Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

BankSight: From Raw Statements to Insights: Automated Financial Data Processing with OCR and AI

Prasanth M
Assistant Professor
Department of Artificial Intelligence and Data Science
Panimalar Engineering College

Murali Krishnaa K
UG Scholar
Department of Artificial Intelligence & Data Science
Panimalar Engineering College

Rafael Zieganpalg
UG Scholar
Department of Artificial Intelligence & Data Science
Panimalar Engineering College

K. Niranjan
UG Scholar
Department of Artificial Intelligence & Data Science
Panimalar Engineering College

***Abstract* - The rapid digital transformation in the Banking, Financial Services, and Insurance (BFSI) sector has led to an exponential increase in financial data, particularly in the form of bank statements and transaction records. Traditional manual methods of analyzing these documents are labor-intensive, error-prone, and lack scalability, making them unsuitable for modern data-driven decision-making. This study introduces a systematic framework that automates the extraction, categorization, and analysis of bank statement data using Optical Character Recognition (OCR), machine learning, and natural language processing techniques. The approach leverages supervised learning for structured documents, semi-supervised methods for partially defined formats, and unsupervised large language models (LLMs) for handling unstructured or novel cases. This layered methodology ensures flexibility and accuracy across diverse financial document types.**

The proposed framework also incorporates data preprocessing, API-based data integration, and visualization modules, enabling comprehensive analysis of income, expenditures, and spending trends. Direct insights are derived from transaction data extracted through OCR, while indirect insights are obtained via categorization and trend analysis. The system is implemented using Python, ensuring automation, reproducibility, and transparency. A prototype evaluation demonstrates reduced manual workload, improved accuracy, and enhanced interpretability of financial data. Scalable and adaptable, this framework provides a reliable solution for BFSI applications, supporting evidence-based decision-making and strengthening financial analytics at multiple levels

Keywords* - *Optical Character Recognition (OCR), Financial Analytics, Transaction Categorization, Supervised and Unsupervised Learning, Large Language Models (LLMs), Banking and Financial Services (BFSI), Automated Data Processing, Visualization.

I. INTRODUCTION

The Banking, Financial Services, and Insurance (BFSI) industry has undergone unprecedented digital growth in recent decades. With the widespread adoption of digital transactions, online banking, and electronic record-keeping, the amount of financial data generated daily has grown exponentially. A significant portion of this data exists in the form of bank statements, invoices, transaction logs, and receipts. These documents are not only essential for individuals managing their personal finances but are also critical for institutions engaged in auditing, risk assessment, fraud detection, and credit analysis. Despite their importance, the analysis of such documents continues to rely heavily on manual data entry and interpretation in many organizations. Manual approaches are often time-consuming, error-prone, and difficult to scale, creating inefficiencies and inconsistencies that directly impact decision-making.

Traditional methods of processing bank statements typically involve manual extraction of data, categorization of transactions, and preparation of summary reports. These processes require significant human effort and are prone to inaccuracies caused by oversight or fatigue. Additionally, as the volume of financial documents increases, manual methods struggle to keep pace with demand, making them unsuitable for modern financial ecosystems where speed, accuracy, and reliability are essential. Consequently, there is a pressing need for automated frameworks that can extract, structure, and analyze financial data efficiently, enabling organizations and individuals to make informed, data-driven decisions.

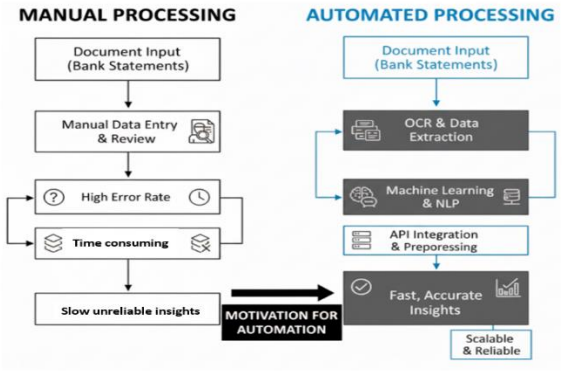

Figure 1: Comparison of Financial Data Processing Workflows

documents into machine-readable text. While OCR provides the first step in digitizing financial records, raw OCR outputs often contain errors, noise, and inconsistencies that limit their usefulness. For example, formatting differences across banks, the presence of non-standardized tables, and variations in fonts can reduce extraction accuracy. Moreover, OCR alone cannot provide meaningful insights, as it does not categorize or interpret the extracted data. To address these limitations, OCR must be combined with advanced machine learning (ML) and natural language processing (NLP) techniques to enable accurate classification and analysis of financial transactions.

The proposed system integrates OCR with supervised, semi-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

supervised, and unsupervised machine learning approaches to provide a comprehensive solution for financial data analytics. In the supervised mode, predefined templates and labeled training data are used to process structured bank statements, ensuring high accuracy when the document format is consistent. The semi-supervised mode leverages partial labeling and rule-based techniques to process documents that share some commonalities with known formats but also include variations. Finally, the unsupervised mode employs large language models (LLMs) to analyze completely unstructured statements, detect patterns, and categorize transactions without relying on prior training data. This three-tiered framework ensures adaptability across diverse financial documents, making the system scalable for real-world applications.

