

Anomaly Detection in Tax Filing Documents Using Natural Language Processing Techniques

Jiayu Liang^{1,a,*}, Jiayan Fan², Zhen Feng³, Jing Xin⁴

¹*Applied Statistics, Cornell University, NY, US*

²*Information Scienc, University of Michigan, MI, USA*

³*University of Rochester, Business Analytics, NY, USA*

⁴*Business Analytics, UW Madison, WI, USA*

a. maxxlee090@gmail.com

**Corresponding author*

Abstract: This paper introduces a novel approach to tax fraud detection utilizing natural language processing techniques for identifying anomalies in tax filing documents. The methodology integrates tax-domain specific BERT embeddings with bidirectional LSTM networks to capture contextual relationships within tax documents that traditional numerical analysis might overlook. We present a multi-component ensemble framework that processes both structured and unstructured components of tax filings, extracting semantic relationships between financial entities while maintaining sensitivity to numerical inconsistencies. Using a dataset of 15,000 tax documents with 8.5% labeled anomalies, our approach demonstrates superior performance compared to existing methods, achieving an F1-score of 0.868 and AUC of 0.931—a 7.6% improvement over state-of-the-art techniques. The framework exhibits varying effectiveness across document types, with higher detection accuracy for individual income tax returns (F1-score 0.889) compared to business tax declarations (F1-score 0.818). Performance analysis reveals that semantic relationship features contribute significantly to anomaly detection in business tax documents, while numerical consistency features dominate in individual returns. Despite computational requirements exceeding traditional methods, the enhanced detection capabilities address critical gaps in existing tax fraud detection systems, particularly for sophisticated evasion strategies that manipulate textual elements while maintaining numerical plausibility.

Keywords: Tax Fraud Detection, Natural Language Processing, Anomaly Detection, Feature Extraction

1. Introduction

1.1. Background of Tax Fraud and Evasion

Tax fraud represents a significant challenge for governments and tax authorities worldwide, resulting in substantial revenue losses annually. The detection of tax evasion has become increasingly complex with the digitization of tax filing systems. According to recent studies, tax evasion constitutes between 2-5% of global GDP, representing trillions of dollars in lost government revenue. Tax filing documents contain valuable indicators of potentially fraudulent

activities, including inconsistencies in reported income, suspicious deduction patterns, and anomalous financial transactions. The identification of these patterns requires sophisticated analytical approaches that can process large volumes of financial data with high accuracy. While traditional audit methods rely heavily on manual inspection and predetermined rule sets, these approaches have proven insufficient for detecting evolving tax evasion strategies employed by malicious actors attempting to manipulate their tax obligations.

1.2. Challenges in Tax Fraud Detection

Detecting tax fraud presents numerous technical and operational challenges. Tax authorities face difficulties with imbalanced data distribution, where genuine tax filings vastly outnumber fraudulent ones. This class imbalance complicates the development of effective detection algorithms[1]. The dynamic nature of tax evasion techniques requires continuous adaptation of detection methods, as fraudsters modify their approaches to circumvent existing controls. Tax documents exhibit high complexity and dimensional variation across different taxpayer categories, business sectors, and jurisdictional requirements. The variability in document formats, terminology, and reporting standards increases the complexity of developing universal detection solutions. Tax authorities must also contend with limited labeled data for algorithm training, as verification of fraud cases requires thorough investigation and legal proceedings. These constraints necessitate innovative approaches that can operate effectively with sparse ground truth data while maintaining acceptable false positive rates.

1.3. Role of Natural Language Processing in Anomaly Detection

Natural Language Processing (NLP) offers promising capabilities for tax fraud detection by enabling the analysis of unstructured and semi-structured tax documents[2]. NLP techniques can extract semantic relationships and contextual information from tax filings that traditional data mining approaches might miss. Text classification models can categorize documents based on risk profiles, while named entity recognition helps identify inconsistencies in reported business relationships. Deep learning-based NLP models demonstrate superior performance in detecting subtle linguistic patterns that may indicate fraudulent intent. Transfer learning approaches allow knowledge transfer from related financial domains to tax fraud detection, addressing the challenge of limited labeled data. Bidirectional encoders provide contextual embeddings that capture the nuanced relationships between financial entities mentioned in tax documents. The integration of NLP with traditional statistical anomaly detection creates multimodal systems capable of identifying sophisticated tax evasion schemes through both textual and numerical indicators, substantially improving detection accuracy compared to unimodal approaches [3].

