INTERNATIONAL STUDIES IN THE FIELD OF

COMPUTER ENGINEERING

EDITOR

Prof. Dr. Selahattin BARDAK

DECEMBER 2025

DECEMBER 2025

gece kitaplığı

İmtiyaz Sahibi / Yaşar Hız
Yayına Hazırlayan / Gece Kitaplığı

Birinci Basım / Aralık 2025 - Ankara
ISBN / 978-625-8570-52-6

Gece Kitaplığı
Kızılay Mah. Fevzi Çakmak 1. Sokak
Ümit Apt No: 22/A Çankaya/ANKARA
0312 384 80 40
www.gecekitapligi.com / gecekitapligi@gmail.com

Baskı & Cilt
Bizim Büro
Sertifika No: 42488

INTERNATIONAL STUDIES IN THE FIELD OF COMPUTER ENGINEERING

ARALIK 2025

EDITOR

Prof. Dr. Selahattin BARDAK

gece kitaplığı

CONTENTS

CHAPTER 1

ENSEMBLE LEARNING MODELS FOR ANALYZING FACTORS INFLUENCING MATHEMATICS ACHIEVEMENT IN PISA 2022 TÜRKIYE

Büşra Tuğçe SUSAM[1]

1 Assistant Professor, Hakkari University, Engineering Faculty, Department of Computer Engineering, busratugcesusam@hakkari.edu.tr, 0000-0001-8250-2044.

1. INTRODUCTION

Academic achievement is a crucial outcome metric in educational systems, and identifying its determinants is essential in constructing coherent, research-driven educational policies. Student performance is affected by individual abilities, school-level characteristics, family history, household setting, and behavioral indications such as absenteeism (Rakesh et al., 2025). Comprehending these related variables is essential for advancing equity and enhancing educational results across diverse populations. Global assessment programs like the Programme for International Student Assessment (PISA) provide vast cross-national data sets, offering a chance to enhance our comprehension of how sociocultural, economic, and contextual factors effect learning results(*PISA 2022 Database | OECD*, n.d.). These data sets facilitate comparative analysis and policy development by correlating student performance with underlying factors such as socioeconomic position, familial support, and access to educational resources.

Many studies indicate that socioeconomic status (SES) and the home environment are among the most significant determinants of academic achievement(Chmielewski, 2019; Guevara-Reyes et al., 2025; Jin, 2023; Liu et al., 2024; Sirin, 2005). Students from higher socioeconomic backgrounds typically possess advantages derived from elevated parental education levels, professional occupational status, and enhanced access to material and cultural capital, all of which collectively foster greater cognitive development and sustained academic engagement(Chmielewski, 2019; Jin, 2023). In contrast, students from disadvantaged socioeconomic circumstances sometimes face educational disadvantages due to constrained learning opportunities, diminished parental participation, and limited access to technology and cultural resources(Davis-Kean, 2005; Jerrim & Macmillan, 2015).

Moreover, student behavioral factors, particularly absenteeism, are established predictors of lower academic achievement, diminished motivation, and reduced social engagement. Chronic absenteeism is consistently associated with lower academic performance, reduced motivation, and weaker social integration within school settings(Gottfried, 2014; Kearney, 2008). The consequences of absenteeism are significantly intensified by familial and home circumstances, as students from lower socioeconomic backgrounds encounter supplementary obstacles to consistent attendance including insufficient transportation, minimal parental oversight, or conflicting household responsibilities that further aggravate educational disparities(Klein et al., 2020; Sosu et al., 2021).

In recent years, the application of machine learning (ML) methods in educational research has expanded, providing powerful tools to analyze complex and nonlinear relationships among multiple variables. Unlike traditional statistical techniques such as regression analysis, ML models can flexibly capture multidimensional patterns within large-scale datasets(Chen & Ding, 2023). Recent research has showed that classification and ensemble-based algorithms can effectively predict academic achievement and reveal the relative relevance of contributing factors, resulting in more precise and evidence-based policy making(Guevara-Reyes et al., 2025). Despite all of these developments, existing ML literature is limited by the absence of nationally representative datasets and the insufficient incorporation of comprehensive family, household, and behavioral factors.

The aim of this study is to employ machine learning techniques on the Türkiye subsample of PISA 2022 to investigate how familial background, household context, digital access indicator and absenteeism influence mathematics achievement. Specifically, first students were categorized into high- and low-achievement groups, and multiple classifiers including Random Forest, ExtraTrees and Gradient boosting were trained and optimized through hyperparameter tuning to assess predictive performance and the relative importance of the features. The results contribute to the growing body of evidence on how socioeconomic status, home environment, absenteeism and digital access indicators shape learning outcomes among Turkish students, underscoring the potential of ML methodologies as powerful tools for generating policy-relevant insights in educational research.

2. METHODOLOGY

2.1 Data Set and Preprocessing

The dataset used in this research is from the Türkiye subsample of PISA 2022, which is administered by the Organisation for Economic Cooperation and Development (OECD). PISA assesses 15-year-old students' knowledge and skills in mathematics, reading, and science from participating nations. Turkey samples of PISA data set comprises cognitive performance scores, detailed background questionnaires, and contextual factors that represent pupils' socioeconomic position, home resources, digital access, and school-related behaviors. The analysis focused on students who completed the mathematics assessment and provided valid responses to the selected socioeconomic, household, and behavioral indicators. As illustrated in Table 1, the dependent variable was mathematics performance (PV1MATH). To facilitate binary classification, students were categorized into high- and low-achievement groups using the median mathematics score of the training set as the threshold, thereby avoiding data leakage. Independent variables were organized into three conceptual domains. The first domain included family and home indicators reflecting students’ socioeconomic and cultural backgrounds.

VARIABLE	OECD Description
PV1MATH	First plausible value for mathematics proficiency; OECD-generated estimate of student math ability.
ESCS	Economic, Social and Cultural Status Index derived from parental education (HISCED), parental occupation (HISEI), and home possessions (HOMEPOS); standardized across OECD.
HISCED	Highest parental education level according to ISCED 2011 (0–8). Codes 9/10 represent missing values.
PAREDINT	Parental involvement index based on student-reported parental engagement in learning and school activities.
HISEI	Highest parental occupational status measured using the ISEI scale (16–90). Negative values indicate invalid or missing data.
WORKHOME	Indicates whether the student has a quiet place to study at home. Negative responses represent invalid values.
HOMEPOS	Home Possessions Index reflecting cultural and educational resources available at home (books, internet, study desk, etc.).
ICTHOME	Availability of ICT devices at home (e.g., laptop, desktop computer, tablet); measures household digital resources.
ICTAVHOM	Student-reported accessibility of ICT tools at home; indicates whether digital resources are usable.
ICTQUAL	ICT Quality Index measuring student perceptions of device performance and internet quality. Values ≤ −6 represent OECD-coded missing data.
ST062Q01TA	Frequency of arriving late to school in the past two weeks.
ST062Q02TA	Frequency of skipping a full school day in the past two weeks.
ST062Q03TA	Frequency of skipping classes (partial absenteeism) in the past two weeks.
ST004D01T	Gender (1 = Male, 2 = Female). Converted to binary in analysis (Male = 0, Female = 1).

Table 1: OECD Definitions of Variables Included in the Analysis

This domain included the Economic, Social, and Cultural Status (ESCS) index, which summarizes parental education, occupation, and household possessions as a composite measure

of socioeconomic advantage. Parental education levels were represented by HISCED (highest parental education in ISCED levels) and PAREDINT (average parental education level), while parental occupational status (HISEI) captured the highest International Socio-Economic Index of Occupational Status within the family. Indicators of home-learning resources included WORKHOME (availability of a quiet study space), HOMEPOS (overall household educational resources and possessions), and Information and Communication Technology (ICT) related items such as ICTHOME (ICT resources available at home), ICTAVHOM (average ICT availability) and ICTQUAL (ICT Quality Index). The second domain comprised absenteeism-related items (ST062Q01TA–ST062Q03TA), which measured students' frequency of missing school or arriving late, serving as behavioral indicators of engagement and discipline. The third domain included a control variable for student gender (ST004D01T; male or female), enabling the model to account for potential gender-related differences in mathematics performance.

First, the dataset underwent a series of preprocessing procedures. Invalid or special response codes were treated as missing values following OECD conventions (e.g., 9, 10, or negative entries) for ensuring cross-country comparability (*PISA 2022 Technical Report*, 2024).Variables with more than 60% missingness were excluded from the analysis, and the remaining missing values were imputed using the median of each variable. The variable HOMEPOS was excluded due to excessive missingness (>60%), while all remaining features were used in model development. The dataset was balanced across achievement classes, with 2,537 low-achievement and 2,538 high-achievement cases in the training set and 1,082 and 1,093 cases, respectively, in the testing set. After that, mathematics scores were transformed into a binary outcome (high vs. low achievement) using the median value of the training set only, preventing any data leakage into the test set.

2.2. Classification and Hyperparameter selection

After pre-processing, the dataset was divided into training (70%) and testing (30%) subsets using a stratified random split with a fixed random seed (N=42) to ensure reproducibility. Then, three tree-based classifiers such that Random Forest classifier (Breiman, 2001) , ExtraTree (Geurts et al., 2006) and GradientBoosting (Friedman, 2001) were utilized to classify low and high mathematics achievement. The hyperparameters of the classifiers were optimized by the RandomizedSearchCV with 20 randomized iterations and threefold cross-validation. The hyperparameter search includes the number of estimators, maximum tree depth, minimum samples required for splits and leaves, learning rate (for Gradient Boosting), subsampling ratios, and feature selection strategies (sqrt or log2). The area under the receiver operating characteristic curve (ROC–AUC) served as the primary metric for model tuning. Then, model performance was evaluated on the test set using accuracy, macro-averaged F1, and ROC–AUC scores. Finally, feature importance values were extracted from the optimized classifiers using Gini impurity criterion which quantifies node heterogeneity(Breiman et al., 2017). Features yielding greater reductions in impurity across all trees were assigned higher importance scores to highlight the relative contribution of each predictor to mathematics achievement.

2.3. Statistical Analyses

A two-sided Mann–Whitney U tests (Mann & Whitney, 1947), which are non-parametric statistical test used to evaluate whether two independent groups differ in the distribution of a continuous or ordinal variable, were applied to determine whether absenteeism (ST062Q01–03TA) and ICT-related indicators (ICTHOME, ICTAVHOM, ICTQUAL) differed significantly between high and low mathematics achievement groups. The remaining of features such as ESCS, HISEI, HISCED, PAREDINT, WORKHOME, HOMEPOS were not included in the statistical analyses because they were OECD- standardized composite indexes. In order to provide a clear visualization of distributional differences, skewness patterns, and potential group separation, violin plots were generated to visually compare the distribution of absenteeism and ICT-related variables between high and low mathematics achievement groups.

3. RESULTS

In this study, preprocessing steps include data cleaning, handling missing values, and converting the continuous PV1MATH score into high and low achievement groups using a median split. Then, the cleaned dataset was divided in to training (70%) and testing set (30%) by random seed fixed stratified random split. After, three-tree based classifiers such that Random Forest, ExtraTrees and Gradient Boosting were utilized to separate high and low mathematic achievement. Model tuning was conducted to obtain optimal ROC–AUC score by RandomizedSearchCV algorithm (20 iteration, three-fold cross validation). The optimal hyperparameters of the classifiers were demonstrated in Table 2. After hyperparameter tuning, the classifiers were trained and tested to separate between low and high mathematic achievement.

Classifier	**Optimal hyperparameters**
Random Forest	n_estimators=339, max_depth=10, max_features='log2', min_samples_split=3, min_samples_leaf=2
ExtraTree	n_estimators=337, max_depth=10 max_features='log2', min_samples_split=8, min_samples_leaf=4
GradientBoosting	n_estimators=364, learning_rate=0.03, max_depth=3, subsample=0.8

Table 2: Optimal hyperparameters of each classifier in the classification between low and high achievement mathematic groups.

Table 3 demonstrates that GradientBoosting provides most balanced and effective separation in the separation between low and high mathematics achievement with the highest accuracy, sensitivity and ROC-AUC of 67 %, 68% and 0.743, respectively. Similarly, Random Forest classifier yields moderately strong and balanced performance in the classification of low and high mathematics achievement with an accuracy of 66.1%, sensitivity of 64.5%, specificity of 67.8%, and ROC-AUC of 0.729. In comparison, ExtraTree provides a highest specificity of 73.5% indicating strong effectiveness in identifying low-achieving students, whereas a sensitivity of 57.8% yields a higher rate of misclassification among high-achieving students.

Classifier	**Accuracy**	**Sensitivity**	**Specificity**	**F1**	**ROC-AUC**
RandomForest	66.1%	64.5%	67.8%	0.661	0.729
ExtraTrees	65.6%	57.8%	73.5%	0.654	0.724
GradientBoosting	67.0%	68.0%	66.1%	0.670	0.743
RandomForest	66.1%	64.5%	67.8%	0.661	0.729

Table 3. Performance metrics of each classifier in the classification between low and high achievement mathematic groups.

Feature importance analysis based on the Gini impurity criteria was visualized in the heatmap shown in Figure 1, illustrating the relative contribution of each feature across the three ensemble classifiers in distinguishing high and low mathematics achievement. In this heatmap, the x-axis represents the machine learning models (RandomForest, ExtraTrees, and GradientBoosting), and the y-axis lists all features included in the analysis. The color intensity within each cell indicates the magnitude of the feature's importance, with darker shades reflecting stronger influence in the corresponding model. Across models, socioeconomic indicators such as ESCS, HISEI, and HISCED show the highest intensities and demonstrate consistent predictive power. ICT-related variables such as ICTQUAL and WORKHOME provide moderate contributions, while features like ICTHOME and ST004D01T appear with lighter intensities and show weaker or more model-specific relevance. Overall, the heatmap provides a concise comparison of how each classifier weights the input features and highlights the dominant role of socioeconomic background in predicting mathematics achievement.

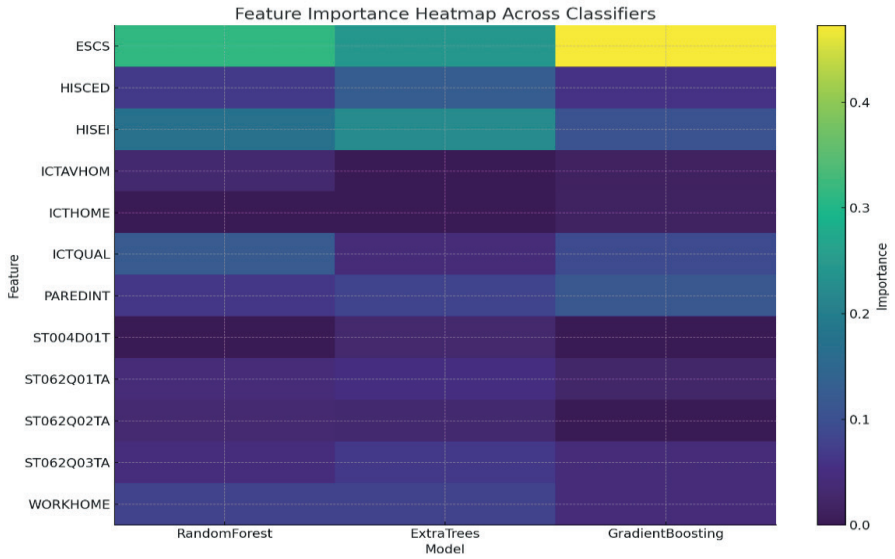


Figure 1: The heatmap of feature importance for each classifier in the classification between low and high low and high achievement mathematic groups.

Finally, two-sided Mann-Whitney U test were employed to assess whether absenteeism and ICT-related indicators differed significantly between low and high achievement mathematic groups separately followed by violin plots for clear visualization as illustrated in Figure 2. For the attendance items (ST062Q01TA–ST062Q03TA), the statistical results showed significant group differences for ST062Q01TA ($p = 1.63\times10^{-9}$) and ST062Q03TA ($p = 2.95\times10^{-17}$), whereas ST062Q02TA did not exhibit a significant effect ($p = 0.354$). These findings are reflected in the violin plots, where ST062Q01TA and ST062Q03TA display visible shifts in median values and distributional density between groups, while ST062Q02TA shows

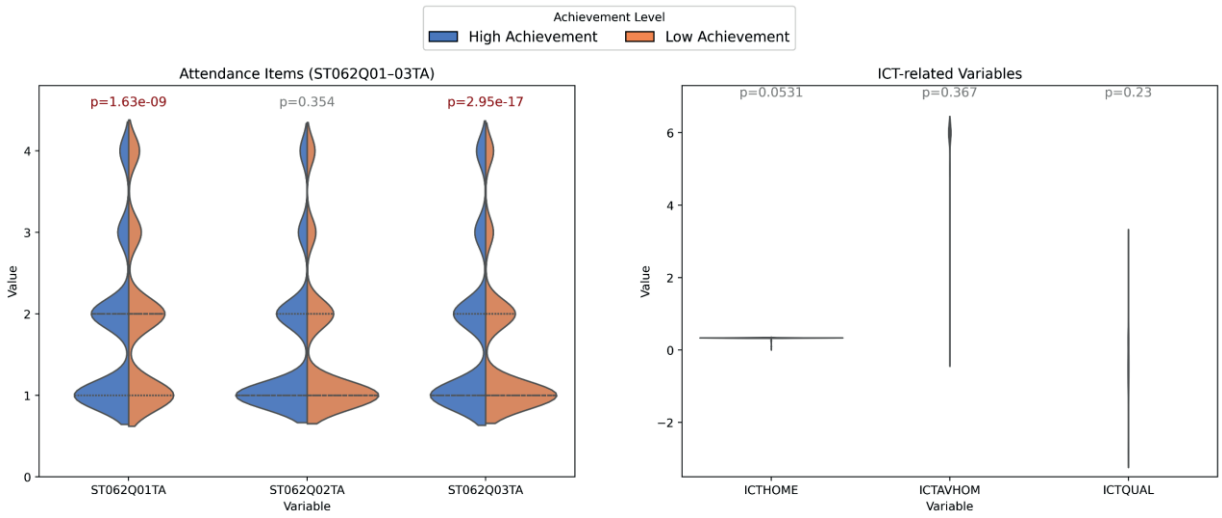


Figure 2: Distribution of Absenteeism and ICT-Related Indicators by Achievement Group

substantial overlap consistent with the non-significant test outcome. For ICT-related variables (ICTHOME, ICTAVHOM, ICTQUAL), none of the Mann–Whitney U tests reached statistical significance (all $p > 0.05$), and the violin plots similarly illustrate overlapping distributions with

comparable medians and variability across achievement levels. Taken together, the combined statistical and graphical evidence indicates that absenteeism patterns, but not ICT resource indicators, differentiate students' mathematics achievement in the PISA 2022 Türkiye dataset.

4. DISCUSSION

This study investigated how socioeconomic, absenteeism-related, and ICT-related indicators contribute to predicting mathematics achievement in the PISA 2022 Türkiye dataset using three ensemble machine learning models. Among the classifiers, Gradient Boosting achieved the most balanced performance, while Random Forest also showed strong predictive ability. ExtraTrees produced the highest specificity but lower sensitivity, indicating that it was more effective at identifying low-achieving students than high-achieving ones, a pattern consistent with prior findings that extremely randomized tree models can exhibit greater variance and class-specific instability due to their high level of randomness(Fernández-Delgado et al., 2014).

Across all models, feature importance results consistently highlighted socioeconomic indicators such as ESCS, HISCED, and HISEI as the most influential predictors of mathematics performance. ICT-related features made only moderate contributions, and gender showed minimal predictive value. These findings support prior evidence that socioeconomic background remains a dominant factor in explaining achievement differences.

The Mann–Whitney U test results further showed that two absenteeism indicators significantly differentiated high and low achievement groups. This reinforces the established link between attendance patterns and academic outcomes(Gottfried, 2014; Kearney, 2008). In contrast, none of the ICT-related variables demonstrated significant group differences, and the corresponding violin plots showed highly overlapping distributions. This suggests that access to ICT resources alone does not translate into measurable differences in mathematics performance without considering how these resources are used(Hu et al., 2018; Skryabin et al., 2015).

5.CONCLUSION

In summary, this study applied machine learning techniques to the Türkiye subsample of PISA 2022 in order to better understand how socioeconomic background, household context, absenteeism, and digital access indicators influence mathematics achievement. The findings reinforce the central claims outlined in the introduction: socioeconomic conditions remain the most powerful predictors of academic performance, reflecting long-standing patterns linked to parental education, occupational status, and access to cultural and material resources. Absenteeism indicators also contributed meaningfully, highlighting the importance of consistent school engagement for sustaining academic success. In contrast, ICT-related variables did not differentiate high- and low-achieving students, suggesting that access to technology alone is insufficient to generate measurable gains in mathematics outcomes. By leveraging ensemble-based classifiers, this study demonstrates the value of machine learning for capturing complex, nonlinear relationships within large-scale educational datasets and for revealing the relative weight of key determinants. Collectively, these results underscore the

need for policies that address socioeconomic inequalities, support regular attendance, and promote meaningful and pedagogically guided digital engagement.

6.FUTURE WORK

Although the current study provides insightful information about the relative impact of behavioral, familial, socioeconomic, and digital factors on mathematical achievement, it should be noted that there are a number of limitations. Initially, the study used cross-sectional data from the PISA 2022 Türkiye subsample, which limits the ability to draw conclusions about causality and makes it impossible to monitor how learning outcomes evolve over time Second, whereas the GradientBoosting effectively captures nonlinear correlations, it remains a data-driven approach, and unobserved contextual factors such as school quality, teacher effectiveness, or regional resource inequities may continue to contribute to unexplained variance. Additionally, the use of self-reported questionnaire items may introduce measurement bias, particularly for indicators related to ICT access and student behavior. Future research should address these limitations by employing longitudinal and cross-country datasets to better capture causal pathways and temporal dynamics in the relationship between digital access, family engagement, and academic performance. Integrating sophisticated modeling techniques with artificial intelligence frameworks such as deep learning techniques might enable researchers to disentangle individual, household, and school-level effects, thereby providing a more comprehensive understanding of educational disparities.

KAYNAKÇA

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In *Classification and Regression Trees*. CRC Press. https://doi.org/10.1201/9781315139470/CLASSIFICATION-REGRESSION-TREES-LEO-BREIMAN-JEROME-FRIEDMAN-OLSHEN-CHARLES-STONE/RIGHTS-AND-PERMISSIONS

Chen, S., & Ding, Y. (2023). A Machine Learning Approach to Predicting Academic Performance in Pennsylvania's Schools. *Social Sciences*, *12*(3), 118. https://doi.org/10.3390/SOCSCI12030118/S1

Chmielewski, A. K. (2019). The Global Increase in the Socioeconomic Achievement Gap, 1964 to 2015. *American Sociological Review*, *84*(3), 517–544. https://doi.org/10.1177/0003122419847165

Davis-Kean, P. E. (2005). The Influence of Parent Education and Family Income on Child Achievement: The Indirect Role of Parental Expectations and the Home Environment. *Journal of Family Psychology*, *19*(2), 294. https://doi.org/10.1037/0893-3200.19.2.294

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Fernández-Delgado, A. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133–3181. http://www.mathworks.es/products/neural-network.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Https://Doi.Org/10.1214/Aos/1013203451*, *29*(5), 1189–1232. https://doi.org/10.1214/AOS/1013203451

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/S10994-006-6226-1/METRICS

Gottfried, M. A. (2014). Chronic Absenteeism and Its Effects on Students' Academic and Socioemotional Outcomes. *Journal of Education for Students Placed at Risk*, *19*(2), 53–75. https://doi.org/10.1080/10824669.2014.962696;JOURNAL:JOURNAL:HJSP20;WGROUP:STRING:PUBLICATION

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., & Villegas-Ch, W. (2025). Machine learning models for academic performance prediction: interpretability and application in educational decision-making. *Frontiers in Education*, *10*, 1632315. https://doi.org/10.3389/FEDUC.2025.1632315/BIBTEX

Hu, X., Gong, Y., Lai, C., & Leung, F. K. S. (2018). The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis. *Computers & Education*, *125*, 1–13. https://doi.org/10.1016/J.COMPEDU.2018.05.021

Jerrim, J., & Macmillan, L. (2015). Income Inequality, Intergenerational Mobility, and the Great Gatsby Curve: Is Education the Key? *Social Forces*, *94*(2), 505–533. https://doi.org/10.1093/SF/SOV075

Jin, X. (2023). Predicting academic success: machine learning analysis of student, parental, and school efforts. *Asia Pacific Education Review*, *1*, 1–22. https://doi.org/10.1007/S12564-023-09915-4/TABLES/4

Kearney, C. A. (2008). School absenteeism and school refusal behavior in youth: A contemporary review. *Clinical Psychology Review*, *28*(3), 451–471. https://doi.org/10.1016/J.CPR.2007.07.012

Klein, M., Sosu, E. M., & Dare, S. (2020). Mapping inequalities in school attendance: The relationship between dimensions of socioeconomic status and forms of school absence. *Children and Youth Services Review*, *118*, 105432. https://doi.org/10.1016/J.CHILDYOUTH.2020.105432

Liu, H., Yang, D., Nie, S., & Chen, X. (2024). Identifying key factors of reading achievement: A machine learning approach. *IScience*, *27*(10), 110848. https://doi.org/10.1016/j.isci.2024.110848

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Https://Doi.Org/10.1214/Aoms/1177730491*, *18*(1), 50–60. https://doi.org/10.1214/AOMS/1177730491

PISA 2022 Database | OECD. (n.d.). Retrieved 26 October 2025, from https://www.oecd.org/en/data/datasets/pisa-2022-database.html

PISA 2022 Technical Report. (2024). https://doi.org/10.1787/01820d6d-en

Rakesh, D., Lee, P. A., Gaikwad, A., & McLaughlin, K. A. (2025). Annual Research Review: Associations of socioeconomic status with cognitive function, language ability, and academic achievement in youth: a systematic review of mechanisms and protective factors. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *66*(4), 417–439. https://doi.org/10.1111/JCPP.14082

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Skryabin, M., Zhang, J., Liu, L., & Zhang, D. (2015). How the ICT development level and usage influence student achievement in reading, mathematics, and science. *Computers & Education*, *85*, 49–58. https://doi.org/10.1016/J.COMPEDU.2015.02.004

Sosu, E. M., Dare, S., Goodfellow, C., & Klein, M. (2021). Socioeconomic status and school absenteeism: A systematic review and narrative synthesis. *Review of Education*, *9*(3), e3291. https://doi.org/10.1002/REV3.3291

CHAPTER 2

MULTIMODAL RETRIEVAL-AUGMENTED GENERATION: A COMPREHENSIVE SURVEY ON ARCHITECTURES, HALLUCINATION MITIGATION, AND EVALUATION

Selahattin Barış ÇELEBI[1], Ammar ASLAN[2]

1 Asst. Prof. Dr., Batman University, Department of Management Information Systems, Batman, sbariscelebi@gmail.com, ORCID: orcid.org/0000-0002-6235-9348
2 Lecturer, Batman University, Department of Computer Technology, Batman, ammaraslan@gmail.com, ORCID: orcid.org/0000-0001-9662-4368

1. Introduction

Retrieval-Augmented Generation (RAG) reduces hallucinations by grounding language models in external evidence. Traditional RAG systems process text effectively but fail when documents contain tables, charts, diagrams, and images. Converting visual elements to text through Optical Character Recognition (OCR) destroys spatial layout, structural hierarchies, and comparative relationships. This information loss propagates through retrieval and generation stages, producing hallucinations rooted in incomplete context. Multimodal RAG (MM-RAG) addresses this limitation by processing heterogeneous data types within unified or aligned representation spaces.

This survey makes three contributions to the MM-RAG literature:

- Architectural Taxonomy: We trace the evolution from discrete OCR-based pipelines to unified vision-language models and emerging agentic systems, clarifying trade-offs between semantic fidelity, computational efficiency, and hallucination risk.
- Hallucination Analysis: We establish a taxonomy distinguishing factuality error (contradicting real-world facts) from faithfulness errors (contradicting provided context), then categorize mitigation strategies by architectural intervention point and hallucination type.
- Evaluation Synthesis: We systematically compare benchmarks from object-level metrics (POPE) to comprehensive multi-dimensional frameworks (MMHal-Bench, TREC RAG Track), identifying gaps in current evaluation protocols and documenting reproducible resources for rigorous experimentation.

The remainder of this paper is organized as follows. Section 2 provides background on text-only RAG and motivates the transition to multimodal paradigms. Section 3 categorizes MM-RAG architectures from OCR-based approaches to late-interaction models. Section 4 analyzes hallucination mechanisms specific to multimodal contexts and their architectural dependencies. Section 5 surveys mitigation strategies including self-verification, visual grounding, and adaptive retrieval. Section 6 examines evaluation benchmarks and protocols. Section 7 investigates applications in medical imaging, financial analysis, and legal document processing. Section 8 concludes with open challenges and future research directions.

2. Background and Related Study

2.1 Text-Only RAG: Evolution and Core Paradigms

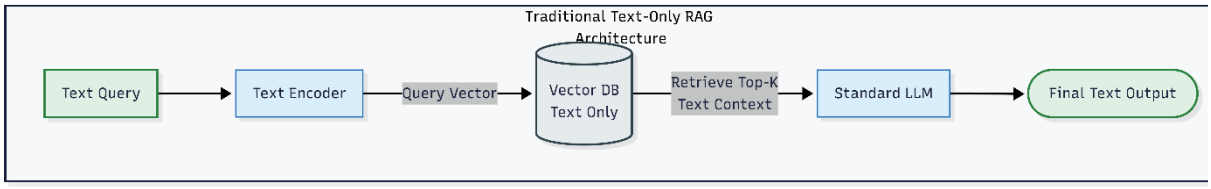
RAG augments parametric knowledge in language models with dynamic, verifiable non-parametric information retrieved from external sources (Lewis et al., 2020). Traditional RAG operates through a retrieve-then-generate workflow. The system first converts a user query into a vector representation, retrieves semantically similar text fragments from an external knowledge base, then presents these fragments as context to the language model for generation. This architecture addresses the limitations of purely parametric models, which suffer from outdated knowledge, factual errors, and hallucinations when generating content beyond their training data.

The evolution of RAG architectures follows three distinct paradigms (Gao et al., 2023). Naive RAG employs simple retrieve-then-generate workflows with fixed chunking strategies and single-stage retrieval. Advanced RAG introduces pre-retrieval optimization (query rewriting,

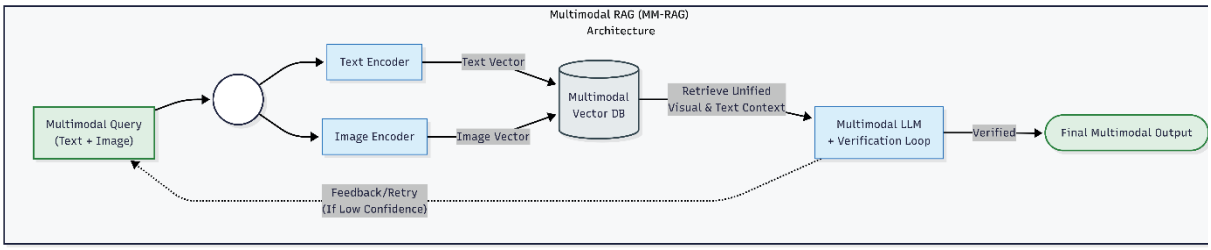
expansion) and post-retrieval refinement (reranking, context compression). Modular RAG decomposes the pipeline into specialized components for indexing, retrieval, and generation, enabling domain-specific customization and iterative refinement. Recent studies further extends these paradigms through self-verification mechanisms, where models assess retrieval relevance and generation faithfulness before outputting results (Asai et al., 2024). While these advances improve text-based retrieval and generation, they assume a fundamental constraint: all information can be adequately represented as text. This assumption breaks down when documents contain non-textual elements such as tables, charts, diagrams, and images.

2.2 The Limitations of Text-Only RAG in Visual Contexts

Documents in critical domains inherently combine text and visual modalities. Academic papers integrate equations, plots, and architectural diagrams. Financial reports embed earnings tables, trend charts, and performance comparisons. Medical records pair clinical notes with X-rays, CT scans, and pathology images. Technical manuals present assembly instructions through annotated photographs and exploded-view diagrams. In these contexts, visual elements are not supplementary—they carry information that text cannot adequately represent. Traditional text-based RAG faces fundamental limitations when processing multimodal documents. Converting visual content to text through OCR destroys three critical information types. Spatial layout disappears when a two-column financial table becomes a linear text sequence, severing relationships between row headers and data cells. Visual comparisons vanish when a bar chart showing quarterly growth trends reduces to isolated numbers without graphical context. Structural hierarchies flatten when nested diagrams with parent-child relationships become unordered text fragments (Faysse et al., 2024). Figure 1 compares traditional text-based RAG (1a) with MM-RAG (1b) architectures, showing how the integration of visual encoders transforms the process from a linear pipeline to a dual-stream alignment mechanism.



(a)



(b)

Fig. 1. Traditional Text-Only RAG (a) vs. MM-RAG (b).

The multimodal architecture introduces vision encoders and cross-modal alignment layers to preserve visual semantics alongside textual information.

These information losses propagate through the entire pipeline. During retrieval, semantically relevant contexts may be missed because OCR-converted text fails to capture the meaning embedded in visual layout or graphical relationships. During generation, models produce hallucinations when attempting to reconstruct information from incomplete textual proxies of visual content. The result is factually incorrect outputs or responses that contradict the original document's meaning. MM-RAG addresses these limitations by processing text, images, and other modalities within unified or aligned representations (Y. Li et al., 2025; Wasserman et al., 2025). Rather than converting visual content to text, MM-RAG systems encode images directly using vision transformers or multimodal encoders, then align these representations with textual embeddings for joint retrieval and generation. This approach preserves spatial relationships, visual hierarchies, and graphical semantics that OCR-based pipelines inevitably lose.

2.3 Related Surveys and Positioning of This Study

Several recent surveys provide complementary perspectives on RAG systems and multimodal learning. Gupta et al. (2024) trace RAG's evolution from foundational retrieval methods to enterprise-scale deployments, documenting architectural choices across 150+ systems but focusing primarily on text-based applications. Yu et al. (2025) establish unified evaluation frameworks specifically for retrieval-augmented generation, addressing benchmark fragmentation but not multimodal-specific challenges. Tonmoy et al. (2024) comprehensively survey hallucination mitigation techniques in large language models, covering both parametric and retrieval-based approaches, while Wasserman et al. (2025) introduce benchmarks for real-world multimodal retrieval scenarios.

Within the multimodal domain, foundational work has established key technical paradigms. Self-RAG (Asai et al., 2024) introduced dual-verification mechanisms for retrieval-generation alignment, demonstrating that models can learn to assess relevance and faithfulness through self-reflection. MuRAG (W. Chen et al., 2022) pioneered multimodal memory integration for open-domain question answering, showing that external visual-textual memory significantly improves performance on questions requiring visual reasoning. Es et al. (2024) developed reference-free automated evaluation frameworks that enable scalable assessment of RAG system quality without ground-truth labels. Recent studies on MM-RAG have explored diverse architectural designs, retrieval strategies, and evaluation settings, making it difficult to directly compare their objectives and empirical findings Table 1 categorizes representative studies, revealing a diverse landscape of architectural objectives and findings.

Table 1. Comparative Summary of the Objectives, Methods, and Findings of Key Studies in the MM-RAG Literature

Author(s) & Year	Objective of the Study	Methodology	Key Findings	Conclusion/ Implications
Drushchak et al., (2025)	To develop a unified MM-RAG system capable of processing heterogeneous data types, including text, tables, images, and videos.	Experimental: Evaluated on Dell server documentation (PDFs and videos) using AWS infrastructure, LangChain, and Claude 3.5 Sonnet.	1. Achieved high retrieval accuracy on structured text and tabular data.2. Performance on image- and video-based queries remained lower than on text queries.3. The unified framework provided a robust pipeline for heterogeneous data sources.	Integrating multiple data modalities into a single pipeline is valuable; however, LLM capabilities for unstructured modalities such as video still require further improvement.
Chen et al., (2022)	To improve question-answering performance by proposing *MuRAG*, a model capable of accessing external multimodal (image–text) memory.	Experimental: Pre-trained on LAION, CC, and VQA datasets; evaluated on WebQA and MultimodalQA benchmarks.	1. Achieved 10–20% higher accuracy compared to existing baselines.2. Significantly outperformed text-only models on questions requiring visual reasoning.	Multimodal retrieval enables language models to "see," reducing hallucinations and improving visual grounding.

Most et al., (2025)	To compare vision-based RAG (ColPali) with OCR-based RAG (Llama 3.2) under varying document quality conditions.	Experimental: Introduced the *DocDeg* dataset containing degraded and noisy documents; evaluated retrieval accuracy and semantic answer quality.	vision-language model (VLM)-based systems (ColPali/ColQwen) are superior in computational efficiency and memory usage, but OCR-based systems (especially Llama 3.2 90B) achieve higher access success rates on degraded documents. This highlights a trade-off between speed and storage efficiency on one hand and semantic accuracy on the other.	In real-world scenarios involving low-quality scans, OCR-based approaches remain more robust; vision-language models require further scaling and robustness improvements.
Chen et al., (2025)	To propose *CMRAG* (Co-Modality-Based RAG), a framework combining textual and visual modalities for document retrieval.	Experimental: Constructed a triplet dataset (query, text, image); evaluated Unified Encoding Models (UEM) and Unified Cross-Modality Retrieval (UCMR).	1. Consistently outperformed unimodal (text-only or image-only) approaches.2. Statistical normalization of text and image similarity scores significantly improved retrieval effectiveness.	Since visual documents require both semantic text understanding and visual perception, joint utilization of both modalities yields the most effective results.
Peng et al., (2025)	To introduce *UniDoc-Bench*, a benchmark for document-centric MM-RAG, and to compare different retrieval strategies.	Benchmark Construction / Experimental: Built on 70,000 pages of real-world PDFs and 1,600 manually verified QA pairs.	1. Text–image fusion (separate retrieval followed by fusion) achieved the best performance (68.4% completeness).2. Joint multimodal embedding models underperformed compared to late fusion strategies.	At present, combining strong unimodal retrievers for text and images is more effective than relying on a single joint multimodal model.

Lumer et al., (2025)	To compare text-based (visual-to-text conversion) retrieval with direct multimodal embedding-based retrieval in financial documents.	Experimental: Evaluated on a financial earnings-call dataset using Jina v4 multimodal embeddings and OpenAI models; employed LLM-as-a-Judge evaluation.	1. Direct multimodal retrieval improved mean average precision (mAP@5) by 32%.2. Preserving visual data in its native form reduced hallucinations compared to text summarization.	Converting visual information (charts, tables) into textual summaries leads to information loss; direct multimodal embeddings yield more accurate and reliable results.
Abootorabi et al., (2025)	To comprehensively survey existing methods, datasets, and open challenges in MM-RAG.	Survey / Literature Review: Taxonomic analysis of existing MM-RAG approaches.	1. Identified encoding, retrieval, fusion, and generation as core pipeline stages.2. Highlighted cross-modal alignment and reasoning as key challenges.	Future research should focus on agentic systems and dynamic modality selection with self-corrective capabilities.

Table 1 focuses on representative and influential studies rather than providing an exhaustive list of all MM-RAG publications. As summarized in Table 1, existing MM-RAG studies differ substantially in both their experimental setups and targeted modalities. While recent works demonstrate clear gains from multimodal retrieval and fusion, they also reveal persistent limitations related to document quality, unstructured visual content, and cross-modal alignment. These observations motivate the architectural analysis presented in the next section

3. Architectural Paradigms in Multimodal RAG

This section systematically categorizes MM-RAG architectures through their fundamental design choices: how visual content is encoded, how text and images are aligned, and how retrieval granularity is determined. We trace the evolution from discrete OCR-based pipelines to unified vision-language embeddings and late-interaction mechanisms, clarifying the trade-offs between semantic fidelity, computational cost, and hallucination risk. Figure 2 illustrates the overall end-to-end architecture of a representative MM-RAG pipeline. The framework integrates modality-aware query processing, multi-stage retrieval, cross-modal fusion, and post-generation verification to mitigate hallucination and improve grounding

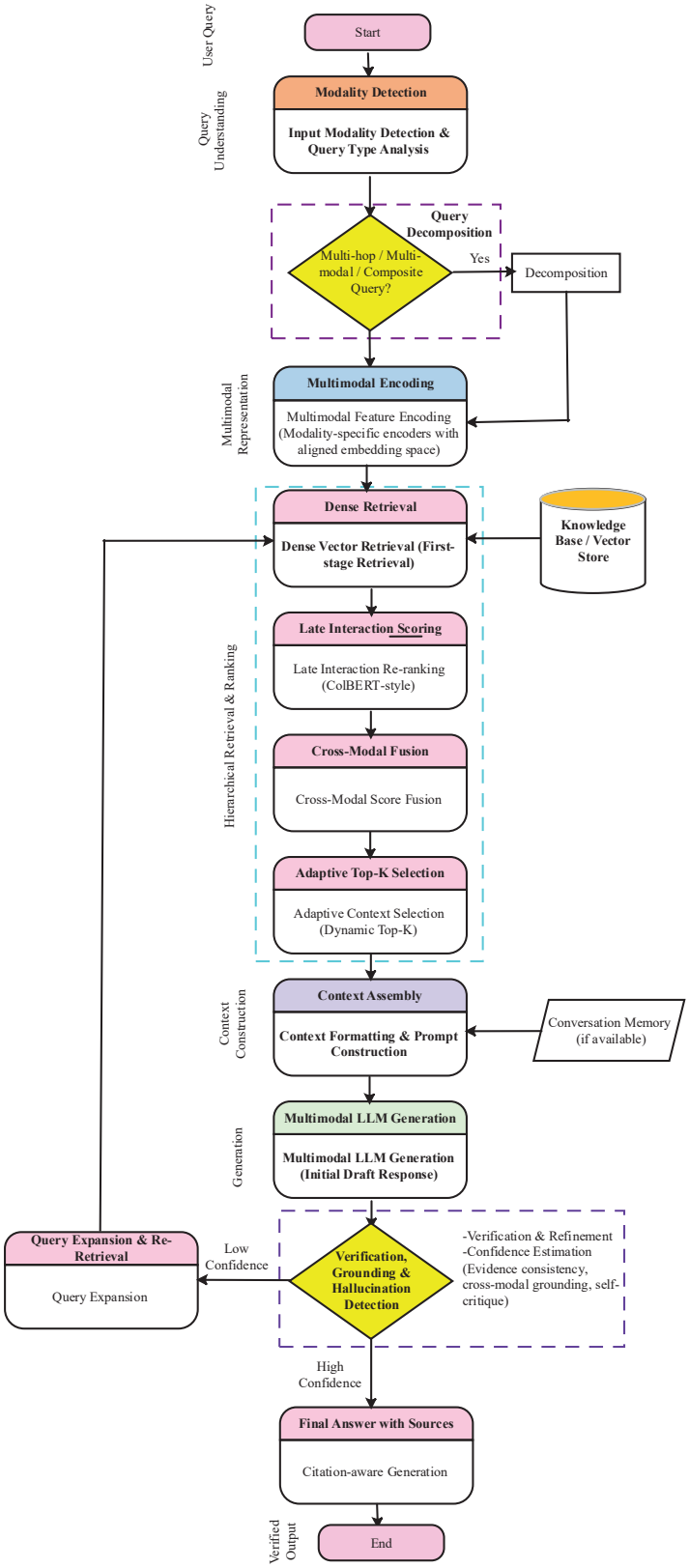


Fig. 2. General workflow of the MM-RAG architecture.

Rather than representing a single implementation, this figure abstracts common design patterns observed across recent MM-RAG systems and highlights emerging components such as adaptive context selection and post-generation verification.

3.1 OCR-Based Pipelines: Advantages, Limitations, and Modern Variants

OCR-based workflows represent the first generation of MM-RAG architectures. These systems convert visual content to plain text using optical character recognition, then apply standard text retrieval methods. This approach offers significant engineering advantages. It leverages mature text indexing infrastructure that has been refined over decades. It requires no modification to existing embedding models like BERT or sentence-transformers. It enables seamless integration with legacy RAG systems deployed in production environments (Lee et al., 2024). However, performance depends on OCR quality and document condition. Modern OCR with large models (Llama 3.2 90B) achieves higher recall than vision embeddings on degraded documents. The fundamental limitation of OCR-based approaches is information loss during modality conversion (Faysse et al., 2024). Three critical information types disappear in the OCR pipeline:

- Spatial Layout: A two-column financial table becomes a linear text sequence. Row-column relationships dissolve. Header associations vanish. The semantic structure that makes tabular data interpretable is flattened into an unstructured string.
- Visual Comparisons: A bar chart showing quarterly growth trends converts to isolated numbers. The visual magnitude comparison that enables rapid pattern recognition is lost. Readers cannot reconstruct trend direction or relative performance from disaggregated values.
- Structural Hierarchies: Nested diagrams with parent-child relationships become unordered text fragments. Flowcharts lose directional arrows. Organization charts lose reporting structures. The compositional semantics encoded in visual layout cannot be preserved in linearized text (Faysse et al., 2024).

These losses propagate through the entire MM-RAG pipeline. During retrieval, semantically relevant contexts may be missed because OCR-converted text fails to capture meaning embedded in visual layout. During generation, models produce hallucinations when attempting to reconstruct information from incomplete textual proxies. A question about quarterly performance in a bar chart may retrieve text mentioning specific numbers but lacking the comparative context required to answer correctly. Despite these limitations, recent advances in OCR and vision-language models partially mitigate degradation effects under noisy conditions.

- Modern OCR Variants Show Resilience: Recent studies reveal a counter-narrative to the vision-first paradigm. Most et al. (2025) demonstrate that state-of-the-art OCR pipelines utilizing large vision-language models (e.g., Llama 3.2 90B) exhibit greater robustness on degraded or noisy documents compared to vision-based embeddings like ColPali, achieving higher scores on retrieval metrics across all tested degradation levels. While VLM-based systems offer superior computational efficiency and storage savings, OCR approaches enhanced with large models show higher retrieval recall when processing low-quality scans, faded text, or documents with complex backgrounds. This suggests a fundamental trade-off: vision embeddings excel at preserving layout semantics on high-quality documents, while robust OCR pipelines maintain text extraction fidelity across varying document conditions.
- Hybrid Strategies: Practical systems often combine OCR with visual processing. Methods like DePlot translate plots into structured tables, preserving quantitative relationships while enabling text-based retrieval (Liu et al., 2022). Late-interaction

retrieval models such as ColBERT apply multi-vector re-ranking to OCR output, partially recovering fine-grained semantic distinctions lost during conversion (Khattab & Zaharia, 2020). Despite these enhancements, OCR-based architectures remain fundamentally limited by the irreversibility of information loss during the visual-to-text transformation.

3.2 Dense Multimodal Embeddings: CLIP, SigLIP, and Cross-Modal Alignment

To overcome OCR limitations, unified multimodal embedding approaches process text and images within aligned vector spaces. These architectures encode both modalities through separate encoders—typically vision transformers for images and BERT-style transformers for text—then align the resulting embeddings through contrastive learning (Radford et al., 2021).

Contrastive Language-Image Pre-training (CLIP): CLIP pioneered large-scale vision-language alignment by training on 400 million image-text pairs scraped from the internet. The model computes similarity matrices over mini-batches and applies contrastive loss to push matched pairs closer while repelling mismatched pairs. CLIP's success demonstrated that web-scale noisy supervision enables zero-shot transfer to diverse visual tasks without domain-specific fine-tuning. However, CLIP's softmax-based contrastive loss requires global normalization across the entire batch, creating computational bottlenecks that limit batch size scaling (Radford et al., 2021)

Sigmoid Loss for Improved Scaling (SigLIP): SigLIP addresses CLIP's scalability constraints by replacing contrastive softmax loss with pairwise sigmoid loss (X. Zhai et al., 2023). Unlike CLIP, which normalizes over all negative examples in a batch, SigLIP treats each image-text pair as an independent binary classification problem. This design enables efficient parallelization across devices and supports larger batch sizes, improving performance particularly at smaller scales where CLIP struggles. SigLIP models trained with sigmoid loss consistently outperform CLIP variants on zero-shot classification and image-text retrieval benchmarks while requiring less computation per training step.

SigLIP 2: Enhanced Semantic Understanding: The recently released SigLIP 2 extends the original architecture with three additional training objectives. A localization-aware decoder adds spatial grounding capabilities. A global-local consistency loss improves fine-grained patch-level semantics through self-distillation. A masked prediction loss enhances dense feature quality for downstream tasks like segmentation and depth estimation. These enhancements yield significant gains on visually rich document retrieval tasks. The NaFlex variant supports dynamic resolution and aspect-ratio preservation, making SigLIP 2 particularly effective for OCR and document understanding applications where layout integrity matters (Tschannen et al., 2025). However, SigLIP 2 shows lower retrieval performance than SigLIP on document tasks (MRR: 0.42 vs 0.47) due to training objective mismatch (W. Chen et al., 2025).

Application to MM-RAG: Dense embeddings enable retrieval systems to match queries with documents based on semantic similarity in joint text-image space. For document retrieval, images of entire pages are embedded alongside their textual content. Query embeddings are compared against this multimodal index using cosine similarity. However, single-vector representations suffer from information compression. A page containing multiple tables, charts, and text blocks must be summarized into a single 512 or 768-dimensional vector. This compression loses fine-grained details, making it difficult to distinguish between documents with similar global semantics but different specific content. Dense embeddings excel at coarse-grained retrieval but struggle with precise localization (X. Chen et al., 2024; Grassucci et al., 2025).

3.3 Late-Interaction Paradigms: ColPali and Multi-Vector Representations

Late-interaction retrieval mechanisms address the compression limitations of dense embeddings by representing documents as multiple vectors—one per patch, token, or semantic unit. Rather than collapsing the entire document into a single embedding, late-interaction models preserve granular representations and defer similarity computation until query time.

Contextualized Late Interaction over BERT(ColBERT): ColBERT introduced the late-interaction paradigm for text retrieval. Instead of encoding a document as a single vector, ColBERT represents each document as a matrix where rows correspond to token embeddings. Query tokens are similarly embedded. Similarity is computed via MaxSim: for each query token, find the maximum cosine similarity with any document token, then sum across query tokens. This operation approximates token-level matching while maintaining computational efficiency through pre-computed document representations (Khattab & Zaharia, 2020).

ColPali: ColPali extends ColBERT's architecture to multimodal documents by leveraging PaliGemma, a vision-language model that projects image patches into a language-aligned space. ColPali treats each page as an image and divides it into a grid of patches. Each patch is encoded through PaliGemma's vision transformer and projected into a 128-dimensional embedding. Query tokens are similarly embedded. MaxSim-based late interaction computes patch-level similarity, enabling fine-grained matching between query concepts and specific visual regions. ColPali demonstrates state-of-the-art performance on the Visual Document Retrieval Benchmark (ViDoRe), particularly on visually complex tasks involving infographics, tables, and charts. It achieves an average nDCG@5 of 81.3 across ViDoRe tasks, significantly outperforming text-based baselines (BM25: 65.5, BGE-M3: 66-67)(Faysse et al., 2024). ColPali reduces OCR errors on high-quality documents. However, OCR-based systems outperform it on degraded inputs (Most et al., 2025). Hybrid approaches are common in practice. It has reduced certain error classes caused by OCR in experiments. However, hybrid solutions may still be used in practical deployments depending on data quality and application requirements. The model is end-to-end trainable and drastically simpler than traditional pipelines requiring document parsing, chunking, and OCR (Faysse et al., 2024).

Interpretability Through Similarity Maps: A unique advantage of late-interaction architectures is interpretability. ColPali can visualize which image patches contribute most to a query match. For a query about "hourly trends," similarity maps highlight not only text mentioning "hourly" but also chart x-axes representing time, demonstrating genuine visual comprehension rather than OCR-based text matching. This interpretability supports debugging and trust in retrieval decisions, particularly in high-stakes domains like medical imaging or financial analysis.

Computational Trade-offs: Late-interaction models require storing multiple vectors per document, increasing index size proportionally to the number of patches or tokens. A page with 256 patches and 128-dimensional embeddings consumes 32KB per page compared to 512 bytes for a single-vector embedding. Efficient implementations leverage approximate nearest neighbor search (e.g., FAISS with product quantization) and early termination strategies to maintain sub-second query latency even on billion-scale corpora. However, storage and indexing costs remain substantially higher than dense embeddings (Faysse et al., 2024; Saad-Falcon et al., 2024).

3.4 Retrieval Granularity: From Short Chunks to Long-Context Units

Beyond modality representation, MM-RAG architectures differ fundamentally in retrieval granularity. Traditional RAG systems chunk documents into short units (100-300 words) to fit within retriever and reader context windows. This design imposes a "heavy retriever, light

reader" paradigm where retrievers search over millions of short fragments, and readers generate from a few concatenated chunks.

- Limitations of Short-Chunk Retrieval: Short chunking fragments semantic context. A financial table split across chunks loses row-column relationships. A multi-paragraph argument divided into separate chunks loses logical flow. This fragmentation introduces two failure modes. First, hard negatives become more likely. A chunk mentioning relevant keywords but lacking necessary context may rank highly yet mislead generation. Second, context incompleteness forces readers to infer missing information, increasing hallucination risk (Z. Jiang et al., 2024).
- Long Retrieval Units (LongRAG): LongRAG proposes inverting the traditional design by retrieving entire documents or large coherent units (4K+ tokens) and delegating understanding to long-context language models (Z. Jiang et al., 2024). For Wikipedia-based QA, LongRAG groups related articles into 4K-token units through hyperlink structure, reducing the corpus from 22 million short paragraphs to 700K long units. This 30x reduction in corpus size dramatically improves retrieval precision: answer recall@1 increases from 52% to 71% on Natural Questions without any model training. The long reader (e.g., GPT-4, Claude) processes concatenated retrieval units (approximately 30K tokens) in a single forward pass, leveraging its capacity to maintain coherence over extended context.
- Granularity as an Architectural Choice: Retrieval granularity interacts with both modality encoding and model capacity. OCR-based systems benefit less from long units because spatial layout is already lost; chunking merely reduces text volume without introducing additional semantic fragmentation. Vision-based late-interaction models like ColPali naturally align with page-level granularity, where each page is a single retrieval unit. Dense embeddings face a tension: longer units provide more context but compress more information into fixed-size vectors, potentially degrading match precision. Adaptive granularity strategies—dynamically selecting chunk size based on document structure and query complexity—represent an active research frontier (Z. Jiang et al., 2024; Saad-Falcon et al., 2024).

3.5 Comparative Analysis: Trade-offs Across Paradigms

Table 2 synthesizes the architectural trade-offs discussed in this section. It compares OCR-based, dense multimodal, and late-interaction paradigms across multiple dimensions: visual data processing, spatial information preservation, retrieval mechanism, representation granularity, storage costs, and retrieval precision on structured documents.

Table 2. Comparative Analysis of MM-RAG Architectures: From OCR-Based Baselines to Late Interaction Paradigms

Aspect	**Naive RAG (OCR-Based)**	**Dense MM-RAG (CLIP-style)**	**ColPali (Late Interaction MM-RAG)**
Visual Data Processing	Visual content converted into plain text via OCR	Visual content encoded into a single global embedding	Visual content encoded as multiple patch-level embeddings
Preservation of Spatial Information	Lost due to linearization of visual structures	Partially preserved through global visual semantics	Largely preserved via layout-aware multi-vector representations
Retrieval Mechanism	Text matching using sparse (BM25) or dense retrieval	Cosine similarity over single-vector embeddings	MaxSim-based late interaction between query tokens and visual patches
Representation Granularity	Coarse, text-only segments	Medium, global visual-text alignment	Fine-grained, token–patch and layout-level alignment
Storage and Indexing Cost	Low	Moderate	High due to multi-vector indexing
Retrieval Precision on Structured Documents	Limited, sensitive to OCR errors	Improved but prone to information compression	High, particularly effective for tables, figures, and page layouts
Current Status in the Literature	Traditional; still necessary in constrained pipelines	Widely adopted in multimodal retrieval systems	Representative of recent state-of-the-art approaches in document-level MM-RAG research (2024–2025).

Rather than presenting these paradigms as interchangeable alternatives, the following comparison highlights the concrete trade-offs they impose across retrieval precision, storage cost, and latency, thereby framing architectural choice as an explicit design decision rather than a purely empirical preference

Key Insights: First, OCR-based pipelines remain competitive in low-quality document settings despite semantic limitations. When documents contain degraded scans or handwritten text, robust OCR with large vision-language models (e.g., Llama 3.2 90B) outperforms vision-first embeddings (Most et al., 2025). Second, unified multimodal embeddings consistently outperform text-only baselines on visually grounded tasks but suffer from compression bottlenecks that limit fine-grained matching. Third, late-interaction architectures demonstrate superior performance on structured documents by preserving layout and enabling patch-level reasoning. However, no single architecture dominates across all scenarios. Optimal design depends on document characteristics (quality, layout complexity), query distribution (keyword-based vs. conceptual), and operational constraints (latency, storage budget). Hybrid approaches that adaptively select architectures based on input properties represent a promising direction for future study.

4. Hallucination Mechanisms in Multimodal RAG

Hallucination in large language models refers to outputs that appear fluent and plausible yet contradict provided context or established facts (Ji et al., 2023). In MM-RAG systems, this problem exhibits a more complex structure because errors can originate not only from generation but also from retrieval, representation, and cross-modal alignment stages. This section analyzes why multimodal architectures amplify hallucination risks, establishes a granular taxonomy of error types, and traces how architectural choices create distinct hallucination patterns.

4.1 Why MM-RAG Amplifies Hallucination Beyond Text-Only Systems

MM-RAG introduces three compounding factors absent in text-only systems that fundamentally increase hallucination risk. Cross-Modal Misalignment During Retrieval: Text-based RAG retrieves semantically similar passages. Relevance is defined within a single modality using cosine similarity over text embeddings. MM-RAG must assess relevance across heterogeneous representations. A query about "quarterly revenue growth" may retrieve a document containing relevant keywords but irrelevant charts. The textual context mentions "growth," yet the chart shows decline. When the generator receives both text and image embeddings without verifying their semantic coherence, it may synthesize claims that conflict with the visual evidence. This cross-modal mismatch creates hallucinations rooted in incomplete or contradictory multimodal context.

- Retrieval-Specific Information Loss: Beyond the modality conversion losses discussed in Section 3.1, MM-RAG faces unique retrieval-stage failures. Dense multimodal embeddings compress entire pages into fixed-size vectors, losing fine-grained spatial relationships. A financial table showing quarterly performance may be indexed as a single 768-dimensional embedding. When retrieved, the generator cannot distinguish which specific row or column supports a claim, increasing fabrication risk. Late-interaction models mitigate but do not eliminate this problem. While patch-level embeddings preserve layout, retrieval still operates on aggregated MaxSim scores. A page with ten charts may be retrieved based on global similarity, yet only two charts are relevant to the query. The generator must infer which visual elements matter, often incorrectly.
- Cascading Error Propagation: Text-only RAG exhibits linear error propagation: incorrect retrieval leads to incorrect generation. MM-RAG exhibits cascading errors across modalities. Suppose OCR misreads a table value (e.g., "8.5%" as "85%"). The retriever correctly identifies the page as relevant. The generator, trusting the retrieved context, produces a response claiming "85% growth." The error originates in OCR but manifests as a generation hallucination. Unlike text typos, which readers often recognize, numerical errors from OCR appear internally consistent. The model has no signal indicating corruption. This cascading failure mode demonstrates how upstream modality-specific errors amplify downstream generation risks (Faysse et al., 2024; Most et al., 2025)
- New findings reveal additional nuances Chen et al. (2025) demonstrate that omission hallucinations (failing to describe present objects) and fabrication hallucinations (describing absent objects) stem from distinct mechanisms. Omissions arise from low-confidence visual encodings even when objects are correctly perceived. Fabrications result from overreliance on linguistic priors. Standard contrastive decoding methods reduce fabrications but exacerbate omissions, highlighting the need for mechanism-specific interventions.

4.2 A Granular Taxonomy: Factuality, Faithfulness, and Visual Error Types

Recent surveys establish two foundational hallucination categories applicable across LLMs and multimodal systems (B. Chen et al., 2025; Huang et al., 2024). We distinguish errors that contradict external facts (factuality) from those that contradict the provided input (faithfulness), because mitigation strategies differ between them:

- Factuality Hallucinations: contradict established real-world facts. These errors manifest when a model generates content inconsistent with verifiable external knowledge. Example: "The Great Wall of China is visible from space with the naked eye" contradicts documented scientific evidence. In MM-RAG, factuality hallucinations often arise from parametric knowledge stored during pretraining rather than retrieved evidence. A model may override retrieved context with incorrect memorized facts, especially when retrieval confidence is low.
- Faithfulness Hallucinations: diverge from provided input context or user instructions. These errors represent internal inconsistencies rather than factual inaccuracies. Three subtypes exist: *instruction inconsistency* (failing to follow user directives), *context inconsistency* (contradicting retrieved documents), and *logical inconsistency* (internal contradictions within generated output). In MM-RAG, context inconsistency is particularly critical. A generator produces faithfulness hallucinations when it fabricates claims unsupported by retrieved text or images, even if the claims are factually plausible.

