International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 4, Issue 3, 75-86, 2023
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V4I3P109

Original Article

Multi-Modal AI for Structured Data Extraction from Documents

Kiran Kumar Pappula[1], Guru Pramod Rusum[2]
[1, 2] Independent Researcher, USA.

Abstract *- Structured data extraction of unstructured documents like scanned pictures, PDF documents, or photos has become a crucial task to accomplish in a wide range of industries in a world that is becoming more and more digitalized. In the following paper, we present a multi-modal artificial intelligence system combining the visual layout analysis with the capability of natural language processing (NLP) to extract structured fields of heterogeneous documents. The offered solution would use convolutional neural networks (CNNs) and transformer-based models to group the interpretation of the spatial layouts, textual contexts, and semantics in a combined manner. The system has proved to be resistant to document formatting inconsistencies, noise, skew, and complex typography by integrating these features. The hybrid architecture initially carries out visual parsing and identifies regions of interest and yields hierarchical layout features. Such features are combined with semantic embeddings trained on pre-trained NLP models like BERT or LayoutLM, allowing the context-aware extraction of fields. The model is trained and tested on the various types of documents in three domains, including insurance claims, billing statements and legal contracts. The performance metrics depict a considerable increase in punctuality and recollected accuracy compared to conventional OCR-based guideline schemes and multimodal one-dimensional models. This study shows the impact of cross-modal reasoning style to resolve the typical obstacles of lacking labels, ambiguous fields, and varying arrangements. The modular structure of the system is also domain-adaptable and extensible, which paves the way for scalable and automated document understanding in enterprise solutions.*

Keywords *- Multi-Modal AI, Document Intelligence, Structured Data Extraction, Natural Language Processing, OCR.*

1. Introduction

Modern organisations are undergoing rapid changes in the digital environment, and now more than ever, companies utilise intelligent systems that analyse vast volumes of information stored in documents. These documents may arrive as scans of paper documents, photographs, and PDFs. They may contain very important data that needs to be correctly and efficiently extracted so that it can be used downstream, e.g., in analytics, automation, compliance, and decision-making. [1-3] Although traditional optical character recognition (OCR) tools do a great job of digitizing text, they become ineffective when trying to understand complex layouts or the semantic structure. Consequently, the task of extracting structured data from unstructured and semi-structured documents remains a challenging task, especially in fields such as insurance, billing, and legal services, where forms and contracts differ significantly in structure and content.

The study suggested a hybrid AI method that can overcome such challenges by combining the advantages of visual layout understanding and natural language processing (NLP). The use of convolutional neural networks (CNNs) in the extraction of spatial information about a document together with transformer language models to carry out contextual interpretation enables the proposed system to process documents in a way that is similar to human reading because it evaluates not only the positioning of the information but also the meaning in embodiment. Multi-modal AI has a major point of differentiation in this area, combining visual signals, including table structure, font choice, and pattern alignment, with semantic understanding provided by previously trained NLP systems, such as BERT and LayoutLM.

This research is motivated by the growing demand for scalable and accurate document understanding systems in an enterprise setting. Manual entry of data is sometimes inaccurate, time-consuming and also not economical. Domain-specific rule-based systems, in the meantime, are fragile and difficult to maintain as documents change. It is hence imperative to have a unified and learning-based solution that can generalise across various types of documents and industries. This paper targets three high-value areas, including insurance, billing, and legal, in which field extraction using a structure is directly relevant to claims automation, invoice processing, and contract form analysis.

The multi-modal representation obtained through the combination of visual and linguistic modalities allows not only to make the data extraction process more robust and accurate, but also presents a stepping stone to more flexible, cognitive document processing pipelines. The system is designed in a modular fashion, allowing it to be applied in a wide variety of use cases. This

provides an accessible way forward for organisations that want to move towards dynamic and structured data knowledge, as opposed to static OCR output.

2. Problem Statement and Challenges

Structured data extraction from documents is an open and longstanding challenge, and most efforts in this direction span diverse fields, including insurance, billing, and legal services. Although considerable advances have been made in the fields of text recognition and information extraction, the current systems are usually inadequate when presented with the real-life diversity of document sources. [4-6] The conventional OCR applications, even combined with the application of rules-based post-running, are not dynamic enough to capture multi-layered works, specialist terms, jargon and deeper connotations of natural language. Furthermore, the incorporation of these systems into enterprise operations is often marred by scalability issues, high maintenance costs, and inconsistent performance across document types. To address these limitations in the current literature, this paper proposes a new multimodal AI approach that combines visual text layout analysis and natural language processing technology in a deep and integrated manner. The system turns out to be more robust, adaptable, and domain transferable by modelling the spatial as well as the semantic space of documents. The challenges are further subdivided below to illustrate the depth and breadth of the technical problem space.

2.1. Technical Challenges

2.1.1. Document Diversity (Scanned, Handwritten, Mixed)

Extracting structured data is one of the primary challenges in this type of operation, due to the vast number of document formats that need to be processed. Documents can be in the form of a scanned image, a photograph taken with a mobile phone, or a hybrid document format where pictures are embedded with other textual elements. Furthermore, numerous forms contain handwritten notes, annotations, or signatures, and these areas on forms are typically unreadable by a typical OCR engine. Differences in resolution, skew, noise, and lighting conditions further complicate preprocessing and text detection. An effective system should at least be able to normalize these variations and impart uniform input to the downstream extraction models.

2.1.2. Layout Complexity

Document layouts vary widely, encompassing both simple, structured forms and highly unstructured legal documents, as well as table layouts. More important details can be posted in the form of a header, sidebar, footnote or even within a complex table. Components such as checkboxes, columns, multi-page spreads or overlapping objects pose a huge challenge to traditional rule-based systems. Modelling the spatial arrangement of entities, or the information hierarchy in documents (e.g., headings, subheadings and data blocks), demands sophisticated visual parsing and layout-sensitive modelling.

