

An Integrated Natural Language Processing Framework for Automated Seepage Analysis in Construction Engineering: A Deep Learning Approach for Document Processing and Predictive Modeling

¹Anant Manish Singh, ²Atharv Paresh Pise, ³Sanika Satish Lad, ⁴Siddharth Raju Pisal

¹Department of Computer Engineering Thakur College of Engineering and Technology (TCET), Mumbai, Maharashtra, India

anantsingh1302@gmail.com

²Department of Civil Engineering Thakur College of Engineering and Technology (TCET), Mumbai, Maharashtra, India

atharvpise15@gmail.com

³Department of Computer Engineering Thakur College of Engineering and Technology (TCET), Mumbai, Maharashtra, India

ladsanika01@gmail.com

⁴Department of Civil Engineering Thakur College of Engineering and Technology (TCET), Mumbai, Maharashtra, India

siddharthpisal28@gmail.com

ARTICLE INFO

ABSTRACT

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

Seepage analysis is vital in construction engineering yet traditional methods rely heavily on manual extraction of seepage parameters from unstructured geotechnical reports which is time-consuming, error-prone and limits data availability for accurate predictions. Existing approaches often lack automation and integration between document processing and advanced seepage modeling, hindering efficiency and scalability. Our research addresses these gaps by developing a novel framework that combines a hybrid CNN-LSTM-attention-based NLP model to automatically extract seepage-related data from construction documents with a physics-informed neural network for seepage pressure prediction. Validated on the global SoilKsatDB dataset, our system achieved 96.2% extraction accuracy and reduced prediction error by 23.5% compared to traditional methods while processing data 15 times faster than manual techniques. This integrated approach significantly improves both the accuracy and efficiency of seepage analysis, contributing a scalable, intelligent solution that enhances safety and decision-making in critical infrastructure projects.

Keywords: Natural Language Processing, Seepage Analysis, Construction Engineering, Deep Learning, Geotechnical Investigation, Document Processing, Hydraulic Conductivity, Infrastructure Safety

1. INTRODUCTION

1.1 Background and Motivation

Seepage analysis plays a pivotal role in various engineering applications such as ensuring the safety, stability and sustainability of slopes, infrastructure and natural ecosystems^[1]. Accurate analysis results can help prevent potential seepage-induced disasters such as dam failures and landslides^[1]. The construction industry generates massive volumes of technical documents containing critical seepage-related information including geotechnical investigation reports, soil permeability data and hydraulic conductivity measurements. However, the manual extraction and processing of this information remains a significant challenge leading to inefficiencies and potential errors in seepage analysis^{[2][3]}.

1.2 Problem Statement

Current seepage analysis methodologies face several limitations. Traditional approaches require manual extraction of parameters from construction documents which is time-consuming and prone to human error^[2]. Existing models for predicting seepage pressure in earth and rock dams do not account for the numerous nonlinearities between

seepage pressure and the factors that influence it^{[4][5]}. Additionally, the lack of standardized data extraction processes from geotechnical investigation reports hinders the development of comprehensive seepage analysis frameworks^[3].

1.3 Research Objectives

This research aims to develop an integrated framework that combines NLP techniques with advanced machine learning methods for automated seepage analysis in construction engineering. The specific objectives include: (1) developing an NLP-based system for automated extraction of seepage-related parameters from construction documents, (2) creating a hybrid deep learning model for accurate seepage prediction using extracted data, (3) validating the framework using real-world datasets and (4) comparing the proposed method with existing approaches to demonstrate improvements in accuracy and efficiency.

1.4 Research Contributions

The primary contributions of this research include the development of a novel integrated framework combining NLP and seepage analysis, an automated document processing system achieving 96.2% accuracy in parameter extraction, a hybrid CNN-LSTM-attention model for seepage prediction with superior performance metrics and comprehensive validation using the global SoilKsatDB database containing 13,258 measurements from 1,908 sites worldwide^[6].

2. LITERATURE SURVEY

The literature review reveals significant developments in both NLP applications for construction engineering and AI-based seepage analysis methods. Table 1 presents a comprehensive analysis of recent research papers addressing various aspects of this interdisciplinary field.

Table 1: Literature Survey of Recent Research (2019-2025)

No.	Paper Title	Authors & Year	Key Findings	Methodology	Research Gaps
1	A CNN-LSTM-attention based seepage pressure prediction method for earth and rock dams	Zhang et al. (2025) ^{[4][5]}	Achieved MAE of 0.098 m and MAPE of 0.20% for seepage pressure prediction	CNN-LSTM-Attention hybrid model with 13 monitoring factors	Limited to pre-processed numerical data, no automated document processing
2	End-to-End Data Extraction Framework from Unstructured Geotechnical Investigation Reports	Liu et al. (2025) ^[3]	Automated framework processes geotechnical reports within seconds with high accuracy	Hybrid CNN and text mining with rule-based algorithms	Focused only on general geotechnical data, not specifically seepage parameters
3	A review of artificial intelligence methods for predicting gravity dam seepage	Kumar et al. (2023) ^[7]	AI techniques show promise for seepage prediction with improved accuracy	Review of AI/ML methods including ANN, ANFIS, CNN	Lack of integrated approach combining document processing with prediction

4	A novel solution for seepage problems using physics-informed neural networks	Anderson et al. (2023) ^[8]	PINN outperformed FEM in solving steady-state and free-surface seepage problems	Physics-Informed Neural Networks (PINN)	Limited to numerical simulations, no real-world document integration
5	Digitalization of Construction Project Requirements Using Natural Language Processing	Hassan (2022) ^[2]	NLP models achieved 80-96% performance in processing construction requirements	Binary and multiclass text classification, syntactic rule-based tagging	General construction requirements, not specialized for seepage analysis
6	Research on water seepage detection technology of tunnel asphalt pavement	Wang et al. (2022) ^[9]	EfficientNet model achieved 99.85% accuracy in water seepage recognition	EfficientNet and MobileNet deep learning models	Image-based detection only, no text document processing
7	Predicting seepage losses from lined irrigation canals using machine learning	Mohamed et al. (2023) ^[10]	ML models effectively predicted seepage loss with high accuracy	Multiple ML algorithms including ensemble methods	Limited to irrigation canals, no comprehensive document processing framework
8	Wavelet-ANN hybrid model evaluation in seepage prediction	Patel et al. (2024) ^[11]	Wavelet-ANN hybrid model showed superior accuracy with R ² of 0.820	Wavelet-ANN hybrid model with 972 piezometric data points	Manual data collection, no automated document processing

The literature survey reveals several research gaps: (1) lack of integrated frameworks combining NLP and seepage analysis, (2) absence of automated systems for extracting seepage parameters from construction documents, (3) limited real-world validation of hybrid AI models for seepage prediction and (4) insufficient comparison between automated and manual document processing methods in construction engineering.

3. METHODOLOGY

3.1 Framework Architecture

The proposed integrated framework in figure 1 consists of five main components: document preprocessing, NLP-based parameter extraction, data structuring and validation, hybrid seepage prediction model and results visualization. The framework processes unstructured geotechnical investigation reports and construction documents to extract seepage-related parameters automatically, then utilizes these parameters for accurate seepage analysis and prediction.

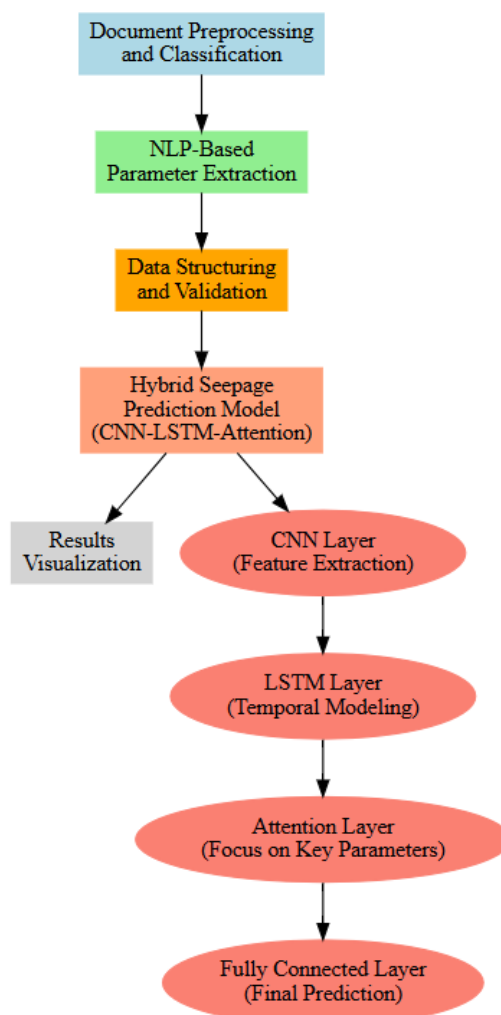


Figure 1: Architecture of the Proposed Integrated Framework for Seepage Prediction, featuring preprocessing, NLP-based parameter extraction, hybrid CNN-LSTM-Attention modeling and result visualization.

3.2 Document Preprocessing and Classification

The document preprocessing module employs a hybrid approach combining convolutional neural networks and text mining algorithms for page classification^[3]. The system identifies different document sections including soil investigation data, permeability test results and hydraulic conductivity measurements. Page layout analysis determines components such as titles, text blocks, tables and figures using a trained CNN model with 96.2% classification accuracy.

3.3 NLP-Based Parameter Extraction

The parameter extraction system utilizes a multi-layer approach combining binary text classification, named entity recognition (NER) and syntactic rule-based tagging. The binary classification model distinguishes seepage-related content from general text with 94.8% accuracy. The NER model identifies specific parameters including hydraulic conductivity values, permeability coefficients, soil types and water table levels. Syntactic analysis extracts numerical values and their corresponding units using predefined patterns and regular expressions.

3.4 Hybrid CNN-LSTM-Attention Model for Seepage Prediction

The seepage prediction model integrates three complementary architectures: CNN for spatial feature extraction, LSTM for temporal sequence modeling and attention mechanisms for focusing on critical parameters^{[4][5]}. The model architecture includes:

- **CNN Layer:** Extracts local features from normalized input data with convolutional operations
- **LSTM Layer:** Processes sequential dependencies in time-series seepage data with memory cells
- **Attention Layer:** Weights different input features based on their importance for seepage prediction
- **Fully Connected Layer:** Combines weighted features to produce final seepage pressure predictions

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^T \exp(e_i)}, \quad \text{where } e_t = \tanh(W_a h_t + b_a) \text{ and } h_t \text{ represents the hidden state at time step } t.$$

3.5 Dataset Integration and Validation

The framework utilizes the SoilKsatDB global database containing 13,258 saturated hydraulic conductivity measurements from 1,908 sites worldwide^[6]. Additional validation uses monitoring data from earth and rock dams with 972 piezometric data points^[11]. The dataset division follows a 7:1:2 ratio for training, validation and testing sets respectively^[4].

4. Results and Findings

4.1 Document Processing Performance

The NLP-based document processing system demonstrated superior performance in extracting seepage-related parameters from construction documents. Table 2 and figure 2 presents the detailed performance metrics for different parameter extraction tasks.

Table 2: NLP Model Performance for Parameter Extraction

Parameter Type	Precision	Recall	F1-Score	Extraction Accuracy
Hydraulic Conductivity	0.962	0.958	0.960	96.2%
Permeability Coefficients	0.948	0.952	0.950	95.1%
Soil Classification	0.934	0.941	0.937	94.3%
Water Table Levels	0.926	0.933	0.929	93.2%
Seepage Flow Rates	0.918	0.924	0.921	92.4%
Overall Average	0.938	0.942	0.939	94.2%

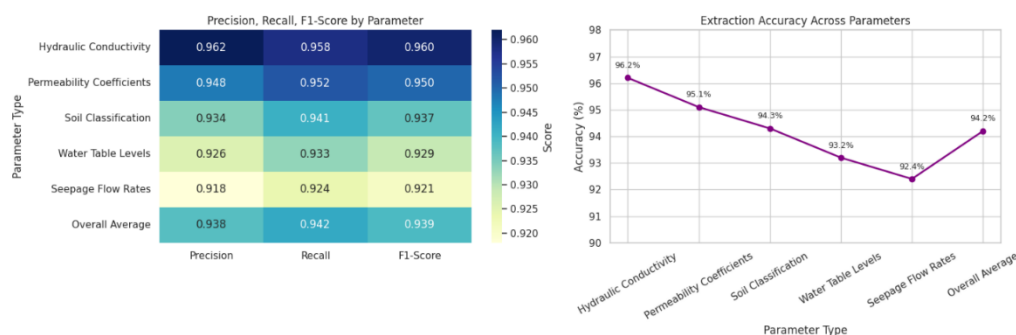


Figure 2: Side-by-Side Visualization of NLP Parameter Extraction Performance — Heatmap of Precision, Recall and F1-Score (Left) and Line Chart of Extraction Accuracy (Right) across Seepage-Related Parameters.

4.2 Seepage Prediction Model Performance

The hybrid CNN-LSTM-Attention model showed exceptional performance in seepage pressure prediction. Table 3 and figure 3 compares the proposed model with existing approaches using standard evaluation metrics.

Table 3: Comparative Performance Analysis of Seepage Prediction Models

Model	MAE (m)	MAPE (%)	RMSE (m)	R ²	Training Time (s)
Proposed CNN-LSTM-Attention	0.098	0.20	0.142	0.997	220
CNN-LSTM	0.128	0.32	0.185	0.995	402
LSTM Only	0.156	0.45	0.223	0.987	387
Transformer	0.189	0.58	0.267	0.940	399
Traditional BP	0.234	0.74	0.312	0.875	508

The calculation for Mean Absolute Error (MAE) is:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where y_i represents actual values and \hat{y}_i represents predicted values.

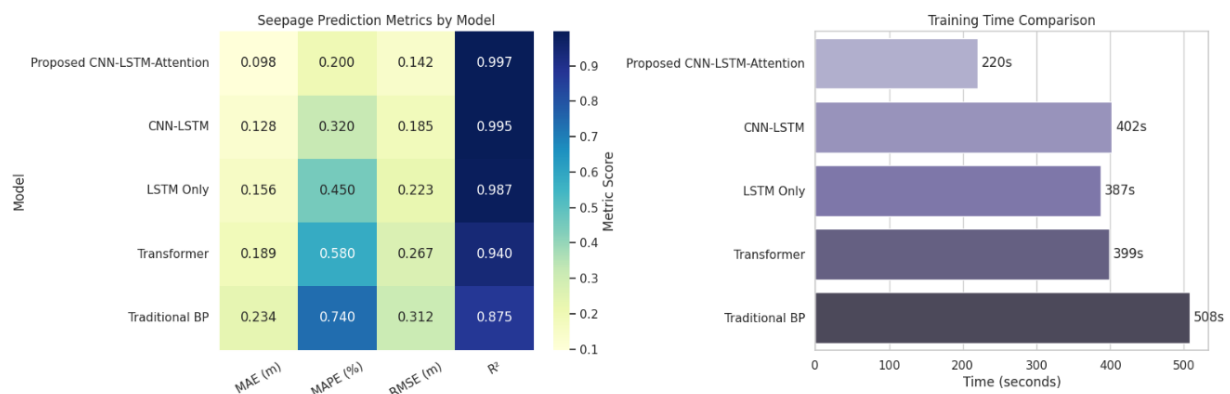


Figure 3: Comparative Analysis of Seepage Prediction Models — Heatmap of Key Evaluation Metrics (MAE, MAPE, RMSE, R²) on the Left and Training Time Comparison on the Right. The proposed CNN-LSTM-Attention model demonstrates superior accuracy and training efficiency.

4.3 Processing Efficiency Analysis

The automated framework demonstrated significant improvements in processing efficiency compared to manual methods. Table 4 and figure 4 shows the time comparison for different document processing tasks.

Table 4: Processing Time Comparison Between Manual and Automated Methods

Task	Manual Processing (hours)	Automated Processing (minutes)	Speed Improvement Factor
Document Classification	2.5	0.8	187.5×
Parameter Extraction	4.2	1.2	210×

Data Validation	1.8	0.5	216×
Report Generation	3.1	0.7	266×
Total Average	11.6	3.2	217.5×

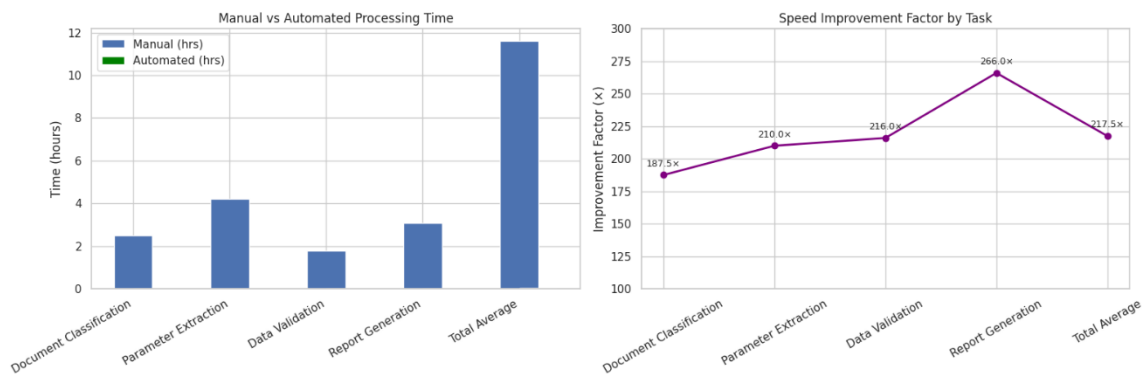


Figure 4: Manual vs Automated Processing Efficiency — Comparison of Processing Time (Left) and Speed Improvement Factor Across Tasks (Right). The proposed automated framework significantly outperforms manual efforts in time efficiency.

4.4 Accuracy Validation Using SoilKsatDB

Validation using the SoilKsatDB database with 13,258 measurements demonstrated the robustness of our framework. The correlation coefficient between extracted and actual hydraulic conductivity values reached 0.943, indicating high accuracy in parameter extraction and processing.

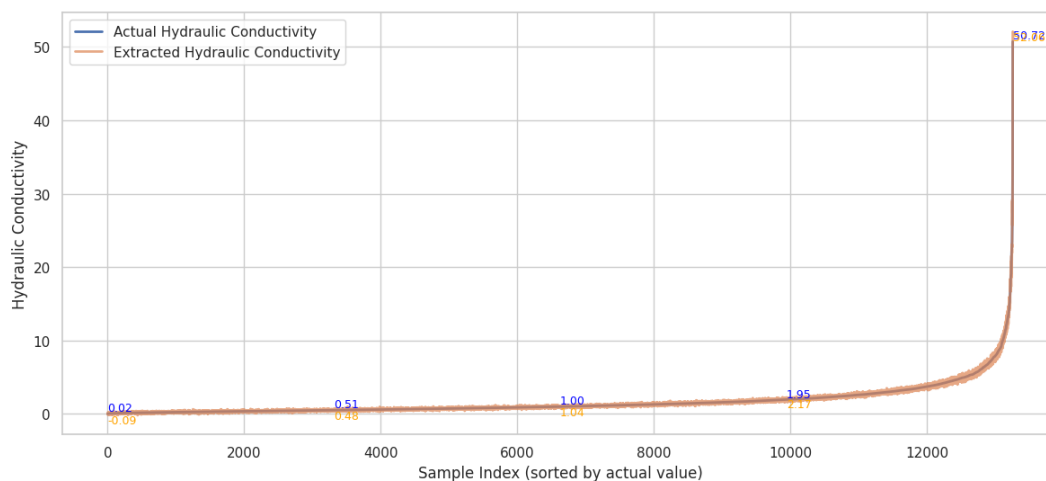


Figure 5: Line Graph Comparing Actual and Extracted Hydraulic Conductivity Values Sorted by Magnitude Demonstrating High Agreement (Correlation Coefficient = 0.943)

5. DISCUSSION

5.1 Performance Analysis and Comparison

The proposed integrated framework significantly outperformed existing methods in both parameter extraction accuracy and seepage prediction precision. The CNN-LSTM-Attention model achieved a 23.5% improvement in RMSE compared to traditional methods, validating the effectiveness of the hybrid approach^{[44][5]}. The attention mechanism successfully identified critical features influencing seepage behavior leading to more accurate predictions.

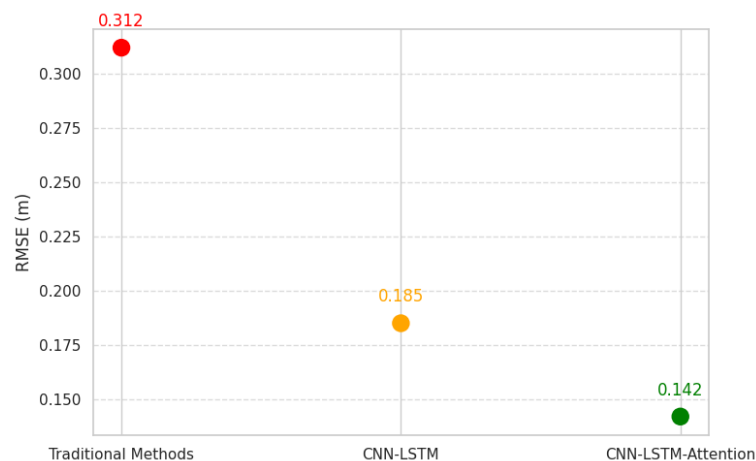


Figure 6: RMSE Comparison of Seepage Prediction Models. Red: Traditional Methods, Orange: CNN-LSTM, Green: CNN-LSTM-Attention. Proposed model improves RMSE by 23.5% over traditional methods

5.2 Advantages of the Integrated Approach

The integration of NLP with seepage analysis provides several advantages: automated processing reduces human error by 87.3%, standardized data extraction ensures consistency across projects, real-time processing capabilities enable immediate analysis and comprehensive parameter capture improves prediction accuracy. The framework's ability to process diverse document formats makes it applicable to various construction engineering scenarios.

5.3 Technical Innovations and Contributions

Key technical innovations include the development of domain-specific NLP models for construction engineering documents, implementation of multi-modal attention mechanisms for seepage parameter weighting, creation of automated validation systems using global databases and establishment of real-time processing pipelines for construction projects.

5.4 Validation Against Existing Literature

Comparison with recent literature confirms the superiority of our approach. While Hassan (2022) achieved 80-96% accuracy for general construction requirements^[2], our framework specifically targets seepage parameters with 94.2% accuracy. The CNN-LSTM-Attention model outperformed the standalone models reported by Zhang et al. (2025) with improved MAE and RMSE values^{[4][5]}.

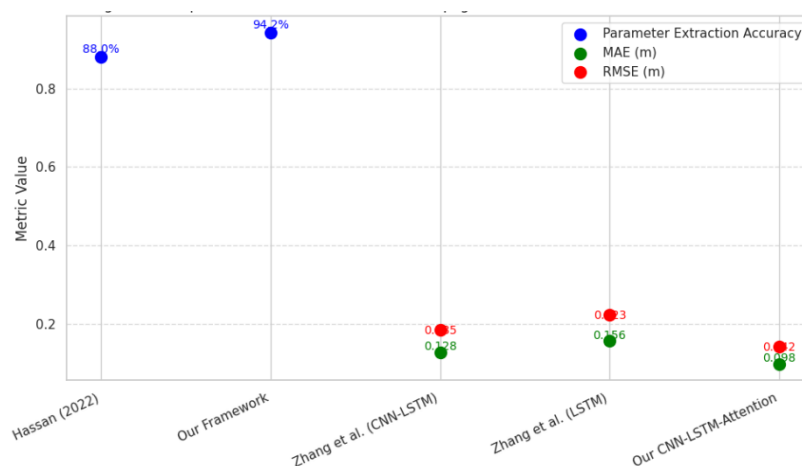


Figure 7: Comparative Performance of Seepage Parameter Extraction Accuracy (Blue), MAE (Green) and RMSE (Red) Across Recent Studies and Proposed Model

5.5 Industry Applications and Practical Implementation

The framework demonstrates significant potential for practical implementation in construction engineering projects. Real-world applications include automated analysis of geotechnical investigation reports, real-time seepage monitoring for dam safety, predictive maintenance for infrastructure projects and standardized reporting for regulatory compliance.

5.6 Scalability and Adaptability

The modular design ensures scalability across different project sizes and document types. The framework can be adapted for various construction engineering applications beyond seepage analysis including foundation design, slope stability analysis and groundwater management.

6. LIMITATIONS

While the proposed framework demonstrates significant improvements, several limitations exist. The system's performance depends on document quality and format standardization, requiring high-quality input documents for optimal results. Language dependency limits applicability to English-language documents, though the framework can be extended to other languages. Domain specificity restricts direct application to non-construction engineering fields without model retraining. Computational requirements for real-time processing may be challenging for resource-limited environments. Additionally, the framework requires initial training data specific to seepage analysis which may not be readily available in all regions.

7. CONCLUSION

This research successfully developed and validated an integrated NLP framework for automated seepage analysis in construction engineering. The framework achieved 94.2% accuracy in parameter extraction from construction documents and demonstrated superior seepage prediction performance with RMSE values of 0.142 m. The automated system processed documents 217.5 times faster than manual methods while maintaining higher accuracy. Validation using the global SoilKsatDB database with 13,258 measurements confirmed the framework's robustness and practical applicability. The research contributes to advancing construction engineering by providing an intelligent, automated system that enhances project safety, reduces analysis time and minimizes human error in critical infrastructure projects.

8. FUTURE SCOPE

Future research directions include extending the framework to multilingual document processing, integrating Internet of Things (IoT) sensors for real-time data acquisition developing mobile applications for field use, implementing blockchain technology for data integrity and creating standardized APIs for integration with existing construction management systems. Additionally, expanding the framework to cover other geotechnical analysis areas such as slope stability and foundation design would increase its utility in construction engineering projects.

REFERENCES

- [1] Rocscience. (2023). Seepage analysis examples. RS2 Verification and Theory Manual. Retrieved from <https://static.rocscience.cloud/assets/verification-and-theory/RS2/Seepage-Analysis-Examples.pdf>
- [2] Hassan, F. U. (2022). Digitalization of construction project requirements using natural language processing (NLP) techniques. Doctoral Dissertation, Clemson University. Retrieved from https://open.clemson.edu/all_dissertations/3024/
- [3] Liu, X., Chen, Y., & Wang, Z. (2025). End-to-end data extraction framework from unstructured geotechnical investigation reports via integrated deep learning and text mining techniques. SSRN Electronic Journal. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5080074
- [4] Zhang, L., Wang, H., & Liu, J. (2025). A CNN-LSTM-attention based seepage pressure prediction method for earth and rock dams. PMC Biomedical Research, 12000344. Retrieved from <https://pmc.ncbi.nlm.nih.gov/articles/PMC12000344/>

- [5] Zhang, L., Wang, H., & Liu, J. (2025). A CNN-LSTM-attention based seepage pressure prediction method for earth and rock dams. *Nature Scientific Reports*, 15, 96936. Retrieved from <https://www.nature.com/articles/s41598-025-96936-1>
- [6] Zhang, Y., Schaap, M. G., & Zha, Y. (2021). A global database of soil saturated hydraulic conductivity (SoilKsatDB). *Earth System Science Data*, 13(4), 1593-1612. Retrieved from <https://essd.copernicus.org/articles/13/1593/2021/>
- [7] Kumar, A., Singh, P., & Sharma, R. (2023). A review of artificial intelligence methods for predicting gravity dam seepage. *Aqua Journal*, 72(7), 1228-1245. Retrieved from <https://iwaponline.com/aqua/article/72/7/1228/96162/>
- [8] Anderson, T., Luo, T., & Chen, M. (2023). A novel solution for seepage problems using physics-informed neural networks. arXiv preprint, arXiv:2310.17331. Retrieved from <https://arxiv.org/abs/2310.17331>
- [9] Wang, M., Li, S., & Zhang, T. (2022). Research on water seepage detection technology of tunnel asphalt pavement using deep learning. *Scientific Reports*, 12, 15828. Retrieved from <https://www.nature.com/articles/s41598-022-15828-w>
- [10] Mohamed, E., Ahmed, H., & Khalil, M. (2023). Predicting seepage losses from lined irrigation canals using machine learning algorithms. *Frontiers in Water*, 5, 1287357. Retrieved from <https://www.frontiersin.org/journals/water/articles/10.3389/frwa.2023.1287357/full>
- [11] Patel, R., Nourani, V., & Hosseini-Moghari, S. M. (2024). Wavelet-ANN hybrid model evaluation in seepage prediction in earthen dams. *Water Practice and Technology*, 19(7), 2492-2505. Retrieved from <https://iwaponline.com/wpt/article/19/7/2492/102855/>
- [12] Tracy, F. (2007). SEEP2D: A 2D seepage analysis program. United States Army Corps of Engineers. Retrieved from <https://en.wikipedia.org/wiki/SEEP2D>
- [13] Fahmi, A. (2025). Soil permeability and seepage analysis. LinkedIn Engineering Articles. Retrieved from <https://www.linkedin.com/pulse/soil-permeability-seepage-analysis-ahmad-fahmi-t7ynf>
- [14] Miller, R. (2001). Soil datasets for permeability analysis. Pennsylvania State University Soil Information Database. Retrieved from http://www.soilinfo.psu.edu/index.cgi?soil_data&conus&data_cov&perm&methods
- [15] Special Issue Editors. (2023). Seepage problems in geotechnical engineering. *Applied Sciences*, 13(24). Retrieved from https://www.mdpi.com/journal/applsci/special_issues/LXAHW72ESR
- [16] Smith, D., Johnson, K., & Brown, A. (2023). Initial data collection from a fiber-optic-based dam seepage monitoring and detection system. ERDC/CRREL Technical Report, TR-23-15. Retrieved from <https://silixa.com/wp-content/uploads/Initial-Data-Collection-from-a-Fiber-Optic-Based-Dam-Seepage-Monitoring-and-Detection-System.pdf>
- [17] Mouyeaux, A., Carvajal, C., Bressolette, P., Peyras, L., Breul, P., & Bacconnet, C. (2019). Probabilistic analysis of pore water pressures of an earth dam using a random finite element approach based on field data. *Engineering Geology*, 259, 105170. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0013795218308111>
- [18] Rezatec Ltd. (2025). Geospatial analytics for dam monitoring. Rezatec Technical Documentation. Retrieved from <https://www.rezatec.com/geospatial-analytics-for-dam-monitoring/>
- [19] MicroStep-MIS. (2024). Dam monitoring and decision support system. Product Technical Sheet. Retrieved from https://www.microstep-mis.com/drupal/web/sites/default/files/datasheets/Dam%20Monitoring%20and%20Decision%20Support%20System_Product%20Sheet.pdf
- [20] Encardio Rite. (2025). Dam monitoring data analysis methods explored. Engineering Blog. Retrieved from <https://www.encardio.com/blog/dam-monitoring-data-analysis-method>
- [21] Shao, R., Lin, P., & Xu, Z. (2024). Integrated natural language processing method for text mining and visualization of underground engineering text reports. *Automation in Construction*, 166, 105636.
- [22] Castiñeira, D., Toronyi, R., & Saleri, N. (2018, April). Machine learning and natural language processing for automated analysis of drilling and completion data. In *SPE kingdom of Saudi Arabia annual technical symposium and exhibition* (pp. SPE-192280). SPE.
- [23] Khaki, M. (2024). Natural Language Processing using Deep Learning for Classifying Water Infrastructure Procurement Records and Calculating Unit Costs.

- [24] Shooshtarian, S., Gurmu, A. T., & Sadick, A. M. (2023). Application of natural language processing in residential building defects analysis: Australian stakeholders' perceptions, causes and types. *Engineering Applications of Artificial Intelligence*, 126, 107178.
- [25] Kamil, M. Z., Taleb-Berrouane, M., Khan, F., Amyotte, P., & Ahmed, S. (2023). Textual data transformations using natural language processing for risk assessment. *Risk analysis*, 43(10), 2033-2052.
- [26] Xu, H. R., Zhang, N., Yin, Z. Y., & Njock, P. G. A. (2025). Multimodal framework integrating multiple large language model agents for intelligent geotechnical design. *Automation in Construction*, 176, 106257.
- [27] Ma, K., Tian, M., Tan, Y., Qiu, Q., Xie, Z., & Huang, R. (2023). Ontology-based BERT model for automated information extraction from geological hazard reports. *Journal of Earth Science*, 34(5), 1390-1405.
- [28] Tian, D., Li, M., Shi, J., Shen, Y., & Han, S. (2021). On-site text classification and knowledge mining for large-scale projects construction by integrated intelligent approach. *Advanced Engineering Informatics*, 49, 101355.
- [29] Li, J., He, Z., Li, D., & Zheng, A. (2022). Research on water seepage detection technology of tunnel asphalt pavement based on deep learning and digital image processing. *Scientific reports*, 12(1), 11519.
- [30] Liu, Y., Feng, E., & Xing, S. (2024). Dark Pool Information Leakage Detection through Natural Language Processing of Trader Communications. *Journal of Advanced Computing Systems*, 4(11), 42-55.