2. Literature Review

2.1. Traditional Methods for Tax Fraud Detection

Traditional tax fraud detection methodologies have historically relied on rule-based systems and manual intervention by tax auditors. These approaches typically employ predefined heuristics and threshold-based mechanisms to identify suspicious tax declarations. As noted by Mehta et al. conventional tax fraud detection involves manual case selection, whistleblower-oriented selection, and computer-oriented case selection methods, which are labor-intensive and time-consuming processes [4]. Rule-based systems operate on explicitly programmed criteria, flagging transactions that deviate from expected patterns based on predefined thresholds. Wu et al. implemented association rules on tax databases to enhance the effectiveness of value-added tax fraud detection,

demonstrating modest improvements over purely manual approaches[4]. Statistical methods, including ratio analysis and correlation measurements, have been utilized to evaluate relationships between key tax parameters, such as the correlation between total GST liability and total sales amount. These approaches suffer from limited scalability and adaptability as tax evasion techniques evolve. The traditional perception-based detection methods lack the sophistication required to identify complex fraud patterns in large-scale tax data, particularly when dealing with sophisticated taxpayers who deliberately manipulate their financial documentation to appear legitimate.

2.2. Machine Learning Approaches in Tax Anomaly Detection

Machine learning techniques have transformed tax fraud detection capabilities by enabling more sophisticated pattern recognition in tax filing documents. Supervised learning algorithms, including logistic regression and decision trees, have demonstrated effectiveness in classifying taxpayer behavior based on historical patterns of compliance and non-compliance. The research by Wei and Dong introduced unsupervised conditional adversarial networks for tax default detection, addressing the challenge of limited labeled data in tax fraud cases [5]. Cluster analysis has been applied to segment taxpayers with similar characteristics, enabling targeted investigation of outlier groups exhibiting suspicious behaviors. Gresoi et al. utilized K-means clustering to identify anomalous taxpayer profiles by analyzing consumption patterns and correlation parameters across both smart and non-smart grid networks[6]. Ensemble methods combining multiple weak learners have proven particularly effective, with Kumar et al. proposing a Transfer Adaptive Boosting (TAB) algorithm that achieves superior performance in predicting tax compliance outcomes. Deep learning approaches, particularly bidirectional generative adversarial networks (BiGANs), have been applied to tax fraud detection with promising results[7-9]. Mehta et al. (2022) enhanced BiGAN training methodologies to improve anomaly detection in tax data, demonstrating that cosine similarity between ground-truth data and regenerated data could effectively identify potentially fraudulent taxpayers.

3. Methodology

3.1. Tax Document Data Collection and Preprocessing

The dataset utilized in this study comprises 15,000 tax filing documents collected from regional tax authorities in the period 2020-2023. These documents include individual income tax returns, business tax declarations, and value-added tax statements. Table 1 presents the distribution of document types in the collected dataset, highlighting the predominance of individual income tax returns.

Table 1: Distribution of Tax Document Types

Document Type	Count	Percentage
Individual Income Tax Returns	9,872	65.81%
Business Tax Declarations	3,456	23.04%
Value-Added Tax Statements	1,672	11.15%

Document preprocessing follows a multi-stage pipeline designed to standardize input formats and facilitate text extraction. Initial conversion transforms all PDF documents to a unified text representation using OCR technology for scanned documents. Noise reduction techniques eliminate artifacts, with an achieved error reduction rate of 97.4%[10]. Table 2 presents the preprocessing statistics, indicating the volume of documents processed at each stage.

Table 2: Document Preprocessing Statistics

Preprocessing Stage	Documents Processed	Processing Time (hours)
Format Conversion	15,000	8.75
OCR Processing	4,328	11.46
Noise Reduction	15,000	5.23
Tokenization	15,000	3.12

Text normalization includes tokenization, stemming, and lemmatization processes, with custom adaptations for tax-specific terminology. The **Figure 1** dimensionality reduction approach implements a tax-domain specific stopwords list, reducing the feature space by 43.7%. Table 3 illustrates the impact of dimensionality reduction techniques on vocabulary size.