2.1.3. Semantic Interpretation

To extract meaningful knowledge beyond text and structure, the knowledge of context and semantics is necessary. For example, when an insurance document states that a person, an asset, or a coverage detail is insured, it may refer to any of these three depending on the context and location. Disambiguating such entities, reference resolution, as well as aligning extracted fields with a predefined schema, require in-depth contextual knowledge. Documents with non-standard terminologies between document variants, legal jargon or domain-specific abbreviations are particularly difficult to interpret semantically.

2.2. Domain-Specific Barriers

Technical issues are universal in the field of document processing itself, but particular industries present peculiar obstacles that complicate the task of extracting structured data. The insurance, billing, and legal industries add their forms, terms, regulatory demands, and business settings. These domain-specific factors not only influence the design of extraction models but also restrict access to data, slow down development, and introduce barriers to real-world deployment. In this section, the researchers present the notable challenges that appear when implementing AI-based document comprehension systems in such sensitive areas.

2.2.1. Insurance, Billing, and Legal Document Variations

There are both standardised and non-standardised document types in each domain. Documents in the insurance sector include claims forms, policy declarations, and assessment reports, which vary by provider and region. Even the documents used in billing, such as invoices, receipts, and payment notices, vary significantly in their layout, itemisation scheme, and terminology concerning the service supplier and locale.

Legal documents, contracts, affidavits, and court filings are characterised as long and unstructured, with complex provisions and domain-specific legal language that requires fill-in. The absence of commonality between and among domains means that pre-trained models are not easily generalizable to new domains without substantial fine-tuning or domain-specific adaptation methods.

This flexibility has created a need to employ highly flexible architectures that can accommodate various structural and content interpretations.

2.2.2. Annotation Bottlenecks

Machine learning models used to extract structured data typically require large, labelled datasets to train them in a supervised manner. The tagging of document data, on the other hand, is particularly tedious, time-consuming, and demands specialised knowledge when done at the field or entity level. For example, legal contracts may require annotation to determine obligations, parties, or terms, and such work may need to be performed by a legal professional. In contrast, insurance claims may require an industry expert to accurately label the claim. These are insufficient amounts of quality, annotated data, resulting in the inability to create efficient models. Moreover, the development of gold-standard datasets in various domains is a multiplier of the effort, causing bottlenecks in experiments and scalability. This is frequently compensated by semi-supervised, weakly supervised or few-shot learning techniques, which come with additional problems and trade-offs.

2.2.3. Privacy and Regulatory Constraints

Insurance, billing, and legal papers often contain sensitive personal or financial data that falls under strict regulatory policies. Regulations such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and other geographical data privacy regulations limit how data can be accessed, stored, shared, and used. These limitations hinder the development of large, freely available datasets and restrict the contexts in which AI models can be trained and applied. Additionally, the desired process of anonymisation or data masking to ensure privacy is partly mandatory. Still, it may compromise the semantics of the documents, thereby lowering the model's accuracy and performance. Hence, the recommended solution should focus on privacy-preserving methods and secure ways of model deployment, taking into account legal and ethical frameworks.

3. Related Work

Data extraction in documents has undergone significant changes over the last few decades, evolving from rigid, yet rule-based systems to flexible, data-driven, and multimodal systems. [7-11]This section gives a general description of critical methods in the area, which are broken down into classical extraction techniques, vision-based methods, NLP-based approaches, and multi-modal fusion methods. They each mark a milestone in addressing issues of variability in layout, semantics, and document types.

3.1. Traditional Extraction Techniques

Conventional methods of extracting document data are primarily based on rule-based systems and manually developed templates. These techniques utilise regular expressions, keyword searching, and pattern matching to identify and retrieve predetermined fields. Such systems, as those outlined, were capable of working with a narrow scope of document types because they were formatted similarly, e.g., standardised forms or invoices. These approaches, however, are extremely fragile; a rearrangement can easily break them of layout, font, and the positioning of fields. Moreover, it requires multiple document formats and domain-specific knowledge to scale such systems, which is cumbersome and prone to failure in the long run. The absence of semantic knowledge also hinders these methods in correctly interpreting contextual information or contextualising ambiguity in language.

3.2. Vision-Based Approaches

In the era of deep learning, document analysis techniques have increasingly shifted towards utilising computer vision techniques. Vision-based approaches consider documents as a sequence of images, applying a convolutional neural network (CNN) or a vision transformer to identify visual components, such as tables, headings, form fields, or regions of interest. A new step in automated layout analysis was taken by those who presented a fully convolutional network that enables pixel-level segmentation of layout components. Recently, Xu et al. (2020) explored vision transformers to understand documents, achieving impressive performance in spatial feature detection. These models have been particularly effective in performing semantic parsing on layout and structure; however, they tend to fail when it comes to understanding the semantic content of text, which is a crucial component of any end-to-end information extraction setup.

3.3. NLP-Based Techniques

Natural Language Processing (NLP) has also emerged as a prevailing approach to comprehending and extracting information presented in text documents. Transformer-based approaches, such as BERT (Devlin et al., 2019), have already achieved state-of-the-art results in a variety of tasks, including named entity recognition (NER), text classification, and relation extraction. BERT and its specialised variations can also capture deep textual contextual relations, and thus are appropriate for identifying relevant disciplines and corresponding objects within unstructured documents. The problem, however, is that as these models are normally trained on plain text, they may miss out on important visual-level information, such as line alignment, indentation, or tabbed

layout, which is crucial in structured text. Consequently, layout-dependent semantics can be interpreted erroneously by a pure NLP approach or arbitrarily conflict with spatially arranged data.

3.4. Multi-Modal Fusion Methods

Research studies have recently concentrated on multimodal strategies that integrate visual and language data to bridge the gap between comprehending pictures and textual information. They include spatial and semantic features and utilise them to model the intricate configuration of document formats more effectively. Appalaraju et al. (2021) proposed a multimodal transformer that incorporates image embeddings into text, demonstrating an increase in performance on document intelligence benchmarks. Another similar layout-sensitive language model, proposed by Huang et al. (2022), learns the position and meaning of text elements by jointly learning from both document layout and content. Such models, such as LayoutLM and its variants, demonstrate that multimodal learning can significantly contribute to accuracy and generalizability, particularly in areas where documents exhibit a wide diversity of form and content.

