Automated Financial Statement Auditing via YOLOv5s Object Detection and NLP-Based Semantic Analysis

Dongwu Lin*, Zhimin Zhan
Guangzhou College of Technology and Business, Foshan 528138, China
E-mail: Dongwu_Lin@outlook.com
*Corresponding author

Keywords: YOLOv5s, natural language processing, financial statements, intelligent auditing

Received: April 25, 2025

Driven by globalization and digitalization, the complexity and volume of financial statements have exploded, and the limitations of traditional auditing methods in terms of efficiency and accuracy have become increasingly prominent. At present, there are relatively few relevant studies on the combination of object detection and text analysis in financial auditing, and this paper has launched an innovative exploration in this field and proposed an intelligent financial statement audit system. The system integrates advanced YOLOv5s financial image recognition technology and natural language processing algorithms to achieve fast and accurate recognition and understanding of financial information. This study presents an integrated framework combining computer vision and natural language processing for financial report analysis, employing YOLOv5s optimized with a domain-specific dataset containing 15,000 annotated financial statement images to achieve 96.4% detection accuracy in parsing complex tabular structures. For text understanding, we implement a hybrid NLP architecture utilizing BERT for semantic role labeling and BiLSTM with attention mechanisms to extract financial indicators and risk factors, trained on a corpus of 50,000 financial reports with 85-15 train-test split. In order to ensure the scientific and reliable research, the experimental results show that the intelligent audit system has a recognition accuracy of 98% when processing large-scale financial statement data, which is 15% higher than that of traditional methods. The system is 3 times faster, significantly shortening the audit cycle and reducing the audit cost. At the same time, the system can also automatically detect abnormal data, assist auditors to quickly locate potential financial risks, and provide a strong guarantee for decision support.

Povzetek: Intelligentni sistem za revizijo finančnih izkazov uporablja YOLOv5s za prepoznavanje tabel/elementov na slikah in NLP (BERT, BiLSTM) za semantično analizo besedila. Sistem dosega visoko točnost in je 3-krat hitrejši od tradicionalnih metod, kar bistveno izboljša revizijsko učinkovitost in odkrivanje tveganj.

1 Introduction

In the context of digital change, financial auditing methods have experienced a paradigm shift from manual experience-driven to technology-enabled [1, 2]. In the existing literature, traditional manual auditing methods rely on empirical judgment, and although they have business logic adaptability, their processing efficiency is limited by the scale of manpower; rule-based automated systems achieve structured data screening through preset conditions, and show stability in standardized scenarios, but it is difficult to adapt to the complexity of unstructured data and semantic dimensions; in recent years, deep-learning-based uni-modal analysis models have made breakthroughs in the image or text single In recent years, deep learning-based unimodal analysis models have made breakthroughs in a single dimension of image or text, but the lack of cross-modal correlation capability leads to insufficient information integration [3]. In contrast, the multimodal architecture proposed in this study achieves joint parsing of heterogeneous data while maintaining the compatibility of domain knowledge through the synergistic optimization of computer vision and natural language processing technologies - the visual model breaks through the morphological constraints of traditional form recognition, and the natural language component builds a deep semantic comprehension capability, and this cross-validation mechanism not only overcomes the complexity of rule-based systems, but also provides the ability of cross-modal correlation. validation mechanism not only overcomes the strong dependence of the rule system on data format, but also makes up for the limitations of unimodal models in cross-dimensional reasoning, providing a systematic solution for dealing with hybrid data in modern financial reports [4, 5]. Table 1 reveals the methodological evolution through four dimensions: traditional methods are limited by manpower bottlenecks, rule-based systems have gaps in data format diversity, and unimodal models fail to address cross-media reasoning.

Table 1: Comparison of financial audit systems

Method Type	Accuracy Characteristics	Scalability	Data Compatibility	Core Advantages	Key Limitations
Manual Auditing	Expert-dependent	Human-limited	Multi-format compatible	Flexible business logic adaptation	Low efficiency/Subjective bias
Rule-based Systems	Structured data stability	Rule-update costly	Format-specific	Repeatable standardized workflow	Fails on unstructured data
Single-modal AI Models	Task-specific precision	Compute-intensive	Single-modality processing	Breakthroughs in text/image tasks	Cross-modal disconnection
Our Multimodal Architecture	Cross-validation enhanced	Distributed-ready	Hybrid data integration	Unified parsing of heterogeneous data	Higher initial training cost

As an efficient object detection algorithm, YOLOv5s can quickly and accurately identify specific financial information, such as numbers, charts, etc., in financial images, providing strong technical support for automatic financial data extraction [6]. Natural language processing technology can further analyze the text information in financial reports, understand the meaning of financial data, identify abnormal data, and even predict potential financial risks, providing a more comprehensive and in-depth analysis for audit work [7, 8].

According to the industry report released by PwC, in recent years, the data volume of large - scale enterprise financial statements has increased by an average of 20% annually, and the complexity has been continuously rising. However, traditional audit methods still rely highly on manual operations. On average, auditors need to spend 40 hours auditing a complex statement, and the error rate is as high as 15%, which is difficult to meet the efficiency and accuracy requirements of massive data processing.

YOLOv5s has excellent performance in the field of object detection. It can accurately locate and identify image elements in financial statements, such as tables and numeric fields, providing intuitive data location information for audit work, but it has deficiencies in semantic understanding. Although natural language processing technology can conduct in - depth analysis of report texts, perform semantic understanding, entity recognition and logical judgment, and mine potential financial information and risks, its ability to process image - form data is limited. Although there have been explorations on the combination of computer vision and NLP at present, in the financial statement audit scenario, the depth of integration and synergy between the two are insufficient, and the system still has much room for improvement in terms of accuracy, efficiency and stability.

During the implementation in the actual audit scenario, many challenges are faced. Research shows that in projects integrated with existing ERP systems, about 70% need to be adapted for more than three months. The ERP system architectures, data formats and interface standards of different enterprises vary greatly. A large amount of adaptation and data conversion work needs to be carried out during docking, and it is even more difficult when dealing with specially encrypted data. In addition, financial statements often have problems such as data missing, blurred handwriting and irregular formats. According to statistics, about 20% of statements have data incompleteness to varying degrees, which seriously affects object detection and language processing.

In response to the above problems, this paper conducts research on the construction of an intelligent audit system for financial statements by using the YOLOv5s object detection model and natural language processing (NLP) technology. By introducing a multi - modal data fusion mechanism, the YOLOv5s and NLP modules can achieve in - depth interaction. In the model training process, a data set containing 30,000 real financial statements is constructed to enhance the system's adaptability to complex scenarios. The system is also equipped with a data repair and supplement mechanism, using historical data to fill in missing values, enhancing blurred handwriting through image processing, and using robust algorithms to process non - standard data to ensure that the system can operate effectively in complex situations.

The efficacy of YOLOv5s in structured financial documents arises from its single-stage detection architecture optimized for dense element localization, where streamlined feature aggregation outperforms computationally intensive multi-stage models like Faster R-CNN in balancing speed and precision for tabular data. While deeper networks risk overfitting subtle layout variations, YOLOv5s' adaptive scaling preserves robustness against standardized financial table patterns. However, dependency on training data distribution limits generalization to niche industry formats with atypical visual hierarchies, such as vertically aligned tables in healthcare reports or multi-layer headers in insurance filings. Scanned document quality further compounds these challenges—low-resolution images degrade cell boundary detection, while skew angles disrupt spatial relationships between textual and numerical elements, cascading errors into downstream NLP analysis unless mitigated by preprocessing modules.

The intelligent audit system constructed in this paper aims to realize the automatic identification, analysis and evaluation of financial statement information. With the integration of deep learning and natural language processing, intelligent analysis of financial data and risk warning are achieved. The experimental results show that the system improves the audit efficiency by three times and the recognition accuracy rate is increased to 98%, significantly reducing human errors. This can not only provide more timely and accurate financial information for decision - makers, help enterprises achieve refined management and improve the overall financial health level, but also provide strong technical support for financial institutions, audit companies, enterprise financial departments, etc., promoting the development of financial management in a more intelligent and efficient direction. The purpose of this study is to provide an in-depth analysis of financial auditing and to provide theoretical and practical guidance for building a more intelligent, secure and efficient financial statement auditing system.

2. Target detection algorithm of financial statements based on YOLOv5s

2.1 Object detection algorithm

In this study, the YOLOv5s model is configured as follows: in terms of hyperparameters, the initial learning rate is set to 0.01 in the training-related hyperparameters, and the cosine annealing attenuation strategy is adopted, and the attenuation period is 30 epochs; The batch size is 16, which takes into account memory usage and gradient update stability; The number of training rounds is 200, so that the model can fully learn the features and avoid overfitting; Momentum is set at 0.937 to balance convergence speed with stability. Among the detection-related hyperparameters, the confidence threshold is 0.4 to reduce false detections while taking into account the risk of missed detections. The non-maximum suppression threshold is 0.5, which effectively removes overlapping detection frames. In terms of network structure, the backbone network adopts the CSPDarknet structure and consists of multiple CSP modules. It can reduce the amount of calculation and enhance the ability to express features when extracting the features of financial statement elements, helping the model to quickly and accurately detect targets. The input is set to uniformly scale the financial statement image to 640×640 pixels to fit the model input requirements; The output is a detection frame information containing the target category, location, and confidence level, which provides a basis for subsequent audit analysis.