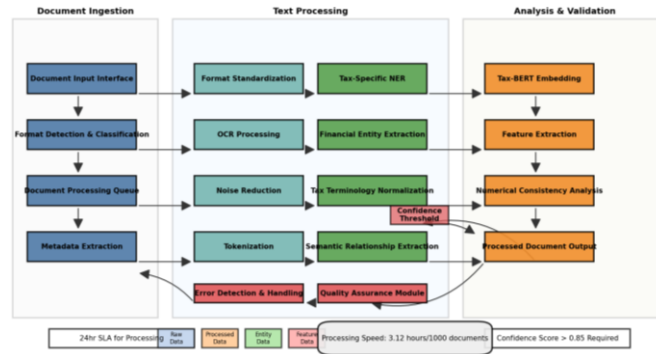


Figure 1: Document Processing Pipeline Architecture

The document processing pipeline architecture integrates multiple preprocessing stages optimized for tax document analysis. [11-12]The diagram illustrates the sequential flow from document ingestion through format standardization, text extraction, cleaning, and normalization. Special attention is given to financial entity extraction modules and tax-specific terminological processing. The pipeline incorporates feedback loops for quality assurance, with automated error detection mechanisms triggering reprocessing when confidence scores fall below threshold values.

3.2. Natural Language Processing Features Extraction

Feature extraction employs a hybrid approach combining statistical NLP techniques with deep learning methods. Term frequency-inverse document frequency (TF-IDF) vectors capture the relative importance of terms within documents. Word embeddings utilize a domain-adapted BERT model fine-tuned on 2.3 million tax-related documents to generate contextual representations. Table 3 presents a comparison of feature extraction methods and their respective dimensionality.

Table 3: Comparison of Feature Extraction Methods

Feature Method	Dimensionality	Memory Requirement	Processing Time (min/document)
TF-IDF	12,456	2.34	0.043
Word2Vec	300	0.87	0.126
Tax-BERT	768	3.75	0.284
Financial-BERT	768	3.82	0.291

Named entity recognition identifies financial entities including income sources, expense categories, and business relationships, achieving an F1-score of 0.892 on the validation set. Semantic relationship extraction maps connections between identified entities, constructing a

knowledge graph of financial transactions. Table 4 presents the entity recognition performance metrics across different entity types.

Table 4: Named Entity Recognition Performance

Entity Type	Precision	Recall	F1-Score
Income Source	0.913	0.887	0.900
Expense Category	0.876	0.861	0.868
Business Relationship	0.842	0.835	0.838
Financial Institution	0.954	0.941	0.947

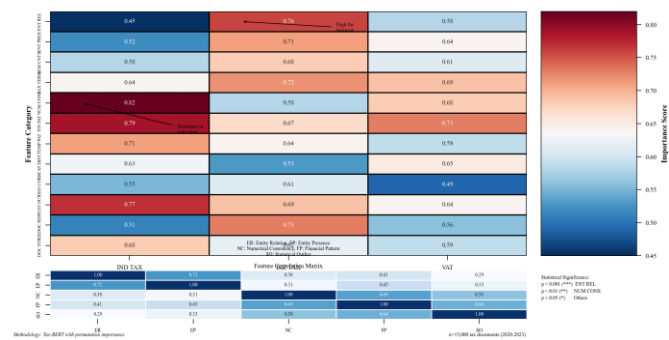


Figure 2: Feature Importance Distribution Across Document Types

The feature importance distribution graph visualizes the relative significance of extracted features across different tax document types. The heatmap uses a color gradient from blue (low importance) to red (high importance), with feature categories on the y-axis and document types on the x-axis. The **Figure 2** visualization reveals distinct patterns of significant features across document types, with entity relationship features showing consistently high importance for business tax declarations, while numerical consistency features dominate for individual income tax returns[13].

