Volume 11, Issue 1, January – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

A Hybrid Retrieval-Generative AI Framework for FinTech Document Handling and Compliance Tracking

Sudeshna Dey[1]; Soumitra De[2]; Jonti Deuri[3]; Siddhartha Chatterjee[4*]

[1]Department of Computer Science and Engineering, Regent Education and Research Foundation, Barrackpore, Telini Para, Kolkata - 700121, West Bengal, India

[2]Department of Computer Science and Engineering, College of Engineering and Management Kolaghat, Purba Medinipur - 721171, West Bengal, India
[3]Faculty of Engineering and Technology, Sharda University 73, Andijan, Boborshah Prospekt, Uzbekistan
[4]Department of Computer Science and Engineering, College of Engineering and Management Kolaghat, Purba Medinipur – 721171, West Bengal, India

Corresponding Author: Siddhartha Chatterjee[4*]

Publication Date: 2026/02/04

Abstract: Large language models (LLMs) have been quickly adopted in the financial services industry, allowing for sophisticated automation in document analysis, client service, compliance monitoring, and decision support. However, hallucinations, explainability issues, privacy concerns, and restricted access to valuable institutional information limit their use in regulated financial situations. Financial organizations need systems that ensure factual accuracy, traceability, and regulatory compliance in addition to producing fluid replies. This chapter describes an expert system for FinTech document intelligence based on Retrieval-Augmented Generation (RAG) that combines conventional term-based search with meaningful vector retrieval to guarantee dependable and auditable autonomous reasoning. The entire architecture, document ingesting pipeline, regulated prompt development, hybrid retrieval mechanism, variable structuring approaches, embedding generation, encryption model, and assessment methodology are all described. A thorough discussion is given of practical applications in financial services, identifying fraud, credit risk assessment, and adherence to regulations. A strategy plan for future study and establishment of policies is also presented, along with ethical issues and regulatory harmonization. Financial specialists can examine, confirm, and override artificially generated insights when needed thanks to the system's provision for human during validation. The platform facilitates regulatory examinations and improves transparency by keeping thorough derivation information and audit recordings for each generated answer. This method bridges the gap between the strict governance necessities of real-world economic ecosystems and advanced generative intelligence.

Keywords: *Generative AI, FinTech, Retrieval-Augmented Generation, Document Intelligence, Regulatory Compliance, Vector Databases, Hybrid Retrieval, Explainable AI, Auditability.*

How to Cite: Sudeshna Dey; Soumitra De; Jonti Deuri; Siddhartha Chatterjee (2026) A Hybrid Retrieval-Generative AI Framework for FinTech Document Handling and Compliance Tracking. *International Journal of Innovative Science and Research Technology*, 11(1), 2779-2790. https://doi.org/10.38124/ijisrt/26jan1585

I. INTRODUCTION

The financial technology (FinTech) industry works in a highly controlled and sensitive data environment. Regulator circulars, corporate compliance regulations, customer transactions, transaction statements, audit documentation, evaluation of risk documents, and contractually materials are just a few of the many digital documents that financial institutions oversee [1,2]. In areas including credit authorization, the identification of fraud, regulatory submission, customer orientation, and managing a portfolio, these documents serve as the basis for operational decision-making. Recent advances in artificial intelligence, particularly large language models (LLMs), have demonstrated strong capabilities in natural language understanding, summarization, question answering, and conversational interaction. These models offer significant potential for automating document-centric tasks in financial institutions, including drafting compliance reports, assisting auditors, answering policy-related queries, and supporting customer

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

service operations [3]. Consequently, many banks and FinTech organizations are actively exploring generative AI as a means to improve efficiency and reduce operational costs.

However, there are significant obstacles to the direct application of LLMs in financial systems. Since these algorithms are trained on freely available data, they are unable to access confidential contracts, client transaction histories, internal credit policies, or regulatory interpretations unique to a particular organization [2]. Furthermore, LLMs may produce answers that seem linguistically accurate but are unverifiable or factually incorrect. Such mistakes can result in inaccurate compliance advice, faulty risk assessment, monetary losses, and harm to one's reputation or legal standing in regulated economic environments [4].

A potential remedy for these constraints is Retrieval-Augmented Generation (RAG). RAG systems retrieve pertinent document sections from trusted resources at demand time and integrate the resulting data into the process of creation instead of depending exclusively on knowledge stored in model parameters [5]. By reducing hallucinations, enabling explicit source attribution, and grounding replies in verified sources, this architecture aligns generative AI with financial organizations demands for responsibility and openness.

This paper presents a retrieval-augmented generative AI expert system specifically designed for FinTech document intelligence and regulatory compliance [2]. The proposed framework integrates document ingestion, adaptive chunking, semantic embeddings, hybrid retrieval techniques, and controlled prompt engineering to deliver accurate, explainable, and auditable responses to financial and regulatory queries. The objective is to enable the safe and scalable adoption of generative AI in financial environments while satisfying technical, legal, and governance constraints [6].

II. BACKGROUND AND RELATED CONCEPTS

This section outlines the key technological and conceptual foundations that motivate the proposed retrieval-augmented generative AI system for FinTech document intelligence [1]. It reviews document intelligence, generative AI in finance, associated compliance risks, information retrieval techniques, and retrieval-augmented generation.

➢ *Document Intelligence in FinTech*

Document intelligence refers to the automated processing, understanding, indexing, and retrieval of information from unstructured or semi-structured documents. In financial institutions, such documents include regulatory circulars, compliance manuals, contracts, customer disclosures, transaction summaries, audit reports, and risk assessment documents [7].

These documents form the backbone of many operational workflows, including customer onboarding, credit evaluation, fraud investigation, regulatory reporting, and internal audits. Unlike conventional enterprise search systems, FinTech document intelligence platforms must satisfy additional requirements such as high accuracy, traceability of information sources, and legal defensibility[8].Regulators and auditors often require institutions to demonstrate which documents were consulted when making specific decisions[8]. As a result, modern FinTech systems increasingly rely on AI-driven document intelligence to improve efficiency while maintaining compliance standards[9].

➢ *Financial Systems and Generational Artificial Intelligence*

Significant advancements in automated text generation and comprehension have been made possible by generative artificial intelligence, especially big language models. These algorithms may extract important terms from contracts, summaries lengthy reports, provide organized explanations, and respond to complicated queries [12].