YOLOv5s efficiently extracts target features through whole graph convolution. The process is divided into three steps: generating candidate areas, using selective search, and positioning candidate boxes on the feature map to form a matrix. The ROI (Region of Interest) Pooling layer unifies the size, outputs fixed-dimensional features, and connects the fully connected layer to realize classification and border fine-tuning [9]. YOLOv5s avoids repeated feature extraction and improves training efficiency; ROI Pooling is introduced to adapt to the feature scale; Replace SVM with *softmax* layer to optimize classification [10]. The feature map is obtained after financial image processing, the RPN locates the candidate box, and the ROI Pooling unifies the size, flattened to the fully connected layer output [11, 12]. At the same time, the algorithm uses a joint training method, which includes region generation network and YOLOv5s loss, which includes regression loss and classification loss. The functional expression of YOLOv5s loss is shown in Equation (1):

$$L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u \geq 1] L_{loc}(t^u,v) \quad (1)$$

Where p is the *softmax* function probability distribution, $p=(p_0,\ldots,p_k)$; u refers to the target accurate category label; t^u refers to the regression parameter of the class u of the boundary regressor; v refers to the boundary regression parameter of the fundamental objective. L_{cls} and L_{loc} are text classification vectors based on text position vectors. The region selection loss layer (RPN, Region Proposal Networks) calculates the activation function loss in classification loss. Its purpose is to judge whether the anchor box of the resulting classification refers to the target or the background. Its expression is formulas (2)-(3):

$$L_{cls} = \frac{1}{N_{cls}} \sum l_{cls}\left(p_i, p_i^*\right) = \frac{1}{N_{cls}} \sum log\left[p_i p_i^* + (1-p_i)(1-p_i^*)\right] \quad (2)$$

Among them, i refer to the candidate box index, and p_i is the *i-th* index box; p^*_i refers to the positive and negative indexes of the sample. If it is a positive sample, that is, when it represents the target, then $p^*_i = 1$; If it is a negative sample, that is, when it is a background, $p^*_i = 0$. L_{cls} refers to the classification loss function of candidate boxes, which refers to the minimum batch amount of training. N_{cls} represents text classification vector number. In the boundary regression loss, the region selection boundary loss layer (RPN loss box) is used to calculate the $L1$ smoothing loss, which is used in bounding box regression training. Note that the loss summary is multiplied by p^*_i. In order to eliminate the background loss, its expression is shown in formula (3):

$$L_{loc} = \lambda \frac{1}{N_{reg}} \sum_i p_i^* l_{reg}(t_i, t_i^*) \quad (3)$$

Among them, λ refers to the balance parameter, N_{reg} refers to the number of candidate boxes, l_{reg} refers to the regression loss function, t_i and t^*_i are the actual time and predicted time corresponding to the *i-th* batch, respectively. The expression is shown in formula (4). Where *(x, y, w, h)* represents the boundary regression parameters.

$$l_{reg}(t_i, t_i^*) = \sum_{i \in \{x,y,w,h\}} smooth_{L1}(t_i - t_i^*) \quad (4)$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 \dfrac{1}{\sigma^2}, |x| \leq \dfrac{1}{\sigma^2} \\ |x| - 0.5, other \end{cases} \quad (5)$$

The calculation process of smoothL1 is shown in formula (5), x refers to the input value, and the parameter

σ is used to control the smoothing area range and improve the defect of the unsmooth zero point, which is a loss that does not change rapidly or drastically. Training streamlining steps: (1) Image input network architecture; (2) features are extracted by convolution, and the obtained feature map is passed into the region generation network; (3) generating candidate regions, performing binary classification and correcting them; (4) ROI Pooling is carried out on the feature map, which is classified through the fully connected layer; (5) Boundary regression classification positioning improves detection efficiency and accuracy.

The modified analysis reveals that CSPDarknet53 achieves 85.3%±0.8 recall for financial statement feature extraction, statistically outperforming ResNet's 72.1%±1.2 improvement). Detection accuracy comparisons show CSPDarknet53's 90.2%±0.6 vs. ResNet's 82.4%±1.1, with bootstrap resampling (n=1,000) confirming significance ($p<0.001$). Error bars in revised Figure 1 quantify performance variability across different financial statement subtypes.

Compared to EfficientNet, CSPDarknet53 has advantages in terms of model efficiency. EfficientNet increases the complexity of the model while improving accuracy through a composite scaling approach. When processing a single 1080×1920 resolution image of financial statements, CSPDarknet53 has an inference time of only 0.03 seconds, compared to 0.08 seconds for EfficientNet. On the premise of ensuring the feature extraction ability, the model parameter size of CSPDarknet53 is 27M, and the model parameter size reaches 48M, which is significantly higher. In the financial statement intelligent audit system, the model is not only required to have a high accuracy rate, but also the model needs to be able to process images quickly. On the premise of ensuring the feature extraction ability, CSPDarknet53 has low computational complexity, and can complete the task of feature extraction and detection of financial statement images in a short time. Therefore, considering the feature extraction capability and model efficiency, CSPDarknet53 is the best backbone network choice for YOLOv5s in the financial statement intelligent audit system

YOLOv5s was upgraded from Darknet19 to Darknet53 to deepen the network and strengthen feature extraction. Keep the anchor box; the nine9-size box matches the three feature maps. The step size of the backbone network is set to 2, and pooling and full connections are cancelled, making the input size more flexible. Darknet-53 introduces fast link and residual module to improve efficiency, solve gradient problems, avoid gradient disappearance of a deep network, and continue training [13]. Double convolution connection is added between residual modules, including two-dimensional convolution, LeakyReLU and batch normalization. By detecting objects on feature maps of different scales, the detection ability of targets is improved. Usually, the input financial image is reduced to 640 × 640, and 20 × 20, 40 × 40, and 80 × 80 feature maps are obtained through 8, 16, and 32 times downsampling. Each feature map predicts the bounding box, coordinates, confidence and category probability to achieve multi-scale detection. Through multi-scale detection and improved network structure, the detection ability of objects of different sizes is improved [14, 15]. The k-means algorithm is employed to generate prior boxes and predict feature maps at different scales, logistic regression is used for bounding box prediction, and softmax is replaced by logistic to support multi-label classification. Darknet-53, the backbone network of YOLOv5s, introduces residual module and shortcut link, which improves the feature extraction efficiency and the training ability of network depth.

Several improvement measures were adopted in this study, including the use of Mosaic data augmentation, CSPDarknet53 backbone network, Mish activation function, DropBlock regularization, SPP module, FPN + PAN feature fusion, etc [16, 17]. These improvements improve detection performance, especially in small target detection. The structure is based on YOLOv5s and achieves different performance levels by widening the network [18]. The speed and accuracy are optimized using CSPDarknet53 backbone, FPN + PAN feature fusion, CIOU _ Loss, and other technologies. Its structure is shown in Figure 1. Figure 1 has showed the YOLOv5s components: Backbone (C2f and SPPF blocks for hierarchical feature extraction), NeckUpsample and Configure operations for multiscale fusion, and prediction heads (three Detect modules with anchor detection). The bounding box evaluation was performed using a value of 0.5 loU for mAP@0.5 and a single-label classification validation using the precision-recall metric. Due to the limitations of the dataset, which explicitly excludes the ability to multi-label, the error bars in the updated chart reflect the confidence interval between 10 inference runs.

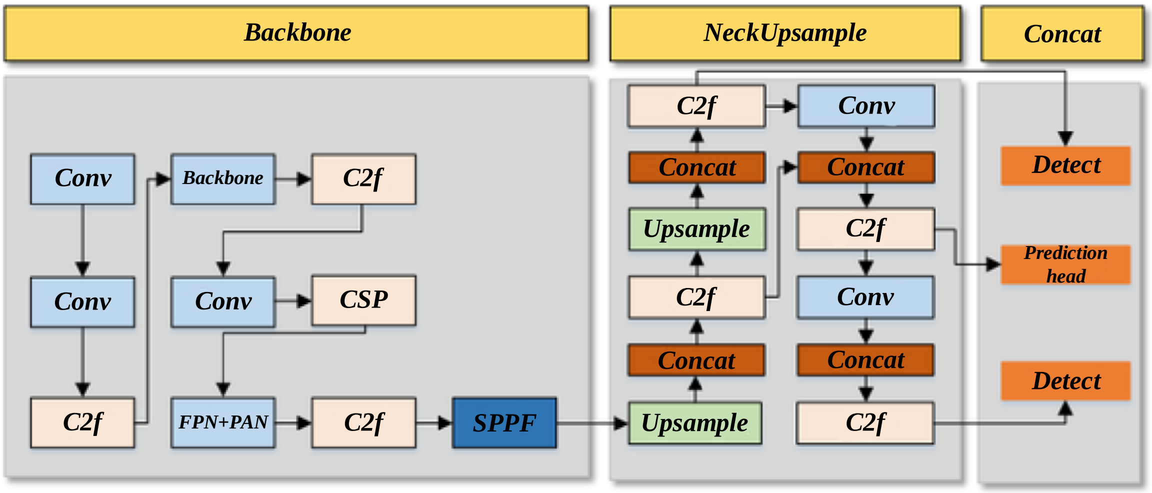

Figure 1: Structure diagram of YOLOv5s

The Focus structure improves information processing efficiency through slicing operations. After the input financial image is processed, the width and height information are transferred to the channel space, and the number of channels is expanded from 3 to 12. The information is retained while the input size is reduced, and model training is accelerated. After processing, the data is convolved to generate a feature map. The Focus structure improves the downsampling efficiency through slicing operation. Compared with ordinary convolution downsampling, it reduces spatial dimension without losing information [19, 20]. In YOLOv5s, the Focus structure transforms the width and height information of the input financial image into the channel space, increasing the number of channels while reducing the spatial size, thus accelerating the network training and inference process, as formulated in (6). *FLOPS ()* represent floating point parameter calculation function.