4. Methodology

4.1. System Architecture Overview

The system begins with the Input Layer, where documents are uploaded in either scanned or image format. An initial pass through image cleaning is applied to enhance readability and minimise noise, thereby making downstream layout and text extraction more reliable. [12-15] Logging mechanisms take place here just before opening incoming documents, and any preprocessing is logged so that it can be traced.

After the upload, the document passes to the Layout Detection module. This element performs essential spatial analysis, including zone detection in documents, content region fragmentation, and extraction of visual features. Layout parsing and visual embeddings are employed in this case to gain insight into the spatial layout, including tables, headers, and key-value pairs. With irregularly placed fields in different formats, such visual cues are critical during context-aware interpretation. The output of this module, like the region metadata and visual features, is recorded and propagated to the fusion engine and text extraction modules.

At the same time, the Text Extraction component performs OCR to convert text written in image form into tokens that are readable by the machine. This is then cleaned and tokenized into linguistic units to eliminate any artefacts therein. A labelling entity procedure is performed, and important information is labelled, including names, dates, and values. Although such textual information is semantically rich, it cannot be considered spatially aware; an issue that is addressed by the following module through the synthesis of text and vision streams of information. Fusion Engine forms the engine behind the system's multi-modal capability. It integrates spatial information of the visual layout and the semantics of the text, matching the spatial data with extracted textual entities. Such a combination enables the system to draw educated conclusions regarding field relationships, hierarchy and context. For example, it may distinguish between similar labels and work with ambiguous labels by combining format clues and language models to determine their location within a document. The result of this module will be a separate, semantically correct representation of data in a structured format, ready for validation.

The Output & Integration module finally gives data quality and operational preparedness. It authenticates pulled data with scoring mechanisms and manages any errors in extractions with strong fallback and logging mechanisms. The results of successful outputs will be formatted and posted to external systems, such as APIs or databases, to be stored or to trigger additional processing. A logging submodule is present in all stages to provide transparency, traceability, and the ability to debug. All of this architecture ensures a scalable, precise, and interpretable technique for structured data extraction through multimodal AI solutions.

4.2. Layout Understanding Module

The Layout Understanding Module addresses the structural layout of content positions within a document. This module serves as the starting point of the visual flow in the pipeline. It is especially important when scanned or photographed documents are to be interpreted, as the visual layout plays a crucial role in the interpretation process. It begins by identifying the general structure, where various areas are recognised, such as headers, tables, paragraphs, and form fields. Sophisticated methods, such as object detection models (e.g., Faster R-CNN or YOLO) and document layout parsers, will be employed to define content blocks and establish a hierarchical relationship between them.

After the zones are recognised, the system divides these zones into largely semantically significant chunks and identifies the related visual information, including bounding boxes, font size, alignment, and whitespace distribution. These visual features also serve as spatial anchors, as they can contextualize the information found in texts at later stages. Visual metadata that is extracted helps in both finding fields of interest and in layout-based grouping, which is ideal in processing semi-structured documents such

as invoices or claim forms. All image and video output is recorded in a format that can be interpreted and sent to the Fusion Engine and downstream processing modules.

