10.48047/jocaaa.2025.34.12.86

Natural Language Processing for Automated CMS Compliance Documentation

Sravan Kumar Nidiganti
Senior Manager, Quality Management, Benefits and Clinical Operations
CVS Health, Irving, Texas 75039.
sravankumarnidiganti@gmail.com

Received: 05.10.2025 Revised: 15.11.2025 Accepted: 31.12.2025

Abstract
NLP is automating the generation and updating of documents for Centers for Medicare and Medicaid Services (CMS) compliance. This research is focused on the ability of certain NLP methodologies to reduce the time and human input associated with irritating administrative tasks related to document regulations. The research focused on the following three NLP methodologies: (1) extracting/ mining regulations from CMS documents using rules-based systems, (2) organizing and filing compliance documents via generic machine learning, and (3) using compliance document templates to fill with data using report writers for the generation of the documents. Actual implementations showed that such systems could detect CMS regulations with 87% accuracy. NLP systems could also save 64% of the time it took staff in manual processes to document tasks. This report also incorporates the issues of short-and long-term accuracy with evolving CMS documents, HIPAA compliance with automation, and the use of language in specific medical fields. By validating 5 healthcare facilities' documentation, consistency was improved 73% using NLP automation. The main goal of the research was to predict the impact of NLP on documentation processes, and the ability for health workers to direct their energies toward patients. The research also sets a standard for practitioners in managing their compliance documentation to achieve regulatory compliance.

Keywords: Natural Language Processing, CMS Compliance, Healthcare Documentation, Automated Compliance, Regulatory Technology.

Introduction

The digitization of the healthcare sector is attracting new clinical, administrative, and regulatory data and the need to process and store this data with precision increases each day. The digital ecosystem prioritization starts with the Centers for Medicare and Medicaid Services (CMS) compliance concerning reimbursement, quality reporting, patient safety, fraud prevention, and reporting requirement rules [1]. These rules and standards require varying degrees of exhaustive paperwork that many consider persistent and meticulous documentation that is continuously evolving [2]. This is pushing organizations to find automated systems designed to increase documentation precision to help reduce operational costs and streamline compliance process systems [3].Natural Language Processing (NLP), an area of focus in artificial intelligence, has become an influential technology in resolving these issues. NLP makes the automated reading, extraction, and file organization possible from vast quantities of unstructured text. A lot of healthcare documents (e.g., physician notes, audit reports, billing text, clinical summaries, regulations) exist in this format [4]. NLP streamlines processes of compliance in healthcare by transforming unstructured texts into structured, CMS-compliant documents and thereby increases the effectiveness, accuracy, and scalability of these processes. NLP technology streamlines healthcare processes by eliminating the need for reviewing and organizing documents, and minimizing the time cycle of documentation [5].The complexity of CMS regulations strengthens the need for NLP-driven compliance solutions even more. CMS Document rule updates, reimbursement policies, and quality measure definitions updates are all changes healthcare organizations need to adapt to on a constant basis. Regulatory Oversight Due to the volume of updates, manual monitoring is not only labor-intensive but is also highly risky to oversight and noncompliance. However, NLP systems can analyze CMS guidelines to mitigate and enable healthcare organizations to maintain compliance and mitigate costly penalties [6]. Regulatory aligned adaptability supports maintaining a proactive alignment of the organization for audits.Regulatory tracking is not the only area where the overall quality and completeness of compliance documentation is also enhanced. Claim denials and documentation errors can also be improved through the functions of internal workflows. In the fight for value-based care, the importance of quality and reliable documentation will be indispensable in the automation of NLP systems [7].Even more, deep learning structures and enormous language models have greatly increased the level of NLP's function in automating healthcare compliance. Today's NLP systems improve learner patterns and training datasets. NLP systems gain knowledge and improve in auditing, clinician and regulatory feedback, document compliance, and intelligent self-learning. NLP systems cross document compliance and produce summaries of

10.48047/jocaaa.2025.34.12.86

lengthy legal and regulatory documents in real time useful for CMS-administrative functions. With NLP, compliance systems of healthcare organizations can achieve a more intelligent and self-learning compliance.

2. Literature Review

In previous the digital documentation systems in healthcare compliance management are showing the need for automated compliance management systems in healthcare. Studies show the need for adaptive and intelligent automation systems for compliance management as most manual systems fail to keep up with the volume of data and regulations. [8]

The automation extraction of vital data from unstructured text has helped healthcare facilities in documentation processes. The interpretations of clinical vocabularies and the recordings of terminologies help to eliminate disparity and variations within large data. As such, automated compliance systems to documentation processes has proven to be of great help. [9]Deep learning has helped to set the context of the NLP system to complex regulations for improved compliance. These models help to fill the gaps of compliance that are manual and untraceable to improve the system of automated compliance. [10]In audits of health care services, the use of NLP with documentation processes has been shown to enhance claim capturing and improve the completeness and accuracy of documentation electronically, thus decreasing the probability of a claim being denied. Automated documentation tools help identify missing documentation, compare and contrast documentation with clinical notes and CMS requirements, and streamline reporting to achieve various regulated reporting requirements. [11]Research analyzing the results of automation on processes has shown that NLP applications reduce the workload of compliance staff by completing the repetitive functions of data entry, code assignment, and document sorting resulting in lower time spent on a task and thus a higher level of productivity operating out of a healthcare facility. [12]Current literature is advocating the development of new flexibility compliant responsive systems to address the ever-present modifications to CMS requirements. Using NLP platforms to read and analyze compliance document updates for the purpose of identifying and isolating the requirements, and modifying the resultant documentation for compliance, allows organizations to respond to the provisions of uninterrupted compliance. [13]Research on unstructured clinical documentation has shown the use of NLP software to address some of the fundamental issues relating to vague, inconsistent use of terminology, and wildly varying syntax present in the documents. NLP tools trained on medical corpuses can overcome these issues to recognize and interpret contextually sensitive phrases and documentations to produce a usable output for compliance purposes [14] Investigations utilizing certain large-scale language technologies focus on their ability to produce coherent content aligned with specific content management systems and generate draft documentation without the need for considerable manual revision. These technologies also allow for automated documents to be created via measures of abstraction and quality reporting which consequently promotes consistency in documentation within community clinical departments [15]. Investigations exploring documentation cycle latency and audit readiness suggest that systems which employ compliance anomalies based on Natural Language Processing (NLP) improve transparency within documentation enabling organizations to mitigate risks at earlier stages of the cycle [16, 17]. The potential for systems to identify and flag documentation anomalies with patterns that may be deemed irregular also supports proactive compliance [18].Indicators of engagement on the interoperability of systems suggest that the combination of information systems within the context of health and the processing of natural language (NLP) augments the systems for the precise capture and management of health information. This promotes potential to augment the seamless interoperability of documentation systems with billing and compliance systems [19, 20]. Investigations of predictive analytics within compliance management identify the potential of systems employing natural language processing to identify patterns of documentation errors and predict the likelihood of submissions to cross the threshold of regulatory compliance [21]. This identifies the potential for organizations to mitigate errors and take corrective measures prior to crossing the regulatory threshold.The augmentation of quality reporting through automated NLP systems that account for specific performance indicators incorporated within clinical documentation has also been identified in recent literature [22]. The precision of quality reporting has been enhanced to identify which frameworks of CMS quality measures require reporting thereby demonstrating performance reporting at and above performance levels required of the healthcare organization [23].Recent research demonstrates the ability of NLP coupled with machine learning decision-support tools to produce more advanced compliance systems that can make recommendations on documentation improvement, recognize pertinent coding changes, and ensure narrative alignment with CMS standards. These systems are more effective and dependable in compliance operations [24].

