iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VIII **Month of publication:** August 2026

DOI: https://doi.org/10.22214/ijraset.2026.83151

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

A Multimodal Deep Learning Framework for Scientific Insight Generation Using Transformer Based NLP and VIT

Ramansh, Sulaksh Malan, Apoorv Verma, Aman Kumar
Dept. of Computer Science & Engg. Lovely Professional University Punjab, India

Abstract: ***Researchinthesciencescontinuestoproducedata in various forms, in forms such as dense technical prose, experimentalfigures,andstructuredmeasurementtables,but themajorityofAI-assistedanalysistoolscanonlyprocessone of these forms simultaneously. This fragmentation constrained the richness of insight that may be provided by automatedsystems as, in many cases, scientific discoveriesof significance are made by patterns that cut across modalities. Our system has a multimodal framework which directly encodes domain specific scientific language with SciBERT, matches visual content with textual context with CLIP(ViT- B/32), and purpose-built neural encoder of structured numericaldata. Instead offusingsuchrepresentationswitha basicconcatenationstep,weproposeaCross-ModalAttention mechanism and a Gated Fusion layer which dynamically scalestheweightingofeachmodalitybasedonhowrelevantit is to the query under consideration.***

Index Terms- Multimodal Learning, Scientific Insight Generation,Cross-ModalAttention,GatedFusion,SciBERT, CLIP,FAISS, Explainable AI, SHAP, ArXivDataset,Natural LanguageProcessing,Computer Vision,TransformerModels

I. INTRODUCTION

Scientific research produces knowledge in varied forms at the same time.Ascientist involved inthe researchofprotein folding needs to read and digest written abstracts inhundreds ofarticles, analyse microscopy pictures and structural schemes and examinations refine-grained products of laboratory tests. Although scientific discovery can be inherently multimodal, current AI-assisted analysis tools have applied the modalities independently of one another. An analyst requires external software and computer vision software as well as statistical softwaretocombineinsightsmanuallyusingcurrenttools,which istime-consuming,error-prone,anditdemand sagreatdealof domain expertise [1]. The amount of scientific literature that is being generated around the world has never been this high. The number of peer-reviewed research papers published on all fields of science is more than two million annually. It is impossible to survey this literature fully by a single researcher or a single researchgroup, especially withinthecaseofneedtoscopeacross avarietyoftextualdescriptions,visualexperimentalfindings,and coordinate data retrieval on this topic. This flood of information is one of the most urgent practical issues of the contemporary scientific practice [4].

A. Shortcoming of Unimodal Approaches

These limitations of unimodal approaches are mentioned in the literature. The visual evidence found in scientific figures cannot beencodedintextonlysystemslikestandardvariantoftheBERT version[13].Imageonlysystemsdonothave thecontextualizing coveragethatispresentedwithrelatedtextualdescriptions.Those systems that work purely with tabular experimental data lack the context of the research findings in general. Above all, currently available multimodal systems serving a broad range of activities like social media or ecommerce do not have the vocabulary and semantic knowledge needed in the case of a science text, where strict terminology is utilized with meaning substantial meaning that marks out very similar concepts [4].

Recentresearchhastriedtomakemultimodalintegrationsinthree majordirections.EarlyFusionusesaconcatenationmethodatthe input of all modalities but has issues of information loss and interference between modalities. The models are divided among late fusion trains and their output is synthesized, without taking intoaccountthetrans-modalrelationships,althoughthe modality is still relevant. Attention based Fusion trains the weights of interaction but progresses at expensive computationally trying to improve. None of these methods take the advantage of domain specific encoders or embed knowledge base to gain interpretability [6]

B. TheExplainabilityProblem

AnotherproblemisthatmodernAIsystemsareopaque.Scientific investigations require reproducibility and verifiability. In instanceswherebyanAIsystemcategorizesapaperaseitherpart of the computational biology or a specific hypothesis, the researchers must have a clear-cut explanation on how the reasoning occurred. Black-box AI despite its performance is essentiallyincompatiblewiththeepistemologicalrequirementsof scientificpractice,asallpropositionsfromthesciencehavetobe traceabletoevidenceandallitsfindingsverifiableinisolation[6]. Techniques like SHAP [7] and LIME [8] have been created to deliver post-hoc interpretability, but they are not integrated into multimodal scientific models as much.