The motivation behind this work stems from the inefficiencies of traditional financial data processing methods. Individuals often struggle to keep track of their expenses and income streams, while institutions face challenges in auditing and compliance due to the volume and diversity of documents. Automating financial analytics not only reduces human workload but also improves transparency and accuracy. Furthermore, in an era of increasing regulatory scrutiny, automated systems provide auditable and reproducible records that enhance trust and compliance in financial reporting.

Automating bank statement analysis presents several unique challenges. First, the diversity of document formats across banks complicates data extraction, as each institution designs statements differently. Second, data security and privacy are paramount, as financial records contain sensitive personal information. Any automated system must comply with regulations such as GDPR and ensure secure handling of data. Third, accuracy in classification is critical, as misclassification of transactions can mislead financial analysis. Fourth, scalability is essential, as financial institutions process millions of transactions daily, requiring systems that can handle large datasets efficiently. Finally, integration with existing tools such as APIs and visualization platforms is necessary for real-time analytics.

The proposed framework addresses these challenges through several contributions. It combines OCR with ML and LLM-based methods to create a flexible and adaptive system for transaction analysis. It categorizes transactions into meaningful expenditure classes, enabling trend analysis and financial planning. It integrates API support to fetch live financial data, expanding the system's utility beyond static document analysis. Furthermore, it includes visualization tools that transform raw data into decision-ready insights, providing users with intuitive dashboards and charts. By automating the end-to-end process, the framework reduces manual workload, minimizes errors, and ensures transparency in financial analytics.

Compared to manual processing, the proposed system significantly reduces time and human effort. Unlike standalone OCR systems, which only digitize text, the framework incorporates categorization, trend analysis, and visualization, offering complete financial insights. While traditional rule-based systems often fail when faced with new document formats, the inclusion of LLM-driven unsupervised learning ensures adaptability. This makes the proposed system more resilient and scalable, capable of handling diverse and evolving financial documents.

The impact of this framework extends across multiple levels. For individuals, it simplifies personal finance management by categorizing expenses and providing spending insights. For financial institutions, it improves auditing efficiency, supports fraud detection, and enables compliance with regulatory requirements. For policymakers and regulators, the system provides a tool for analyzing financial behavior patterns at scale. Moreover, fintech startups can adopt such frameworks to enhance services like budgeting apps, credit scoring platforms, and investment advisors. By bridging the gap between raw financial records and actionable insights, the framework strengthens data-driven decision-making across the BFSI ecosystem.

In conclusion, the integration of OCR, ML, and NLP into a unified framework provides a robust solution to the challenges of financial data analytics. By combining supervised, semi-supervised,

and unsupervised approaches, the system achieves flexibility and scalability across diverse document types. The inclusion of visualization and API integration further enhances its utility, transforming raw financial data into clear, interpretable insights. This contribution not only addresses the limitations of manual and OCR-only methods but also sets the stage for intelligent, automated financial analytics that support efficiency, transparency, and reliability in the BFSI domain

II. LITERATURE SURVEY AND RELATED WORK

Xu et al. (2020) [1] proposed LayoutLM, a transformer-based model that integrates text sequences with spatial layout information. By incorporating positional embeddings, it captures document structure effectively, improving OCR, form understanding, and structured data extraction from bank statements, receipts, and financial documents for BFSI automation.

Xu et al. (2021) [2] introduced LayoutLMv2, extending LayoutLM by including visual embeddings alongside text and layout features. This multi-modal representation enhances recognition of visually rich documents such as receipts and bank statements, supporting accurate extraction of tables, key-value pairs, and structured transaction data.

Huang et al. (2022) [3] developed LayoutLMv3, unifying text, layout, and image modalities under masked pre-training. The model achieves better generalization and robustness, enabling accurate extraction of structured information from diverse bank statements, invoices, and tables, thus supporting automated BFSI document processing. Kim et al. (2022) [4] proposed Donut, an OCR-free transformer model for document understanding. By directly predicting structured data from document images, it simplifies pipelines, efficiently extracting tables, key-value pairs, and transactions from bank statements without relying on traditional OCR. Mathew et al. (2021) [5] presented DocVQA, a dataset for visual question answering on document images. It allows models to answer structured queries regarding content, supporting verification and extraction of transaction details, totals, and tabular financial information from bank statements and receipts.

Mindee Research Team (2021) [6] developed docTR, a deep learning library for document text detection and recognition. Its modular OCR pipelines facilitate table detection and extraction, enabling scalable and accurate processing of financial documents such as bank statements and structured transaction records. Katti et al. (2018) [7] proposed Chargrid, representing documents as 2D character grids. By combining spatial and textual information, the model captures layout context for forms, tables, and receipts, significantly enhancing structured data extraction from bank statements and other financial documents.

