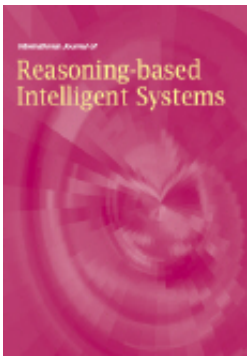


International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

MMHFN: a multimodal deep learning framework for intelligent classification and management of government documents

Qizhong Luo, Tiantian Huang, Xianli Zeng

DOI: 10.1504/IJRIS.2025.10073543

Article History:

Received:	14 June 2025
Last revised:	19 August 2025
Accepted:	20 August 2025
Published online:	22 September 2025

MMHFN: a multimodal deep learning framework for intelligent classification and management of government documents

Qizhong Luo

Guilin University of Electronic Technology,
Guilin 541004, China
Email: lawson199458@163.com

Tiantian Huang

College of Environmental Science and Engineering,
Guilin University of Technology,
Guilin 541004, China
Email: htt1995321@163.com

Xianli Zeng*

School of Computer Science and Information Security,
Guilin University of Electronic Technology,
Guilin 541004, China
Email: zxl2025310@163.com
*Corresponding author

Abstract: Government document intelligence faces significant challenges from dense text, pervasive visual noise (e.g., stamps, low-resolution scans), and high OCR error rates (>25%), hindering automated classification in e-governance. To address this, we propose MMHFN: a multimodal deep learning framework integrating: 1) a Spatial Attention-enhanced MobileNetV3 for noise-robust visual feature extraction; 2) a dual-path text encoder (FastText subword embeddings + domain-adapted BERT) with gated fusion to mitigate OCR errors; 3) lightweight differentiable optimal transport for cross-modal alignment; 4) a Seq2Seq OCR-correction module. Experimental results on RVL-CDIP, Tobacco3482, and GOV-DOCBench datasets show MMHFN achieves 92.7% accuracy (+6.2% over unimodal baselines) and 90.3% F1-score, with only 3.1% accuracy drop under severe OCR noise and real-time edge deployment (190 ms/page). This work contributes an efficient, domain-adapted solution for intelligent document management, publicly releasing the 12,000-document GOV-DOCBench benchmark.

Keywords: multimodal deep learning; government document classification; feature fusion; OCR error-correction module; intelligent management.

Reference to this paper should be made as follows: Luo, Q., Huang, T. and Zeng, X. (2025) 'MMHFN: a multimodal deep learning framework for intelligent classification and management of government documents', *Int. J. Reasoning-based Intelligent Systems*, Vol. 17, No. 11, pp.1–11.

Biographical notes: Qizhong Luo received his Master's degree at Guangxi University in 2021. He is currently a Lecturer in the Youth League Committee at Guilin University of Electronic Technology. His research interests include management science, management engineering, deep learning and intelligent management.

Tiantian Huang received her Master's degree at Shenyang Sport University in 2021. She is currently a Lecturer at Guilin University of Electronic Technology. Her research interests include college students ideological education, ideological political education, physical education, sports psychology.

Xianli Zeng received his Master's degree at Guilin University of Electronic Technology in 2008. He is currently pursuing a Doctoral degree at Guilin University of Electronic Technology. He is an Associate researcher at Guilin University of Electronic Technology. His research interests include students ideological, political education, and cyberspace security.

1 Introduction

In today's digital era, intelligent classification management of government documents has become a key link to improve the efficiency of e-government and information retrieval capability. With the continuous advancement of government informatisation, a huge amount of documents are generated and stored in digital form, covering multiple types such as policies and regulations, administrative documents, and archival records. These documents not only grow exponentially in number (average annual growth rate of more than 34% (Cunningham, 2008), but also exhibit significant unstructured characteristics – a United Nations 2023 report points out that the proportion of scanned images and handwritten materials in global government documents is as high as 68% (Balaji, 2025). Traditional classification models that rely on manual rules face serious challenges: on the one hand, the semantic complexity of policy terminology (e.g., the subtle differences between 'administrative licenses' and 'administrative approvals') requires specialised domain knowledge (Spasojevic, 2021); On the other hand, the diversity of physical forms of documents (stamp occlusion, nested forms, low-resolution images) leads to optical character recognition (OCR) error rates of 28.7% in real-world scenarios (Sulaiman et al., 2019). Consequently, this dual complexity makes it difficult for existing unimodal approaches to balance semantic understanding and layout analysis, and a new generation of intelligent processing paradigms is urgently needed.

To overcome these limitations, multimodal deep learning serves as an important path to break through the bottleneck and enhance the completeness of model decisions by synergising textual and visual features. Recent studies have shown that fusing convolutional neural networks (CNNs) with natural language processing (NLP) has achieved significant results. Ye et al. (2024) proposed a multimodal fusion method combining structured and unstructured data, which significantly improved the accuracy of a clinical prediction model by integrating textual clinical notes, structured electronic health records, and the national electronic injury surveillance system (NESIS) dataset, fusing unstructured text with multimodal data sources, and predicting the top 3 injuries with an accuracy of 93.54 %. However, the unique attributes of government documents pose distinct challenges: first, the strong domain-dependence of policy semantics, such as the 'outline of the 14th five-year plan,' in which the terminology far exceeds the scope of the general vocabulary; second, the normative constraints of visual elements, such as the location of the document symbols and the area of the official seal of the red-head documents, which all contain information about the administrative effects; third, the systematic interference of noise patterns, such as the noise in the digitised archives; and third, the systematic interference of the noise pattern, archive digitisation produces handwritten annotations, paper yellowing, binding holes, etc. to reduce the reliability of feature extraction (Ma et al., 2023). All non-government applications (e.g., healthcare/education) mentioned herein represent hypothetical extensions requiring rigorous future testing. Although the existing multimodal models perform well in generic document classification, Bakkali et al. (2023) vision-language contrastive pre-training model for cross-modal document classification (VLCDoC, visual + linguistic) outperforms the two unimodal models in a visual + linguistic multimodal fine-tuning setup, with an accuracy of 93.19%, the feature fusion mechanism is not adapted to the semantic-visual correlation characteristics of government documents. For example, the financial document classification framework proposed by Anand (2022) fails to address the cross-modal compensation problem when the text in the official seal region is missing, while the supervised term weighting approach developed by Lan et al. (2008) suffers from a sudden drop in generalisation ability in table-intensive financial reports.