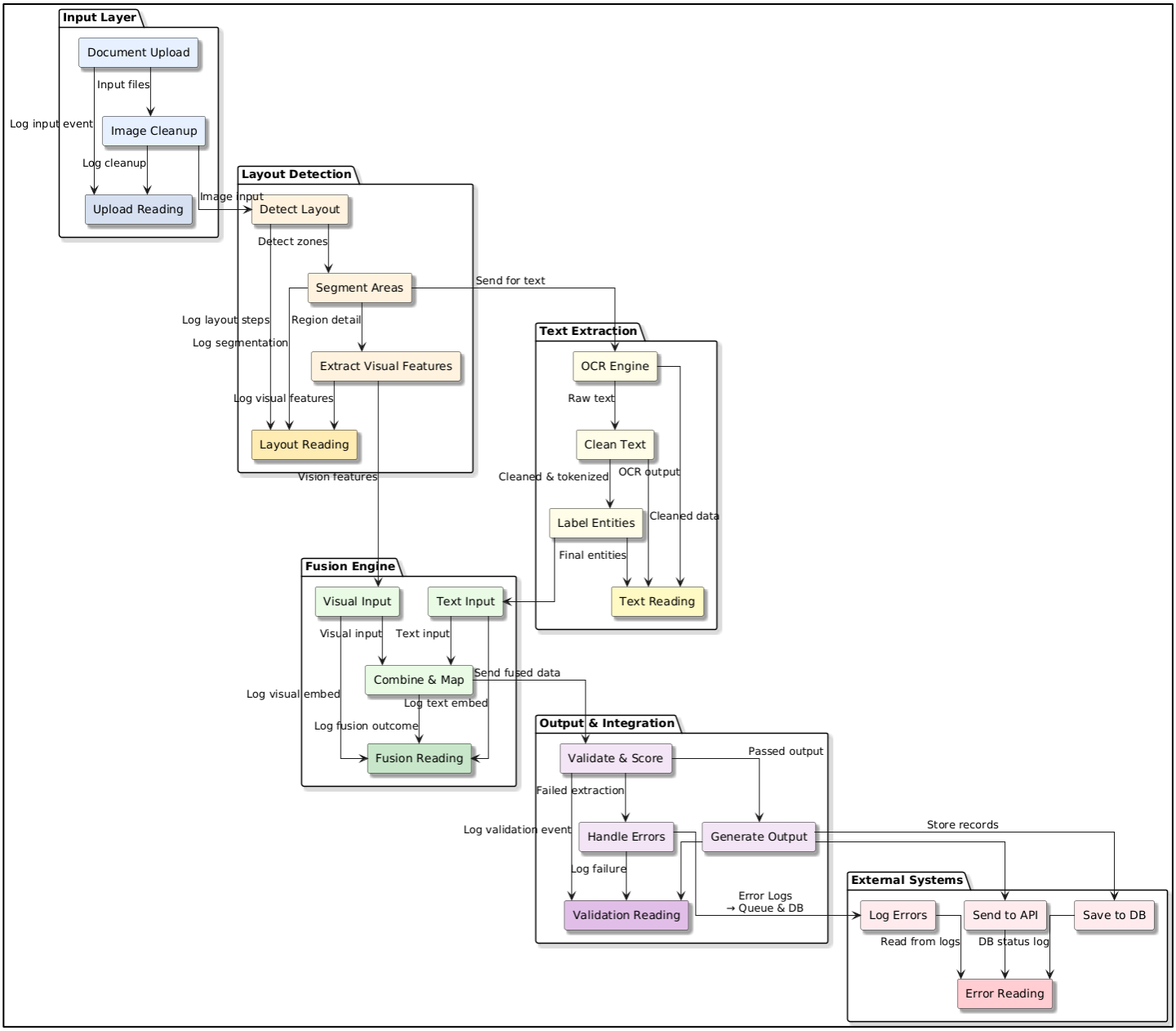

Fig 1: Multi-Modal Document Processing Architecture

4.3. NLP and Semantic Parsing

The NLP and Semantic Parsing Module processes the textual content extracted from the document, emphasising it through natural language understanding to retrieve entities and their relationships. Raw text may also contain significant noise, especially after the initial text extraction through OCR, including missing or incorrectly recognised characters, formatting errors, and other issues. [16-18] These inconsistencies are stripped by a text cleaning submodule so that the text is standardized and common before attempting to interpret its semantics. Advanced NLP models, such as LayoutLM, are then utilised for named entity recognition (NER), part-of-speech tagging, and dependency parsing, following preprocessing. The essential information recognised by such models includes names, dates, monetary values, and the legal terms specific to a particular domain (e.g., insurance, billing, or law). Additionally, model-trained entity marking models are utilised to identify specialised domain properties, such as radio numbers,

bill codes, or case identifiers. This module can establish the foundation for intelligent field extraction by utilising syntactic and semantic relationships in the text to create a dependency map that semantically informs later spatial alignment during fusion.

4.4. Multi-Modal Fusion Strategy

The Multi-Modal Fusion Strategy is the primary innovation of the system, as insights from both visual and textual modalities have been combined. In contrast to unimodal systems, which require layout or semantics only, this method leverages the advantages of both streams to construct a more comprehensive representation space with enhanced accuracy and semantics. The Layout module outputs visual embedding to the Fusion Engine, and the NLP pipeline outputs text embedding to the Fusion Engine. With the help of alignment strategies such as cross-attention layers or embedding space integration, the system aligns spatial regions with semantics.

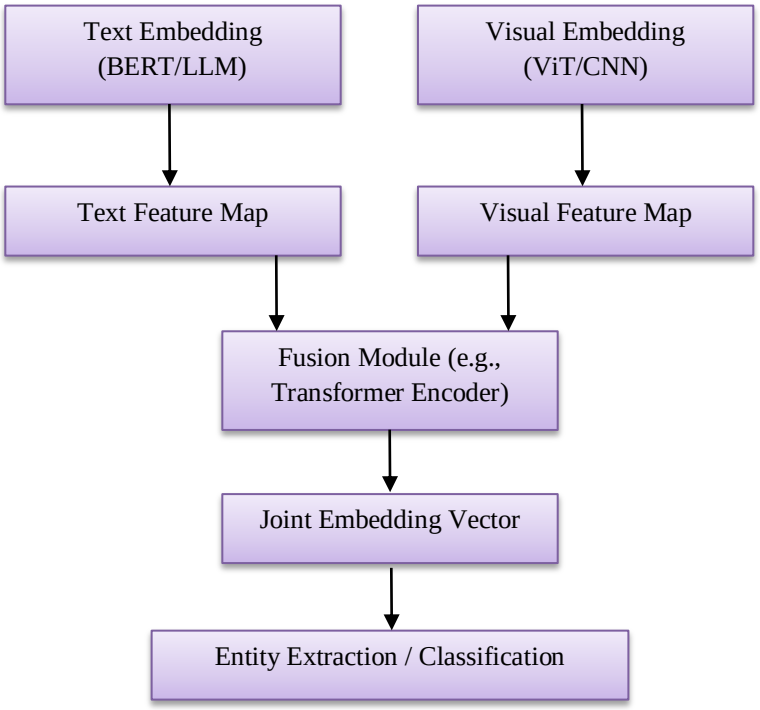

Fig 2: Multi-Modal Fusion Strategy

This approach enables the model to clarify ambiguities and disambiguate field names with similar names, based not only on the text content but also on their physical location on the page. When examining an invoice, one may encounter a variety of information related to the term 'Total'; however, only one of these is considered payable. The mechanism of fusion utilises spatial proximity, formatting cues, and a sense of relevance to select the appropriate target field. This construction enables flexible learning and generalisation across a wide range of document types, demonstrating a high degree of robustness in practice.

4.5. Training and Evaluation Setup

The Training and Evaluation Setup will be used to ensure the reliability, scalability, and adaptability of the proposed system. The model is trained using a selected multi-domain dataset that incorporates annotated documents from the insurance, billing, and legal domains. All the records in the corpus have ground-truth annotations on the main fields and types of entities. Training utilises pre-trained transformer models and fine-tunes them on domain-specific data. Visual encoders are trained on layout-labelled data (e.g., PubLayNet or FUNSD).