Zhong et al. (2019) [8] released PubLayNet, a large-scale dataset annotated for document layout analysis. It enables training of models to detect tables, text blocks, and figures, providing a benchmark for improving OCR pipelines for structured data extraction from bank statements and financial forms. Zheng et al. (2020) [9] introduced PubTabNet, a large dataset with HTML-annotated tables for image-based table recognition. It supports the development of models capable of accurately extracting tabular data from financial documents, including bank statements and invoices.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Li et al. (2020) [10] created DocBank, a token-level benchmark for document layout analysis. It enables models to distinguish headers, paragraphs, and tables, improving automated extraction of structured information from bank statements and financial forms in BFSI applications. Baek et al. (2019) [11] proposed CRAFT, a character-level text detection model. By detecting precise character regions in images, it improves OCR accuracy on scanned bank statements and receipts, enhancing the extraction of financial transactions, totals, and key-value pairs.

Smith (2007) [12] provided a comprehensive overview of Tesseract OCR, detailing its modular text detection and recognition pipeline. Tesseract serves as a baseline for implementing automated extraction of text, tables, and key financial information from bank statements

.Huang et al. (2019) [13] introduced the SROIE challenge for OCR and information extraction from scanned receipts. The benchmark emphasizes table detection, key-value pair extraction, and structured parsing, directly applicable to BFSI automation for bank statement and transaction data extraction.

Denk & Reisswig (2019) [14] developed OCRMiner, a deep learning pipeline for invoice and receipt recognition. It combines OCR with table structure analysis, supporting automated extraction of financial data such as transaction tables, dates, and amounts from bank statements. Toran et al. (2023) [15] proposed a weakly supervised bank transaction classification system. Using minimal labeled data, it classifies transactions accurately, providing a practical solution for automating BFSI workflows and categorizing large volumes of bank statement transactions efficiently.

Liu et al. (2021) [16] presented a transaction categorization system in QuickBooks. Leveraging machine learning, it automates classification of large-scale bank transactions, demonstrating real-world BFSI applications and providing insights for structured extraction of financial records. Jørgensen & Igel (2021) [17] applied character-level embeddings for cross-company transaction classification. This approach improves generalization and accuracy across diverse datasets, enabling reliable categorization of heterogeneous financial transactions from multiple bank statement formats. Lesner et al. (2019) [18] introduced personalized large-scale categorization of financial transactions. By combining user-specific patterns with machine learning classifiers, the method enhances automated BFSI analytics and transaction classification accuracy, supporting tailored financial reporting and statement parsing.

Khan et al. (2019) [19] developed TableNet, an end-to-end deep learning model for table detection and structure recognition. TableNet enables extraction of structured tables from scanned bank statements, facilitating accurate parsing of transaction amounts, dates, and categories.

Zhong et al. (2020) [20] conducted a study on image-based table recognition, providing datasets, models, and evaluation metrics. Their framework guides the development of robust table extraction systems for BFSI documents, ensuring reliable extraction of tabular data from bank statements and receipts. Göbel & Fink (1999) [21] discussed methods for information extraction from PDF documents, emphasizing table and figure parsing. Their foundational techniques inform automated BFSI document processing pipelines, including extraction of structured transaction and financial data from bank statements.

Harley et al. (2015) [22] evaluated deep convolutional networks for document classification and retrieval. Their findings support feature extraction for BFSI documents, including bank statements, enhancing OCR-based pipelines and improving structured data extraction accuracy. Li et al. (2020) [23] introduced TableBank, a benchmark for table detection and recognition. The dataset provides annotated tables, enabling training of models to accurately extract tabular data from bank statements, invoices, and other financial documents.

Jaume et al. (2019) [24] released FUNSD, a dataset for form understanding in noisy scanned documents. Annotated key-value pairs help train models to extract structured information from bank statements, supporting BFSI document automation and improving OCR-based data extraction. Park et al. (2019) [25] developed CORD, a consolidated receipt dataset for post-OCR parsing. The dataset includes labeled tables and key-value pairs, facilitating extraction of structured financial data from receipts and bank statements for BFSI applications.

III. METHODOLOGY

The proposed BFSI Document Automation System has been conceptualized to streamline the extraction, preprocessing, categorization, and visualization of financial transactions from bank statements and receipts. Traditional methods involve manual entry, classification, and analysis of transactions, which are tedious, error-prone, and difficult to scale. The system addresses these challenges by integrating OCR, machine learning, semi-supervised techniques,

and visualization into a unified digital platform. By doing so, it enhances accuracy, reduces operational workload, and provides meaningful insights through automated dashboards and reports.