C. ParticipationofRetrieval-AugmentedGeneration

With the advent of large-scale dense retrieval systems, there are nownewpossibilitiesofbasingAIoutputonverifiedknowledge. The FAISS library [9] allows performing approximate nearest neighbour search over millions of embeddings of documents efficiently, thus allowing semantic searches to be performed in real-time, which are computationally feasible. Lewis et al. [14] showed that dense retrieval combined with language generation, namelytheRetrieval-AugmentedGeneration(RAG)paradigmis significantly more effective when performing knowledge intensive tasks as it bases responses on what is retrieved as opposed to when a purely generative model is used that entirely relies on the memory of the parametric model. This fact directly inspires the inclusion of FAISS-based retrieval in FLAN-T5 willpower to our framework, where generated academic knowledge is framed in the context of already existing literature of 1.7 million ArXiv papers [11].

D. NovelContributions

The paper presents the AI-Driven Multimodal Framework of Scientific Insight Generation as an attempt to solve these limitations in four major contributions:

1) Unified Multimodal Encoding: It is a pipeline that transforms scientific text, images, and tabular data to similarembeddingrepresentationscreatedbypretraininga domain-specific model. SciBERT [4] reads research text with scientific vocabulary in mind, CLIP [5] encodes images in an image-text semantic space of direct cross-modalcomparison,andaMLPencoderattentionencoderis a system that takes a tabular experimental data, which is projected into a shared 512-dimensional space. Intelligent Fusion: Cross-modal Attention mechanism (8 heads) with a dynamically-weighted (contextually based) GatedFusioncoverquasi-diversityofmodalitiesin contribution. In contrast to mere concatenation or averaging, the technique achieves use of the modalities giving most useful signals when presented with a specific input and changes weighting policy according to content and quality of available data.

2) Insight Generation: Generation of structured scientific insightssuchassummaries,researchgapidentificationand hypothesis proposals via the integration of FLANT5 [10]. This generation power is supported more by FAISS-based retrieval [9] on 1.7 million arXiv papers [11] with the provision of grounded references that contextualize generated insights in the already existing literature.

3) Complete explainability: A combination of three explanationmechanismsdescribedbelow,namelySHAP [7] to find the relative attribution of tabular features using cooperativegametheory,[LIME8]tofindtheattributionof text tokens using linear approximation.

The whole configuration is supported by a Flask REST API backend, allowing it to use inexpensive commodity hardware.

E. PaperOrganization

The rest of this paper will be structured in the following way. Section II reviews the related work in multimodal learning, domain-specific language models, domain understanding vision- languagealignment,efficientretrievalandexplainableAI.Section III describes the entire methodology which consists of system architecture, encoder design, fusion mechanisms, downstream tasks, and the explainability framework. Section IV provides results and discussion. Section V contains conclusion and future work directions.

II. LITERATURE REVIEW

A. MultimodalLearninginScientificDomains

Processing heterogeneous scientific data through a single unified system has been a long-standing challenge. Ngiam et al. [1] provided early evidence that jointly trained cross modal models outperform independently trained unimodal counterparts, motivating the use of shared representation spaces for mixed- modalityinput.

Subsequentworkinthescientificdomainexplored figurecaptioning[2]andbiomedicalvisualtextalignment[3],yet bothsystems werepurpose-builtforasinglepairingofmodalities rather than arbitrary combinations of text, images, and experimentaltables.Akeydifficultydistinguishingscientificdata from general web content is that scientific figures encode quantitativeinformationtrends,comparisons,errormarginsthat cannot be recovered from captions alone. Equally, a caption withoutitsfigurelacksthevisualevidencethatsupportstheclaim. Any system targeting scientific understanding must therefore model both within-modality semantics and cross-modality dependencies [12].

B. Domain-SpecificLanguageModels

Standardpre-trainingcorporausedformodelssuchasBERT[13] consist primarily of encyclopaedia articles and digitised books, which share little vocabulary or stylistic structure with peer- reviewed research. Scientific prose is dense with multi-word technical terms, nested clauses, and implicit assumptions that require domain grounding to interpret correctly. To address this, Beltagy et al. [4] released SciBERT, built on the BERT architecture but trained from scratch on approximately 1.14 million full papers drawn from ArXiv and Semantic Scholar. A custom vocabulary constructed from that corpus keeps domain specific compound terms intact during tokenisation, which reduces sequence fragmentation and preserves meaning in downstream tasks such as relation extraction and multi-label document classification. In the proposed framework, a Sentence Transformer variant is used for throughput-sensitive batch encoding, while SciBERT serves as the high-fidelity option for single-document analysis; both apply mean pooling across all token positions rather than relying on the [CLS] representation, sincerelevant contentinscientificabstractsisnotconfinedtothe opening tokens. Raffel et al. [10] introduced T5, a sequence-to- sequence modelthatcastseveryNLPproblemasconditionaltext generation. The FLAN-T5 variant, further fine-tuned with instruction templates across hundreds of tasks, retains strong generalisation to unseen prompts. These properties make FLANT5 suitable for structured scientific output such as concise summaries, identification of open problems, and candidate hypotheses.