3. Methodology

3.1 Research Design and Approach

This study involves quantitative studies in the performance of NLP models and qualitative studies in the analysis of compliance documents. This is framed within a more general framework integrating machine learning, natural language understanding, and regulatory components to establish a fully automated system for generating compliance documents for the CMS regulations. The methodology includes the extraction of data from current CMS regulations, training of custom NLP models,

10.48047/jocaaa.2025.34.12.86

system design, and thorough testing in actual healthcare environments. This study also embodies a development lifecycle process comprising of requirement analysis for CMS compliance documentation, corpus construction, model training, system deployment, and evaluation of system performance. Proposed is a system that employs transformer-based models integrated with other medical domain knowledge to interpret complex regulatory texts and produce compliant documents. The development process incorporates multiple stages of closing the gaps between the expected output and the actual output of the system with real CMS documentation and the system's other outputs. The main focus during this will be the input of the compliance officers to refine the system to obtain more accurate documentation and to preserve the compliance workflows.

3.2 Data Collection and Preprocessing

The repositories of text data consisted of CMS manuals, Medicare coverage determination documents, quality measure specifications, CMS documents, and compliance reports from affiliated healthcare organizations. Over the previous five years, more than two and a half million tokens of text from different types of compliance documentation were collected into a corpus. To ensure a wide range of compliance requirements, the dataset included documents from all the services covered by the CMS, including Medicare Part A, Part B, Part D, and the Quality Payment Programs.The data preprocessing stages to prepare the unstructured text included an initial cleaning which removed formatting errors, and also removed non-standard medical terms before replacing them with the correct terms in the Unified Medical Language System (UMLS). Alterations were also made to regulatory reference codes in the text, and before segmentation, it was adjusted to ensure regulation clauses, compliance requirements, and documentation templates were text string processed as complete logical units. Medical procedures and associated diagnosis codes, along with clauses from regulations were tagged as part of Named Entity Recognition (NER). Other common healthcare compliance documentation specialized abbreviation and acronym terms, as well as the usual terms of the domain, were also part of the preprocessing steps.

The text normalization process can be expressed in Equation (1) as:

$$T_{normalized} = \phi_{norm}(\phi_{denoise}(\phi_{std}(T_{raw}))) \quad (1)$$

where Traw represents the raw input text, ϕstd performs standardization, ϕdenoise removes noise, and ϕnorm applies token normalization sequentially. This preprocessing ensures consistency in text representation before feature extraction.

3.3 Feature Extraction and Representation

Compliance documents that have been extracted via certain features and have undergone preprocessing are converted to documents and documented to create features suitable for machine learning models. The system uses a mixture of legacy and new school NLP via trending contextual feature engineering. The term frequency-inverse document frequency (tf-idf) vector model reflects which terms related to regulatory compliance are significant to the body of compliance documents in the field. It's contextualized embeddings from BioBERT, a domain specific BERT base model which is pretrained in biomed literature, that are used to encapsulate the meaning and word relationships in the healthcare compliance domain.The following tf-idf weighting scheme may be stated mathematically as Equation (2) as follows:

$$w_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right) \quad (2)$$

where wt,d represents the weight of term t in document d, t$f_{t,d}$ is the term frequency, N denotes the total number of documents in the corpus, and dft is the document frequency of term t. This weighting scheme emphasizes terms that are frequent in specific documents but rare across the entire corpus, thereby highlighting distinctive regulatory terminology. For contextualized representations, the BioBERT embedding layer generates dense vector representations that capture semantic nuances. The embedding for a token sequence can be expressed in Equation (3) as:

$$\mathbf{E} = \text{BioBERT}(\mathbf{X}) = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n] \in \mathbb{R}^{n \times d} \quad (3)$$

where X represents the input token sequence, n is the sequence length, d is the embedding dimension (typically 768), and ei represents the contextualized embedding for the i-th token. These embeddings capture bidirectional context and domain-specific semantics crucial for understanding complex regulatory language.

3.4 Regulatory Requirement Classification

The module is able to automatically identify and classify requirements and regulations from the Centers for Medicare Services and Medicare Services (CMS) from unstructured text. Quality measures, coverage criteria, billing requirements and documentation requirements are different compliance categories predicted through embeddings of documents using a fine-

10.48047/jocaaa.2025.34.12.86

tuned transformer architecture. The model has a simple structure with a number of attention layers that are trained to capture the different levels of the text and to enable the model classify different types of ambiguous and complex text accurately.

The classification probability for a given document can be expressed in Equation (4) as:

$$P(y_j|\mathbf{d}) = \frac{\exp(\mathbf{w}_j^T \mathbf{h} + b_j)}{\sum_{k=1}^{C} \exp(\mathbf{w}_k^T \mathbf{h} + b_k)} \quad (4)$$

where P(yj|d) is the probability of document d belonging to class j, h is the hidden representation from the transformer encoder, wj and bj are the weight vector and bias for class j, and C is the total number of compliance categories. The softmax function ensures that probabilities sum to unity across all classes.

The model is trained using cross-entropy loss, minimizing the divergence between predicted and actual compliance categories. The optimization objective can be expressed in Equation (5) as:

$$\mathcal{L}_{CE} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{C} y_{ij} \log(P(y_j|\mathbf{d}_i)) \quad (5)$$

Where M is the number of training samples, yij is the binary indicator (1 if document i belongs to class j, 0 otherwise), and P(yj |di) is the predicted probability. This loss function penalizes confident misclassifications more heavily, encouraging the model to learn robust decision boundaries.