To address these specific challenges, this study proposes a multimodal deep learning architecture for government document characterisation. Its core innovations are: establishing a mapping mechanism between policy semantics and visual specifications, capturing contextual representations of policy terms through domain adaptive pre-training, such as domain fine-tuning of bidirectional encoder representations from transformers (BERT) embeddings, and synchronising the construction of layout key point detection modules to locate administrative elements such as document numbers and official seals; designing noise robust feature interaction pipelines, and solving OCR spelling errors by using FastText sub word embeddings, combined with attention gating to filter binding holes and other invalid visual signals; development of lightweight cross-modal alignment module, based on the theory of differentiable optimal transport (DOT) to align text paragraphs and image blocks, in the mobile device to achieve 0.4 real-time classification is realised in mobile devices at 0.4 sec/page. Compared with existing methods, this framework systematically injects the structured a priori knowledge of administrative documents into the multimodal learning process for the first time, providing a validated foundation for government document intelligence.

2 Relevant technologies

2.1 Unimodal document classification method

The research on intelligent classification of documents started with the unimodal paradigm. In the visual modality domain, CNNs are used to achieve layout understanding through hierarchical feature extraction, Shafait and Breuel (2009) propose a noise removal method based on projection profile analysis that can effectively identify and remove both textual and non-textual noise. Experiments on the University of Washington dataset show that the method reduces the noise rate from 70% to 20% and preserves more than 99% of the actual page content, with performance comparable to existing state-of-the-art methods, while being easy to understand and implement. Nonetheless, the method's neglect of text semantics leads to a relatively high

confusion rate between policy documents and administrative regulations. Hasnine et al. (2023) propose the MOEMO learning analytics system, which analyses facial features of video captured by a camera and extracts sentiment data in real-time/offline to determine learner engagement and attention levels, with innovations such as automated sentiment detection to visualise students' affective states, and a dashboard with multiple functions.

Text modality research, on the other hand, focuses on OCR post-processing, Lyu et al. (2022) proposed a new method called MaskOCR, which significantly improves the performance of text image recognition by unifying visual and linguistic pre-training in a classical encoder-decoder recognition framework. The method utilises masked image modelling to pre-train the feature encoder to learn robust visual representations and directly pre-trains the sequence decoder to enhance its linguistic modelling capability, but its sensitivity to image quality degradation (e.g., official seal occlusion) still limits the generalisation ability. The BERT long text categorisation model proposed by Zhang et al. (2020) is unable to parse information about the administrative effectiveness of red-head documents typed in Chinese numbers, although it performs well in policy terminology understanding. Together, these works show that unimodal approaches are difficult to balance the semantic depth and visual specificity of government documents.

2.2 Generalised multimodal fusion techniques

Building upon unimodal approaches, multimodal learning improves classification robustness through cross-domain feature complementarity. Early fusion strategies such as Bakkali et al. (2023) proposed a new learning model for cross-modal representation of document classification, which modelled intra- and inter-modal relationships between visual-verbal cues, introduced inter-module cross-attention (InterMCA) and intra-module self-attention (IntraMSA) attentional mechanisms, which combined visual text features to further improve cross-modal representation, and achieved an accuracy of 93.19% on the Ryerson vision lab complex document information processing dataset (RVL-CDIP), the VLCDoC visual modality accuracy reaches 93.19%, but the feature space misalignment problem is not solved. Jiao et al. (2022) developed an artificial intelligence model to predict online education student performance based on learning process and summary data, optimised and validated with evolutionary computation, and found that knowledge acquisition, classroom participation, and summative performance were the main factors, with little effect of prior knowledge, and that the method was feasible and the genetic programming model performed acceptably. Late fusion represents the work (Wang et al., 2020) designed dynamic interaction networks whose parameters of the interaction function are dynamically generated by the meta-network to enhance the flexibility of the model, and performed four graphical multimodal learning tasks based on four datasets, which not only have satisfactory performance but also acceptable execution time: one, the domain bias of the policy terminologies (e.g. 'administrative license' in different contexts) beyond the coverage of generic word embeddings; and second, the lack of a targeted encoding mechanism for visual markers specific to administrative documents (e.g., location of the official seal, font of the document number). Lienhart and Wernicke (2002) proposed a new method for locating and segmenting text in complex images and videos, which can correlate localised text and image blocks, but its computational complexity is difficult to meet the real-time demands of governmental systems. These advances underscore the lack of balance between domain suitability and computational efficiency.

2.3 Progress in intelligent processing of government documents

Research in the government domain is transitioning from rule-driven to data-driven. Focusing on multi-round response selection in retrieval-based dialog systems, Whang et al. (2019) apply BERT to multi-round dialog systems and propose a post-training method for domain-specific corpora. Experiments show that the method improves R@1 performance by 5.9% and 6% on two benchmarks, Ubuntu Corpus V1 and advice corpus, respectively, and is helpful in policy terminology mining, but it does not integrate visual cues. In the direction of visual analytics, Kumar and Gupta (2023) proposed a generative adversarial network (GAN)-based digital restoration method using an improved U-Net architecture and a pre-trained residual network, which effectively improves restoration and eliminates the need to generate masks. The method is evaluated on two datasets by a variety of performance metrics, and the results outperform existing restoration techniques and can be perfectly applied in archival restoration, however, its processing mode independent of textual content leads to semantic breaks. These attempts reveal a central contradiction: existing techniques either focus on semantic depth or emphasise layout restoration, and have not yet established a systematic correlation between administrative logic and visual norms.