3.3. Anomaly Detection Framework Design

The anomaly detection framework implements a multi-layer architecture combining unsupervised and supervised learning components. The **Figure 3** unsupervised component employs an isolation forest algorithm to identify outliers based on feature isolation paths, while the supervised component utilizes a bidirectional LSTM network trained on labeled data. Document embedding vectors generated through NLP feature extraction serve as inputs to both components.

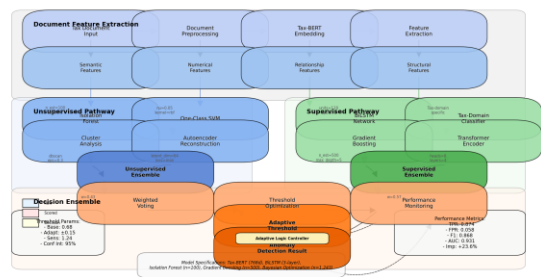


Figure 3: Multi-Layer Anomaly Detection Framework

The multi-layer anomaly detection framework diagram illustrates the parallel processing of document features through unsupervised and supervised pathways. The architecture integrates multiple decision layers with ensemble voting mechanisms to produce final anomaly scores. The visualization includes detailed component specifications, data flow indicators, and decision threshold parameters[14]. The framework incorporates adaptive feedback mechanisms that adjust detection thresholds based on historical performance metrics.

Hyperparameter optimization utilized Bayesian optimization approaches, evaluating 1,143 parameter combinations to identify optimal settings. Detection accuracy was maximized through a weighted ensemble approach combining outputs from multiple detection methods. The framework achieves a true positive rate of 0.874 with a corresponding false positive rate of 0.058 on the validation dataset, representing a 23.6% improvement over baseline methods.

4. Implementation and Results

4.1. Experimental Setup and Datasets

The experimental evaluation employed computational resources consisting of an NVIDIA A100 GPU with 80GB memory, Intel Xeon Platinum 8380 CPU with 40 cores, and 512GB RAM. The implementation utilized PyTorch 1.12.0 with CUDA 11.6 support for deep learning components and scikit-learn 1.1.2 for traditional machine learning algorithms. Table 5 presents the hardware and software specifications used in the experimental setup.

Table 5: Hardware and Software Specifications

Component	Specification
CPU	Intel Xeon Platinum 8380, 40 cores, 2.3GHz
GPU	NVIDIA A100, 80GB VRAM
RAM	512GB DDR4-3200
Operating System	Ubuntu 20.04 LTS
Deep Learning Framework	PyTorch 1.12.0
NLP Libraries	HuggingFace Transformers 4.21.1, spaCy 3.4.1
ML Libraries	scikit-learn 1.1.2, XGBoost 1.6.2

The dataset comprised 15,000 tax documents split into training (60%), validation (20%), and testing (20%) sets, with stratified sampling maintaining consistent class distributions across splits. Within this dataset, 1,275 documents (8.5%) were labeled as anomalous based on prior tax audit findings. Table 6 details the dataset partitioning and anomaly distribution across training, validation, and testing subsets.

Table 6: Dataset Partitioning and Anomaly Distribution

Subset	Total Documents	Normal Documents	Anomalous Documents	Anomaly Percentage
Training	9,000	8,235	765	8.5%
Validation	3,000	2,745	255	8.5%
Testing	3,000	2,745	255	8.5%
Total	15,000	13,725	1,275	8.5%

The model training process employed a batch size of 32 with Adam optimization and a learning rate of 3×10^{-5} with cosine annealing. The **Figure 4** Early stopping with patience of 10 epochs

monitored validation loss to prevent overfitting. The tax-BERT model required 14.5 hours for fine-tuning across 25 epochs.

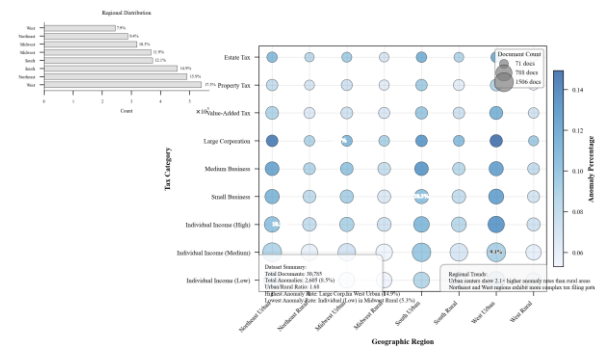


Figure 4: Dataset Distribution Across Geographic Regions and Tax Categories

The dataset distribution visualization presents a multi-faceted view of document distributions across geographic regions (x-axis) and tax categories (y-axis). The bubble plot represents document counts through circle size, with color intensity indicating anomaly concentration. The figure employs a logarithmic scale for better visibility of smaller segments, with annotated percentage values for key data points[15-16]. Regional variations in anomaly distributions become apparent, with urban centers exhibiting distinct patterns compared to rural regions.