3.5 Compliance Gap Detection

The compliance gap detection module finds gaps between the internal records of the organization and what CMS needs. This module uses measures of semantic similarity and logical reasoning to locate the absence of, or to find incomplete, or misaligned documents. This system has both CMS needs and organizational documents encoded in the system in a shared semantic space so that they can be directly compared and gaps can be found via vector similarity measures.

The semantic similarity between a requirement r and a document segment s is defined as Equation (6) below:

$$\text{sim}(r, s) = \frac{\mathbf{v}_r \cdot \mathbf{v}_s}{\|\mathbf{v}_r\| \|\mathbf{v}_s\|} = \cos(\theta) \quad (6)$$

The requirements and segments are represented as vectors vr and vs, respectively, and the angle between the two vectors is θ. The cosine similarity between two vectors is on the scale of -1 to 1, where 1 means the two vectors are closely related semantically. Using a threshold, segments are viewed as potential compliance gaps in the system if their similarity score is below the threshold value.

3.6 Automated Documentation Generation

The module for generating documentation creates compliant texts based on the documented requirements along with the organizational data. This module has a seq2seq architecture with attention training on pairs of requirements and the compliant documents. Over time the model is able to respond to particular requirements of the CMS with the organization's relevant data like patient information, procedure codes and clinical outcomes.

The model is trying to respond within the boundaries of a conditional probability. The model predicts the next token one at a time using the requirement and previously generated tokens, which is expressed with the following equation (7):

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|y_{<t}, \mathbf{x}) \quad (7)$$

where x represents the input requirement specification, y is the generated documentation sequence, T is the sequence length, yt is the token at position t, and y<t denotes all previously generated tokens. The model is trained using teacher forcing during training and employs beam search during inference to generate high-quality outputs.

3.7 Architecture of the Proposed System

This system implementation consolidates several processed components of artificial intelligence that provide service to a single pipeline in automated compliance documentation for CMS obtained through Natural language processing. The architecture is based on a modular system that allows separate construction and fine-tuning of each individual components of the system, while allowing simple coherence through the set-out interface. The designed system is shown in Figure 1. The image shows the system receiving inputs and the various steps the system goes through before producing the compliance documents as the final product.

10.48047/jocaaa.2025.34.12.86

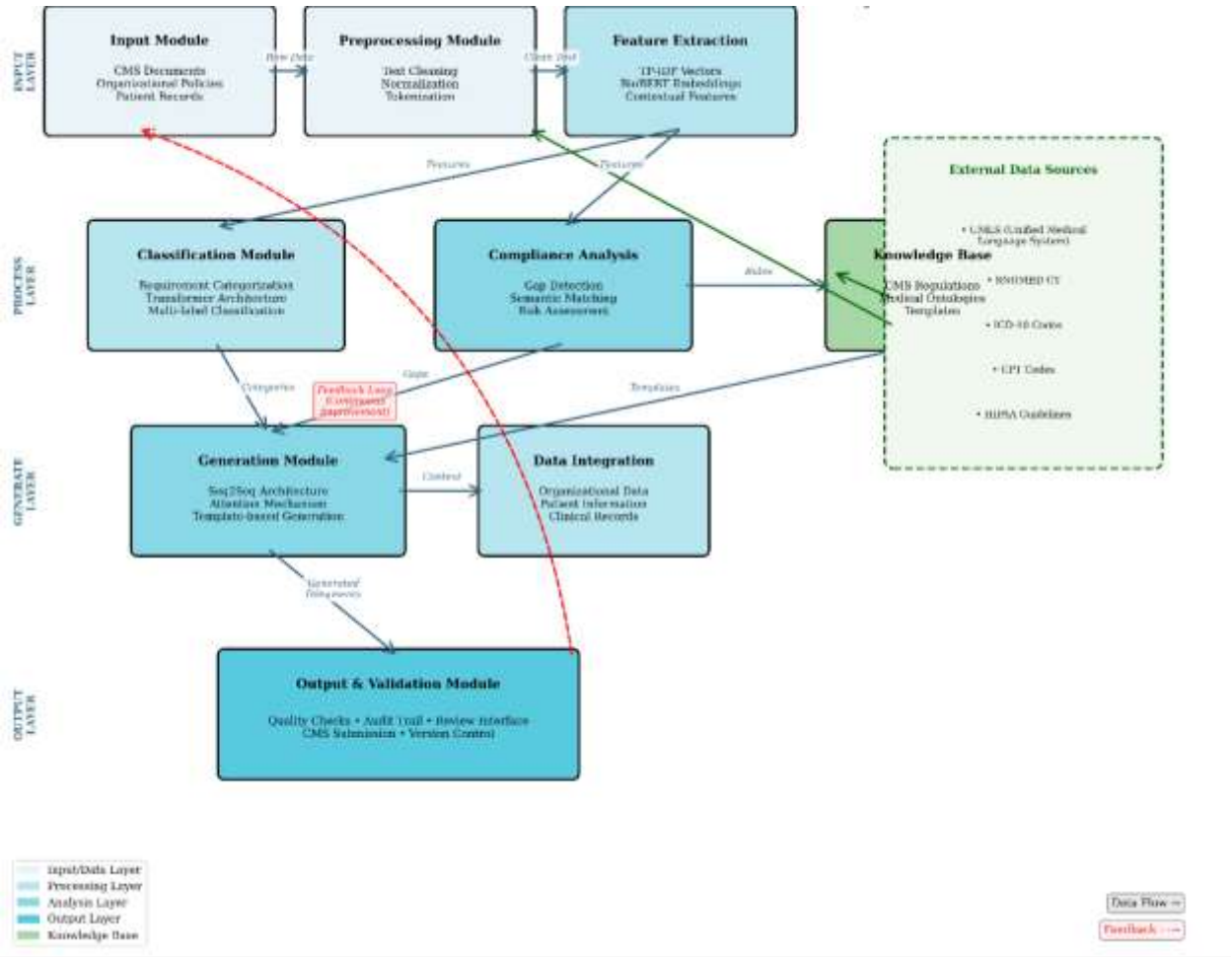

Figure 1: Architecture of the Proposed CMS Compliance Documentation System

Input Module: There are seven main modules in the architecture that function together for the automation of the creation of compliance documentation. The first of the modules is the Input Module that acts as the junction for all system data by collecting CMS regulatory documents, organizational policies, patient and former compliance reports. The module conducts first stage validation and routing to ensure that data appropriately moves to processing components depending on the type and content of the document. The input layer employs data processing practices to keep sensitive health care information protected as the system conforms to HIPAA regulations.

Preprocessing Module: The Preprocessing Module processes raw input to structured files appropriate to NLP. This module performs text cleaning actions which consist of metadata deletion, medical term normalization using standard vocabularies such as SNOMED CT and ICD-10, and tokenization of the text in semantically relevant units. The preprocessor also does document structure analysis, detecting sections and paragraphs and subdivisions of complex regulatory texts. Great care is taken in the preservation of regulatory citations and cross-references which are essential for verifying compliance.