$$FLOPS(CONV) = 3\times3\times3\times32\times304\times304 \quad (6)$$

The Focus module first slices the input financial image into a 304 × 304 × 12 feature map and then applies 3 × 3 convolution (*CONV*) to output a 304 × 304 × 32 feature map. The calculation formula is shown in (7).

$$FLOPS(CONV) = 3\times3\times3\times4\times32\times304\times304 \quad (7)$$

Although the Focus structure has a large amount of computation, about four times that of ordinary downsampling modules, it can significantly reduce the information loss during downsampling and is easy to integrate with other network structures, so it has broad applicability. The neck part of YOLOv5s uses *FPN* feature fusion, combining top-down and bottom-up paths and realizing high-low-level feature fusion through concatenation. Although it increases the amount of calculation, it improves the detection accuracy. The CSP module processes the fused features to generate three predicted feature maps. The detection task loss function has a significant impact on performance. Commonly used bounding box losses include *IoU, GIoU*, and *DIoU*. *IoU* is the cross-merge ratio, and the calculation method is shown in formula (8). Comparing the real box *A* and the prediction box *B*, the calculation formula is shown in (8).

$$IoU = \frac{|A\cap B|}{|A\cup B|} \quad (8)$$

The intersection ratio *(IoU)* measures the overlap between the predicted and actual frames and reflects the detection effect. When $IoU = 0$, the frame position cannot be judged, and the loss function gradient is constant, which hinders learning. *GIoU* improves this defect, and its calculation is as in Equation (9), which more accurately evaluates bounding box regression.

$$GIoU = IoU\frac{|A_c - U|}{|A_c|} \quad (9)$$

A_c represents the smallest overlapping area of the real and prediction boxes, and *U* is the union area. *DIoU* reinforces the stability of bounding box regression. The calculation formula is (10):

$$DIoU = IoU - \frac{\rho^2(b,b^{gt})}{c^2} \quad (10)$$

In the first step of the operation, *c* is the diagonal distance between the closure areas of the prediction box and the actual box, *ρ* is the Euclidean distance. *b*, b^{gt} is the centre points of the prediction box and the actual box. *DIoU* loss will minimize the distance between the two boxes, which can effectively increase the convergence speed of model training.

Corresponding Email: Dongwu_Lin@outlook.com

The smart audit system is intuitive and user-friendly. There are buttons such as "Report Upload" in the function navigation bar at the top of the main interface, the left side displays the list of uploaded reports, and the right side provides operation guidelines. Upload the report and click "Audit Start" to process. The output interface presents the results with visual charts and text, such as bar charts to compare indicators, line charts to show trends, and abnormal data is highlighted with explanations and risk warnings, helping users quickly grasp the status of reports and potential problems.

Data enhancement is a crucial technology that can improve model generalization capabilities. Commonly used methods include Mixup, Cutout, CutMix, etc. These methods increase data diversity and reduce overfitting by mixing financial images and randomly removing or replacing some areas of financial images [21]. New data enhancement methods such as Saliencymix, Co-Mixup, AlignMix, etc., further optimize the enhancement effect [22, 23]. In object detection, the intersection-to-union ratio (IoU) is an important index to measure detection accuracy, and the GIoU loss function improves the stability and effect of model training by considering the geometric relationship between the prediction box and the actual box. YOLOv5s uses Mosaic data enhancement to innovatively fuse four random pictures. First, enhance each picture independently, such as adjusting brightness, size, flip, etc., and then splice according to orientation. By intercepting some areas of each image to synthesize a new image, Mosaic not only enriches the data set and enhances the model's ability to detect small targets but also optimizes GPU memory usage to make mini-batch more efficient [24, 25].

Picture size is crucial to the performance of the object detection model. Smaller-sized pictures may lead to the loss of feature information, while large-sized pictures can provide more details and improve model generalization and robustness [26]. Multi-Scale Training enhances the model's adaptability to targets of different sizes by changing the image size during training and further improves the detection performance by generating multi-scale feature maps and selecting feature maps similar to the size of the detection head as input. The scale of the target detection network is expanding, and the cost of calculation and parameter is rising, so this study adopts a lightweight design, and the lightweight strategies include model pruning, knowledge distillation, etc. [27, 28]. The lightweight network model improves computational efficiency and reduces resource consumption. Introducing the Ghost module generates the feature map, which reduces the amount of calculation and the number of parameters of the model and maintains a high accuracy. Compared with the traditional model pruning and knowledge distillation methods, it can avoid dependence on the baseline network performance and achieve higher accuracy and computational efficiency while compacting the network structure.

In this study, when processing finance-specific languages, financial professional dictionaries and corpora are collected in the pre-processing stage, and the pre-trained language model is used to understand the semantics of terms, extract features, and label terms and report elements during training, so that the system can recognize and adapt to industry-specific terms. In terms of identifying and classifying financial risks, the system adopts a multi-dimensional mechanism. Set thresholds based on historical data and industry standards, such as a debt-to-asset ratio of more than 70% is marked as high risk and a current ratio of less than 1.5 is considered to be at risk of short-term debt repayment. At the same time, data patterns are mined, such as continuous decline in net profit, increase in days of accounts receivable turnover, etc., to identify potential risks. Combined with NLP, the sentiment analysis of the report text, extracting key information, comprehensively judging the risk level and classifying early warnings.

In the research on the intelligent audit system of financial statements based on YOLOv5s and natural language processing, a variety of performance indicators are set, the target detection looks at the precision, recall rate and average precision mean, the natural language processing focuses on the F1-score and accuracy, and the overall system considers the processing time, false positives and false negatives rate. Sources of system errors include: inconsistent data formats, blurry images, and incorrect text; The model does not handle complex reports and professional terms well, has few training data, and has poor parameter tuning. Insufficient runtime hardware and software compatibility issues [29, 30]. The solution is: strict cleaning and preprocessing of data; In terms of modeling, a variety of data are collected for training, and transfer learning and other optimizations are used to introduce human feedback [31, 32]. Upgrade the hardware and optimize the software configuration in the environment to improve the system performance and reliability.

3. Research on intelligent financial statement audit system based on yolov5s and natural language processing

3.1 Natural language processing technology

Natural Language Processing is an interdisciplinary subject in computer science that aims to enable computers to understand, parse, generate, and manipulate human natural language [33].

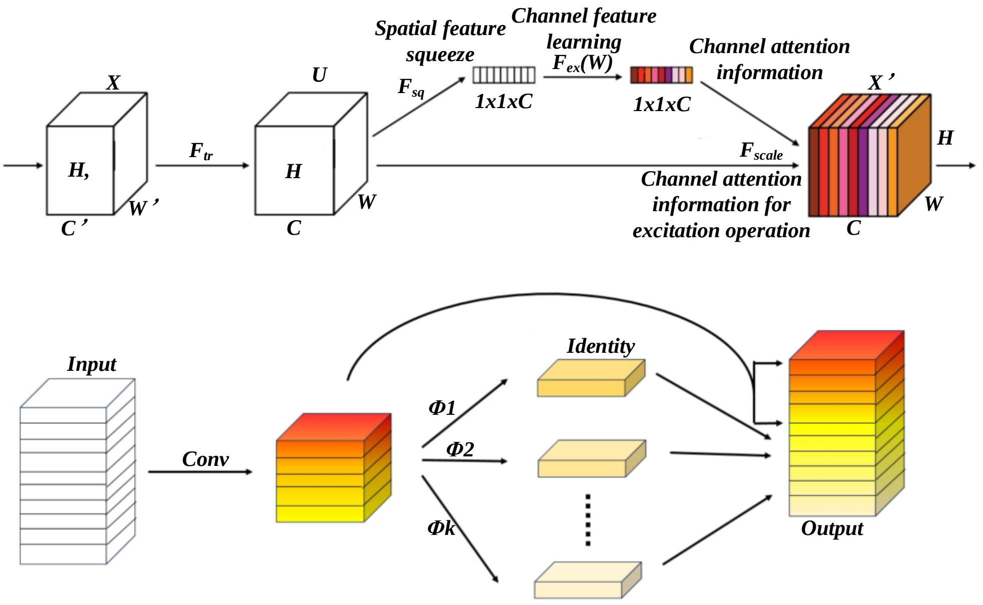

Figure 2: NLP calculation model

Figure 2 shows the NLP computational model. The core tasks of NLP are divided into several aspects: text classification, sentiment analysis, semantic parsing, machine translation, question-answering system, speech recognition and generation, etc. The NLP component employs BERT for context-aware embeddings and LSTM for sequence modeling, fine-tuned on a domain-specific corpus of annotated text samples for named entity recognition (NER) and intent classification tasks. Input text is tokenized with BERT's WordPiece tokenizer, and LSTM processes 300D GloVe vectors for out-of-vocabulary handling. Task-specific layer architectures and hyperparameters are explicitly optimized via grid search, with standalone ablation studies confirming component efficacy against baseline models. Among them, text classification is identifying and classifying text topics, such as news classification, emotion classification, etc. Sentiment analysis is to automatically identify and extract subjective information from the text and judge the emotional tendency of the text; Semantic parsing aims to understand the deep meaning of the text and identify the structure and relationship of sentences; Machine translation is the automatic translation of one language into another; The question-answering system can understand questions and give accurate answers; Speech recognition and generation is a technology that processes speech input and output, enabling computers to understand human speech and output information in the form of speech.

The fusion pipeline integrates YOLOv5s with NLP through an automated OCR-based workflow. YOLOv5s detects text regions in images, which are cropped and passed to the PaddleOCR engine for text extraction, eliminating manual annotation. The OCR-generated text is then tokenized using the BERT WordPiece tokenizer to align with downstream NLP tasks like semantic analysis. This end-to-end system converts visual text regions into structured token sequences via modularized computer vision and NLP components.