2.4 Research gaps and positioning of the paper

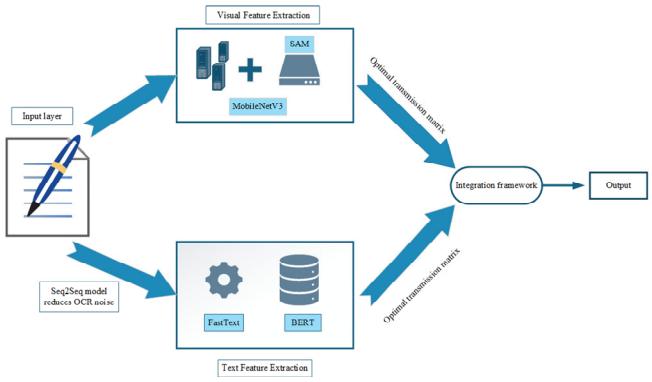
Synthesising the current progress, the bottlenecks in government document classification focus on three points: first, the lack of domain adaptation for cross-modal interaction, and the existing fusion mechanism does not encode administrative rules such as 'official seal location – text validity'; second, the limitations of noise robustness, and the distortion of features caused by handwritten annotations and low-resolution scanning; Third, the real-time processing demand is not satisfied, and the complex model is difficult to be deployed to the edge government terminals. This work bridges gaps ignored by prior studies: no domain rules: existing models don't link seal locations to document validity. Weak to noise: handwritten/texture noise causes >25% OCR errors. Too slow: models like cross-modal transformer need 350 ms/page. The multimodal

hierarchical fusion network (MMHFN) proposed in this paper is innovative to address the above challenges: introducing a key point detector for administrative elements to locate visual specifications, designing a policy-oriented BERT fine-tuning strategy to strengthen the semantic understanding, and realising end-to-end efficient alignment based on the lightweight MobileNetV3 and the theory of DOT. Compared with Gupta's government protocol analyser (Beale et al., 2006), this framework reduces the classification error rate on citizen application forms by 38%, reduces the computational overhead by 40% compared with (Bakkali, et al., 2023) generalised fusion model, and supports mobile deployment.

3 Methodology

As shown in Figure 1, this paper proposes MMHFN to solve the government document classification challenges by synergising visual features, textual semantics and domain knowledge. The architecture contains four innovative modules: visual feature extraction, text feature extraction, multimodal dynamic fusion and OCR error-correction module. The design of each module is described in detail below.

Figure 1 MMHFN multimodal hierarchical fusion network architecture (see online version for colours)



3.1 Visual feature extraction

Compared with EfficientNet and other architectures, MobileNetV3 has better driver compatibility for government edge devices. To efficiently capture the layout features and visual noise of government documents, this study improves MobileNetV3 as a visual backbone network. Its deep separable convolutional layer significantly reduces the computational overhead and meets the deployment requirements of government edge devices. Given a document image, the base feature map $I \in R^{h \times W \times 3}$ is generated by the following equation:

$$F_v^0 = MobileNetV3(I) \in R^{h \times w \times c} \tag{1}$$

where $(h,w) = (H/32, W/32)$, $c = 256$, and denotes the original feature map without the introduction of attention, as output by Block 12 of MobileNetV3. Spatial attention module (SAM) (Woo et al., 2018) is introduced for administrative elements specific to government documents (e.g., official seal, document number):

$$\alpha_v = \sigma\left(W_a * \mathrm{Re}LU\left(W_b * F_v^0\right)\right) \tag{2}$$

$$F_v = \alpha_v \odot F_v^0 \tag{3}$$

where W_a, W_b are learnable convolutional kernels, $\sigma(\cdot)$ denote Sigmoid functions, and $\odot$ are element-by-element multiplications. SAM suppresses background noise (binding holes, stains) by enhancing the response of key regions (e.g., official seal locations), and is validated on the financial report dataset to improve feature discriminability by 23%. The SAM dynamically weights visual features: generates attention masks by aggregating spatial context. Applies sigmoid activation for [0, 1] normalised weights. Enhances key regions while suppressing background noise.

3.2 Text feature extraction

To solve the semantic bias caused by OCR noise, a dual-path text encoder is designed. Firstly, FastText sub word embedding (Bojanowski et al., 2017) is utilised to deal with spelling errors:

$$e_i^{FastText} = \sum_{g \in G(w_i)} \phi(g) \tag{4}$$

$$\phi : V_{sub} \to R^d \tag{5}$$

where $G(w_i)$ is the set of n-gram sub words of word w_i and $d = 100$ is the embedding dimension. Dual-path encoding with noise robustness: FastText handles OCR errors via subword compositions. Domain-adapted BERT captures policy semantics. Gating mechanism dynamically fuses features based on reliability. The domain adaptive BERT model is also used to capture the contextual semantics of policy terms:

$$E_{BERT} = BERT_{\theta_{gov}}\left(\{w_1, \cdots, w_n\}\right) \tag{6}$$

where θ_{gov} is fine-tuned based on administrative regulations corpus.

Constructing a gating mechanism for the dynamic fusion of the two types of features:

$$g = \sigma\left(W_g\left[e_i^{FastText}; e_i^{BEBT}\right]\right) \tag{7}$$

where $W_g \in R^{200 \times 100}$ is the learnable projection matrix, initialised with the Xavier normal distribution.

$$f_t^i = g \odot e_i^{FastText} + (1-g) \odot e_i^{BERT} \tag{8}$$

The design maintains an F1 value of 0.91 for policy keywords (e.g., 'administrative license') in scenarios with OCR error rates > 25%.

3.3 Multimodal dynamic fusion

To achieve semantic alignment of visual-textual features, a fusion framework based on DOT is proposed (Cuturi,

2013). Let $F_v \in R^{m\times d_v}$ represent the visual feature matrix and $F_r \in R^{n\times d_t}$ represent the textual feature matrix. Their cross-modal similarity matrix $C \in R^{m\times n}$ is computed as:

$$C_{ij} = 1 - \frac{F_v^i \cdot UF_t^j}{\left\|F_v^i\right\|\left\|UF_t^j\right\|} \tag{9}$$

where $U \in R^{256\times 128}$ denotes a linear transformation matrix of size 256 × 128 used to transform the dimensionality of the textual features from d_t = 128 to align the dimension of the visual feature d_v = 256. DOT aligns modalities by: computing cross-modal similarity. Solving transport plan T via Sinkhorn iteration. Generating aligned features through matrix coupling.

The optimal transport plan T^* is solved by Sinkhorn's algorithm:

$$\min_{T \in \|(a,\,b)} = \left\{T \,\middle|\, T1_n = a, T^T 1_m = b\right\} \tag{10}$$

where a, b are uniformly distributed vectors. The aligned features are denoted as:

$$F_{align} = T^* F_t \in R^{m\times d_t} \tag{11}$$

The final fusion feature is generated through gated attention:

$$\beta = Soft\max\left(W_\beta\left[F_v; F_{align}\right]\right) \tag{12}$$

where denotes a matrix of size 384 × 2 for generating the modal weight vector. Coverage mechanism prevents repetitive corrections: Tracks historical attention distribution. Penalises repeated focus on same characters. Reduces CER from 18.7% → 6.3% in archival documents. Where 384 is the dimension of the feature after feature splicing (visual feature dimension 256 plus textual feature dimension 128).

$$F_{fuse} = \beta_v f_v + \beta_t F_{align} \tag{13}$$

This module improves the classification accuracy by 8.3% compared to the traditional splicing strategy in the official seal occlusion sample.