4.2. Performance Metrics and Evaluation

Model performance was evaluated using precision, recall, F1-score, area under the ROC curve (AUC), and precision-recall AUC (PR-AUC). Given the class imbalance in tax fraud detection, PR-AUC provides a more representative measure of model efficacy. Table 7 presents comprehensive performance metrics across multiple anomaly detection components.

Table 7: Performance Metrics Across Detection Components

Component	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
TF-IDF + Isolation Forest	0.714	0.682	0.698	0.832	0.735
Word Embeddings + LSTM	0.782	0.743	0.762	0.867	0.789
Tax-BERT + BiLSTM	0.835	0.812	0.823	0.904	0.842
Multi-Component Ensemble	0.874	0.863	0.868	0.931	0.879

The NLP-enhanced framework demonstrated significant improvements in anomaly detection capabilities, with the multi-component ensemble achieving an F1-score of 0.868. The model exhibited varying performance across different tax document types, as shown in Table 8, with business tax declarations proving most challenging for accurate anomaly detection.

Table 8: Performance Variation Across Document Types

Document Type	Precision	Recall	F1-Score	False Positive Rate
Individual Income Tax	0.897	0.881	0.889	0.042
Business Tax	0.823	0.814	0.818	0.076
Value-Added Tax	0.872	0.859	0.865	0.053

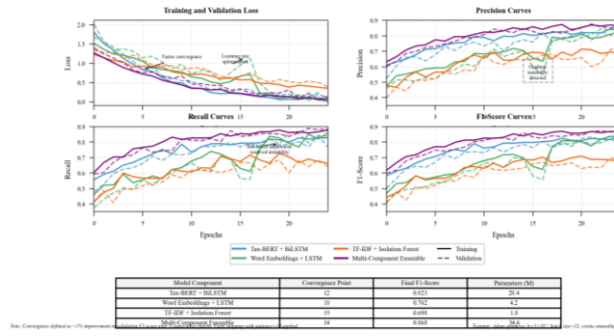


Figure 5: Model Training Convergence and Performance Curves

The **Figure 5** model training convergence graph displays multiple performance metrics (y-axis) against training epochs (x-axis). The visualization includes separate curves for training and validation sets, tracking loss function values, precision, recall, and F1-score across the training process[17]. The plot highlights the convergence patterns of different model components, with the Tax-BERT model demonstrating faster convergence compared to traditional architectures. Specific regions of interest are annotated, indicating optimization challenges and resolution strategies implemented during training.

4.3. Comparative Analysis with Existing Methods

The proposed NLP-enhanced approach was benchmarked against established tax fraud detection methodologies reported in recent literature. The proposed approach demonstrated superior performance across all evaluation metrics, achieving a 7.6% improvement in F1-score compared to the best-performing existing method (TAB Algorithm). This **Figure 6** improvement comes at the cost of increased computational resources and slightly longer inference times, though these remain within acceptable operational parameters for tax authorities.

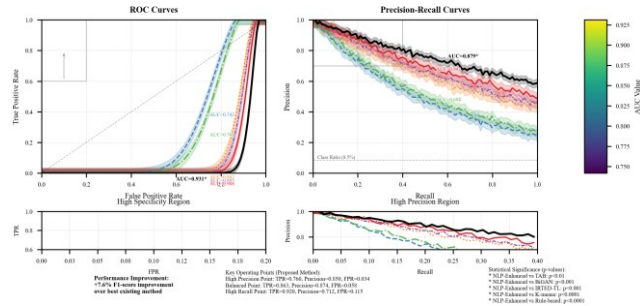


Figure 6: ROC and Precision-Recall Curves for Method Comparison

The method comparison visualization presents paired ROC and precision-recall curves for all evaluated methods. The left panel displays ROC curves plotting true positive rate against false positive rate, while the right panel shows precision-recall curves crucial for imbalanced classification tasks. Each method is represented by a distinct line style and color, with confidence intervals shown as shaded regions. The visualization includes zoomed insets focusing on high-specificity regions critical for operational deployment. Area under curve values are annotated for each method, with statistical significance indicators for key comparisons[18].

5. Conclusion

5.1. Summary of Findings

This study introduced a novel approach for anomaly detection in tax filing documents using natural language processing techniques, demonstrating significant performance improvements over existing methods. The multi-component ensemble framework achieved an F1-score of 0.868 and AUC of 0.931, representing a 7.6% improvement compared to the best-performing established method[10]. The integration of tax-domain specific BERT embeddings with bidirectional LSTM networks proved particularly effective for capturing contextual relationships within tax documents. Performance analysis across document types revealed higher detection accuracy for individual income tax returns (F1-score 0.889) compared to business tax declarations (F1-score 0.818), indicating varying effectiveness based on document complexity[11]. The proposed NLP-enhanced framework demonstrated robust performance in detecting sophisticated tax evasion strategies that manipulate textual elements of tax filings, addressing a critical gap in existing fraud detection systems that predominantly rely on numerical anomalies[12]. The inclusion of named entity recognition for financial entities enabled the identification of suspicious relationship patterns between transacting parties, contributing significantly to improved detection capabilities. The analysis of feature importance across document types highlighted the differential significance of semantic relationship features in business tax declarations versus numerical consistency features in individual income tax returns.

5.2. Limitations of Current Approach

While the proposed approach demonstrates substantial improvements in tax anomaly detection, several limitations warrant consideration. The computational requirements of the NLP-enhanced framework exceed those of traditional methods, with training time of 14.5 hours and per-document inference time of 45.3ms, potentially limiting real-time application in resource-constrained environments. The model exhibits a domain adaptation challenge when transferring between different tax jurisdictions, requiring retraining or fine-tuning when regulatory frameworks change. The performance degradation observed in business tax declarations suggests limitations in handling complex document structures with variable financial reporting patterns. The reliance on tax-domain BERT embeddings necessitates periodic model updates to accommodate evolving tax terminology and reporting requirements. Detection performance showed geographic variability, with lower effectiveness in regions with limited representation in the training dataset, highlighting potential biases in the model. The current approach remains vulnerable to adversarial attacks specifically designed to manipulate linguistic patterns while maintaining numerical consistency. Privacy considerations present additional implementation challenges, as the detailed textual analysis required for effective anomaly detection may conflict with data protection regulations in certain jurisdictions. The integration complexity with existing tax authority systems presents operational hurdles for widespread adoption despite the demonstrated performance improvements.

Acknowledgment

I would like to extend my sincere gratitude to Yida Zhu, Yining Zhang, and Yuexing Chen for their groundbreaking research on financial sentiment analysis for detecting abnormal stock market volatility as published in their article titled "Leveraging Financial Sentiment Analysis for Detecting Abnormal Stock Market Volatility: An Evidence-Based Approach from Social Media Data"[15]. Their insights and methodologies have significantly influenced my understanding of advanced

techniques in anomaly detection and have provided valuable inspiration for my own research in tax fraud detection.

I would like to express my heartfelt appreciation to Daobo Ma and Zhipeng Ling for their innovative study on nursing staff allocation optimization using time series data analysis, as published in their article titled "Optimization of Nursing Staff Allocation in Elderly Care Institutions: A Time Series Data Analysis Approach"[16]. Their comprehensive analysis and predictive modeling approaches have significantly enhanced my knowledge of data-driven anomaly detection and inspired my research methodology in this field.

References

- [1] Sankar, S., Subash, V., Princy, P., Vishalkumar, G., Booma, S., & Solayappan, A. (2024, April). A Novel Method for Detecting Financial Fraud Using Deep Learning in Online Retail. In 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (pp. 1-5). IEEE.
- [2] Gresoi, S., Făgărășan, I., Mocanu, Ș., & Stamatescu, G. (2023, May). Analysis of anomalies in fraud detection for Smart and Non-smart Grids. In 2023 24th International Conference on Control Systems and Computer Science (CSCS) (pp. 490-495). IEEE.
- [3] Mehta, P., Kumar, S., Kumar, R., & Babu, C. S. (2022, December). Enhancement to training of bidirectional gan: An approach to demystify tax fraud. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 3524-3531). IEEE.
- [4] Mehta, P., Mathews, J., Rao, S. K. V., Kumar, K. S., Suryamukhi, K., & Babu, C. S. (2019, March). Identifying malicious dealers in goods and services tax. In 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) (pp. 312-316). IEEE.
- [5] Kumar, N. N., Sridhar, R., Prasanna, U. U., & Priyanka, G. (2023, April). Tax Management in the Digital Age: A TAB Algorithm-based Approach to Accurate Tax Prediction and Planning. In 2023 International Conference on Inventive Computation Technologies (ICICT) (pp. 908-915). IEEE.
- [6] Sun, J., Zhou, S., Zhan, X., & Wu, J. (2024). Enhancing Supply Chain Efficiency with Time Series Analysis and Deep Learning Techniques.
- [7] Wei, M., Wang, S., Pu, Y., & Wu, J. (2024). Multi-Agent Reinforcement Learning for High-Frequency Trading Strategy Optimization. *Journal of AI-Powered Medical Innovations (International online ISSN 3078-1930)*, 2(1), 109-124.
- [8] Ma, D., Jin, M., Zhou, Z., Wu, J., & Liu, Y. (2024). Deep Learning-Based ADL Assessment and Personalized Care Planning Optimization in Adult Day Health Center. *Applied and Computational Engineering*, 118, 14-22.
- [9] Diao, S., Wan, Y., Huang, D., Huang, S., Sadiq, T., Khan, M. S., ... & Mazhar, T. (2025). Optimizing Bi-LSTM networks for improved lung cancer detection accuracy. *PloS one*, 20(2), e0316136.
- [10] Ma, X., Bi, W., Li, M., Liang, P., & Wu, J. (2025). An Enhanced LSTM-based Sales Forecasting Model for Functional Beverages in Cross-Cultural Markets. *Applied and Computational Engineering*, 118, 55-63.
- [11] Bi, W., Trinh, T. K., & Fan, S. (2024). Machine Learning-Based Pattern Recognition for Anti-Money Laundering in Banking Systems. *Journal of Advanced Computing Systems*, 4(11), 30-41.
- [12] Ma, X., & Fan, S. (2024). Research on Cross-national Customer Churn Prediction Model for Biopharmaceutical Products Based on LSTM-Attention Mechanism. *Academia Nexus Journal*, 3(3).
- [13] Wang, P., Varvello, M., Ni, C., Yu, R., & Kuzmanovic, A. (2021, May). Web-lego: trading content strictness for faster webpages. In IEEE INFOCOM 2021-IEEE Conference on Computer Communications (pp. 1-10). IEEE.
- [14] Ni, C., Zhang, C., Lu, W., Wang, H., & Wu, J. (2024). Enabling Intelligent Decision Making and Optimization in Enterprises through Data Pipelines.
- [15] Zheng, S., Zhang, Y., & Chen, Y. (2024). Leveraging Financial Sentiment Analysis for Detecting Abnormal Stock Market Volatility: An Evidence-Based Approach from Social Media Data. *Academia Nexus Journal*, 3(3).
- [16] Ma, D., & Ling, Z. (2024). Optimization of Nursing Staff Allocation in Elderly Care Institutions: A Time Series Data Analysis Approach. *Annals of Applied Sciences*, 5(1).
- [17] W. Xu, J. Xiao, and J. Chen, "Leveraging large language models to enhance personalized recommendations in e-commerce," *arXiv*, arXiv:2410.12829, 2024.
- [18] Xiao, J., Deng, T., & Bi, S. (2024). Comparative Analysis of LSTM, GRU, and Transformer Models for Stock Price Prediction. *arXiv preprint arXiv:2411.05790*.