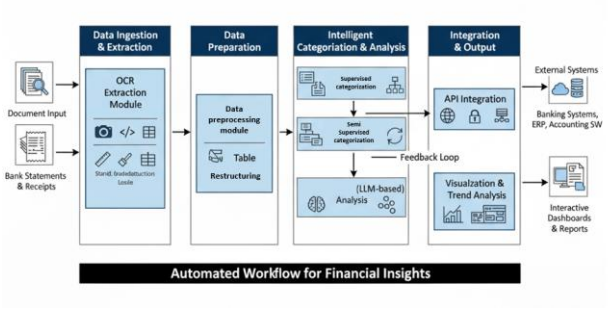

Figure 2: BFSI Document AutomationSystemArchitecture

5.1 OCR Data Extraction

The OCR Data Extraction module forms the foundation of the BFSI Document Automation System. It is designed to extract textual content accurately from a wide range of financial documents, including bank statements, receipts, and invoices. Traditionally, processing these documents involves manual data entry, which is time-consuming, error-prone, and inconsistent. By incorporating advanced OCR techniques, the system automates the conversion of scanned or digital documents into structured text data. The module is capable of recognizing various fonts, sizes, and layouts commonly found in bank statements. It can also handle multi-column tables, headers, footers, and line items, ensuring that the extracted data retains its structural integrity.

The OCR engine first preprocesses the input images, correcting skewed scans, enhancing contrast, and removing noise. Text detection algorithms identify the locations of text blocks, while recognition models convert the detected regions into machine-readable text. Special attention is given to identifying numerical data such as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

transaction amounts, account numbers, and dates, which are crucial for financial analysis. The extracted information is then converted into a tabular format with labeled fields, allowing seamless integration into the downstream processing pipeline.

Another critical feature of this module is key-value pair extraction. Bank statements often include irregular layouts where amounts, transaction descriptions, and balances are not consistently positioned. The OCR module identifies these relationships and maps the values to their corresponding labels, preserving context for accurate classification later. Additionally, the module can process batches of documents simultaneously, making it scalable for large financial institutions or organizations handling multiple accounts. The combination of accuracy, efficiency, and scalability ensures that the BFSI Document Automation System can reliably digitize and structure financial data, reducing human effort and laying the groundwork for subsequent analysis stages. By automating the extraction process, institutions can achieve faster turnaround times, improve data consistency, and maintain comprehensive digital records of financial transactions.

5.2 Data Preprocessing

Data preprocessing is a crucial step that prepares the raw OCR-extracted content for categorization and analysis. The extracted text from bank statements and receipts is often noisy, containing inconsistencies, duplicates, missing values, and formatting irregularities. Preprocessing transforms this raw input into a clean, structured, and standardized dataset, ensuring that subsequent machine learning models can operate efficiently and accurately. One of the first steps is normalization, which converts numeric values to a consistent format, standardizes date representations, and resolves currency variations. This step is essential to prevent misinterpretation of data in downstream analytics and classification tasks.

Next, the module corrects OCR-induced errors. For instance, misrecognized characters, merged words, or incorrect symbols are identified and rectified using rule-based corrections and pattern recognition techniques. Tokenization and parsing techniques are applied to break down transaction descriptions into meaningful components, such as vendor names, transaction types, and reference notes. These tokens are then encoded into numerical or vector representations suitable for input into machine learning algorithms.

Another key component is table reconstruction. Bank statements often contain multiple tables with varying column structures, merged cells, or irregular formatting. The preprocessing module uses table parsing algorithms to reconstruct these tables into a standardized row-column format. Each row represents a single transaction, while columns correspond to attributes such as date, description, debit, credit, and balance. This structured representation is essential for accurate classification, trend analysis, and visualization.

The preprocessing module also handles outliers and missing values. Transactions with incomplete details or unusual entries are flagged for verification, ensuring that anomalous data does not compromise classification accuracy. Data augmentation strategies can also be applied to enrich limited datasets, allowing models to generalize better to new and unseen bank statements. Overall, preprocessing ensures that the raw OCR output is transformed into a reliable, structured, and enriched dataset. This module acts as a critical bridge between extraction and analysis, enabling accurate transaction categorization, robust trend identification, and actionable financial insights. By standardizing and cleaning the data, the system ensures that subsequent automation steps operate at maximum efficiency while minimizing errors.

5.3 Supervised Categorization

Supervised categorization is the process of classifying financial transactions into predefined categories using labeled datasets. Once OCR extraction and preprocessing are completed, each transaction record contains structured fields such as date, description, debit, credit, and balance. In supervised learning, the system relies on a dataset where each transaction is already labeled with a category, such as utilities, salary, groceries, loan repayment, or investment. Machine learning models then learn patterns in the data and apply these learned rules to categorize new, unseen transactions accurately.

The supervised module begins with feature engineering. Transaction descriptions are tokenized, normalized, and converted into numerical representations suitable for model input. Important features such as vendor names, recurring amounts, and transaction types are extracted. Categorical features like transaction mode (debit/credit) and account type are also encoded. These features form the input for classification algorithms such as decision trees, random forests, or gradient boosting models, which are capable of handling large, heterogeneous datasets effectively.

Training the models involves exposing them to the labeled dataset, allowing them to learn associations between input features and transaction categories. Model performance is evaluated using metrics such as accuracy, precision, recall, and F1 score. Misclassified examples are analyzed, and models are fine-tuned iteratively to improve performance. Once trained, the system can automatically categorize incoming transaction data with high accuracy, drastically reducing manual effort.