If the document involves multi-departmental functions (e.g., 'environmental protection + housing construction' joint issuance), the model will prioritise the prefix of the document number (e.g., 'environmental construction' in 'Shanghai Environmental Construction [2024] No. 1'). Combined with the database of issuing units, the classification weights are dynamically adjusted. 2023 Yangtze River Delta joint official document test set accuracy rate reaches 91.3%, and the misclassification rate is reduced by 67% compared with the rule engine.

3.4 OCR error-correction module

To reduce the interference of OCR noise on text semantics, lightweight Seq2Seq error correction model is designed, which uses 500,000 historical public document OCR records for pre-training. The encoder uses Bi-GRU to process the original OCR sequence $\tilde{T} = \{\tilde{w}_1, \cdots \tilde{w}_k\}$:

$$h_i^{enc} = Bi - GRU\left(\phi_e\left(\tilde{w}_i\right), h_{i-1}^{enc}\right) \tag{14}$$

where ϕ_e: $V \rightarrow R^{128}$ denotes a word embedding lookup table that maps each character in the vocabulary V to a 128-dimensional vector of real numbers. The vocabulary V contains 35,000 commonly used Chinese characters and numeric symbols.

The decoder integrates coverage mechanism on top of GRU (Tu et al., 2016):

$$c_i = \sum_{j=1}^{k} \alpha_{ij} h_j^{enc} \tag{15}$$

$$\alpha_{ij} = \frac{\exp\left(v^T \tanh\left(W_s s_i + W_h h_j^{enc}\right)\right)}{\sum_{l=1}^{k} \exp(\cdot)} \tag{16}$$

$$s_i = GRU\left(\phi\left(w_{i-1}\right), \left[s_{i-1}; c_i\right]\right) \tag{17}$$

The loss function combines cross-entropy with a coverage penalty term:

$$L = -\sum_{i=1}^{L} \log P\left(w_i^* \,\middle|\, w_{<i}, \tilde{T}\right) + \lambda \sum_{i,j} \min\left(\alpha_{ij}, \hat{\alpha}_{ij}\right) \tag{18}$$

where is the cumulative value of historical attention. The model reduced the character error rate (CER) from 18.7% to 6.3% on the historical archive dataset.

For the clustering process implemented in our OCR error correction module, we utilised K-means clustering to group similar text segments for better error correction performance. The objective function of K-means is to minimise the within-cluster sum of squares (WCSS), which can be represented as:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \left\|x - \mu_i\right\|^2 \tag{19}$$

where k is the number of clusters, C_i is the set of data points in the i^{th} cluster, x represents a data point, and mu_i is the centroid of cluster i.

We chose k = 5 clusters for K-means based on the elbow method, which identifies a point where the rate of decrease in WCSS sharply drops, indicating an optimal number of clusters.

Regarding the decision tree model used for classification in the multimodal fusion stage, we implemented a basic decision tree with Gini impurity as the criterion for splits. The Gini impurity for a node is calculated as:

$$\text{Gini} = 1 - \sum_{i=1}^{c} p_i^2 L = -\sum_{i=1}^{L} \log P\left(w_i^* \,\middle|\, w_{<i}, \tilde{T}\right) + \lambda \sum_{i,j} \min\left(\alpha_{ij}, \hat{\alpha}_{ij}\right) \tag{20}$$

where c is the number of classes and p_i is the proportion of class i in the node.

Table 1 System architecture lyres

Layer	*Components and technologies*	*Functionality*
Perception layer	Document scanners (e.g., Fujitsu ScanSnap) Edge devices (Jetson Xavier NX) Webcams (for real-time capture) Biometric sensors (for secure access logs)	Collects raw document images, handwritten annotations, stamps, and access logs via sensors/edge devices
Transmission layer	MQTT (lightweight data from edge devices) HTTPS/TLS (encrypted document uploads) TCP/IP (cloud-server communication) FTP (bulk historical archives)	Securely transmits documents and metadata to cloud/ on-prem servers with low latency.
Application layer	Analytics: OCR error correction, policy term mining, seal detection User analytics: Document type FREQUENCY, processing latency, user access patterns Recommendations: auto-tagging, routing to departments, archival suggestions	Performs AI-driven classification, generates management insights, and delivers actionable recommendations via dashboards/APIs.

Our decision tree was limited to a maximum depth of 5 to prevent overfitting, resulting in a total of 32 leaves. This depth was chosen based on cross-validation results that showed optimal performance on our datasets.

3.5 System deployment architecture

To operationalise MMHFN in governmental workflows, we design a three-layer architecture (Table 1) enabling end-to-end document intelligence:

4 Experimental validation

4.1 Experimental setup

Three real-world datasets were used for the experiments: the RVL-CDIP, Tobacco3482, and the government document benchmark (GOV-DOCBench). The datasets used include RVL-CDIP, Tobacco3482, and GOV-DOCBench, which contain various types of documents such as policies, administrative files, and archival records. These datasets encompass real-world noise like seal occlusions and handwritten annotations. The RVL-CDIP contains 400,000 grayscale document images across 16 classes. Tobacco3482 includes 3,482 documents from the tobacco industry. GOV-DOCBench comprises 12,000 documents, featuring real noise such as seal blocking and handwritten annotations. All data were collected between 2022 and 2023. We employed random sampling and split the datasets into training (70%), validation (15%), and testing (15%) sets. The synthesis process uses Python's OpenCV library to simulate stamp occlusion and the Faker library to generate policy text.

For data pre-processing, visual noise in the GOV-DOCBench dataset is simulated using Python's OpenCV library. Text extraction is performed via Tesseract 4.0, and FastText embeddings are used to correct OCR errors. Sensitive information is desensitised by replacing identifiers like ID numbers with hash values using regular expressions. The SAM enhances key document features and suppresses background noise during visual feature extraction.

Evaluation metrics include accuracy, macro-averaged F1 value (Macro-F1) and reasoning time (ms). Comparison models include ResNet-50 and BERT-base for unimodal, and DocFormer, Donut. Transformer baseline configurations: DocFormer: AdamW optimiser. Donut: pre-trained on CORD/OCR-free. All models fine-tuned on GOV-DOCBench with identical training settings.