Financial organizations have started investigating generative AI for jobs including creating compliance documents, helping analysts interpret policies, automating customer service, and assisting with internal information management. These features promise quicker accessibility to data and less manual labour [13]. However, generative models are trained on general-purpose datasets and do not inherently possess knowledge of proprietary institutional documents or organization-specific regulatory interpretations [14]. This limitation reduces their reliability in regulated environments.

The use of standalone generative models in financial compliance introduces several critical risks that must be mitigated:

- Hallucination: Generative models are probabilistic engines that prioritize linguistic fluency over factual accuracy [10]. In a financial context, this often leads to the fabrication of plausible-sounding but non-existent regulatory clauses, policy mandates, or case precedents. Such errors can result in severe compliance violations if decision-makers act upon these "hallucinated" directives without verification [3].
- Lack of source attribution: Standard Large Language Models (LLMs) generate answers based on internal parametric knowledge rather than referencing specific documents. This creates a "black box" effect where it is impossible to trace a generated claim back to an authoritative source, making it extremely difficult for auditors to verify the accuracy of the advice or demonstrate regulatory diligence [5].
- Outdated knowledge: Financial regulations and internal corporate policies are dynamic, with frequent updates, circulars, and amendments. Since standalone models are limited to the data available at their training cutoff date, they cannot account for real-time regulatory changes, leading to advice that may be legally obsolete or non-compliant with current standards [11].
- Privacy concerns: Using public or third-party generative models raises significant data governance issues. There is a risk that sensitive financial data, personally identifiable information (PII), or proprietary trading strategies included in a prompt could be exposed, logged by the model

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

provider, or inadvertently memorized and regurgitated in future model iterations [15].

➢ *Retrieval-Augmented Generation*

Retrieval-Augmented Generation integrates information retrieval with generative language models. Instead of relying solely on knowledge encoded in model parameters, RAG systems retrieve relevant document segments from trusted repositories and provide them as context during response generation [16].

This approach ensures that answers are grounded in verified documents, reduces hallucination, and enables explicit citation of sources. It also allows systems to incorporate newly issued regulations or updated internal policies without retraining the language model. Financial documents contain formal legal language, numerical constraints, structured clauses, and domain-specific terminology [18]. Relying on a single retrieval method is often insufficient. Hybrid retrieval combines semantic similarity search with keyword-based ranking to handle both conceptual queries and exact regulatory references [19].

- *Hybrid Retrieval in Financial Contexts:* Financial documents contain formal legal language, numerical constraints, structured clauses, and domain-specific terminology. Relying on a single retrieval method is often insufficient. Hybrid retrieval combines semantic similarity search with keyword-based ranking to handle both conceptual queries and exact regulatory references [20].

This strategy improves retrieval reliability and is particularly effective for compliance analysis, contract interpretation, and regulatory question answering. The proposed expert system adopts document intelligence pipelines, hybrid retrieval mechanisms, and retrieval-augmented generation to address the limitations of standalone generative models. By grounding responses in authoritative documents and enabling source attribution, the system provides a practical foundation for accurate, explainable, and regulation-compliant document intelligence in FinTech environments[22].

- *System Requirements in FinTech:* A FinTech document intelligence system based on retrieval-augmented generation must satisfy a wide range of technical, operational, and regulatory constraints. Unlike general conversational AI systems, such platforms operate in environments where errors can result in regulatory violations, financial losses, and legal consequences. This section outlines the key functional, non-functional, privacy, and governance requirements considered in the design of the proposed system.

➢ *Functional Requirements*

The system must support the end-to-end processing and intelligent querying of financial documents across multiple operational domains.

First, it should enable multi-format document ingestion, supporting commonly used financial document types such as PDF (digital and scanned), DOCX, plain text files, CSV datasets, and HTML-based regulatory publications. This ensures compatibility with regulatory circulars, internal policy manuals, contracts, audit reports, and transaction summaries[13].

Second, the system must perform reliable text extraction and normalization. Financial documents often contain complex layouts including tables, multi-column structures, headers, and footnotes. Accurate extraction, logical ordering of text, encoding normalization, and graceful handling of corrupted or low-quality files are essential to ensure high-quality indexing and retrieval.

Third, the system should implement adaptive document segmentation. Documents must be divided into meaningful chunks that preserve legal clause boundaries and contextual coherence while remaining compatible with language model context limitations. Different chunking strategies may be required for regulatory texts, contracts, and operational manuals[18].

Fourth, the platform must provide hybrid retrieval capability, combining semantic vector-based search with keyword-based ranking techniques. This enables both conceptual understanding of user queries and precise matching of regulation numbers, legal terms, and financial identifiers [23].

Finally, the system must generate evidence-grounded responses with explicit source attribution and traceability. Each answer should reference the documents used, including document identifiers and section-level information where available. In addition, comprehensive logging and audit trails must be maintained, capturing user queries, retrieved content, generated outputs, system configurations, and timestamps to support regulatory audits and internal governance[15].

➢ *Non-Functional Requirements*

TBeyond core functionality, the system must meet several operational quality standards.

The platform should deliver high accuracy and consistency, minimizing hallucinations and ensuring stable behavior across repeated queries and document updates. Output reliability is essential for compliance-critical applications.

It must also support scalability and performance, enabling fast responses even when operating over large document repositories containing millions of text segments. Horizontal scaling and efficient indexing strategies should be supported to accommodate organizational growth[18].

Availability and fault tolerance are equally important. The system should tolerate partial failures of individual components, support graceful degradation, and provide recovery mechanisms to avoid operational disruptions.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

It should allow easy replacement of embedding models, migration between vector databases, modification of prompt templates, and integration of new regulatory logic without requiring major architectural redesign[21].

Finally,The system should integrate seamlessly with existing document management systems, compliance platforms, risk management tools, and identity and access management infrastructure.

FinTech systems handle sensitive financial and personal data and must therefore adhere to strict security and regulatory standards.

Dataset governance, which includes monitoring document provenance, change history, indexing metadata, and deletion procedures, must also be supported by the platform [24][26].

The system should guarantee reproducibility so that previous outputs can be recreated using configured snapshots and stored documentation to aid in investigations and audits.

Lastly, the system should provide for human monitoring, including routines for manual reviews, feedback systems, and escalation protocols for questions that are unclear or high-risk[34].