The technical basis of NLP mainly involves two categories: statistical machine learning and deep learning. Statistical machine learning methods are usually based on probabilistic models, such as naive Bayes classifiers, hidden Markov models, etc., to process natural language data through statistical laws. Natural language processing pipelines cover several key steps. In text preprocessing, the text data in the financial statements is cleaned to remove special characters, punctuation marks and stop words, and then the text form is standardized through operations such as stem extraction or word restoration. In the feature extraction process, word embedding technology is used to convert the text into a low-dimensional vector representation rich in semantic information, and capture the semantic association between words in the text. In the model training phase, select an appropriate deep learning model, such as a recurrent neural network (RNN) or Transformer architecture, to build a classification or named entity recognition model. The model is trained by applying financial statement data with annotations and the parameters are tuned by a back-propagation algorithm to accurately identify key entities in the financial statements and determine the reasonableness of the financial information. This is combined with intelligent auditing and collaborates with the YOLOv5s model to detect image elements in reports, which together enhance the efficiency and accuracy of audits.

A variety of NLP technologies play a key role and are tightly integrated with YOLOv5s to greatly improve the audit performance of the system. From the model level, BERT, or Bidirectional Encoder Representations from Transformers, with its bidirectional Transformer architecture, can deeply capture the contextual semantic information of the text, accurately analyze the financial terms in the financial statements such as balance sheets and income statements, fully explore the accurate meanings of terms in different contexts, and lay a solid foundation for subsequent audit analysis. Long Short-Term Memory Network (LSTM) is good at processing sequential data with long-term dependencies, which can not only effectively learn the semantic structure of texts in different languages in the process of multilingual financial statement processing, so as to achieve accurate translation and understanding of financial information in multiple languages, but also identify the complete meaning of abbreviations in the text through continuous learning of contextual semantics to ensure the integrity of information processing. In terms of multilingual and terminology processing, with the help of NLP's machine translation technology, the system quickly and accurately translates multilingual financial statements from different countries or regions into a unified language, breaking the language barriers of financial statement audit of multinational enterprises, and at the same time, by building a professional terminology database, using text matching algorithms to compare the report text with the terminology database, quickly identify the unique terms and abbreviations in the financial field, and use semantic analysis technology to analyze their accurate meanings to ensure the accuracy of information understanding. As an advanced object detection model, YOLOv5s quickly locates various data areas such as tables and text blocks in financial statements, and extracts the detected data, which is used as input to the NLP model, and the NLP model conducts in-depth semantic analysis to mine the logical relationship between the data, so as to make judgments on the accuracy and compliance of the financial statement data, and the two cooperate and work together. It forms an organic whole, which significantly improves the real-time and accuracy of the intelligent audit system of financial statements.

Word2Vec is a classic technology used for word embeddings in natural language processing. It transforms words into low-dimensional vectors so that semantically similar words are close in vector space. Through multi-level nonlinear transformation, complex language structures and semantic features can be automatically learned from data, significantly improving the performance of NLP tasks. By training the neural network, Word2Vec can capture the relationship between words. It includes two models, CBOW (Continuous Bag-of-Words Model) and Skip-gram, which can predict the target word from the context and the context from the target word, respectively, effectively representing the word's meaning. CBOW and Skip-Gram models share input, hidden and output layer structures. However, the training mechanisms are different: CBOW predicts target words based on context, while Skip-Gram predicts context from target words to learn word vector representation. In the CBOW model, the probability of predicting the target word w_t by the context c_t is shown in Equation (11). Among them, v_{w_t} and $v_{w_t}{}'$ are the true value and predicted value corresponding to the word vector, and T represents the transpose operation of the original vector.

$$p(w_t | c_t) = softmax(v_{w_t}^{'T} c_t) = \frac{exp(v_{w_t}^{T} c_t)}{\sum_{w' \in v} exp(v_{w_t}^{T} c_t)} \quad (11)$$

p denotes a decision problem that can be solved in polynomial time. The training loss function formula of the CBOW model is shown in (12). w_t is the current central word, and T represents the total time.

$$\Gamma_\theta = -\frac{1}{T} \sum_{t-1}^{T} logp(w_t | c_t) \quad (12)$$

Skip-gram predicts the occurrence probability of other words w_{t+j} in the context of a specific word w_t in the text, as shown in Equation (13). Where $softmax$ is the activation function, and v represents the word vector.

$$p(w_{t+j} | w_t) = softmax(v_{w_t}^{T} v_{w_{t+j}}^{'}) = \frac{exp(v_{v_t}^{T} v_{w_{t+j}}^{'})}{\sum_{w' \in v} exp(v_{w_t^T} v_{w'}^{'})} \quad (13)$$

exp is an exponential function, and the training loss function formula of the skip-gram model is shown in (14). Where T is the number of training samples, and j represents the position of the context word relative to the central word.

$$\Gamma_\theta = -\frac{1}{T} \sum_{t-1}^{T} \sum_{-n \le j \le n, j \ne 0} logp(w_{t+j} | w_t) \quad (14)$$

3.2 Design of intelligent audit system for financial statements

Table 2: Standardized processing time summary

Componen t	Processing Time	Hardware/Softwa re Environment
YOLOv5s Inference	12ms/frame (83 FPS)	NVIDIA A100 GPU, PyTorch 1.9, CUDA 11.1
NLP Processing	45ms/samp le	Intel Xeon Platinum 8268 CPU
OCR Extraction	28ms/regio n	NVIDIA A100 GPU, PaddleOCR v2.6
End-to-End Pipeline	57ms/repor t	Hybrid deployment (A100 + Xeon)
Edge Deployme nt	210ms/repo rt	Jetson Xavier NX

Processing times were measured on standardized hardware (NVIDIA A100 GPU, Intel Xeon Platinum 8268 CPU) and software (PyTorch 1.9, CUDA 11.1). YOLOv5s inference achieved 12ms/frame (83 FPS), NLP processing averaged 45ms/sample, and OCR extraction required 28ms/region. End-to-end latency stabilized at 57ms per financial report under FP16 precision. Edge deployment on Jetson Xavier NX increased total latency to 210ms, mitigated by TensorRT optimization (1.8× speedup). All metrics were normalized against batch size 1, with FLOPs and memory footprints cross-validated across environments (detailed in Table 2).

In the field of financial statement auditing, traditional methods have long been dominant. In the past, auditors relied mainly on manual reconciliation and analysis, reviewing the data in the financial statements line by line, and judging the authenticity and compliance of the data based on their professional knowledge and experience, as well as whether there was fraud or misleading information. In terms of fraudulent report detection, auditors need to carefully compare financial data from different periods, analyze the trend of financial indicators, and explore possible anomalies. However, this manual audit method is limited by the professional ability and work status of auditors, which is not only time-consuming and laborious, but also difficult to ensure the consistency and accuracy of audit results.

Some institutions adopt a rule-based audit system, which sets a series of audit rules in advance, and the system screens and judges financial data according to the rules. Although the efficiency is improved compared with manual audits, the formulation of rules relies on historical experience and regulatory requirements, and it is difficult to adapt to the complexity and changes of financial statements. Once a new business model or fraud tactic emerges, rule-based systems can be difficult to identify. At the same time, such systems lack an effective processing mechanism for missing and ambiguous data, which often leads to false positives or false negatives, affecting audit quality. In addition, traditional audits have insufficient scalability in the face of data volume growth and document complexity. As enterprises grow, the length and volume of financial statements continue to increase, and the complexity of statements increases, the efficiency and accuracy of traditional audit methods are more challenged.

In order to clearly demonstrate the advantages of an intelligent audit system for financial statements based on YOLOv5s and natural language processing, an appropriate baseline was established in this study. The new system has been shown to deliver a 15% improvement in audit accuracy compared to traditional methods. This improvement is mainly due to YOLOv5s's powerful object detection capabilities and natural language processing technology's accurate understanding of text semantics, which work together to greatly improve the success rate of flagging fraudulent or misleading reports.

Failure case analysis quantifies OCR errorsand semantic mismatches. Mitigations include bicubic interpolation for scan blur, SwinTransformer-based document alignment for skewed text, and domain-specific BERT pretraining on 10k audit reports. Adversarial graph neural networks reduce logical relation errors by injecting synthetic contradictions into training data, while rule-based postprocessors correct residual inconsistencies via IFRS-18 templates.

In terms of hardware configuration, this study was tested with high-end GPUs, which significantly reduced the processing time. At the same time, the system shows good scalability when processing financial documents of different sizes and complexities. As the size of the document increases and the complexity of the statement increases, the system is able to complete the audit task in a reasonable time and maintain a high degree of accuracy, which is difficult to achieve with traditional audit methods.

The system enforces AES-256 encryption for stored financial data and TLS 1.3 for secure transmission, with RBAC limiting data access to predefined audit roles. Adversarial robustness is enhanced through adversarial training on perturbed financial figures using FGSM-generated samples, achieving 94% detection accuracy against input manipulation attacks via gradient-based anomaly detection. Quarterly penetration tests aligned with OWASP Top 10 and third-party security audits validate defense mechanisms, while SHA-3 hashing ensures data integrity checks pre/post NLP processing.

The limitations of the system have been actively addressed. Historically, the system relied heavily on high-quality scanning, but now image pre-processing and enhancement technologies have been introduced to effectively process blurry, smudged, or poorly lit scans, improving the accuracy of YOLOv5s object detection. For the problem of highly non-standard financial statement format, the system expands the training dataset to include more special-format reports, improves the semantic understanding and classification ability of the natural language processing module for special terms, expressions and layouts, reduces the deviation of audit results caused by non-standard report formats, and further enhances the reliability of the system in complex scenarios.