C. Vision-LanguageAlignment

Radford et al. [5] introduced CLIP, which trains dual image and text encoders using a contrastive objective over 400 million web sourced image-text pairs. The resulting 512dimensional joint space places conceptually related images and text in close geometric proximity, enabling cross-modal search without any task-specific labelling. For scientific content this is valuable because a natural-language description of a figure type and the figure itself can be directly compared, supporting retrieval and zero-shot categorisation of charts, diagrams, and microscopy images alike. The image encoder within CLIP adopts the Vision Transformer architecture [12], which divides each input into a grid of 32 × 32 pixel tiles and applies multi-head self-attention acrossalltilessimultaneously.Thisstandsincontrastto convolutional approaches, where receptive fields grow only graduallyacrosslayers. For scientific figures, where a data point inonecornerofaplotmaybedirectlyqualifiedbyalegendinthe opposite corner, the ability to model arbitrary spatial dependencies in a single attention step is practically important.

D. ApproximateNearestNeighbourRetrieval

Scaling semantic search to corpora of millions of documents requires sub-linear retrieval strategies. The FAISS toolkit [9] addressesthisthroughseveralindextypes;theInvertedFile(IVF) variantquantisestheembeddingspaceintoVoronoiregionsusing k-meanscentroidsandrestrictseachquerytoitsnearestpartitions. Ona50,000-documentindexwith256dimensionalvectorsand
100 centroids, the IVF structure achieves 0.02-second query latency compared to roughly 2.0 seconds for an exhaustive flat search, while recovering over 99% of the ground-truth nearest neighbours. Lewis et al. [14] showed that prepending retrieved passages to a generation prompt the Retrieval-Augmented Generation pattern substantially reduces factual errors compared with relying on parametric model memory alone. Adopting this principle, the proposed framework retrieves the top-five semantically similar papers from a 1.7-milliondocument ArXiv index[11] and suppliesthemasgroundingcontextto the FLAN- T5 generator.

E. ExplainabilityinHigh-StakesAI

Lipton [6] draws a practical distinction between model transparency, which concerns the interpretability of weights and operations, and post-hoc explanation methods, which produce human-readable accounts of individual predictions without exposing internal mechanics. Both are argued to be necessary when AI-generated outputs inform consequential decisions, as they do in scientific research where a misclassified paper or a spurious hypothesis can waste substantial effort. SHAP [7] grounds attribution in Shapley values from cooperative game theory.Eachfeaturereceivesascoreequaltoitsaveragemarginal contribution across all possible feature orderings, satisfying consistency, efficiency, and symmetry properties.

The KernelSHAP variant estimates these values through weighted least squares without requiring model gradients, making it applicable to the tabular encoder regardless of its internal structure. LIME [8] takes a complementaryapproach: it perturbs the input locally, evaluates the model on each perturbation, and fits a sparse linear surrogate whose coefficients serve as token-levelimportancescores.Together,SHAPandLIMEcoverfeature level and token-level explanation respectively, while gate values extracted directly from the fusion layer provide an additional modality-levelaccountofwhichdatasourcemostinfluencedeach prediction.

F. ResearchGap

Taken individually, each component discussed above has been validated in prior work. What has not been attempted is their integration into a single cohesive system: domain-adapted encoders for text [4], image [5], and tabular data; cross-modal attention and gated fusion for adaptive modality weighting; million-scale semantic retrieval via FAISS [9]; instruction tuned insightgenerationwithFLAN-T5[10];andsimultaneousposthoc explanation through SHAP [7], LIME [8], and fusion gate analysis. The framework proposed in this paper fills that gap.

III. METHODOLOGY

This section presents the complete methodology of the proposed framework, beginning with the overall system architecture, followed by detailed descriptions of each encoding module, the fusion mechanism, downstream tasks, and the explainability framework. The overall architecture is illustrated in the figure below. The framework employs a six-layer modular architecture progressingfromrawmultimodalinputstoexplainablescientific insights,servedthroughaFlaskRESTAPIbackend.Eachlayeris independently testable and replaceable. The six layers are:

Layer1(Input):Ittakesinthreekindsofdatatext(likeresearch abstracts), images (those classic scientific figures), and tables in CSV format (think experimental results). Each type gets its own custom cleanup step.

