Research Article

Gulnar Balakayeva, Mukhit Zhanuzakov*, Uzak Zhapbasbayev, and Kalamkas Nurlybayeva

An intelligent enterprise system with processing and verification of business documents using big data and AI

https://doi.org/10.1515/jisys-2024-0446
received November 14, 2024; accepted June 18, 2025

Abstract: The increasing demand for operational efficiency and data integrity has led enterprises to prioritize the digital transformation of internal workflows. This is done through automation of document-related business processes. This study proposes an intelligent enterprise system that integrates artificial intelligence and big data technologies for the automated generation, validation, and approval of business documents. The motivation behind this work derives from the need to reduce human error, enhance accuracy, and accelerate document turnaround times in enterprise environments. The authors employ large language models to automatically generate document templates and a fine-tuned bidirectional encoder representations from transformer-based classifier for validating document content. Big data tools such as Apache Spark are used for processing and cleaning large volumes of enterprise documents. Additionally, low-confidence predictions are handled through a human-in-the-loop mechanism to ensure high reliability. The research process involves system design, data collection from over 9,000 real enterprise documents, model training, and integration into a business process management system. Experimental results show that the proposed approach improves document processing efficiency while maintaining data quality. This article presents a unified framework and implementation methodology that can be adapted for broader enterprise automation needs. The authors use data from thermal grid enterprises as an example for testing the developed intelligent models.

Keywords: enterprise, digitalization, document generation, business processes, automation, big data, artificial intelligence

1 Introduction

Document creation and authenticity verification are key components of corporate systems that provide operational efficiency and data integrity. Enterprises utilize these components to produce, manage, and safeguard a wide range of documents, such as contracts, invoices, reports, and certificates. This article provides the development of intellectual models for effective document coordination, generation, and verification in the context of enterprise systems. As a testing environment, the authors use a heat-supply digital system.

*** Corresponding author: Mukhit Zhanuzakov**, Department of Computer Science, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan, e-mail: zhanuzakov_mukhit2@live.kaznu.kz
Gulnar Balakayeva: Department of Computer Science, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan, e-mail: gulnar.balakaeva@kaznu.kz
Uzak Zhapbasbayev: Laboratory “Modeling in Energy Sector”, Satbayev University, Almaty 050013, Kazakhstan, e-mail: u.zhapbasbayev@satbayev.university
Kalamkas Nurlybayeva: Department of Computer Science, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan, e-mail: kalamkas.nurlybayeva@kaznu.kz

Coordination of documents is an important part of the work of an enterprise. The main goal of managing this process is to optimize decisions, reduce risks, improve the final result, and control and track the timing (dates) of approvals [1].

Automation of business process management (BPM) is essential for any organization because automated BPM provides the following features:

1. The BPM system acts as a transporter of information between services [2].
2. The BPM system forms a clear scheme of information flow between different departments, systems, etc. [3]
3. In rare cases, the BPM system requires human decision-making.
4. The BPM system allows one to know the current stage of the process at any moment.

1.1 Contributions of this work

This study introduces an intellectual system that uses artificial intelligence (AI) and big data technologies to streamline document generation, validation, and approval processes. One contribution is the integration of large language models (LLMs) for automatic template generation. This allows users to produce standardized business documents from simple natural language descriptions. The system also incorporates a multilingual bidirectional encoder representations from transformer (BERT)-based classifier, fine-tuned specifically to detect errors in documents written in Kazakh and Russian languages. This enables accurate validation and minimizes the risk of incorrect or inconsistent documentation in enterprise operations. Additionally, the study demonstrates the application of Apache Spark for scalable data preprocessing, using different documents to train and evaluate the model. These contributions collectively present a robust and adaptable framework for enhancing digital transformation initiatives across various enterprise contexts.

1.2 Related studies

Usage of AI and big data tools is increasing in many areas of human activities. The following authors analyze different activities dedicated to this field.

The study by Qanbar and Algamal [4] addresses the challenges of imbalanced datasets and influential observations in support vector machine (SVM) classification by integrating with the pigeon optimization algorithm. Experimental results on three real-world datasets demonstrate improved classification accuracy, suggesting potential applications in biological, chemical, and medical fields.

Al-kababchee et al. [5] proposed an improved K-means clustering method using an equilibrium optimization approach. This approach dynamically selects both the number of clusters and relevant attributes. The results across five datasets demonstrate superior clustering performance.

Another study by Al-Kababchee et al. [6] enhanced penalized regression-based clustering using a nature-inspired algorithm to achieve more accurate estimations. Applied to gene expression data, the proposed method demonstrates significant improvements over existing approaches.

Al-Thanoon et al. [7] proposed an enhanced binary crow search algorithm by incorporating an opposition-based learning strategy (OBL-BCSA). This aims to optimize the flight length parameter for effective feature selection. Experiments on two datasets show that OBL-BCSA achieves superior classification accuracy and computational efficiency compared to traditional algorithms.

Al Kababchee et al. [8] introduced an improved penalized regression-based clustering algorithm using a hybrid black hole algorithm to enhance data fusion in clustering tasks. Evaluated on gene expression data, the proposed approach outperforms existing methods.

Esmaeili et al. [9] presented the ResMorCNN model, which combines 3-D convolutional layers with morphological mathematics to enhance feature extraction and classification accuracy in hyperspectral image (HSI) analysis. Tested on four diverse datasets, the model outperforms existing deep learning (DL) methods, with an average accuracy improvement of 3.37%.

Akhtarmanesh et al. [10] introduced an enhanced UNet model with attention blocks for accurate road extraction from aerial images, using a carefully preprocessed and augmented dataset from DeepGlobe. This model achieves 98.33% accuracy, while a detailed precision–recall analysis highlights challenges from dataset bias and informs directions for future improvements.

Marzvan et al. [11] used 30 years of Landsat time-series data and the spectral angular mapper method to analyze the expansion of Azolla in Iran's Anzali Lagoon and its environmental impact. The results revealed fluctuating growth patterns linked to rainfall, water volume, and temperature, highlighting the need for strategic ecosystem management to preserve lagoon biodiversity.

Felegari et al. [12] integrated remote sensing with machine learning to map cadmium (Cd) concentrations using multitemporal Landsat-9 imagery and soil samples from northeastern Kazakhstan. The results showed that support vector regression (SVR) with original band features offers the most accurate estimation, outperforming other models and highlighting the advantages of multitemporal data for heavy metal monitoring.