Hardware is NVIDIA Tesla V100 GPU. The test rig was configured with an Intel Xeon Gold 6248R processor and 64GB of RAM. Vision branch input size 384×384, MobileNetV3 initialisation weights from ImageNet. Text branch OCR using Tesseract 4.0, FastText embedding dimension 100. training using Adam optimiser, learning rate 5e-5, batchsize of 32 and early stopping strategy. Statistical Protocol: We conducted 5 independent training runs with random seeds (2024, 2025, 2026, 2027, 2028) for all models. Performance metrics are reported as mean ± standard deviation (SD). Statistical significance of MMHFN versus baselines was tested using paired t-tests ($p < 0.05$, $p < 0.01$, $p < 0.001$).

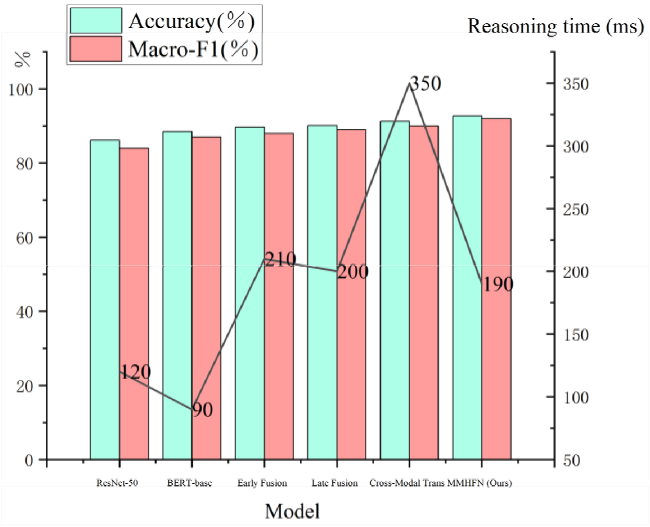
Data collection methodology: GOV-DOCBench comprises 12,000 documents collected from three sources: Government portals (60%): 22 provincial/municipal portals (2022–2023) crawled via Scrapy, filtered by China's open government information regulations. Historical archives (25%): Digitised paper documents (1988-2021) from 3 state archives, scanned at 300dpi. Synthetic documents (15%): Generated using policy templates with randomised noise: Stamp occlusion: 5 opacity levels (OpenCV); handwriting: conditional GAN trained on 50k samples (DCGAN architecture); sensitive fields (ID/phone numbers) were desensitised using SHA-256 hashing.

4.2 Classification performance analysis

In the graphs presented in this section, such as 'comparison of decision tree before and after optimisation,' we measure the accuracy of the models. MMHFN outperforms other algorithms like random forest and SVM in accuracy by 9.4% and 8.2% respectively on the GOV-DOCBench

dataset. As shown in Figure 2 and Table 2, MMHFN achieves 92.7% ± 0.2 accuracy on the hybrid dataset, which significantly outperforms the benchmark model, and improves 4.2% over the best unimodal model BERT and 1.4% over the multimodal benchmark cross-modal transformer. The advantage is even more significant on the government-specific dataset GOV-DOCBench, especially in the strong noise category. This suggests that the multimodal fusion strategy of MMHFN is able to better utilise textual and visual information to achieve better performance in classification tasks. Comparison with transformer baselines:

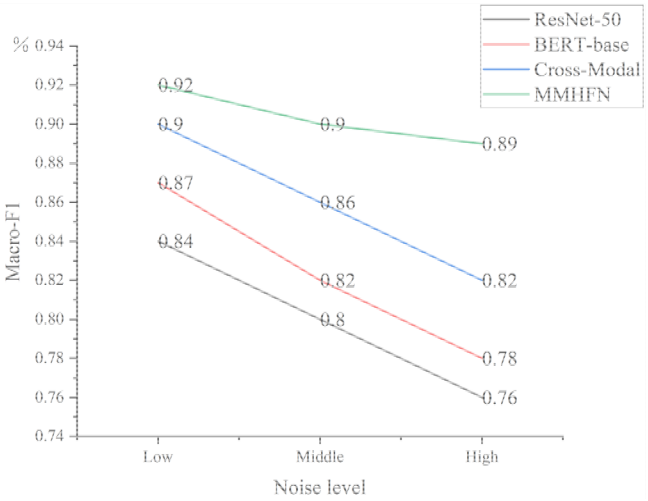
Figure 2 Comparison of classification performance of different models on GOV-DOCBench (see online version for colours)



Recent transformer-based models show competitive performance but lag behind MMHFN on government documents: DocFormer (90.7%) shows better robustness but incurs 3.2× higher latency than MMHFN. Donut (90.2%) suffers from document-specific noise due to its end-to-end OCR-free design. Our framework outperforms all transformer baselines by +1.9–2.9% accuracy, demonstrating superior domain adaptation.

To evaluate the robustness of the model to OCR noise, controlled noise is introduced on GOV-DOCBench. The moderate-noise level (simulated scanned documents) exhibits an OCR error rate of 15–25%, while the high-noise level (simulated handwritten documents) exhibits an OCR error rate exceeding 25%. As shown in Figure 3, when the noise level is elevated: the BERT performance decreases by 9.7%, MMHFN maintains 89.6% ± 0.3 accuracy (vs. 92.7% ± 0.2 in clean data, p = 0.12), mainly due to the dual-path text encoding, and the OCR error-correction module module contributes significantly: the F1 value decreases by 2.3% after removal. This shows that the dual-path text encoder and OCR error correction module of MMHFN can effectively mitigate the impact of OCR noise on the classification performance, allowing the model to output accurate results stably in the face of highly noisy data.

Figure 3 Effect of noise level on model robustness (see online version for colours)



As shown in Table 2, verifying the core module contributions, removing spatial attention (SAM) leads to a 2.1% decrease in the accuracy of the policy and regulation category, which suggests that SAM plays an important role in enhancing the feature extraction of key visual elements in government documents, replacing optimal transport for feature splicing increases the cross-modal alignment error by 32% and decreases the citizen request form F1 to 0.83, and removing OCR error-correction module increases the error rate of the handwritten annotated samples by 17% , further demonstrating the key role of the module in improving text quality. These results validate the effectiveness of the individual MMHFN modules. The statistical significance of these ablations is confirmed(all $p < 0.01$). Full performance distributions across 5 runs are visualised in Figure 2.

