0	0.01	0.01	0.01	0.01	0.03
6	0.01	0	0.07	0	0	0	0	0	0.01	0	0	0	0.02	0.01	0	0.01	0.01	0	0	0	0.02	0	0	0	0.01	0	0	0	0.01
7	0	0	0.01	0	0	0	0	0.03	0.02	0	0	0.02	0	0	0.02	0.01	0.01	0.01	0	0	0	0.01	0.01	0	0	0.03	0	0.04	0.02
8	0	0.02	0	0.01	0.01	0	0	0	0.02	0	0.04	0	0	0.01	0	0.01	0	0	0.01	0.01	0	0.03	0	0	0	0.01	0	0	0
9	0.16	0.01	0	0.02	0.01	0.01	0.04	0.03	0	0	0	0.02	0.02	0.01	0.03	0.01	0.03	0.01	0	0.01	0.01	0.02	0	0	0	0.03	0.02	0.07	0.03
10	0.01	0.01	0.01	0.04	0	0	0.01	0	0	0	0.03	0.03	0	0.03	0	0.01	0	0.04	0.02	0.04	0.01	0.03	0.08	0	0.08	0	0.04	0	0.01
11	0.01	0.02	0	0	0.01	0	0	0.08	0.01	0.03	0	0	0.01	0.03	0.02	0.03	0	0	0.01	0.03	0.02	0.03	0.1	0	0.02	0	0.01	0	0.01
12	0	0	0.03	0.01	0	0	0.08	0.01	0.03	0.06	0.01	0	0.01	0.03	0	0	0	0.12	0	0.02	0	0.04	0.01	0	0.09	0	0.01	0	0.01
13	0.07	0.15	0.17	0.1	0.17	0.2	0.08	0.08	0.11	0	0.08	0.02	0	0.14	0.12	0.18	0.23	0.06	0.19	0.12	0.26	0.12	0.07	0	0.03	0.15	0.14	0.16	0.02
14	0.02	0.1	0.25	0.07	0.15	0.22	0.12	0.1	0.06	0.21	0.17	0.23	0.18	0	0.09	0.15	0.12	0.21	0.15	0.24	0.17	0.15	0.15	0	0.27	0.13	0.12	0.05	0.14
15	0.03	0.03	0.03	0.04	0.13	0.01	0.03	0	0.01	0.02	0.03	0	0.02	0.02	0	0.01	0.01	0	0.02	0	0.03	0.02	0	0	0.02	0.05	0.03	0.04	0.03
16	0.04	0.05	0.01	0.05	0.03	0.01	0.01	0.02	0.04	0.02	0.08	0	0.06	0.07	0.02	0	0.04	0	0.13	0.06	0.04	0.03	0.09	0	0.02	0.02	0.07	0.06	0.08
17	0	0.02	0	0	0	0.01	0.01	0	0.02	0	0.02	0	0.03	0.01	0.01	0.02	0	0	0.01	0.02	0.05	0.01	0.03	0	0	0.04	0.01	0.06	0.03
18	0.01	0.01	0.01	0.01	0	0.02	0.02	0	0	0.03	0	0.11	0.01	0.03	0	0	0.02	0	0	0.04	0	0.03	0.01	0	0.06	0	0	0	0.01
19	0.01	0.03	0.01	0.02	0.01	0.01	0.02	0.06	0	0.02	0.02	0.01	0.04	0.03	0.01	0.07	0.01	0.01	0	0.01	0	0.06	0	0	0.01	0	0.03	0.03	0.05
20	0	0.03	0.03	0.01	0	0.01	0.02	0.04	0.01	0.08	0.03	0.04	0.05	0.07	0	0.05	0.05	0.07	0.02	0	0.01	0.02	0.09	0	0.09	0.05	0.01	0.03	0.03
21	0	0.06	0.02	0.01	0.01	0	0	0	0	0	0.01	0	0.05	0.02	0.03	0.02	0.1	0	0	0	0	0.02	0.02	0	0	0.04	0.01	0.02	0.03
22	0.03	0.11	0.05	0.11	0.08	0.04	0.13	0.3	0.1	0.1	0.11	0.12	0.06	0.07	0.04	0.07	0.03	0.08	0.18	0.06	0.05	0	0.06	0	0.11	0.09	0.08	0.02	0.02
23	0	0.03	0	0	0	0	0.01	0	0	0.09	0.1	0.02	0.03	0.05	0	0.07	0.05	0.02	0.01	0.08	0.03	0.03	0	0	0.08	0.01	0	0.01	0.02
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04
25	0.01	0.02	0.16	0.04	0.01	0.07	0.08	0.02	0.01	0.21	0.05	0.21	0.02	0.11	0.02	0.01	0.01	0.21	0.03	0.12	0.02	0.09	0.12	0	0	0.02	0.04	0.02	0.02
26	0.01	0	0.01	0	0	0.02	0	0	0.02	0	0	0	0	0	0.03	0	0.01	0	0	0	0.01	0	0	0	0	0	0	0.02	0.02
27	0.02	0.02	0	0.11	0.02	0	0.03	0	0.02	0	0	0	0.03	0	0.03	0.01	0.01	0	0	0.01	0	0.01	0	0	0	0	0	0.01	0.01
28	0.03	0.01	0.01	0.02	0.01	0.02	0.09	0.03	0.07	0	0.01	0.01	0.03	0.01	0.05	0.04	0.04	0	0	0.01	0	0.01	0.02	0	0	0.04	0.01	0	0.06
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0.08	0.06	0.05	0.04	0.1	0.08	0.09	0.01	0.09	0.02	0.13	0.06	0.14	0.09	0.17	0.16	0.17	0.06	0.14	0.04	0.16	0.03	0.05	0	0.03	0.17	0.07	0.2	0