Layer2(Pre-processing):Textrunsthroughregexfilterstowipe out LaTeX math, citation markers, and links, plus some extra whitespace scrubbing. For images the system sharpens contrast without messing with colour by running CLAHE on the luminance channel split into 8x8 blocks and clipped at 3.0, so details pop but colours stay true.

TablesgetfilledinusingKNNformissingdata(withk=5),allthe categories turn numeric with label encoding, and it normalizes everything with Standard Scaler.

Layer 3 (Encoding): Each data type then goes into its own encoder. For text, AIMMF uses either SciBERT or a Sentence Transformer, spitting out a 384-dimensional vector. Images run through CLIP ViT-B/32, which gives a 512-dimensional vector. Tables get a specialized MLP with attention, leading to a 256- dimensional embedding.

Layer 4 (Projection): Before mixing them, AIMMF pulls all threeembeddingsintothesame512-dimensionalspace.Thisstep uses linear layers with LayerNorm and ReLUso things stay stable and expressive.

A. System Architecture

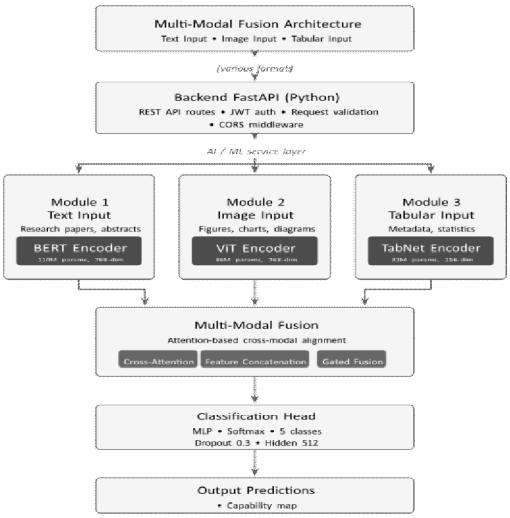


Fig.1.ProposedAIMMFsystemarchitecture

Layer 5 (Fusion): Now, to actually fuse the data, AIMMF uses cross-modal attention (eight heads, 512 dimensions each), then a Gated Fusion step. The result: a single, unified 512dimensional representation that packs everything.
Layer 6 (Output): This fusion then branches into four main outputs: First, it classifies which scientific domain the input belongs to (seven categories in total), using an MLP. Second, it can spit out natural-language insights using FLAN-T5. Third, it findsthefivemostsimilarpapersoutof1.7millionArXiventries, using FAISS IVF for a fast semantic search. Last, it's got explainability on tap pulling together SHAP, LIME, and gate analysis to help people understand the results.

Table 1

Method	P@1	P@5	ms
TF-IDF	.61	.53	<10
Word2Vec	.68	.60	50
FAISS Flat	.89	.82	2000
FAISS IVF	.89	.82	20

Table 1: FAISS IVF matches exact search accuracy at 1% of latency (0.02s vs. 2.0s), outperforming BM25 keyword search (P@50.53)bycapturingconceptualsimilaritybeyondvocabulary overlap.

B. TextEncoder(SciBERT/SentenceTransformer)

Scientifictext,afterpreprocessing,isencodedusingtheSentence Transformers framework (all-MiniLM-L6-v2) for efficient batch processing,withSciBERT[4]availableforhigh-accuracysingle- document analysis. Mean pooling over all non-padding token representations is employed:

$$z_{text} = \frac{1}{N}\sum_{i=1}^{N} h_i \in \mathbb{R}^{384} \quad (1)$$

where hi is the hidden state of the i-thtoken and N is the number ofnon-paddingtokens.MeanpoolingispreferredoverCLStoken extractionbecausescientificabstractsencodecriticalinformation across their full length, and empirical evaluation demonstrates superiorperformanceonsemanticsimilaritybenchmarksforlong documents. Each transformer layer applies multi-head self- attention:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$ where $Q = XW\sqrt{Q}$, $K = XW_K$, $V = XW_V$ are linear