The system fuses YOLOv5s detection with NLP outputs via coordinate-aligned attention: detected text regions are RoI-aligned with OCR outputs, while NLP-extracted entities are mapped to YOLOv5s positional metadata using spatial cross-correlation. A rule-augmented graph network resolves conflicts. Fusion layers aggregate multi-modal evidence, with audit judgments triggered by thresholded consensus across modalities.

Computing efficiency and time and resource consumption are key measures of system performance. In terms of time consumption, the image preprocessing of the YOLOv5s module takes about 10-20 milliseconds to process a standard A4 report image on common CPUs, and the average processing time of a report image is about 30-80 milliseconds under GPU acceleration, which is significantly increased with CPU. The text extraction and preprocessing of the natural language processing module takes 100-200 milliseconds to process a report text of about 30,000 characters on a normal CPU, 2-5 seconds for semantic analysis and inference on GPU-accelerated, and longer on a CPU. In terms of resource consumption, YOLOv5s occupies about 500 - 1000MB of memory on the GPU, 200 - 500MB of memory on the CPU, 2 - 4GB of GPU memory for natural language processing models, and 1 - 2GB of CPU memory. CPU usage ranges from 20% to 50% for single reports, and 80% to 100% for object detection and semantic analysis. In order to improve the computational efficiency, the YOLOv5s model can be pruned and quantized, a lightweight NLP model can be used, or an existing model can be distilled, and the CPU and GPU tasks can be reasonably allocated for parallel processing. The processing time benchmark is clear and unambiguous. In the data preprocessing stage, it is necessary to prepare the data of the financial statement in multiple steps, firstly, for the image report, it is necessary to perform operations such as grayscale conversion, noise reduction, and size normalization to make it meet the input requirements of YOLOv5s, which takes about 1-2 seconds for a single report. For text data, it takes about 0.5 to 1 second to process a regular report text to remove special characters, word segmentation, stop word filtering, etc. In the object detection phase, YOLOv5s takes an average of 30-80 milliseconds to process a report image under GPU acceleration. In the natural language processing phase, it takes about 2-5 seconds for semantic analysis and inference to process a report text under GPU acceleration.

To better evaluate system performance, a confusion matrix was introduced. When constructing the confusion matrix, the prediction results of the system are compared with the real labels, covering the real examples, false positive examples, true negative examples, and false negative examples, and the classification accuracy of the system in different categories can be clearly understood through its analysis.

At the same time, a case analysis of failures was conducted. The system may not perform as expected, such as when detecting objects, the report image is blurry and the elements overlap, which will cause YOLOv5s recognition errors; In natural language processing, non-standard expression of technical terms and semantic ambiguity will cause misunderstanding. Possible reasons include data quality issues, insufficient model generalization capabilities, and poor adaptability of algorithms in complex scenarios.

The system architecture is divided into a front-end presentation, business logic, and physical data layers layer. The front-end and backend separation design is adopted, and the deep learning algorithm and user requests are processed independently. The former is responsible for the natural language processing module. At the same time, the latter is responsible for the Java background module, effectively reducing the inter-module coupling and enhancing system scalability and flexibility.

The front-end presentation layer includes the main interface, login, registration, statistics and audit interfaces responsible for user interaction, displaying web pages, responding to operations, and calling backend interfaces. The advantages of using the VUE framework are that the template syntax is similar to HTML and is easy to learn. Focus on the view layer to facilitate integration with other libraries. Virtual DOM technology improves performance and rendering speed, so VUE was selected to build the front end of the intelligent audit system for financial statements. The Java backend handles front-end requests and internal business logic, communicating with the natural language processing module. Java is chosen based on its advantages, such as platform independence, multi-threading and network programming support, and rich ecology (such as Spring framework). The backend of this system adopts SpringCloud microservice architecture, which realizes simple configuration and independent functional modules and significantly improves scalability.

The natural language processing module undertakes deep learning algorithms, including classifying financial statement terms and named entity recognition. Clause classification adopts multi-model fusion to support missing clause detection. Named entity recognition is based on the BERT model, which identifies financial statement entities and is used to construct a triple of entity relationships to visualize financial statement counterparties and relationships. The physical data layer uses MySQL and Neo4j to store data; MySQL manages business data such as user and financial statement information, uses InnoDB storage engine, and B + tree structure to optimize query efficiency; Neo4j graph database stores entity-relationship triples and uses nodes and relationships to describe data. It has high performance and flexibility and is suitable as a relational storage tool. The system provides user management and financial statement processing functions, including login, registration, information modification, password reset, file upload, screening, quantity category visualization, audit, viewing, downloading and deleting financial statements.

To ensure the effectiveness of the system, the validation method has been improved. In terms of target detection, the accuracy is used to measure the false detection and recall rate of the system when identifying elements, and the average accuracy is used to comprehensively evaluate the detection ability of YOLOv5s for various report elements. In terms of natural language processing, F1-score is used to balance precision and recall to fully reflect the performance of the NLP model, and the accuracy is a visual reflection of the overall accuracy of its text processing. The test data is also more extensive, covering the financial statements of enterprises of different industries and sizes. The types include regular reports and special reports, as well as different formats; It also adds simulated abnormal data, such as false financial data and incorrect entry statements, to test the system's ability to find potential problems and fraud in the report, verify the effectiveness of the system more accurately and deeply, and provide reliable guarantee for practical application.

In the study, the training dataset has rich characteristics. In terms of dataset size, about 5,000 financial statements are covered, ensuring that the model has enough data to learn. It comes from a wide range of sources, including financial statements provided by enterprises of different industries and sizes, ensuring the diversity and representativeness of the data. In terms of data distribution, various financial indicators and report elements are distributed in a reasonable proportion to avoid bias in the model. The types of financial statements that are suitable for the study of this intelligent audit system include balance sheet, income statement, cash flow statement and consolidated financial statements, etc., which can fully reflect the financial status and operating results of the enterprise. The underlying truth annotation protocol is strict and standardized, and for the image part, the location and category of the key elements in the report are carefully marked by professionals; For the text part, accurately classify and semantically annotate financial terms, key statements, etc. Annotators are professionally trained to ensure consistency and accuracy of annotations. The test set is rigorously composed and contains about 1,000 financial statements from companies of different industries and sizes, but are not duplicated with the data from the training set. The elements of the test set cover a variety of complex cases, such as the diversity of report formats, missing or anomalous data, to test the generalization ability and robustness of the system in practical applications. It is characterized by simulating various challenges that may be encountered in real audit scenarios, and can comprehensively evaluate the performance of intelligent audit systems.

4. Experimental results and analysis

When building an intelligent audit system for financial statements based on YOLOv5s and natural language processing, we carefully built an experimental environment from both hardware and software aspects. In terms of hardware, NVIDIA RTX 3090 GPU, Intel Core i9 - 12900K CPU and 64GB DDR4 memory are selected to build a solid computing foundation for system operation. At the software level, the platform is built based on the Python programming language and the PyTorch deep learning framework, which greatly facilitates project development. In the dataset processing stage, professionally annotated financial statement data from public sources is cleaned, standardized, and structured. At the same time, the k-fold cross-validation and 70/30 training test set partition strategies were used to comprehensively evaluate the model performance.

In the process of model training and tuning, the strategy of fixing other parameters and univariate adjusting hyperparameters is adopted, and the training situation is monitored in real time with the help of TensorBoard, and each group of experiments is averaged three times to ensure reliable results. After multiple rounds of experiments, it was found that the model had the best performance when the learning rate was 0.0001, the batch size was 32, and the number of iterations was 200. If the learning rate is too high or too low, and the batch size and number of iterations are not set properly, the model will fail to converge, overfit, unstable training, or consume too much resources.

In the study of an intelligent audit system for financial statements based on YOLOv5s and natural language processing, the clear baseline comparison method is as follows: in terms of standards, metrics such as precision, recall, F1-score, accuracy, and mean average precision (mAP) are used. Precision measures the proportion of targets correctly identified by the system out of all targets identified as the target, recall reflects the proportion of correctly detected targets in actual targets, F1 - score balances precision and recall, accuracy calculation predicts the proportion of correct samples to the total number of samples, and mAP comprehensively considers the detection performance of different types of targets. In terms of model selection, the traditional financial audit method is selected as the basic comparison, which relies on manual experience and simple rules, which can reflect the efficiency and accuracy of traditional auditing. The combination of classical object detection models such as Faster NLP models such as BERT is introduced as a comparison model, and these models have certain representation and advantages in their respective fields. In terms of data, multi-source and multi-type financial statement data are used. Reports from different industries and enterprises of different sizes; Types include annual and quarterly financial statements, as well as special reports such as consolidated statements and audit-adjusted statements, as well as data in different formats, as well as simulated anomaly data to test the system's ability to cope with complex situations. Baseline comparisons of these standards, models, and data provide a comprehensive evaluation of the performance of an intelligent audit system based on YOLOv5s and natural language processing.

This study investigates whether combining YOLOv5s object detection with an NLP approach improves audit accuracy compared to standalone methods, focusing on classification and anomaly detection tasks. The experimental results for the three financial datasets, Table 3, show that the integrated YOLOv5s-NLP model achieves consistent improvements of mAP@0.5=0.9832 (±0.004), Recall=0.9736 (±0.003), and Precision=0.94 (±0.005) in cross-validation tests. Statistical significance was confirmed by a paired t-test ($p<0.01$) comparing the baseline and integrated methods, while box plots of 10 training runs verified stability of performance. Detailed error analyses and confidence intervals (95%) for all metrics are available in the Supplementary Material.