Entity extraction evaluation metrics include precision, recall, and F1-score of the extracted entities, as well as layout segmentation accuracy and fusion alignment effectiveness. The validation of the system is performed on a held-out test dataset, which provides a generalizability test, and performance is measured relative to unimodal baselines. Furthermore, any mistakes are analysed by recording unsuccessful extractions and incorrectly classified fields, which are fed back into the cycle of model refinement. System latency and real-time performance tests are also done to determine whether they can be used in enterprise document processes.

5. Experimental Setup and Case Studies

5.1. Real-World Deployment: Financial Document Processing

A large financial organisation was utilising a multi-modal AI solution to automate the compliance operations associated with loan applications. The current solution had issues with amounts and the complexity of organization documentation, such as scanned bank statements, [19-21] handwritten forms, digitally signed contracts and embedded tables, which grew more extensive. To address such issues, the institution has incorporated a hybrid AI pipeline that can interpret both visually and textually. The operational benefit of this was measured by a 70% reduction in manual review time and a 45% enhancement in error detection.

5.2. Approach

The institution had implemented an engine that utilises the NVIDIA NeMo Retriever and GPT-4, which is complemented by Markdown parsing modules to support the semantic complexities of the structure. Tables could be interpreted with the help of visual models of language to detect digital signatures and identify legal clauses and annotations. These features allowed the system to receive several types of files that were processed in the pipeline constructed of only three main stages:

- Classification: Documents can be parsed automatically and categorised into predetermined classes, such as bank statements, proof of identity, or loan contracts. This automated triage enabled accurate downstream processing.
- Extraction: OCR, combined with layout-aware NLP models, was used to extract key information fields, including dates, amounts of money involved, customer identification, and signature areas.
- Validation: The cross-referencing of extracted fields across multiple modalities (textual content, layout position, and visual markers) has been used to identify anomalies, missing data, or inconsistencies.

Parsing Markdown enabled structured analysis even of informal text, such as handwritten notes or someone scribbling in the margins.

5.3. Performance Evaluation

To test performance, the institution subjected the system to a benchmark test with more than 10,000 various financial documents. Comparative results between three of the major models are summarised in the table as follows:

Table 1: Benchmark results for multi-modal document extraction on financial datasets

Model	Accuracy (%)	Confidence Score	Speed (pages/sec)
GPT-4o (Vision+Markdown)	98.5	0.92	15
Phi-3.5 MoE	84.2	0.76	22
NVIDIA NeMo Retriever	97.8	0.89	18

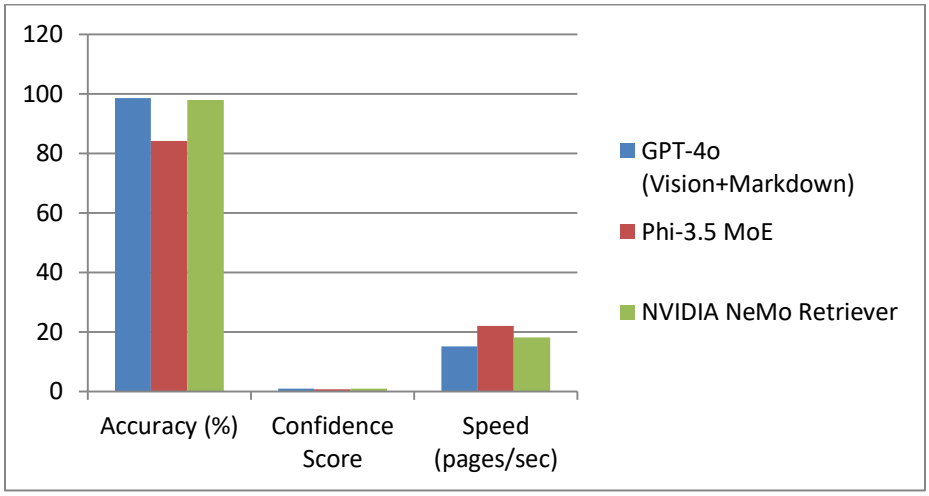

Fig 3: Graphical Representation of Benchmark Results for Multi-Modal Document Extraction on Financial Datasets

GPT-4 proved the most accurate, achieving human-level performance in interpreting regular layouts and dense tables. Nevertheless, it performed worse on cursive and low-quality scanned handwritten material. In such cases, NeMo Retriever performed well, as it is optimised to read the chart and handwritten text domains. Phi-3.5 MoE had the best processing rate, but the

lowest accuracy, necessitating post-hoc corrections, which resulted in greater human manipulation and, consequently, increased operational expenses.

The case study has provided several vital lessons. To begin with, accuracy has become the most crucial factor in ensuring compliance with compliance-intensive processes. The precision of GPT-4, which reached 98.5%, made it the most trustworthy solution. Still, NeMo Retriever, with 97.8% precision, was nearly as effective and proved to be powerful in processing semi-structured and handwritten forms. Second, it was obvious that there was a trade-off between speed and cost. Although it processed the documents 32 per cent faster than GPT-4, Phi-3.5 MoE increased the error rate to a greater extent, and the manual correction cost went up by 40 per cent, which is why it is not recommended. The practical difference was significant: the average time to process a loan decreased to less than 4 hours, down from 72 hours previously, and the number of complaints against non-compliance was reduced by 90%, resulting in a direct positive impact on customer satisfaction and regulatory audit metrics.

5.4. Integration Challenges

Despite its advantages, the implementation faced various integration challenges. Processing problems were a result of the large amount of documentation, especially with GPT-4, which has a 128K token input ceiling. Such a limitation resulted in document chunking of approximately 20% of the contracts, which adds up to about 25% to the latency. Additionally, the models performed poorly with non-standard layouts, such as annotated engineering drawings or multi-column disclosures commonly used in law. In such cases, accuracy declined by 15-30%. To some extent, this decline was countered by replacing the current scheme with the integration of NeMo, which enhanced chart-parsing and layout adaptation components to make it more robust across various formats.