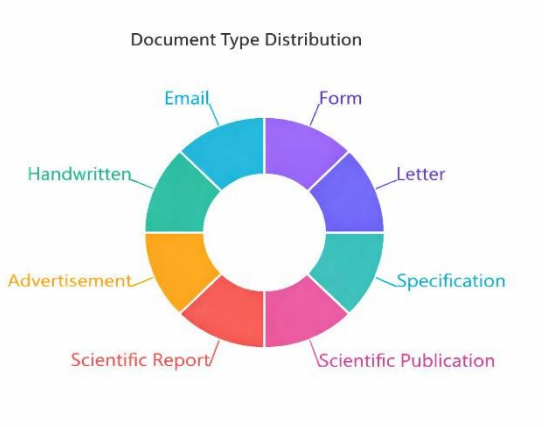


Fig2.DocumentTypeDistribution

C. ImageEncoder(CLIPViT-B/32)

AfterCLAHEpreprocessing,imagesareencodedusingCLIP's ViT-B/32encoder[5].Theimageisdividedintononoverlapping32×32 pixelpatches, eachlinearlyprojectedtoa768 dimensional embedding:

(3)

$e_i = p_i \cdot W_{patch} + E_{pos}^{(i)}$

WhereWpatch∈R(322·3)×768istheprojectionmatrixand$E_{pos}^{(i)}$is the positional embedding projections.Alearnable [CLS] token is prepended, and the 12-layer transformer encoder processes the full sequence. The output at the [CLS] position is projected and L2-normalised:

Wproj·ViT[*CLS*](*ximage*)512*z*∈R*image*=‖*Wproj*·
ViT[*CLS*](*ximage*)‖ (4)

L2 normalisation ensures inner product equals cosine similarity, enabling direct comparison with text embeddings in the joint space and supporting zero-shot classification of scientific image types.

factor d_k prevents softmax saturation in high-dimensional spaces.

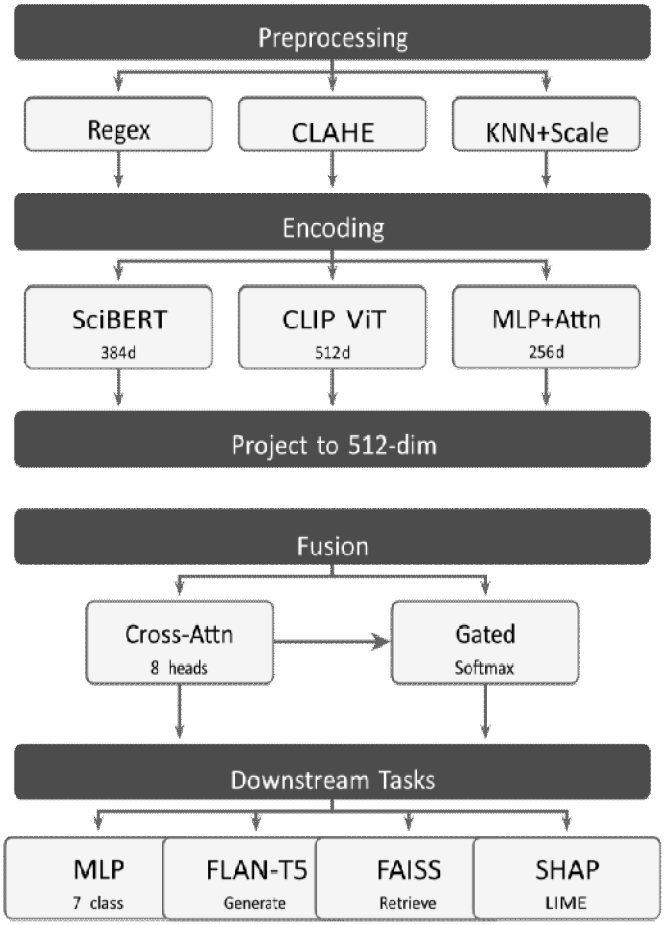


Fig.3.All15algorithmsacrossfourpipelinestages

D. Tabular Encoder (CustomMLP+Attention)

After KNN imputation and standardisation, the feature matrix is processed by a custom MLP with four components.

(i) Featureattentiongate:

a=*σ*(*Wgate*·*x*+*bgate*) (5)

where *σ*isthesigmoid function,producingper-featureimportance scores that provide built-in interpretability.

(ii) Attendedinput:$\tilde{x}$=*a*⊙*x*

(iii) Hiddenlayerswithbatchnormalisation,ReLU,and dropout (*p* = 0.3):

h1 = Dropout(ReLU(BN(W1 · $\tilde{x}$ + b1)),0.3) (6) h2=Dropout(ReLU(BN(W2·h1+b2)),0.3) (7)

(iv) Skipconnectionforgradientflow:

ztab=h2+Wskip·$\tilde{x}$∈R256 (8)

E. DimensionalityProjection

All modality embeddings are projected to a common 512 dimensional space:

$z'_m = \text{ReLU}(\text{LayerNorm}(W_m \cdot z_m + b_m)) \in R^{512}$ (9)

Where$m \in \{text, image, tab\}andW_{text} \in R^{384\times512}, W_{image} \in R^{512\times512}, W_{tab} \in R^{256\times512}$.