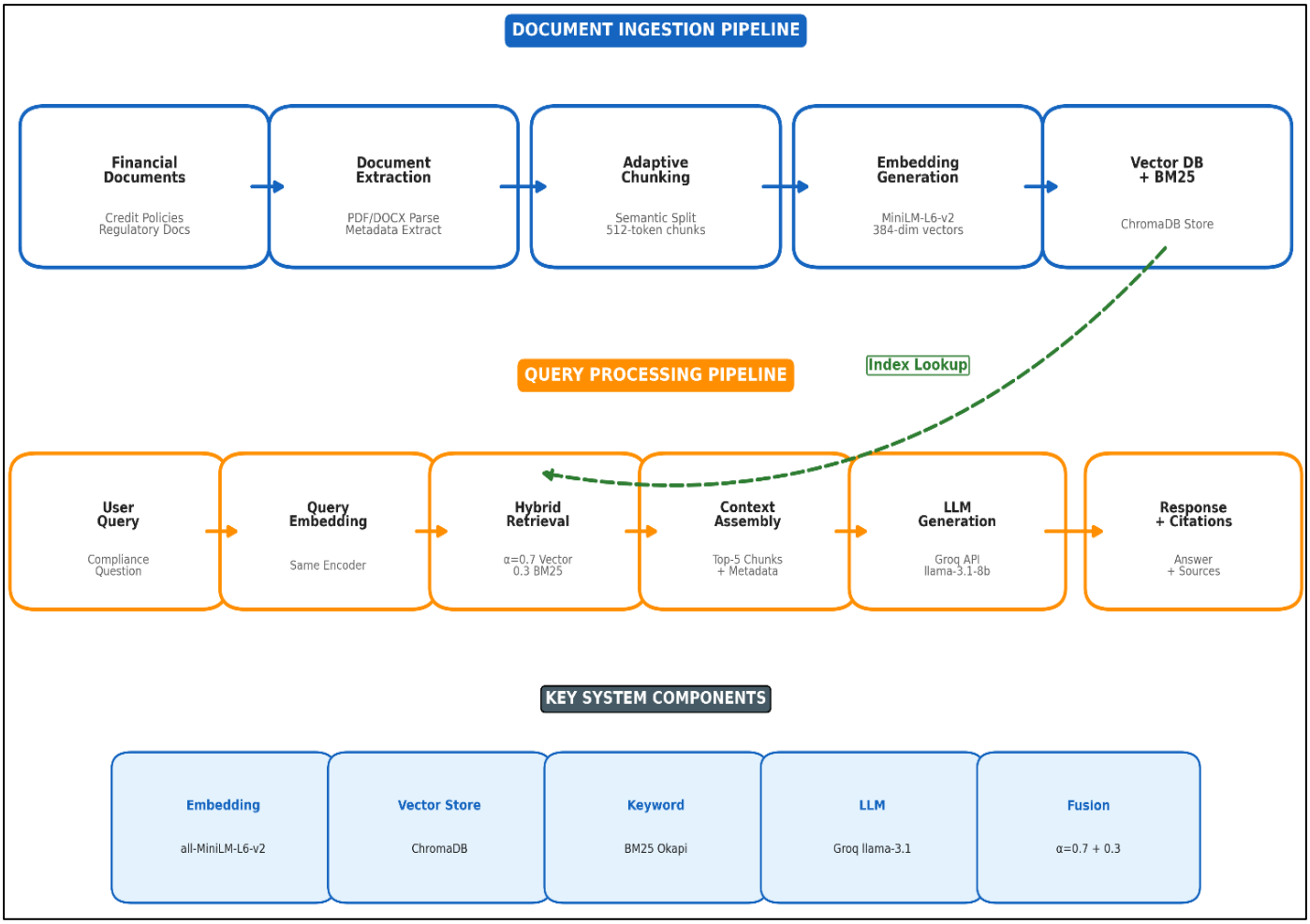

Fig 1 High-Level Architecture of the FinTech RAG Expert System.

The diagram illustrates the dual-pipeline approach: the Document Ingestion Pipeline (top) handles extraction, adaptive chunking, and vector indexing, while the Query Processing Pipeline (bottom) manages user queries through hybrid retrieval and context-aware LLM generation.

III. ARCHITECTURAL STRUCTURE

➢ *The System Consists of the Following Core Layers:*

- Data Layer: The safe and scalable preservation of every system's asset is the responsibility of this basic layer. It handles a diverse range of data types, such as extracted text segments, normalized metadata, and raw financial records (PDFs, DOCX). Importantly, it houses the inverted indices needed for keyword search as well as the vector database (like ChromaDB) for high-dimensional embeddings. This layer employs encryption-at-rest for all stored assets to meet financial data governance standards and upholds stringent data residency rules to guarantee adherence to regional banking laws.
- Processing Layer: The processing layer functions as the transformation engine of the architecture. It manages the entire ingestion pipeline, starting with the use of OCR (Optical Character Recognition) for scanned files and the parsing of intricate document layouts. The crucial

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

"adaptive chunking" procedure, which divides documents into logically coherent parts based on conceptual boundaries rather than mechanical word counts, is carried out by this layer. After segmentation, it coordinates the creation of embeddings, transforming text segments into retrieval-optimized vector representations [26].

- Intelligence Layer: This layer serves as the system's cognitive center, bridging the gap among generative thinking and data retrieval. In order to maximize pertinent context, it performs the hybrid retrieval logic by dynamically combining the outcomes of keyword-based enquiries and semantic vector searches. Additionally, it oversees the "centralized prompt construction," putting together the user inquiry, evidence that has been retrieved, and stringent system commands into a logical prompt payload. Additionally, it controls how the Large Language Model (LLM) interacts with it, imposing limitations to avoid hallucinations and making sure the model stays rooted in the given context [27][39].
- Application Layer: The application layer provides the interface through which end-users—such as compliance officers, risk analysts, and auditors—interact with the system. It abstracts the underlying technical complexity, offering intuitive web interfaces and API endpoints for domain selection and query submission. A key feature of this layer is response visualization, which presents the generated answer alongside interactive citations. This allows users to hover over or click on specific claims to instantly view the source document and page number, facilitating immediate verification [28].
- Governance Layer: The governance layer operates as the oversight mechanism, ensuring that the system adheres to strict security and regulatory protocols. It enforces Role-Based Access Control (RBAC) to ensure that users can only retrieve documents they are authorized to view. Additionally, it maintains comprehensive, immutable audit logs that record every interaction, including the query, the specific document chunks retrieved, and the generated response. This layer is essential for post-hoc auditing and ensures the system's operations are transparent, traceable, and legally defensible.

This layered design enables independent scaling, maintenance, and upgrading of individual components while preserving end-to-end system integrity.