Another study by Mahdipour et al. [13] addressed inherent uncertainty in land-cover segmentation by modeling image formation and preprocessing processes using high-resolution panchromatic satellite images. Through the proposed "ultra fusion" method and fuzzy C-means clustering, the approach improves segmentation accuracy and efficiency, with notable gains in the overall accuracy, kappa, and $F1$-score of 0.86, 0.52, and 1.03%, respectively.

Safari et al. [14] introduced EddyNet, a DL framework based on a modified U-Net architecture for the automatic identification and classification of oceanic eddies using sea surface temperature data from Copernicus Marine Service. By integrating various convolutional neural network (CNN) backbones and a pixel-wise classification layer, the model achieves strong segmentation accuracy and computational efficiency, supporting real-time oceanographic analysis.

Mirhoseini Nejad et al. [15] presented a hybrid DL model combining ConvLSTM, 3D-CNN, and Vision Transformer to predict the soybean yield using multispectral remote sensing data. The model demonstrates superior accuracy and robustness over existing methods and offers valuable insights for precision agriculture and sustainable crop management across diverse regions.

Farmonov et al. [16] presented HypsLiDNet, a DL framework that integrates HSI with light detection and ranging (LiDAR) data to enhance the crop classification accuracy. By combining spectral and structural information through attention mechanisms and morphological feature extraction, the model outperforms traditional and recent DL methods.

Vafaeinejad et al. [17] introduced a fully automated AI-based system for updating and digitizing agricultural cadastral maps using photogrammetric images. Leveraging the segment anything model (SAM), the system achieves a high segmentation accuracy with a 92% intersection over union (IoU) score, reducing the processing time by 40% and eliminating manual input.

Recent work by Sharifi and Safari [18] presented a transformer-based DL model designed to enhance the spatial resolution of Sentinel-2 satellite imagery. By integrating multiheaded attention with spatial and channel attention mechanisms, the model effectively reconstructs fine details from low-resolution inputs. Evaluated on the Sentinel-2, aerial image dataset, and UC-Merced datasets, the model outperforms leading methods such as ResNet, Swin Transformer, and ViT.

Similar work was done by Dwivedi et al. [19], where the authors explain the development of a neural network for finance department that is intended for amount detection and verification; additionally, it can also extract various entities which contribute to largely performing analytics. The application rewards business in reducing the turnaround time and human errors.

Another work by Baviskar et al. [20] reviewed AI-based techniques for automatic information extraction from unstructured documents. The authors found that the existing methods lack the capability to tackle complex document layouts in real-time situations, and the datasets available publicly are of low quality. This raises the importance of gathering quality datasets.

Table 1 shows the overall key results of comparative analyses of related studies.

In this work, the authors provide different methods for generating and validating documents in BPM systems using these tools.

Table 1: AI and big data approaches in data processing

No.	Authors	Objective	Key techniques/models	Application domain	Key results
1	Qanbar and Algamal [4]	Improve SVM classification on imbalanced data	SVM + pigeon optimization algorithm	Biology and medicine	Improved classification accuracy
2	Al-kababchee et al. [5]	Enhance clustering quality and feature selection	Improved K-means + equilibrium optimization	Multidomain	Better clustering and dynamic attribute selection
3	Al-Kababchee et al. [6]	Refine penalized regression-based clustering	Nature-inspired algorithms	Genomics	More accurate cluster estimations
4	Al-Thanoon et al. [7]	Efficient feature selection	Binary crow search + opposition-based learning	Classification tasks	Higher accuracy and computational efficiency
5	Al Kababchee et al. [8]	Enhance data fusion in clustering	Hybrid black hole algorithm	Gene expression data	Outperformed existing clustering methods
6	Esmaeili et al. [9]	Improve HSI classification	ResMorCNN (3D CNN + morphological ops)	Remote sensing	+3.37% accuracy over baseline models
7	Akhtarmanesh et al. [10]	Road extraction from aerial imagery	Enhanced U-Net + attention blocks	Aerial image analysis	98.33% accuracy; dataset bias identified
8	Marzvan et al. [11]	Monitor plant expansion and environmental impact	Landsat Time Series + SAM	Environmental studies	Revealed seasonal growth patterns
9	Felegari et al. [12]	Map heavy metal concentrations (Cd)	SVR + multitemporal Landsat-9	Environmental monitoring (Kazakhstan)	SVR with original bands performed best
10	Mahdipour et al. [13]	Address uncertainty in land-cover segmentation	Ultrafusion + fuzzy C-means clustering	Satellite imagery	Improved accuracy, kappa, and *F*1-score
11	Safari et al. [14]	Identify and classify ocean eddies	EddyNet (modified U-Net)	Oceanography	High segmentation accuracy and real-time capable
12	Mirhoseini Nejad et al. [15]	Predict soybean yield	ConvLSTM + 3D-CNN + vision transformer	Precision agriculture	Outperformed existing models in accuracy
13	Farmonov et al. [16]	Enhance crop classification	HypsLiDNet (HSI + LiDAR)	Agricultural analytics	Superior to recent DL and classical methods
14	Vafaeinejad et al. [17]	Update and digitize cadastral maps	SAM	Land mapping	92% IoU and 40% reduction in processing time
15	Sharifi and Safari [18]	Super-resolve satellite imagery	Transformer + spatial/channel attention	Remote sensing	Outperformed ResNet, Swin, and ViT
16	Dwivedi et al. [19]	Finance document automation	NN for amount detection + NER	Financial departments	Reduced turnaround time and errors
17	Baviskar et al. [20]	Extract information from unstructured docs	Survey of AI techniques	Document analytics	Highlighted lack of robust real-time datasets
18	Our work	Document generation and validation	BERT	Various enterprise systems and BPM systems	82% validation accuracy and document generation using AI

2 Methods

2.1 Assessing the reliability of information, identifying erroneous data: Anomalies, duplicates, contradictions, empty values, and correcting identified errors

In this section, the authors present methods for information reliability and identification of erroneous data. To support this, the system includes several built-in forms. One key example is the user registration and authorization process [21]. During registration and login, the data entered by users are validated on both the client and server sides. If incorrect information is provided (such as an invalid email or name), the server returns an error message. The following are the main rules used to detect data errors:

- Full name – must consist of at least two characters and cannot be empty;
- Email – using a regular expression, the authenticity of the email is checked. To check the entered text, a certain regular expression RegEx was used, which returns false if the text is not an actual e-mail;
- Password – the password must be a minimum of 8 characters and must contain capital letters and special characters. This is one of the main points of implementing information system security. If users have too light passwords, they can be hacked by guessing the password, which will lead to data leakage;
- Other fields – we managed to create conditions so that most fields were not empty and were not created and edited in the database with empty values. But there were still some fields that, when created, stored type null or an empty string.