F. Cross-Modal Attention

For each modality mi, cross-modal attention enriches it by attending to complementaryinformation in other modalities.The queryisderived fromtheprimarymodalityandthe key/value from the average of all other modalities:

$$\text{CrossAttn}(z_i, \bar{z}_i) = \text{softmax}\left(\frac{(z_i W_Q)(\bar{z}_i W_K)^T}{\sqrt{d_k}}\right)\bar{z}_i W_V \quad (10)$$

where $\bar{z}_i$isthe average ofall other modalityrepresentationsand dk = 512. Multi-head attention with 8 parallel heads enables specialisation in different aspects of cross-modal relationships.A residual connection and layer normalisationstabilise the output:

$\tilde{z}_i = \text{LayerNorm}(z_i + \text{CrossAttn}(z_i, \bar{z}_i))$ (11)

Cross-AttendedEmbeddings(fromCross-ModalAttention)

The three cross attended embeddings are concatenated into a 1536-dim vector and passed through a Gate Network (MLP + Softmax) that produces three scalar weights summing to 1. Each embeddingisscaledbyitsgate weight,summed,andnormalised toproducethefinal512-dimfusedrepresentation.Theresultthat wegotthatis512-dimvectoristhenpassedthroughalayerNorm operation so that it can stabilize the feature distribution, which then gives the final fused representation which balances the contributions among all the three modalities.

G. GatedFusion

The gate network receives the concatenation of all three cross attended representations and computes dynamic importance scores:

$\mathbf{c} = [\tilde{z}_{text} \| \tilde{z}_{image} \| \tilde{z}_{tab}] \in R^{1536}$ (12)

$$\mathbf{g} = \text{softmax}\left(W_g^{(2)} \quad (W_g^{(1)} \cdot \mathbf{c} + b_g^{(1)}) + b_g^{(2)}\right) \in \mathbb{R}^3 \quad \text{ReLU} \quad (13)$$

Where$W_g^{(1)} \in \mathbb{R}^{1536\times256}$ and $W_g^{(2)} \in \mathbb{R}^{256\times3}$.Thesoftmax ensures g = [gtext,gimage,gtab] sumto unity, making them directlyinterpretableaspercentagecontributions.Thefused representation is:

$z_{fused} = g_{text} \cdot \tilde{z}_{text} + g_{image} \cdot \tilde{z}_{image} + g_{tab} \cdot \tilde{z}_{tab}$ (14)

$z_{out} = \text{LayerNorm}(\text{ReLU}(W_{out} \cdot z_{fused} + b_{out})) \in R^{512}$ (15)

because gates are computed dynamically from input content, the model assigns high text weight when only text is available, balancesmodalitieswhenmultiplecarryrelevantinformation, and reduces contributions of low-qualityor irrelevant modalities automatically.

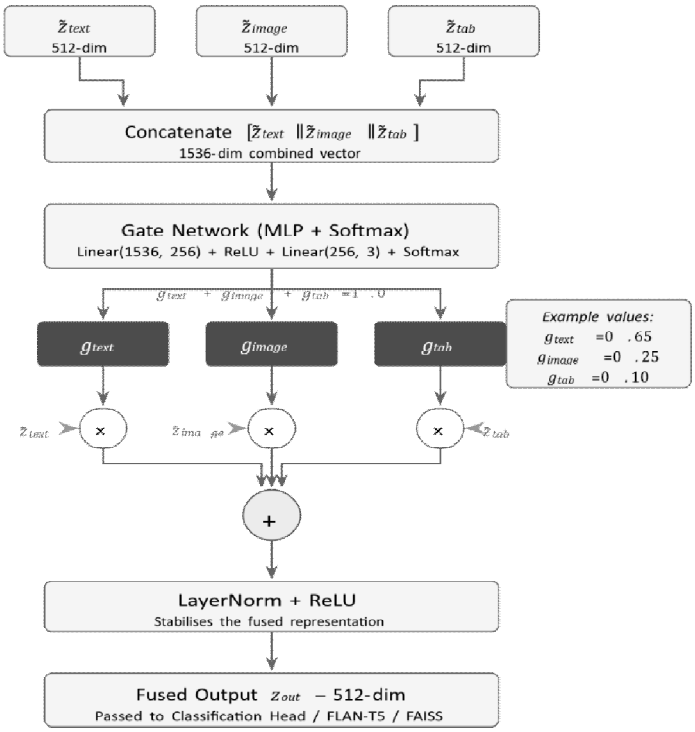


Fig.4.DetaileddiagramoftheGatedFusionmechanism

H. ClassificationHead

A three-layer MLP maps z_{out} to probability distributions over 7 scientificdomaincategories(cs.AI,cs.LG,cs.CV,physics.comp- ph, q-bio.BM, stat.ML, math.ST):