IV. DOCUMENT INGESTION AND PREPROCESSING

Document ingestion and preprocessing form the foundation of the proposed retrieval-augmented generative AI system. Since retrieval accuracy and response reliability depend directly on the quality of indexed data, this stage is designed to ensure robustness, consistency, and compliance with financial data governance standards [32].

➢ *Document Acquisition and Format Handling*

The ingestion process begins with the secure acquisition of documents from internal repositories, compliance management systems, and enterprise document management platforms. The system supports commonly used financial document formats, including digitally generated and scanned PDFs, DOCX files, plain text documents, CSV datasets, and HTML-based regulatory publications.

This multi-format capability ensures seamless integration with regulatory circulars, internal policy manuals, contracts, audit reports, transaction summaries, and risk assessment documentation that are routinely used in FinTech environments.

➢ *Text Extraction and Content Normalization*

Once documents are uploaded, automated text extraction is performed using a combination of rule-based parsers and optical character recognition (OCR) engines for scanned content. Financial documents frequently contain complex structures such as tables, multi-column layouts, headers, footnotes, and embedded annotations [30].

To address this, the extraction pipeline includes:

- Layout normalization to preserve logical reading order
- Removal of redundant formatting artifacts
- Encoding correction and character normalization
- Detection of missing or corrupted text segments

Documents that fail predefined quality thresholds are flagged for manual review to prevent unreliable data from entering the knowledge base [31].

Following extraction, normalization operations are applied, including whitespace correction, sentence boundary detection, removal of boilerplate headers and footers, and standardization of numerical and currency formats. These steps improve semantic consistency and ensure stable embedding generation during later stages [37].

➢ *Metadata Enrichment and Governance Support*

After normalization, metadata is generated and attached to each document and its derived text segments. Typical metadata fields include:

- Document source and department
- Creation or publication date
- Regulatory authority or jurisdiction
- Document category (e.g., compliance policy, contract, audit report)
- Confidentiality level and access classification

This metadata plays a critical role in domain-aware retrieval, access control enforcement, regulatory auditability, and lifecycle management of indexed documents.

➢ *Chunk Preparation and Secure Indexing Readiness*

Before indexing, the cleaned text is passed to the adaptive chunking module, which segments the content into logically coherent units suitable for embedding and retrieval. Each chunk inherits the metadata of its parent document and is assigned a unique identifier to support precise source attribution during response generation[28].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

Finally, all processed text segments and associated metadata are encrypted and securely stored, preparing them for downstream embedding generation and indexing within the vector database and keyword retrieval system.

This structured ingestion and preprocessing pipeline ensures that the knowledge base remains accurate, searchable, secure, and continuously updatable as new financial documents and regulatory updates become available.

V. CHUNKING STRATEGYETHICAL ISSUES AND RESPONSIBLE AI IN FINTECH

Document chunking is a critical step in transforming large financial documents into manageable units suitable for semantic indexing and retrieval. Since modern language models operate under fixed context window limitations, long documents must be segmented while preserving semantic integrity and legal coherence. An effective chunking strategy directly influences retrieval precision, response accuracy, and citation reliability in compliance-oriented applications [19].

➢ *Rationale for Chunking in Financial Documents*

Financial and regulatory documents often contain lengthy sections, nested clauses, tables, and cross-references. Indexing entire documents as single units leads to:

- Reduced retrieval precision
- Context dilution
- Inefficient embedding representations
- Increased hallucination risk during generation

Therefore, documents are divided into smaller, logically consistent segments that balance contextual completeness with computational efficiency.

➢ *Chunking Approaches Adopted*

The system employs a hybrid chunking strategy that adapts to the structure and content type of each document:

Fixed-length chunking is used for unstructured operational documents where consistent segmentation is sufficient. This approach enforces uniform chunk sizes with overlapping boundaries to maintain continuity.

Semantic chunking groups sentences based on topical similarity using linguistic cues and embedding-based similarity analysis. This method preserves conceptual coherence and is especially effective for regulatory texts and policy documents.

Question–answer-oriented chunking is applied to structured compliance manuals and regulatory FAQs, where clauses naturally follow a question–answer or rule–exception format.

➢ *Metadata-Aware Chunk Construction*

Each generated chunk is enriched with metadata inherited from the source document, including document identifiers, regulatory category, jurisdiction, and confidentiality level. This enables:

- Domain-specific retrieval
- Fine-grained access control
- Accurate source attribution
- Regulatory audit traceability

Unique chunk identifiers are also assigned to support citation mapping during answer generation.

➢ *Impact on Retrieval and Compliance Reliability*

Semantic chunking significantly improves retrieval relevance for legal and compliance documents where clause boundaries and regulatory context are critical. Smaller, coherent chunks reduce ambiguity, improve vector similarity matching, and enable precise citation of regulatory provisions [38].

By combining structural awareness with semantic segmentation, the system ensures that retrieved evidence is both contextually meaningful and legally defensible, supporting high-confidence deployment in financial compliance environments.

VI. HYBRID RETRIEVAL MECHANISM & SEMANTIC RETRIEVAL COMPONENT

Accurate information retrieval is central to the reliability of retrieval-augmented generative systems, particularly in regulated financial environments where incomplete or incorrect evidence can lead to compliance failures. To address the limitations of relying on a single retrieval strategy, the proposed system employs a hybrid retrieval mechanism that combines semantic vector-based search with traditional keyword-based ranking

Financial documents exhibit unique characteristics such as formal legal language, structured regulatory clauses, numerical constraints, domain-specific terminology, and jurisdiction-dependent references. Queries submitted by users may range from conceptual questions (e.g., compliance obligations under a regulation) to highly specific requests (e.g., clause numbers, reporting thresholds, or transaction identifiers) [43].

The semantic component uses vector similarity search over document embeddings generated during the indexing phase. This enables the system to identify passages that are conceptually similar to the user query even when exact keywords do not match.

➢ *Keyword-Based Retrieval Component*

In parallel, the system applies keyword-based retrieval using ranking algorithms such as BM25. This component is optimized for:

- Regulation identifiers and clause numbers
- Legal and technical terminology

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

- Transaction references
- Numerical thresholds and reporting codes

Keyword-based retrieval ensures high precision for exact matches and structured queries, which are common in compliance and audit workflows.