The Builders<Statement>.Filter() method was used to identify the existing data. Before creating or updating any records in the database, the service first checks whether the element already exists. If a matching element is found, the system prompts the user to navigate to the existing object. If not, an error message is returned. Since objects may vary, it is necessary to implement duplicate checks across all POST, PATCH, and PUT requests. This was achieved by passing the object from the request body to the Filter() method [22].

Fault tolerance and system safety were implemented using .NET Core. A dedicated subsystem was developed to manage the organization's personnel activities in line with the overall system architecture. A user interface was also created to support these activities, with key components identified and presented in a user-friendly format [23]. For the analysis and processing of large-scale personnel data, Apache Spark was employed. It ensures efficient processing even in cases of instability or failures, with the processed data subsequently stored in the database. Horizontal scalability is fully supported by Apache Spark. The system includes components for both large-scale and streaming data storage. Finally, the output of the enterprise activity management subsystems was processed and analyzed.

Based on the classification of criteria for assessing the quality of information, we highlight the following criteria: authenticity, usefulness, completeness, unambiguity, significance, correctness, consistency, timeliness, source, format, and content [24].

The effectiveness of analyzing the information quality of a digital corporate system requires effective ways to obtain reliability metrics implemented during system testing, which will have a positive impact on improving the quality of reliability assessment in the future. We will divide methods for assessing the reliability of information into two groups: heuristic – used by auditors – and formal – operating with resources, processes, and the information system as a whole. When assessing the reliability of information in the databases of a digital enterprise system, the analysis focuses on the quality of the entity-relationship graph. In this context, relationships are evaluated based on the paths they form within the graph. With this approach, primitive defects, such as non-existent indexes, are eliminated simply using data cleaning tools [25].

The general model for assessing the reliability of a digital enterprise system focuses on analyzing data flows. It evaluates several parameters across different stages, including data collection, input, processing, storage, transmission, and delivery. This model is designed to detect and prevent various types of errors, both in intermediate processes and final outputs.

2.2 Development of a business process for coordination and approval of documents

Coordination of documents is an integral part of the work of an enterprise. The main goal of managing this process is to optimize decision-making, reducing risks, and improving the final result.

Document approval is the most frequently used business process in any organization. The main goal of the business process for the organization is to control and track the timing of approvals.

Examples of approvals include the following:

(1) The sales department is preparing a draft contract for the supply of a large batch of products manufactured by the enterprise. The legal director denies the allegation. This not only improves the quality of the contract but also insures the company against long-term troubles: claims from the counterparty and other [26].
(2) Preparation of tender documentation. Each interested department of the organization must approve or reject the draft documents that form the contractual basis.
(3) Coordination of the procurement process. The purchase of goods and services by any organization requires a vote by the tender commission. Each committee member must vote to agree or disagree with the selection of the supplier company. The result of the approval process is shown to the accounting department for the payment process.
(4) Each organization approves or re-approves internal documents, charters, regulations, and strategies that require mandatory approval by the top management and shareholders.

Most companies are faced with the task of approving a document. The following stages of the document approval process are proposed:

(1) Start of the process – approval of a document indicating the type of document. The initiator fills out the main form for approval.
(2) Document preparation – the form is completed based on a predefined template specific to the document type. For example, a reference document includes fields such as reference number, responsible executor, reference text, sender, recipient, and execution date. A contract, on the other hand, requires the contract number, date of conclusion, agreement text, and an electronic version of the document. The system automatically sets the time frame for review and approval based on the document type: agreement – 5 working days, order – 3 working days, instruction – 1 working day, and memo – 6 working days.
(3) Identification of the initiator's supervisor for approval – once the document is submitted, it is assigned the status "Under approval by the manager." The initiator's supervisor then reviews the document and either approves it or rejects it with comments. If rejected, the document is returned to the initiator for revision. If approved, it proceeds to the next stage and is assigned the status "Under agreement with the performers."
(4) Approval by the Head of the Contractor's Department – once the performer is selected, the head of the contractor's department reviews and approves the document. At the same time, an execution subprocess is initiated in parallel for all assigned executors.
(5) After the functions of all participants are completed, the result is saved in the database.
(6) Notifying managers about execution.
(7) Registration of documents. Assigning a number in accordance with the classifier and nomenclature of the organization's affairs.
(8) Ending the process.

All stages of the document approval business process are subject to auditing. Once approved, documents are archived, and the digital archive system enables efficient search and retrieval of all relevant information. A key feature of the system is that, at every stage of the process, users have the option to download, print, or return the document for revision [27–31].

2.3 Process of validation and submission of documents

To submit a document for approval and signing, it is necessary to fill in the minimum set of document details such as title, file, and executors. After that, document preparation and execution will include the following processes: validation, download, and sending depending on the selected type of approval (sequential/parallel).

Each document can also have its own approval route. Advantages of using document routes are as follows:

- Document movement is programed. Each document goes through the necessary stages of approval, passing from one user to another;
- Control of document movement. It is always possible to track where the document is and what is its current stage;
- Many stages of the document lifecycle can be automated, for example, change of statuses, sending for approval or review, etc.

Document generation contains the following fields to be filled in:

- Name – the name of the document to be displayed in further reports and stored in the MongoDB database, in the process documentation, and process regulations. The same name will be assigned to the user's button for switching to the document, if no other name is defined for switching in the process;
- Description – additional description of the document. It is displayed in the document generation.

A new template can be added by clicking the "generate" button. Since the subsystem has the functionality of template generation using AI (see Section 2.4) in the system, the template will be generated immediately. A template file can be of the following formats: DOC and DOCX. One can also upload an existing document in the following formats: TXT, HTML, XML, DOC, XLS, DOCX, and XLSX.

2.4 Generating a document template using LLMs

The developed system implements the generation of document templates using LLMs. This function is recommended only for generating internal documents of the enterprise.

To generate, one needs to click the "Generate" button, then a window will appear with the "Describe the document" field. After entering a description of the document, user has to click "Generate", then a document template will appear below (Figure 1).

This function allows one to generate a document template from a description using the AI model ChatGPT 3.5. This speeds up the document creation process and is one of the advantages of the developed system.

To connect to OpenAI, we have the API, the key that was used in the connection configuration. Next, when entering and clicking the "Generate" button, we sent the following parameters:

{
model: "text-davinci-003",
prompt: ' Create a document template in html format for employee ${user.fullName} using the following description: ${description}.
max_tokens: 1,000,
temperature: 0.4
}

The model parameter specifies the model type. In our case, the "text – davinci -003" model is perfect. "text-davinci-003" includes the following improvements:

- It ensures high-quality writing and helps apps deliver clearer, more engaging, and compelling content.
- It can handle more complex instructions.
- It is better suited for creating content in longer forms, allowing us to accomplish tasks that would previously have been too difficult to achieve.