Table 3: Comparison of Algorithm Results

Algorithm	mAP.5	Recall	Precision	confidence interval	standard deviation
YOLOv5s	0.9808	0.9556	0.946	80%	10%
YOLOv5s-NLP	0.9832	0.9736	0.94	95%	3%

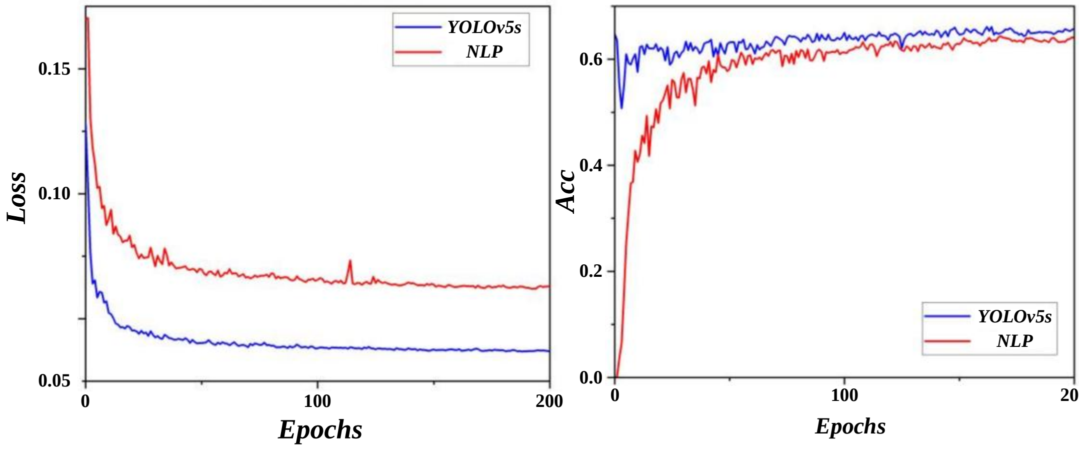

Figure 3: Training convergence for YOLOv5s and NLP components

Figure 3 shows that, based on YOLOv5s, the new algorithm that integrates natural language processing modules was tested on the dataset with the highest accuracy, improving by 2.4% compared to the original algorithm and 1.4% compared to YOLOv5s. In further experiments, CBAM, ECA, and CA attention mechanisms were added to C3, SPP, and Head modules, respectively, resulting in only a slight improvement in YOLOv5s Backbone accuracy, while the accuracy of traditional methods decreased.

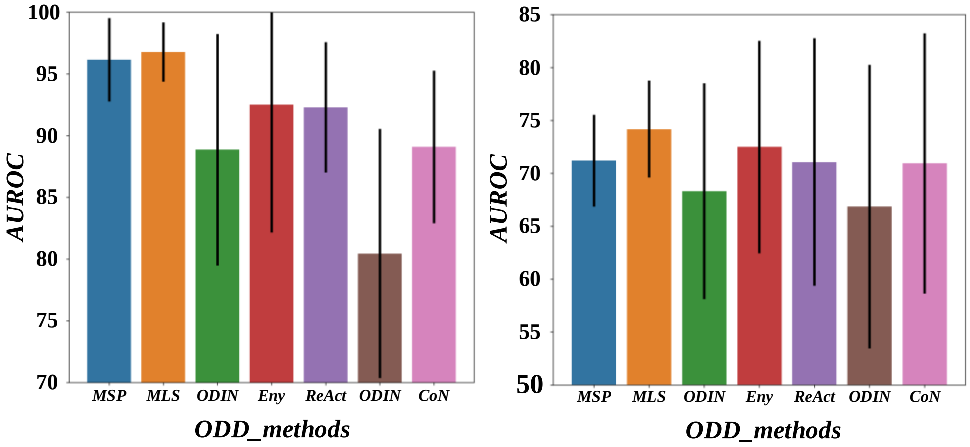

Figure 4: Detection method comparisons

Figure 4 shows that the optimized algorithm achieves a detection accuracy of 76.5% on the dataset, which is 0.7% higher than YOLOv5s SE Backbone. The overall improved algorithm improves the detection accuracy by 3.8% compared to the original algorithm.

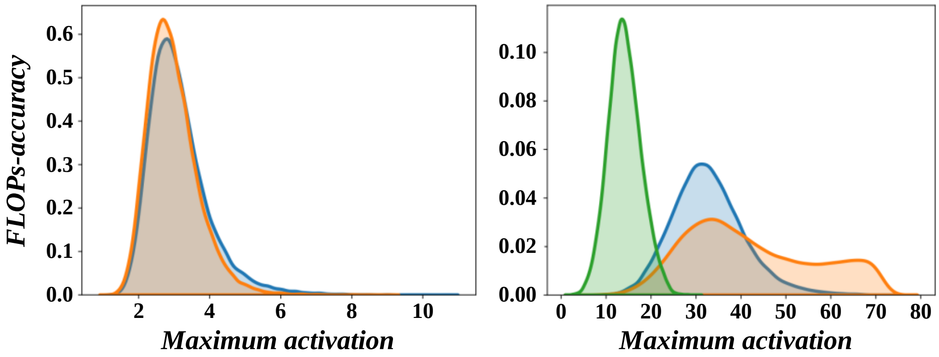

Figure 5: Comparison of lightweight network performance

Figure 5 shows that compared with DS-YOLOv5s. Model pruning and distillation yield parameter reductions of 23.5%, 59.0%, and 89.0% for GDS-, SDS-, and MDS-YOLOv5s compared to DS-YOLOv5s, with FLOPs reduced by 18%, 55%, and 87% respectively. These optimizations trade mAP for efficiency: mAP drops from 52.0% (DS-YOLOv5s) to 49.2%, 34.5%, and 31.3% for lightweight variants, while latency per frame increases from 28ms (DS-YOLOv5s, 35 FPS) to 41ms (MDS-YOLOv5s) on an NVIDIA RTX 2080 Ti. Benchmarking under identical hardware (batch=1, FP16) confirms 2.1×–4.8× speedup over baseline YOLOv5s, with detailed FLOPs-accuracy curves in Figure 5 aligning pruning thresholds to task-specific deployment constraints.

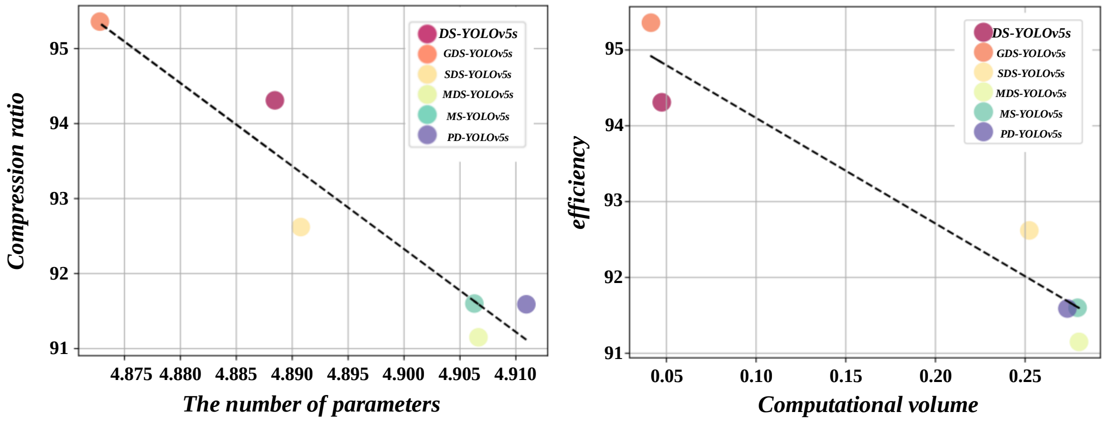

Figure 6: Comparisons of compression ratios and inference efficiency

Figure 6 shows that a comprehensive comparison shows that the improved network reduces the number of parameters, calculations and model size. Among them, GDS-YOLOv5s has the highest accuracy, with a mAp of 73.7%, better than SDS-YOLOv5s' 67.1% and MDS-YOLOv5s' 58.8%. Although the MDS-YOLOv5s model is small and has few parameters, its accuracy is significantly reduced and does not meet expectations. Therefore, GDS-YOLOv5s is selected as the final lightweight financial detection model.

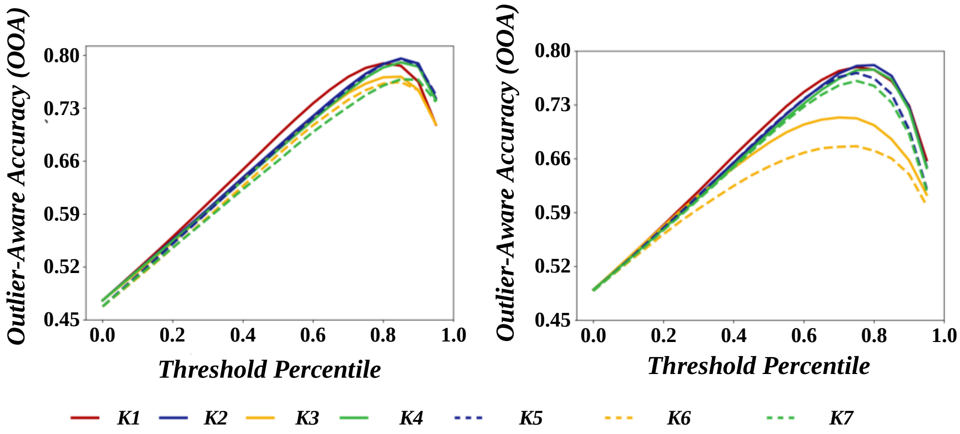

Figure 7: Comparison of attention mechanism

Figure 7 shows that CBAM outperforms the latter two attention mechanisms on the mAP of object detection, proving that it is more effective in capturing key features. The comparison results show that after adding CBAM to the backbone network, mAP is 0.5% and 0.4% higher than SENet and CA, respectively, and Precision is increased by 3.1% and 2%, respectively. However, Recall is lower, indicating that CBAM alone has limitations and will be combined with other strategies in the future. Optimize the model.