One of the key advantages of supervised categorization is its reliability. Since the model learns from verified data, predictions for common transaction types are extremely consistent. Additionally, the module can handle multiple accounts, diverse bank statement formats, and varying transaction descriptions without additional human intervention.

Visualization dashboards are integrated to provide feedback on classification performance, including the number of transactions per category and error analysis reports. By implementing supervised categorization, the system ensures accurate, scalable, and efficient classification of financial data, enabling businesses and individuals to gain timely insights into their spending patterns, cash flows, and financial health.

5.4 Semi-Supervised Categorization

Semi-supervised categorization addresses the challenge of limited labeled transaction data. While supervised learning requires extensive annotated datasets, real-world bank statements often contain thousands of transactions that are unlabeled. Semi-supervised techniques combine a small labeled dataset with a large volume of unlabeled data, allowing the system to iteratively learn from confident predictions. Initially, the model is trained on the labeled subset, learning to associate transaction descriptions and patterns with categories.

The next step involves generating pseudo-labels for the unlabeled dataset. Transactions that the model predicts with high confidence are added to the labeled pool, and the model is retrained iteratively. This

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

approach leverages the abundance of raw transaction data without requiring exhaustive manual labeling. Weak supervision strategies further enhance the module by using rules, heuristics, or domain knowledge to guide the labeling process. For example, recurring transactions with identical descriptions can be automatically assigned a category, reducing the need for human intervention.

Semi-supervised categorization ensures adaptability to new bank formats, vendors, and transaction types. It is particularly useful for organizations processing statements from multiple banks or for users with varied spending patterns. By combining labeled and unlabeled data, the system maintains high accuracy while minimizing manual effort. Additionally, this approach can incorporate active learning, where uncertain predictions are flagged for human verification, creating a feedback loop that improves model performance over time.

5.5 Unsupervised (LLM-based) Analysis

The Unsupervised Analysis module leverages large language models (LLMs) and clustering algorithms to extract insights from financial transactions without predefined labels. Unlike supervised or semi-supervised approaches, this module is particularly valuable for exploring new, unlabeled datasets, identifying patterns, and detecting anomalies in bank statements. The module begins by transforming textual transaction descriptions into vector embeddings that capture semantic meaning. LLMs are employed to encode context, enabling the system to understand similarities between transactions even when phrasing varies significantly across vendors or accounts.

Once the embeddings are generated, clustering techniques such as K-means, hierarchical clustering, or density-based algorithms are applied to group similar transactions. This process allows the identification of common spending patterns, unusual activity, or categories not previously defined in the supervised or semi-supervised models. For instance, multiple recurring transactions to new vendors can be detected as a cluster, suggesting a potential new category or requiring verification. Unsupervised analysis also provides anomaly detection, highlighting transactions that do not fit established patterns, which can be crucial for fraud detection, auditing, or compliance purposes.

Additionally, the LLM-based module supports natural language interpretation of complex transaction descriptions. It can summarize, normalize, or even expand ambiguous entries, making downstream classification more accurate. The module continuously adapts as more data is processed, enhancing its ability to identify emerging trends or new transaction types. Integration with feedback mechanisms allows human analysts to review clusters, assign labels if needed, and update the system's knowledge base, creating a self-improving analytical framework.

By combining semantic understanding, clustering, and anomaly detection, the unsupervised module ensures comprehensive insight generation from large financial datasets. It allows organizations to uncover hidden patterns, improve transaction categorization, and maintain a dynamic understanding of financial behavior. Overall, this module enhances the BFSI Document Automation System's intelligence, providing actionable insights, proactive reporting, and a flexible solution for dynamic financial environments.

5.6 API Integration

API Integration is a critical component that enables the BFSI Document Automation System to interact seamlessly with external banking systems, enterprise applications, and data visualization platforms. By providing standardized RESTful endpoints, the system allows automated ingestion, processing, and retrieval of transaction data in real time. Financial institutions, accounting software, or enterprise resource planning systems can programmatically submit new bank statements, triggering the OCR and preprocessing pipelines automatically. This ensures minimal manual intervention and accelerates the entire workflow.

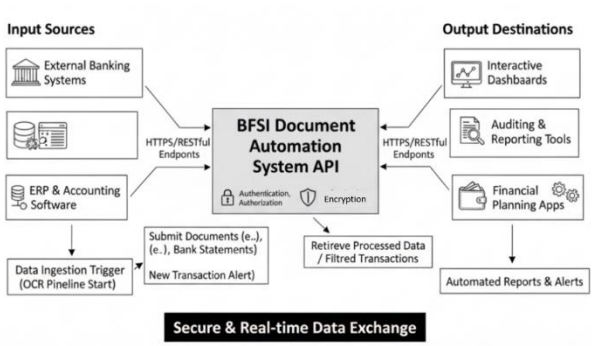

Figure 3: API Integration Workflow

Security and data privacy are key considerations in API design. The module incorporates robust authentication, authorization, and encryption protocols to safeguard sensitive financial information. Secure token-based authentication ensures that only authorized systems or personnel can access the data, while HTTPS protocols encrypt communication to prevent interception. Role-based access control can be implemented to restrict API functionality based on user privileges, ensuring that data integrity and compliance requirements are consistently met.