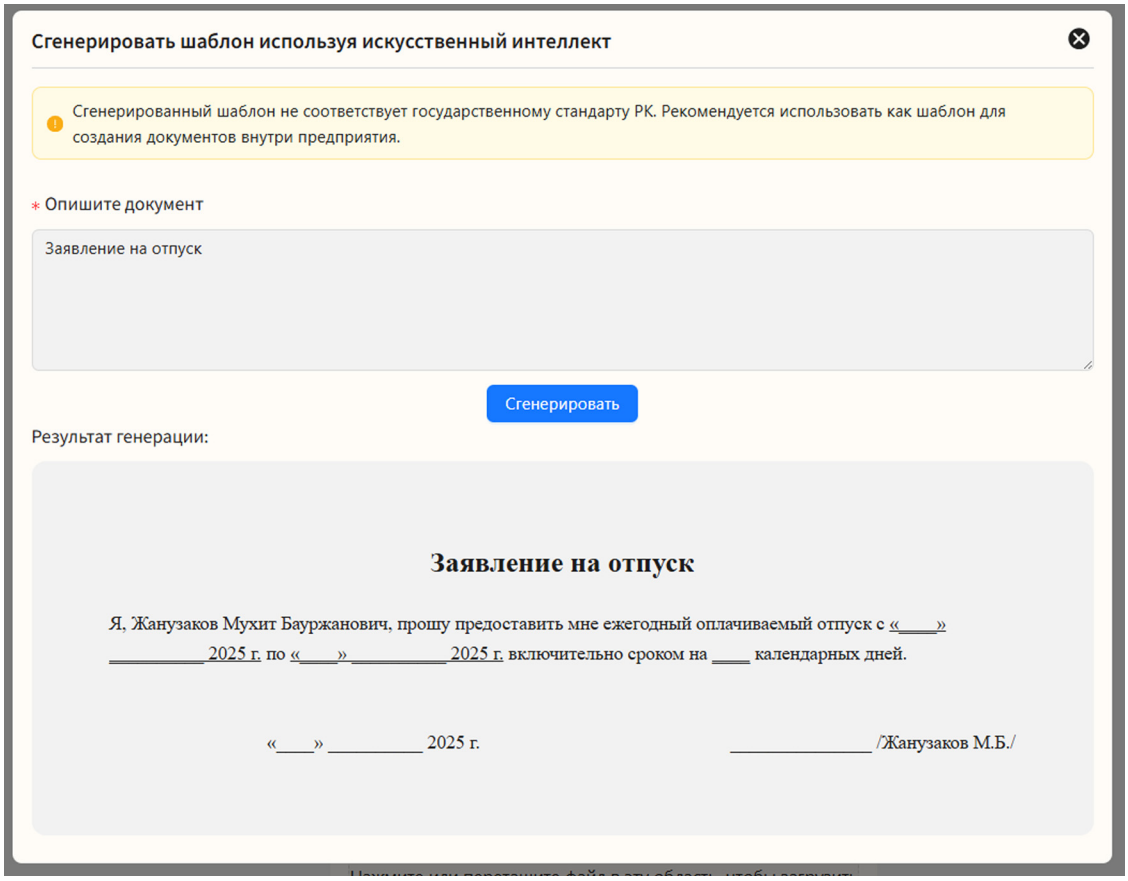

Figure 1: Interface for generating a document template using AI in a heat network enterprise. Source: Created by the authors.

To ensure the model generates the template correctly, we included a predefined template in the prompt field. The “max_tokens” parameter defines the maximum number of tokens the model can generate. In our case, a limit of 1,000 tokens is sufficient for producing a complete template.

The “temperature” field for this API specifies which sample temperature to use, from 0 to 2. Higher values, such as 0.8, will make the output more random, while lower values, such as 0.2, will make it more targeted and deterministic. The intermediate value of 0.4 was chosen to generate documents.

2.5 Error validation using BERT and big data processing

Validation is necessary because it is aimed at reducing the number of erroneous generation of templates and attaching incorrect templates for approval.

2.5.1 Model selection

In order to validate if the document does not contain errors, we use BERT to classify documents into “with errors” – 1 and “no errors” – 0. As a pre-trained language model, BERT is able to recognize syntax errors. However, as the official documents are written in a business writing style, we need to train the model to recognize stylistic errors. The language of these documents is either Kazakh or Russian. That is why BERT multilingual was selected as a pre-trained model.

2.5.2 Data gathering and cleaning using big data technologies

When using AI to predict error-prone document, there are several ways to improve the model:

- Quality of data sources: Reliable and credible data sources should be selected [31]. This may include checking the authority of the source, reputation of its creator, and using trusted data sources such as official statistical databases or academic publications.
- Diversity of sources: It is important to use information from a variety of sources to reduce the likelihood of misrepresentation or incorrect data [32–35]. A variety of sources will help provide a complete and more objective picture of the topic.
- Fact-checking and clarification: One need to use AI capabilities to fact-check and clarify information. There are tools that can automatically check the validity of data and provide recommendations on how to use it [36–38].
- Training a model on reliable data: When using machine learning models or neural networks to generate a document, it is important to train the model on reliable and validated data. The better the model is trained, the less likely it is to present incorrect data.
- Human validation: A human validation and editing stage should be included to correct possible errors or inaccuracies that may arise from the use of AI.

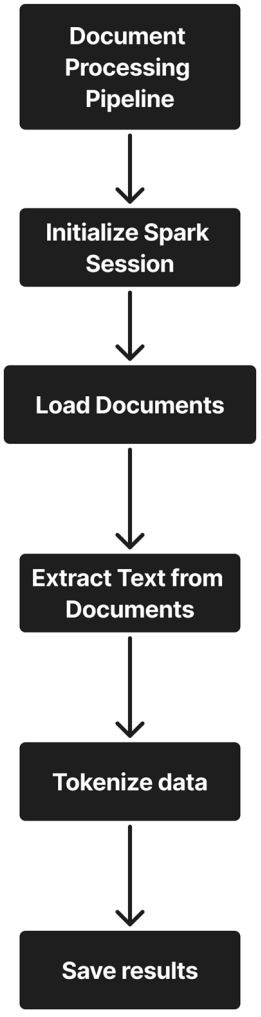

Figure 2: Data processing using Apache Spark. Source: created by the authors.

– Feedback and improvement: There is a need to collect feedback from users and correct errors or inaccuracies that may be identified during the process of using AI to generate a document. Continuous improvement of the system will help reduce the risks of presenting incorrect data in the future.