Predicted Class ID ↓

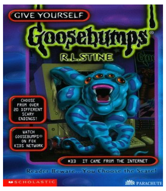
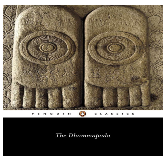
Table 4.10: Misclassification samples

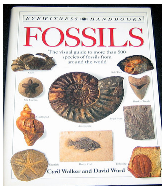
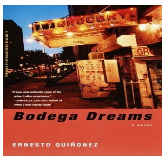
(a) Misclassification due to genre ambiguity

<table>
<tr><td colspan="3">***(a) 9780140296419:*** Gladiator; **Author:** Dewey Gram; **Publisher:** Penguin Putnam</td></tr>
<tr><td rowspan="2"> </td><td>**Actual genre**</td><td>Fiction: History, Mystery & Thriller & Suspense & Horror, Press & Media, Sci-Fi & Fantasy</td></tr>
<tr><td>**Predicted genre**</td><td>Fiction: History, Mystery & Thriller & Suspense & Horror, Sci-Fi & Fantasy</td></tr>
<tr><td colspan="3">***(b) 9781569711293:*** Star Wars: Battle of the Bounty Hunters; **Author:** Ryder Windham; **Publisher:** Dark Horse Comics</td></tr>
<tr><td rowspan="2"> </td><td>**Actual genre**</td><td>Fiction: Comics & Graphics, Press & Media, Sci-Fi & Fantasy</td></tr>
<tr><td>**Predicted genre**</td><td>Fiction: Comics & Graphics, Sci-Fi & Fantasy</td></tr>
<tr><td colspan="3">***(c) 9780446908047:*** Psycho II; **Author:** Robert Bloch; **Publisher:** Warner Books</td></tr>
<tr><td rowspan="2">

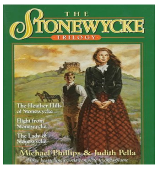
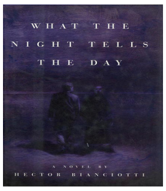
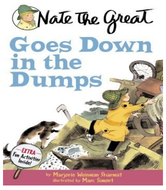
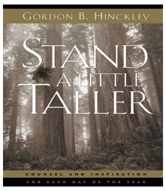
PSYCHO II
Bloch turns the shower on again—and your blood runs cold!
ROBERT BLOCH

</td><td>**Actual genre**</td><td>Fiction: Mystery & Thriller & Suspense & Horror</td></tr>
<tr><td>**Predicted genre**</td><td>Fiction: Mystery & Thriller & Suspense & Horror, Sci-Fi & Fantasy</td></tr>
<tr><td colspan="3">***(d) 9780590516655:*** It Came from the Internet; **Author:** R.L. Stine; **Publisher:** Scholastic Inc.</td></tr>
</table>