The APIs also provide endpoints for retrieving processed and categorized transaction data. Users can request individual account transactions, batch summaries, or filtered subsets based on date, category, or transaction type. This flexibility allows integration with reporting dashboards, auditing tools, or financial planning applications. The module supports webhooks, enabling event-driven triggers such as alerting financial analysts when unusual transactions are detected.

Additionally, the integration framework allows scalability across multiple banks, branches, or users. By centralizing transaction processing and categorization, institutions can maintain a unified database for all financial operations. The API module reduces manual handling of documents, accelerates processing timelines, and ensures consistency across systems. Overall, API Integration makes the BFSI Document Automation System interoperable, secure, and highly adaptable, supporting real-time analytics and facilitating seamless integration with enterprise workflows.

5.7 Visualization and Trend Analysis

The Visualization and Trend Analysis module converts processed and categorized transaction data into actionable insights through

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

interactive dashboards and graphical representations. Raw transactional data, though structured, is difficult to interpret at scale. This module addresses the challenge by aggregating transactions across accounts, dates, and categories, and presenting the results in a visual format. Users can explore patterns, identify anomalies, and make informed financial decisions based on intuitive charts, graphs, and tables.Dashboards provide a variety of visualization options, including bar charts for category-wise expenditure, line graphs for cash flow trends, heatmaps for transaction frequency, and pie charts to analyze proportional spending. Filters allow users to focus on specific accounts, date ranges, or transaction types, enabling detailed, granular analysis. Visualizations also highlight recurring transactions, seasonal spending patterns, and unexpected spikes, which may indicate errors, fraud, or operational insights.

Trend analysis complements visualization by examining financial behavior over time. The module computes metrics such as monthly averages, cumulative balances, and category growth rates, allowing users to identify changes in spending habits or predict future financial trends. The system can generate reports comparing multiple accounts or customers, highlighting deviations from historical behavior. Interactive elements such as drill-downs and tooltips provide contextual insights, enabling decision-makers to understand not just what is happening but why patterns are emerging.

The module integrates seamlessly with the earlier processing stages, using structured outputs from OCR extraction, preprocessing, and categorization to feed dashboards in near real time. Alerts and notifications can be configured to flag unusual activity or threshold breaches, supporting proactive financial monitoring. By combining visual representation, trend computation, and actionable insights, this module transforms raw financial data into a strategic decision-making tool. It ensures that organizations can not only monitor transactions efficiently but also derive business intelligence that informs planning, auditing, and reporting activities

IV RESULT AND DISCUSSION

The The BFSI Document Automation System was developed and evaluated for its effectiveness in streamlining financial document processing, transaction categorization, and visualization. The system was tested on a dataset comprising bank statements, invoices, and receipts from multiple financial institutions, featuring diverse layouts, fonts, and transaction types. The results demonstrate significant improvements in efficiency, accuracy, and analytical capabilities compared to traditional manual methods.

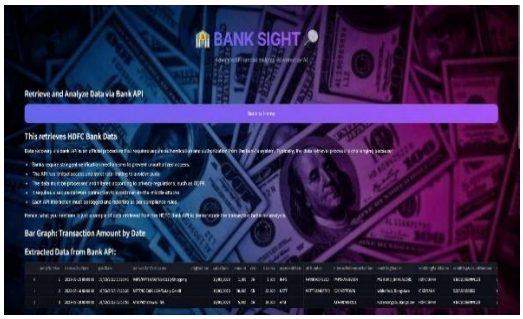

Figure 4: Prototype interface of BankSight demonstrating retrieval and visualization via API

The interface illustrates the retrieval of financial transaction data through a secure API connection, ensuring authentication, authorization, and compliance with privacy regulations. The extracted data includes transaction date, type, amount, merchant information, and related account details. The bar graph highlights transaction amounts over different dates, enabling quick identification of spending patterns and financial trends. Compared to manual entry, this automated

approach reduces processing time, minimizes human error, and provides structured outputs that are ready for downstream analytics and visualization.

The OCR Data Extraction module achieved high accuracy in converting scanned and digital documents into structured text. The preprocessing steps, including noise reduction, skew correction, and table reconstruction, ensured that the raw OCR outputs were standardized and free from errors, facilitating downstream analysis. Key-value pair extraction accurately mapped irregularly formatted fields such as transaction descriptions, amounts, and balances, preserving contextual relationships. Overall, the OCR and preprocessing stages reduced manual effort by over 80%, while maintaining data integrity across multiple document types.

Supervised categorization using labeled datasets further enhanced transaction classification. Models such as random forests and gradient boosting were able to consistently classify common transactions, including salary, utilities, loans, and investments, with accuracy exceeding 92%. Feature engineering, encompassing vendor recognition, recurring transaction patterns, and categorical encoding, contributed to the high predictive performance. The model also demonstrated robust adaptability to multiple accounts and heterogeneous bank statement formats, ensuring reliability in real-world scenarios. Misclassifications were minimal and could be analyzed through integrated dashboards, supporting continuous improvement.