In order to train the model, the authors prepared various text samples in business writing style. A total of 9,257 documents were gathered by authors from different Kazakhstani Enterprise Document systems. All documents were .docx/.doc files. Before using it for training, all texts were extracted using "python-docx" library, which allows parsing word documents. Documents were cleaned using Apache Spark (Figure 2). After processing, Apache found 255 faulty documents that were removed from the list.

Documents with errors were labeled manually. The parsed data are turned into "pandas" data-frame for further usage (Figure 3). After extraction, the text is encoded using label encoder and split into training and testing sets (80 and 20%, respectively).

text	label
Прошу предоставить очередной оплачиваемый отпу	0
На основании приказа №127-ЛС от 15.04.2024 про...	0
В связи с производственной необходимостью напр...	0
На основании распоряжения №45 от 20.03.2024 пр...	1
Настоящим подтверждаю получение материальных	0

Figure 3: First five rows of gathered data. Source: created by the authors.

2.5.3 Model training

The hyperparameters of model are displayed in Table 2. The parameters chosen for the BERT classifier are carefully selected to balance the performance and resource efficiency. The "evaluation_strategy" is set to epoch. This means the model is evaluated at the end of each training epoch, which is a practical frequency for most fine-tuning tasks. A "learning_rate" of 2×10^{-5} is commonly recommended for BERT. It allows gradual fine-tuning without drastically altering pretrained weights. Both "per_device_train_batch_size" and "per_device_eval_batch_size" are set to 8 to accommodate memory constraints while ensuring sufficient gradient updates. The training runs for 5 epochs, which is typically enough for the model to learn meaningful patterns without overfitting. Finally, a "weight decay" of 0.01 is applied to help regularize the model and improve generalization by penalizing large weights.

Table 2: Hyperparameters for training BERT

Parameter	Value
Output_dir	./results
Evaluation_strategy	Epoch
Learning_rate	2×10^{-5}
Per_device_train_batch_size	8
Per_device_eval_batch_size	8
Num_train_epochs	5
Weight_decay	0.01
Logging_dir	./logs

2.5.4 Model architecture

Figure 4 illustrates the architecture of a BERT-based binary classification model, divided into three main stages: input processing, BERT processing, and classification head. In the input processing phase, raw text is first tokenized into subword units and then converted to numerical token IDs that the BERT model can understand. Special tokens such as CLS and SEP are added to signal the beginning and end of input sequences. Padding is then applied to ensure that all inputs are of the same length, and an attention mask is created to help the model ignore padded positions during processing. The prepared input is passed into the BERT processing block, where it first goes through an embedding layer, followed by 12 transformer layers that generate contextual hidden states for each token. From these, the pooled output, typically derived from the CLS token, is extracted to represent the whole sequence. This pooled output is sent into the classification head, which includes a dropout layer with a dropout rate of 0.3 to prevent overfitting, a linear layer that maps features to output dimensions, and a softmax function that converts the output into probability scores. The final result is a binary classification output indicating either an error or no error, which can be applied to tasks such as error detection or sentiment analysis.

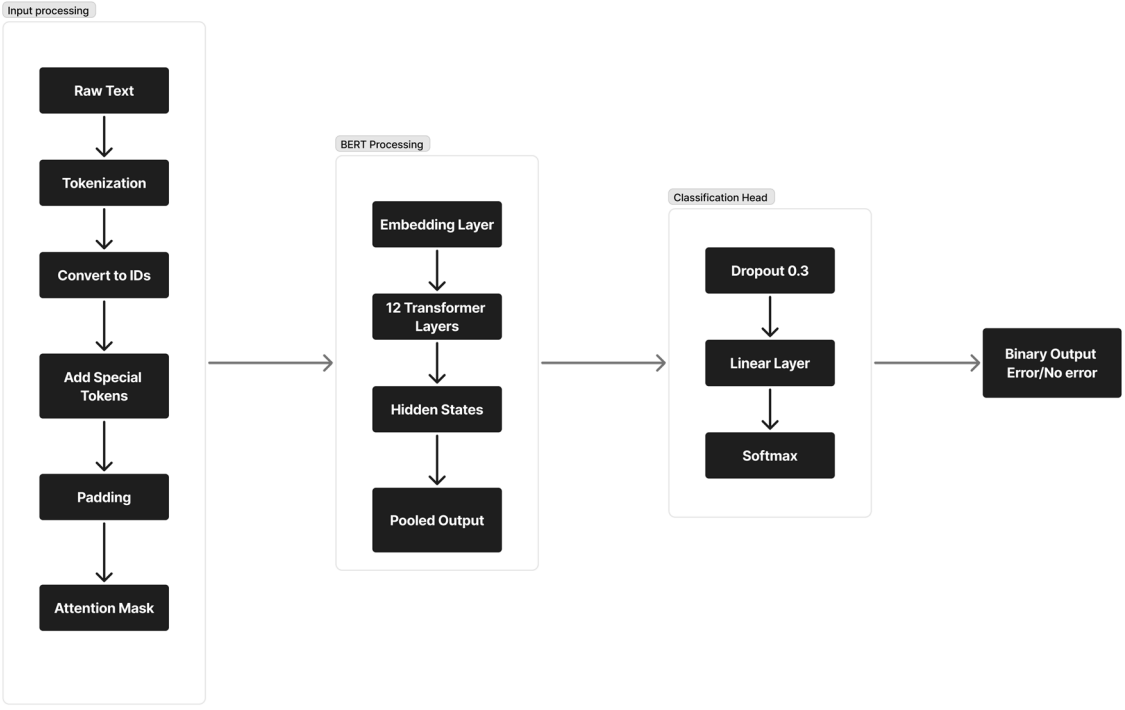

Figure 4: Model architecture. Source: created by the authors.

2.6 System architecture

Overall, the developed system architecture is shown in Figure 5. When a user logs into the system, their identity is validated, if successful, the user is granted access to the web interface for working with the BPM system. BPM consists of the route manager, approval tracker, and workflow engine. Workflow engine is used to generate, validate, and process documents. Route manager simply submits the successfully generated documents to other users of the system. Approval tracker controls the approval and signing processes of the documents. Documents are stored in a MongoDB database for better performance as documents are stored in unstructured format (base64).

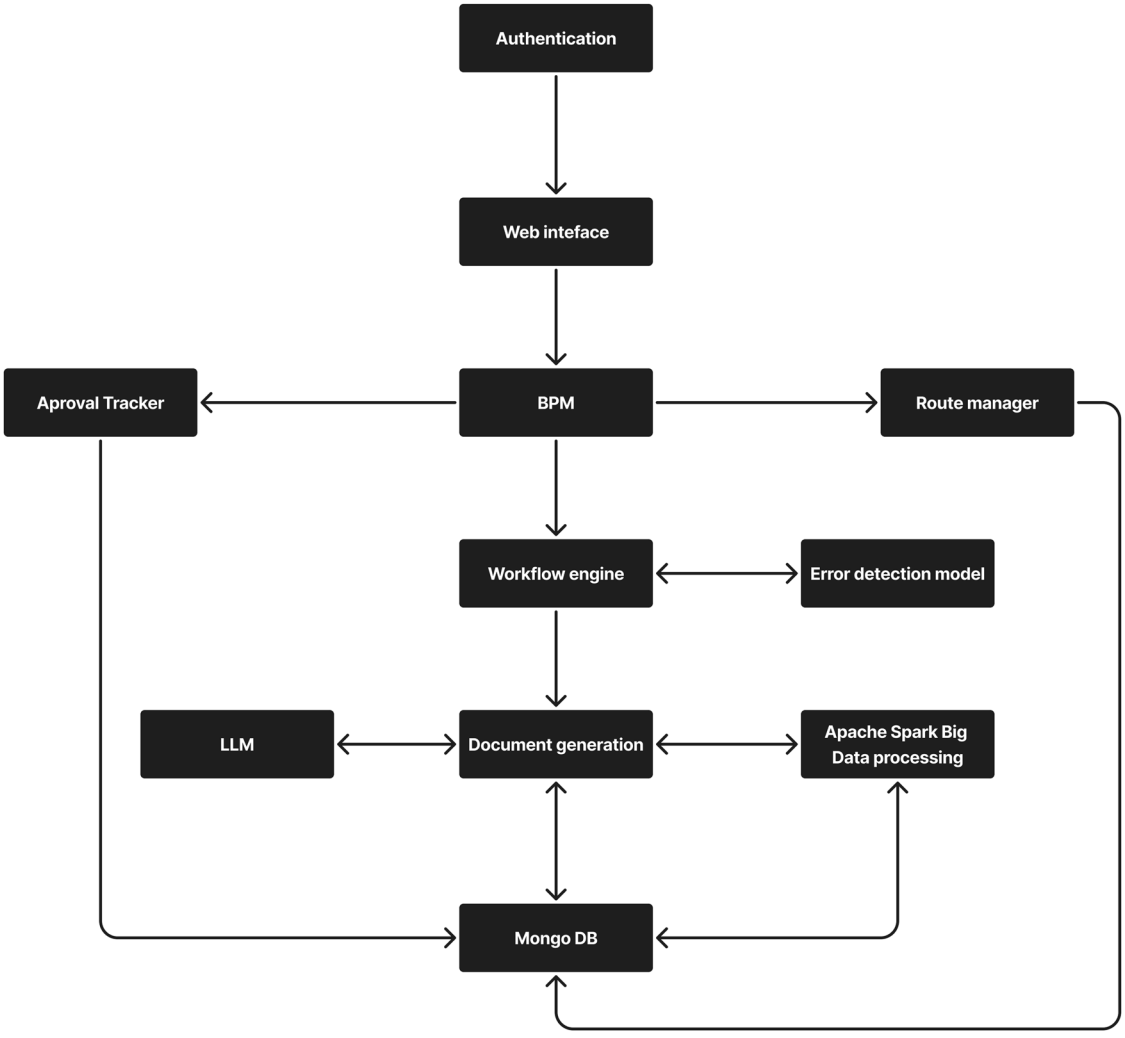

Figure 5: System architecture. Source: created by the authors.

3 Results and discussion

3.1 Performance evaluation

The analysis of the model's performance results indicates strong potential for practical application, particularly in detecting errors within documents. After five training epochs, the model achieved a training accuracy of about 95% and a validation accuracy of 82%, with an average training loss of 0.15. This demonstrates effective learning with a relatively low risk of overfitting. After training the model, we prepared a classification report that can be seen in Table 3.

Table 3: Classification report

	Precision	Recall	*F*1
Incorrect (0)	0.77	0.91	0.83
Correct (1)	0.94	0.55	0.69
Macro avg	0.88	0.77	0.79
Weighted avg	0.86	0.82	0.81

The classification report further supports the model's capability: for class 0, which represents incorrect documents, it achieved a precision of 0.77, recall of 0.91, and *F*1-score of 0.83. These values show that the model is especially reliable at identifying erroneous documents.

For class 1 (correct documents), the precision is high at 0.94, but the recall drops to 0.55, resulting in a lower *F*1-score of 0.69. This suggests that while the model is confident in labeling correct documents, it sometimes fails to recognize them.

The macro average *F*1-score of 0.79 and the weighted average *F*1-score of 0.81 confirm the overall balanced performance, although there is room for improvement in recall for class 1.

These results suggest that the model is better at flagging errors than confirming correctness, which is acceptable in quality assurance contexts where it is preferable to over-flag than to miss actual errors.

The integration of this model as a microservice ensures it can be deployed effectively in enterprise environments, automatically analyzing newly generated or edited documents to enhance reliability and efficiency in document validation workflows. Now, in order to utilize the model, we added the model to our application as a microservice that triggers as soon as a new document is generated or created manually.

3.2 Handling low-confidence predictions

To address low-confidence predictions and ambiguous data patterns, the system applies a confidence threshold mechanism. Each document classified by the BERT model is assigned a confidence score (accuracy). If this score is lower than 0.7, the document is flagged as "uncertain." Then, these documents with ambiguous

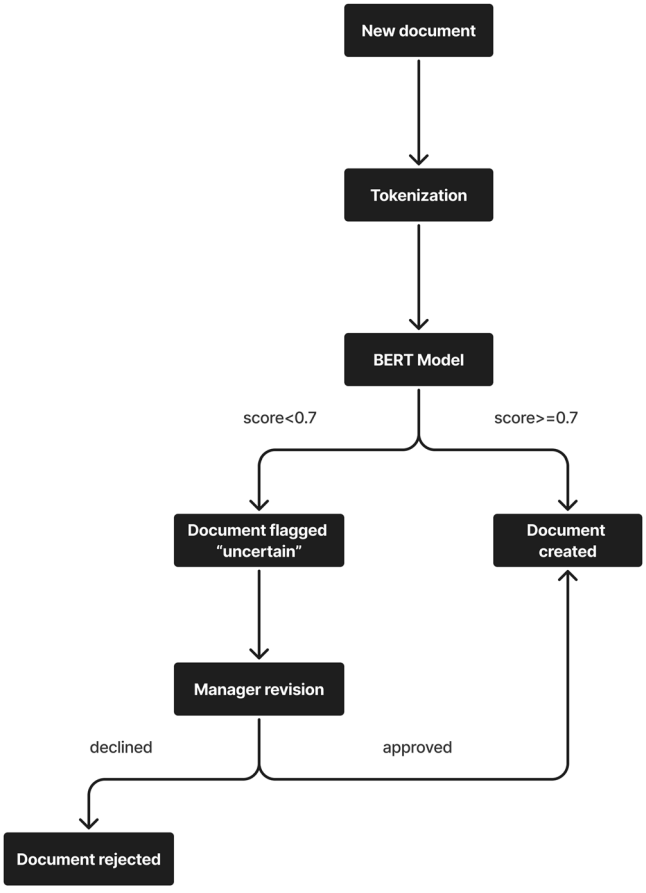

Figure 6: Document validation flow. Source: created by the authors.

content or low model confidence are forwarded to human managers. Their feedback ensures continuous improvement of the model and confidence of the predictions. Using this, newly generated documents are processed and checked for errors (Figure 6).

4 Conclusions

Considering the transformative impact of digitalization on enterprise productivity, this study proposes intelligent models that leverage AI and big data tools. The key innovation lies in applying AI techniques to automate and validate document generation processes. A BERT model was trained to detect errors in text, achieving over 82% accuracy on the validation set. Big data technologies were employed to handle the large-scale data required for AI training. It is important to acknowledge that AI systems are not flawless and may still produce errors [37,38]. Therefore, the proposed methods are designed to reduce these risks and enhance the quality of enterprise documentation. The overall system design is presented, detailing the various modules that interact with BPM workflows. The system is built using modern technologies and is tailored to enterprise operations, with the potential to be adapted for both internal and external business processes.

Limitation of the purposed model is the multilingual challenges when dealing with idiomatic expressions or mixed-language content. This issue will be the future focus for the authors. In order to solve this, the authors plan to further expand and diversify the training dataset with a broader range of documents that include formal, informal, and idiomatic language in both languages.

Funding information: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant #BR24992907).

Author contributions: Gulnar Balakayeva: conceptualization, formal analysis, writing – review and editing, supervision. Mukhit Zhanuzakov: methodology, data curation, writing – original draft, visualization, formal Analysis. Uzak Zhapbasbayev: investigation, validation, writing – review and editing. Kalamkas Nurlybayeva: data curation, resources, project administration, writing – review and editing.

Conflict of interest: The authors declare that there is no conflict of interest regarding the publication of this article.

Data availability statement: At the time of publication, the data used in this research was closed source due to its integral role in an ongoing scientific project. Once the project is completed in 2026, the authors will make the data available upon reasonable request.

References

[1] Sutherland Global. What is BPM. USA: Sutherland. Retrieved May 27, 2024. https://www.sutherlandglobal.com/insights/technology/what-is-bpm.

[2] Ravesteijin P, Zoet M. A BPM-systems architecture that supports dynamic and collaborative processes. J Int Technol Inf Manag. 2010;19(3):1. doi: 10.58729/1941-6679.1083.

[3] Jung J, Choi I, Song M. An integration architecture for knowledge management systems and business process management systems. Comput Ind. 2007;58(1):21–34. doi: 10.1016/j.compind.2006.03.001.

[4] Qanbar AA, Algamal ZY. Improving Support vector machine for Imbalanced big data classification. J Intell Syst Internet Things. 2024;11(2):22–9. doi: 10.54216/JISIoT.110202.

[5] Al-kababchee S, Algamal Z, Qasim O. Enhancement of K-means clustering in big data based on equilibrium optimizer algorithm. J Intell Syst. 2023;32(1):20220230. doi: 10.1515/jisys-2022-0230.

[6] Al-Kababchee SGM, Qasim OS, Algamal ZY. Improving penalized regression-based clustering model in big data. In Journal of Physics: Conference Series. Vol. 1897, No. 1, IOP Publishing; 2021. p. 012036. doi: 10.1088/1742-6596/1897/1/012036.