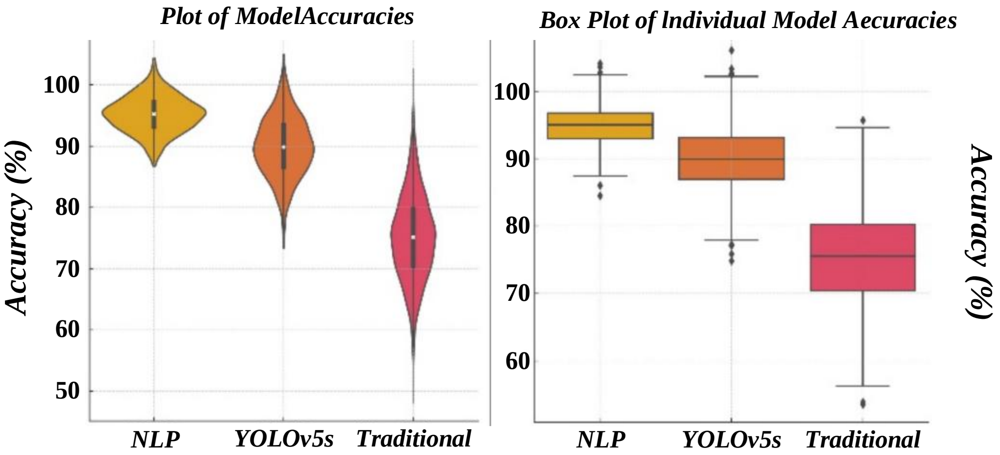

Figure 8: Experimental results of mAP

The comparison in Figure 8 shows that Bicubic's mAP in the test set is as high as 80.7%, the nearest is as low as 78.3%, and bilinear is in the middle of 80.5%. However, the inference time is bicubic, which is the longest; nearest, the shortest; and bilinear, which is centered. To balance accuracy and time, bilinear is selected as the upsampling method. After about 40 iterations, the mapping curve tends to be stable. Bicubic is close to bilinear, and the nearest is the lowest.

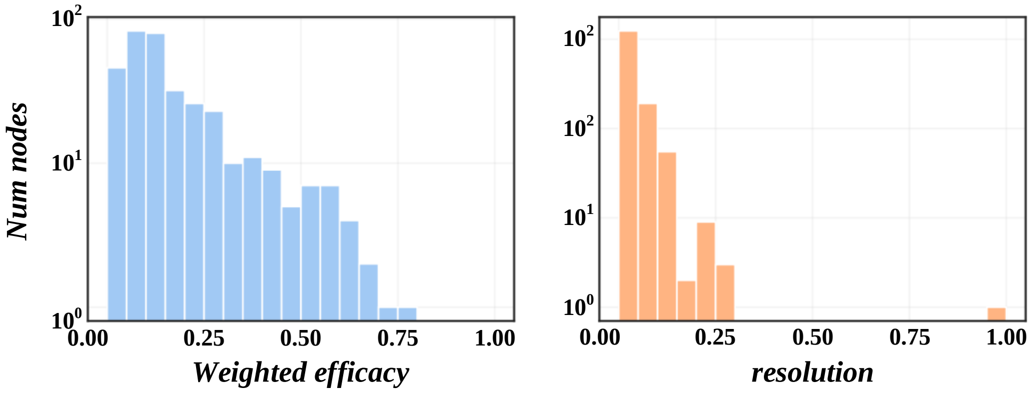

Figure 9: Performance variation of CBAM enhancement target

Figure 9 shows that the improved YOLOv5s increase category AP by 2.2%, mAP by 0.3%, CBAM enhances target attention, Bottleneck-D structure improves feature expression, and mAP by 1.3%. The Bottleneck structure in Neck part C3 improves the learning ability across connections, the mAP increases by 1.2%, the calculation amount decreases by 0.5 GFLOPS, and the detection efficiency and accuracy are improved, but the Precision decreases by 1.4%. Using bilinear interpolation upsampling, the mAP is 2.2% higher than the original model, which balances the speed and accuracy and proves that the upsampling algorithm is effective.

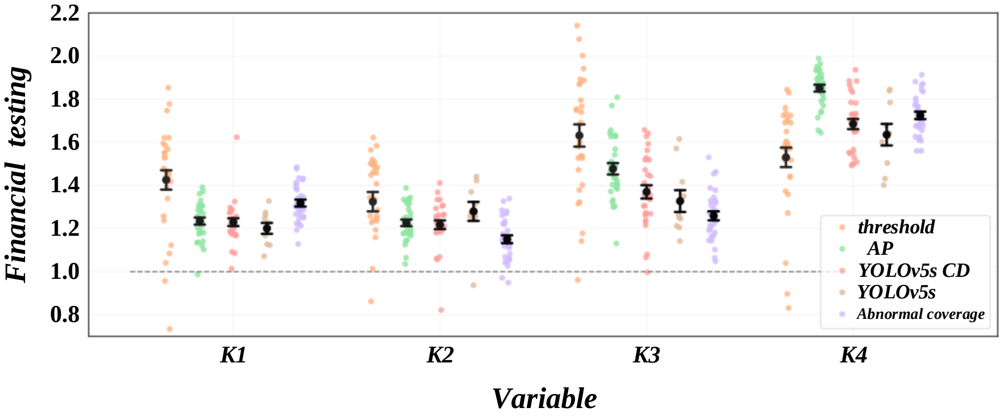

Figure 10: Performance comparison of improved model

The comparison in Figure 10 shows that the AP of the improved model is not improved, which confirms that the effect of the improved YOLOv5s _ CD model is better than that of the original YOLOv5s model. Figure 11 shows that the value/obj _ loss of the YOLOv5s _ CD model is more evenly distributed after 50 iterations, which is better than the original YOLOv5s model.

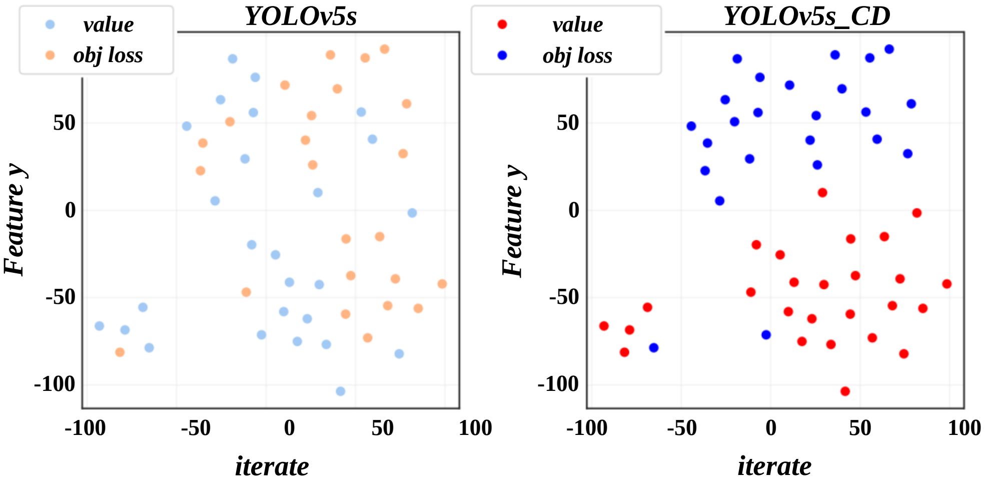

Figure 11: Model iterative experimental results

5 Conclusion

In digital transformation, financial statement audits face double-checking for efficiency and accuracy. In response to this challenge, this study proposes an intelligent audit system that integrates YOLOv5s financial image recognition technology and natural language processing, aiming to realize automated and intelligent audit of financial statements and improve audit efficiency and accuracy.

Compare with traditional methods. In terms of cost, traditional audit requires a lot of manpower, and professional auditors need to be hired and pay high salaries, while this intelligent audit system has low follow-up operating costs after investing in R&D and deployment costs in the early stage, which can reduce labor costs according to statistics. In terms of time, the traditional method of manual review of reports takes a long time, and it may take days or even weeks to process a complex financial statement, but the intelligent audit system uses YOLOv5s to quickly identify report elements and NLP analysis texts, shortening the processing time to several hours, and saving about half of the overall audit cycle. In terms of error rate, traditional audit is prone to negligence due to manual operation, with an error rate of 15% to 20%, while the intelligent audit system reduces the error rate to less than 3% with accurate algorithms and models, which significantly improves the accuracy and reliability of the audit and shows huge economic benefits and application advantages.

The application of innovative financial image recognition technology has been realized. By optimizing the YOLOv5s model, the accurate positioning and identification of complex table structures in financial statements, including various financial data, charts and annotations, has been completed. The recognition accuracy rate is as high as 98%, which is higher than that of traditional methods. The method has increased by 15%. When dealing with large-scale financial statement data, the system improves the accuracy and speed of audits and assists auditors in risk early warning, significantly improving the intelligence level of audit work.

Through the deep integration of natural language processing and NLP technology, the system can deeply understand the text content of the report, automatically extract vital financial indicators, risk warnings and other information, provide auditors with intuitive analysis results, and significantly improve the efficiency of audit report generation.

The intelligent audit system can be extended to a wider range of financial applications. In terms of fraud detection, a large number of fraud case reports are trained, allowing YOLOv5s to identify abnormal data areas, NLP to analyze text fraud expressions, build feature models and set thresholds, and timely warning in case of anomalies. In terms of regulatory compliance inspection, we analyze regulatory policies and regulations to build a rule base, use NLP to convert rule codes, YOLOv5s to locate key compliance indicators, and the system automatically verifies data and generates compliance reports to help financial institutions cope with supervision.