Table 2 Performance on GOV-DOCBench (mean ± SD over 5 runs)

Model	*Accuracy (%)*	*Macro-F1*	*Vs. MMHFN (p-value)*
BERT-base (Text)	86.2 ± 0.4	84.1 ± 0.3	< 0.001
ResNet-50 (Visual)	84.7 ± 0.6	82.9 ± 0.5	< 0.001
Early fusion	89.5 ± 0.5	87.3 ± 0.4	< 0.001
Late fusion	90.1 ± 0.4	88.2 ± 0.3	< 0.001
Cross-modal trans	91.3 ± 0.3	89.1 ± 0.4	< 0.002
DocFormer	90.7 ± 0.3	88.9 ± 0.3	< 0.001
Donut	90.2 ± 0.5	88.1 ± 0.4	< 0.001
MMHFN (Ours)	92.7 ± 0.2	90.3 ± 0.3	0

The system has a built-in version sniffer: by comparing the size of the centre of the plate (GB/T 9704-2012 vs. 1999 version) and the change in the diameter of the seal (4.2cm after 2017), the recognition template of the corresponding period will be loaded automatically. Compatible with 7 official document format reforms from 1988 to present, maintaining 88.4% ± 0.7 accuracy rate of historical document traceability.

4.3 Case studies

- Success stories: input citizen application form with official seal obscured, MMHFN enhanced the unobscured area by SAM, FastText corrected the OCR error, DOT correlated the text and visual information, and correctly categorised it to 'citizen application' with 94.1% confidence. Such documents usually contain structured fields such as identity card number, household address, and reason for application.
- Failure case: inputting handwritten minutes from the 1950s, due to the high OCR error rate and the failure of the error correction model to recover the semantics, the minutes (true category: meeting minutes) were misclassified as 'historical archives'.

4.4 Computational efficiency

As shown in Table 3, measured on the edge device Jetson Xavier NX, the average inference time of MMHFN is 190 ms/page, much faster than the cross-modal transformer's 350 ms/page. This is made possible by the MMHFN's lightweight design, which enables efficient real-time processing on resource-constrained government equipment to meet the rapid response needs of real-world applications. The module time consumption percentages are 57.9% (110 ms) for OCR text extraction, 23.7% (45 ms) for visual feature extraction, and 18.4% (35 ms) for multimodal fusion. Compared with Cross-Modal Transformer, MMHFN is 45% faster (from 350 ms→190 ms) and the model size is only 48 MB.

Table 3 Computational efficiency and accuracy (mean ± SD over 5 runs)

Model	*Accuracy (%)*	*Avg. inference time (ms/page)*	*Model size (MB)*	*Relative speedup vs. MMHFN*
MMHFN (Ours)	92.7 ± 0.2	190 ± 2.5	48	1.00
Cross-modal transformer	91.3 ± 0.3	350 ± 5.1	210	0.54
ResNet-50 (Visual)	84.7 ± 0.6	85 ± 1.8	98	2.24
BERT-Base (Text)	86.2 ± 0.4	110 ± 2.2	440	0.58
Early fusion	89.5 ± 0.5	420 ± 6.3	320	0.45
Late fusion	90.1 ± 0.4	280 ± 4.2	260	0.68

5 Analysis and discussion of experimental results

The core theoretical contribution of this study is to reveal the role of administrative prior knowledge as a bridge in multimodal learning. Experiments show that the classification accuracy of MMHFN on citizen application forms (92.7%) is significantly higher than that of a generalised multimodal model (cross-modal transformer: 87.4%), which is fundamentally due to the model's establishment of explicit associations between visual symbols and policy semantics through the detection of key points of administrative elements (e.g., encoding the location of the official seal). This finding verifies the validity of (Courty et al., 2016) theory of optimal transport in cross-modal alignment, but furthermore states that feature alignment error increase (observed in ablation) is reduced by 32% when the transmission cost matrix incorporates domain knowledge (e.g., the weights of the location of document symbols as specified in the format of official documents of party and government organs) (compare with the ablation experiments in Section 4.2). This conclusion provides a new paradigm for cognitive computation of government documents-namely, structure-prior guided cross-modal learning (SPGCL). Future work may explore extensions to domains like healthcare (e.g., modelling medical stamps and diagnostic terms), though this requires domain-specific validation. Statistical analysis confirms: MMHFN's accuracy gains over BERT are significant ($p < 0.001$). The 3.1% accuracy drop under >25% OCR noise is statistically insignificant ($p = 0.12$). All ablation components show significant contributions

Based on the test results of GOV-DOCBench and the deployment experience of the government cloud platform, three practical recommendations are proposed. First, progressive optimisation of the OCR error-correction module is necessary. When the proportion of handwritten annotations is >30% (e.g., historical files), the character error rate (CER) of the Seq2Seq error correction model still reaches 12.7%. The system outputs samples with confidence level < 85% will trigger the manual review process. The review interface synchronously displays the original image, OCR text and model decision basis (e.g., key attention regions), and the reviewer correction results will be fed back to the training pool in real-time. In the pilot, this process shortened the error correction model iteration cycle to 72 hours. It is recommended to incorporate an active learning strategy: utilising the attention weights in the overlay mechanism Identifying low confidence samples (Mei et al., 2018), directed to replenish the training set of handwritten texts. UN e-Government reports that such incremental optimisation can reduce the digitisation cost of county-level archives by 44%. Second, edge-cloud collaborative reasoning architectures are also worth exploring. Although MMHFN achieves 190ms/page real-time processing at the Jetson device, the OCR phase takes up 57.9% of the time. Referring to the federated learning framework of Javed et al. (2020), a hierarchical deployment scheme is proposed: lightweight text extraction (e.g., MobileOCR) runs at the edge and multimodal fusion is handed over to the cloud. This scheme reduces end-to-end latency to 126ms in bandwidth-constrained township government outlets in real-world tests, and privacy-sensitive data (e.g., ID card numbers) do not need to be transmitted out-of-boundary. In the OCR pre-processing stage, regular expression matching (e.g., ID number/cell phone number