[7] Al-Thanoon NA, Algamal ZY, Qasim OS. Feature selection based on a crow search algorithm for big data classification. Chemom Intell Lab Syst. 2021;212:104288. doi: 10.1016/j.chemolab.2021.104288.
[8] Al Kababchee SG, Algamal ZY, Qasim OS. Improving penalized-based clustering model in big fusion data by hybrid black hole algorithm. Fusion: Pract Appl. 2023;11(1):70–6. doi: 10.54216/FPA.110105.
[9] Esmaeili M, Abbasi-Moghadam D, Sharifi A, Tariq A, Li Q. ResMorCNN model: hyperspectral images classification using residual-injection morphological features and 3DCNN layers. IEEE J Sel Top Appl Earth Obs Remote Sens. 2023;17:219–43. doi: 10.1109/JSTARS.2023.3328389.
[10] Akhtarmanesh A, Abbasi-Moghadam D, Sharifi A, Yadkouri MH, Tariq A, Lu L. Road extraction from satellite images using attention-assisted UNet. IEEE J Sel Top Appl Earth Obs Remote Sens. 2023;17:1126–36. doi: 10.1109/JSTARS.2023.3336924.
[11] Marzvan S, Moravej K, Felegari S, Sharifi A, Askari MS. Risk assessment of alien Azolla filiculoides Lam in Anzali Lagoon using remote sensing imagery. J Indian Soc Remote Sens. 2021;49:1801–9. doi: 10.1007/s12524-021-01362-1.
[12] Felegari S, Sharifi A, Khosravi M, Sabanov S. Using experimental models and multitemporal Landsat-9 images for cadmium concentration mapping. IEEE Geosci Remote Sens Lett. 2023;20:1–4. doi: 10.1109/LGRS.2023.3291019.
[13] Mahdipour H, Sharifi A, Sookhak M, Medrano CR. Ultrafusion: Optimal fuzzy fusion in land-cover segmentation using multiple panchromatic satellite images. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:5721–33. doi: 10.1109/JSTARS.2024.3360648.
[14] Safari MM, Sharifi A, Mahmood J, Abbasi-Moghadam D. Mesoscale eddy detection and classification from sea surface temperature maps with deep neural networks. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:10279–90. doi: 10.1109/JSTARS.2024.3402823.
[15] Mirhoseini Nejad SM, Abbasi-Moghadam D, Sharifi A. ConvLSTM-ViT: A deep neural network for crop yield prediction using Earth observations and remotely sensed data. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:17489–502. doi: 10.1109/JSTARS.2024.3464411.
[16] Farmonov N, Esmaeili M, Abbasi-Moghadam D, Sharifi A, Amankulova K, Mucsi L. HypsLiDNet: 3D-2D CNN model and spatial–spectral morphological attention for crop classification with DESIS and LiDAR data. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:11969–96. doi: 10.1109/JSTARS.2024.3418854.
[17] Vafaeinejad A, Alimohammadi N, Sharifi A, Safari MM. Super-resolution AI-based approach for extracting agricultural cadastral maps: form and content validation. IEEE J Sel Top Appl Earth Obs Remote Sens. 2025;18:5204–16. doi: 10.1109/JSTARS.2025.3530714.
[18] Sharifi A, Safari MM. Enhancing the spatial resolution of sentinel-2 images through super-resolution using transformer-based deep learning models. IEEE J Sel Top Appl Earth Obs Remote Sens. 2025;18:4805–20. doi: 10.1109/JSTARS.2025.3526260.
[19] Dwivedi A, Vijayan P, Gupta R, Ramdasi P. Enhancing enterprise business processes through AI based approach for entity extraction–an overview of an application. In Recent Trends in Image Processing and Pattern Recognition: Third International Conference, RTIP2R 2020, Aurangabad, India, January 3–4, 2020, Revised Selected Papers, Part I 3. Springer Singapore; 2021. p. 373–80. doi: 10.1007/978-981-16-0507-9_32.
[20] Baviskar D, Ahirrao S, Potdar V, Kotecha K. Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. IEEE Access. 2021;9:72894–936, doi: 10.1109/ACCESS.2021.3072900.
[21] Rhem AJ. Ethical use of data in AI applications. In Ethics – Scientific Research, Ethical Issues, Artificial Intelligence and Education [Working Title]. London, UK: IntechOpen; 2023. doi: 10.5772/intechopen.1001597.
[22] DeVerna MR, Yan HY, Yang K-C, Menczer F. Fact-checking information generated by a large language model can decrease news discernment (v1). Proc Natl Acad Sci USA. 2023;121:1–48. doi: 10.1073/pnas.2322823121.
[23] Meng F, Wang W. The impact of digitalization on enterprise value creation: An empirical analysis of Chinese manufacturing enterprises. J Innov Knowl. 2023;8:100385. doi: 10.1016/j.jik.2023.100385.
[24] Gupta A, Tung YA, Marsden JR. Digital signature: use and modification to achieve success in next generational e-business processes. Inf Manag. 2004;41(5):561–75. doi: 10.1016/S0378-7206(03)00090-9.
[25] Chang SE, Chen YC, Wu TC. Exploring blockchain technology in international trade: Business process re-engineering for letter of credit. Ind Manag Data Syst. 2019;119(8):1712–33. doi: 10.1108/IMDS-12-2018-0568.
[26] Pérez-Álvarez JM, Gómez-López MT, Eshuis R, Montali M, Gasca RM. Verifying the manipulation of data objects according to business process and data models. Knowl Inf Syst. 2020;62(7):2653–83. doi: 10.1007/s10115-019-01431-5.
[27] Balakayeva G, Ezhilchelvan P, Makashev Y, Phillips C, Darkenbayev D, Nurlybayeva K. Digitalization of enterprise with ensuring stability and reliability. Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska. 2023;13(1):54–7. doi: 10.35784/iapgos.3295.
[28] Viriyasitavat W, Da Xu L, Bi Z, Pungpapong V. Blockchain and internet of things for modern business process in digital economy – the state of the art. IEEE Trans Comput Soc Syst. 2019;6(6):1420–32, doi: 10.1109/TCSS.2019.2919325.
[29] Zacharewicz G, Diallo S, Ducq Y, Agostinho C, Jardim-Goncalves R, Bazoun H, et al. Model-based approaches for interoperability of next generation enterprise information systems: state of the art and future challenges. Inf Syst e-Bus Manag. 2017;15:229–56. doi: 10.1007/s10257-016-0317-8.
[30] Chauhan P, Verma JP, Jain S, Rai R. Blockchain based framework for document authentication and management of daily business records. In Blockchain for 5G-Enabled IoT: The new wave for Industrial Automation. Cham: Springer; 2021. p. 497–517. doi: 10.1007/978-3-030-67490-8_19.
[31] Deshpande A, Kumar M. Artificial intelligence for big data: Complete guide to automating big data solutions using artificial intelligence techniques. Birmingham: Packt Publishing Ltd; 2018. doi: 10.5555/3265030.

[32] Mahmood HS. Conducting in-depth analysis of AI, IoT, web technology, cloud computing, and enterprise systems integration for enhancing data security and governance to promote sustainable business practices. J Inf Technol Inform. 2024;3(2):297–322. https://www.researchgate.net/profile/Dildar-Abdulqadir/publication/383087255_Conducting_In-Depth_Analysis_of_AI_IoT_Web_Technology_Cloud_Computing_and_Enterprise_Systems_Integration_for_Enhancing_Data_Security_and_Governance_to_Promote_Sustainable_Business_Practices/links/66bdb090311cbb094939611f/Conducting-In-Depth-Analysis-of-AI-IoT-Web-Technology-Cloud-Computing-and-Enterprise-Systems-Integration-for-Enhancing-Data-Security-and-Governance-to-Promote-Sustainable-Business-Practices.pdf.
[33] Azman NA, Mohamed A, Jamil AM. Artificial intelligence in automated bookkeeping: a value-added function for small and medium enterprises. JOIV: Int J Inform Vis. 2021;5(3):224–30. doi: 10.30630/joiv.5.3.669.
[34] Mah PM, Skalna I, Muzam J. Natural language processing and artificial intelligence for enterprise management in the era of industry 4.0. Appl Sci. 2022;12(18):9207. doi: 10.3390/app12189207.
[35] Yathiraju N. Investigating the use of an artificial intelligence model in an ERP cloud-based system. Int J Electr Electron Comput. 2022;7(2):1–26. doi: 10.22161/eec.72.1.
[36] Bharadiya JP. A comparative study of business intelligence and artificial intelligence with big data analytics. Am J Artif Intell. 2023;7(1):24. doi: 10.11648/j.ajai.20230701.14.
[37] Balakayeva G, Zhanuzakov M, Kalmenova G. Development of a digital employee rating evaluation system (DERES) based on machine learning algorithms and 360-degree method. J Intell Syst. 2023;32(1):20230008. doi: 10.1515/jisys-2023-0008.
[38] Helo P, Hao Y. Artificial intelligence in operations management and supply chain management: An exploratory case study. Prod Plan Control. 2022;33(16):1573–90, doi: 10.1080/09537287.2021.1882690.