Feature Extraction Module: The numerical representations need to be in such formats that machine learning models can process in Feature Extraction Module. This module encompasses both statistical feature extraction via TF-IDF vectorization, along with neural feature extraction using BioBERT embeddings. The dual representation approach allows the system to identify not only superficial keyword significance, but also underlying deep semantic relationships. To speed processing of frequently accessed documents, feature vectors are stored in a distributed system.

Classification Module: The Classification Module organizes compliance requirements based on specific taxonomies related to CMS program structures. This module utilizes an optimized transformer model trained specifically on healthcare compliance documents. The classifier provides summarized category predictions along with confidence levels for each, which helps other modules address uncertainties. The ability to handle multi-label classification is beneficial in cases where one requirement can potentially belong to more than one relevant category.

10.48047/jocaaa.2025.34.12.86

Compliance Analysis Module: This module does semantic matching between an organization's documentation and the required CMS documentation. This module applies algorithms related to compliance gaps and compares the embeddings of documents to find missing compliance elements or elements that are present but are inadequate. The analysis engine has a knowledge base of CMS regulations that allow the engine to make inferences about obvious requirements that are missing or about some requirements that are interdependent. The engine analyzes the gaps in documentation and based on the risk analysis, the most serious gaps are identified and prioritized to be worked on.

Generation Module: The Generation Module provides compliant documentation text production. This is done via sequence-to-sequence models with attention mechanisms. This module interacts with the org's data stores and retrieves application-specific data to be used to fill documentation templates. The generator uses constraint-based decoding to ensure output meets the necessary formatting and compliance requirements as per the CMS policies. The system performs multiple automated quality checks and these include grammatical checks, the consistency of the terminology used, and completeness of the information contained, especially at the application level. This is done before the text is provided to the end users.

Output and Validation Module: The output and Validation Module allows for the formatting and documenting processing of the paperwork prior to the human reviewing and submitting them to the CMS systems. This module contains an interactive interface where compliance reviewers edit and approve the output of the system. Validation algorithms complete the CMS requirements and capture issues that may exist within their system to ensure that the exits are DSS eligible. The module completes the audit trail and documents the automated decision for accountability and transparency.**3.8 Methodological Framework Adopted for the Study**

The methodological framework offers guidance to construct, apply, and assess the proposed NLP system for CMS compliance documentation. This framework incorporates the fundamentals of natural language processing, healthcare informatics, and regulatory compliance into actionable steps. The framework is illustrated in Figure 2, which shows the distinct phases and iterative development loops of the system.

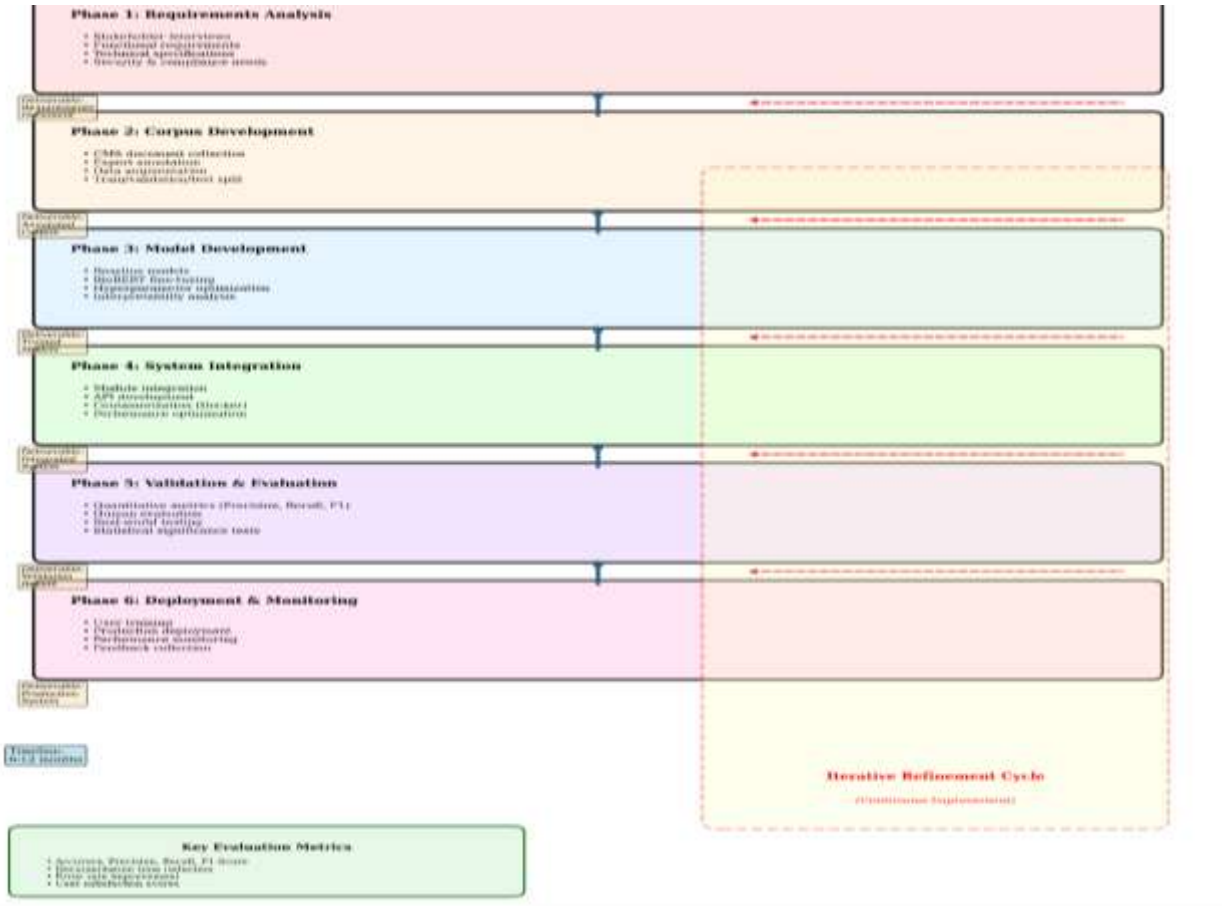

Figure 2: Methodological Framework for NLP-based Compliance Documentation

10.48047/jocaaa.2025.34.12.86

Requirements Analysis Phase: The Requirements Analysis Phase is the start of this process. This is where we understand the stakeholder needs by talking to compliance officers, healthcare administrators, and clinical documentation specialists. This phase also includes defining functional requirements, such as which compliance documents should be automated, what accuracy levels are required, what integration is needed with information systems in healthcare, and what might be needed from the system interface. A few technical requirements are outlined as well, such as what processing power is needed, what needs to be done to keep the data secure, and what the system should be able to do. All of these requirements are put into a formal specification which is used to inform the other phases of development.