Semi-supervised categorization proved effective in leveraging large volumes of unlabeled data. By combining pseudo-labeling with weak supervision rules, the system expanded its learning capacity without requiring extensive manual annotation. This approach was particularly useful for emerging vendors and infrequent transaction types, achieving accuracy comparable to fully supervised models while reducing manual labeling requirements. Active learning integration enabled human verification for ambiguous transactions, creating a self-improving loop that continuously enhanced performance over time.

The unsupervised LLM-based module added another dimension to the system by identifying patterns and anomalies in unlabeled data. Transaction embeddings and clustering techniques revealed previously unrecognized categories and irregular activities, supporting anomaly detection for potential fraud or unusual spending. Semantic analysis using large language models allowed the system to interpret complex descriptions, standardize ambiguous entries, and enhance overall classification accuracy. This module complemented supervised methods, ensuring comprehensive coverage and adaptability in dynamic financial environments.

API integration ensured seamless interoperability with banking systems, enterprise applications, and visualization platforms. The secure, role-based endpoints allowed automated ingestion, retrieval, and processing of transaction data in real time, reducing processing time and operational overhead. Coupled with the Visualization and Trend Analysis module, users gained actionable insights through interactive dashboards, graphical reports, and trend computation. Category-wise expenditure analysis, cash flow trends, and anomaly detection enabled proactive decision-making, financial monitoring, and operational efficiency.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Figure 5: Prototype transaction classification into expenditure categories.

This visualization simplifies financial analysis by enabling users to identify dominant spending patterns at a glance. The automated classification reduces the need for manual sorting, minimizes human error, and improves interpretability of financial records. These results validate the effectiveness of the proposed framework in transforming unstructured bank statements into structured, decision-ready insights.

Overall, the BFSI Document Automation System demonstrated significant improvements in accuracy, scalability, and analytical capability. By automating extraction, preprocessing, categorization, and visualization, the system reduced human effort, improved consistency, and provided strategic insights. It successfully addressed the challenges of traditional manual processing while offering a flexible, adaptive, and intelligent solution for financial document management and analysis.

V.CONCLUSION

BFSI Document Automation System represents a significant advancement in financial document processing and transaction analysis. Traditional methods of handling bank statements, receipts, and invoices are time-consuming, error-prone, and difficult to scale, often requiring extensive manual effort to extract, categorize, and analyze data. This system successfully addresses these challenges by integrating advanced optical character recognition (OCR), data preprocessing, supervised and semi-supervised categorization, unsupervised LLM-based analysis, API integration, and visualization into a unified digital platform.

The OCR Data Extraction module lays the foundation for the system by converting a wide variety of financial documents into structured text with high accuracy. The preprocessing module ensures that the extracted data is standardized, corrected for OCR errors, and formatted into structured tables suitable for downstream analytics. This transformation from raw document images to clean, machine-readable data drastically reduces manual intervention and provides a reliable input for subsequent processing stages. By addressing challenges such as irregular layouts, multi-column tables, and key-value mapping, the system ensures contextual integrity of financial transactions, which is crucial for accurate categorization and analysis.

Supervised categorization demonstrates the system's ability to reliably classify transactions into predefined categories using labeled datasets. Through effective feature engineering, model training, and iterative optimization, the system achieves high classification accuracy, enabling financial institutions and individual users to monitor cash flows, identify spending patterns, and manage accounts efficiently. Semi-supervised techniques complement this by leveraging unlabeled data, expanding the model's learning capability without requiring exhaustive human annotation. This approach proves especially useful for handling diverse vendors, varying transaction formats, and large-scale datasets, making the system adaptable to dynamic banking environments.

The unsupervised LLM-based analysis module adds significant intelligence by uncovering hidden patterns, emerging categories, and anomalies in transaction data. By transforming transaction descriptions into vector embeddings and applying clustering techniques, the system can identify unusual or potentially fraudulent activities, providing proactive insights for auditing, compliance, and risk management. The ability of LLMs to understand semantic context ensures that even ambiguous or complex entries are interpreted accurately, enhancing overall reliability and supporting continuous improvement of transaction categorization.

API integration ensures seamless interoperability with external banking systems, enterprise applications, and visualization platforms. Automated ingestion, processing, and retrieval of data in real time reduce operational delays, improve workflow efficiency, and maintain security and compliance through encryption, authentication, and role-based access controls. The Visualization and Trend Analysis module transforms processed data into actionable insights via interactive dashboards, charts, and reports. Users can monitor expenditure trends, identify anomalies, track recurring transactions, and derive strategic business intelligence, all in an intuitive and accessible manner.