➢ *Result Fusion and Ranking Strategy*

The outputs of the semantic and keyword retrieval components are combined using a weighted scoring function:

$$\text{Score} = \alpha \times \text{VectorScore} + (1 - \alpha) \times \text{BM25Score}$$

Where α controls the balance between semantic relevance and lexical precision. In the current implementation, $\alpha = 0.7$ was found to provide an effective trade-off [33].

Additional ranking heuristics are applied to further refine results, including:

- Boosting recent or updated regulatory documents
- Filtering by document category and jurisdiction

The fusion process is illustrated in Figure 2

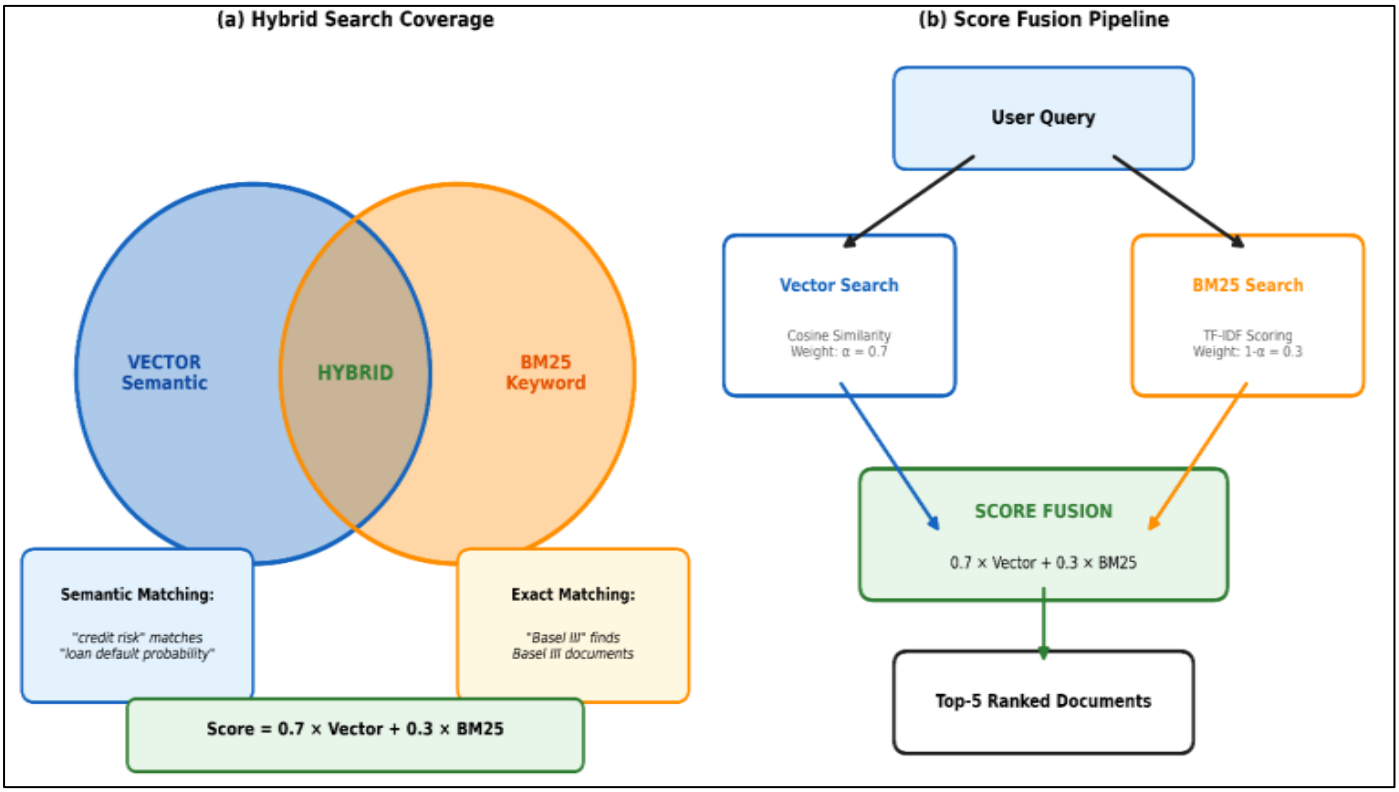

Fig 2 Hybrid Retrieval and Score Fusion Mechanism.

VII. EVALUATION METHODOLOGY

Evaluating a retrieval-augmented generative AI system for FinTech applications requires assessing both the quality of information retrieval and the reliability of end-to-end response generation. Since the system is intended for compliance-critical environments, evaluation focuses not only on technical accuracy but also on transparency, citation correctness, and operational efficiency [40][41].

➢ *Retrieval Performance Evaluation*

The first stage of evaluation measures how effectively the system retrieves relevant document segments for a given query. Standard information retrieval metrics are employed, including Recall@k, Precision@k, and Mean Reciprocal Rank (MRR). These metrics quantify the proportion of relevant regulatory or policy clauses retrieved within the top-k results and the ranking quality of the retrieval engine [42].

Latency is also measured at multiple percentiles to ensure that retrieval remains efficient even as document repositories grow. Low and predictable response times are essential for real-time compliance analysis and customer support scenarios.

➢ *End-to-End System Evaluation*

Beyond retrieval, the complete pipeline—from query submission to final answer generation—is evaluated to assess practical usability and regulatory suitability. Key evaluation dimensions include:

- Citation accuracy is evaluated to ensure that all factual statements within a response are explicitly supported by correct document references. This verification eliminates "black box" reasoning, allowing auditors to trace claims back to authoritative sources and guaranteeing the system does not fabricate information [45].
- This check ensures the language model acts strictly as a reasoning engine grounded in provided institutional data,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

rather than relying on potentially outdated or irrelevant internal training data.

➢ *Compliance-Oriented Validation*

To assess regulatory readiness, domain experts evaluate system outputs against real compliance scenarios such as regulatory interpretation queries, internal audit checks, and policy clarification tasks. Responses are examined for traceability, completeness, and alignment with official documentation [44].

This qualitative evaluation complements quantitative metrics by validating that the system behaves predictably under realistic operational conditions.

➢ *Experimental Results Overview*

Experimental results demonstrate that hybrid retrieval significantly improves recall and citation reliability compared to standalone semantic or keyword-based methods. The system maintains stable performance across increasing document volumes while preserving low latency and high factual accuracy.