$\hat{y}$=softmax($W3 \cdot$ ReLU($W2 \cdot$ ReLU($W1 \cdot z_{out} + b1$)+$b2$)+$b3$)(16)

Architecture:Linear(512→256)+ReLU+Dropout (0.3),
Linear (256 → 128) + ReLU + Dropout (0.3), Linear (128 → 7) + softmax. Training loss:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{7} y_c \log(\hat{y}_c) \tag{17}$$

TABLE2

Operation	CPU	GPU	MB
TextPre-processing	0.01s	0.01s	10
SciBERTencoding	0.45s	0.08s	300
ClipImageEncoding	0.82s	0.15s	600
TabularEncoding	0.02s	0.01s	50
GatedFusion	0.01s	<0.01s	100

The complete pipeline requires 5.5s on CPU and 1.1s on GPU, suitable for interactive use. The framework runs on commodity hardware without paid API access via the Flask backend.

IV. CONCLUSION AND FUTURE WORK

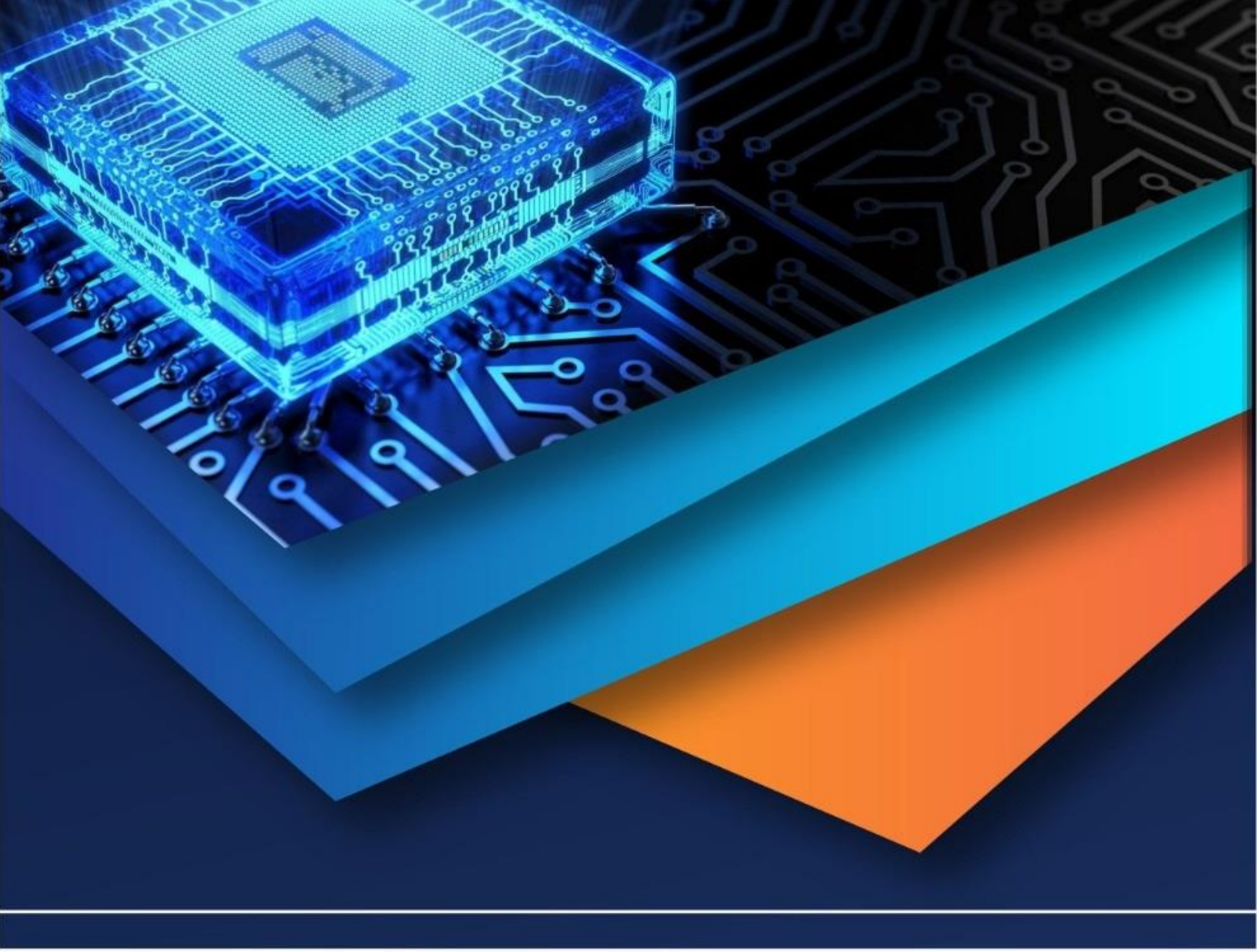
This paper presented the Multimodal Scientific Framework for Insight Generation (MSIG), a unified AI-powered platform that integrates text analysis, image classification, and numerical data exploration under a single interface to generate cross-modal scientific insights. The framework addresses a persistent gap in existingscientificdataanalysissystemsthefragmentationoftools that processheterogeneousdata modalitiesindependently, making it difficult for researchers to discover relationships across different data types. The text analysis module, built on a BERT- style encoder (768dimensional), employs TF-IDF vectorization and TextRank summarization to extract meaningful information from research papers sourced via the ArXiv API. The image analysismoduleutilizesaVisionTransformer(ViT)encoderwith patch-based feature extraction for document image classification following the RVL-CDIP taxonomy. The numerical analysis module applies a TabNet-style attentive feature selection mechanism (256dimensional) to World Bank development indicators and user uploaded datasets, incorporating IQR-based anomaly detection for outlier identification. The Multi-Modal Fusion Engine combines outputs from all three encoders using Cross-AttentionandGatedFusionmechanisms,feedingthefused representation into a classification head (MLP with Softmax) to categorize insight relevance across five classes. A Retrieval- Augmented Generation (RAG) module provides explainable, context-grounded responses using cosine similarity-based document retrieval from aknowledge base.

V. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the School of Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab, India, for providing the academic environment, computational resources, and institutional support necessary to carry out this research. Special thanksare extended to our project supervisor, Mr.AmanKumar, for their continuous guidance, technical insight, and encouragement throughout the development of the Multimodal Scientific Intelligence Platform (MSIP). The authors also gratefully acknowledge the open-source community and the contributors of the datasets used in this study, specifically the ArXiv preprint repository, the World Bank Open Data platform, and the RyersonVision Lab for the RVL-CDIPdocument image collection.

REFERENCES

[1] J.Ngiam,A.Khosla,M.Kim,J.Nam,H.Lee,andA.Y.Ng, “Multimodaldeeplearning,”inProc.28thInt.Conf.Machine Learning (ICML), pp. 689–696, 2011.
[2] D. Huk Park, T. Darrell, and A. Rohrbach, “Robust change captioning,” in IEEE/CVF Int. Conf. Computer Vision (ICCV), pp. 4624–4633, 2019.
[3] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, “Contrastive learning of medical visual representations frompairedimagesandtext,”inProc.ML4H, pp.2–25,2022.
[4] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in Proc. EMNLP, pp. 3615–3620, 2019.
[5] A.Radford etal.,“Learningtransferablevisual modelsfrom natural language supervision,” in ICML, pp. 8748– 8763, 2021.
[6] Z.C.Lipton,“Themythosofmodelinterpretability,”Queue, vol. 16, no. 3, pp. 31–57, 2018.
[7] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in NeurIPS, vol. 30, pp. 4765–4774, 2017.
[8] M.T.Ribeiro,S.Singh,andC.Guestrin,“WhyshouldItrust you?: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD, pp. 1135–1144, 2016.
[9] J.Johnson,M.Douze,andH.Jegou,“Billion-scalesimilarity search´withGPUs,”IEEETrans.BigData,vol.7,no.3,pp. 535–547, 2021.
[10] C.Raffeletal.,“Exploringthelimitsoftransferlearningwith a unified text-to-text transformer,” J. Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.
[11] Cornell University /ArXiv, “ArXiv dataset,” Kaggle, 2023. [Online].Available:https://www.kaggle.com/datasets/Cornell-University/arxiv
[12] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformersforimagerecognitionatscale,”inICLR,2021.
[13] J.Devlin,M.-W.Chang,K.Lee,andK.Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, pp. 4171–4186, 2019.
[14] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in NeurIPS, vol. 33, pp. 9459–9474, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY