EUROPEAN JOURNAL OF EMERGING DATA SCIENCE AND MACHINE LEARNING

Automated Seepage Characterization in Geotechnical Engineering Via Integrated NLP And Deep Learning: Enhancing Document Analysis and Predictive Capabilities

Dr. Sara M. Lopez
Department of Computer Science, University of Cambridge, United Kingdom

Dr. Fatima Al-Dossari
Department of Computer Science, King Saud University, Saudi Arabia

V0LUME01 ISSUE01 (2024)
Published Date: 14 December 2024 // Page no.: - 25-41

ABSTRACT

Seepage analysis is a critical aspect of geotechnical and hydraulic engineering, essential for ensuring the stability and longevity of civil infrastructure such as earth dams, tunnels, retaining walls, and deep excavations. Traditional methods for seepage assessment heavily rely on manual extraction and interpretation of parameters from vast amounts of unstructured geotechnical reports, monitoring logs, and design specifications. This manual process is inherently time-consuming, highly prone to human error, and severely limits the real-time availability of data for accurate predictions and proactive decision-making. This article presents a novel, integrated framework that leverages cutting-edge Natural Language Processing (NLP) and deep learning techniques to automate the extraction of crucial geotechnical and seepage-related information from diverse construction-related documents and to develop highly accurate predictive models for complex seepage behavior. The proposed methodology encompasses advanced NLP techniques, including custom-trained Named Entity Recognition (NER), sophisticated relation extraction, and detailed event extraction, designed to convert raw, unstructured textual data into a structured, machine-readable, and actionable knowledge base. These intelligently extracted features, combined with historical sensor monitoring data, are subsequently fed into robust deep learning architectures, specifically hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) models augmented with advanced attention mechanisms. This sophisticated model is engineered to predict critical seepage parameters such as pore water pressure and flow rates with enhanced precision. Validated extensively on the global SoilKsatDB dataset and real-world dam monitoring data, this research demonstrates a significant leap towards enhancing the efficiency, accuracy, and real-time capabilities of seepage analysis. It offers a scalable, intelligent, and robust solution for proactive monitoring, early anomaly detection, and comprehensive risk management in large-scale and complex civil infrastructure projects, thereby contributing substantially to infrastructure safety and operational sustainability.

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

INTRODUCTION

1.1 Background and Motivation

Seepage, defined as the flow of water through porous geological formations and engineered structures, is a ubiquitous and profoundly influential phenomenon in geotechnical and hydraulic engineering [15]. Its accurate assessment is not merely an academic exercise but a practical imperative for ensuring the structural integrity, long-term stability, and operational safety of a wide array of civil infrastructure [1]. From the foundational design of earth dams and levees that retain vast water bodies, to the complex construction of tunnels beneath urban landscapes or through challenging geological strata, and the stability of deep excavations and retaining walls, uncontrolled or unpredicted seepage can lead to catastrophic consequences. These include increased pore water pressures that reduce effective stresses and shear strength, the initiation of piping and internal erosion, slope instability, and ultimately, the risk of structural collapse and environmental damage [15].

Historically, seepage analysis has relied on a combination of theoretical fluid mechanics principles, empirical correlations, and numerical modeling techniques such as Finite Element Analysis (FEA) and Finite Difference Method (FDM) [1]. Specialized software like RS2 and SEEP2D are commonly employed to simulate groundwater flow and estimate seepage quantities and pore water pressures within complex geometries and varying geological conditions [1, 12]. The reliability of these numerical simulations, however, hinges critically on the accurate and precise input of fundamental geotechnical parameters, prominently including soil permeability, hydraulic conductivity, and well-defined boundary conditions [13, 14]. These parameters are traditionally derived from laboratory tests, in-situ field investigations,

and expert geological interpretations.

A pervasive and often underestimated challenge in the practical application of these sophisticated analytical and numerical methods lies in the management of project data. Modern construction projects, particularly large-scale infrastructure developments, generate an immense volume of technical documentation. This includes, but is not limited to, comprehensive geotechnical investigation reports, detailed soil boring logs, laboratory test results, daily construction logs, instrumental monitoring records from embedded sensors, design specifications, and post-construction maintenance reports. The overwhelming majority of this critical information resides in unstructured or semi-structured formats – such as PDF documents, scanned images, word processing files, or handwritten notes – making automated data access and processing extremely difficult [2, 3].

The manual extraction, interpretation, and transcription of this disparate data into structured formats suitable for numerical models or predictive analytics is a laborious, time-consuming, and highly error-prone process. This is particularly true for projects spanning several years or covering vast geographical areas, where tens of thousands of pages of documentation may accumulate [2, 21]. This inherent manual bottleneck creates several significant limitations: it delays the availability of crucial information for real-time analysis, impedes the development of comprehensive predictive models, hinders the proactive identification of potential seepage issues, and ultimately restricts agile decision-making processes essential for effective risk management.

1.2 Problem Statement

The limitations inherent in current seepage analysis methodologies can be distilled into several key problems:

- Manual Data Extraction Bottleneck: Traditional approaches necessitate extensive manual extraction of geotechnical and seepage-related parameters from unstructured construction documents. This process is exceedingly time-consuming, subject to significant human error, and poses a major impediment to efficiency, especially in large-scale projects [2]. The heterogeneity of document formats and the variability in reporting styles further exacerbate this issue, making consistent data collection a persistent challenge.

- Inadequate Account for Nonlinearities in Prediction Models: Existing models for predicting seepage pressure, particularly in structures like earth and rock dams, often struggle to fully capture the complex, nonlinear relationships between seepage pressure and its myriad influencing factors, such as fluctuating water levels, soil heterogeneity, and environmental conditions [4, 5]. While various machine learning models have been applied [7, 10, 11], a comprehensive approach that deeply integrates contextual textual information to enhance predictive accuracy remains largely unexplored.

- Lack of Standardized Data Extraction Processes: The absence of standardized, automated data extraction protocols from geotechnical investigation reports hinders the development of comprehensive and scalable seepage analysis frameworks [3]. This fragmentation prevents the creation of large, consistent datasets necessary for training advanced data-driven models.

- Limited Integration between Document Processing and Predictive Modeling: Despite the recognition of the value of textual information, there is a distinct research gap in integrated frameworks that seamlessly combine automated document processing with advanced predictive modeling for seepage analysis. Most existing studies tend to focus either on information extraction or predictive analytics in isolation, neglecting the powerful synergy that arises from their integration [21].

- Insufficient Real-World Validation and Comparative Analysis: While individual NLP and AI techniques have shown promise in sub-domains of construction [22, 23, 24, 25], comprehensive real-world validation of integrated AI models specifically for seepage prediction, particularly comparing automated versus manual document processing methods, is lacking. This gap restricts the confidence and adoption of such advanced systems in practical engineering applications.

1.3 Research Objectives

This research aims to address the aforementioned problems by developing a novel, integrated framework that combines cutting-edge Natural Language Processing (NLP) techniques with advanced deep learning methods for automated seepage analysis in construction engineering. The specific objectives are:

1. Develop an NLP-driven System for Automated Parameter Extraction: To design and implement a robust NLP-based system capable of efficiently and accurately extracting a wide range of relevant geotechnical and seepage-related parameters (e.g., hydraulic conductivity, soil type, pore water pressure, flow rates, event descriptions) from diverse unstructured construction documents (e.g., geotechnical reports, monitoring logs). This system will overcome the limitations of manual data extraction by converting qualitative textual information into structured, machine-readable data.

2. Construct Robust Hybrid Deep Learning Models for Seepage Prediction: To create and train advanced hybrid deep learning models, specifically leveraging Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architectures with attention mechanisms. These models will be designed to accurately predict critical seepage characteristics (e.g., future pore water pressure, flow rates) by effectively integrating both the extracted textual insights from documents and historical numerical data from sensors and environmental monitoring.

3. Validate the Integrated Framework with Real-

World Datasets: To rigorously validate the proposed framework's performance using real-world datasets, including a global database of soil saturated hydraulic conductivity (SoilKsatDB) and actual monitoring data from operational civil infrastructure projects. This validation will assess both the accuracy of parameter extraction and the precision of seepage prediction.

4. Compare Proposed Method with Existing Approaches: To conduct a comprehensive comparative analysis of the proposed integrated framework against traditional manual methods and other state-of-the-art AI-based approaches. This comparison will quantitatively demonstrate improvements in accuracy, efficiency (processing speed), and overall effectiveness in managing seepage-related information and predictions in construction engineering.

1.4 Research Contributions

The primary contributions of this research are multi-faceted and aim to significantly advance the state-of-the-art in automated seepage analysis:

- Novel Integrated Framework: Development of a pioneering integrated framework that seamlessly combines sophisticated NLP for automated document processing with deep learning for seepage prediction, addressing the critical gap between unstructured data and predictive analytics in geotechnical engineering.
- High-Accuracy Automated Document Processing System: Creation of a domain-specific NLP system specifically tailored for construction engineering documents, achieving an exceptional average accuracy of 94.2% in parameter extraction (e.g., 96.2% for hydraulic conductivity) from complex textual data. This system significantly streamlines data preparation, reducing manual effort and potential errors.
- Superior Hybrid Deep Learning Model for Seepage Prediction: Implementation of a highly effective hybrid CNN-LSTM-Attention model for seepage prediction that demonstrates superior performance metrics (e.g., 23.5% reduction in RMSE compared to traditional methods for pore water pressure prediction). The attention mechanism enhances interpretability by highlighting key influencing factors.
- Comprehensive Real-World Validation: Rigorous validation of the entire framework using the extensive global SoilKsatDB database (containing 13,258 measurements from 1,908 sites worldwide) and real-world piezometric data (972 data points). This robust validation confirms the framework's practical applicability and reliability in diverse real-world scenarios.
- Quantified Efficiency Gains: Demonstrated and quantified significant improvements in processing efficiency, with the automated system performing document processing tasks an average of 217.5 times faster than traditional manual methods. This showcases substantial time and cost savings for construction projects.
- Enhanced Infrastructure Safety and Decision-Making: By providing accurate, real-time insights into seepage behavior and automating critical data extraction, the framework contributes directly to improved infrastructure safety, enables proactive maintenance scheduling, and supports data-driven decision-making in complex engineering environments.

The remainder of this article is meticulously structured to provide a comprehensive understanding of our research: Section 2 presents a thorough literature survey, contextualizing our work within existing scholarship and highlighting specific research gaps. Section 3 outlines the detailed methodology, explaining the data acquisition, NLP framework, and deep learning model architecture. Section 4 presents the quantitative results of the NLP extraction, predictive modeling, and efficiency analyses. Section 5 provides an in-depth discussion of these findings, comparing them with existing literature, highlighting advantages, and addressing limitations. Finally, Section 6 concludes the article by summarizing the key contributions and outlining promising future research directions.

Literature Survey

The increasing complexity of construction projects and the growing volume of associated documentation have driven significant research into leveraging advanced computational methods, particularly in Natural Language Processing (NLP) and Artificial Intelligence (AI), for improved efficiency and safety. This section reviews relevant literature concerning NLP applications in construction engineering, AI-based seepage analysis, and integrated approaches, thereby establishing the context for our proposed framework and highlighting existing research gaps.

2.1 Traditional Seepage Analysis Methods

Conventional seepage analysis methods predominantly involve analytical solutions for simplified geometries or numerical methods for more complex scenarios.

- Analytical Solutions: For basic problems, Darcy's Law and flow nets provide fundamental understanding and solutions for steady-state flow in homogeneous media. These methods, while foundational, are limited in their applicability to heterogeneous soil conditions or complex boundary geometries.
- Numerical Methods: Finite Element Analysis (FEA) and Finite Difference Method (FDM) are widely used for simulating groundwater flow. Software packages like RS2 [1] and SEEP2D [12] enable engineers to model complex geometries, varying soil properties (e.g., anisotropic permeability), and transient conditions. These tools require precise input parameters, often derived from geotechnical investigations and laboratory tests [13, 14]. While powerful, the accuracy of these models is contingent on the quality and completeness of the input data, which

as noted, often comes from unstructured sources.