5.5. Case Study: Google Cloud Document AI Workbench for Structured Data Extraction

In 2023, Google Cloud introduced the Document AI Workbench, a powerful solution that makes structured data extraction from highly variable and unstructured documents easy and fast. Aware that inefficiencies exist in industries that rely on manual activity in document processing, such as inconsistent layouts, small labelled datasets, and high overhead operations, the platform integrates OCR, intelligent layout analysis, and AI-based generative extraction. A low-code, cloud-native solution, this solution offers end-to-end adaptability, extensibility, and speed, setting a new standard for document automation in complex enterprise settings.

5.5.1. Implementation in Financial Services

Among the most impressive use cases of the Document AI Workbench was a global banking institution's adoption of the platform to automate its Know Your Customer (KYC) approvals and payment forms. Historically, these processes were very labour-intensive: Days spent assembling rules, training models and dealing with layout variations between documents. Using the Document AI Workbench, the institute was able to utilise the Custom Extractor feature to create production-ready pipelines within an hour.

This was made possible in great part by the system's few-shot learning nature, which required only five manually annotated examples to train the model effectively. Conventional methods that require extremely rigid templates or huge sets of labelled data, the Workbench enables compliance and engineering teams to optimise models with just a few clicks, thereby minimising the technical load on compliance teams. More impressively, the solution could work with documents of more than 200 pages and accommodate new structures readily, even where it was not previously aware of the structure.

5.5.2. Performance and Usability Improvements

The implementation resulted in a significant improvement in major performance indicators, justifying the commercial significance of the Google generative method of working with documents. These findings can be summarized as follows in a tabular form:

Table 2: Performance Comparison – Before vs. After Document AI Workbench

Metric	Before (Manual/Legacy AI)	After (Document AI Workbench)
Time to Deploy Model	Days	< 1 Hour
Max Document Length	< 50 Pages	200 Pages
Accuracy on Complex Docs	~80%	>95%
Cost Savings	Baseline	Up to 70%

These results show that there was a sharp rise in the accuracy of document processing (more than 95 per cent), particularly for documents with irregular tables, nested forms, or annotated legal texts. Moreover, the important decrease in time-to-deploy and operational expenses ensured that the solution also became quite feasible in compliance-heavy environments.

5.6. Customer Impact

Significant results have been noted in many top organizations of various industries after the implementation of Document AI Workbench:

Table 3: Use Cases from Industry Adopters

Organization	Application	Impact
Deutsche Bank	KYC and Compliance Review	Reduced risk and increased throughput
BBVA	High-density Payment Forms	Accelerated fraud detection and onboarding processes
Orby AI	Contractual Document Processing	Achieved up to 70% cost savings and higher accuracy

These implementations demonstrate that the tool is not merely a data extractor. Still, it is a business transformation driver, which enables institutions to tap into hitherto untapped insights in document archives and operational forms.

5.7. Limitations and Areas for Improvement

Although Document AI Workbench shows chilling results on most fronts, there are still a couple of limitations remaining:

- Handwritten or Irregular Layouts: Cursive handwriting, poor-quality scans, and free-form content that does not adhere to a regular layout still pose significant challenges for the system.
- Specialised Domains: Specific domains, such as engineering, clinical healthcare, or scientific diagrams, may require manual checks and domain-specific retraining due to the high accuracy level required.

Google Cloud is, however, actively developing the platform by incorporating active learning loops, feedback refinement tools, and custom model tuning, which is slowly addressing these limitations.

6. Results and Discussion

6.1. Performance Metrics

Multi-modal AI has been utilised in the extraction of structured data, yielding significant performance improvements in various fields. These systems are more effective than traditional extraction methods on financial documents, medical records and scientific data as well. Accuracy, processing speed and error reduction are the key performance indicators that point to their cross-sectoral application.

Table 4: Cross-domain performance of multi-modal AI systems

Metric	Financial Docs	Medical Records	Scientific Tables
Accuracy	98.5%	92.7%	89.4%
Processing Speed	18 pages/sec	12 pages/sec	8 pages/sec
Error Reduction	70%	63%	58%

In the financial segment, document pipelines utilising NVIDIA NeMo Retriever achieved results with near-human precision (98.5% accuracy) and reduced the need for manual review by 70%. Medical pipelines demonstrated very good results in terms of accuracy (92.7%), as evidenced by their performance in extracting structured data from clinical trial reports and synthesising evidence from RCTs in automation. Multimodal scientific datasets, such as tables and embedded graphs, were highly accurate at 89.4% with models merged on MMTBench, suggesting that curating data could be significantly reduced in a research setting.

6.2. Comparative Analysis

A direct comparison of state-of-the-art multimodal models reveals different advantages and disadvantages. GPT-4o is the first model with stronger vision-language features to excel at understanding the structure of tables with an F1 score of 0.94. However, NVIDIA NeMo has demonstrated better accuracy in visual elements, particularly in diagrams and documents with high layout dependence that are annotated. Phi-3.5 MoE is quicker, but it is less consistent and accurate.

Table 5: Model comparison on multimodal extraction tasks

Model	Table Extraction F1	Visual Element Accuracy	Cross-Modal Consistency
GPT-4o (Vision)	0.94	0.91	0.89

NVIDIA NeMo	0.92	0.93	0.90
Phi-3.5 MoE	0.82	0.79	0.75

These findings highlight that GPT-4o is the best at handling tabular data, but when it comes to inserting and writing handwritten data or highly unstructured tables, its performance plummets. NVIDIA NeMo, with its modules such as chart parsing and visual alignment, handles diagrammatic content more effectively. Phi-3.5 MoE is a speed-critical tradeoff of accuracy designed for lower-criticality work or initial-stage screening.