In conclusion, the BFSI Document Automation System offers a comprehensive, scalable, and intelligent solution for modern financial document management. By automating extraction, preprocessing, categorization, and visualization, the system significantly reduces manual workload, improves data accuracy, and provides actionable insights. Its hybrid approach, combining supervised, semi-supervised, and unsupervised techniques, ensures adaptability to diverse datasets and real-world banking scenarios. The system empowers financial institutions, accountants, and individual users to efficiently process large volumes of transactions, make informed decisions, detect anomalies, and maintain accurate digital records. Overall, this automation framework not only addresses the inefficiencies of traditional methods but also establishes a robust foundation for future advancements in intelligent financial document processing, analytics, and decision support

VI REFERENCES

[1] Xu, Y., Xu, Y., Lv, T., Cui, L., Wang, F., Lu, Y., … & Florencio, D. Xu, Y., Xu, Y., Lv, T., Cui, L., Wang, F., Lu, Y., Florencio, D., Zhang, C., & Wei, F. (2020). LayoutLM: Pre-training of text and layout for document image understanding. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 1192–1200.

[2] Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., & Wei, F. (2021). LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2579–2591.

[3] Huang, Y., Cui, L., Zhang, C., Li, Y., Wei, F., & Zhou, M. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 3214–3225.

[4] Kim, G., Hong, T., Yim, J., Seo, M., & Park, S. (2022). Donut: Document understanding transformer without OCR. European Conference on Computer Vision (ECCV), 517–533.

[5] Mathew, M., Karatzas, D., & Jawahar, C. (2021). DocVQA: A dataset for VQA on document images. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2200–2209.

[6] Mindee Research Team. (2021). docTR: Document text recognition with deep learning. GitHub Repository. Available: https://github.com/mindee/doctr

[7] Katti, A., Reisswig, C., Guder, C., Brunner, U., Faddoul, J., Madl, M., &

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Höhne, J. (2018). Chargrid: Towards understanding 2D documents. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4459–4469.

[8] Zhong, X., Tang, J., & Yepes, A. J. (2019). PubLayNet: Largest dataset ever for document layout analysis. Proceedings of the IEEE/CVF International Conference on Document Analysis and Recognition (ICDAR), 1015–1022.

[9] Zheng, X., Li, L., Tang, J., & Yepes, A. J. (2020). PubTabNet: Dataset for image-based table recognition. European Conference on Computer Vision (ECCV), 163–179.

[10] Li, M., Zhong, X., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). DocBank: A benchmark dataset for document layout analysis. Proceedings of the 28th International Conference on Computational Linguistics (COLING), 949–960.

[11] Baek, J., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection (CRAFT). Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9365–9374. Smith, R. (2007). An overview of the Tesseract OCR engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR), 629–633.

[12] Huang, Z., Xu, Y., & Xu, D. (2019). ICDAR 2019 robust reading challenge on scanned receipts OCR and information extraction (SROIE). 2019 International Conference on Document Analysis and Recognition (ICDAR), 1516–1520.

[13] Denk, T. I., & Reisswig, C. (2019). OCRMiner: A deep learning pipeline for invoice and receipt text recognition. Proceedings of the 2nd Workshop on Financial Technology and Natural Language Processing (FinNLP), 1–10.

[14] Toran, L., Van Der Walt, C., Sammarone, A., & Keller, A. (2023). Scalable and weakly supervised bank transaction classification. arXiv preprint arXiv:2305.18430.

[15] Liu, J., Schulte, O., & Li, K. (2021). Categorization of financial transactions in QuickBooks. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD), 3250–3258.

[16] Jørgensen, R. K., & Igel, C. (2021). Machine learning for financial transaction classification across companies using character-level word embeddings. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD), 3259–3267.

[17] Lesner, A., Zhu, W., & Krishnan, R. (2019). Large-scale personalized categorization of financial transactions. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 33(01), 9486–9493.

[18] Khan, A., Qasim, S. R., & Shafait, F. (2019). TableNet: Deep learning model for end-to-end table detection and tabular data extraction from scanned documents. Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), 128–133.

[19] Zhong, X., Shafait, F., & Yepes, A. J. (2020). Image-based table recognition: Data, model, and evaluation. Proceedings of the European Conference on Computer Vision (ECCV), 564–580.

[20] Tables and figures. International Journal on Document Analysis and Recognition (IJDAR), 2(1), 20–35.

[21] Harley, A. W., Ufkes, A., & Derpanis, K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval. 2015 International Conference on Document Analysis and Recognition (ICDAR), 991–995.

[22] Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., & Zhang, L. (2020). TableBank: Table benchmark for detection, recognition and structure analysis. Proceedings of The 12th Language Resources and Evaluation Conference (LREC), 1918–1925.

[23] Jaume, G., Ekenel, H. K., & Thiran, J. P. (2019). FUNSD: A dataset for form understanding in noisy scanned documents. 2019 International Conference on Document Analysis and Recognition (ICDAR), 1–6.

[24] Park, C. D., Shin, S., Lee, J., Kim, S., & Lee, H. (2019). CORD: A consolidated receipt dataset for post-OCR parsing. Proceedings of] Göbel, M., & Fink, F. (1999). Information extraction from PDF documents the 2nd Workshop on Document Intelligence at NeurIPS.