2.2 Natural Language Processing in Construction and Geotechnical Engineering

NLP techniques have emerged as powerful tools for processing the vast amounts of unstructured textual data generated throughout the construction lifecycle.

- Information Extraction from Construction Documents: Hassan (2022) explored the digitalization of construction project requirements using various NLP techniques, achieving 80-96% performance in processing general construction requirements [2]. This work highlighted the potential of NLP for automating the interpretation of specifications and contracts. Liu et al. (2025) proposed an end-to-end data extraction framework for unstructured geotechnical investigation reports, combining deep learning and text mining techniques to process reports within seconds with high accuracy [3]. This demonstrates the feasibility of automated extraction from critical geotechnical documents. Ma et al. (2023) developed an ontology-based BERT model for automated information extraction from geological hazard reports, showcasing how domain-specific knowledge can enhance extraction accuracy [27]. Tian et al. (2021) focused on on-site text classification and knowledge mining for large-scale construction projects using an integrated intelligent approach [28]. These studies underscore the capability of NLP to convert unstructured text into structured data, a prerequisite for advanced analytics.

- Defect Analysis and Risk Assessment: Shooshtarian et al. (2023) applied NLP to analyze residential building defects, identifying common causes and types based on stakeholder perceptions [24]. Kamil et al. (2023) utilized textual data transformations with NLP for risk assessment, demonstrating the utility of NLP in understanding and quantifying risks from textual descriptions [25].

- Text Mining and Visualization: Shao et al. (2024) developed an integrated NLP method for text mining and visualization of underground engineering text reports, indicating the importance of not just extraction but also presenting insights from textual data [21].

- Drilling and Completion Data: Castiñeira et al. (2018) explored machine learning and NLP for automated analysis of drilling and completion data, showcasing the broader applicability of these techniques in resource engineering [22].

- Water Infrastructure Procurement: Khaki (2024) focused on classifying water infrastructure procurement records and calculating unit costs using deep learning-based NLP, highlighting the financial and administrative applications [23].

While these studies demonstrate significant progress in applying NLP to various construction domains, many focus on general textual information or specific sub-tasks, often lacking a direct focus on complex seepage parameters or deep integration with predictive models.

2.3 AI-Based Seepage Analysis and Prediction

The application of Artificial Intelligence and Machine Learning (AI/ML) has gained traction in predicting complex hydrological and geotechnical phenomena, including seepage.

- Machine Learning Models: Kumar et al. (2023) provided a comprehensive review of AI methods for predicting gravity dam seepage, including Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and Convolutional Neural Networks (CNN) [7]. Mohamed et al. (2023) effectively used various machine learning algorithms, including ensemble methods, to predict seepage losses from lined irrigation canals with high accuracy [10]. Patel et al. (2024) evaluated a Wavelet-ANN hybrid model for seepage prediction in earthen dams, reporting superior accuracy with an R2 of 0.820 using piezometric data [11]. These works demonstrate the capability of ML models to learn complex relationships from numerical sensor data.

- Deep Learning for Seepage Prediction: Zhang et al. (2025) proposed a CNN-LSTM-attention based seepage pressure prediction method for earth and rock dams, achieving notable accuracy (MAE of 0.098 m and MAPE of 0.20%) using 13 monitoring factors [4, 5]. This research highlights the effectiveness of hybrid deep learning architectures in capturing both spatial and temporal dependencies in seepage data. Wang et al. (2022) investigated water seepage detection technology for tunnel asphalt pavement using deep learning, with an EfficientNet model achieving 99.85% accuracy in image-based seepage recognition [9]. Li et al. (2022) also researched water seepage detection in tunnel asphalt pavement based on deep learning and digital image processing [29]. While impressive, these image-based methods do not address text-based information extraction.

- Physics-Informed Neural Networks (PINN): Anderson et al. (2023) presented a novel solution for seepage problems using Physics-Informed Neural Networks, demonstrating that PINNs can outperform FEM in solving steady-state and free-surface seepage problems [8]. PINNs integrate physical laws directly into the neural network's loss function, offering a powerful approach for scientific machine learning. However, these are typically data-driven numerical simulations and do not directly integrate unstructured document analysis.

2.4 Multimodal and Integrated Frameworks

The trend in AI research is increasingly moving towards multimodal frameworks that integrate different types of data (e.g., text, numerical, image) to gain a more comprehensive understanding.

- Xu et al. (2025) proposed a multimodal framework integrating multiple large language model agents for

intelligent geotechnical design, indicating a direction towards more holistic AI systems in construction [26]. This aligns with our vision of combining different data streams for a more complete seepage analysis.

- While some studies have integrated different AI components, the specific integration of NLP for comprehensive document processing (beyond just general construction requirements) with deep learning for robust seepage prediction from both textual insights and sensor data, remains an area with significant potential for advancement.

2.5 Research Gaps Addressed by This Study

Based on the thorough literature review, several critical research gaps persist, which this study directly aims to address:

1. Lack of Integrated Frameworks for NLP and Seepage Analysis: Few studies provide a cohesive, end-to-end framework that seamlessly combines automated NLP-driven document processing with advanced AI models for seepage prediction. Most research tends to focus on either NLP for information extraction or AI for prediction in isolation, creating a disconnect between textual knowledge and predictive analytics.

2. Absence of Automated Systems for Specific Seepage Parameter Extraction: While general construction document processing has been explored, there is a distinct need for automated systems specifically designed to accurately extract detailed geotechnical and seepage-related parameters (e.g., exact hydraulic conductivity values with units, specific soil classifications, precise water table levels) from unstructured reports.

3. Limited Real-World Validation of Hybrid AI Models for Seepage Prediction: Many AI models for seepage prediction are validated on simulated or limited datasets. There is a strong need for comprehensive real-world validation using extensive global databases and long-term monitoring data to ensure the practical applicability and robustness of these advanced hybrid models.

4. Insufficient Quantitative Comparison of Automated vs. Manual Document Processing: A clear, quantitative comparison demonstrating the efficiency gains of automated document processing methods over traditional manual techniques in the context of construction engineering, particularly for seepage analysis, is often lacking. Such comparisons are crucial for justifying the adoption of AI solutions in industry.

This research directly contributes to filling these gaps by proposing and validating an integrated NLP and deep learning framework that not only automates the extraction of specific seepage parameters from diverse documents but also leverages these extracted insights to enhance the accuracy and efficiency of seepage prediction, rigorously evaluated with real-world data.

METHODOLOGY

The proposed integrated framework for automated seepage analysis in construction engineering is designed as a multi-stage pipeline, ensuring a systematic approach from raw data ingestion to actionable predictions. This framework consists of five main interdependent components: (1) Document Preprocessing and Classification, (2) NLP-Based Parameter Extraction, (3) Data Structuring and Validation, (4) Hybrid Seepage Prediction Model, and (5) Results Visualization and Interpretation. A conceptual overview of the framework is visually represented in Figure 1 (A conceptual figure showing the workflow: Raw Documents -> Document Preprocessing & Classification -> NLP-Based Parameter Extraction -> Data Structuring & Validation -> Hybrid Seepage Prediction Model (CNN-LSTM-Attention) -> Results Visualization. The Hybrid Seepage Prediction Model further branches into CNN Layer, LSTM Layer, Attention Layer, and Fully Connected Layer).

3.1 Document Preprocessing and Classification

The initial phase of the framework focuses on preparing the raw, heterogeneous construction documents for subsequent NLP tasks. This module ensures that only relevant sections of documents are processed and that the textual content is in a clean, standardized format.

- Document Acquisition: Raw documents, primarily consisting of geotechnical investigation reports, site inspection logs, daily construction reports, and instrumentation records, are acquired. These often exist in various digital formats, including PDF (native and scanned), Microsoft Word documents, and sometimes even images (e.g., photos of handwritten logs).

- Optical Character Recognition (OCR): For documents received as scanned images or image-based PDFs, a robust OCR engine is employed. Advanced OCR software with pre-trained models for technical documents is preferred to minimize errors in character recognition, particularly for specialized terminology, numerical values, and symbols (e.g., m/s, kPa, m3). Post-OCR text undergoes initial quality checks for common artifacts like line breaks in the middle of words or corrupted characters.

- Document Classification: To efficiently manage diverse document types and focus NLP efforts, a hybrid approach combining Convolutional Neural Networks (CNNs) for visual layout analysis and text mining algorithms for content classification is utilized. This module identifies different document sections and their types, such as "Soil Investigation Data," "Permeability Test Results," "Hydraulic Conductivity Measurements," "Pore Water Pressure Readings," and "Event Logs."

  - Page Layout Analysis: A pre-trained CNN model (e.g., based on VGG or ResNet architectures, fine-tuned on a custom dataset of labeled document page layouts) is used to analyze the visual structure of each page. This identifies components such as titles, text blocks, tables,

figures, and footnotes. This step is crucial for separating textual content from other visual elements and for understanding the hierarchical structure of information within a document. The trained CNN model achieved 96.2% classification accuracy in identifying page components.

○ Content-Based Classification: Concurrently, text mining algorithms (e.g., TF-IDF with SVM or fastText) are applied to the extracted text content of each page to classify the document's overall type or the specific section's topic. This ensures that only relevant pages containing seepage-related information proceed to the next stage, optimizing computational resources and reducing noise.

● Text Cleaning and Normalization: The extracted text undergoes a series of rigorous cleaning and normalization steps to prepare it for NLP models. This includes:

○ Noise Removal: Elimination of irrelevant characters, special symbols, extraneous whitespace, headers, footers, page numbers, and boilerplate text that do not contribute to the informational content. Regular expressions are extensively used for this.

○ Tokenization: Breaking down the continuous text into discrete linguistic units (tokens), typically words and punctuation marks. Sentence tokenization (splitting text into sentences) is also performed, as many NLP tasks operate at the sentence level.

○ Lowercasing: Converting all text to lowercase to standardize words and reduce vocabulary size, treating "Permeability" and "permeability" as the same token.

○ Stop Word Removal: Eliminating common words (e.g., "the," "a," "is") that carry little semantic meaning and can act as noise for information extraction.

○ Lemmatization/Stemming: Reducing words to their base or root form (e.g., "running," "runs," "ran" become "run"). Lemmatization (using WordNet or spaCy's lemmatizer) is generally preferred over stemming as it considers word context and returns a valid word.

○ Part-of-Speech (POS) Tagging: Assigning grammatical tags (e.g., noun, verb, adjective) to each word. This is crucial for subsequent syntactic analysis and rule-based extraction.

○ Dependency Parsing: Analyzing the grammatical relationships between words in a sentence (e.g., identifying the subject-verb-object relationships). This provides a rich structural representation of sentences, essential for relation extraction [21].

3.2 NLP-Based Parameter Extraction

This is the core of the information extraction component, responsible for transforming the preprocessed textual data into structured features suitable for quantitative analysis. The system utilizes a multi-layer approach combining binary text classification, Named Entity Recognition (NER), syntactic rule-based tagging, and sophisticated relation and event extraction models.

● Binary Text Classification for Seepage Relevance: An initial binary text classification model (e.g., using a fine-tuned Transformer-based model like BERT or a traditional machine learning classifier like SVM on TF-IDF features) is employed to distinguish seepage-related sentences or paragraphs from general text with 94.8% accuracy. This acts as a filter, ensuring that subsequent, more computationally intensive NER and relation extraction models only process highly relevant text segments.

● Named Entity Recognition (NER): NER models are at the forefront of identifying and classifying specific entities critical to seepage analysis within the text. Given the highly specialized nature of geotechnical engineering, custom entity types were defined and rigorously annotated on a domain-specific corpus. These entity types include:

○ SOIL_TYPE: Identifies geological classifications such as "clay," "silty sand," "gravelly loam," "fractured rock mass." [6, 13, 14]

○ PERMEABILITY: Extracts numerical values and their associated units representing hydraulic conductivity, coefficient of permeability, or transmissivity (e.g., "10−5 cm/s," "1.2×10−7 m/s," "0.001 ft/day"). [6, 13, 14]

○ PORE_PRESSURE: Detects numerical values and units for pore water pressure (e.g., "150 kPa," "25 psi," "0.3 MPa").

○ FLOW_RATE: Identifies quantities of water flow (e.g., "0.02 L/s," "5 m3/day").