	**Actual genre**	Fiction: Children's Book, Mystery & Thriller & Suspense & Horror, Sports & Outdoors
	Predicted genre	Fiction: Children's Book, Mystery & Thriller & Suspense & Horror, Sports & Outdoors, Sci-Fi & Fantasy
(e) 9780140442847: The Dhammapada: The Path of Perfection; **Author:** Anonymous; **Publisher:** Penguin Classics		
	**Actual genre**	Nonfiction: Literature, Mythology & Religion & Spirituality, Humanities
	Predicted genre	Nonfiction: Literature, Mythology & Religion & Spirituality, History
(f) 9780812561876: Cat in an Indigo Mood; **Author:** Carole Nelson Douglas; **Publisher:** Forge Books		
CAROLE NELSON DOUGLAS CAT IN AN INDIGO MOOD A MIDNIGHT LOUIE MYSTERY	**Actual genre**	Fiction: Animals & Wildlife & Pets, Mystery & Thriller & Suspense & Horror, Romance
	Predicted genre	Fiction: Mystery & Thriller & Suspense & Horror, Romance
(g) 9781564580719: Fossils; **Author:** David J. Ward; **Publisher:** DK ADULT		

	**Actual genre**	Nonfiction: Animals & Wildlife & Pets, History, Reference & Language, Science & Math & Mathematics
	Predicted genre	Nonfiction: History, Reference & Language, Science & Math & Mathematics
(h) 9780375705892: Elephant Man; **Author:** Christine Sparks; **Publisher:** Ballantine Books		
	**Actual genre**	Fiction: History, Literature, Teen & Young Adult
	Predicted genre	Fiction: History, Literature
(i) 9780345260680: Rocket Ship Galileo; **Author:** Robert A. Heinlein; **Publisher:** Del Rey		
	Actual genre	Fiction: Mystery & Thriller & Suspense & Horror, Sci-Fi & Fantasy, Teen & Young Adult
	Predicted genre	Fiction: Mystery & Thriller & Suspense & Horror, Sci-Fi & Fantasy
False positives are marked in red, and False negatives are marked in blue.		

(b) Misclassification due to level-1 misprediction

(a) 9780884861331: The Stonewycke Trilogy; **Author:** Michael R. Phillips, Judith Pella; **Publisher:** Bbs Pub Corp		
	**Actual genre**	Fiction: History, Mythology & Religion & Spirituality, Romance
	Predicted genre	Nonfiction: Sci-Fi & Fantasy
(b) 9781565842410: What the Night Tells the Day: A Novel; **Author:** Hector Bianciotti, Linda Coverdale; **Publisher:** The New Press		
 WHAT THE NIGHT TELLS THE DAY A NOVEL BY HECTOR BIANCIOTTI	**Actual genre**	Fiction: Fashion & Lifestyle
	Predicted genre	Nonfiction: History, Mythology & Religion & Spirituality
(c) 9780440404385: Nate the Great Goes Down in the Dumps; **Author:** Marjorie Weinman Sharmat, Marc Simont; **Publisher:** Yearling		
 Nate the Great Goes Down in the Dumps by Marjorie Weinman Sharmat illustrated by Marc Simont	**Actual genre**	Fiction: Arts & Photography, Children's Book, Mystery & Thriller & Suspense & Horror
	Predicted genre	Nonfiction: Humor & Entertainment, Sci-Fi & Fantasy
(d) 9781570087677: What the Night Tells the Day: A Novel; **Author:** Hector Bianciotti, Linda Coverdale; **Publisher:** The New Press		
 GORDON B. HINCKLEY STAND A LITTLE TALLER COUNSEL AND INSPIRATION FOR EACH DAY OF THE YEAR	**Actual genre**	Nonfiction: Mythology & Religion & Spirituality, Self-Help & Motivation
	Predicted genre	Fiction: History, Humanities, Literature, Biographies & Memoir

4.13 Genre-wise 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 behavior and guiding future improvements in dataset balancing and architectural tuning.

Table 4.12 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 4.13 provides a comprehensive performance breakdown for non-fiction 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 4.12: 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 4.13: 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

4.14 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* and *NonFiction*) 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 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 multi-modal classification by seamlessly integrating visual and textual elements.

Chapter 5

Book Genre Identification from Unreliable Reviews Refined 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 Nonfiction, 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 Nonfiction, and subsequently map it to its subgenre using the enriched textual features.

Mathematical Representation: We define the mapping as follows:

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

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

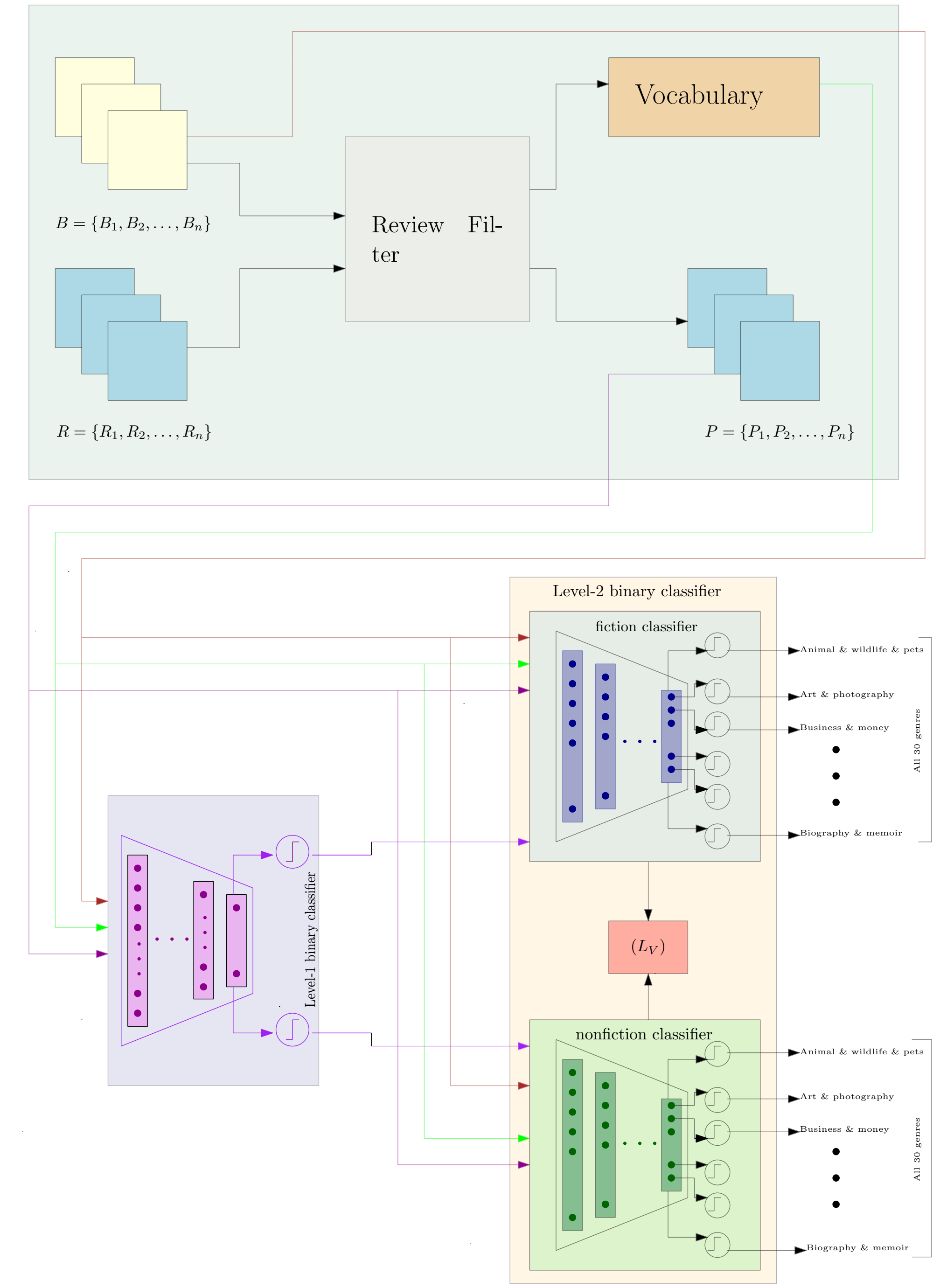


Figure 5.1: Proposed Architecture

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 $\mathcal{R}_i$:

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

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

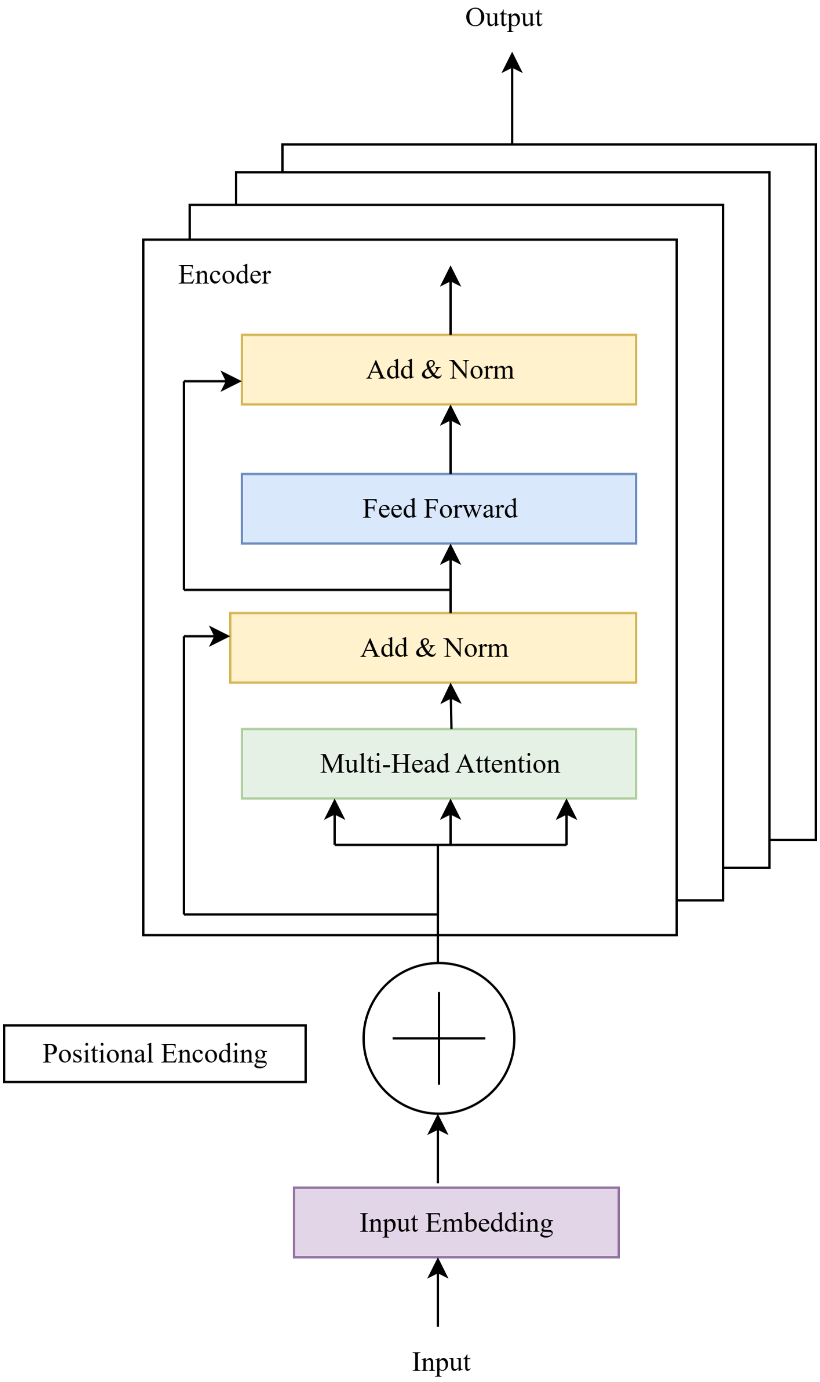


Figure 5.2: BERT encoder architecture

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:

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

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

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

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 and nonfiction.

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 Nonfiction, 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 Nonfiction 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 **Nonfiction**, it is processed by the *Nonfiction 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 Nonfiction, 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) [40]:

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

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 provides 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:

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

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 Experimental 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. Nonfiction) and Level-2 (sub-genre classification within Fiction and Nonfiction). Metrics used for evaluation include the Macro F1-Score ($\mathcal{F}_M$), Accuracy ($\mathcal{A}$), and Balanced Accuracy ($\mathcal{BA}$) for both Fiction and Nonfiction 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 Nonfiction 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: Results of hierarchical classification for book crowd source reviews and blurbs