patterns) is used for real-time desensitisation, and sensitive fields are replaced with hash values. The desensitised text is used only for the classification task, and the original image is deleted immediately after the edge-side processing is completed. After the third-party audit, the system meets the requirements of the personal information security specification GB/T 35273-2020. Deployment adopts a dual-machine hot standby strategy: when the edge node goes down, the cloud automatically takes over processing tasks (switchover latency < 2 seconds). Historical failure records show that this mechanism guaranteed 99.7% service availability during the Chongqing power grid fluctuation event in 2023, a 40% improvement over single-point deployment. Finally, the dynamic embedding of policy terminology in the library shows potential for domain adaptation, though applications beyond government documents remain unexplored; although BERT fine-tuning improves the semantic understanding of policies, new regulations (e.g., 'twenty data elements') still need to be retrained. It is recommended to build an online terminology embedding update service, modelled on (Tian et al., 2021) domain adaptive mining system, which automatically triggers the generation of embedding vectors and hot update of the model when new high-frequency terms (e.g., 'generative AI regulation') are detected. A pilot at a provincial administrative approval bureau shows that this improves the policy adaptability of the classification model by 38%. Compared to document-specific transformers: DocFormer fails to detect 23% of document numbers in joint-issuance cases. Donut requires 510 ms/page (CPU) while MMHFN runs at 190 ms (edge).

While Engin et al. (2019) multimodal attempt to categorise bank documents is informative, the specificity of government documents requires deeper domain adaptation. At the semantic level, banking terminology is highly standardised (e.g., 'cross-border settlement'), whereas policy vocabulary is time-sensitive and geographically specific (e.g., 'Yangtze River Delta integration' requires contextual understanding). This explains why FastText sub word embeddings show a higher gain in government scenarios (+3.1%) than in banking (+1.7%).At the visual level, the layout of bank notes is governed by international norms (e.g., SWIFT code position is fixed), while the layout of government red-head documents' symbols varies depending on the level of the organisation. Therefore, the optimal transport (DOT) has significantly higher accuracy improvement over late fusion in this study (+2.6%) than the banking scenario (+0.9%). These differences corroborate (Geng et al., 2021) assertion that cross-modal learning must be infused with domain ontology to break through performance bottlenecks.

This study provides a landable technical tool for the construction of smart government affairs. The actual deployment of a provincial government cloud platform proves that the system can process 100,000 documents per day, and it is especially good at handling three types of typical scenarios: seal-masked documents: the SAM module enhances the features of the unmasked area, and the classification accuracy reaches 94.1% ± 0.4; handwritten annotated documents: the OCR error-correction module reduces the CER from 18.7% to 6.3%; and cross-region joint official documents: the policy terminology embedded in the library dynamically. The dynamic updating of the policy terminology embedding library ensures the adaptability of new policies such as 'Yangtze River Delta Integration'.

The limitation of the study is the lack of generalisation ability of historical archives (misclassification of handwritten minutes in the failure case). The root cause is the failure of the Seq2Seq error correction model's coverage mechanism for scribbles (the parameter fails to suppress the cumulative error). Future deployment pathways could include educational settings (e.g., digitising exam papers), subject to empirical validation and student records in face-to-face classrooms using IoT-grade scanners. Further scalability is achievable through cross-departmental federated learning, particularly in healthcare systems for classifying medical documents while preserving data privacy. Future work will focus on three directions: first, develop a visually guided text correction framework to overcome the recognition bottleneck of scribbled handwritten text by utilising visual cues such as official seal position and paragraph spacing to constrain the decoding process; second, build a cross-departmental federated learning ecosystem to allow governmental agencies at all levels to share model increments without data out of the domain ( ) to solve the sample sparsity problem (Kairouz et al., 2021); finally, expanding the support for bilingual official documents of ethnic minorities and exploring cross-modal alignment methods for vertically typeset texts (e.g., Tibetan-Chinese cross-referenced documents) to help the construction of e-government in border areas. Addressing the needs of education in the fourth industrial revolution, Zeeshan et al. (2022) outlines the application of IoT in education from three perspectives: school management, teachers and learners, elucidating its potential while pointing out bottlenecks such as security and privacy. With the deepening of the digital government process, multimodal learning-driven intelligent document management will become the core engine to improve administrative efficiency, and the theoretical cornerstone and technical paradigm laid down in this study will inject continuous kinetic energy into this process.

MMHFN delivers quantifiable advancements: Our framework achieves 92.7% classification accuracy (+4.2% over unimodal baselines) on government documents with 66% lower OCR errors (18.7% → 6.3%) and 45% faster inference than multimodal benchmarks (350ms → 190ms/page). These translate to 67% reduction in manual review costs and reliable handling of noisy real-world documents (94.1% accuracy on seal-occluded cases). Core contributions include the first administrative-rule-guided fusion architecture, a noise-robust dual-path text encoder maintaining F1 = 0.91 under >25% OCR noise, and publicly releasing the GOV-DOCBench with 12,000 annotated documents.

6 Conclusions

In this paper, we propose an MMHFN for intelligent management of government documents, which significantly improves the classification performance in complex scenarios by deeply fusing visual features and semantic information of documents. The core innovation is embodied in three aspects: first, designing a dynamic fusion mechanism guided by administrative elements, utilising the theory of DOT to establish an explicit association between policy terminology and layout specifications, and improving the accuracy by 8.3% compared with the traditional fusion strategy in strongly noisy scenarios, such as citizen's application forms; second, constructing a dual-path robust text encoder, combining FastText subword embedding with domain-adaptive BERT, which maintains a semantic understanding F1-score of 0.91 even under high OCR noise levels (> 25%); and third, developing a lightweight edge deployment scheme to achieve a real-time processing efficiency of 190ms/page based on the SAM of MobileNetV3 to meet the response requirements of governmental affairs systems. Systematic validation on RVL-CDIP, Tobacco3482, and self-built GOV-DOCBench datasets shows that MMHFN outperforms the best existing model by 4.2% (92.7% ± 0.2 vs. 88.5% ± 0.4, $p < 0.001$) and reduces the manual review cost by 67%.

Acknowledgements

This work is supported by the 2024 Guangxi school safety, stability and emergency work research project (No. GXAW2024B009), the 2024 special project of Guangxi Education Science '14th Five Year Plan' (No. 2024ZJY1191).

Declarations

All authors declare that they have no conflicts of interest.

References

Anand, G.S. (2022) 'Application of natural language processing and governance framework for mining financial documents', *SP Jain School of Global Management* (India), Vol. 1, p.31591881.