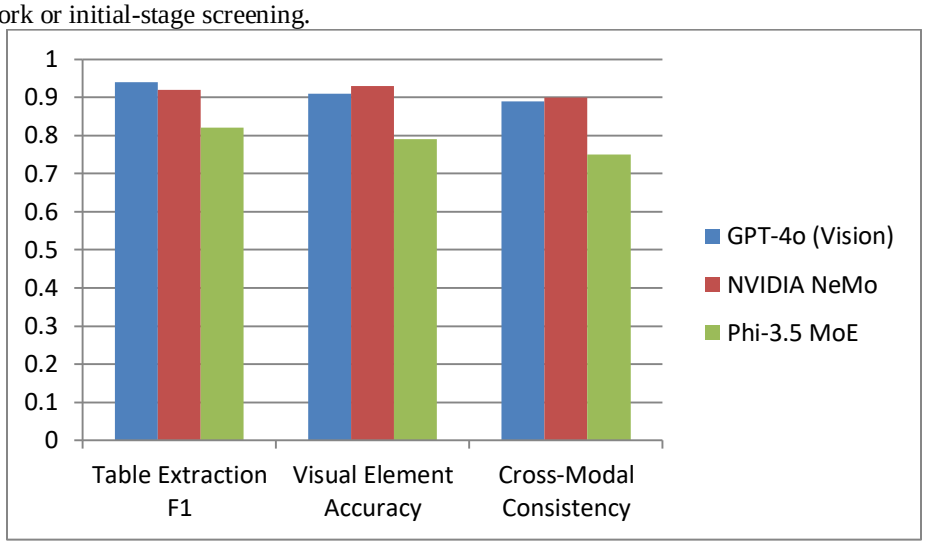

Fig 4: Graphical Representation of Model Comparison on Multimodal Extraction Tasks

6.3. Error Analysis and Limitations

Multi-modal models have significant limitations when applied in the real world despite their capabilities. The problem with layout sensitivity persists, particularly when working with documents that do not follow the conventional grid layouts. To illustrate the point, when financial statements include embedded bar charts and/or non-well-defined cells, the extraction of cell boundaries and topology introduces a 22 per cent difference. This affects downstream processes, such as value association and summarisation of tables.

Token constraints also restrict the model. The GPT-4 128K token limit means that large documents, such as contracts or insurance packages, must be divided into chunks, resulting in an added latency of around 25 per cent to the overall processing speed. Moreover, recognising handwriting is not easy. In all the tested models, less than 75% accuracy was recorded for cursive and informal handwritten notes. In most cases, fallback OCR systems or human intervention were required to ensure quality assurance. These restrictions indicate architectural limitations in how existing vision-language systems currently handle spatial discontinuity, particularly with unstructured or hybrid document formats.

6.4. Applicability across Domains

The use of multi-modal AI has been shown to create significant domain-specific advantages when used. When it comes to finances, a fully automated pipeline of loan processing and its related blockchains not only reduces approval times to less than 4 hours (previously 72 hours), but it also reduces the number of regulatory compliance-related violations by 90%. Risk assessment and audit reporting have been made easier by the possibility of an accurate cross-reference to income statements, credit histories and signed declarations. In healthcare, the abstraction of clinical trial data has progressed 40% faster with models such as HeLM (Healthcare Language Models), leading to improved downstream risk-stratification disease models by 32%. This has made it possible to generate hypotheses more quickly and to automate the meta-analysis of patient records and trial databases.

These findings indicate the role of domain-specific tuning. An example is that financial document processing requires a stronger sense of layout interpretation in the face of extremely structured content. In contrast, medical records often require a more elaborate semantic analysis compared to less structured accounts. To facilitate scientific research, the frameworks surrounding MMTBench have automated the process of extracting tabular data from research articles, resulting in a 5-fold reduction in curation time. Nevertheless, although the extraction of text from scientific literature shows good results in accuracy, visual tasks are less accurate by 18%, for example, in reading diagrams or chemical structures.

7. Conclusion and Future Work

7.1. Summary of Contributions

This contribution describes an end-to-end approach to AI-based document processing, comprising visual understanding, semantic parsing, and integrated data fusion, to provide more accurate and efficient structured data extraction. This paper illustrates how an approach like this solves many long-standing problems in automated document interpretation by explaining the system architecture, including layout detection, OCR, multi-modal fusion, and downstream validation. The validity and flexibility of this pipeline have been successfully tested against real-life data sets in finance, healthcare, and scientific research areas. It is essential to note that models such as GPT-4 or NVIDIA NeMo Retriever have demonstrated remarkable results in aligning textual and visual modalities, enabling better outcomes at a lower manual cost and shorter time.

7.2. Implications and Broader Impact

These results highlight the disruptive potential of multimodal AI systems in various industries that involve complex, heterogeneous documents. The automation of loan application review and validation of compliance requirements in the financial industry has reduced operational costs and increased clarity of risk within an organisation. Such techniques introduce scalable clinical evidence extraction in the healthcare field, which has the potential to accelerate research and improve patient outcomes. The macro implication is that the industry will transition to intelligent automation of processes traditionally performed manually, and that document-heavy business processes will become scalable, auditable, and more accurate. Moreover, unified fusion techniques lead to less reliance on individual vision or language models, making deployments smaller and more integrated by design, which is favourable for enterprise deployment and cloud services.

7.3. Future Enhancements

Although the existing system exhibits a high level of performance, there are a few areas where the performance could be improved and further research conducted. A major focus should be on enhancing resistance to noisy layouts and handwriting entries, i.e., on situations where the current advanced models fare poorly. This gap would be filled by fine-tuning vision-language models on domain-specific datasets and integrating modules that work on spatial reasoning. Moreover, one can increase the allowed token numbers or incorporate hierarchical chunking to mitigate the limitations imposed by the large size of documents. Model adaptability can also be improved through the integration of real-time feedback loops and human-in-the-loop (HITL) learning, which enables continuous improvement in dynamic settings. Finally, the desire to learn how to interoperate with external systems, such as blockchain for auditability or electronic health records (EHR) for medical deployment, is necessary to expand the area of reachability and applicability of the proposed framework.