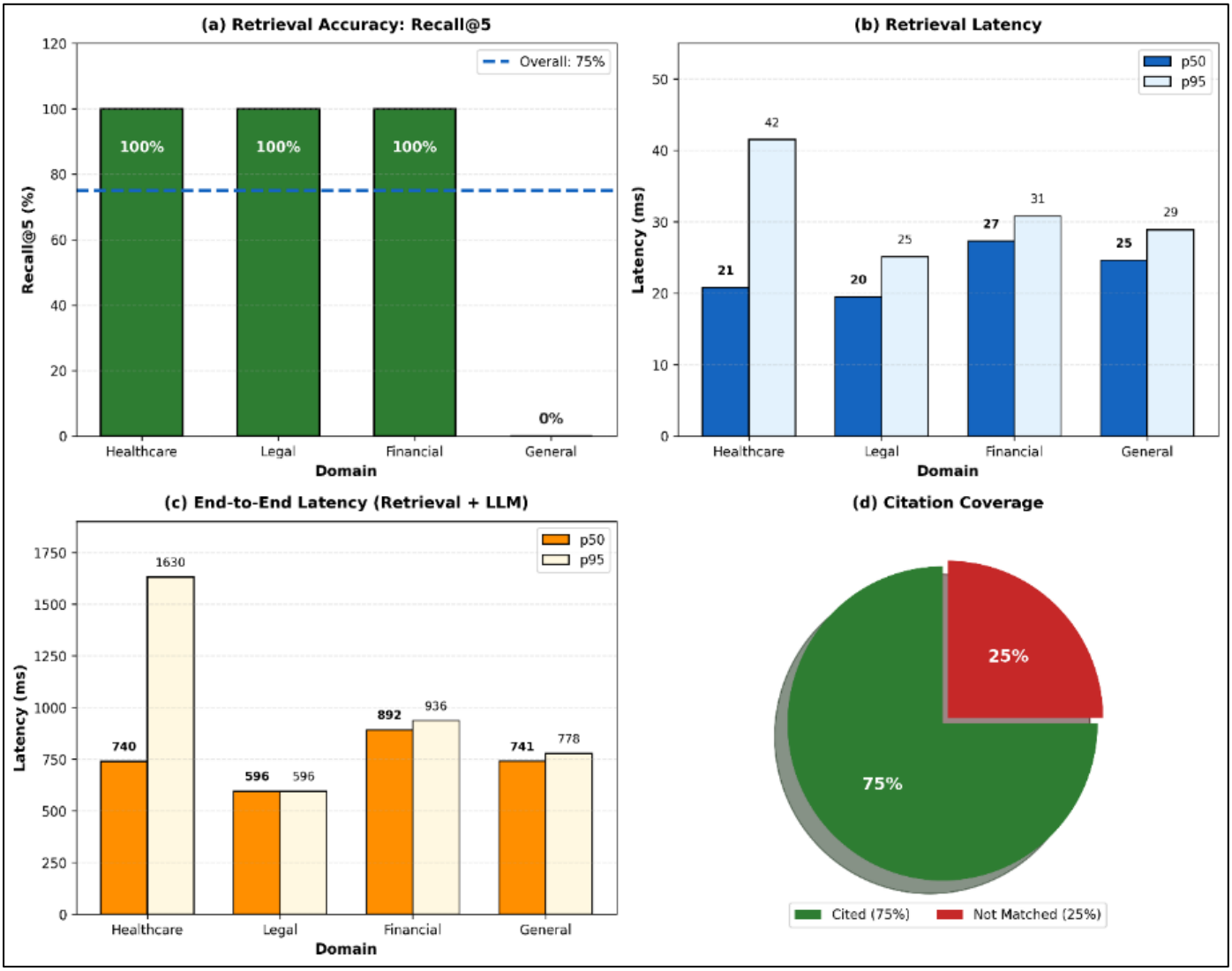

Fig 3 Quantitative Performance Evaluation. (a) Retrieval Recall@5 Across Different Domains (Legal, Financial, General). (b) System Latency at p50 and p95 Percentiles. (c) End-to-End Latency Including Generation. (d) Citation Coverage Distribution, Highlighting the System's High Reliability in Source Attribution.

VIII. ETHICAL ISSUES AND RESPONSIBLE AI IN FINTECH

The deployment of generative AI systems in financial services raises important ethical considerations related to transparency, fairness, accountability, and user trust. In regulated environments, ethical compliance is not optional but a fundamental requirement for system acceptance and long-term sustainability.

A primary concern is explainability. Financial decisions and compliance interpretations must be justifiable to regulators, auditors, and affected stakeholders. The proposed retrieval-augmented framework addresses this requirement by grounding all responses in verifiable documents and providing explicit source citations. This enables human reviewers to trace system outputs to authoritative evidence and assess their validity [46].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

Another key issue is bias and fairness. Training data and document repositories may reflect historical or institutional biases that could influence automated reasoning. To mitigate this risk, the system relies on curated and institution-approved document sources, continuous monitoring of response patterns, and periodic audits of retrieval and generation behavior. Incorporating diverse regulatory sources and regularly updating policy documents further reduces the likelihood of systematic bias.

Privacy and data protection also constitute major ethical obligations. Financial documents often contain sensitive personal and corporate information. The system enforces strict access controls, encryption mechanisms, and data minimization policies to ensure that sensitive content is not exposed beyond its intended scope.

IX. STRATEGIC ROADMAP FOR FUTURE RESEARCH AND POLICY

While retrieval-augmented generative systems represent a significant advancement for FinTech document intelligence, several research and policy directions remain open for further development.

From a technical perspective, future work may focus on the creation of standardized evaluation benchmarks for RAG systems in financial and regulatory domains. Such benchmarks would enable objective comparison of retrieval accuracy, citation reliability, and compliance alignment across different architectures.

The integration of financial knowledge graphs with retrieval-based pipelines presents another promising direction. Knowledge graphs can encode structured relationships between entities such as institutions, regulations, financial instruments, and risk categories, further enhancing reasoning capabilities and cross-document inference [36].

On the governance side, the development of AI certification and compliance frameworks specific to financial services would help standardize best practices for deployment, auditing, and lifecycle management of generative systems. Regulatory authorities may also introduce formal guidelines for explainability thresholds, data usage boundaries, and accountability mechanisms in AI-assisted financial decision-making.

Cross-border financial operations highlight the need for regulatory harmonization, as institutions increasingly operate across jurisdictions with differing legal requirements. Retrieval-augmented systems can support this process by enabling comparative analysis of international regulatory documents and facilitating consistent policy interpretation [34].

Finally, the adoption of real-time regulatory monitoring—where new policy updates are automatically ingested, indexed, and incorporated into the knowledge base—represents a key step toward fully adaptive compliance systems capable of responding immediately to evolving legal environments.