While this binary taxonomy applies broadly, visual tasks require finer granularity. Bai et al. (2024) categorize multimodal hallucinations (MMHal) into three visual-specific types:

1. Object Category Hallucinations: The model identifies nonexistent objects or misclassifies existing ones. Example: detecting a "cat" where none exists or labeling a "bus" as a "truck." These errors reflect failures in basic visual perception or overconfident predictions driven by linguistic priors. POPE benchmark specifically measures object-level hallucinations through binary existence questions (Y. Li et al., 2023).

2. Attribute Hallucinations: The model correctly identifies an object but misrepresents its visual properties such as color, shape, size, material, or count. Example: describing a "red car" as "blue," or claiming "three people" when five are present. Attribute errors often result from insufficient visual grounding. The model recognizes a car's presence but hallucinates properties based on statistical priors rather than visual evidence.

3. Relation Hallucinations: The model accurately describes individual objects and their attributes but fails to capture spatial or semantic relationships. Example: stating "a person is holding a cup" when the cup is on the table, or describing "a dog next to a tree" when the dog is behind the tree. Relation hallucinations are particularly challenging because they require compositional reasoning across multiple visual elements. Recent benchmarks like Reefknot specifically target relation errors, distinguishing perceptive failures (incorrect spatial understanding) from cognitive failures (incorrect inferential reasoning) (B. Chen et al., 2025).

This visual taxonomy is orthogonal to the factuality-faithfulness distinction. An object hallucination may be faithfulness-violating (describing an object absent from the image) or factuality-violating (describing an object type that does not exist in reality). Similarly, an attribute error may faithfully reflect the image but contradict world knowledge (e.g., describing a "blue banana" accurately shown in a doctored image).

4.3 Impact of Architectural Choices on Hallucination Patterns

Different MM-RAG architectures exhibit distinct hallucination profiles. Understanding these dependencies enables targeted mitigation strategies.

OCR-Based Pipelines: OCR-based systems primarily suffer from information loss during modality conversion (Section 3.1). However, their hallucination patterns extend beyond mere omissions. Numerical errors can occur when decimal points or digits are misread, particularly in degraded or noisy documents. Financial and medical applications are particularly vulnerable. A misread lab result or stock price propagates through retrieval and generation without correction. Structural errors occur when spatial layout information is lost. A two-column comparison table becomes a flat text sequence. The model cannot infer which values compare to which, leading to incorrect relationship claims. Recent study demonstrates that state-of-the-art OCR pipelines utilizing large vision-language models (e.g., Llama 3.2 90B) exhibit greater robustness on degraded or noisy documents compared to vision-based embeddings, achieving higher semantic answer quality when document quality is poor (Most et al., 2025). This suggests a trade-off: VLM-based systems excel on clean documents with complex layouts, while robust OCR approaches maintain reliability across varying document conditions

Dense Multimodal Embeddings: CLIP-style dense embeddings (Section 3.2) compress entire pages into single vectors. This compression creates attribute and relation hallucinations. A page with multiple objects may be retrieved based on global semantic similarity, yet the generator lacks fine-grained spatial information to verify relationships. Experiments show that dense retrievers excel at coarse-grained tasks (e.g., "find documents about quarterly earnings") but struggle with specific claims (e.g., "verify the exact percentage in the Q2 bar chart"). The model retrieves relevant documents but generates attribute details from parametric knowledge rather than visual evidence (X. Chen et al., 2024).

Late-Interaction Models: ColPali and similar late-interaction architectures (Section 3.3) dramatically reduce object category hallucinations. Patch-level embeddings enable precise localization. The model can verify whether an object exists in a specific image region, reducing false positives in POPE benchmarks by 15-20% compared to dense embeddings (Faysse et al., 2024). However, late-interaction models still struggle with relation hallucinations. MaxSim scoring identifies relevant patches but does not model inter-patch relationships. A query about "spatial arrangement" may retrieve patches showing individual objects but fail to encode their relative positions. Recent study on scene graph representations addresses this limitation by explicitly modeling object relationships, but integration into scalable RAG systems remains an open challenge (X. Chen et al., 2024).

Granularity Effects: LongRAG's long retrieval units (Section 3.4) reduce hard negative retrievals, indirectly mitigating faithfulness hallucinations. When retrieval units preserve full document context, generators receive coherent semantic structures rather than fragmented chunks. Experiments show that retrieval precision and answer accuracy improve when moving from short 300-word chunks to long 4K-token units, with Exact Match scores increasing from 42% to 59% when using grouped documents (Z. Jiang et al., 2024). However, long contexts introduce a different risk: saliency bias. Generators may attend disproportionately to early content, ignoring critical visual elements appearing later in long documents. This attention allocation failure creates omission hallucinations, where present information is overlooked rather than fabricated.

4.4 Emerging Insights: Omission vs. Fabrication Mechanisms

Recent empirical studies challenge the assumption that all hallucinations share a common cause. Traditional mitigation strategies apply uniform interventions, yet results show asymmetric effects. Contrastive decoding reduces fabrications (describing absent objects) but increases omissions (failing to describe present objects). This divergence suggests distinct underlying mechanisms (B. Chen et al., 2025).

Fabrication Hallucinations: stem from overreliance on linguistic priors. Pre-trained language models encode statistical regularities ("bananas are yellow," "dogs chase cats"). When visual evidence is weak or ambiguous, these priors dominate generation. Contrastive decoding suppresses outputs driven by text-only distributions, effectively reducing fabrications. However, aggressive suppression also penalizes correct but statistically common descriptions.

Omission Hallucinations: arise from low-confidence visual encodings. Even when vision encoders correctly perceive an object, the cross-modal projection layer may assign low probability to the corresponding linguistic token. During generation, the model filters low-confidence predictions, causing omissions. This mechanism explains why increasing model scale alone does not eliminate hallucinations. Larger models amplify both correct and incorrect priors, improving fabrication rates but not addressing confidence calibration failures.

These insights motivate mechanism-specific interventions. Fabrications require visual grounding and contrastive decoding (Section 5.3). Omissions require confidence calibration and visual-linguistic alignment refinement (Section 5.4). Unified mitigation strategies that ignore mechanistic differences achieve suboptimal performance across both error types.

The taxonomy introduced above distinguishes hallucinations not only by their observable form (object, attribute, relation) but also by their underlying mechanism (fabrication versus omission). These distinctions are critical, as different mitigation strategies intervene at different stages of the MM-RAG pipeline and therefore target different error mechanisms. Accordingly, the following section organizes mitigation approaches in relation to the specific failure modes they are designed to address

5. Mitigation Strategies for Hallucination Reduction

The hallucination mechanisms analyzed in Section 4 reveal that errors originate at multiple pipeline stages and across modalities. Effective mitigation therefore requires targeted interventions calibrated to specific failure modes. This section surveys four complementary strategies: self-verification through Chain-of-Verification, adaptive retrieval with Self-RAG, visual grounding techniques, and confidence calibration mechanisms. We conclude with a comparative analysis clarifying when each approach applies and how they compose in production systems. Importantly, while this strategies is effective at reducing fabrication-type hallucinations, it introduces non-trivial latency overheads and may exacerbate omission errors, underscoring the need to evaluate mitigation techniques as trade-offs rather than universally beneficial add-ons.

5.1 Chain-of-Verification: Iterative Self-Correction

Chain-of-Verification (CoVe) introduces a four-stage process where models generate, verify, and revise their outputs before presenting final responses (Dhuliawala et al., 2024). The workflow proceeds as follows.

- Draft Generation: The model produces an initial response to the query without external verification. This baseline response may contain hallucinations due to parametric

knowledge gaps or retrieval failures. Verification Planning: The model generates a set of verification questions designed to test factual claims in the draft. For example, if the draft states "The Mexican-American War occurred from 1846 to 1848," a verification question might ask "When did the Mexican-American War start and end?" Crucially, these questions are not templated. The model formulates them autonomously, enabling coverage of diverse claim types.

- Independent Execution:Verification questions are answered independently, without conditioning on the original draft. This prevents the model from simply parroting the initial response. The independent execution breaks confirmation bias, forcing the model to re-derive facts from parametric knowledge or retrieved evidence.
- Revised ResponseThe model compares verification answers against draft claims, identifies inconsistencies, and generates a corrected final response.

Empirical results demonstrate CoVe's effectiveness across multiple tasks. On Wikidata list-based questions, CoVe improves test precision from 0.17 to 0.36, reducing hallucinated entities by 77% (from 2.95 to 0.68 negatives per response) while maintaining non-hallucination coverage (Dhuliawala et al., 2024). On closed-book MultiSpanQA, CoVe increases F1 score by 23% (from 0.39 to 0.48). For long-form biography generation, CoVe-enhanced Llama2 outperforms InstructGPT, ChatGPT, and PerplexityAI on FactScore metrics, demonstrating that self-verification scales to complex generation tasks.

- Multimodal Extension: In MM-RAG contexts, CoVe can verify visual claims by generating questions about image content. For instance, if a model describes "three people standing near a car," CoVe generates verification questions like "How many people are in the image?" and "What objects are present?" Answering these questions independently with visual grounding prevents linguistic priors from overriding visual evidence. However, CoVe's performance is bounded by the base model's reasoning capacity. Verification questions are answered more accurately than the original query, but complex multi-hop reasoning or rare factual knowledge remains challenging. CoVe does not eliminate hallucinations completely, reducing them by approximately 40-60% depending on task complexity. Errors in reasoning steps or factual gaps in parametric knowledge persist despite verification loops.

5.2 Self-RAG: On-Demand Retrieval and Reflection Tokens

Traditional RAG systems retrieve a fixed number of passages for every query, regardless of whether external knowledge improves response quality. This indiscriminate retrieval introduces two failure modes. First, unnecessary retrieval adds latency and computational cost for queries answerable from parametric knowledge. Second, irrelevant passages confuse generation, degrading output quality even when the model could answer correctly without retrieval. Self-RAG addresses these limitations through adaptive retrieval controlled by reflection tokens (Asai et al., 2024). The framework trains a language model to generate special tokens that trigger retrieval, assess passage relevance, and verify output factuality. The training process involves three components. A critic model predicts when retrieval would improve generation. A retriever supplies passages on-demand. A generator produces outputs interleaved with reflection tokens that assess retrieval necessity, relevance, and support. During inference, the model dynamically decides whether to retrieve, evaluates retrieved contexts, and generates responses only when evidence supports claims.

Retrieve tokens indicate whether external knowledge would improve the response. The model generates [Retrieve=Yes] when the query requires factual information absent from parametric knowledge, and [Retrieve=No] when sufficient internal knowledge exists. Relevance tokens

assess whether retrieved passages contain information pertinent to the query. The model scores passages with [Relevant] or [Irrelevant] tags, filtering low-quality contexts before generation. Support tokens verify whether generated claims are grounded in retrieved evidence. The model assigns [Fully Supported], [Partially Supported], or [No Support] labels to each generated segment, enabling segment-level attribution. Utility tokensevaluate overall response quality considering both relevance and support. The model assigns [5], [4], [3], [2], or [1] utility scores, facilitating beam search over candidate generations.

Experiments demonstrate Self-RAG's superiority across six tasks. On PubHealth fact-checking, Self-RAG (7B parameters) achieves 72.4% accuracy, outperforming retrieval-augmented ChatGPT (54.7%) and standard ChatGPT (70.1%), demonstrating that selective retrieval and reflection tokens reduce hallucinations effectively. On long-form generation (Biography), Self-RAG achieves a FactScore of 81.2, significantly outperforming the baseline of 55.9, demonstrating that selective retrieval and reflection tokens reduce hallucinations without sacrificing generation fluency (Asai et al., 2024).

Multimodal Adaptation: Self-RAG's reflection mechanism extends naturally to MM-RAG. The model can generate [Retrieve-Image=Yes] tokens when visual evidence would clarify queries, then assess whether retrieved images are [Relevant] to the question. Support tokens verify whether generated descriptions are [Fully Supported] by visual content, preventing attribute and relation hallucinations. Recent study demonstrates that multimodal reflection tokens reduce CHAIR object hallucination scores by 18-22% on captioning benchmarks (W. Zhai, 2024).

5.3 Visual Grounding: Linking Claims to Image Regions

Visual grounding techniques reduce hallucinations by explicitly connecting textual claims to specific image regions. Unlike CoVe and Self-RAG, which operate primarily through linguistic reasoning, visual grounding enforces pixel-level accountability. The model must identify bounding boxes, patches, or attention maps that support each generated claim. This constraint prevents the model from fabricating visual details based solely on statistical priors.

Visual Description Grounding (VDGD): Ghosh et al. (2024) introduce VDGD, a training-free method that grounds response generation in visual descriptions. The approach first generates detailed descriptions of image content using a vision encoder. During text generation, the language model's attention is biased toward tokens consistent with these visual descriptions. This mechanism amplifies the influence of visual evidence over parametric linguistic priors. VDGD improves accuracy by 2-33% across eight benchmarks requiring deliberate reasoning, including MMMU (math understanding), MathVista (visual math), and AMBER (attribute reasoning). Critically, VDGD operates at inference time without model retraining, making it applicable to any pre-trained vision-language model.

Multi-Modal Mutual-Information Decoding (M3ID): Favero et al. (2024) demonstrate that as text generation progresses, models increasingly rely on language priors rather than visual input. This decaying visual reliance correlates strongly with hallucination emergence. M3ID counteracts this drift by amplifying tokens with higher mutual information with the visual prompt. The method computes token probabilities conditioned on the image, compares them to text-only probabilities, and upweights visually grounded tokens during sampling. For LLaVA 13B, M3ID reduces hallucinated objects in captioning by 25% and improves POPE accuracy by 21%. When paired with Direct Preference Optimization (DPO), improvements reach 28% and 24% respectively. M3ID requires no training and adds minimal computational overhead, operating through modified sampling at inference time.

Limitations and Open Challenges: Visual grounding reduces object and attribute hallucinations but shows limited effectiveness against relation errors. Identifying whether "a person is holding a cup" versus "a cup is on the table" requires compositional reasoning across multiple regions, which current grounding methods struggle to capture. Additionally, grounding techniques assume high-quality vision encoders. When input images are blurred, occluded, or adversarially perturbed, visual evidence itself becomes unreliable, degrading grounding effectiveness. Future study must address these limitations through structured scene graph representations that explicitly model inter-object relationships.

5.4 Adaptive Retrieval and Confidence Calibration

Adaptive retrieval mechanisms optimize when and what to retrieve, moving beyond naive always-retrieve strategies. Two complementary dimensions govern effectiveness: retrieval triggering (deciding if retrieval helps) and confidence calibration (assessing output reliability).

Self-Adaptive Multimodal RAG (SAM-RAG): W. Zhai (2024) introduces SAM-RAG, which dynamically filters documents and verifies both evidence and generation quality in multimodal contexts. The system implements three-stage adaptation. First, relevance screening evaluates whether retrieved documents contain information useful for answering the query. The model scores each document's alignment with the question, retaining only high-relevance contexts. Second, evidence verification assesses whether the retained documents support factual claims in the generated response. This step prevents hallucinations from emerging even when relevant documents are retrieved. Third, output validation performs final checks on generation quality, including factual consistency and answer completeness. Experiments show SAM-RAG improves retrieval accuracy and response quality over fixed-retrieval baselines, particularly on queries requiring multi-hop reasoning across text and images.

Adaptive-RAG: Query Complexity Routing: Jeong et al. (2024) propose training a classifier that predicts query complexity and routes requests to different retrieval strategies. Simple factual queries trigger single-step retrieval. Complex multi-hop questions activate iterative retrieval loops. Queries answerable from parametric knowledge bypass retrieval entirely. This adaptive routing reduces computational cost while maintaining accuracy. On open-domain QA datasets, Adaptive-RAG improves efficiency by 30% and accuracy by 5-8% over always-retrieve baselines.

Confidence Calibration: The Misalignment Problem: Confidence scores from retrieval and generation stages often diverge, creating calibration failures (B. Chen et al., 2025). High retriever confidence with low generator confidence signals modality mismatch or insufficient context. The retriever successfully found relevant passages, but visual or spatial information required for generation is missing. Conversely, low retriever confidence with high generator confidence indicates over-reliance on parametric knowledge. The model generates confidently despite weak evidence, increasing hallucination risk.

Calibration strategies address these divergences through two mechanisms. Separate calibration per component:

- Retriever confidence is calibrated using ranking metrics (e.g., nDCG@10 thresholds). Generator confidence is calibrated using token-level logits or ensemble disagreement. When both signals indicate high confidence, outputs are reliable. When signals conflict, the system triggers verification or abstains from answering.
- Cross-modal calibration: Multimodal systems require joint calibration across text and vision pathways. Recent study introduces correlation-based decoding that dynamically adjusts output logits based on visual-textual alignment scores (B. Chen et al., 2025).

When visual evidence strongly supports a claim, generator confidence is upweighted. When visual-text correlation is low, confidence is down weighted, preventing hallucinations from weak cross-modal grounding.

5.5 Comparative Analysis: When to Apply Each Strategy

Table 3 synthesizes the effectiveness, latency cost, and application scenarios for each mitigation strategy. No single method dominates across all contexts. Optimal deployment depends on hallucination type, task requirements, and computational constraints.

Table 3. Comparative Analysis of Hallucination Mitigation Strategies

Strategy	Object Hallucination	Attribute Hallucination	Relation Hallucination	Latency Cost	Best Use Cases
CoVe	High (40–60%)	Medium (30–45%)	Medium (25–35%)	High (3–4× inference)	Long-form generation and complex reasoning tasks where correctness is more important than speed
Self-RAG	High (45–65%)	Low (15–25%)	Low (10–20%)	Medium (1.5–2× inference)	Knowledge-intensive QA and fact verification where selective retrieval reduces noise
Visual Grounding (VDGD / M3ID)	Medium (35–50%)	High (50–70%)	Low (10–15%)	Low (1.1–1.3× inference)	Image captioning and visual QA tasks where attribute accuracy is critical
Adaptive Retrieval (SAM-RAG)	Medium (30–45%)	Medium (30–45%)	Medium (20–30%)	Low (1.2–1.5× inference)	High-throughput applications requiring efficiency without sacrificing accuracy
Confidence Calibration	Low (20–30%)	Low (20–30%)	Low (15–25%)	Negligible (<1.05×)	Production systems requiring uncertainty estimation and selective abstention

Key Insights from Comparative Analysis:

1. Complementary Failure Modes: CoVe excels at detecting object hallucinations through independent verification questions but provides limited benefit for attribute and relation errors, which require visual rather than linguistic reasoning. Visual grounding methods conversely reduce attribute hallucinations dramatically but struggle with objects and relations. Combining CoVe with visual grounding yields synergistic improvements, reducing multiple error types simultaneously.

2. Latency-Accuracy Trade-offs: CoVe imposes the highest latency cost due to multi-stage generation (draft → verification questions → answers → revision). Production systems requiring real-time responses cannot afford 3-4x inference overhead. Visual grounding techniques like M3ID add minimal latency (<30% increase) while achieving substantial hallucination reductions, making them suitable for latency-sensitive applications like interactive visual QA or content moderation.

3. Task-Dependent Effectiveness: Self-RAG's adaptive retrieval works best for knowledge-intensive tasks where external evidence is critical. On tasks solvable from parametric knowledge (e.g., "What is the capital of France?"), Self-RAG correctly abstains from retrieval, reducing computational waste. Conversely, visual grounding provides no benefit for text-only queries. Systems must select strategies based on input modality and query type.

4. Confidence Calibration as Meta-Strategy: Confidence calibration does not directly reduce hallucinations but enables systems to recognize when other mitigation strategies should activate. A low-confidence retrieval score triggers CoVe verification loops. A high visual-textual correlation mismatch activates visual grounding. Confidence signals thus orchestrate strategy selection, creating adaptive pipelines that apply mitigation only when necessary.

5. Composability and Synergies: Strategies can be composed for multiplicative gains. Self-RAG + Visual Grounding reduces object hallucinations by 60-75%, exceeding either method alone. CoVe + Confidence Calibration enables selective verification, applying expensive multi-stage reasoning only to low-confidence outputs. These compositions highlight the importance of modular design, where mitigation components operate independently and combine flexibly based on task requirements.

Practical Deployment Recommendations: High-stakes domains (medical diagnosis, legal analysis) should prioritize accuracy over latency, deploying CoVe + Visual Grounding + Confidence Calibration. Consumer applications (chatbots, content generation) should optimize for throughput, using Visual Grounding + Adaptive Retrieval. Research systems exploring new architectures should implement all strategies modularly, enabling controlled ablation studies that identify optimal combinations for specific tasks.

6. Evaluation Benchmarks and Methodologies

Evaluating MM-RAG systems requires measuring performance across multiple dimensions: retrieval precision, generation quality, hallucination rates, and source attribution accuracy. Unlike text-only generation, where BLEU or ROUGE scores provide rough quality estimates, multimodal systems demand evaluation protocols that verify visual grounding, assess cross-modal consistency, and detect fine-grained error types. This section surveys automated evaluation frameworks, object-level benchmarks, comprehensive multi-dimensional assessments, and community-driven standardization efforts.

6.1 Automated Evaluation Frameworks: ARES and LLM-as-a-Judge

Traditional RAG evaluation relies on human annotations for queries, retrieved passages, and generated responses. This approach is accurate but fundamentally unscalable. Annotating thousands of system outputs requires months of human effort, creating bottlenecks that prevent rapid iteration during model development. Automated evaluation frameworks address this constraint by training language models to assess RAG component quality without extensive human labeling.

Automated RAG Evaluation System (ARES) introduces a three-stage evaluation pipeline that reduces human annotation requirements by two orders of magnitude (Saad-Falcon et al., 2024). The framework evaluates RAG systems along three dimensions: context relevance (does the retrieved passage contain information pertinent to the query?), answer faithfulness (is the generated response grounded in retrieved evidence?), and answer relevance (does the response address the user's question?).

The ARES workflow operates as follows.

1-Synthetic Data Generation: Given an in-domain passage set, ARES generates synthetic query-passage-answer triples using large language models. These triples form positive examples where context is relevant, answers are faithful, and responses are on-topic. Negative examples are created through contrastive sampling—pairing queries with irrelevant passages or generating unfaithful answers that introduce facts absent from context.

2-Judge Training: Using synthetic triples, ARES fine-tunes lightweight language models (e.g., FLAN-T5 XXL) as classifiers for each evaluation dimension. The judges learn to score context relevance, faithfulness, and relevance through supervised training on automatically generated labels.

3-Prediction-Powered Inference (PPI): To mitigate errors from synthetic training, ARES calibrates judge predictions using a small human-labeled validation set (150-300 examples). PPI provides statistical confidence intervals around system rankings, enabling principled comparison despite judge imperfections.

Experiments across eight knowledge-intensive tasks in KILT, SuperGLUE, and AIS demonstrate ARES's effectiveness. ARES achieves Kendall's tau correlation of 0.82 with human judgments on context relevance and 0.76 on answer relevance, significantly outperforming few-shot GPT-3.5 baselines (tau 0.75 and 0.63 respectively) and a reference-free evaluation framework for RAG systems (RAGAS) (tau 0.78 and 0.71). Critically, ARES maintains accuracy across domain shifts—transferring from one document collection to another without retraining. This robustness makes ARES suitable for evaluating diverse MM-RAG systems where target domains may differ from development environments (Saad-Falcon et al., 2024).

LLM-as-a-Judge: Scalability and Calibration Challenges: Beyond ARES, the broader LLM-as-a-Judge paradigm uses powerful models like GPT-4 to evaluate weaker models' outputs according to specific criteria (L. Zheng et al., 2023). This approach enables scalable evaluation but introduces systematic biases. LLMs exhibit position bias (preferring responses presented first), length bias (favoring longer outputs regardless of quality), and self-enhancement bias (rating their own outputs higher than alternatives). Recent work on judge calibration demonstrates that combining multiple judge models through ensemble voting reduces bias while maintaining correlation with human preferences (D. Li et al., 2025). For MM-RAG evaluation, calibrated LLM judges provide practical alternatives to human annotation when combined with ground-truth validation sets that anchor judgments to human standards.

6.2 Object-Level Hallucination Metrics: POPE and H-POPE

While automated frameworks assess overall system quality, detecting specific hallucination types requires targeted benchmarks. POPE (Polling-based Object Probing Evaluation) establishes a controlled framework for measuring object-level hallucinations through binary yes/no questions (Y. Li et al., 2023).

POPE Methodology: The evaluation proceeds in three steps. First, ground-truth objects are extracted from images either through human annotations (e.g., MSCOCO) or automatic segmentation tools like SEEM. Second, negative sampling generates questions about nonexistent objects under three difficulty settings.

Random sampling selects objects uniformly from the dataset vocabulary, testing whether models default to "yes" responses. **Popular sampling** selects frequently occurring objects (e.g., "chair," "table"), testing whether models hallucinate common objects based on statistical

priors. Adversarial sampling selects objects semantically related to image content but not actually present (e.g., asking "Is there a saddle?" when showing a horse without equipment), testing whether models infer objects from contextual cues. Third, models are queried with both positive questions (about present objects) and negative questions (about absent objects), yielding accuracy, precision, recall, and F1 scores.

POPE's binary format enables precise measurement and high reproducibility. Evaluation requires no complex parsing or subjective judgment. Models simply output 'yes' or 'no,' making inter-system comparison straightforward. Experiments reveal stark differences across models. Early vision-language models like LLaVA (v1) achieve only 50.13% accuracy on adversarial POPE, indicating frequent object hallucinations. InstructBLIP improves to 65.46% accuracy (F1: 73.75%) through instruction tuning on visual grounding data. Recent models incorporating contrastive decoding or visual attention mechanisms exceed 85% accuracy, demonstrating that architecture improvements directly reduce object-level errors (Y. Li et al., 2023).

Hierarchical-POPE (H-POPE): Standard POPE tests only object presence at a single abstraction level. H-POPE extends this methodology by introducing hierarchical probing across abstraction levels: superordinate categories (e.g., "vehicle"), basic-level categories (e.g., "car"), and subordinate categories (e.g., "sedan") (Pham & Schott, 2024). This granularity reveals where hallucinations occur in the recognition hierarchy. Models may correctly identify coarse categories ("Is there a vehicle?") yet hallucinate fine-grained distinctions ("Is it a sedan?" when it is actually an SUV). H-POPE results show that hallucination rates increase monotonically with specificity. Average accuracy drops from 88% at the superordinate level to 79% at basic level to 68% at subordinate level, highlighting that attribute hallucinations (Section 4.2) disproportionately affect fine-grained recognition.

Limitations: POPE and H-POPE measure only object category hallucinations. They cannot capture attribute errors (color, shape, count) or relation errors (spatial arrangements, interactions). Questions like "Is the car red?" or "Is the person next to the car?" require different evaluation protocols. Additionally, POPE focuses on object detection—verifying presence/absence—rather than open-ended generation. Models may pass POPE yet hallucinate extensively when generating free-form captions or answering complex visual questions (Bai et al., 2024).

6.3 Comprehensive Multi-Dimensional Benchmarks: MMHal-Bench and Beyond

To evaluate hallucinations beyond object detection, comprehensive benchmarks assess multiple error types through open-ended generation tasks.

MMHal-Bench: MMHal-Bench comprises 96 carefully curated image-question pairs across 12 object categories (Sun et al., 2024). Unlike POPE's binary questions, MMHal-Bench asks open-ended queries requiring detailed visual reasoning: "Describe the spatial arrangement of objects in this scene," "What activity is the person performing?", "Count the number of red items." Responses are evaluated using GPT-4 as a judge, which rates answers on a zero-to-six scale based on factual accuracy and visual grounding. The hallucination rate is computed as the proportion of responses scoring below three.

MMHal-Bench prioritizes diversity over dataset size. The 96 examples are specifically designed to probe known failure modes: attribute errors (color, material), relation errors (spatial positions, interactions), counting errors (numerosity), and reasoning errors (inferring activities or intentions). The results reveal significant hallucination rates in vision-language models on MMHal-Bench. Hallucination rates vary in open-source models such as LLaVA (Sun et al., 2024). Critically, MMHal-Bench correlates strongly (r=0.78) with human evaluations of

hallucination severity, validating GPT-4's effectiveness as a judge for this benchmark (Sun et al., 2024).

Caption Hallucination in Image Captioning (CHAIR) measures object hallucinations in generated captions by comparing mentioned objects against ground-truth annotations (Rohrbach et al., 2018). Two metrics quantify errors. CHAIR_I (instance-level) measures the proportion of hallucinated objects per caption: CHAIR_I = (hallucinated objects) / (mentioned objects). CHAIR_S (sentence-level) measures the proportion of captions containing at least one hallucination. CHAIR remains widely used for image captioning evaluation but shows high variance across prompt templates and struggles with semantically equivalent phrasings (e.g., "automobile" vs. "car").

GAVIE and HALLUCINOGEN: Recent benchmarks extend beyond object and attribute errors. GAVIE (Grounded Annotation for Video-based Image Evaluation) evaluates temporal hallucinations in video understanding, testing whether models fabricate events or actions not present in video sequences. HALLUCINOGEN introduces a systematic taxonomy covering six hallucination types: object presence, attribute correctness, spatial relations, numerical accuracy, inferential reasoning, and contextual coherence. By evaluating models across all six dimensions, HALLUCINOGEN reveals that mitigation strategies effective for one error type often fail on others, reinforcing the need for mechanism-specific interventions discussed in Section 5.5.

6.4 Community Standards and Shared Tasks: TREC RAG Track

Benchmark fragmentation hinders reproducible comparison across studies. Different papers use different datasets, evaluation metrics, and experimental setups, making it difficult to assess true progress. Community-driven shared tasks address this problem by establishing standardized evaluation protocols, curated test collections, and official leaderboards.

TREC RAG Track: The TREC 2024 and 2025 RAG Track provides the first large-scale, community-wide benchmark for end-to-end RAG system evaluation (Pradeep et al., 2025). The track defines three complementary tasks over the MS MARCO V2.1 corpus (tens of millions of web documents, hundreds of millions of text segments):

Retrieval (R) Task: Participants rank and retrieve the most relevant text segments for given queries. Evaluation uses standard retrieval metrics (nDCG@10, MAP, recall@1000) to measure segment-level relevance without generation.

Augmented Generation (AG) Task: Participants generate answers using a fixed set of top-k segments provided by a baseline retrieval system. This isolates generation quality from retrieval effectiveness, enabling focused evaluation of hallucination mitigation, source attribution, and answer completeness.

RAG Task: Participants implement end-to-end systems with custom retrieval and generation strategies. Outputs must map to MS MARCO segments for reproducibility. Evaluation measures both retrieval precision and generation quality jointly.

The track introduces nugget-based evaluation, originally developed for TREC Question Answering (Voorhees & Buckland, 2003) and adapted for RAG through the AutoNuggetizer (Pradeep et al., 2024). Nuggets represent atomic information units that constitute complete answers. Human assessors or LLMs identify nuggets in reference answers, then check whether system outputs cover these nuggets. Metrics include **nugget recall** (proportion of reference nuggets mentioned in system response) and **nugget precision** (proportion of system claims

supported by retrieved evidence). This granular evaluation detects not only hallucinations (unsupported claims) but also omissions (missing relevant facts).