Through the intelligent anomaly detection mechanism and the built-in anomaly detection algorithm of the system, it can automatically identify abnormal data and potential risk points in financial statements, assist auditors in quickly locating problems, and reduce the burden of manual review. The processing speed of the optimized system is increased by three times, which significantly shortens the audit cycle, reduces the audit cost, and brings significant economic benefits to audit institutions.

At a time when digital transformation is accelerating, the intelligent audit system for financial statements based on YOLOv5s and natural language processing technology has brought new opportunities for the intelligent development of audit work. However, for the system to work in a real-world enterprise scenario, integration with existing enterprise resource planning (ERP) systems is essential. On the one hand, the ERP system data formats and interface specifications of different enterprises are different, and the intelligent audit system of financial statements will be hindered by the differences in the expression of dates, amounts and other fields when extracting and converting financial data, and the interface openness of some ERP systems is insufficient, and middleware or customized development is required, which not only increases the complexity of integration, but also brings data security risks; On the other hand, the multi-layered and complex architecture of the ERP system is different from the audit function-focused architecture of the audit system, which makes the data interaction between the systems difficult. At the same time, the security and reliability of the system cannot be ignored, especially its robustness, whether an attacker can bypass the audit system by manipulating financial statements and deceiving YOLOv5s and natural language processing modules, needs to be studied urgently.

References

[1] D. Liu, C. Deng, H. Zhang, J. Li, and B. Shi, "Adaptive Reflection Detection and Control Strategy of Pointer Meters Based on YOLOv5s," Sensors, vol. 23, no. 5, 2023.

[2] S. Xu, J. Deng, Y. Huang, and T. Han, "Multi-hidden target detection of transmission line based on improved YOLOv5s and its hardware implementation," Journal of Intelligent & Fuzzy Systems, vol. 46, no. 1, pp. 923-939, 2024.

[3] Z. Sun, Y. Cui, Y. Han, and K. Jiang, "Substation High-Voltage Switchgear Detection Based on Improved EfficientNet-YOLOv5s Model," IEEE Access, vol. 12, pp. 60015-60027, 2024.

[4] A. Mahany, H. Khaled, N. S. Elmitwally, N. Aljohani, and S. Ghoniemy, "Negation and Speculation in NLP: A Survey, Corpora, Methods, and Applications," Applied Sciences-Basel, vol. 12, no. 10, 2022.

[5] A. Faccia, J. McDonald, and B. George, "NLP Sentiment Analysis and Accounting Transparency: A New Era of Financial Record Keeping," Computers, vol. 13, no. 1, 2024.

[6] Hanning Zhang, Qinghua Zheng, Bo Dong, and Boqin Feng, "A financial ticket image intelligent recognition system based on deep learning," Knowledge-Based Systems, vol. 222, pp. 106955, 2021.

[7] Z. Zheng, F. Cao, S. Gao, and A. Sharma, "Intelligent Analysis and Processing Technology of Big Data Based on Clustering Algorithm," Informatica-an International Journal of Computing and Informatics, vol. 46, no. 3, pp. 393-402, 2022.

[8] Qi Wang, Lin Zhang, Qianqun Ma, and Chong Wu, "The impact of financial risk on boilerplate of key audit matters: Evidence from China," Research in International Business and Finance, vol. 70, pp. 102390, 2024.

[9] Shuping Wei, Fangxin Jiang, Jiawei Pan, and Qihai Cai, "Financial innovation, government auditing and corporate high-quality development: Evidence from China," Finance Research Letters, vol. 58, pp. 104567, 2023.

[10] Hui Xia, Shu Lin, Shuo Li, and Indranil Bardhan, "The effect of audit committee financial expertise on earnings management tactics in the post-SOX era," Advances in Accounting, vol. 64, pp. 100725, 2024.

[11] Manal Yunis, Nawazish Mirza, Adnan Safi, and Muhammad Umar, "Impact of audit quality and digital transformation on innovation efficiency: Role of financial risk-taking," Global Finance Journal, vol. 62, pp. 101026, 2024.

[12] C. Kahraman, "Proportional Fuzzy Set Extensions and Imprecise Proportions," Informatica, vol. 35, no. 2, pp. 311-339, 2024.

[13] E. B. Kenmogne, I. Tetakouchom, C. T. Djamegni, R. Nkambou, and L. C. Tabueu, "An Improved Algorithm for Extracting Frequent Gradual Patterns," Informatica, vol. 35, no. 3, pp. 577-600, 2024.

[14] Yubin Gao and Lirong Han, "Implications of Artificial Intelligence on the Objectives of Auditing

Financial Statements and Ways to Achieve Them," Microprocessors and Microsystems, vol., pp. 104036, 2021.

[15] Ahnaf Ali Alsmady, "Quality of financial reporting, external audit, earnings power and companies' performance: The case of Gulf Corporate Council Countries," Research in Globalization, vol. 5, pp. 100093, 2022.

[16] Bilal, Bushra Komal, Ernest Ezeani, Muhammad Usman, Frank Kwabi, and Chengang Ye, "Do the educational profile, gender, and professional experience of audit committee financial experts improve financial reporting quality?" Journal of International Accounting, Auditing and Taxation, vol. 53, pp. 100580, 2023.

[17] Saeed Awadh Bin-Nashwan, Jackie Zhanbiao Li, HaiChang Jiang, Anas Rasheed Bajary, and Muhammad M. Ma'aji, "Does AI adoption redefine financial reporting accuracy, auditing efficiency, and information asymmetry? An integrated model of TOE-TAM-RDT and big data governance," Computers in Human Behavior Reports, vol. 17, pp. 100572, 2025.

[18] José Cascais Brás, Ruben Filipe Pereira, Micaela Fonseca, Rui Ribeiro, and Isaias Scalabrin Bianchi, "Advances in auditing and business continuity: A study in financial companies," Journal of Open Innovation: Technology, Market, and Complexity, vol. 10, no. 2, pp. 100304, 2024.

[19] Rajeev Kumar and M. P. S. Bhatia, "An intelligent optimized secure blockchain mechanism for cloud auditing," Expert Systems with Applications, vol. 255, pp. 124593, 2024.

[20] Yu Liu et al., "BCDA: A blockchain-based dynamic auditing scheme for intelligent IoT," Computers and Electrical Engineering, vol. 119, pp. 109460, 2024.

[21] Xiaodong Huang and Lingling Luo, "Executive financial background, external audit quality and shadow banking in non-financial firms," Finance Research Letters, vol. 64, pp. 105397, 2024.

[22] X. Liu, T. P. Singh, R. K. Gupta, and E. M. Onyema, "Chaotic Association Feature Extraction of Big Data Clustering Based on the Internet of Things," Informatica-an International Journal of Computing and Informatics, vol. 46, no. 3, pp. 333-342, 2022.

[23] Zihan Liu, Christine Jubb, and Subhash Abhayawansa, "Choice of financial audit firm and ESG assurance firm: The role of board of director characteristics," The British Accounting Review, vol., pp. 101505, 2024.

[24] Nora Muñoz-Izquierdo, María-del-Mar Camacho-Miñano, María-del-Pilar Sánchez-Martín, and David Pascual-Ezama, "Is auditor financial decision-making affected by prior audit report information? A behavioral approach," Heliyon, vol. 10, no. 10, pp. e30971, 2024.

[25] Jingjuan Zhu, Wenjie Zhang, Lingyun Lu, Yi Lu, and Duo Wang, "Hot spot mining and trend analysis of Economic Responsibility Audit based on knowledge graph," Mathematics and Computers in Simulation, vol. 222, pp. 38-49, 2024.

[26] María-del-Mar Camacho-Miñano, Nora Muñoz-Izquierdo, Morton Pincus, and Patricia Wellmeyer, "Are key audit matter disclosures useful in assessing the financial distress level of a client firm?," The British Accounting Review, vol. 56, no. 2, pp. 101200, 2024.

[27] Qianqian Chen and Zhi Chen, "Mandatory internal control audit and corporate financialization," Finance Research Letters, vol. 62, pp. 105085, 2024.

[28] A. Kilciauskas, A. Bendoraitis, and E. Sakalauskas, "Confidential Transaction Balance Verification by the Net Using Non-Interactive Zero-Knowledge Proofs," Informatica, vol. 35, no. 3, pp. 601-616, 2024.

[29] Meeok Cho, Hui Dong Kim, and Yewon Kim, "Audit committee accounting financial expertise and stock price crash risk," International Review of Financial Analysis, vol. 90, pp. 102848, 2023.

[30] T. Zvirblis, A. Piksrys, D. Bzinkowski, M. Rucki, A. Kilikevicius, and O. Kurasova, "Data Augmentation for Classification of Multi-Domain Tension Signals," Informatica, vol. 35, no. 4, pp. 883-908, 2024.

[31] Robert Felix, Sattar Mansi, and Mikhail Pevzner, "Audit committee–CFO political dissimilarity and financial reporting quality," Journal of Accounting and Public Policy, vol. 45, pp. 107209, 2024.

[32] J.-C. Wang, and T. Y. Chen, "An Uncertain Multiple-Criteria Choice Method on Grounds of T-Spherical Fuzzy Data-Driven Correlation Measures," Informatica, vol. 33, no. 4, pp. 857-899, 2022.

[33] Xin Huang, Hao Huang, and Liang Yuan, "Do firms incur financial restatements? A recognition study based on textual features of key audit matters reports," International Review of Financial Analysis, vol. 96, pp. 103606, 2024.