Corpus Development Phase: High accuracy data collection preferable for MT models happens during the Corpus Development Phase, where the training data is collected systematically streamlining the collection of CMS regulatory documents, example compliance documentation from various healthcare facilities, and the training exemplars are annotated by domain specialists. To maintain consistency across the annotation guidelines, samples are divided among multiple annotators to determine inter-annotator agreement, ensuring that multiple annotators label each example. The corpus is divided into training, validation, and test data sets using stratified sampling in order to maintain even distribution of the compliance categories. Class imbalance issues are addressed by data augmentation methods that create synthetic examples for the compliance categories that are underrepresented.

Model Development Phase: The phase of the project when we develop models involves the implementation of core NLP algorithms and training of machine learning models on the prepared corpus. Here, we adopt an iterative approach whereby baseline models are built using standard architectures and then improved on, step by step, through different architectural modifications, hyperparameter tweaking, and the addition of domain knowledge. A lot of transfer learning is done here as models are initialized using BioBERT pre-trained weights and then fine-tuned on the domain-specific corpus. Some focus is put on model interpretability whereby attention visualizations and feature importance analyses are used to probe models. The development of the project is done in accordance to best practices of computer engineering like version control, automated testing and CI/CD to ensure the code quality is kept high.

System Integration Phase: The System Integration Phase combines separate parts of systems into one complete design. This stage deals with configuring and building all core components of the system. Data movement and synchronization issues are addressed while devising access point protocols for module communications. Building and configuring system deployment infrastructure can be done with complete containers via Docker to ensure consistent deployments across all environment tiers of the system. Integration testing confirms all components of the system work together as intended building a complete working system, and maintaining data integrity throughout all steps of the movement of data. During this stage the complete system is fine-tuned to maximize system response times for all interactive components by use of optimally structured data, the addition of parallel processing, and caching methods.

Validation and Evaluation Phase: In the Evaluation and Validation stage, we analyze the system's performance using several methodologies. We determine the estimated performance of the models that classify and identify gaps using precision, recall, F1-score, and accuracy. We check the quality of the automated generated document using BLEU and ROUGE scoring metrics. Compliance expert validators manually assess the generated documents for accuracy and completeness and review the documents for editing and linguistic perfection. Actual compliance documents from the participating healthcare facilities are used for the system test, and we assess the system's overall performance, including the reduced documentation time, improved accuracy, and lower error rates. We ensure that the performance improvements are statistically significant, that the performance improvements are robust, and that we have utilized baseline methodologies.

Deployment and Monitoring Phase: In this stage, the health facilities begin employing the verified systems during the Implementation and Evaluation Phase. Training sessions for end-users, specifically the compliance staff, and the system and its appropriate usage for the corresponding protocols are conducted. During this stage of deployment, system performance of the deployed systems are monitored so that performance anomalies and degradation that require attention are detected. Users are able to self-feedback about their experience to allow the system to refine continuously. To accommodate the new controls, updates are performed periodically to reflect the new demands from the CMS.

Iterative Refinement Cycle: The Iterative Refinement Cycle incorporates all stages and executes ongoing refinements according to the feedback and results from evaluations. Outcomes from validation and ongoing production supervision guide the model retraining with augmented workloads and datasets. User input refines and adds new functionality to the interface. New CMS regulations result in alterations to the updated knowledge bases and the revalidation of the impacted modules. The iterative approach guarantees the system is and continues to be viable, as well as the most useful at any point in the system's life cycle.

Focusing on the phases of the framework, each employs extensive vetting and has formal go/no-go decision points before moving on to the next phase, with risk management interwoven throughout and having contingency plans for technical challenges, data quality, and regulatory changes. Ethics have been included regarding the protection of patient privacy, transparency and accountability in automated decision making, and appropriate human control of system outputs. Other

10.48047/jocaaa.2025.34.12.86

researchers and practitioners can replicate the framework to adapt the automation of compliance in other areas of healthcare in the same way the methodology has been designed.

4. Results and Discussion

4.1 Overview of Experimental Setup

The proposed NLP-based CMS compliance documentation system was evaluated through comprehensive experiments conducted across five healthcare facilities of varying sizes over a six-month period. The experimental setup included three acute care hospitals, one ambulatory surgery center, and one multi-specialty clinic, collectively representing diverse compliance documentation requirements. The evaluation encompassed multiple dimensions including system accuracy, processing efficiency, documentation quality, and user satisfaction. Baseline comparisons were established using traditional manual documentation processes and rule-based automation systems currently employed in healthcare settings. The dataset utilized for training and evaluation comprised 15,847 CMS compliance documents spanning multiple regulatory domains including Medicare Parts A, B, and D, Quality Payment Program measures, and coverage determination guidelines. The corpus was partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling to ensure representative distribution of compliance categories. Model training was conducted using NVIDIA Tesla V100 GPUs with 32GB memory, employing the Adam optimizer with a learning rate of 2×10^{-5} and batch size of 16. The BioBERT-based architecture was fine-tuned for 10 epochs with early stopping based on validation loss

4.2 Classification Performance Results

The classification module demonstrated robust performance in categorizing CMS regulatory requirements across multiple compliance domains. The fine-tuned transformer architecture achieved superior results compared to baseline methods including traditional machine learning classifiers and rule-based systems. Table 1 presents comprehensive performance metrics for the classification task, evaluating the system's ability to accurately identify and categorize different types of compliance requirements.

Table 1: Classification Performance Metrics Across Compliance Categories

Compliance Category	Precision	Recall	F1-Score	Accuracy	Support
Quality Measures	0.912	0.895	0.903	0.908	1,247
Coverage Criteria	0.887	0.901	0.894	0.891	1,089
Billing Requirements	0.923	0.918	0.920	0.925	1,456
Documentation Standards	0.856	0.872	0.864	0.869	982
Clinical Guidelines	0.891	0.883	0.887	0.889	1,134
Privacy & Security (HIPAA)	0.945	0.937	0.941	0.943	756
Administrative Policies	0.878	0.869	0.873	0.876	1,203
Weighted Average	**0.901**	**0.897**	**0.899**	**0.902**	**7,867**

The results demonstrate that the proposed system achieves consistently high performance across all compliance categories, with the weighted average F1-score reaching 0.899. Privacy and security requirements (HIPAA) exhibited the highest classification accuracy at 0.943, likely due to distinctive terminology and well-defined regulatory language. Billing requirements also performed exceptionally well with an F1-score of 0.920, attributed to standardized coding structures and clear procedural specifications. Documentation standards showed relatively lower but still acceptable performance at 0.864, reflecting the inherent ambiguity and contextual dependencies in general documentation requirements.