X. PERFORMANCE AND SCALABILITY ANALYSIS

The performance and scalability of a retrieval-augmented generative AI system are critical for its adoption in real-world FinTech environments, where document volumes are large and query workloads can be highly variable.

From a retrieval perspective, the use of vector databases enables efficient approximate nearest-neighbor search even when the knowledge base contains millions of document chunks. Empirical benchmarks indicate that query latency increases sub-linearly with data volume when appropriate indexing structures and sharding strategies are employed. Hybrid retrieval introduces additional computational overhead due to result fusion and ranking; however, this overhead remains acceptable for interactive use when supported by parallel processing and caching mechanisms.

The document ingestion and embedding generation pipeline is designed to scale horizontally. Batch embedding, asynchronous processing, and distributed task scheduling allow the system to accommodate continuous inflow of regulatory updates and internal documents without disrupting query-time performance. Incremental indexing ensures that new content becomes searchable within minutes rather than hours[37].

At the generation stage, response time is primarily influenced by the size of the retrieved context and the underlying language model architecture. By limiting prompt size through relevance filtering and chunk prioritization, the system maintains predictable latency while preserving citation completeness[33].

Stress testing under simulated enterprise workloads demonstrates that the architecture supports concurrent queries from hundreds of users with stable performance when deployed on moderate cloud or on-premise clusters. These results indicate that retrieval-augmented architectures can scale effectively to meet the operational demands of large financial institutions while maintaining compliance-oriented reliability[31].

XI. CONCLUSION

A full retrieval-augmented generative AI expert system for FinTech document intelligence and regulatory compliance was presented in this chapter. In order to create a cohesive architecture appropriate for regulated financial contexts, the suggested system combines secure content ingestion, dynamic chunking, semantic integration, hybrid extraction, controlled prompt creation, and scientific response generation. The paper illustrated how retrieval augmentation tackles the main drawbacks of standalone language models, such as hallucination, lack of explainability, and out-of-date knowledge, through thorough architectural analysis,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

evaluation methodology, application scenarios, and a real-world case study.

REFERENCES

[1]. T. Brown et al., "Language models are few-shot learners," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 1877–1901, Dec. 2020.

[2]. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 9459–9474, Dec. 2020.

[3]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[4]. C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[5]. S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, Apr. 2009. doi: 10.1561/1500000019.

[6]. A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, Long Beach, CA, USA, Dec. 2017.

[7]. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Hong Kong, China, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[8]. European Commission, "Ethics guidelines for trustworthy AI," High-Level Expert Group on Artificial Intelligence, Brussels, Belgium, Rep., Apr. 2019.

[9]. Basel Committee on Banking Supervision, "Principles for financial market infrastructures," Bank for International Settlements, Basel, Switzerland, Rep., Apr. 2012.

[10]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[11]. Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019. doi: 10.48550/arXiv.1907.11692.

[12]. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. (draft). Upper Saddle River, NJ, USA: Pearson, 2020.

[13]. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, Tech. Rep., 2018.

[14]. S. Arora et al., "Theoretical analysis of retrieval-augmented generation models," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[15]. R. Guo et al., "Accelerating large-scale inference with anisotropic vector quantization," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 3887–3896.

[16]. Pinecone Systems, "Vector databases for production AI," 2022.

[17]. Weaviate Community, "Hybrid search in vector databases," 2022.

[18]. Information technology — Security techniques — Information security management systems — Requirements, ISO/IEC 27001:2013, Int. Org. Standardization, Geneva, Switzerland, 2013.

[19]. M. Kearns and A. Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design. New York, NY, USA: Oxford Univ. Press, 2019.

[20]. Reserve Bank of India, "Guidelines on digital lending," RBI/2022-23/111, Sep. 2022.

[21]. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), European Union, Brussels, Belgium, 2016.

[22]. PCI Security Standards Council, "Payment Card Industry (PCI) Data Security Standard: Requirements and testing procedures," ver. 4.0, Mar. 2022.

[23]. J. Zhang, Z. Zhang, and D. Wang, "Explainable AI in finance: A survey," IEEE Access, vol. 9, pp. 126581–126599, 2021. doi: 10.1109/ACCESS.2021.3110594.

[24]. S. Ruder, "Neural transfer learning for natural language processing," Ph.D. dissertation, Nat. Univ. Ireland, Galway, Ireland, 2019.

[25]. M. Chen et al., "Evaluating large language models trained on code," arXiv preprint arXiv:2107.03374, 2021. doi: 10.48550/arXiv.2107.03374.

[26]. A. Fan et al., "Augmenting transformers with KNN-based composite memory," in Proc. Assoc. Comput. Linguistics (ACL), Online, Aug. 2021, pp. 327–340. doi: 10.18653/v1/2021.acl-long.28.

[27]. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), Virtual Event, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[28]. McKinsey Global Institute, "The value of generative AI in banking," McKinsey & Company, Rep., 2022.

[29]. OECD, "OECD AI policy observatory," 2022.

[30]. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 1st ed., IEEE, New York, NY, USA, 2019.

[31]. Poushali Das, Charanjit Singh, Rituparna Mondal, Dipika Paul, Nitu Saha and Siddhartha Chatterjee "An Intelligent Geofenced Air Quality Monitoring System: Real-Time AQI Detection and Autonomous Location-Based Health Intervention Using Machine Learning" in International Journal of Innovative Science and Research Technology (IJISRT). Vol. 10, Issue. 12, ISSN No. 2456-2165,pp.2376-2389,DOI:https:/doi.org/10.38124/ijisrt/25dec1660 on 2026/01/05.

[32]. Nayan Adhikari, Pallabi Ghosh, Abhinaba Bhattacharyya and Siddhartha Chatterjee "AQIP: Air Quality Index Prediction Using Supervised ML Classifiers" in International Journal of Innovative Science and Research Technology (IJISRT). Vol 10,Issue.7,ISSNNo.2456-2165,pp.835-842,DOI: https://doi.org/10.38124/ijisrt/25jul758 on 2025/07/16.

[33]. Nitu Saha, Rituparna Mondal, Arunima Banerjee, Rupa Debnath and Siddhartha Chatterjee "Advanced DeepLungCareNet: A Next-Generation Framework for Lung Cancer Prediction", in International Journal of Innovative Science and Research Technology (IJISRT), Vol. 10, Issue.6, ISSN No. 2456-2165, pp. 2312-2320, DOI: https://doi.org/10.38124/ijisrt/25jun1801 on 2025/07/02..