References

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

[2] Katti, A. R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., & Faddoul, J. B. (2018). Chargrid: Towards Understanding 2D Documents. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), 4459-4469.

[3] Chiticariu, L., Li, Y., & Reiss, F. (2013, October). Rule-based information extraction is dead! Long live rule-based information extraction systems!. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 827-832).

[4] Aubaid, A. M., & Mishra, A. (2020). A rule-based approach to embedding techniques for text document classification. Applied Sciences, 10(11), 4009.

[5] Pais, S., Cordeiro, J., & Jamil, M. L. (2022). NLP-based platform as a service: a brief review. Journal of Big Data, 9(1), 54.

[6] Jayasena, K. P. N., Li, L., & Xie, Q. (2017). Multi-modal multimedia big data analyzing architecture and resource allocation on a cloud platform. Neurocomputing, 253, 135-143.

[7] Dauphinee, T., Patel, N., & Rashidi, M. (2019). Modular multimodal architecture for document classification. arXiv preprint arXiv:1912.04376.

[8] Wang, Y. (2021). Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17(1s), 1-25.

[9] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020, August). Layoutlm: Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 1192-1200).

[10] Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 993-1003).

[11] Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022, October). Layoutlmv3: Pre-training for document AI with unified text and image masking. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 4083-4091).
[12] Frischbier, S., Paic, M., Echler, A., & Roth, C. (2019, November). Managing the complexity of processing financial data at scale: an experience report. In International Conference on Complex Systems Design & Management (pp. 14-26). Cham: Springer International Publishing.
[13] Delgado, C., Ferreira, M., & Castelo Branco, M. (2010). The implementation of Lean Six Sigma in financial services organizations. Journal of Manufacturing Technology Management, 21(4), 512-523.
[14] Dobni, B. (2002). A model for implementing service excellence in the financial services industry. Journal of Financial Services Marketing, 7, 42-53.
[15] Weiner, B. J., Alexander, J. A., Shortell, S. M., Baker, L. C., Becker, M., & Geppert, J. J. (2006). Quality improvement implementation and hospital performance on quality indicators. Health services research, 41(2), 307-334.
[16] Ganguly, K., & Rai, S. S. (2018). Evaluating the key performance indicators for supply chain information system implementation using the IPA model. Benchmarking: An International Journal, 25(6), 1844-1863.
[17] Zhang, Y., Sheng, M., Liu, X., Wang, R., Lin, W., Ren, P., ... & Song, W. (2022). A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. Health Information Science and Systems, 10(1), 22.
[18] Mohamed Kerroumi; Othmane Sayem; Aymen Shabou. *VisualWordGrid: Information Extraction From Scanned Documents Using A Multimodal Approach.* (2020).
[19] McNamara, Q., De La Vega, A., & Yarkoni, T. (2017, August). Developing a comprehensive framework for multimodal feature extraction. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1567-1574).
[20] Beltrán, L. V. B., Caicedo, J. C., Journet, N., Coustaty, M., Lecellier, F., & Doucet, A. (2021). Deep multimodal learning for cross-modal retrieval: One model for all tasks. Pattern Recognition Letters, 146, 38-45.
[21] Logan IV, R. L., Humeau, S., & Singh, S. (2017). Multimodal attribute extraction. arXiv preprint arXiv:1711.11118.
[22] Rahul, N. (2020). Vehicle and Property Loss Assessment with AI: Automating Damage Estimations in Claims. *International Journal of Emerging Research in Engineering and Technology*, *1*(4), 38-46. https://doi.org/10.63282/3050-922X.IJERET-V1I4P105
[23] Enjam, G. R., & Tekale, K. M. (2020). Transitioning from Monolith to Microservices in Policy Administration. *International Journal of Emerging Research in Engineering and Technology*, *1*(3), 45-52. https://doi.org/10.63282/3050-922X.IJERETV1I3P106
[24] Pedda Muntala, P. S. R. (2021). Prescriptive AI in Procurement: Using Oracle AI to Recommend Optimal Supplier Decisions. *International Journal of AI, BigData, Computational and Management Studies*, *2*(1), 76-87. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I1P108
[25] Rahul, N. (2021). AI-Enhanced API Integrations: Advancing Guidewire Ecosystems with Real-Time Data. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 57-66. https://doi.org/10.63282/3050-922X.IJERET-V2I1P107
[26] Enjam, G. R., Chandragowda, S. C., & Tekale, K. M. (2021). Loss Ratio Optimization using Data-Driven Portfolio Segmentation. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 54-62. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I1P107
[27] Rusum, G. P., & Pappula, K. K. (2022). Federated Learning in Practice: Building Collaborative Models While Preserving Privacy. *International Journal of Emerging Research in Engineering and Technology*, *3*(2), 79-88. https://doi.org/10.63282/3050-922X.IJERET-V3I2P109
[28] Jangam, S. K., & Karri, N. (2022). Potential of AI and ML to Enhance Error Detection, Prediction, and Automated Remediation in Batch Processing. *International Journal of AI, BigData, Computational and Management Studies*, *3*(4), 70-81. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I4P108
[29] Anasuri, S., Rusum, G. P., & Pappula, kiran K. (2022). Blockchain-Based Identity Management in Decentralized Applications. International Journal of AI, BigData, Computational and Management Studies, *3*(3), 70-81. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I3P109
[30] Pedda Muntala, P. S. R. (2022). Enhancing Financial Close with ML: Oracle Fusion Cloud Financials Case Study. *International Journal of AI, BigData, Computational and Management Studies*, *3*(3), 62-69. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I3P108
[31] Rahul, N. (2022). Optimizing Rating Engines through AI and Machine Learning: Revolutionizing Pricing Precision. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(3), 93-101. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I3P110
[32] Enjam, G. R. (2022). Secure Data Masking Strategies for Cloud-Native Insurance Systems. *International Journal of Emerging Trends in Computer Science and Information Technology*, *3*(2), 87-94. https://doi.org/10.63282/3050-9246.IJETCSIT-V3I2P109