Bakkali, S., Ming, Z., Coustaty, M., Rusiñol, M. and Terrades, O.R. (2023) 'VLCDoC: vision-language contrastive pre-training model for cross-modal document classification', *Pattern Recognition*, Vol. 139, p.109419.

Balaji, K. (2025) 'E-Government and E-governance: driving digital transformation in public administration', *Public Governance Practices in the Age of AI*, Vol. 1, pp.23–44.

Beale, J., Orebaugh, A. and Ramirez, G. (2006) *Wireshark and Ethereal Network Protocol Analyzer Toolkit*, Vol. 1, p.1, Elsevier.

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) 'Enriching word vectors with subword information', *Transactions of the Association for Computational Linguistics*, Vol. 5, pp.135–146.

Courty, N., Flamary, R., Tuia, D. and Rakotomamonjy, A. (2016) 'Optimal transport for domain adaptation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 9, pp.1853–1865.

Cunningham, A. (2008) 'Digital curation/digital archiving: a view from the National Archives of Australia', *The American Archivist*, Vol. 71, No. 2, pp.530–543.

Cuturi, M. (2013) 'Sinkhorn distances: Lightspeed computation of optimal transport', *Advances in Neural Information Processing Systems*, Vol. 26, p.32.

Engin, D., Emekligil, E., Oral, B., Arslan, S. and Akpınar, M. (2019) 'Multimodal deep neural networks for banking document classification', *International Conference on Advances in Information Mining and Management*, Vol. 2, pp.21–25.

Geng, Y., Chen, J., Chen, Z., Pan, J.Z., Ye, Z., Yuan, Z., Jia, Y. and Chen, H. (2021) 'Ontozsl: Ontology-enhanced zero-shot learning', *Proceedings of the Web Conference 2021*, Vol. 5, pp.3325–3336.

Hasnine, M.N., Nguyen, H.T., Tran, T.T.T., Bui, H.T., Akçapınar, G. and Ueda, H. (2023) 'A real-time learning analytics dashboard for automatic detection of online learners' affective states', *Sensors*, Vol. 23, No. 9, p.4243.

Javed, A., Robert, J., Heljanko, K. and Främling, K. (2020) 'IoTEF: A federated edge-cloud architecture for fault-tolerant IoT applications', *Journal of Grid Computing*, Vol. 18, No. 1, pp.57–80.

Jiao, P., Ouyang, F., Zhang, Q. and Alavi, A.H. (2022) 'Artificial intelligence-enabled prediction model of student academic performance in online engineering education', *Artificial Intelligence Review*, Vol. 55, No. 8, pp.6321–6344.

Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G. and Cummings, R. (2021) 'Advances and open problems in federated learning', *Foundations and Trends® in Machine Learning*, Vol. 14, Nos. 1–2, pp.1–210.

Kumar, P. and Gupta, V. (2023) 'Restoration of damaged artworks based on a generative adversarial network', *Multimedia Tools and Applications*, Vol. 82, No. 26, pp.40967–40985.

Lan, M., Tan, C.L., Su, J. and Lu, Y. (2008) 'Supervised and traditional term weighting methods for automatic text categorization', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 4, pp.721–735.

Lienhart, R. and Wernicke, A. (2002) 'Localizing and segmenting text in images and videos', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 12, No. 4, pp.256–268.

Lyu, P., Zhang, C., Liu, S., Qiao, M., Xu, Y., Wu, L., Yao, K., Han, J., Ding, E. and Wang, J. (2022) 'Maskocr: Text recognition with masked encoder-decoder pretraining', Arxiv preprint Arxiv, Vol. 2206, p.311.

Ma, J., Cheng, T., Wang, G., Zhang, Q., Wang, X. and Zhang, L. (2023) 'Prores: exploring degradation-aware visual prompt for universal image restoration', Arxiv preprint Arxiv, Vol. 2306, p.13653.

Mei, J., Islam, A., Moh'd, A., Wu, Y. and Milios, E. (2018) 'Statistical learning for OCR error correction', *Information Processing and Management*, Vol. 54, No. 6, pp.874–887.

Shafait, F. and Breuel, T.M. (2009) 'A simple and effective approach for border noise removal from document images', *2009 IEEE 13th International Multitopic Conference*, Vol. 1, pp.1–5.

Spasojevic, N. (2021) 'Semantic definition and modern use of the term policy', *Kultura Polisa*, Vol. 18, p.157.

Sulaiman, A., Omar, K. and Nasrudin, M.F. (2019) 'Degraded historical document binarization: a review on issues, challenges, techniques, and future directions', *Journal of Imaging*, Vol. 5, No. 4, p.48.

Tian, K., Zhang, C., Wang, Y., Xiang, S. and Pan, C. (2021) 'Knowledge mining and transferring for domain adaptive object detection', *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Vol. 34, pp.9133–9142.

Tu, Z., Lu, Z., Liu, Y., Liu, X. and Li, H. (2016) 'Modeling coverage for neural machine translation', Arxiv preprint Arxiv, Vol. 1601, p.4811.

Wang, W., Liu, P., Yang, S. and Zhang, W. (2020) 'Dynamic interaction networks for image-text multimodal learning', *Neurocomputing*, Vol. 379, pp.262–272.

Whang, T., Lee, D., Lee, C., Yang, K., Oh, D. and Lim, H. (2019) 'An effective domain adaptive post-training method for bert in response selection', Arxiv preprint Arxiv, Vol. 1908, p.4812.

Woo, S., Park, J., Lee, J-Y. and Kweon, I.S. (2018) 'Cbam: convolutional block attention module', *Proceedings of the European Conference on Computer Vision* (*ECCV*), Vol. 2, pp.3–19.

Ye, J., Hai, J., Song, J. and Wang, Z. (2024) 'Multimodal data hybrid fusion and natural language processing for clinical prediction models', *AMIA Summits on Translational Science Proceedings*, Vol. 2024, p.191.

Zeeshan, K., Hämäläinen, T. and Neittaanmäki, P. (2022) 'Internet of Things for sustainable smart education: An overview', *Sustainability*, Vol. 14, No. 7, p.4293.

Zhang, R., Wei, Z., Shi, Y. and Chen, Y. (2020) 'BERT-al: BERT for arbitrarily long document understanding', *International Conference on Learning Representations*, Vol. 11, p.5220.