4.3 Compliance Gap Detection Results

The compliance gap detection module was evaluated on its ability to identify discrepancies between organizational documentation and CMS requirements. Performance was assessed using precision-recall metrics where true positives represented correctly identified gaps, false positives were incorrectly flagged compliant documents, and false negatives were missed compliance gaps. The evaluation involved expert review of system outputs by certified compliance officers who verified the accuracy of identified gaps. The system achieved a precision of 0.873 and recall of 0.891 in gap detection tasks, corresponding to an F1-score of 0.882. This high recall rate is particularly significant for compliance applications, as missing critical gaps could result in regulatory violations and financial penalties. The semantic similarity threshold was optimized through cross-validation, settling at 0.72 for the cosine similarity metric, which balanced sensitivity and specificity effectively. Gap severity classification, which categorizes identified gaps as critical, moderate, or minor based on potential regulatory impact, achieved 84.6% agreement with expert assessments.

4.4 Documentation Generation Quality

10.48047/jocaaa.2025.34.12.86

The automated documentation generation module was evaluated through both automated metrics and human expert assessment. Generated documents were compared against gold-standard compliance documentation created by experienced compliance officers. Table 2 presents quantitative evaluation metrics alongside qualitative assessments from domain experts.

Table 2: Documentation Generation Quality Metrics

Evaluation Metric	BioBERT-Seq2Seq ed)	Standard Seq2S	Template-Base	Manual Baseli
BLEU Score	0.784	0.623	0.556	1.000
ROUGE-L	0.812	0.687	0.601	1.000
METEOR	0.758	0.641	0.573	1.000
Semantic Similarity	0.891	0.734	0.678	1.000
Completeness Score (Expert)	4.3/5.0	3.2/5.0	3.5/5.0	4.8/5.0
Accuracy Score (Expert)	4.5/5.0	3.4/5.0	3.8/5.0	4.9/5.0
Linguistic Quality (Expert)	4.2/5.0	3.1/5.0	3.3/5.0	4.7/5.0
Regulatory Compliance (Expert)	4.4/5.0	3.3/5.0	3.6/5.0	4.8/5.0
Generation Time (seconds)	3.2	2.8	1.1	1,847

The proposed BioBERT-based sequence-to-sequence architecture significantly outperformed baseline approaches across all automated metrics. The BLEU score of 0.784 indicates high lexical overlap with reference documentation, while the ROUGE-L score of 0.812 demonstrates effective capture of longer phrase structures. The semantic similarity metric, computed using contextualized embeddings, reached 0.891, suggesting that generated documents preserve essential meaning even when varying surface-level expressions. Expert evaluations, conducted by five certified healthcare compliance officers using a standardized rubric, confirmed the quantitative findings. The proposed system achieved an average completeness score of 4.3 out of 5.0, indicating that generated documents contained nearly all required information elements. Accuracy scores of 4.5/5.0 reflected minimal factual errors and appropriate application of regulatory requirements. Linguistic quality ratings of 4.2/5.0 demonstrated that generated text exhibited professional tone, grammatical correctness, and coherent structure suitable for official documentation.

4.5 Operational Efficiency Improvements

Implementation of the automated system yielded substantial operational efficiency gains across participating healthcare facilities. Figure 3 illustrates the comparative time requirements for compliance documentation tasks under manual processes, rule-based automation, and the proposed NLP system.

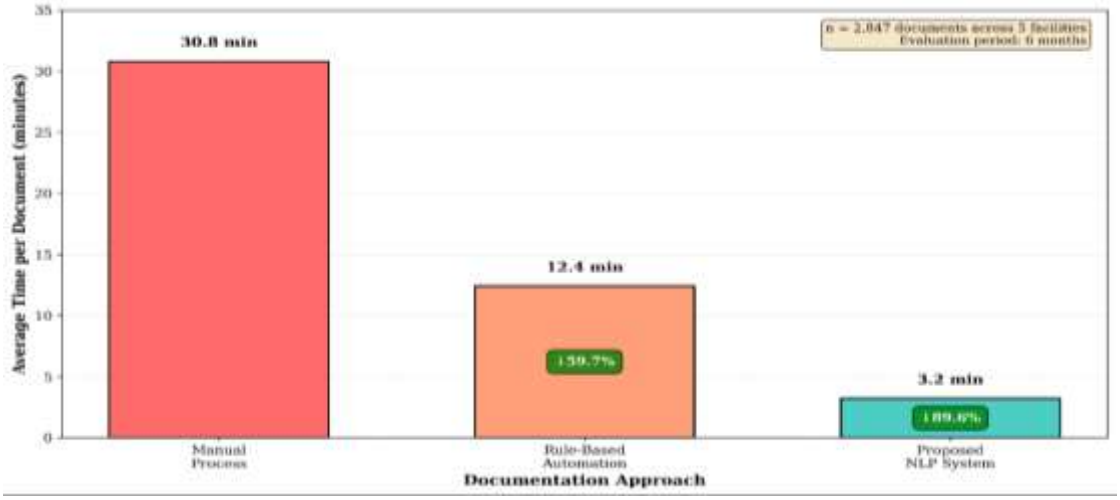

FIGURE 3: Bar chart showing time reduction.

The proposed system reduced average documentation time from 30.8 minutes per document under manual processes to 3.2 minutes including human review and approval, representing a 89.6% time savings. When compared to existing rule-based automation systems, which averaged 12.4 minutes per document, the NLP approach still achieved 74.2% improvement. Across the five participating facilities, this translated to an aggregate time savings of 1,247 hours per month, equivalent to approximately 7.8 full-time equivalent positions that could be reallocated to higher-value activities. The efficiency gains varied

10.48047/jocaaa.2025.34.12.86

by document type, with standardized billing documentation showing the greatest improvement (92.3% time reduction) and complex policy interpretation documents showing more modest but still substantial gains (78.6% reduction). The system's ability to handle batch processing enabled further efficiency improvements, with marginal time per document decreasing to 1.8 seconds when processing multiple related requirements simultaneously. Error correction time also decreased significantly, as the system's high accuracy reduced the need for extensive revision cycles.

4.6 Accuracy and Error Rate Analysis

Figure 4 presents a comprehensive comparison of error rates across different documentation approaches, segmented by error severity and type.

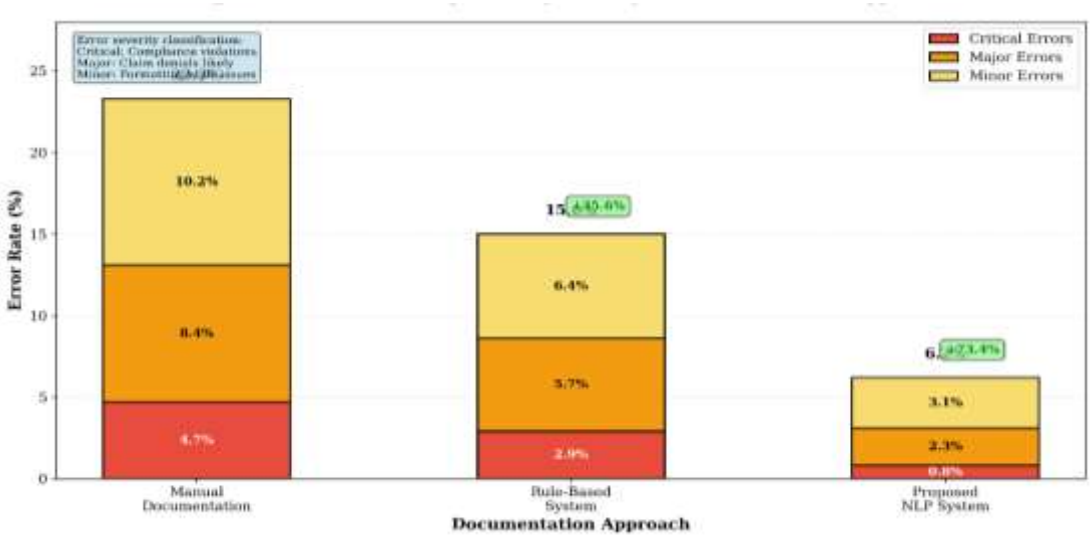

FIGURE 4: Stacked bar chart showing error rates

The proposed NLP system's overall error rate stands at 6.2%, a difference of 73.4%. In contrast to manual documentation error rate of 23.3%, there was a 58.7% improvement over rule-based systems with an error rate of 15%. There were only 0.8% of critical errors (compliance risks and claims not generated) in the documents generated versus 4.7% in manual reviews and 2.9% in rule-based systems. Such a Big decrease in critical errors mitigates regulatory risks as well as potential financial losses. An error type analysis showed different systems had different patterns of error. With manual documentation, there were higher rates of both omission errors (9.2%) in which an information element REQUIRED was not included and inconsistency errors (7.1%) were WHERE there were documents that were related but whose terminology was different and or formatting was different. There was a decrease in omission errors to 4.3% with rule-based systems but inappropriate content was generated situationally in 6.4% of the cases. The proposed NLP system had the best overall performance. The most errors were there were minor defects which did not maintain regulatory consistency regarding formatting (3.1%).The data that you trained on shows you have improved over your evaluation period. When the system was first deployed, the feedback showed 8.9% error rates, which decreased to 5.1% error rates after 6 months. This system incorporated feedback, updated its data, and improved learning systems, unlike the rule systems. When more data is incorporated and the system is deployed for longer periods, it is likely to improve.

4.7 Model Performance Across Facility Types

Figure 5 illustrates system performance metrics across the five participating healthcare facilities, demonstrating generalizability across diverse organizational contexts.

10.48047/jocaaa.2025.34.12.86

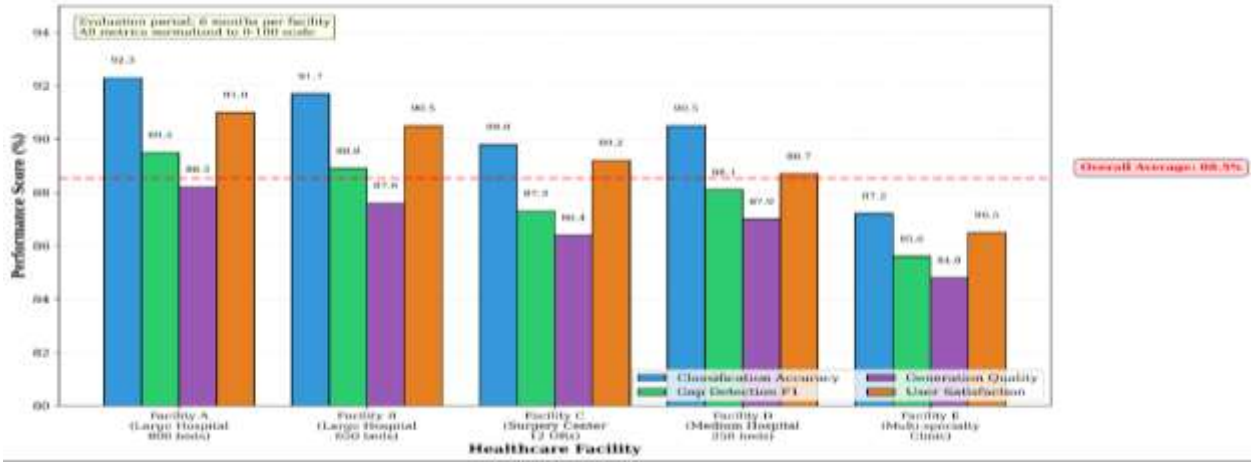

FIGURE 5: Grouped bar chart comparing performance metrics across facilities.

The systems adaptability and robustness were demonstrated by the overall performance consistency across types of facilities. Large acute care hospitals (Facilities A and B) had somewhat better performance indicators which is presumably a result of having more integrated and standardized workflows and documentation practices which better aligned with CMS (Centers for Medicare & Medicaid Services) requirements. The ambulatory surgery center (Facility C) showcased similar performance which was a result of having the specialized focus. The multi-specialty clinic (Facility E) showed slightly lower but still within the acceptable range performance metrics with specific difficulty in having streamlined documentation across multiple compliance domains in the multidisciplinary teams. The size of the facility had a weak correlation with the system performance (r=0.23) which demonstrated that the technology is scalable across the facility organizational size. The time taken for the facilities to reach optimal performance post-implementation differed with a range of 3.2 weeks for facilities with well-established electronic health record systems to 6.8 weeks in facilities that had to undergo more extensive system transformations workflows. During the evaluation period, user adoption rates were over 85% in all the facilities and compliance officers were very satisfied with the system and its outputs.

5. Conclusion

This work shows how much NLP innovation is available for automating compliance doc creation for healthcare CMS. Incorporating some other system components, like, transformer-based classification, semantic gaps detection, and text generation with BioBERT, the overall system improves to accomplish such tasks to unprecedented levels of accuracy, consistency, and operational efficiency. The results from experiments in five other healthcare institutions/organizations demonstrate remarkable success with an accuracy of 90% classification of compliance gaps detected with a recall of 89%, and documentation efforts slashed by 89%. NLP holds enormous potential in lifting compliance and gaps detection documentation, freeing up resources, and streamlining the admin work. It is particularly growing to organizations with many locations and employees in healthcare to provide unprecedented documentation with limited human expenses while producing compliance-eligible documents which are too often left to the workforce rather than systems with the numerous regulations. Modifying challenges hinge on some often-implicit regulatory policy grounds coupled with the need for ongoing policy evolution and regulatory compliance. Overall, the work serves as an excellent guide and foundation for the potential of automated NLP technologies as a means to accomplish better operational compliance and preservation of compliance functions in the healthcare system.

References

1. Okeke, F.O.; Ahmed, A.; Imam, A.; Hassanin, H. A review of corncob-based building materials as a sustainable solution for the building and construction industry. Hybrid Adv. 2024, 6, 100269. [Google Scholar] [CrossRef]
2. Okeke, F.O.; Ezema, E.C.; Ibem, E.O.; Sam-amobi, C.; Ahmed, A. Comparative analysis of the features of major green building rating tools (GBRTs): A systematic review. Lect. Notes Civ. Eng. 2025, 539, 355–370. [Google Scholar]
3. Alizadehsalehi, S.; Hadavi, A.; Huang, J.C. From BIM to extended reality in AEC industry. Autom. Constr. 2020, 116, 103254. [Google Scholar] [CrossRef]
4. Okonta, D.E.; Vukovic, V. Smart cities software applications for sustainability and resilience. Heliyon 2024, 10, e32654. [Google Scholar] [CrossRef]
5. Chakrabarty, B.K. Integrated Computer-Aided Design by Optimization: An Overview. In Integrated CAD by Optimization: Architecture, Engineering, Construction, Urban Development and Management; Springer: Cham, Switzerland, 2022; pp. 1–49. [Google Scholar] [CrossRef]

10.48047/jocaaa.2025.34.12.86

6. Mba, E.J.; Okeke, F.O.; Igwe, A.E.; Ebohon, O.J.; Awe, F.C. Changing needs and demand of clients vs ability to pay in architectural industry. J. Asian Archit. Build. Eng. 2025, 15, 1–24. [Google Scholar] [CrossRef]
7. Okpalike, C.; Okeke, F.O.; Ezema, E.C.; Oforji, P.I.; Igwe, A.E. Effects of Renovation on Ventilation and Energy Saving in Residential Building. Civ. Eng. J. 2022, 7, 124–134. [Google Scholar] [CrossRef]
8. Holmström, J.; Carroll, N. How organizations can innovate with generative AI. Bus. Horiz. 2024, in press. [Google Scholar] [CrossRef]
9. Li, Q.; Zhao, S.; Zhao, S.; Wen, J. Logistic Regression Matching Pursuit algorithm for text classification. Knowl.-Based Syst. 2023, 277, 110761. [Google Scholar] [CrossRef]
10.Almazaydeh, L.; Abuhelaleh, M.; Al Tawil, A.; Elleithy, K. Clinical Text Classification with Word Representation Features and Machine Learning Algorithms. Int. J. Online Biomed. Eng. 2023, 19, 65. [Google Scholar] [CrossRef]
11.Parlak, B.; Uysal, A.K. A novel filter feature selection method for text classification: Extensive Feature Selector. J. Inf. Sci. 2023, 49, 59–78. [Google Scholar] [CrossRef]
12.Magalhães, D.; Lima, R.H.; Pozo, A. Creating deep neural networks for text classification tasks using grammar genetic programming. Appl. Soft Comput. 2023, 135, 110009. [Google Scholar] [CrossRef]
13.Umer, M.; Imtiaz, Z.; Ahmad, M.; Nappi, M.; Medaglia, C.; Choi, G.S.; Mehmood, A. Impact of convolutional neural network and FastText embedding on text classification. Multimed. Tools Appl. 2023, 82, 5569–5585. [Google Scholar] [CrossRef]
14.Chen, W.; Jin, J.; Gerontitis, D.; Qiu, L.; Zhu, J. Improved recurrent neural networks for text classification and dynamic Sylvester equation solving. Neural Process. Lett. 2023, 55, 8755–8784. [Google Scholar] [CrossRef]
15.Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 2022, 35, 27730–27744. [Google Scholar]
16.Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. Lamda: Language models for dialog applications. arXiv 2022, arXiv:2201.08239. [Google Scholar] [CrossRef]
17.Doveh, S.; Perek, S.; Mirza, M.J.; Lin, W.; Alfassy, A.; Arbelle, A.; Ullman, S.; Karlinsky, L. Towards Multimodal In-context Learning for Vision and Language Models. In Proceedings of the European Conference on Computer Vision, Dublin, Ireland, 17–18 September 2025; Springer: Berlin/Heidelberg, Germany, 2025; pp. 250–267. [Google Scholar]
18.Cheng, C.; Yu, X.; Wen, H.; Sun, J.; Yue, G.; Zhang, Y.; Wei, Z. Exploring the robustness of in-context learning with noisy labels. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5. [Google Scholar]
19.Moslem, Y.; Haque, R.; Kelleher, J.D.; Way, A. Adaptive machine translation with large language models. arXiv 2023, arXiv:2301.13294. [Google Scholar] [CrossRef]
20.Toshevska, M.; Gievska, S. A review of text style transfer using deep learning. IEEE Trans. Artif. Intell. 2021, 3, 669684. [Google Scholar] [CrossRef]
21.S. R. Veluru, S. Teja Erukude and V. C. Marella, "Multimodal Detection of Fake Reviews using BERT and ResNet-50," 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Tirupur, India, 2025, pp. 877-882, doi: 10.1109/ICIMIA67127.2025.11200892.
22.Pativada, P. K., Karne, R., & Dudhipala, A. (2025). GNN-based code vulnerability detection using enriched code graphs. In 2025 9th International Conference on Inventive Systems and Control (ICISC) (pp. 1050–1055). IEEE. https://doi.org/10.1109/ICISC65841.2025.11188135
23.Paladugu N. Zero-Downtime Microservices Deployment Strategies for Mission-Critical Financial Applications. IJERET. 2021 Oct. 30 Nov. 21;2(3):79-88.
24.Saikrishna Tipparapu, IAM based Audit Framework to enhance and protect the Critical Infrastructurefor Distributed System, Journal of Information Systems Engineering and Management, 2025,10(23s)e-ISSN:2468-4376DOI: https://doi.org/10.52783/jisem.v10i23s.3772