Modality	**Model**	**Level-1**		**Level-2**							
				Fiction				**Nonfiction**			
		$\mathcal{FM}$	$\mathcal{A}$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$	$\mathcal{FM}_\mu$	$\mathcal{BA}_\mu$	$\mathcal{FM}_m$	$\mathcal{BA}_m$
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 [41]	92.36	90.85	51.08	72.71	45.46	70.12	48.68	72.60	44.48	72.72
	RoBERTa [42]	90.90	89.42	53.80	75.01	39.69	66.04	53.61	75.84	50.27	73.97
	XLNet [43]	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 [41]	92.25	90.91	52.45	73.26	45.48	68.81	57.31	76.48	39.47	68.14
	RoBERTa [42]	92.99	91.65	51.47	73.59	41.29	67.57	52.84	75.30	50.19	74.72
	XLNet [43]	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 Ablation Studies

5.4.1 Performance Analysis on Non-Augmented Data

Figure 5.3 presents the Level-1 results for BERT on the coarse *Fiction vs. Non-Fiction* classification under two settings: *non-augmented* and *augmented* training data. The results show a clear improvement with augmentation. On the non-augmented dataset, BERT achieves an F1 score of 83.27% and an accuracy of 80.81%. After applying augmentation, performance increases to 88.22% F1 and 91.86% accuracy, indicating that augmentation helps BERT learn more robust decision boundaries and reduces misclassification between the two high-level classes.

Figure 5.5 reports the Level-2 ablation results for *fiction* and *non-fiction* under *non-augmented* vs. *augmented* training, using FM-macro (macro F1) and BA-macro (macro

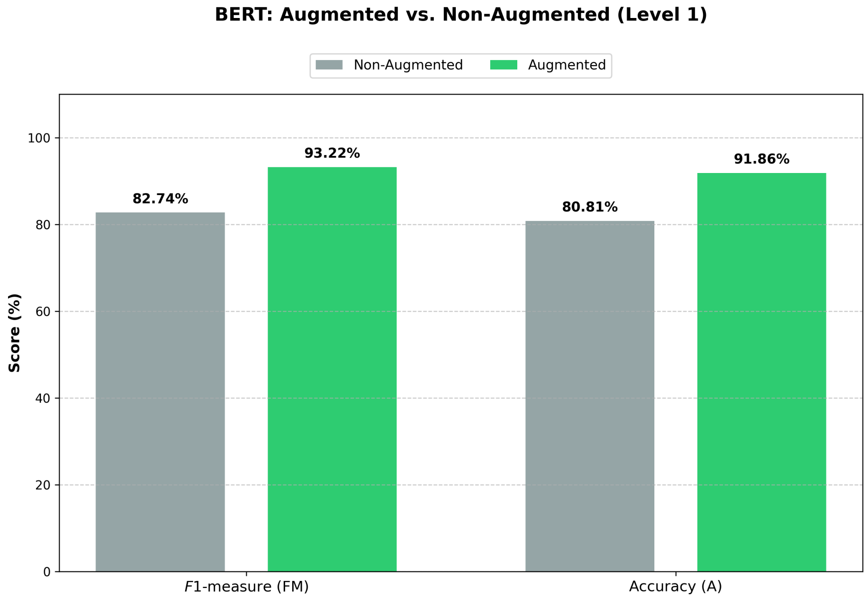


Figure 5.3: Level-1 performance analysis of BERT between augmented and non-augmented training.

balanced accuracy). For both classes, augmentation consistently improves the scores, yielding higher macro F1 and macro balanced accuracy. Overall, the gains suggest that augmentation strengthens feature learning at the finer level, reducing confusion among Level-2 categories and improving balanced performance across classes.

5.4.2 Performance on Flat Multi-label Classification

In this ablation, we compare flat (single-level) multi-label classification with the proposed multi-level (hierarchical) formulation at Level-2, reporting macro F1 (FM) and accuracy (A) for both fiction and non-fiction. The multi-level setup consistently performs better than the flat baseline for both classes, indicating that exploiting the hierarchical structure helps the model learn more discriminative representations. In particular, the gains in F1 suggest improved handling of class imbalance and fewer misclassifications among fine-grained labels, while the higher accuracy confirms an overall improvement in prediction correctness. Overall, these results support the effectiveness of multi-level classification over a flat approach for Level-2 categorization.

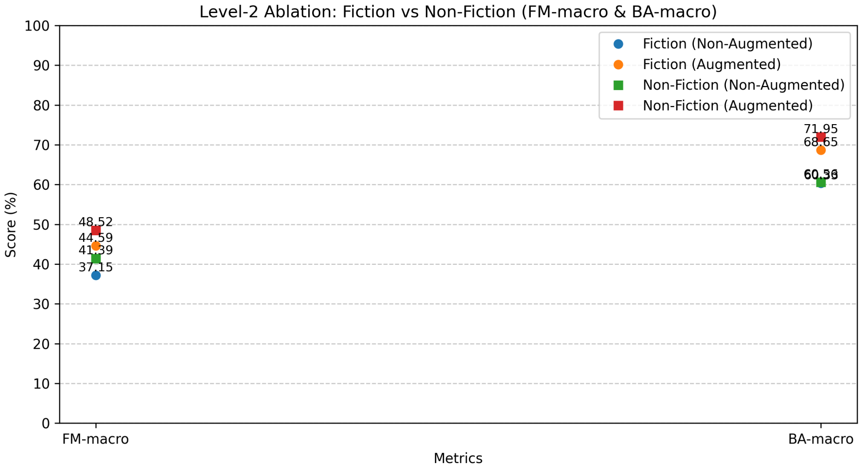


Figure 5.4: Level-2 performance analysis (FM-macro and BA-macro) for fiction and non-fiction under augmented and non-augmented training.

5.4.3 Inference Time

To evaluate the practical deployability of the proposed framework, we conducted a rigorous analysis of inference latency and model complexity. This evaluation is critical for understanding the trade-off between the high-fidelity genre mining achieved by transformer-based architectures and the real-time processing requirements of large-scale digital libraries.

Table 5.3: Inference Latency Analysis for Hierarchical Genre Classification

Model	**Parameters (Millions)**	**Latency per Level (ms)**		**Total Inference Time (ms)**
		Level-1 (Binary)	**Level-2 (Multi-label)**	
BERT[44]	110.0	132.4	198.6	331.0
BLIP [26]	224.0	232.4	348.6	581.0
DistilBERT [41]	66.0	78.5	117.8	196.3
RoBERTa-base [42]	125.0	141.2	211.8	353.0
XLNet-base [43]	110.0	165.5	248.3	413.8
ALBERT-base [39]	12.0	68.2	102.3	170.5

5.5 Summary

In this work, we proposed a robust framework for book genre classification leveraging textual information from user reviews and blurbs. Unlike traditional metadata-

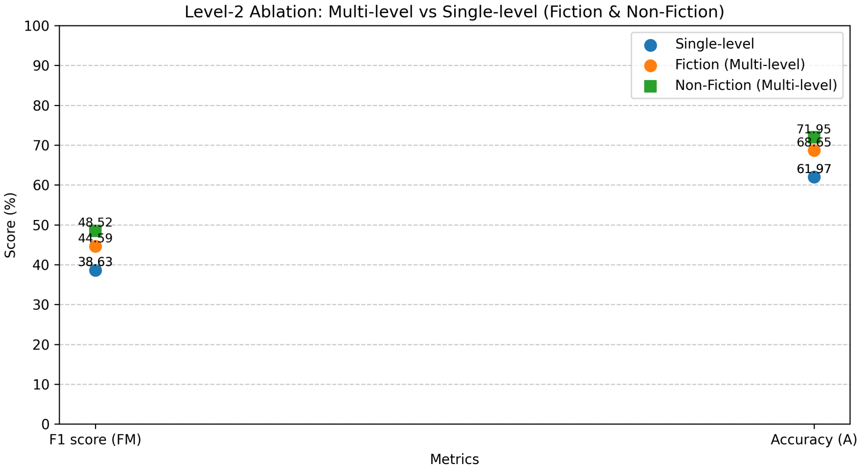


Figure 5.5: Level-2 performance analysis (FM-macro and BA-macro) for fiction and non-fiction under augmented and non-augmented training.

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 Nonfiction genres, followed by finer subgenre classification.

The experimental results demonstrate that integrating user perceptions and narrative-driven insights from reviews significantly enhances genre prediction accuracy. This work sets a foundation for multi-modal 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. Nonfiction) 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 Multimedia (Under review)
   https://arxiv.org/html/2505.03839v1

2. Suraj Kumar, Utsav Kumar Nareti, Soumi Chattopadhyay, Chandranath Adak, Prolay Mallick have submitted a research paper to an A* ranked international conference (communicated)
   https://arxiv.org/abs/2512.21076

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 multi-modal 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] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011.

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

[21] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

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

[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] 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.

[41] 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.

[42] 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.

[43] 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.

[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.