[34]. Rajdeep Chatterjee, Siddhartha Chatterjee, Saikat Samanta and Suman Biswas "AI Approaches to Investigate EEG Signal Classification for Cognitive Performance Assessment" In the 6th International Conference on Computational Intelligence and Networks (CINE 2024), IEEE Conference Record#63708, IEEE Computer Society, IEEE CTSoc, IEEE Digital Explore indexed by SCOPUS and Web ofScience(WoS),pp.1-23,February,2025,DOI: 10.1109/CINE63708.2024.10881208.

[35]. Sima Das, Siddhartha Chatterjee, Altaf Ismail Karani and Anup Kumar Ghosh, "Stress Detection while doing Exam using EEG with Machine Learning Techniques", In the Proceedings of Innovations in Data Analytics (ICIDA 2023, Volume 2), Lecture Notes in Networks and Systems (LNNS, Volume 1005), ISSN Electronic: 2367-3389, ISBN (eBook): 978-981-97-4928-7, pp.177-187, 10th Sept. 2024, DOI: http://doi.org/10.1007/978-981-97-4928-7_14, Springer Singapore.

[36]. Ahona Ghosh, Siddhartha Chatterjee, Soumitra De and Atindra Maji,"Towards Data-Driven Cognitive Rehabilitation for Speech Disorder in Hybrid Sensor Architecture", 2022 IEEE International Conference on Electronics, Computing and CommunicationTechnologies(CONNECT),2022,DOI: 10.1109/CONNECT55679.2022.9865794,pp. 1-6.

[37]. Mauparna Nandan, Siddhartha Chatterjee, Antara Parai and Oindrila Bagchi, "Sentiment Analysis of Twitter Classification by Applying Hybrid Based Techniques", Lecture Notes in Electrical Engineering (LNEE), ICCDC 2021, vol 51, pp 591-606, http://doi.org/10.1-1007/978-981-16-9154-6_1, Springer on 2022.

[38]. Sudipta Hazra, Siddhartha Chatterjee, Rituparna Mondal and Anwesa Naskar "Analysis and Comparison Study of Cardiovascular Risk Prediction using Machine Learning Approaches" in the Proceedings of International Conference on Advanced Computing and Systems (AdComSys2024), Springer Nature Book Series, Singapore, "Algorithm for Intelligent Systems" – SCOPUS, Web of Science Indexed, DOI: https://doi.org/10.1007/978-981-97-9532-1_11 pp. 125-134 on 23rd July 2025.

[39]. Anudeepa Gon, Sudipta Hazra, Siddhartha Chatterjee and Anup Kumar Ghosh "Application of Machine Learning Algorithms for Automatic Detection of Risk in Heart Disease" In IGI Global, Book Name – Cognitive Cardiac Rehabilitation Using IoT and AI Tools, DOI: 10.4018/978-1-6684-7561-4, pp. 166-188, ISBN13: 9781668475614, EISBN13: 9781668475621.

[40]. Sudipta Hazra, Swagata Mahapatra, Siddhartha Chatterjee and Dipanwita Pal, "Automated Risk Prediction of Liver Disorders Using Machine Learning" In the proceedings of 1st International conference on Latest Trends on Applied Science, Management, Humanities and Information Technology (SAICON-IC-LTASMHIT-2023) on 19th June 2023, ISSN: 978-81-957386-1-8, pp. 301-306, In Association with Alpha-LPHA Scientific work, IQAC, Department of Science, Computer Science and Application, Sai College.

[41]. Payel Ghosh, Sudipta Hazra and Siddhartha Chatterjee, "Future Prospects Analysis in Healthcare Management Using Machine Learning Algorithms" In the International Journal of Engineering and Science Invention (IJESI), ISSN (online): 2319-6734, ISSN (print): 2319-6726, Vol.12, Issue 6, pp. 52-56, Impact Factor – 5.962, UGC Sl. No.- 2573, Journal No.-43302, DOI: 10.35629/6734-12065256, June 17, 2023.

[42]. Mauparna Nandan, Siddhartha Chatterjee, Antara Parai and Oindrila Bagchi, "Sentiment Analysis of Twitter Classification by Applying Hybrid Based Techniques", Lecture Notes in Electrical Engineering (LNEE), ICCDC 2021, vol 51, pp 591-606, http://doi.org/10.1-1007/978-981-16-9154-6_1, Springer on 2022.

[43]. Sangita Bose, Siddhartha Chatterjee, Bidesh Chakraborty, Pratik Halder and Saikat Samanta, "An Analysis and Discussion of Human Sentiment based on Social Network Information", In International Journal of HIT Transaction on ECCN, Online at http://hithaldia.in/paper/7_1a/J7_1A_05.pdf, Print ISSN: 0973-6875, vol. Issue 1A (2021), pp. 62-71, DOI: 10.5281/zenodo.5892855, 2021 at Haldia Institute of Technology Publishing (ECCN Transaction).

[44]. Rajdeep Chatterjee, Siddhartha Chatterjee, Ankita Datta and Debarshi Kumar Sanyal, "Diversity Matrix based Performance Improvement for Ensemble Learning Approach", In Hybrid Computational Intelligence: Research and Applications, 2019, CRC Press, Taylor and Francis Group on October 1, 2019, ISBN-97811-3832-0253-CAT#K391719.

[45]. Sutirtha Kumar Guha, Somasree Bhadra, Sudipta Hazra, Siddhartha Chatterjee and Abhinaba Bhattacharyya "Classical Optimization Problem Solution using Nature Inspired Algorithm", in IEEE 4th International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication and Computational Intelligence (RAEEUCCI-2025), IEEE Xplore Digital Library, IEEE Madras Section, pp.DOI:10.1109/RAEEUCCI63961.2025.11048319, ISBN: 979-8-3503-9266-1,SCOPUS & DBLP Indexed on 28th June, 2025 organized by SRMIST, Tamil Nadu, India.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1585

[46]. Arunima Banerjee, Nitu Saha, Arijita Washim Akram, Saundarya Biswas and Siddhartha Chatterjee "Handwritten Digit Pattern Recognition by Hybrid of Convolutional Neural Network (CNN) and Boosting Classifier", in International Journal of Innovative Science and Research Technology (IJISRT), Vol. 10,Issue.7,ISSNNo.2456-2165,pp.1012-1025,DOI:https://doi.org/10.38124/ijisrt/25jul782 on 2025/7/17.