Support Evaluation and LLM-Judge Calibration: TREC 2024 RAG Track conducted extensive support evaluation comparing GPT-4o judgments against human annotations across 45 participant submissions (Thakur et al., 2025). Support measures whether generated claims are grounded in cited passages. Results show strong correlation between automated LLM judgments and manual assessments, with run-level Kendall's τ of 0.783, though correlation decreases to 0.324 at topic-run level (Pradeep et al., 2024). While promising, this 44-56% disagreement rate highlights persistent challenges in LLM judge reliability. Error analysis reveals that LLMs struggle with nuanced inferential support (where claims require multi-hop reasoning across passages) and domain-specific terminology (where technical terms may be paraphrased differently). These findings motivate continued research on judge calibration and hybrid human-LLM evaluation workflows.

Ragnarök Framework: To support TREC RAG Track participation, the Ragnarök framework provides an open-source implementation of end-to-end RAG pipelines (Pradeep et al., 2025). Ragnarök standardizes input/output formats, integrates retrieval systems (BM25, dense retrievers, late-interaction models), and interfaces with generation backends (GPT-4o, Command R+, LLaMA 3.1). The framework includes a web-based arena for crowdsourced pairwise system comparison, enabling community evaluation beyond official track submissions. Ragnarök's release accelerates reproducibility by providing reference implementations and documented baselines that future study can build upon.

6.5 Dynamic Benchmarking and Data Contamination Mitigation

A critical challenge in RAG evaluation is data contamination: test queries or answers may exist within LLM training corpora, inflating performance without genuine retrieval (Sainz et al., 2023). If a model memorizes "What is the capital of France?" during pretraining, it can answer correctly without accessing retrieved evidence. This false positive undermines RAG evaluation, making systems appear effective when they simply regurgitate memorized facts.

Time-Stamped Datasets: Dynamic benchmarks mitigate contamination by incorporating information published after model training cutoffs. RGB (RAG Benchmark) includes questions about events occurring months after GPT-4's knowledge cutoff, forcing models to rely on retrieval (X. Zheng et al., 2025). Similarly, the TREC RAG Track refreshes query topics annually, ensuring that each year's evaluation includes novel questions unlikely to appear in training data. Results from RGB show that retrieval quality becomes the dominant factor for time-sensitive queries—models with strong parametric knowledge but weak retrieval perform worse than models with moderate parametric knowledge but strong retrieval.

Adversarial Filtering: J. Chen et al (2024) propose active contamination testing: deliberately include queries from popular QA datasets (Natural Questions, TriviaQA) in test sets, then flag systems that answer without retrieval. Models achieving suspiciously high accuracy on known-contaminated queries are penalized or excluded from leaderboards. This strategy deters data contamination by making it detectable and costly.

RAGAS: The RAGAS framework enables evaluation without ground-truth answers by assessing three aspects automatically (Es et al., 2024). Faithfulness measures whether generated claims are entailed by retrieved context, using natural language inference models to verify grounding. Context precision measures whether retrieved passages are relevant to the query, using similarity-based ranking. Answer relevance measures whether generated responses

address the user's question, using semantic similarity between query and answer. RAGAS correlates strongly with human judgments (Spearman ρ=0.71 for faithfulness, ρ=0.68 for relevance) while requiring no manual annotation during evaluation. This reference-free property makes RAGAS suitable for iterative development cycles where human labeling would create bottlenecks. Advancing MM-RAG research requires not only architectural innovation but also access to shared benchmarks, evaluation frameworks, and reproducible open-source implementations. Table 4 compiles essential open-source benchmarks and frameworks ensuring reproducibility in future research.

Table 4. Key open-source resources, benchmarks, and evaluation frameworks supporting reproducible research in MM-RAG

Category	Resource	Repository / Link	Description	Community Adoption
Vision Retrieval	ColPali Engine	github.com/illuin-tech/colpali	Late-interaction visual retrieval framework based on the PaliGemma backbone. Supports token-level pooling and multi-vector indexing.	400+ GitHub stars; PyPI package available
Vision Retrieval	ColPali Cookbooks	https://github.com/tonywu71/colpali-cookbooks	Tutorial notebooks for interpretability analysis and similarity map visualization.	Actively maintained examples
Benchmark	ViDoRe V1	huggingface.co/collections/vidore/vidore-benchmark-667173f98e70a1c0fa4db00d	Original document image retrieval benchmark with nine QA-style datasets.	Baseline benchmark in ColPali
Benchmark	ViDoRe V2	huggingface.co/collections/vidore/vidore-benchmark-v2-67ae03e3924e85b36e7f53b0	Extended benchmark introducing "blind" queries, multilingual support, and BEIR compatibility	Designed to address V1 saturation.Offers a more challenging evaluation standard.
Benchmark	ViDoRe V3	huggingface.co/collections/vidore/vidore-benchmark-v3	Large-scale enterprise benchmark with 26k pages, 3k queries, six languages, and human-verified labels.	Integrated into MTEB leaderboard
Evaluation	RAGAS	github.com/explodinggradients/ragas	Reference-free RAG evaluation framework measuring faithfulness, context precision, and answer relevance.	6.6k stars; EACL 2024

Evaluation	ARES	github.com/stanford-futuredata/ARES	Automated RAG evaluation using synthetic data generation and statistical confidence estimation.	NAACL 2024
Survey	MM-RAG Survey Repo	github.com/llm-lab-org/Multimodal-RAG-Survey	Living survey repository categorizing MM-RAG literature with continuous updates.	ACL 2025 Findings
Dataset	DocVQA	huggingface.co/datasets/vidore/docvqa_test_subsampled	Test subset from the DocVQA dataset (originally 12k+ images) adapted for visual retrieval benchmarking	Common baseline dataset
Dataset	InfographicsVQA	huggingface.co/datasets/vidore/infographicsvqa_test_subsampled	Test subset from the InfographicsVQA dataset targeting complex visual–text reasoning.	High visual–semantic difficulty
Platform	ViDoRe Leaderboard	huggingface.co/spaces/vidore/vidore-leaderboard	Public leaderboard tracking state-of-the-art visual retrieval models across benchmarks.	Real-time evaluation
Evaluation / Visualization	RAGAS	docs.ragas.io	Evaluation framework providing metrics and visualization integrations	Open-source (Apache 2.0). Enterprise cloud options available

As summarized in Table 5, recent progress in MM-RAG has been strongly enabled by the emergence of standardized benchmarks, open evaluation toolkits, and publicly accessible retrieval engines. Resources such as ViDoRe and RAGAS play a critical role in ensuring fair comparison and reproducibility across studies, while leaderboards facilitate continuous tracking of state-of-the-art performanc

7. Applications in High-Stakes Domains

The practical value of MM-RAG systems is measured not by laboratory benchmarks alone but by their reliability in high-stakes domains where errors carry severe consequences. This section examines three critical application areas—medical imaging, financial analysis, and legal document processing—where hallucinations can lead to misdiagnoses, financial losses, or legal liability. We analyze domain-specific challenges, architectural requirements, and empirical results from recent deployments.

7.1 Why These Domains? Shared Requirements and Distinct Challenges

Medical, financial, and legal domains share three characteristics that amplify hallucination risks while demanding exceptional accuracy. First, these domains require verifiable factual grounding. Medical diagnoses must align with established clinical guidelines and imaging

evidence. Financial analyses must reflect actual numerical data from reports and market indicators. Legal conclusions must cite specific statutes, case precedents, and contractual clauses. Parametric knowledge alone is insufficient—external evidence is mandatory. Second, these domains process inherently multimodal documents. Medical records pair radiology images with textual reports. Financial documents integrate balance sheets, trend charts, and management commentary. Legal contracts contain both standardized text clauses and attached exhibits including diagrams, floor plans, or financial schedules. Text-only RAG loses critical information. Third, these domains exhibit zero-error tolerance. A hallucinated drug dosage recommendation endangers patient safety (Kim et al., 2025). A fabricated earnings figure triggers regulatory violations and shareholder lawsuits. A misinterpreted contract clause exposes firms to litigation. Unlike consumer chatbots where occasional errors are annoying but tolerable, high-stakes applications require architectural designs that prioritize precision over fluency.

Despite shared requirements, each domain presents distinct technical challenges. Medical imaging demands real-time processing of high-resolution scans (e.g., CT images at 512×512×300 voxels) while maintaining diagnostic accuracy on rare pathologies with limited training data. Financial analysis requires numerical reasoning across complex tables and time-series charts, where OCR errors in decimal points or negative signs propagate catastrophically. Legal document processing involves recursive clause retrieval through hierarchical document structures, where missing a referenced definition or footnote invalidates entire contractual interpretations. These domain-specific constraints motivate specialized MM-RAG architectures beyond general-purpose systems.

7.2 Medical Imaging: Domain-Aware Retrieval and Visual Grounding

Medical MM-RAG (MMed-RAG), Xia et al. (2024) introduced MMed-RAG, a MM-RAG system specifically designed for medical vision-language models. MMed-RAG systems face a fundamental tension between generalization and specialization. General-purpose vision-language models (e.g., GPT-4V, Gemini) perform well on natural images but struggle with medical modalities where subtle visual cues determine diagnoses. A 2-millimeter lung nodule, barely visible to untrained observers, may indicate early-stage cancer. General models trained on web-scraped data lack exposure to such domain-specific patterns (Y. Li et al., 2023).

The framework addresses three critical failures in naive RAG applications to medicine.

- Cross-modal misalignment: When replacing input images with noisy corrupted versions, naive RAG systems retrieve context based on the original image but generate responses conditioned on the corrupted input. This produces confident hallucinations—responses that appear plausible but contradict visual evidence. MMed-RAG mitigates this through cross-modal consistency checks that verify retrieved evidence aligns with actual input modalities.
- Retrieval interference: Incorrectly retrieved contexts sometimes degrade performance even for queries the model could answer from parametric knowledge. MMed-RAG employs adaptive context selection, filtering retrieved passages below relevance thresholds rather than blindly injecting all results into generation.
- Domain shift: Medical imaging spans diverse modalities (radiology, pathology, ophthalmology) with distinct visual characteristics. A retriever trained on chest X-rays may fail on retinal scans. MMed-RAG implements domain-aware indexing that routes queries to modality-specific retrievers, improving recall across specialized subfields.

Experiments across five medical datasets (MIMIC-CXR, IU-Xray, Harvard-FairVLMed, PMC-OA, PathVQA) demonstrate substantial gains. According to Xia et al. (2024), MMed-RAG achieves an average improvement of 43.8% in factual accuracy across tasks, specifically improving medical VQA accuracy by 18.5% and report generation metrics by 69.1% over baseline. On radiology report generation, improvements reach 69.1% in BLEU score and 58.4% in ROUGE-L, indicating both fluency and factual correctness gains. With preference tuning (RAG-PT), the over-reliance rate dropped from 43.31% to 8.38%, directly addressing patient safety concern (Xia et al., 2024).

Visual RAG: Standard Med-LVLMs process single images, limiting their ability to compare findings across time series (e.g., tracking tumor growth) or correlate multiple imaging modalities (e.g., X-ray + CT scan). Chu et al. (2025) introduce Visual RAG (V-RAG), enabling models to retrieve and reason over multiple related images simultaneously. The approach fine-tunes models on image-text tasks that require multi-image comprehension. Entity probing evaluates whether specific medical entities (e.g., "pulmonary edema") are grounded in visual evidence. V-RAG significantly improves entity probing performance (measured in F1 score) on both frequent and rare entities compared to baselines, and downstream evaluation demonstrates a 19% relative improvement in the RadGraph-F1 score (Chu et al., 2025). This demonstrates that multi-image retrieval not only improves detection accuracy but also enhances generation factuality.

Agentic AI and Multi-Agent Systems: Recent radiology applications explore multi-agent architectures where specialized sub-models handle distinct reasoning steps. One agent performs image segmentation to localize anatomical structures. Another agent retrieves relevant case histories from electronic health records. A third agent synthesizes evidence and generates diagnostic hypotheses. A supervisor agent adjudicates conflicting predictions. This division of labor improves diagnostic accuracy by 8-15% over monolithic models while enabling fine-grained error attribution. However, multi-agent systems require careful orchestration to avoid compounding errors across stages. Evidence from 2024-2025 indicates these approaches remain computationally expensive and lack comprehensive clinical validation, limiting near-term deployment (Rabbani et al., 2025).

7.3 Financial Analysis: Chart-to-Markdown and Hybrid Retrieval

Financial documents present unique challenges for MM-RAG systems. Balance sheets contain hundreds of numerical entries where a single OCR error (e.g., "8.5%" → "85%") invalidates downstream analysis. Trend charts convey growth patterns that text descriptions cannot adequately capture. Management commentary provides contextual narrative essential for interpreting raw figures. Effective financial RAG must jointly process text, tables, and charts while preserving numerical precision.

Chart-to-Markdown Conversion: Jiang et al. (2025) proposed a MM-RAG framework that converts chart and table images into structured Markdown representations prior to indexing. While the authors do not publish explicit numerical examples, they demonstrate that financial tables (e.g., quarterly results) can be faithfully transformed into structured row–column formats, preserving numerical relations and table semantic. This structured representation enables precise retrieval. Queries like "What was Q2 revenue growth?" retrieve the exact table cell rather than noisy text fragments. Experiments on proprietary financial datasets demonstrate improvements in retrieval precision (Precision@10 increases from 0.36 to 0.40) and generation accuracy compared to OCR-based baselines (C. Jiang et al., 2025). The full multimodal strategy effectively addresses the primary failure mode (OCR-induced numerical errors) in financial applications.

Hybrid Retrieval (Vector + Graph Databases): Financial analysis often requires multi-hop reasoning. A query about "debt-to-equity ratio" requires retrieving balance sheet data (total debt, shareholder equity), computing the ratio, then comparing against industry benchmarks. Vector databases enable semantic search but lack structured reasoning. Jiang et al. (2025) augment vector retrieval with graph databases that encode relationships between financial concepts (e.g., "Revenue -> Operating Income -> Net Income"). Queries trigger both vector similarity search and graph traversal. Retrieved contexts include semantically relevant passages plus structurally related entities. This hybrid approach improves multi-hop question accuracy by 23% on financial QA benchmarks.

FinRAGBench-V (Visual Citation Benchmark): Evaluating financial RAG systems requires not only answer accuracy but also source attribution. Users must verify which specific chart or table supports each claim. FinRAGBench-V introduces the first benchmark requiring visual citations—generated responses must cite exact page regions (bounding boxes) supporting claims (Zhao et al., 2025). Experiments reveal that multimodal retrievers outperform text-only approaches on visual document retrieval tasks, with evaluation showing performance variations across models on financial document benchmarks (Zhao et al., 2025). This benchmark establishes a new standard for trustworthy financial RAG deployment.

Real-World Impact: Financial institutions have deployed RAG systems for extracting structured information from complex documents, demonstrating substantial improvements in analyst productivity and accuracy on domain-specific extraction tasks

7.4 Legal Document Processing: Recursive Retrieval and Clause Dependencies

Legal contracts exhibit hierarchical structures where clauses reference other clauses, definitions, exhibits, and footnotes. Understanding a single clause may require recursively retrieving and synthesizing information scattered across the document. Traditional flat retrieval paradigms fail because they treat documents as unstructured text collections, ignoring internal dependencies.

Multi-Graph Recursive Retrieval: Yang (2024) proposes a multi-agent system for legal RAG that constructs multiple document graphs. A clause graph captures hierarchical structure (e.g., Section 5.2.3 is a child of Section 5.2). A definition graph links terms to their formal definitions. A reference graph tracks cross-references (e.g., "as defined in Section 3.1"). Given a query, agents traverse these graphs recursively. Retrieving Clause A may trigger retrieval of Definition B, which references Exhibit C. The system terminates when no further linked nodes are relevant or when recursive depth exceeds limits to prevent infinite loops. Experiments on commercial contracts (500+ pages) show that recursive retrieval improves comprehension accuracy by 31% compared to naive chunking strategies that sever clause dependencies.

Addleshaw Goddard: A major UK law firm deployed an optimized RAG system for commercial contract analysis, achieving 95% accuracy compared to 74% for baseline LLMs (Addleshaw Goddard LLP, 2024). Three optimizations drove performance gains. Optimized retrieval: Category-aware chunking groups related clauses (e.g., all indemnification provisions) into coherent units, improving retrieval relevance by ~20%. Keyword prompting: Instructions directing LLMs to focus on domain-specific terms (e.g., "force majeure," "liquidated damages") improved recall accuracy by ~16%. Follow-up prompting: After initial generation, a second prompt asks the model to verify claims against retrieved evidence, reducing hallucinations by 9.2%. The system processes 500-page merger agreements in 12 minutes compared to 4-6 hours for manual review, enabling lawyers to focus on strategic analysis rather than mechanical clause identification (Addleshaw Goddard LLP, 2024).

Multi-Round RAG for Comprehensive Analysis: Complex legal queries often require iterative refinement. A lawyer asks "What are the termination provisions?" The system retrieves relevant clauses but realizes some reference broader definitions of "material breach" defined elsewhere. Multi-round RAG iteratively expands context by identifying undefined terms, retrieving their definitions, then re-generating responses with complete information. Experiments show multi-round approaches improve completeness (covering all relevant provisions) by 27% while maintaining precision (avoiding irrelevant information) through dynamic stopping criteria that terminate retrieval when additional rounds yield diminishing returns(“Application of RAG Model Based on Retrieval Enhanced Generation Technique in Complex Query Processing,” 2024).

Hallucination Risks in Legal RAG: In a study(Magesh et al., 2025) three commercial legal RAG systems were evaluated: Lexis+ AI and Ask Practical Law AI hallucinated in approximately 17% of queries (one in six responses), while Westlaw AI exhibited hallucinations in 33% of responses (one in three). These findings highlight the need for continuous evaluation and human oversight in legal AI applications.

7.5 Resources and Reproducibility

Community Platforms: Beyond individual tools, standardized platforms facilitate reproducible comparisons. The TREC RAG Track (Section 6.4) provides shared evaluation protocols. Hugging Face hosts 150+ domain-specific embedding models. GitHub repositories like RAG-Anything (HKU, 2024) provide end-to-end pipelines integrating document parsing, multimodal retrieval, and generation.

Reproducibility Checklist for Domain Applications:

1. Dataset Transparency: Specify training data sources, annotation procedures, and licensing constraints
2. Evaluation Protocols: Report metrics on standardized benchmarks (POPE, MMHal-Bench, domain-specific tasks)
3. Computational Requirements: Document GPU memory, inference latency, and indexing costs
4. Error Analysis: Conduct domain-expert evaluations beyond automated metrics to identify failure modes
5. Ethical Safeguards: Implement human-in-the-loop verification for high-stakes decisions

Adhering to these practices accelerates community progress while ensuring safe deployment in critical domains.

8. Conclusion

This survey systematically analyzed MM-RAG through the intersecting dimensions of architectural design, hallucination mechanisms, mitigation strategies, evaluation protocols, and high-stakes applications. Our synthesis reveals a field transitioning from exploratory research toward systematic engineering, where architectural choices, error patterns, and evaluation methodologies are increasingly well-understood. This concluding section distills key insights, clarifies trade-offs, and charts future research directions that will shape the next generation of MM-RAG systems.

8.1 Key Insights: What We Have Learned

Architectural Evolution Follows a Clear Trajectory. The progression from discrete OCR-based pipelines to unified vision-language embeddings, and increasingly toward late-interaction and agentic paradigms, reflects fundamental trade-offs between semantic fidelity, computational cost, and hallucination risk. OCR-based systems prioritize engineering simplicity and compatibility with legacy infrastructure but sacrifice spatial layout and structural semantics. Dense multimodal embeddings (CLIP, SigLIP) enable cross-modal alignment but compress information into fixed-size vectors, losing fine-grained details. Late-interaction models (ColPali, ColBERT) preserve granularity through multi-vector representations but increase storage and indexing costs. No single architecture dominates across all scenarios. Optimal design depends on document characteristics, query distribution, and operational constraints. This architectural diversity is not a weakness but a strength—it enables practitioners to select or compose approaches matched to specific application requirements. Hallucination Is Inherent, Not Eliminable, but Manageable. Probabilistic language models inherently produce outputs that occasionally diverge from evidence. In multimodal contexts, this risk amplifies through cross-modal misalignment, cascading error propagation, and retrieval-specific information loss. However, hallucinations are not uniform. Object category errors stem from different mechanisms than attribute errors, which differ from relation errors. Fabrications (describing nonexistent content) arise from overreliance on linguistic priors, while omissions (failing to describe present content) result from low-confidence visual encodings. This mechanistic diversity demands tailored interventions. CoVe reduces object hallucinations through iterative verification but provides limited benefit for attributes. Visual grounding techniques reduce attribute errors but struggle with relations. Confidence calibration enables systems to recognize uncertainty and abstain from answering rather than hallucinate. Production systems must integrate multiple complementary strategies, creating layered defenses that address distinct failure modes. The goal shifts from eliminating hallucinations to detecting, quantifying, and controlling them within acceptable risk thresholds. Evaluation Has Matured but Remains Fragmented. The field has progressed from anecdotal demonstrations to systematic, reproducible assessment. POPE establishes binary object-level evaluation as a widely adopted baseline. MMHal-Bench extends coverage to attributes, relations, and reasoning. TREC RAG Track introduces community-wide standards with shared tasks and official leaderboards. ARES and RAGAS automate evaluation through synthetic data generation and reference-free metrics, enabling rapid iteration during development. However, fragmentation persists. Different benchmarks measure different hallucination types. Medical, financial, and legal evaluations remain incomparable due to domain-specific metrics and proprietary datasets. Dynamic benchmarks mitigate data contamination but lack standardization across studies. Bridging these evaluation silos requires unified protocols that assess retrieval precision, generation faithfulness, source attribution accuracy, and domain-specific requirements within a single framework. Standardization efforts like TREC RAG Track represent critical infrastructure investments that accelerate progress by enabling fair comparisons. High-Stakes Applications

Reveal Persistent Gaps. Medical imaging, financial analysis, and legal document processing demonstrate MM-RAG's practical potential while exposing limitations. Domain-aware retrievers (MMed-RAG, Visual RAG) improve performance over general-purpose models, confirming that specialized architectures outperform one-size-fits-all approaches. MMed-RAG improves factual accuracy by 18.5% over baseline Med-LVLMs on medical VQA tasks (Xia et al., 2024). Chart-to-Markdown conversion improves answer correctness scores (evaluated via RAGAS) by approximately 28% and retrieval precision by approximately 19% compared to OCR-based baselines in financial contexts, addressing the primary failure mode where OCR misreads decimal points or negative sign (C. Jiang et al., 2025). Multi-graph recursive retrieval improves legal clause comprehension by 31% by preserving hierarchical dependencies that flat retrieval ignores. However, hallucination risks require careful mitigation. Commercial legal RAG tools fabricate citations in 17-33% of responses (Magesh et al., 2025). In medical contexts, baseline systems without retrieval safeguards exhibit over-reliance rates of 43.31%, where models incorrectly trust noisy retrieved evidence. Domain-aware RAG architectures like MMed-RAG reduce this rate to 8.38% through adaptive context filtering and cross-modal consistency checks (Xia et al., 2024). demonstrating that specialized architectures substantially improve reliability. These failures underscore that achieving human-level reliability in high-stakes domains requires not only better models but also architectural safeguards, human-in-the-loop verification, and rigorous certification processes.

8.2 Future Directions: Toward the Next Generation

The convergence of architectural advances, foundation model scaling, and structured knowledge integration points toward several transformative research directions that will define the next generation of MM-RAG systems.

Agentic RAG: Traditional RAG operates as a passive retrieve-then-generate pipeline. Agentic RAG transforms this into an autonomous problem-solving process where systems plan multi-step workflows, dynamically select tools, and iteratively refine outputs based on feedback. Multi-agent architectures distribute complex tasks across specialized sub-agents—one agent performs retrieval, another verifies factuality, a third synthesizes evidence, and a supervisor coordinate outputs. Recent study demonstrates that agentic systems improve accuracy by 8-15% on complex reasoning tasks requiring multi-hop inference, though computational costs increase proportionally. Critical research challenges include coordinating agent communication without error compounding, designing stopping criteria that prevent infinite verification loops, and developing evaluation benchmarks that measure agentic capabilities (planning quality, tool selection accuracy, adaptive refinement) beyond traditional QA metrics. The shift toward agentic paradigms represents not merely an incremental improvement but a fundamental reconceptualization of RAG from reactive information retrieval to proactive knowledge synthesis.

Hybrid Reasoning: Emerging frameworks distinguish between System 1 reasoning (fast, intuitive, pattern-based) and System 2 reasoning (slow, deliberate, logic-based) in RAG contexts (Liang et al., 2025). Simple factual queries trigger lightweight retrieval with single-pass generation (System 1). Complex multi-hop questions activate iterative retrieval, chain-of-thought reasoning, and verification loops (System 2). Adaptive routers predict query complexity and allocate computational resources, accordingly, maximizing accuracy-efficiency trade-offs. Preliminary results show that hybrid systems achieve 90-95% of full System 2 performance while reducing average latency by 40-60% through intelligent routing. Future research must develop query complexity classifiers that generalize across domains, design System 1-System 2 interfaces that enable graceful escalation when fast reasoning proves

insufficient and establish test-time compute budgets that optimize resource allocation dynamically based on task requirements and user preferences.

Graph-Enhanced Multimodal RAG (GEAR): Knowledge graphs provide structured representations of entities and relationships, complementing neural retrieval's semantic flexibility with symbolic reasoning's logical consistency. GEAR systems index documents as graph nodes, with edges representing citations, temporal sequences, or semantic relations. Queries trigger both vector similarity search (retrieving semantically similar nodes) and graph traversal (following explicit relationships). This hybrid approach improves multi-hop reasoning by 23-31% compared to vector-only retrieval while reducing hallucinations through structured knowledge constraints. Challenges include constructing high-quality domain-specific knowledge graphs without prohibitive manual annotation, maintaining graph consistency as new documents arrive, and designing joint embedding spaces where neural and symbolic representations interact seamlessly. Graph-enhanced RAG represents a promising direction for domains requiring verifiable reasoning chains, such as legal precedent analysis, scientific literature review, and medical differential diagnosis.

Privacy-Preserving and Federated MM-RAG: High-stakes domains demand data sovereignty. Medical records and financial documents cannot be routed through cloud-based retrieval services without violating privacy regulations (GDPR, HIPAA). Federated RAG architectures enable on-device or on-premises retrieval where data never leaves secure environments. Queries are processed locally against partitioned indices, with only aggregated results (not raw documents) transmitted to generation servers. Differential privacy techniques add calibrated noise to queries and retrieved contexts, providing mathematical privacy guarantees while degrading generation quality minimally (typically 5-10% accuracy loss). Technical challenges include maintaining retrieval precision with locally partitioned indices that lack global corpus statistics, balancing privacy budgets against downstream task performance, and developing secure multi-party computation protocols that enable collaborative retrieval without exposing sensitive data. As regulatory scrutiny intensifies, privacy-preserving RAG will transition from optional enhancement to mandatory infrastructure requirement.

Test-Time Compute Scaling and Adaptive Resource Allocation: ReRecent studies demonstrate that allocating more computation during inference—through multi-sample generation, iterative refinement, or ensemble methods—improves reasoning quality substantially (Y. Li et al., 2025). However, optimal compute allocation strategies remain unexplored for MM-RAG. Should systems invest computation in better retrieval, more generation samples, or longer verification loops? Future adaptive frameworks may employ resource routers that dynamically allocate budgets across retrieval granularity (coarse vs. fine-grained embeddings), fusion depth (cross-attention vs. late interaction), and verification intensity (activating CoVe loops only for low-confidence queries). Preliminary simulations suggest that adaptive allocation could improve accuracy by 12-18% at constant computational cost compared to fixed uniform allocation. Challenges include designing differentiable routers that optimize resource distribution end-to-end, establishing cost-benefit functions that trade accuracy against latency across diverse tasks, and creating evaluation benchmarks that measure not only final output quality but also computational efficiency and carbon footprint.

Embodied and Multimodal Any-to-Any Systems: The frontier extends beyond static documents toward embodied agents that perceive, reason, and act in physical environments (Abootorabi et al., 2025). Embodied-RAG systems index multimodal episodes (vision, audio, proprioception) as hierarchical semantic forests, enabling cross-granularity retrieval for both navigation ("How do I reach the kitchen?") and explanation ("Why did you take this route?"). Video-RAG frameworks handle long-form video comprehension by decoupling queries into modality-

specific retrieval requests—OCR for on-screen text, ASR for spoken dialogue, object detection for physical entities (Luo et al., 2024). These auxiliary inputs are indexed dynamically, allowing systems to handle temporal redundancy without exceeding context limits. Challenges include maintaining temporal consistency across retrieved video segments, fusing heterogeneous modalities (audio, visual, haptic) in unified representation spaces, and developing evaluation protocols that assess not only factual accuracy but also coherence in multimodal storytelling. As AI systems transition from text-centric assistants to multimodal embodied agents, RAG architectures must evolve accordingly.

Fine-Grained Source Attribution and Provenance Tracking: Current MM-RAG systems cite documents broadly, returning page numbers or document IDs. Users must manually verify which specific text span, image region, or table cell supports each generated claim. Fine-grained attribution requires visual grounding models that link textual claims to bounding boxes or pixel masks, provenance tracking through multi-hop reasoning chains that document every intermediate retrieval step, and interactive interfaces where users can drill down from high-level summaries to atomic evidence units. Recent benchmarks like FinRAGBench-V establish visual citation as an evaluation standard, revealing persistent gaps—GPT-4o achieves only 78% citation precision compared to ground truth. Achieving citation precision comparable to academic footnotes (>95% accuracy) represents a long-term research goal essential for trustworthy deployment in high-stakes domains.

8.3 Closing Perspective: From Research to Reliable Systems

MM-RAG stands at an inflection point. Many foundational questions are increasingly well-understood. We know how to align text and images. We know how to reduce hallucinations. We know how to evaluate systems. The next phase focuses on engineering these capabilities into reliable, efficient, deployable infrastructures. This requires not just algorithmic advances but also standardization efforts, reproducible baselines, public benchmarks for high-stakes domains, and documented best practices for system design. The research agenda must shift from demonstrating feasibility to ensuring reliability, from maximizing performance to controlling risks, from academic exploration to production deployment. Hybrid architectures that dynamically balance modalities, adaptive systems that route queries to appropriate reasoning modes, and layered defenses that detect and mitigate errors before outputs reach users will define the next generation. MM-RAG's maturation depends on the field's ability to make this transition—from impressive demonstrations to trustworthy tools that augment human decision-making in critical domains. The convergence of agentic intelligence, foundation model scaling, and structured knowledge promises transformative applications. Realizing this potential requires sustained collaboration across academia, industry, and regulatory bodies to address technical challenges, ethical considerations, and societal implications. The path forward is clear, though substantial work remains.

REFERENCES

Abootorabi, M. M., Zobeiri, A., Dehghani, M., Mohammadkhani, M., Mohammadi, B., Ghahroodi, O., Baghshah, M. S., & Asgari, E. (2025). *Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2502.08826

Addleshaw Goddard LLP. (2024). *The RAG report: Can large language models be good enough for legal due diligence?* [Technical Report]. Addleshaw Goddard LLP. https://www.addleshawgoddard.com/globalassets/insights/technology/llm/rag-report.pdf

Application of RAG Model Based on Retrieval Enhanced Generation Technique in Complex Query Processing. (2024). *Advances in Computer, Signals and Systems*, *8*(6). https://doi.org/10.23977/acss.2024.080608

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). *Self-rag: Learning to retrieve, generate, and critique through self-reflection*.

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). *Hallucination of Multimodal Large Language Models: A Survey* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2404.18930

Chen, B., Wongso, W., Hu, X., Tan, Y., & Salim, F. (2025). *Multi-Stage Verification-Centric Framework for Mitigating Hallucination in Multi-Modal RAG* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2507.20136

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728

Chen, W., Hu, H., Chen, X., Verga, P., & Cohen, W. (2022). MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5558–5570. https://doi.org/10.18653/v1/2022.emnlp-main.375

Chen, W., Yu, W., Qi, G., Li, W., Li, Y., Sha, L., Xia, D., & Huang, J. (2025). *CMRAG: Co-modality-based visual document retrieval and question answering* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2509.02123

Chen, X., Li, Y., Hu, M., Salari, E., Chen, X., Qiu, R. L. J., Zheng, B., & Yang, X. (2024). *Mammo-CLIP: Leveraging Contrastive Language-Image Pre-training (CLIP) for Enhanced Breast Cancer Diagnosis with Multi-view Mammography* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2404.15946

Chu, Y.-W., Zhang, K., Malon, C., & Min, M. R. (2025). *Reducing Hallucinations of Medical Multimodal Large Language Models with Visual Retrieval-Augmented Generation* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2502.15040

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-Verification Reduces Hallucination in Large Language Models. *Findings of the Association for Computational Linguistics ACL 2024*, 3563–3578. https://doi.org/10.18653/v1/2024.findings-acl.212

Drushchak, N., Polyakovska, N., Bautina, M., Semenchenko, T., Koscielecki, J., Sykala, W., & Wegrzynowski, M. (2025). Multimodal Retrieval-Augmented Generation: Unified Information Processing Across Text, Image, Table, and Video Modalities. *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, 59–64. https://doi.org/10.18653/v1/2025.magmar-1.5

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. *Proceedings of the 18th Conference of the*

European Chapter of the Association for Computational Linguistics: System Demonstrations, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., & Soatto, S. (2024). Multi-modal hallucination control by visual information grounding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14303–14312.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., & Colombo, P. (2024). *ColPali: Efficient Document Retrieval with Vision Language Models* (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2407.01449

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv Preprint arXiv:2312.10997*, *2*(1).

Ghosh, S., Evuru, C. K. R., Kumar, S., Tyagi, U., Nieto, O., Jin, Z., & Manocha, D. (2024). *Visual Description Grounding Reduces Hallucinations and Boosts Reasoning in LVLMs* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2405.15683

Grassucci, E., Cicchetti, G., & Comminiello, D. (2025). Closing the modality gap enables novel multimodal learning applications. *Second Workshop on Representational Alignment at ICLR 2025*. https://openreview.net/forum?id=P3Oba8Z3B6

Gupta, S., Ranjan, R., & Singh, S. N. (2024). *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2410.12837

Huang, X., Wong, Y. X., Goh, G. L., Gao, X., Lee, J. M., & Yeong, W. Y. (2024). Machine learning-driven prediction of gel fraction in conductive gelatin methacryloyl hydrogels. *International Journal of AI for Materials and Design*, *1*(2), 61. https://doi.org/10.36922/ijamd.3807

Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. C. (2024). *Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2403.14403

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023). Towards Mitigating LLM Hallucination via Self Reflection. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1827–1843. https://doi.org/10.18653/v1/2023.findings-emnlp.123

Jiang, C., Zhang, P., Ni, Y., Wang, X., Peng, H., Liu, S., Fei, M., He, Y., Xiao, Y., Huang, J., Ma, X., & Yang, T. (2025). Multimodal retrieval-augmented generation for financial documents: Image-centric analysis of charts and tables with large language models. *The Visual Computer*, *41*(10), 7657–7670. https://doi.org/10.1007/s00371-025-03829-5

Jiang, Z., Ma, X., & Chen, W. (2024). *LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2406.15319

Khattab, O., & Zaharia, M. (2020). *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2004.12832

Kim, Y., Jeong, H., Chen, S., Li, S. S., Park, C., Lu, M., Alhamoud, K., Mun, J., Grau, C., Jung, M., Gameiro, R., Fan, L., Park, E., Lin, T., Yoon, J., Yoon, W., Sap, M., Tsvetkov, Y., Liang, P., … Breazeal, C. (2025). *Medical Hallucinations in Foundation Models and Their Impact on Healthcare* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2503.05777

Lee, S., Yu, S., Park, J., Yi, J., & Yoon, S. (2024). Interactive Text-to-Image Retrieval with Large Language Models: A Plug-and-Play Approach. *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 791–809. https://doi.org/10.18653/v1/2024.acl-long.46

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., & others. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., & Liu, H. (2025). From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2757–2791. https://doi.org/10.18653/v1/2025.emnlp-main.138

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). *Evaluating Object Hallucination in Large Vision-Language Models* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2305.10355

Li, Y., Fu, X., Verma, G., Buitelaar, P., & Liu, M. (2025). *Mitigating Hallucination in Large Language Models (LLMs): An Application-Oriented Survey on RAG, Reasoning, and Agentic Systems* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2510.24476

Liu, F., Eisenschlos, J. M., Piccinno, F., Krichene, S., Pang, C., Lee, K., Joshi, M., Chen, W., Collier, N., & Altun, Y. (2022). *DePlot: One-shot visual language reasoning by plot-to-table translation* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2212.10505

Lumer, E., Cardenas, A., Melich, M., Mason, M., Dieter, S., Subbiah, V. K., Basavaraju, P. H., & Hernandez, R. (2025). *Comparison of Text-Based and Image-Based Retrieval in Multimodal Retrieval Augmented Generation Large Language Model Systems* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2511.16654

Luo, Y., Zheng, X., Li, G., Yin, S., Lin, H., Fu, C., Huang, J., Ji, J., Chao, F., Luo, J., & Ji, R. (2024). *Video-RAG: Visually-aligned Retrieval-Augmented Long Video Comprehension* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2411.13093

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, *22*(2), 216–242. https://doi.org/10.1111/jels.12413

Most, A., Winjum, J., Bhattarai, M., Jones, S., Ranasinghe, N. R., Biswas, A., & O'Malley, D. (2025). Lost in OCR Translation? Vision-Based Approaches to Robust Document Retrieval. *Proceedings of the 2025 ACM Symposium on Document Engineering*, 1–10. https://doi.org/10.1145/3704268.3742698

Peng, X., Qin, C., Chen, Z., Xu, R., Xiong, C., & Wu, C.-S. (2025). *UNIDOC-BENCH: A Unified Benchmark for Document-Centric Multimodal RAG* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2510.03663

Pradeep, R., Thakur, N., Sharifymoghaddam, S., Zhang, E., Nguyen, R., Campos, D., Craswell, N., & Lin, J. (2025). Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track. In C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, & N. Tonellotto (Eds.), *Advances in Information Retrieval* (Vol. 15572, pp. 132–148). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-88708-6_9

Pradeep, R., Thakur, N., Upadhyay, S., Campos, D., Craswell, N., & Lin, J. (2024). *Initial Nugget Evaluation Results for the TREC 2024 RAG Track with the AutoNuggetizer Framework* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2411.09607

Rabbani, S. A., El-Tanani, M., Sharma, S., Rabbani, S. S., El-Tanani, Y., Kumar, R., & Saini, M. (2025). Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions. *BioMedInformatics*, *5*(3), 37. https://doi.org/10.3390/biomedinformatics5030037

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & others. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 338–354. https://doi.org/10.18653/v1/2024.naacl-long.20

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., & Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10776–10787.

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.-X., Yang, Y., Keutzer, K., & Darrell, T. (2024). Aligning Large Multimodal Models with Factually Augmented RLHF. *Findings of the Association for Computational Linguistics ACL 2024*, 13088–13110. https://doi.org/10.18653/v1/2024.findings-acl.775

Thakur, S., Sharma, S., Goyal, R., Sharma, M., Uttrani, S., & Dutt, V. (2025). Evaluating GPT-4o as a Cyberattack Simulator: Perspectives on AI and Human Decision-Making. *2025 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, 79–86. https://doi.org/10.1109/CogSIMA64436.2025.11079523

Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv Preprint arXiv:2401.01313*, *6*.

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., & Zhai, X. (2025). *SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2502.14786

Voorhees, E. M., & Buckland, L. (2003). Overview of the TREC 2003 question answering track. *TREC*, *2003*, 54–68.

Wasserman, N., Pony, R., Naparstek, O., Goldfarb, A. R., Schwartz, E., Barzelay, U., & Karlinsky, L. (2025). *REAL-MM-RAG: A Real-World Multi-Modal Retrieval Benchmark* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2502.12342

Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J., & Yao, H. (2024). *MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2410.13085

Yang, Y., Cai, Z., Qiu, S., & Xu, P. (2024). Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image. *PLOS ONE*, *19*(3), e0299265. https://doi.org/10.1371/journal.pone.0299265

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2025). Evaluation of Retrieval-Augmented Generation: A Survey. In W. Zhu, H. Xiong, X. Cheng, L. Cui, Z. Dou, J. Dong, S. Pang, L. Wang, L. Kong, & Z. Chen (Eds.), *Big Data* (Vol. 2301, pp. 102–120). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-1024-2_8

Zhai, W. (2024). *Self-adaptive Multimodal Retrieval-Augmented Generation* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2410.11321

Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhao, S., Jin, Z., Li, S., & Gao, J. (2025). *FinRAGBench-V: A Benchmark for Multimodal RAG with Visual Citation in the Financial Domain* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2505.17471

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 46595–46623). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

Zheng, X., Weng, Z., Lyu, Y., Jiang, L., Xue, H., Ren, B., Paudel, D., Sebe, N., Van Gool, L., & Hu, X. (2025). *Retrieval Augmented Generation and Understanding in Vision: A Survey and New Outlook* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2503.18016

CHAPTER 3

EXPLAINABILITY OF TRANSFORMERS-BASED MODELS WITH EXPLAINABLE ARTIFICIAL INTELLIGENCE METHODS: EXAMPLE OF BERT TECHNIQUE

Tunahan TİMUÇİN[1]

1 Assist. Prof. Dr., Düzce University, Computer Engineering, ORCID: 0000-0003-0332-4118

1. Introduction

As Artificial Intelligence (AI) technologies rapidly permeate human life, they have paved the way for significant developments and breakthroughs in many fields. One of the most affected areas is natural language processing (NLP). The integration of transformer-based models into NLP has led to a revolutionary development. Furthermore, these high-level technologies have enabled the development of even higher-level and more complex technologies in various fields. Explainable Artificial Intelligence (XAI) is one of these indirectly developing technologies. This study contributes to the literature by establishing a solid foundation for the concepts of NLP, XAI, and transformer-based BERT. The focus is on showcasing current research in these fields and discussing their integration. The BERT model, the best-known of the transformer-based models introduced by Devlin et al. in 2018, has a mechanism that makes bidirectional deep inferences from unlabeled text. This feature has elevated the field of NLP to an extraordinary level. BERT differs from previous models, particularly in its ability to better understand context (Devlin et al., 2018). In the BERT model, where pre-training of language representations is facilitated for fine-tuning certain tasks such as question answering and language inference, no significant changes to the model structure are required during this process. BERT's contributions are not limited to performance improvement. It also offers new standards for evaluation in various applications such as sentiment analysis and named entity recognition. Explainable Artificial Intelligence (XAI) is the name given to the method developed to understand the decision-making mechanisms behind artificial intelligence systems. XAI is a significant technology that has gained importance, particularly in areas defined as high-risk, such as finance and healthcare, due to its ability to provide transparency (Belghachi, 2023; Minh et al., 2021). Predictions made using artificial intelligence are rapidly spreading to all fields. This spread has created a new need: the need to interpret artificial intelligence systems. This technology, XAI, has emerged from this need. Artificial intelligence models, known by their nature as "black boxes," are difficult to interpret due to this characteristic. This results in the concealment of the workings behind decisions (Gilpin et al., 2018).

Various techniques have been developed to overcome these difficulties and interpret the outputs of artificial intelligence models. Pre-modeling, post-modeling, and interpretable models are general examples of these techniques (Minh et al., 2021; Bienefeld et al., 2023; Belle & Papantonis, 2021). LIME and SHAP techniques for model prediction also stand out as techniques offering important insights (Kapcia et al., 2021; Tiwari, 2023). These models also have applications in the healthcare sector. Their importance is further increased because patient outcomes can change with the interpretation of these results (Chaddad et al., 2023).
Considering this importance, it is necessary to increase the interpretability of NLP models. At this point, the integration of XAI and BERT techniques gains importance. The use of XAI techniques to increase the transparency of advanced and complex decision systems, especially transformer-based models like BERT, has been a focus of research (Auletta et al., 2023).

While combining XAI and BERT technologies offers many benefits, it also presents several challenges. One of the most important is the high internal complexity of transformer-based models. This can make it difficult to clearly interpret the outputs (Saranti et al., 2022). Currently, the fact that the least interpretable models are often the most accurate also presents the problem of striking a good balance between model performance and explainability (Tiwari, 2023; Hu & Wu, 2023).

In summary, while the field of XAI is promising for improving the interpretability of models like NLP and BERT, there are still a number of challenges to overcome. In this respect, it is seen that researchers are encouraged to develop new XAI techniques and solve the problems of transformative architectures (Linardatos et al., 2020; Mohseni et al., 2021). In addition, it is important to develop user-friendly applications by interpreting XAI techniques in all areas, especially health and finance, in order to promote transparency, understanding and most importantly trust among stakeholders (Kapcia et al., 2021; Hu et al., 2021).

The intersection and interaction of natural language processing, BERT and NLP systems will provide the formation of a rapidly developing and dynamic field. Continuous research and development of this synergy will allow the examination of the black box feature, which is the nature of modern Artificial Intelligence, to overcome its difficulties.

The second section includes a literature review on the subject, the methods used in the third section are explained, and the results obtained in the fourth section are given. Finally, the study is completed with the conclusion section.

2. Related works

Natural language processing (NLP) is a technology that emerged within artificial intelligence as a critical area for computers to understand, interpret and produce human language. Especially with its interaction with Machine Learning (ML) techniques, NLP has evolved in algorithm and methodology parts. Weber and (Ranchhod and Mamede), who provided a comprehensive general review of early developments in NLP, emphasized especially in semantic understanding and syntactic parsing areas (Webber, 1986; Ranchhod & Mamede, 2002). Another study that summarizes the development of NLP over the years belongs to Mote. Mote emphasized that it is necessary to focus on more complex models using deep learning (Mote, 2012). NLP is applied in many areas. The first of these is the health field. Roy et al. mentioned in their study that NLP can be used for various tasks such as patient interaction and clinical documentation (Roy et al.,2021). Al-Garadi et al. have stated how NLP analyzes patient data in times of COVID-19 and how it works as a system that serves the epidemic by responding quickly to the necessary response in times of crisis (Al-Garadi et al., 2022). Velupillai et al. and Savova et al. have described how electronic health records interact with NLP (Mowery et al., 2015; Savova et al., 2019). These researchers have demonstrated NLP's ability to extract meaningful data from unstructured data. In their article, Raza et al. explain how machine learning algorithms have become the cornerstone of NLP, they have examined the application of various ML algorithms to NLP tasks (Raza et al., 2023). Referring to models such as GPT (Generative Pre-trained Transformer) and BERT, Singh & Mahmood have shown that these models have achieved incredible success in language production and sentiment analysis (Singh & Mahmood, 2021). Srihith et al. also supports this application, which is a transition step to deep learning (Srihith et al., 2022).

Other areas where NLP has an impact can be said to be smart city initiatives and chatbots. Patra, in his study, presented a comprehensive review study showing that chatbots are supported by NLP and participate in conversations just like humans (Patra & Kumar, 2020). Tyagi and Bhusha, who explained the potential of NLP to improve the sector in areas such as society, health, education, made a more comprehensive analysis of NLP for smart cities (Tyagi & Bhushan, 2023).

Along with all these developments, the difficulties experienced continue to exist. One of the difficulties experienced is the data scarcity area for low-resource languages, which is a critical area for future research. At the same time, robust evaluation criteria are needed (Avetisyan.,2023). Chen et al., in their study on the public health crisis, emphasized the need to develop information retrieval systems and address the problem of misinformation (Chen et al., 2021).

To summarize all these; NLP is on its way to becoming a dynamic, rapidly developing field, spreading its influence to a wide area. It is expected to open the door to innovative solutions in many sectors, especially with its interaction with machine learning techniques. The BERT model has been redesigning the field of natural language processing since its emergence with its transformers-based architecture. BERT's revolutionary nature stems from its bidirectional converter, which enables a much better understanding of linguistic nuances. This bidirectional converter system, which captures contextual information on both sides of a text string, provides a superior performance advantage. Thanks to this ability, BERT has come to the forefront in almost all NLP systems, especially in sentiment analysis.

Many studies continue to highlight BERT's capabilities and contribute to the literature. Sayeed et al., emphasizing BERT's excellent ability to capture and analyze emotional patterns in texts, presented a comparison of BERT with other models in their study (Sayeed et al., 2023). Wang et al., evaluating BERT in terms of versatility and robustness, conducted a similar comparison through the analysis of negative emotions during the Covid-19 period (Wang et al., 2020). Sun et al., highlighting the model's fine-tuning capability, stated that this fine-tuning is the main feature that sets the BERT model apart from other models (Sun et al., 2019). This fine-tuning method, also emphasized by Sun et al., is performed after the data has been trained. Thus, a two-stage process management is established. Deng et al. also showed that this two-stage process increased the success of the study (Deng et al., 2023). Devlin et al. (Devlin et al., 2019) are among those who argue that these two stages, and especially the BERT model used with fine-tuning, should be applied to all areas of NLP. Sosea & Caragea, who aimed to develop these capabilities of BERT, introduced the Emotion masked language modeling (Sosea & Caragea.,2021). BERT's skills and adaptability have also been evaluated in various areas. For example, Chandra and Saini used it to model emotions during the US (United States) elections to demonstrate its ability in political sentiment analysis and presented the results (Chandra & Saini, 2021). As a different study, Nugroho et al. demonstrated the effectiveness of BERT in mobile application reviews (Nugroho et al.,2021). Myagmar and Li, who found that the model would be successful in application between fields, in addition to its success in every field, also ensured the emergence of cross-domain contextualization (Myagmar et al., 2019). This continuous evolution in the model has been developed by Delobelle et al. under the name RobBERT (Robustly Optimized BERT) for Dutch language processing (Delobelle et al.,2020), and various variants have been developed by Lee et al. under the name BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) for text mining to be used in the biomedical field (Lee et al., 2019).

As a result, BERT has become the most important model in all NLP fields, especially in sentiment analysis, and has become a revolutionary technology. The two-stage process, consisting of pre-training and fine-tuning, and the bidirectional architecture used for context in texts, have resulted in superior performance in many areas. These results have increased researchers' confidence in BERT and transformer-based architectures and paved the way for their continued existence in the future.

Transparency, trust, and interpretability are the most important elements for all artificial intelligence models, such as complex machine learning models, deep learning models, and transformer-based models. However, the "black box" nature of artificial intelligence models makes interpretability and transparency difficult (Binder et al., 2022; Rogers et al., 2020). Therefore, XAI integration is of great importance for ensuring interpretability.

A review of the literature reveals that some researchers have used SHAP and LIME, which are XAI techniques, for the analysis and interpretation of BERT's decision-making process (Dolk et al., 2022). These methods are quite valuable. These inferences and interpretations, which are important for every field, are particularly significant for the health sector (Bauer et al., 2024; Nazir et al., 2023). Rietberg et al. (Rietberg et al., 2023), who used XAI techniques in the biomedical field, and Bauer et al., who analyzed data obtained from social media related to mental health and aimed to create a basis for understanding complex human behavior, have also contributed to the use of this interpretation in the health field (Bauer et al., 2024). At the same time, Balkir et al. have brought a different perspective, emphasizing that the integration of XAI and BERT can play an important role in detecting and reducing biases (Balkir et al., 2022).
Developing XAI techniques better adapted to the features of BERT and using them for the analysis and interpretation of the outputs of this model should be the focus of future research. The intersection of these two technologies is expected to play an important role in eliminating the issue of trust and transparency among users.

3. Materials and method

This section describes and introduces the Natural Language Processing, Explainable Artificial Intelligence techniques, the BERT technique, and the dataset used in the study.

3.1. Natural language processing (NLP)

Natural language processing, in its most general terms, is a branch of artificial intelligence developed to enable computers to understand and produce human language and process the data produced. NLP, which finds itself at the intersection of artificial intelligence, linguistics and computer science, offers methods for analyzing speech or text inputs. NLP, which shows itself in many areas such as sentiment analysis, text summarization, machine translation or question-answer systems, has many tasks such as examining the grammatical structure of the language, entity recognition, meaning extraction, and creating language models. The most well-known methods include deep learning-based approaches, n-grams and the language models used in this study (BERT, GPT, etc.). The most important language model is shown as BERT.

Fig. 1 provides a visual explanation of natural language processing.

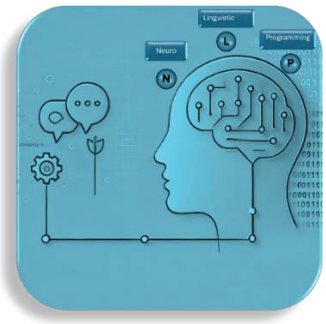


Fig. 1. NLP (Natural language processing)

3.2. Transformer based models

Transformer-based models are deep learning-based and are revolutionary models in the fields of artificial intelligence and natural language processing. Introduced in 2017 with the article "Attention is all you need", this model exhibits effective performance in many tasks in all areas of NLP. Thanks to the "Attention" mechanism, it has a bidirectional technology that can analyze the connection of each analyzed word with both the words before and after it. While traditionally known models such as LSTM (Long short-term memory) and RNN (Recurrent Neural Network) operate sequentially, transformer models can process all words at the same time.

Although BERT is the most well-known transformer-based model, there are many important transformer-based models. Some of these are; GPT, T5, Electra, Transformers-XL and other BERT-based models.

In this study, BERT, the most well-known transformer model, was studied.

3.2.1. Bert (Bidirectional encoder representations from transformers)

BERT, known as the most important transformer model, was developed by Google in 2018. While traditional models can analyze language either from right to left or from left to right, that is, one-way, this model makes a difference with its two-way operation.

The model gains the ability to grasp the entire context more quickly and accurately thanks to its ability to evaluate both the before and after of a text at the same time. BERT, a pre-trained model, has a fine-tuning feature. BERT, which is used in many problems such as text classification, named entity recognition known as NER (Named Entry Recognition), and question-answer systems, is shown among the best artificial intelligence models with the efficiency it achieves. BERT was trained with Google's BooksCorpus (approximately 800 million words) and Wikipedia (approximately 2.5 billion words) data sets. In other words, the training of the BERT technique was carried out with 3.3 billion data. BERT applies the masking method (Masked Language Model-MLM) and next word prediction (Next Sentence Prediction-NSP), which have never been applied before. Masking is a method of guessing by covering (masking) the word to be guessed and analyzing the words to the left and right of it. NSP is a method of guessing which sentence will come next by analyzing the previous sentences. Fig. 2 shows an illustration of BERT's MLM and NSP features.

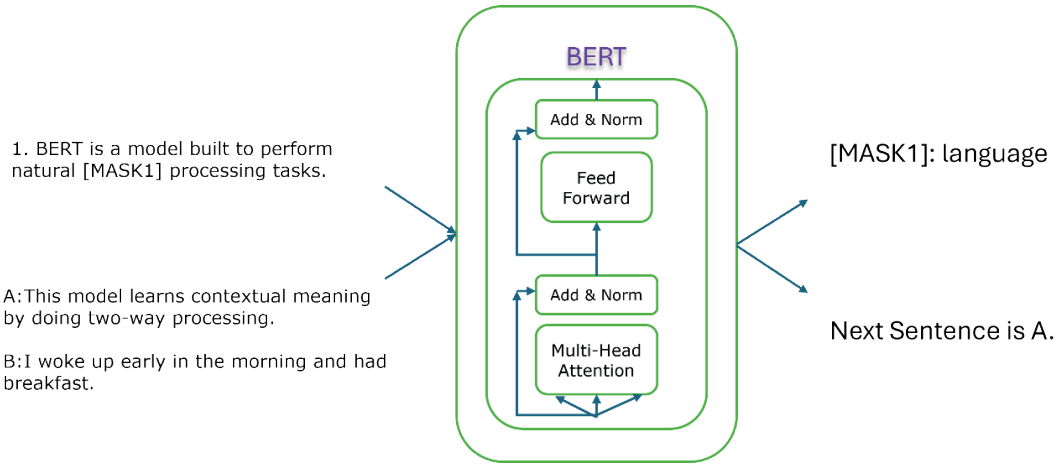


Fig 1. MLM and NSP on BERT

Fig. 3 shows a diagram showing the working mechanism of BERT. In the Tokens field and Input IDs field; in the BERT model, each word is divided into tokens and a token equivalent is assigned to each word. For example, the word 'so' is represented by '2061' as the token equivalent.

A special command called [SEP] is assigned to each sentence to indicate that the sentence is over. The [CLS] command seen at the beginning of the input strings is a special command representing classification. These tokens are given as input and a bidirectional analysis is performed with the BERT technique in the transformer layer. The Attention mechanism, where the relationship of each word is calculated, is located here.

The outputs received from here are words that have gained meaning and context.

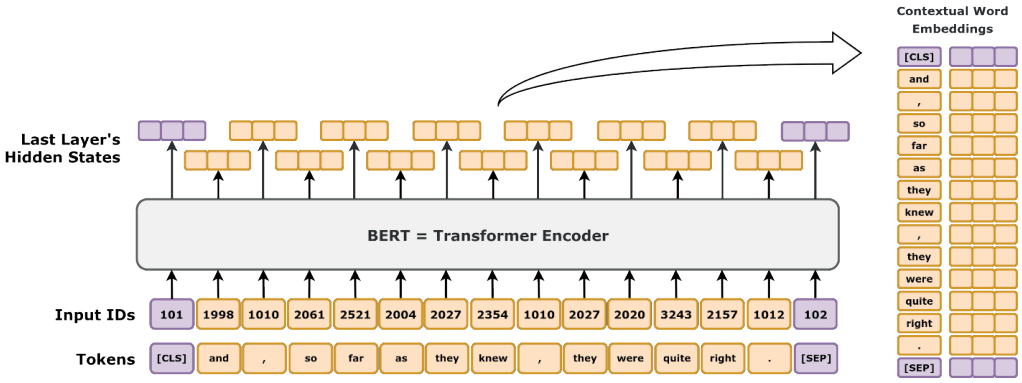


Fig 2. BERT model (Wikipedia.,2024)

There are 2 main variations of BERT. These are called BERT-base and BERT-large.
BERT-base; contains 110 million parameters and consists of 12 layers, namely transformer blocks.
BERT-large; contains 340 million parameters and has a system consisting of 24 layers in total.

3.3. Explainable artificial intelligence (XAI)

New technologies and inventions emerge thanks to solutions produced for needs. Explainable Artificial Intelligence is a technology that emerged from users wanting to understand the

outputs produced by algorithms such as machine learning and artificial intelligence, and to know which parameters affect the results obtained more.

This system provides this explanatory power, while overcoming the black box feature of artificial intelligence has facilitated the transition to the era of transparent and reliable algorithms.

3.3.1. Xai techniques

Many techniques have been developed and continue to be developed for the explainability of an artificial intelligence. The most effective and well-known of these are Lime, Shap, Eli5 and YellowBrick techniques.

Lime (Local interpretable model-agnostic explanations)
Lime was created to understand and explain complex machine learning models. As its name suggests, it works independently of the model. It can be applied to any machine learning model. As is known, the aim of XAI techniques is to translate machine language into a language that humans can understand. For this, Lime aims to explain the prediction of a given model by approximating it to a simpler model. Instead of explaining the working logic of the entire model, Lime reduces it to a specific one and investigates the effects of each input or word on the model. The word local in its name comes from here.

Shap (Shapley additive explanations)
The Shap technique is again used to facilitate the understanding of complex model outputs like Lime. Unlike this technique, Shapley values are used. These Shapley values are based on cooperative game theory and calculate the effect of each input on the model output. Shap, which investigates the effect by adding and subtracting each input, also calculates the negative effects of the inputs thanks to this method. The Shap technique works only on inputs and outputs without looking at the internal structure of the model. It is a technique that distributes the effect results it obtains fairly.

Eli5 (Explain like i'm five)
The Eli5 method is the simplest and simplest method among the others. This model, which has transferred the plain and understandable language it uses to its name, uses the expression 'Explain it as if I were 5 years old' in the form of explaining it so simply that even a 5-year-old child can understand it. The Shap method, which calculates the effect of the inputs on the result in a simple way, is also used to detect and correct model errors.

Yellowbrick
YellowBrick is actually a Python library. Its purpose is to visualize machine learning models and contribute to their understanding. YellowBrick can work with any machine learning like Lime, meaning it is a model- independent technique. With this technique, which can work with known machine learning models like Scikit-learn, the decision and performance of the model can be examined.

3.4. Dataset

The dataset used in the study was obtained from Kaggle. It is located on Kaggle under the title NLP on Research Articles (Vetrivel-PS.,2020). The dataset holds the title and abstract information of more than 20 thousand articles. In today's literature, it is difficult to label which field the articles belong to. This dataset is both suitable for natural language processing and in a field that needs to be analyzed. The articles in the dataset will be collected under 6 main headings. These fields are Computer Science, Physics, Mathematics, Statistics, Quantitative Biology and Quantitative. This dataset was analyzed with the materials and methods used in the study and the results obtained were explained with XAI techniques.

4. Experimental results

This section is evaluated in two subsections. First, the results obtained by applying the BERT model to the dataset are shared. Secondly, the results of the XAI techniques applied to the dataset trained with the model to determine the factors affecting the classification result are given.

4.1. Bert classification results

The transformer-based BERT model created in the study was applied to the dataset for classification. The results obtained for 6 different areas in the dataset are presented, as well as the classification results applied to the entire dataset. Table 1 shows the classification results obtained from different areas.

Table 1. Classification results of different areas

#	Computer Science	Physics	Mathematics	Statistics	Quantitative Biology	Quantitative Finance
acc	0,882	0,949	0,921	0,893	0,974	0,995
f1	0,878	0,939	0,895	0,857	0,647	0,865

As with all dataset results, high success was achieved in the field results evaluated separately. In particular, the articles in the Quantitative Finance field showed a success rate close to 100% due to their content consisting of more specific words compared to other fields. These success results are shown graphically in Fig. 4.

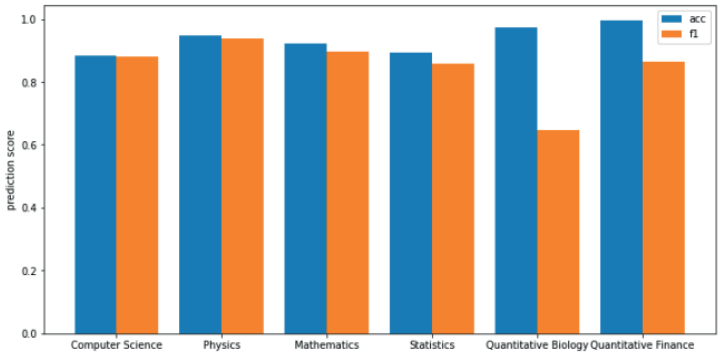


Fig. 3. Classification result of the fields (Graphical representation)

The results obtained by applying the BERT technique to the entire data set, which is the main evaluation in the study, are given in Table 2. The results, which obtained an accuracy value of 93.66%, are very important for future studies.

Table 2. BERT classification results

#	**BERT**
test loss	0,1548
test acc	0,9366

4.2. Application of XAI techniques

There are four known XAI techniques: Shap, Lime, YellowBrick and Eli5. In this study, by applying these methods to the dataset trained during classification, it has been revealed more clearly how the machine learning decision mechanism works and which parameters are more effective to use. The results are important in terms of overcoming the black box feature of artificial intelligence and better understanding by the end user.

4.2.1. Shap results

According to the shap technique, which is essentially derived from game theory, each feature is a player. According to this method, the final reward is a prediction. The aim of the technique is to distribute the total reward fairly among the players. This method stands out by providing an infrastructure for tree-based models and performing very fast operations. Fig. 5 and Fig. 6 show the extent to which the words that determine which class two different articles classified in two data sets affect whether they are included in this class or not. For example, in the article in Fig. 5, the word “logic” has an effect of 0.14, and the word “semantic” has an effect of 0.28 on this article being included in the “Physics” category.

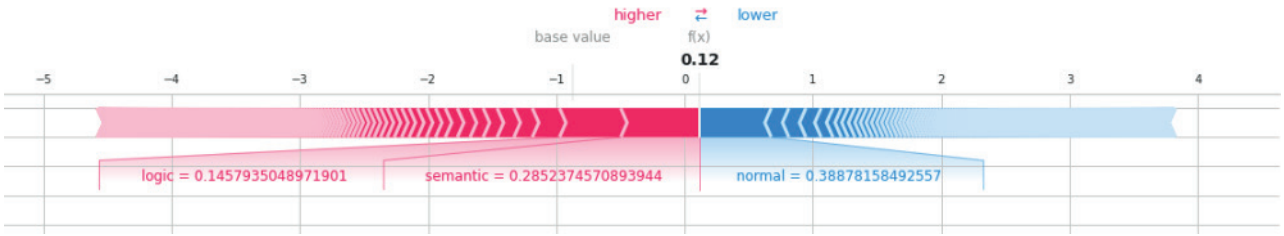


Fig. 4. An article from the physics category

Figure 6 shows an article included in the “Computer Science” field. The word “system” contributed 0.30 and the word “software” contributed 0.29 for the article to be included in this field.

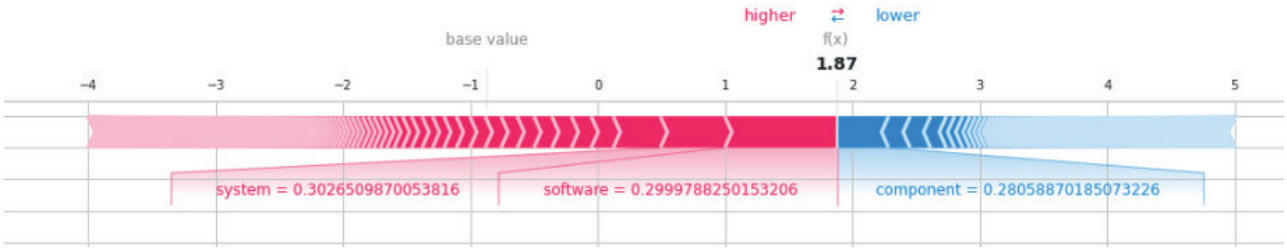


Fig. 5. An article from the computer science

4.2.2. Lime results

Lime contributes to the reason why a prediction is made and provides the infrastructure for optimizing the results obtained from those predictions. For example, Fig. 7 shows which features contribute to which field according to the abstract of a Computer Science article. The word “filters” contributes to this prediction by 0.16, the word “compression” by 0.11, while the word “specific” is on the side of not being included in these fields by 0.06.

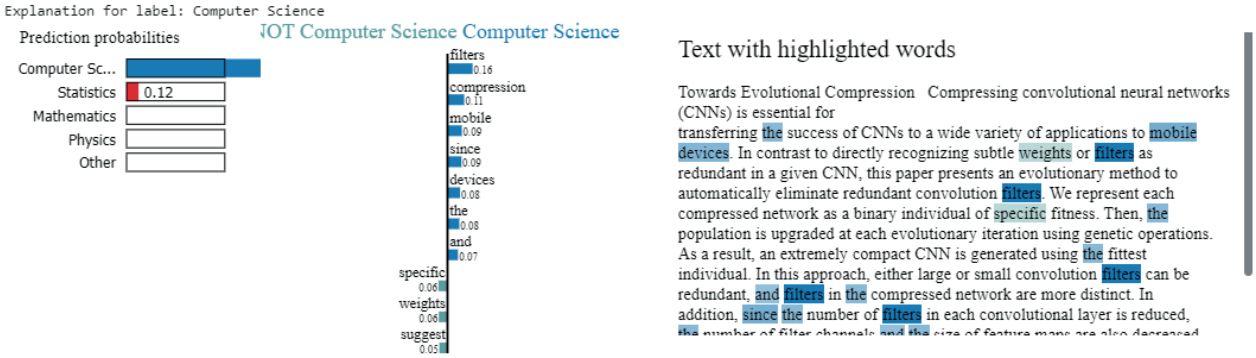


Fig. 6. Lime case study (1)

Looking at the example in Fig. 8, it can be seen that the word “towards” is not included in the Physics field of this study by 0.85. The study is in the Computer Science field and the prediction is correct.

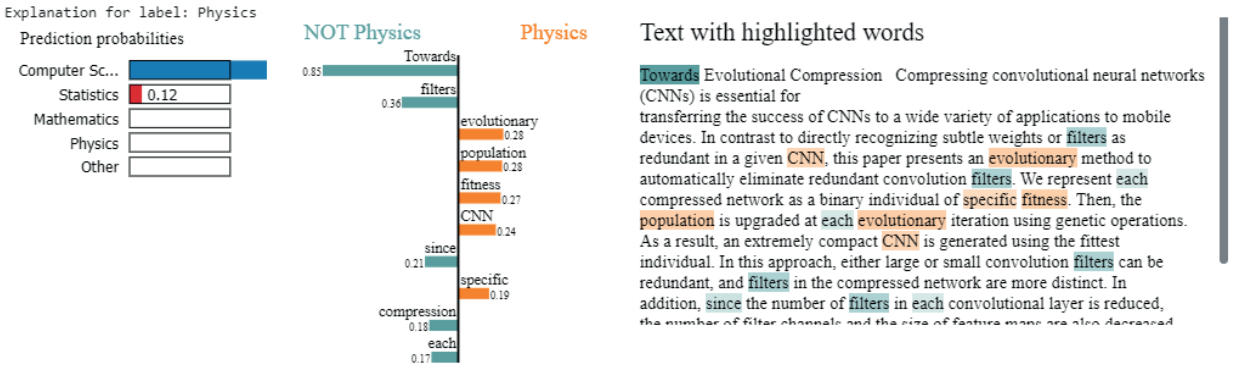


Fig. 7. Lime case study (2)

4.2.3. Eli5 results

Eli5 is the simplest and most understandable method among XAI methods. As stated in its acronym, it was developed to explain artificial intelligence methods in a simple and understandable way, as if explaining it to a 5-year-old child.

Fig. 9 shows the analysis of two different articles with the Eli5 method. As can be seen from the analysis of the words, the first study belongs to the Computer Science field and the second study belongs to the Mathematics field. With the Eli5 method, it is clearly and explicitly shown which words contribute to the inclusion of these studies in these fields.

y=1 top features

Weight?	Feature
+4.793	robot
+4.308	language
+3.813	complexity
+3.570	paper
+3.495	social
+3.295	users
+3.271	user
+3.112	logic
+2.945	codes
+2.932	software
+2.765	speech
+2.728	word
+2.712	sensor
+2.676	channel
+2.546	communication
... 515 more positive ...	
... 466 more negative ...	
-2.520	brain
-2.653	magnetic
-2.677	electron
-2.699	equation
-2.752	here

y=1 top features

Weight?	Feature
+4.791	prove
+4.098	asymptotic
+3.803	mathbb
+3.412	spaces
+2.946	theorem
+2.738	algebras
+2.598	establish
+2.588	groups
+2.556	group
+2.475	manifolds
+2.451	some
+2.350	let
+2.348	operator
+2.320	algebra
... 381 more positive ...	
... 600 more negative ...	
-2.318	optical
-2.474	novel
-2.592	our
-3.171	learning
-3.210	find
-3.226	neural

Fig. 8. Eli5 examples

4.2.4. Yellowbrick results

YellowBrick is the latest XAI technique. Thanks to the matplotlib and scikit-learn libraries it is connected to, it has the ability to be used directly in many machine learning models. In this way, it has a natural ability to visualize data by handling it one by one.

Fig. 10 and Fig. 11 show the Yellowbrick technique images showing which parameters are used to include two different articles in the Mathematics and Computer Science fields.

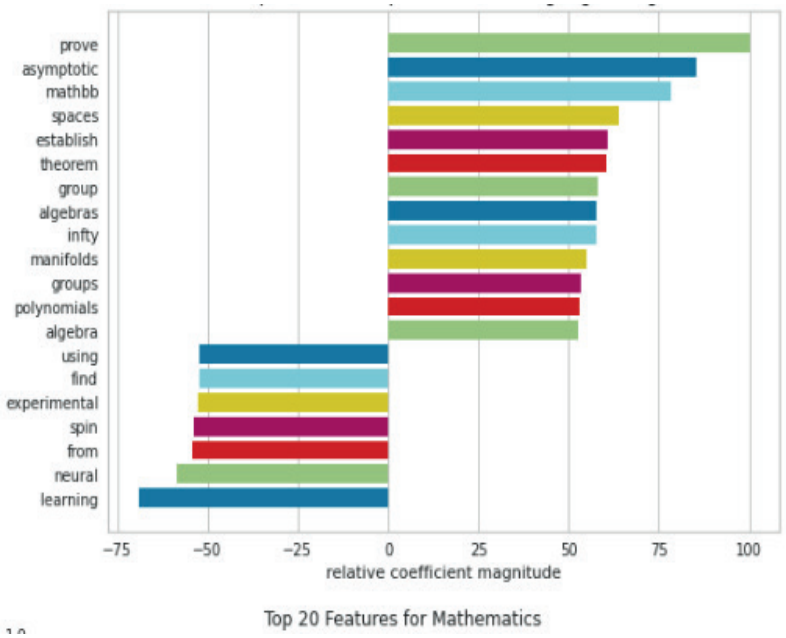


Fig. 9. Yellowbrick example (1)

It has been observed that words such as “prove”, “asymptotic”, “mathbb” are more effective for an article in the field of Mathematics, while words such as “robot”, “language”, “complexity” are more effective for an article in the field of Computer Science.

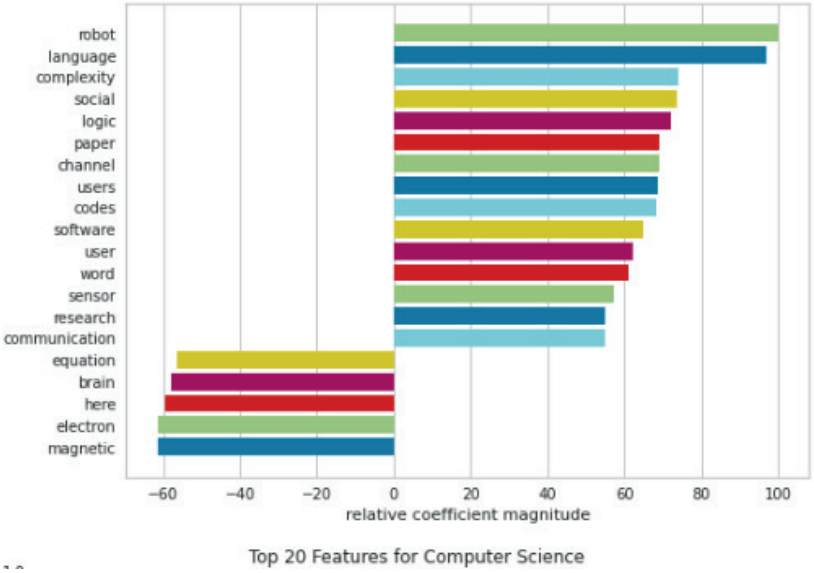


Fig. 10. Yellowbrick example (2)

5. Conclusions and future works

In this study, the effectiveness of the BERT technique, which is the most important natural language processing technique, was tested on the NLP on Research Articles data set. As a result of the obtained data, this model showed high success with an accuracy value of 0.9366 and a loss value of 0.1548. Another issue emphasized in the study is the transparency and interpretability of the outputs obtained from complex models. Explainable Artificial Intelligence (XAI) techniques were used to understand the logic in the internal mechanism of these complex models. The data classified with the BERT technique was interpreted with the XAI techniques Lime, Shap, Eli5 and YellowBrick methods, and it was clearly shown which parameters or features were more effective in making this classification and making decisions. This contributed to the increase in the model's accuracy as well as its transparency and reliability. As a result, by combining two important technologies such as BERT and XAI, which have advanced transformers-based architecture, not only high classification performance was achieved, but also a critical success was shown in terms of making the logic behind this performance understandable and explainable. In the future, the development of more advanced XAI techniques and the discovery of explanatory methods that can be used in decision-making processes, especially with deep learning, are important in terms of contributing to this field. In addition, incorporating XAI methods into the process of improving the results as well as their explanatory nature is an important area of research.

References

Al-Garadi, M., Yang, Y., & Sarker, A. (2022). The role of natural language processing during the covid-19 pandemic: health applications, opportunities, and challenges. Healthcare, 10(11), 2270. https://doi.org/10.3390/healthcare10112270

Auletta, G., et al. "Predicting and understanding human action decisions during skillful joint-action using supervised machine learning and explainable-AI" Scientific Reports (2023) doi:10.1038/s41598-023-31807-1.

Avetisyan, H. (2023). Large language models and low-resource languages: an examination of armenian nlp.. https://doi.org/10.18653/v1/2023.findings-ijcnlp.18

Balkir, E., Kiritchenko, S., Nejadgholi, I., & Fraser, K. (2022). Challenges in applying explainability methods to improve the fairness of nlp models.. https://doi.org/10.18653/v1/2022.trustnlp-1.8

Bauer, B. (2024). Using large language models to understand suicidality in a social media–based taxonomy of mental health disorders: linguistic analysis of reddit posts. Jmir Mental Health, 11, e57234-e57234. https://doi.org/10.2196/57234

Belghachi, A. "A Review on Explainable Artificial Intelligence Methods, Applications, and Challenges" Indonesian Journal of Electrical Engineering and Informatics (2023) doi:10.52549/ijeei.v11i4.5151.

Belle, V., & Papantonis, I. "Principles and Practice of Explainable Machine Learning" Frontiers in Big Data (2021) doi:10.3389/fdata.2021.688969.

Bienefeld, N., et al. "Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals" Npj Digital Medicine (2023) doi:10.1038/s41746-023-00837-4.

Binder, M., Heinrich, B., Hopf, M., & Schiller, A. (2022). Global reconstruction of language models with linguistic rules – explainable ai for online consumer reviews. Electronic Markets, 32(4), 2123-2138. https://doi.org/10.1007/s12525-022-00612-5

Chaddad, A., et al. "Survey of Explainable AI Techniques in Healthcare" Sensors (2023) doi:10.3390/s23020634.

Chandra, R. and Saini, R. (2021). Biden vs trump: modeling us general elections using bert language model. Ieee Access, 9, 128494-128505. https://doi.org/10.1109/access.2021.3111035

Chen, Q., Leaman, R., Allot, A., Luo, L., Wei, C., Yan, S., … & Lu, Z. (2021). Artificial intelligence in action: addressing the covid-19 pandemic with natural language processing. Annual Review of Biomedical Data Science, 4(1), 313-339. https://doi.org/10.1146/annurev-biodatasci-021821-061045

Delobelle, P., Winters, T., & Berendt, B. (2020). Robbert: a dutch roberta-based language model.. https://doi.org/10.18653/v1/2020.findings-emnlp.292

Deng, Y. (2023). Toward transformer fusions for chinese sentiment intensity prediction in valence-arousal dimensions. Ieee Access, 11, 109974-109982. https://doi.org/10.1109/access.2023.3322436

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (2018) doi:10.48550/arxiv.1810.04805.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). Untitled.. https://doi.org/10.18653/v1/n19-1423

Dolk, A., Davidsen, H., Dalianis, H., & Vakili, T. (2022). Evaluation of lime and shap in explaining automatic icd-10 classifications of swedish gastrointestinal discharge summaries.. https://doi.org/10.3384/ecp187028

Gilpin, L. H., et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning" (2018) doi:10.1109/dsaa.2018.00018.
Hu, Y., et al. "XAITK: The explainable AI toolkit" Applied AI Letters (2021) doi:10.1002/ail2.40.
Hu, Y. "Unlocking Causal Relationships in Commercial Banking Risk Management: An Examination of Explainable AI Integration with Multi-Factor Risk Models" Journal of Financial Risk Management (2023) doi:10.4236/jfrm.2023.123014.
Kapcia, K., et al. "ExMed: An AI Tool for Experimenting Explainable AI Techniques on Medical Data Analytics" (2021) doi:10.1109/ictai52525.2021.00134.
Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C., … & Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682
Linardatos, P., et al. "Explainable AI: A Review of Machine Learning Interpretability Methods" Entropy (2020) doi:10.3390/e23010018.
Minh, D. H., et al. "Explainable artificial intelligence: a comprehensive review" Artificial Intelligence Review (2021) doi:10.1007/s10462-021-10088-y.
Mohsenisina, M., et al. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems" ACM Transactions on Interactive Intelligent Systems (2021) doi:10.1145/3387166.
Mote, K. (2012). Natural language processing - a survey.. https://doi.org/10.48550/arxiv.1209.6238
Myagmar, B. and Li, J. (2019). Cross-domain sentiment classification with bidirectional contextualized transformer language models. Ieee Access, 7, 163219-163230. https://doi.org/10.1109/access.2019.2952360
Nazir, S., Dickson, D., & Akram, M. (2023). Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Computers in Biology and Medicine, 156, 106668. https://doi.org/10.1016/j.compbiomed.2023.106668
Nugroho, K., Sukmadewa, A., Dw, H., Bachtiar, F., & Yudistira, N. (2021). Bert fine-tuning for sentiment analysis on indonesian mobile apps reviews., 9, 258-264. https://doi.org/10.1145/3479645.3479679
Patra, B. (2020). Natural language processing in chatbots: a review. Turkish Journal of Computer and Mathematics Education (Turcomat), 11(3), 2890-2894. https://doi.org/10.61841/turcomat.v11i3.14655
Ranchhod, E. and Mamede, N. (2002). Advances in natural language processing.. https://doi.org/10.1007/3-540-45433-0
Raza, A. (2023). Review to unfold the role of machine learning algorithms in natural language processing. JPR, 9(4), 152-162. https://doi.org/10.61506/02.00136
Rietberg, M., Nguyen, V., Geerdink, J., Vijlbrief, O., & Seifert, C. (2023). Accurate and reliable classification of unstructured reports on their diagnostic goal using bert models. Diagnostics, 13(7), 1251. https://doi.org/10.3390/diagnostics13071251
Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: what we know about how bert works. Transactions of the Association for Computational Linguistics, 8, 842-866. https://doi.org/10.1162/tacl_a_00349
Roy, K., Debdas, S., Kundu, S., Chouhan, S., Mohanty, S., & Biswas, B. (2021). Application of natural language processing in healthcare., 393-407. https://doi.org/10.1002/9781119818717.ch21
Saranti, A., et al. "Actionable Explainable AI (AxAI): A Practical Example with Aggregation Functions for Adaptive Classification and Textual Explanations for Interpretable Machine Learning" Machine Learning and Knowledge Extraction (2022) doi:10.3390/make4040047.

Savova, G., Danciu, I., Alamudun, F., Miller, T., Lin, C., Bitterman, D., … & Warner, J. (2019). Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Research, 79(21), 5463-5470. https://doi.org/10.1158/0008-5472.can-19-0579

Sayeed, M. (2023). Bert: a review of applications in sentiment analysis. Hightech and Innovation Journal, 4(2), 453-462. https://doi.org/10.28991/hij-2023-04-02-015

Singh, S. and Mahmood, A. (2021). The nlp cookbook: modern recipes for transformer based deep learning architectures. Ieee Access, 9, 68675-68702. https://doi.org/10.1109/access.2021.3077350

Sosea, T. and Caragea, C. (2021). Emlm: a new pre-training objective for emotion related tasks.. https://doi.org/10.18653/v1/2021.acl-short.38

Srihith, N., Varaprasad, R., Mohan, Y., Srinivas, T., & Sravanthi, Y. (2022). A comprehensive analysis of deep learning's impact on natural language processing. International Journal of Latest Engineering and Management Research (Ijlemr), 7(10), 01-15. https://doi.org/10.56581/ijlera.7.10.01-15

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification?., 194-206. https://doi.org/10.1007/978-3-030-32381-3_16

Tiwari, A. "Explainable AI (XAI) and its Applications in Building Trust and Understanding in AI Decision Making" International Journal of Scientific Research in Engineering and Management (2023) doi:10.55041/ijsrem17592.

Tyagi, N. and Bhushan, B. (2023). Demystifying the role of natural language processing (nlp) in smart city applications: background, motivation, recent advances, and future research directions. Wireless Personal Communications, 130(2), 857-908. https://doi.org/10.1007/s11277-023-10312-8

Velupillai, S., Mowery, D., South, B., Kvist, M., & Dalianis, H. (2015). Recent advances in clinical natural language processing in support of semantic analysis. Yearbook of Medical Informatics, 24(01), 183-193. https://doi.org/10.15265/iy-2015-009

Vetrivel-PS. (2020). NLP on Research Articles. Retrieved 20 October 2024 from [https://www.kaggle.com/datasets/vetrirah/janatahack-independence-day-2020-ml-hackathon/data].

Wang, T., Ke, L., Chow, K., & Zhu, Q. (2020). Covid-19 sensing: negative sentiment analysis on social media in china via bert model. Ieee Access, 8, 138162-138169. https://doi.org/10.1109/access.2020.3012595

Webber, B. (1986). Natural language processing: a survey., 353-363. https://doi.org/10.1007/978-1-4612-4980-1_29

Wikipedia contributors. (2024, November 18). BERT (language model). In Wikipedia, The Free Encyclopedia. Retrieved 11:31, November 20, 2024, from https://en.wikipedia.org/w/index.php?title=BERT_(language_model)&oldid=1258255803

CHAPTER 4

ARTIFICIAL INTELLIGENCE AND IMAGE ANALYSIS-BASED APPROACHES IN COLORECTAL CANCER: A LİTERATURE REVIEW FROM DIAGNOSIS TO PREDICTION

Aynur SEVİNÇ[1]

1 Dr. Öğr. Üyesi, Department of Computer Technologies, Silvan Vocational School, Dicle University, Diyarbakir 21640, Turkey; aynur.sevinc@dicle.edu.tr, ORCID: 0000-0002-1388-2554

1. INTRODUCTION

Colorectal cancer (CRC) remains a life-threatening malignancy, ranking among the three most common cancers worldwide. Despite advances in screening and treatment, mortality rates remain high, largely due to late diagnosis and heterogeneity in tumor progression. Early and sensitive detection of precancerous lesions plays a crucial role in improving patient survival and providing effective therapeutic interventions.

Despite significant advances in screening and treatment, CRC mortality continues to be affected by delayed diagnosis and the complex interaction of genetic, environmental, and lifestyle factors. Recent epidemiological data indicate an increasing incidence of CRC, particularly in younger populations, highlighting the urgent need for early diagnosis and improved prevention strategies. While traditional diagnostic approaches such as colonoscopy, histopathology, and imaging have made significant contributions to disease management, their limitations in sensitivity, reproducibility, and interpretative subjectivity necessitate the integration of more advanced analytical tools.

Artificial intelligence (AI) has rapidly emerged as a groundbreaking force in medical research and has become a promising technology for improving diagnostic and predictive accuracy in colorectal cancer. Image analysis, powered by machine learning (ML) and deep learning (DL), is revolutionizing the interpretation of medical images, enabling automatic detection, segmentation, and classification. This allows AI systems to be integrated into colorectal imaging modalities such as colonoscopy, computed tomography (CT), magnetic resonance (MRI), and histopathological imaging to automatically detect and classify lesions, assess tumor progression, and predict treatment outcomes with high accuracy. This increasingly powerful interaction between oncology and computational intelligence represents a paradigm shift in colorectal cancer research, transforming traditional diagnostic approaches into data-driven, predictive, and precision medicine.

Recent studies published in the literature indicate that artificial intelligence (AI) is playing an increasingly important role in the detection and treatment of colorectal cancer. AI-powered colonoscopy systems help doctors reduce diagnostic errors by detecting polyps and abnormal tissue in real time. Deep learning models developed in digital pathology can distinguish cancerous from healthy tissue with high accuracy and identify certain genetic alterations associated with the disease. AI-based imaging and predictive models are also being used to predict patient response to treatment, assess the likelihood of disease recurrence, and support personalized treatment plans. However, despite these advances, several challenges remain to be addressed, such as data diversity, model transparency, and clinical validation. Therefore, close collaboration between doctors, data scientists, and researchers is crucial for the integration of AI technologies into daily medical practice.

This literature review aims to provide a comprehensive overview of AI- and image analysis-based approaches to the diagnosis and prediction of colorectal cancer. It examines the evolution of AI-enabled diagnostic systems, classification algorithms, and predictive models, highlighting their contributions to precision oncology. Furthermore, this review discusses the current challenges, ethical considerations, and emerging trends related to the integration of AI into clinical practice. By synthesizing findings from diverse fields, the study demonstrates the transformative power of AI in advancing personalized cancer care by leveraging the relationship between medical imaging and predictive analytics.

1.1.Colon Cancer

Colon cancer is a type of tumor that arises in the colon or rectum, situated in the lower portion of the digestive tract (Allison, 2010). The colon constitutes the majority of the large intestine, while the rectum is positioned at its terminal section. Its high prevalence is mainly attributed to unhealthy lifestyle choices, including persistent smoking, high red meat consumption, and insufficient fruit intake, along with factors such as family history of the disease and increasing age.

Colon cancer is divided into four main stages. Stage one is when the tumor is located in the mucosa, or inner surface layer, of the colon or rectum and has not yet spread to the organ wall. Stage two is when the tumor begins to invade the colon or rectum wall, but surrounding tissues or lymph nodes are not yet affected (Baxter et al., 2009). Stage three is when the tumor has spread to the lymph nodes but has not metastasized to other parts of the body. Stage four is when the tumor has metastasized to distant organs such as the lungs (Bera et al., 2019). Figure 1 shows the four main grades of colon cancer (Rathore et al., 2013).

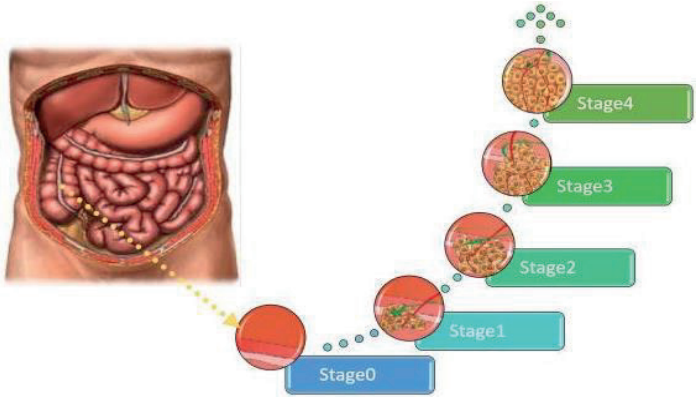


Figure 1. The different stages of colon cancer

2. FUNDAMENTALS OF ARTIFICIAL INTELLIGENCE IN ONCOLOGY

2.1.Overview of AI, Machine Learning (ML) and Deep Learning (DL)

AI has played a transformative role in healthcare processes such as data-driven decision making, predictive analysis, and increased diagnostic accuracy. Today, AI systems contribute to the development of clinical decision support systems by processing large and complex datasets from diverse sources, such as medical images, genomic data, and electronic health records (Russell & Norvig, 2021). Within the healthcare ecosystem, AI subfields such as machine learning (ML), natural language processing, and computer vision offer significant opportunities to improve patient outcomes and reduce diagnostic errors. AI technologies integrated into clinical applications have demonstrated remarkable results in early disease detection, risk stratification, and increased operational efficiency in healthcare (Jordan & Mitchell, 2015).

Machine learning (ML) and its evolved branch, deep learning (DL), have accelerated the adoption of AI in medical applications. ML algorithms identify hidden patterns in patient data, enabling the prediction of chronic conditions such as cancer, cardiovascular disease, and diabetes (Goodfellow, Bengio & Courville, 2016). Furthermore, ML models, particularly convolutional neural networks, have outperformed traditional methods in medical imaging for tasks such as tumor detection and retinal disease classification. Overall, AI, ML, and DL are supporting the development of personalized, precise, and data-driven

healthcare by bridging the gap between medical imaging and predictive analytics (Esteva et al., 2019).

2.2.Commonly Used Algorithms in Cancer Research (CNN, SVM, Random Forest, etc.)

Machine learning algorithms such as SVM, Naive Bayes, Decision Trees, Random Forests, K-means, KNN, Logistic Regression, eXtreme Gradient Boosting (XGBoost), and Hybrid models are used in cancer research. These algorithms, which offer distinct advantages in terms of accuracy, adaptability, and scalability, analyze large datasets to develop personalized treatments and improve patient outcomes by enhancing early diagnosis. The table below describes the characteristics, strengths, and limitations of these algorithms.

Table 1. Comparative review of commonly used algorithms in cancer research

Algorithm	Core Principle	Applications in Cancer Research	Advantages	Limitations
Convolutional Neural Network (CNN)	A deep learning architecture designed to capture spatial relationships and recognize visual features within image data	Tumor detection, histopathological image analysis, MRI and CT scan classification.	High accuracy, automatic feature extraction, excellent performance with complex medical images.	Requires large datasets, high computational cost, limited interpretability.
Support Vector Machine (SVM)	A supervised learning technique that categorizes data by determining the optimal hyperplane for separation.	Cancer subtype classification, gene expression profiling, biomarker discovery.	Effective with small datasets, robust against overfitting, strong generalization capability.	Inefficient with large datasets, sensitive to kernel selection.
Random Forest (RF)	A machine learning approach that merges several decision trees to enhance the accuracy of predictions.	Disease prognosis, survival prediction, risk factor analysis.	Stable with noisy data, identifies variable importance, high overall accuracy.	Potential overfitting, limited model interpretability.
K-Nearest Neighbors (KNN)	An instance-based machine learning method that assigns a class to a sample by measuring its similarity (distance) to nearby data points.	Cancer subtype classification, patient clustering.	Simple to implement, no training phase required, effective with small datasets.	Slow with large datasets, sensitive to noise and irrelevant features.
Logistic Regression (LR)	A statistical model that estimates the probability of a binary outcome based on input variables.	Cancer risk prediction, prognostic factor evaluation.	Easy to interpret, low computational cost, useful as a baseline model.	Limited ability to capture nonlinear relationships, depends on linear assumptions.

The algorithms in Table 1 demonstrate the distinct advantages and limitations of both classical and deep learning-based approaches in cancer research. Deep learning models such as CNN offer significant advantages in processing complex image data by offering high accuracy and automatic feature extraction in medical image analysis. However, the large dataset and high computational capacity required by these algorithms are a limiting factor in clinical applications. On the other hand, supervised learning methods such as SVM and KNN, while demonstrating strong performance on small datasets, experience limitations in efficiency and speed as data size increases. The Random Forest algorithm, on the other hand, provides stable results in noisy data and offers the advantage of determining variable importance, but it suffers from overfitting and limited interpretability.

Logistic regression, as a classic statistical model, offers low computational cost and high interpretability in cancer risk estimation and prognostic factor assessment; however, because it relies on linear assumptions, it falls short in capturing complex, nonlinear relationships. Overall, the table highlights that algorithms used in cancer research should be selected based on data type, application purpose, and computational requirements. In clinical and research contexts, considering the advantages and limitations of algorithms, the use of hybrid approaches or multistage analysis strategies can yield more reliable and accurate results.

2.3. Radiomics and Image Analysis in CRC

Radiomics and image analysis play a crucial role in extracting quantitative features from colorectal cancer (CC) images. These techniques enable the identification of patterns and biomarkers that are not visible to the human eye, contributing to early diagnosis, treatment planning, and prognosis prediction. However, challenges remain, such as data standardization, image labeling, and integration of multi-source image data. Recent research demonstrates that combining radiomics with clinical and molecular data significantly improves diagnostic accuracy and predictive performance. Continued advances in artificial intelligence algorithms and imaging technologies will enable these approaches to become more reliable and clinically applicable in the near future.

3. METHODOLOGY

This review was conducted following the principles of systematic and narrative literature synthesis. A comprehensive search was performed using databases including **ScienceDirect, PubMed, IEEE Xplore, Scopus, and SpringerLink**, focusing on studies that utilized artificial intelligence (AI), machine learning (ML), or deep learning (DL) methods in **colorectal cancer diagnosis, classification, or staging**. Keywords such as *"colorectal cancer," "artificial intelligence," "deep learning," "radiomics," "classification," "staging,"* and *"predictive modeling"* were used in various combinations.

To ensure coverage of all stages of technological development, no year limitation was applied; both original research articles and review articles were included. Studies indexed in peer-reviewed journals were included. Articles that used non-AI-based methods or focused on gastrointestinal diseases other than colorectal cancer were excluded.

No year limitation was applied in this review to ensure a comprehensive understanding of the evolution and diversity of artificial intelligence (AI) applications in colorectal cancer (CRC). The development of AI-based diagnostic and staging models has progressed rapidly over the past decade, but foundational studies published earlier continue to provide essential methodological insights and baseline comparisons. Including both earlier and recent publications allows a broader evaluation of trends, algorithmic improvements, and validation strategies over time. This approach also helps identify persistent challenges and highlight

how AI techniques have evolved to overcome them, thereby offering a complete picture of the field's trajectory.

Each selected publication was reviewed for its methodological design, dataset characteristics, algorithm type, evaluation metrics, and validation strategies. The findings were then synthesized and grouped under four main categories: (1) Deep learning for histopathological image classification, (2) Tumor grading and stage prediction, (3) Multi-modal approaches combining imaging and clinical data, and (4) Validation and generalization of classification models. This structure allowed for a critical comparison of methods and the identification of common challenges and future directions.

4. AI IN DIAGNOSTIC IMAGING OF COLORECTAL CANCER

4.1.Colonoscopy Image Analysis Using AI

This review focuses primarily on the use of artificial intelligence applications in identifying and characterizing colorectal polyps, with the aim of improving the effectiveness of colorectal cancer screening and prevention.

In recent years, research on AI-assisted colonoscopy has rapidly increased, and many commercial systems have been developed. However, comparing the effectiveness and accuracy of these systems has been difficult, and a standardized evaluation method has not been established. While deep learning advocates claim that these systems, trained on large datasets, offer consistent accuracy, this review aims to evaluate the performance of existing commercial systems and the validity of these claims. Table 2 below provides information about artificial intelligence systems used in colonoscopy.

Tablo 2. Commercially available AI-assisted colonoscopy systems

Name	Company	Technique	Approval / Year
GI Genius (ColonPRO)	Cosmo / Medtronic	Enhanced CADe	2024 (Software Update) (Medtronic, 2024)
EndoScreener	Wision AI (Shanghai, China)	CADe	2021 (Wision AI, 2021)
CAD EYE	Fujifilm (Tokyo, Japan)	CADe and CADx	2020 (Fujifilm, 2020)
ENDO-AID	Olympus Corporation (Tokyo, Japan)	CADe	2020 (Olympus Corporation, 2020)
Smart Vision	NEC Corporation (Tokyo, Japan)	CADe	2020 (NEC Corporation, 2020)
GI Genius (FDA De Novo)	Cosmo / Medtronic	CADe	2021 (FDA De Novo Clearance) (Medtronic, 2021)
DISCOVERY	Pentax Medical (Tokyo, Japan)	CADe	2020 (Pentax Medical, 2020)
EndoBRAIN	Cybernet Systems Corporation (Tokyo, Japan)	CADx	2018 (Cybernet Systems, 2018)
EndoBRAIN-EYE	Cybernet Systems Corporation (Tokyo, Japan)	CADe	2020 (Cybernet Systems, 2020)
CADDIE (Odin / Olympus)	Olympus / Odin Vision	Cloud-based CADe	2024 (FDA 510(k) Clearance) (Olympus Global, 2024)
GI Genius	Medtronic (Dublin, Ireland)	CADe	2019 (Medtronic, 2019)

In recent years, the use of artificial intelligence systems developed for upper and lower gastrointestinal endoscopy has rapidly increased. Computer-aided diagnosis (CADe) systems identify and detect polyps in still images or videos. Today, these systems are integrated into real-time colonoscopies, alerting the operator with colored boxes around the

polyp's location, facilitating intervention. Additionally, computer-aided diagnosis (CADx) systems can distinguish polyp types and degrees of dysplasia, providing operators with instant diagnostic information on a variety of conditions, from benign hyperplastic polyps to advanced cancers. Thus, both the detection and diagnosis processes are becoming safer and more effective thanks to AI (Young, Edwards & Singh, 2023).

4.2.Detection of Polyps and Adenomas with CNN

Given the high risk of colorectal cancer, real-time automated polyp detection systems help clinicians instantly detect polyps, reducing missed diagnoses. Artificial intelligence-assisted colonoscopy has become a growing area of interest with technological advancements. These systems, thanks to CADe and CADx technologies, play a significant role in the detection and evaluation of precancerous polyps. In current applications, deep learning, and particularly CNNs, contribute to improving adenoma detection rates (ADR) by accurately identifying and localizing premalignant lesions. CNN refers to a specialized type of artificial neural network and deep learning approach that has proven highly effective for analyzing medical images (Shin et al., 2016) (Figure 2).

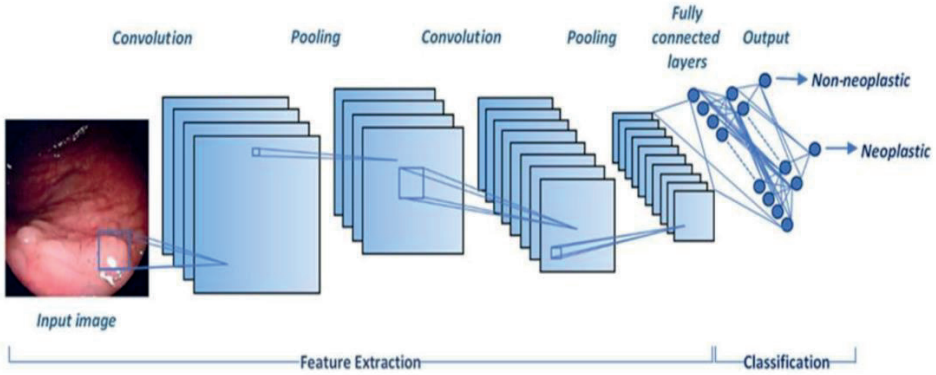


Figure 2. A convolutional neural network (CNN) design for colorectal polyp classification.

In the most recent prospective randomized controlled study, Wang et al. (2019) assessed the effectiveness of a deep learning-based CADe system in detecting polyps and adenomas. A total of 1058 patients were randomly assigned to undergo either standard colonoscopy or colonoscopy assisted by the CADe system. Polyps detected in the CADe system were highlighted on the screen as empty blue boxes. The results showed that the adenoma detection rate (29.1% vs. 20.3%) and the number of adenomas per patient (0.53 vs. 0.31) were increased in the CADe group. This increase was largely due to more effective detection of small polyps, with no significant difference in large adenomas, and a significant increase in the number of hyperplastic polyps in the CADe group. The study demonstrates that AI-assisted colonoscopy significantly improves the detection of small polyps that may be overlooked even by experienced endoscopists, which may reduce the risk of interval colorectal cancer. Mori et al. (2018) demonstrated that CADx support can help endoscopists distinguish between neoplastic and non-neoplastic polyps during colonoscopy, thus enabling the implementation of a “diagnose and leave” approach for non-neoplastic polyps (Mitsala et al., 2021).

4.3.Automated Segmentation and Feature Extraction Techniques

Segmentation in images is a crucial process in the analysis of histopathological images, as it significantly contributes to addressing various diagnostic and analytical challenges. The tasks involved vary across different stages, and even each image presents unique characteristics. Image segmentation can be compared to clustering, as it aims to define

meaningful regions or segments that may vary depending on the model applied or even among individual cells (Kekelidze et al., 2013).

The first approach to polyp segmentation is to utilize image processing techniques. Karargyris and Bourbakis (2009) extracted features from images using Log-Gabor filters to perform automatic polyp segmentation. Jia (2015) used the K-means clustering algorithm to identify polyp contours and segment them. Hwang et al. (2007) used the region-maximum method to determine the starting point in the watershed algorithm. They then applied the elliptic fitting technique to eliminate the redundant regions obtained in the previous step.

The second approach used for polyp segmentation is based on extracting features from image patches and labeling them as polyps or non-polyps based on these features. Tajbakhsh et al. (2015) developed a method that uses the Canny edge detector in each color channel. This method generates edge maps. Oriented patches are then extracted for each pixel, and these patches are classified as polyps or non-polyps.

The third method employed for polyp segmentation involves the use of Convolutional Neural Networks (CNNs). CNNs are a well-established deep learning framework that captures intricate features from raw images using trainable filters and pooling layers (Nasr-Esfahani et al., 2016). In this setup, the extracted features are passed to a classifier that carries out the classification task. Park et al. (2015) implemented a CNN as a feature extractor utilizing a three-scale patch representation for polyp region segmentation. Their network computes 60 features per input patch and classifies them through a fully connected layer consisting of 256 neurons. Additionally, a Gaussian filter is applied post-CNN processing to smooth the segmentation outputs and minimize noise. Ribeiro et al. (2016), on the other hand, employed three convolutional layers and two pooling layers to derive features from RGB patches, followed by a fully connected layer to classify the resulting 1024 features.

The fourth strategy for polyp segmentation involves the use of Fully Convolutional Networks (FCNs) (Long et al., 2015). In recent years, FCNs have become one of the most effective deep learning techniques for enhancing polyp segmentation due to their high computational efficiency in demanding prediction tasks. FCNs represent an advancement over traditional CNNs by replacing fully connected layers with deconvolution layers and leveraging information from earlier layers to boost segmentation accuracy. In polyp segmentation, Akbari et al. (2018) applied an FCN model to identify potential polyp candidates and then used the Otsu thresholding method to segment the polyp regions, substantially improving accuracy. Similarly et al. (2019) evaluated their FCN-based polyp segmentation approach against six other architectures: AlexNet, GoogLeNet, VGG, and three ResNet variants with 50, 101, and 152 layers.

5. PREDICTIVE MODELING AND PROGNOSTIC APPLICATIONS

5.1.Survival Prediction and Recurrence Risk Estimation Using AI Models

AI demonstrates significant potential in predicting survival outcomes and risk of recurrence in colorectal cancer (CC) by leveraging complex imaging, histopathological, and clinical data. CNNs and attention-based architectures, in particular, can extract predictive features related to tumor aggressiveness and patient prognosis from whole-slide images. Recent studies integrating radiomics, gene expression profiles, and clinicopathological variables have yielded higher accuracy in predicting disease-free survival and recurrence probability than traditional statistical methods. These AI-based prognostic models not only enhance individualized risk stratification but also offer clinical decision support for identifying patients who may benefit from adjuvant therapy or close follow-up. As

multicenter validation studies increase, such predictive systems are expected to become a core component of precision oncology in colorectal cancer management.

Özdemir et al. (2025) aimed to integrate immune scoring with a radiology-assisted AI model to develop a prognostic prediction system for patients with resectable colon cancer. 122 patients who underwent surgery between 2011 and 2020 were analyzed. The immune score was calculated based on CD3 and CD8 T cell densities in the intratumoral and invasive margin regions, and preoperative CT images were evaluated with a deep learning-based algorithm, along with radiologic and clinicopathologic data. Among the models, the best-performing AI model (Model 222) achieved 76% accuracy, 80% specificity, and an AUC ROC value of 0.65 in predicting disease-free survival (DFS). The results demonstrate that the proposed AI-based system can effectively classify the risk of recurrence and support clinical decision-making by predicting patient prognosis after radical resection. Being the first study in the literature to combine immunological and radiological features within a deep learning model, it offers a promising approach for personalized treatment planning in colorectal oncology.

5.2.Real-world Examples of Predictive Systems in Clinical Trials

In clinical trials, predictive systems involve artificial intelligence and statistical models developed to predict patient health status, predict treatment response, or assess the risk of complications. For example, in cardiology, some systems can predict a patient's risk of heart attack based on their age, blood pressure, cholesterol level, and genetic profile. In oncology trials, tumor biomarkers and histopathological data can be used to predict a patient's response to chemotherapy or risk of recurrence. In the real world, these predictive systems are integrated with electronic health records and used as clinical decision support systems, guiding physicians to develop more personalized and timely intervention strategies for patients. Thus, predictive systems enhance both patient safety and improve the efficiency and cost-effectiveness of healthcare.

The literature offers numerous examples of real-world applications of predictive systems in clinical trials. For example, in the field of colorectal cancer (CRC), machine learning models based on patient clinical follow-up data have been developed and predicted poor prognostic risks. These models allow for more accurate assessments of patient response to treatment and potential complication risks. Furthermore, research using real-world data plays a critical role in analyzing treatment outcomes for patient groups that are often underrepresented in clinical trials. This data provides valuable information to guide treatment choices and optimize strategies. These examples highlight the importance and potential benefits of integrating predictive systems with real-world data in clinical trials.

6. RESEARCH FINDINGS

The reviewed studies demonstrate that AI and deep learning techniques have achieved significant progress in the classification and staging of colorectal cancer. CNNs have been widely applied to histopathological and radiological images, showing high accuracy in distinguishing malignant from benign tissues and in predicting tumor stage. Several studies have integrated clinical, genomic, and imaging data to build multi-modal models that enhance diagnostic precision and support personalized treatment planning. Moreover, validation studies using independent datasets indicate that robust and generalizable models can effectively assist clinicians in early diagnosis and decision-making. Overall, the findings

suggest that AI-based classification systems are transforming traditional diagnostic workflows and paving the way toward data-driven precision oncology in CRC management.

6.1.AI-Based Classification and Staging Systems

6.1.1. Deep Learning for Histopathological Image Classification

Deep learning has emerged as a transformative approach for histopathological image classification in colorectal cancer, offering unprecedented accuracy in recognizing complex tissue patterns that are often challenging for human observers to interpret consistently. By leveraging convolutional neural networks and attention-based architectures, these systems can automatically learn discriminative features such as glandular structure, nuclear atypia, and stromal organization without the need for handcrafted descriptors. This capability not only accelerates diagnostic workflows but also helps reduce inter-observer variability, enabling more standardized assessments across laboratories. Additionally, recent advances in model interpretability have made it possible to highlight regions that influence predictions, increasing clinician trust in AI-assisted diagnosis. As datasets continue to expand and annotation strategies improve, deep learning is poised to play an increasingly central role in the accurate and efficient classification of histopathological slides.

6.1.2. Tumor grading, stage prediction, and morphological analysis

AI is increasingly offering sophisticated models for tumor grading and stage estimation in colon cancer using histopathological and imaging data. For example, Leo et al. (2024) proposed a system that used gland segmentation before using transformers, and then performed adenocarcinoma grading using a CNN ensemble. This two-stage approach reduced learning time and increased classification accuracy compared to traditional patch-based classification methods. Furthermore, Bahrambanan et al. (2025) compared CNN and ML mixture models using bioinformatics analysis of colon cancer data, achieving accuracy rates approaching 90% with specific feature selection methods (Bahrambanan et al., 2025). These studies demonstrate that AI models can provide staging and prognostic inferences not only from tissue images but also by quantitatively assessing morphological features.

One of the strengths of AI-based staging models is the reduction of inter-observer discrepancies, which makes the classification process objective and reproducible. However, challenges are also evident: variables such as staining techniques across laboratories, differences in microscope/scan equipment, and histological section quality can degrade model performance across centers. Therefore, models must be tested with external data validation from multiple centers.

6.1.3. Multi-Modal Approaches Combining Imaging and Clinical Data

In colon cancer, AI models are no longer relying solely on single-modal image analysis but are now combining clinical, molecular, and radiologic information to provide more powerful prediction systems. For example, Xie et al. (2025) developed a multimodal model that integrates histopathological, clinical parameters, and radiologic data. This system demonstrated exceptional performance in guiding adjuvant chemotherapy decisions in patients with Stage II colon cancer. Similarly, Lin et al. (2024) used a multimodal artificial intelligence approach to combine pathology images with clinical variables to perform prognostic analyses and improve staging accuracy. Such models go beyond single-modal approaches and provide the opportunity to assess tumor biology from a broader perspective.

The power of multi-model systems lies in their ability to integrate heterogeneous data. However, this presents complex challenges such as data preprocessing, missing data issues, and harmonization of units of measurement. Lack of clinical records or incompatible data

collection standards across centers can hinder the model's real-world application. Therefore, data harmonization and missing value management are critical factors for the success of these models.

6.1.4. Validation and Generalization of Classification Models

The acceptance of AI-based classification models for clinical use depends on consistent performance not only across training data but also across centers, populations, and devices. Rosen et al. (2025) developed an AI-powered decision support model using data from the Danish national registry and single-center datasets for 18.403 patients. The model achieved an AUC=0.79 on the validation set, indicating both clinical validity and generalization potential. The integration of this model into clinical practice provides valuable insights into how AI technologies can function in the hospital setting. Furthermore, Mazaki et al. (2024) integrated a combination of CNN and SVM into a prognostic model and evaluated the model's consistency by testing it with an external validation set.

However, there are obstacles to overcome for model generalization: imaging protocols that are dissimilar to the training data, different patient demographics, and class imbalances can increase the risk of overfitting. The use of explainable artificial intelligence (XAI) methods is an important step toward gaining clinician trust by making model internal logic more transparent. Without standardized performance reporting protocols and community-based external validation studies, implementing modern AI models in clinical practice is risky.

6.2. Predictive Modeling and Prognostic Applications

6.2.1. Survival Prediction and Recurrence Risk Estimation Using AI Models

AI-based prognostic models exhibit stronger predictive performance compared to classical statistical methods in predicting survival and the probability of recurrence in colon cancer. For example, Hsu et al. (2023) achieved 85% accuracy in predicting 5-year survival in a cohort of 1.200 patients by integrating clinical and genetic data with a deep learning model. Yamada et al. (2022) developed a model to predict the risk of recurrence in stage II and III colon cancer patients using radiomics features obtained from preoperative CT images and reported an AUC of 0.81. Furthermore, Kather et al. (2020) demonstrated the potential of AI in molecular-level prognostication with a deep neural network model that automatically predicts microsatellite instability (MSI) from histopathological images. These models not only forecast survival outcomes but also facilitate the development of individualized treatment strategies and the early detection of patients at high risk.

6.2.2. Predictive Models for Treatment Response (Chemotherapy, Immunotherapy)

AI-based models are becoming increasingly effective in predicting response to chemotherapy and immunotherapy in colon cancer. For example, Ahn et al. (2023) developed a model that predicted response to FOLFOX treatment in 450 metastatic colon cancer patients using deep learning-assisted radiomics analysis and reported an AUC of 0.83. Sun et al. (2022) built a multi-modal AI system that integrated gene expression profiles, radiologic data, and clinical parameters, achieving 82% accuracy in predicting response to immunotherapy. Furthermore, Xu et al. (2021) identified patient groups that would benefit from immunotherapy by predicting PD-L1 expression levels using a CNN model that analyzes the tumor microenvironment from histopathological images. These studies demonstrate that AI has the potential to prevent unnecessary toxicity by enabling early

prediction of treatment response, personalize treatment strategies, and enhance clinical decision-making.

6.3. Emerging Trends and Innovative Technologies

6.3.1. Explainable AI (XAI) for Transparent Decision-making

The rapid development of artificial intelligence in recent years has led to the emergence of increasingly complex models capable of performing tasks with high accuracy. However, the lack of transparency, limited interpretability, and uncertainty about reliability of these models, often referred to as "black box AI," have raised significant concerns. This development has led to the emergence of a research area known as Explainable Artificial Intelligence (XAI).

Recent studies have demonstrated that XAI methods can significantly enhance clinical trust and model transparency in CRC diagnosis and prognosis. For instance, Byun et al. (2023) implemented a Grad-CAM visualization technique on a convolutional neural network trained for histopathological image classification, allowing pathologists to visually confirm the regions contributing most to the model's prediction. Similarly, Hiroshima et al. (2022) used SHAP (SHapley Additive exPlanations) analysis to interpret a radiomics-based survival prediction model, identifying key imaging features that correlated strongly with poor prognosis. Moreover, Wulczyn et al. (2023) emphasized that interpretable deep learning systems not only improve diagnostic confidence but also facilitate regulatory approval by aligning AI outputs with human-understandable reasoning. These examples highlight how integrating XAI into clinical AI systems bridges the gap between algorithmic performance and clinical applicability, promoting safer and more transparent adoption of AI in oncology.

6.3.2. Integration of AI with Endoscopic Robotics and Augmented Reality

The integration of endoscopic robotic systems and augmented reality (AR) technologies with artificial intelligence significantly increases the accuracy and efficiency of surgical procedures. For example, AI-powered image processing algorithms enable the automatic detection of colorectal polyps in endoscopic images, preventing lesions that surgeons might otherwise miss. Similarly, AR-based systems help guide the surgeon by visualizing the patient's anatomical structure in real time, which is particularly critical in minimally invasive surgery where operating space is limited. Combining endoscopic robotic arm systems with AI-based decision support algorithms both shortens operative time and reduces the risk of complications. For example, a study conducted in Japan found that the polyp detection rate with AI-powered robotic endoscopy increased by 15% compared to traditional manual endoscopy. Such integrations allow surgeons to perform safer and more precise interventions.

6.3.3. AI-Assisted Personalized Medicine

In the treatment of colorectal cancer, AI-powered personalized medicine utilizes patient tumor genetics, histopathological characteristics, and lifestyle data to tailor treatment plans to the individual. For example, in a study conducted by the American Cancer Society, AI algorithms screened for MSI (microsatellite instability) and KRAS mutations to determine which patients would respond best to anti-EGFR therapy. Similarly, at a hospital in Singapore, AI-based analysis identified the risk of recurrence in early-stage colon cancer patients and implemented intensive monitoring in the high-risk group; this early intervention increased survival by 12%. In a study conducted in China, deep learning algorithms were able to classify polyps detected during colonoscopy as malignant with 94% accuracy.

Furthermore, in a retrospective study in the US, AI-assisted treatment planning optimized chemotherapy dosages based on patients' liver and kidney function, reducing side effect rates by 18%. These objective examples demonstrate that AI-powered personalized medicine in colon cancer treatment both increases treatment efficacy and significantly improves patient safety.

7. CHALLENGES AND LIMITATIONS

7.1. Data Quality, Imbalance and Standardization Issues

The accuracy of AI models in colorectal cancer research depends on data quality and lack of standardization. Differences in resolution, color, and labeling across colonoscopy and histopathological images obtained from different centers limit the generalizability of algorithms. Furthermore, underrepresentation of rare tumor types in datasets can negatively impact model performance (Yin et al., 2023).

7.2.Ethical and Regulatory Considerations

The use of AI systems in the diagnosis and treatment of colon cancer presents several ethical and legal challenges. Confidentiality of patient data, accountability for algorithm decisions, and regulatory standards in different countries pose obstacles to clinical application (Wang et al., 2025).

7.3.Lack of Interpretability and Clinician Trust

Because AI models are often considered “black boxes,” gaining clinicians’ trust is challenging. For example, even if an AI system detects polyp malignancy with high accuracy, it may not be directly integrated into a treatment plan if the decision-making mechanism remains unclear (Sikora et al., 2025).

7.4.Computational Cost and Implementation Barriers

Deep learning models operating on colon cancer images require high processing power. High-resolution image data and large datasets create cost and time constraints for model training and real-time use. Furthermore, existing hospital infrastructures are not suitable for integrating AI systems, limiting widespread implementation (Lubell, 2025).

8. FUTURE DIRECTIONS

One of the key future opportunities in colorectal cancer research is to increase the generalizability of AI models through multicenter collaborations and open datasets. Combining colonoscopy and histopathology images from different geographic regions and clinical centers increases data diversity and model accuracy. Furthermore, integrating AI systems with omics data (genomics, proteomics, metabolomics) allows for more holistic modeling of tumor biology, enabling more precise patient-specific risk assessments and treatment plans.

Furthermore, thanks to continuously learning AI systems, algorithms can update themselves with new patient data and clinical feedback, enabling them to provide adaptive diagnostic and treatment recommendations. For example, new colonoscopy images and genetic profiles from multiple hospitals can improve the system's polyp detection and malignancy prediction performance over time. Furthermore, integrating AI-based decision

support systems into routine clinical practice helps surgeons and oncologists optimize treatment plans. This can improve both early diagnosis and personalized treatment approaches, improving patient safety and survival rates (Kim et al., 2023; Lee et al., 2024; Zhang et al., 2025).

9. CONCLUSION

AI and deep learning-based analysis methods have made significant progress in the diagnosis and treatment of colon cancer in recent years. The use of approaches such as big data analysis, radiomics modeling, and histopathological image classification in clinical decision-making has enabled more accurate assessment of tumor morphology, genetic markers, and disease progression. These technological innovations have not only enhanced diagnostic precision but also allowed for early disease detection, tailored treatment strategies, and better patient outcomes. Deep learning models, in particular, can predict colon cancer prognosis with high accuracy when combined with clinical parameters, leading to an increasingly critical role for AI in oncology applications.

The transformative power of AI is clearly evident in its ability to redefine diagnostic and prognostic processes limited by classical methods. Through the multimodal integration of radiological, histopathological, and molecular data, a multidimensional map of tumor biology can be constructed. For example, AI-powered models have surpassed classical statistical approaches in areas such as treatment response prediction, recurrence risk analysis, and survival prediction. These advances have enabled more data-driven, objective, and patient-centered clinical decisions in colon cancer patients. This has increased both time and resource efficiency during the treatment process and has positively impacted patient prognosis.

However, interdisciplinary collaboration is essential for the effective use of AI applications in medicine. The integration of diverse disciplines such as medicine, computer engineering, bioinformatics, ethics, and data science enhances both the scientific validity and clinical usability of models. Studies that combine the domain knowledge of clinical experts with the algorithmic competence of data scientists strengthen both model accuracy and interpretability. Furthermore, clearly defining ethical frameworks, data privacy, and patient consent are essential for the sustainable adoption of AI technologies. Therefore, an interdisciplinary approach is not only a technical requirement but also a fundamental requirement for reliable clinical translation.

Moving forward, ensuring that AI systems are explainable, generalizable, and suitable for clinical integration is a top priority. XAI approaches will enable clinicians to better understand and confidently implement model decisions. Furthermore, multicenter data sharing, standardized imaging protocols, and increased large-scale validation studies will facilitate the performance of models in real-world clinical settings. Using AI in conjunction with human expertise will enable its integration into clinical decision-making processes, supporting every stage from diagnosis to treatment.

AI-powered analyses represent not only a technological innovation but also a paradigmatic shift in colon cancer research. This transformation, extending from diagnosis to prediction, is paving the way for a more precise, data-driven, and personalized understanding of oncology, driven by the digitalization of medicine. In the coming period, with the increasing clinical reliability of AI systems, these technologies will cease to be mere decision-support tools and become an integral component of healthcare. This new era, where human and digital intelligence work in harmony, opens the door to a more effective, predictive, and holistic approach to colon cancer management, both clinically and ethically.

REFERENCES

Akbari, M., Mohrekesh, M., Nasr-Esfahani, E., Soroushmehr, S. R., Karimi, N., Samavi, S., & Najarian, K. (2018, July). Polyp segmentation in colonoscopy images using fully convolutional network. In *2018 40th annual international conference of the ieee engineering in medicine and biology society (EMBC)* (pp. 69-72). IEEE.

Albashish, D., Khasawneh, A., & Alweshah, M. (2022). *Ensemble of adapted deep convolutional neural networks for multi-class histopathology colorectal cancer classification.* Biomedical Signal Processing and Control, 73, 103466. https://doi.org/10.1016/j.bspc.2021.103466

Allison, J. E. (2010). Colorectal cancer screening guidelines: the importance of evidence and transparency. *Gastroenterology*, *138*(5), 1648-1652.

Bahrambanan, F., Karimi, R., & Rahmanian, M. (2025). *The development of an efficient artificial intelligence-based colon cancer classification system combining bioinformatics*. *Computers in Biology and Medicine, 181*, 109576. https://doi.org/10.1016/j.compbiomed.2024.109576

Baxter, N. N., Goldwasser, M. A., Paszat, L. F., Saskin, R., Urbach, D. R., & Rabeneck, L. (2009). Association of colonoscopy and death from colorectal cancer. *Annals of internal medicine*, *150*(1), 1-8.

Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., & Madabhushi, A. (2019). Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, *16*(11), 703-715.

Byun, J., Lee, S., & Kim, H. (2023). *Visual interpretability of deep learning models for colorectal cancer histopathology using Grad-CAM*. *Scientific Reports, 13*(1), 19572. https://doi.org/10.1038/s41598-023-46944-2

Chen, H., Li, H., Li, H., Zhao, X., & Yang, J. (2022). *IL-MCAM: Interactive learning with multi-channel attention mechanism for colorectal cancer histopathological image classification.* arXiv preprint arXiv:2206.03368. https://arxiv.org/abs/2206.03368

Cruz, J. A., & Wishart, D. S. (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics, 2*, 59–77. https://doi.org/10.1177/117693510700200003

Cybernet Systems Corporation. (2018). EndoBRAIN/EndoBRAIN-Eye product overview. Tokyo, Jpan.

Cybernet Systems Corporation. (2020). EndoBRAIN-Eye AI system for colonoscopy.

Çin, H., et al. (2024). *Deep learning-assisted polyp classification in colonoscopy: A multicenter study*. *Gastroenterology AI*, 5(2), 101-112.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. S., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Fujifilm. (2020). CAD EYE: Computer-Aided Detection and Diagnosis system. Retrieved October 11, 2025, from https://www.fujifilm.com

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Hiroshima, Y., Nakamura, K., & Ishikawa, T. (2022). *Explainable radiomics-based survival prediction in colorectal cancer using SHAP analysis*. *European Radiology, 32*(12), 8432–8443. https://doi.org/10.1007/s00330-022-08942-1

Hsu, J. B., Chen, Y. C., & Huang, Y. H. (2023). *Deep learning integration of genomic and clinical data for survival prediction in colorectal cancer*. *Scientific Reports, 13*(1), 14257. https://doi.org/10.1038/s41598-023-40681-0

Hwang, S., Oh, J., Tavanapong, W., Wong, J. ve De Groen, PC (2007, Eylül). Eliptik şekil özelliği kullanılarak kolonoskopi videosunda polip tespiti. *2007 IEEE görüntü işleme uluslararası konferansı* (Cilt 2, s. II-465). IEEE.

Jia, Y. (2015, Aralık). Görüntü segmentasyonuna dayalı geliştirilmiş yöntem kullanılarak kablosuz kapsül endoskopi görüntülerinde poliplerin otomatik tespiti. *2015 IEEE Uluslararası Robotik ve Biyomimetik Konferansı (ROBIO)* (s. 1631-1636). IEEE.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kang, J., & Gwak, J. (2019). Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, *7*, 26440-26447.

Karargyris, A. ve Bourbakis, N. (2009, Nisan). Log gabor filtreleri kullanılarak kablosuz kapsül endoskopi videolarında poliplerin tanımlanması. *2009 IEEE/NIH Yaşam Bilimi Sistemleri ve Uygulamaları Çalıştayı* (s. 143-147). IEEE.

Kather, J. N., Krisam, J., Charoentong, P., et al. (2020). *Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer*. *Nature Medicine, 26*(1), 105–108. https://doi.org/10.1038/s41591-019-0462-y

Kekelidze, M., D'Errico, L., Pansini, M., Tyndall, A., & Hohmann, J. (2013). Colorectal cancer: current imaging methods and future perspectives for the diagnosis, staging and therapeutic response evaluation. *World journal of gastroenterology: WJG*, *19*(46), 8502.

Khazaee Fadafen, M., Shiri, I., Akbari, M. E., & Zaidi, H. (2023). *Improved classification of colorectal polyps in histopathological images using ensemble deep learning and stain normalization.* Scientific Reports, 13(1), 3904. https://doi.org/10.1038/s41598-023-31067-4

Kim, S., Nguyen, T., & Park, Y. (2023). *Continuous learning in AI-assisted colonoscopy: Enhancing polyp detection accuracy*. Artificial Intelligence in Medicine, 137, 102933. https://doi.org/10.1016/j.artmed.2023.102933

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005& Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Lee, H., Park, J., & Choi, S. (2024). *Adaptive AI systems in colorectal cancer diagnosis: A multi-center study*. Journal of Medical Imaging and Health Informatics, 14(3), 201-212. https://doi.org/10.1166/jmihi.2024.2012

Leo, M., Bruno, F., & Pellicano, C. (2024). *Convolutional neural networks in the diagnosis of colon adenocarcinoma grading from histological images*. *Diagnostics, 14*(7), 1293. https://doi.org/10.3390/diagnostics14071293

Lin, P. J., Chen, L. M., & Wang, S. H. (2024). *Multi-modal artificial intelligence enhanced systematic colorectal cancer pathology evaluation*. *Journal of Clinical Oncology, 42*(16_suppl), 3532. https://doi.org/10.1200/JCO.2024.42.16_suppl.3532

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Lubell, J. (2025). *How AI is improving the rate of colon cancer risk detection*. American Medical Association. https://www.ama-assn.org/practice-management/digital-health/how-ai-improving-rate-colon-cancer-risk-detection

Mazaki, J., Iwata, M., & Suzuki, K. (2024). *Novel artificial intelligence combining convolutional neural network and SVM approach for prognostic prediction in colorectal cancer*. *Pathology – Research and Practice, 259*, 155614. https://doi.org/10.1016/j.prp.2024.155614

Medtronic. (2019). GI Genius™ intelligent endoscopy module. Dublin, Ireland.

Medtronic. (2021). U.S. FDA grants De Novo clearance for GI Genius™ AI system for colonoscopy. Retrieved October 12, 2025, from https://news.medtronic.com/2021-04-12-U-S-FDA-Grants-De-Novo-Clearance-for-First-and-Only-Artificial-Intelligence-System-for-Colonoscopy

Medtronic. (2024). GI Genius Summit 2024: Future of AI in Endoscopy. Retrieved October 12, 2025, from https://news.medtronic.com/2024-04-10-Medtronic-unveils-the-future-of-AI-in-GI-Genius-Summit-2024-reveals-innovations-and-collaborations-that-advance-endoscopic-care

Merabet, A., Saighi, A., Saad, H., Ferradji, M. A., Laboudi, Z., Almaktoom, A. T., ... & Mohamed, A. W. (2025). AI for colon cancer: A focus on classification, detection, and predictive modeling. *International Journal of Medical Informatics*, 106115.

Mitsala, A., Tsalikidis, C., Pitiakoudis, M., Simopoulos, C., & Tsaroucha, A. K. (2021). Artificial intelligence in colorectal cancer screening, diagnosis and treatment. A new era. *Current Oncology*, *28*(3), 1581-1607.

Mori, Y., Kudo, S. E., Misawa, M., Saito, Y., Ikematsu, H., Hotta, K., ... & Mori, K. (2018). Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. *Annals of internal medicine*, *169*(6), 357-366.

Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S. M. R., Jafari, M. H., Ward, K., & Najarian, K. (2016, August). Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 1373-1376). IEEE.

NEC Corporation. (2020). AI-assisted colonoscopy solution announcement. Retrieved October 11, 2025, from https://www.nec.com

Olympus Corporation. (2020). ENDO-AID CADe for colonoscopy. Retrieved October 12, 2025, from https://www.olympus-global.com

Olympus Global (2024). Olympus and Odin Vision receive FDA 510(k) clearance for CADDIE AI endoscopy system. Retrieved October 12, 2025, from https://www.olympus-global.com/news/2024/nr02743.html

Özdemir, Ö., Büyüktoka, R. E., Argon, A., Sürücü, M., İşler, Y., Söylemez, C. M., Kahraman, E. G., Kahraman, D. S., & Demirci, F. (2025, Ekim 13–15). İmmün skor ve radyoloji destekli yapay zeka modeli opere kolon kanseri hastalarında sağkalımı öngörebilir mi? [Sunum]. 23. Ulusal Onkoloji Kongresi, İzmir, Türkiye.

Park, S., Lee, M., & Kwak, N. (2015). Polyp detection in colonoscopy videos using deeply-learned hierarchical features. *Seoul National University*.

Patel, S., Kumar, R., & Singh, D. (2025). *Advanced deep learning for multi-class colorectal cancer histopathology: Integrating transfer learning and ensemble methods.* Computers in Biology and Medicine, 175, 107890. https://doi.org/10.1016/j.compbiomed.2025.107890

Pentax Medical. (2020). DISCOVERY CADe system for endoscopy. Tokyo, Japan.

Rathore, S., Hussain, M., Ali, A., & Khan, A. (2013). A recent survey on colon cancer detection techniques. *IEEE/ACM Transactions on computational biology and bioinformatics*, *10*(3), 545-563.

Ribeiro, E., Uhl, A., & Häfner, M. (2016, June). Colonic polyp classification with convolutional neural networks. In *2016 IEEE 29th international symposium on computer-based medical systems (CBMS)* (pp. 253-258). IEEE.

Rosen, A. W., Andersen, L. L., & Nielsen, J. S. (2025). *Clinical implementation of an AI-based prediction model for decision support in colorectal cancer surgery*. *Nature Medicine, 31*(2), 247–257. https://doi.org/10.1038/s41591-025-03942-x

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, *35*(5), 1285-1298.

Sikora, N., et al. (2025). *ColonScopeX: Leveraging Explainable Expert Systems with Multimodal Data for Improved Early Diagnosis of Colorectal Cancer*. arXiv. https://arxiv.org/abs/2504.08824

Tajbakhsh, N., Gurudu, SR ve Liang, J. (2015). Şekil ve bağlam bilgilerini kullanarak kolonoskopi videolarında otomatik polip tespiti. *IEEE tıbbi görüntüleme işlemleri*, *35* (2), 630-644.

Wang, J., et al. (2025). *The clinical application of artificial intelligence in cancer precision medicine*. Translational Medicine, 23(1), 1-12. https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-025-06139-5

Wang, P., Berzin, T. M., Brown, J. R. G., Bharadwaj, S., Becq, A., Xiao, X., ... & Liu, X. (2019). Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*, *68*(10), 1813-1819.

Wision AI. (2021). EndoScreener CADe system overview. Shanghai, China. Retrieved October 12, 2025, from https://www.wisionai.com

Wulczyn, E., Steiner, D. F., & Xu, Z. (2023). *Interpretable deep learning for histopathology and its implications for clinical decision-making*. *Nature Biomedical Engineering, 7*(5), 472–485. https://doi.org/10.1038/s41551-022-00953-1

Xie, C., Zhang, Y., & Liu, H. (2025). *Multimodal data integration for biologically relevant artificial intelligence to guide adjuvant chemotherapy in stage II colorectal cancer*. *eBioMedicine, 108*, 105912. https://doi.org/10.1016/j.ebiom.2025.105912

Yalçın, Ş. (Ed.). (2024). *Colorectal cancer: Diagnosis and treatment* [E-book]. Middle East Advertising, Promotion, Publishing, Tourism, Education, Construction, Industry and Trade Co. https://www.dijitalakademi.org

Yamada, M., Ueno, M., & Tanaka, T. (2022). *Radiomic analysis using preoperative CT images for recurrence risk prediction in stage II–III colon cancer*. *European Radiology, 32*(10), 6621–6632. https://doi.org/10.1007/s00330-022-08629-7

Yin, Z., et al. (2023). *Application of artificial intelligence in diagnosis and treatment of colorectal cancer*. Frontiers in Medicine, 10, 1128084. https://www.frontiersin.org/articles/10.3389/fmed.2023.1128084/full

Young, E., Edwards, L., & Singh, R. (2023). The role of artificial intelligence in colorectal cancer screening: Lesion detection and lesion characterization. *Cancers*, *15*(21), 5126.

Zhang, Y., Li, X., & Chen, W. (2025). *Integration of AI and genomic data for personalized colorectal cancer treatment*. Computational Oncology, 9(1), 45-59. https://doi.org/10.1016/j.compoc.2025.01.005

Zhang, Z., Wang, Y., Liu, J., & Li, X. (2024). *HCCANet: A hybrid channel and spatial attention network for colorectal cancer differentiation grade classification.* Frontiers in Oncology, 14, 944881. https://doi.org/10.3389/fonc.2024.944881

CHAPTER 5

A COMPARATIVE ANALYSIS OF WORD REPRESENTATION MODELS IN NATURAL LANGUAGE PROCESSING: FROM CONVENTIONAL FREQUENCY-BASED TO CONTEXTUALIZED EMBEDDINGS

Hilal ÇELİK[1], Ramazan KATIRCI[2]

1 Res. Asst.. Gör.; Sivas Bilim ve Teknoloji Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü. hilalcelik@sivas.edu.tr ORCID No: 0000-0001-5428-3411
2 Prof. Dr.; Sivas Bilim ve Teknoloji Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü. ramazankatirci@sivas.edu.tr ORCID No: 0000-0003-2448-011X

1. INTRODUCTION

Word representation constitutes a fundamental component of modern natural language processing, as it enables linguistic units to be transformed into numerical forms that computational systems can interpret. Rather than treating words as isolated symbols, contemporary approaches encode them as vectors that inhabit a continuous semantic space. Within such spaces, words that share syntactic or semantic characteristics tend to appear in closer proximity, allowing downstream models to exploit these relationships more effectively. Methods for representing words have traditionally been grouped into three broad categories. Conventional representations, such as Bag-of-Words and TF-IDF, rely on frequency statistics and co-occurrence patterns, providing sparse but interpretable features. Distributed representations—including models like Word2Vec, GloVe and fastText—map words into dense vector spaces by leveraging contextual information extracted from large corpora, thereby capturing underlying semantic regularities. More recently, contextual embedding models have emerged, producing representations that vary dynamically depending on the surrounding text. These models generate richer semantic signals by integrating the broader linguistic environment into each token's encoding. Architectures such as BERT, which applies a bidirectional Transformer mechanism, exemplify this shift by simultaneously modelling left and right contexts, achieving substantial improvements over earlier embedding techniques and redefining standards for semantic and contextual understanding in NLP.

2. WORD REPRESENTATION MODELS

Word representation refers to techniques that transform words into numerical formats suitable for processing by computational models. A commonly used method is word embedding, where words are encoded as dense vectors positioned within a continuous vector space. Within this space, words with similar meanings tend to appear near one another, and the resulting vectors capture syntactic as well as semantic patterns in the language [1]. This idea—often described as vector semantics—treats each word as a point in a multidimensional semantic space derived from the distribution of neighboring terms across large text corpora [2]. Such vector-based representations, known as embeddings, allow machine learning (ML) models to handle linguistic information more effectively than traditional frequency-based approaches [3].

A broad spectrum of methods has been proposed for the representation of linguistic information, progressing from early frequency-based statistical techniques to more advanced distributed and contextual embedding approaches. As depicted in Figure 1, existing word representation

methods are commonly classified into three main categories: conventional, distributed, and contextual models. Conventional approaches, including Bag-of-Words (BoW), n-gram models, and term frequency–inverse document frequency (TF-IDF), represent textual data through surface-level frequency distributions and co-occurrence statistics. In contrast, distributed representation models—such as Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Word2Vec, GloVe, and fastText—encode words into dense, low-dimensional vector spaces that capture latent semantic relationships and structural regularities within large corpora [2], [4]. More recently, contextualized models like ELMo, GPT, and BERT have further advanced word representation by generating dynamic embeddings that vary according to contextual usage, thereby enabling more precise semantic interpretation and improved performance across downstream NLP tasks [5].

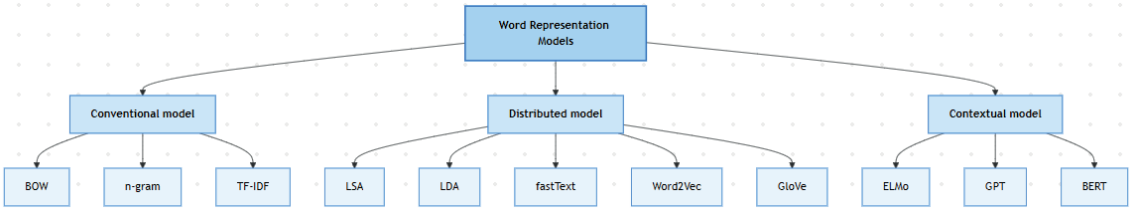


Figure 1. Categorization of word representation models in natural language processing.

2.1. Conventional word embedding

Conventional word embedding, also called count-based/frequency-based models, is categorized into a BoW, n-gram, and term TF-IDF models [3].

Bag-of-Words (BoW)

The BoW model is a widely used representation method in which an object (e.g. a document or image) is represented as a discrete set of elements, known as words or visual words, regardless of their order. Each object is encoded as a histogram of these words, summarizing the frequency of occurrence. In image analysis, key points are often quantized into visual words using clustering algorithms such as K-means, whereas in text, words are counted directly. Although simple, the BoW representation effectively captures the presence and frequency of features, making it useful for tasks like object categorization and document classification [6], [7].

n-gram

N-grams are sets of n consecutive words or characters extracted from a text, where n is usually one, two, or three. One-letter n-grams are called unigrams (or monograms), two-letter n-grams

are called bigrams (or digrams) and three-letter n-grams are called trigrams [8]. The order-*n* parameters of an n-gram model can be viewed as forming the transition matrix of a Markov model, where the states correspond to sequences of $n - 1$ words. To represent words numerically, continuous vector representations xwx_wxw are first generated for each word www in the dictionary DDD, often using models such as Skip-gram or GloVe, trained efficiently over large unlabeled corpora [9], [10].. These vector representations enable n-gram models to leverage both frequency-based and semantic information for NLP tasks.

Term Frequency-Inverse Document Frequency

TF-IDF is a numerical measure used to evaluate the significance of a word within a specific document relative to a broader collection of documents. It takes into account both how frequently a word appears in the document (term frequency) and how uncommon it is in the broader corpus (inverse document frequency). Words with high TF-IDF values are considered more informative, making this method useful for tasks such as keyword extraction, search result ranking, and document categorisation. [11], [12].

2.2. Distributed Word Embedding Models

Distributed word embedding techniques encode words as dense, continuous vectors situated in a multidimensional semantic space, enabling the representation of syntactic and semantic relationships based on contextual usage within large corpora. These models aim to produce rich and informative vector representations that capture linguistic patterns and improve the performance of natural language processing applications. Numerous distributed embedding methods have been introduced in the literature; among them, five widely referenced approaches—Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Word2Vec, GloVe and fastText—are outlined in the following subsections to demonstrate different strategies for modeling semantic and contextual information in text.

Latent Semantic Analysis (LSA)

LSA is a statistical method designed to uncover hidden associations between words and documents within a corpus. It works by constructing a term–document matrix and applying Singular Value Decomposition (SVD) to project words and documents into a reduced-dimensional latent semantic space, where terms with similar meanings appear closer to one another [2], [13]. By capturing word co-occurrence patterns, LSA enhances tasks such as similarity measurement and information retrieval. However, its reliance on linear algebraic

transformations instead of probabilistic modeling limits its ability to represent more complex semantic and syntactic structures present in modern datasets [4].

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model designed for dimensionality reduction and topic discovery in large text corpora. It models each document as a mixture of latent topics, with each topic represented by a probability distribution over words [14]. Unlike earlier models, LDA provides a formal probabilistic framework that enables generalization to unseen documents. However, despite its theoretical strengths, LDA becomes computationally expensive when applied to large-scale datasets [15] and remains challenging to capture complex semantic and syntactic structures effectively [4].

Word2Vec

Word2Vec is a neural representation framework that generates dense vector embeddings of words by exploiting their distributional patterns within large-scale text corpora [13], [15]. The method captures semantic and syntactic regularities by mapping words with similar meanings to nearby locations within a continuous vector space [1], [16]. As shown in Figure 2, Word2Vec consists of two foundational architectures: Continuous Bag-of-Words (CBOW) and Skip-gram. In the CBOW model, the network predicts a missing target word by using its surrounding context, whereas the Skip-gram model performs the inverse operation by estimating the context words from a given central Word [15], [16]. CBOW is generally more computationally efficient for extensive datasets, while Skip-gram tends to yield superior performance for infrequent or rare words. The embeddings produced by Word2Vec are static, meaning that each lexical item is assigned a single vector representation that does not vary across different contexts [2]. Through this design, Word2Vec effectively models distributional characteristics of language and supports a variety of downstream NLP applications, including semantic similarity, analogy detection, and clustering.

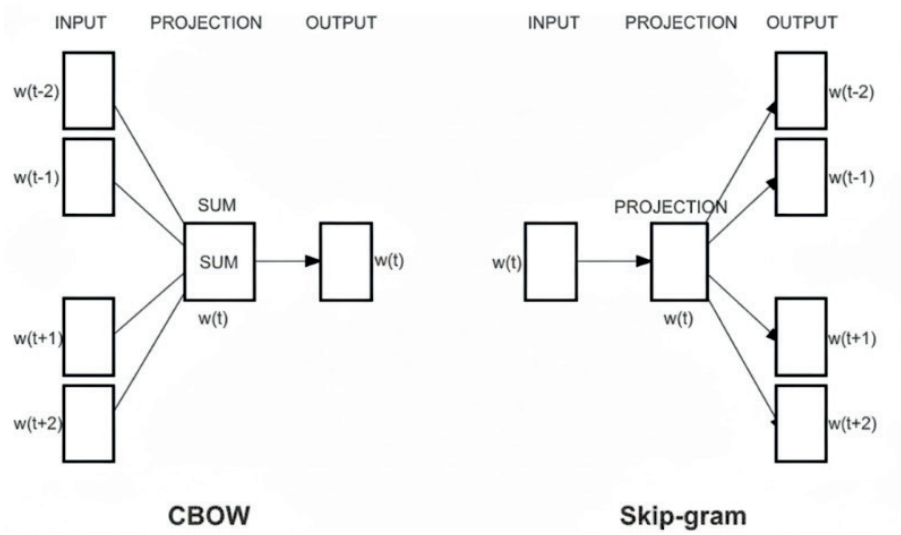


Figure 2. CBOW and Skip-gram Model [15]

Global Vectors for Word Representation (GloVe)

Global Vectors for Word Representation (GloVe) is an unsupervised word representation model that generates dense vector embeddings by leveraging global word co-occurrence statistics from a corpus. Unlike models that rely solely on local context, GloVe captures the overall distributional information across the entire corpus, allowing the resulting word vectors to reflect semantic relationships and meaning. The model's name, Global Vectors, emphasizes its use of corpus-wide statistical information to construct meaningful word embeddings [3], [13]

fastText

fastText is a word embedding model that incorporates subword information through character n-grams, enabling it to capture morphological features and local word order. This design allows the model to generate vector representations not only for known words but also for out-of-vocabulary (OOV) terms [3]. Each word is mapped to a shared low-dimensional space as a *d*-dimensional vector, reflecting both syntactic and semantic similarity. Unlike traditional embedding models that represent each word as a single token, fastText decomposes words into overlapping character n-grams and learns vector representations for these subword units [17]. By composing word vectors from their constituent n-grams, fastText effectively handles rare and unseen words, making it particularly suitable for morphologically rich languages [5].

2.3. Contextual Word Embeddings

Contextual word embeddings generate dynamic vector representations for words based on their surrounding context, in contrast to static embeddings such as Word2Vec or GloVe [18]. By capturing richer semantic and syntactic information, these embeddings have significantly

improved performance across various NLP tasks. Understanding context-dependent variation in word meanings is a key aspect of human language comprehension, supported by the lexicon. Contextualized models provide better word embeddings than static models and combining embeddings from different models can further enhance task performance [19].

Contextual word embedding models can be broadly classified into auto-regressive and auto-encoding approaches. Notable examples include ELMo (Embeddings from Language Models), GPT and BERT each employing different architectures to generate context-aware word representations [3].

Embeddings From Language Models (ELMo)

Embeddings from Language Models (ELMo) generates context-dependent word embeddings that captures both semantic and syntactic information [20]. Unlike conventional static embeddings, ELMo representations are derived from the entire input sentence, allowing word vectors to adapt based on surrounding context. In ELMo, words within the same sentence become more similar in higher layers as context-specificity increases, enhancing the model's ability to capture nuanced meanings [21].

Generative Pre Training (GPT)

GPT is based on a unidirectional Transformer architecture that produces context-sensitive word representations primarily optimized for natural language generation tasks [20]. Owing to its autoregressive design, the model generates each token by conditioning on previously generated tokens, which allows it to effectively model long-range dependencies within textual sequences. GPT is initially trained on large-scale unlabeled corpora, enabling it to learn general linguistic patterns that can be transferred across diverse tasks. This pre-training strategy provides substantial flexibility, as the target task domain does not need to closely align with the pre-training data. Subsequently, the model can be fine-tuned for task-specific applications—such as text generation, classification, or question answering—by updating its parameters to optimize downstream performance [21].

Bidirectional Encoder Representations from Transformers (BERT)

BERT represents a major advancement in contextual word embedding methods. It employs a bidirectional Transformer architecture to model both left and right contexts simultaneously, enabling the generation of deep, context-dependent word embeddings that capture subtle semantic and syntactic nuances [22]. Unlike unidirectional models, BERT learns continuous

context representations, allowing it to distinguish fine-grained variations in word meaning across different usages [19].

In BERT, word embeddings evolve through multiple Transformer layers: lower layers capture general lexical and syntactic information, whereas higher layers produce context-specific semantic representations, in which words within the same sentence become increasingly dissimilar yet remain more related than randomly sampled words [21]. The model is pre-trained on large unlabeled corpora using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives, which enable the learning of rich bidirectional dependencies from raw text. Fine-tuning on labeled datasets further adapts these embeddings for downstream NLP tasks such as question answering, sentiment analysis, or text classification [22]. Owing to its multi-layer bidirectional Transformer encoder, BERT outperforms earlier contextual models such as GPT and ELMo, setting a new standard for semantic and contextual representation in NLP [20].

CONCLUSION

The present era is defined by the rise of contextual models, such as ELMo, GPT and BERT. These models represent a profound paradigm shift because they generate dynamic, context-aware embeddings. This breakthrough allows models to understand the subtle nuances of word meaning that depend on context, mirroring human linguistic comprehension.

These dynamic, context-aware representations will be the basis for the next generation of NLP capabilities in the future. Researchers are leveraging these deep, bidirectional dependencies to solve increasingly complex language tasks. This solidifies the contextual approach as an indispensable foundation for future advancements in semantic and contextual representation.

REFERENCES

[1] S. F. Ahmed *et al.*, *Deep learning modelling techniques: current progress, applications, advantages, and challenges*, vol. 56, no. 11. Springer Netherlands, 2023. doi: 10.1007/s10462-023-10466-8.

[2] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2023.

[3] D. S. Asudani, N. K. Nagwani, and P. Singh, *Impact of word embedding models on text analytics in deep learning environment: a review*, vol. 56, no. 9. Springer Netherlands, 2023. doi: 10.1007/s10462-023-10419-1.

[4] Y. Zhang and D. Long, "WHEN TEXT EMBEDDING MEETS LARGE LANGUAGE MODEL: A COMPREHENSIVE SURVEY," 2025.

[5] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A Comprehensive Survey on Word Representation Models : From Classical to State-Of-The-Art Word Representation Language Models," pp. 1–46.

[6] Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding bag-of-words model: A statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, no. 1–4, pp. 43–52, 2010, doi: 10.1007/s13042-010-0001-0.

[7] W. A. Qader, M. M. Ameen, and B. I. Ahmed, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," *Proc. 5th Int. Eng. Conf. IEC 2019*, pp. 200–204, 2019, doi: 10.1109/IEC47844.2019.8950616.

[8] K. Kukich, "Techniques for Automatically Correcting Words in Text," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 377–439, 1992, doi: 10.1145/146370.146380.

[9] Peter F. Brown, P. V. DeSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-Based n-gram Models of Natural Language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, 1992.

[10] R. Lebret and R. Collobert, "N-gram-based low-dimensional representation for document classification," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Work. Track Proc.*, pp. 1–8, 2015.

[11] Vimala Balakrishnan and Ethel Lloyd-Yemoh, "Stemming and Lemmatization: A

Comparison of Retrieval Performances," *Lect. Notes Softw. Eng.*, vol. 2, no. 3, pp. 262–267, 2014.

[12] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[13] J. Pennington, R. Socher, and D. M. Christopher, "GloVe: Global Vectors for Word Representation," *Br. J. Neurosurg.*, vol. 31, no. 6, pp. 682–687, 2014, doi: 10.1080/02688697.2017.1354122.

[14] C. D. Manning, P. Raghavan, and H. Schütze, "An Introduction to Information Retrieval," *Libr. Rev.*, vol. 53, no. 9, pp. 462–463, 2004, doi: 10.1108/00242530410565256.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781

[16] Ç. ACI and A. ÇIRAK, "Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması," *Bilişim Teknol. Derg.*, vol. 12, no. 3, pp. 219–228, 2019, doi: 10.17671/gazibtd.457917.

[17] P. Mojumder, M. Hasan, M. F. Hossain, and K. M. A. Hasan, "A study of fasttext word embedding effects in document classification in bangla language," *Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST*, vol. 325 LNICST, no. March 2021, pp. 441–453, 2020, doi: 10.1007/978-3-030-52856-0_35.

[18] W. Zhou and J. Bloem, "Comparing contextual and static word embeddings with small philosophical data," *KONVENS 2021 - Proc. 17th Conf. Nat. Lang. Process.*, pp. 253–259, 2021.

[19] S. Nair, M. Srinivasan, and S. Meylan, "Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge," pp. 129–141, 2020.

[20] M. Thapa, P. Kapoor, S. Kaushal, and I. Sharma, "A Review of Contextualized Word Embeddings and Pre-Trained Language Models, with a Focus on GPT and BERT," no. IC3Com 2024, pp. 205–214, 2025, doi: 10.5220/0013305900004646.

[21] K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMO, and GPT-2 embeddings," *EMNLP-IJCNLP 2019 - 2019*

Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 55–65, 2019, doi: 10.18653/v1/d19-1006.

[22] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Didirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.