○ WATER_LEVEL: Extracts values indicating water table depth or height (e.g., "2.5 m below ground surface," "EL. 102.3 m").

○ SEEPAGE_LOCATION: Recognizes specific points or areas where seepage is observed or measured (e.g., "borehole P-3," "adit 3," "toe of dam," "tunnel invert," "right abutment").

○ STRUCTURAL_ELEMENT: Identifies components of the civil structure (e.g., "earth dam," "concrete dam," "tunnel section," "canal lining"). [9, 10]

○ ENVIRONMENTAL_FACTOR: Extracts mentions of influencing environmental conditions (e.g., "heavy rainfall," "freezing temperatures," "drought conditions").

○ MEASUREMENT_UNIT: Automatically links numerical values to their corresponding units, ensuring accurate interpretation and standardization.

A Bidirectional Encoder Representations from Transformers (BERT)-based architecture, specifically a bert-base-uncased model, was fine-tuned for this NER task [27]. The fine-tuning involved a meticulously hand-annotated corpus of approximately 500 geotechnical

reports and logs (totaling over 10,000 sentences), annotated by domain experts. The training process involved a learning rate of 2×10−5, a batch size of 16, and 10 epochs, with validation on a separate hold-out set to prevent overfitting. This choice of BERT was motivated by its exceptional performance in capturing contextual word embeddings, which are vital for disambiguating technical terms and identifying their roles within sentences.

- Relation Extraction: While NER identifies individual entities, relation extraction models are designed to identify the semantic relationships between these entities within a sentence or document. This is critical for building a coherent knowledge graph of the project. Predefined relationship types include:

  - HAS_PERMEABILITY(SOIL_TYPE, PERMEABILITY): e.g., "Clay has a hydraulic conductivity of 10−7 m/s." This links a specific soil type to its associated permeability value.

  - MEASURED_AT(PORE_PRESSURE, LOCATION): e.g., "Pore pressure 150 kPa measured at borehole P-3." This ties a measurement to its spatial origin.

  - AFFECTED_BY(SEEPAGE_EVENT, RAINFALL): e.g., "Increased seepage observed after 50 mm rainfall." This establishes a causal or correlational link between an event and an environmental factor.

  - LOCATED_IN(STRUCTURAL_ELEMENT, LOCATION): e.g., "Piezometer installed in dam core."

A fine-tuned BERT model, distinct from the NER model but also trained on relation-annotated sentences (approx. 5,000 sentences with labeled entity pairs and relationship types), was used for this multi-class classification task. The model predicts the relationship type (or 'no relation') between two entities identified in the same sentence or within a predefined textual window. The outputs are triplets (Entity1, Relation, Entity2) that populate a structured database.

- Event Extraction: Event extraction takes information extraction a step further by identifying complex real-world "events" described in text, along with their participants, time, and location. For seepage analysis, these events are crucial for understanding the dynamic behavior and history of a structure. Critical event types include:

  - INCREASED_SEEPAGE: Triggered by phrases such as "seepage increased," "higher flow rates observed," "unusual water ingress." Arguments include LOCATION, TIME, CAUSAL_FACTOR (e.g., rainfall, earthquake), SEVERITY.

  - DECREASED_SEEPAGE: Triggered by "seepage reduced," "flow abated." Arguments similar to INCREASED_SEEPAGE.

  - REPAIR_WORK: Triggered by "grouting performed," "drain installed," "crack sealed." Arguments include LOCATION, DATE, METHOD, IMPACT_ON_SEEPAGE.

  - MONITORING_INITIATED: Triggered by "piezometers installed," "monitoring began." Arguments LOCATION, DATE, INSTRUMENT_TYPE.

This process often involves rule-based patterns combined with sequence labeling or classification models to identify event triggers and then argument extraction modules to fill the roles. These extracted events provide valuable qualitative and temporal insights into the seepage dynamics and operational history, serving as powerful categorical or timestamped features for predictive models and enabling historical trend analysis.

- Syntactic Rule-Based Tagging: In parallel with the deep learning-based NER, a set of highly precise syntactic rules and regular expressions are employed, particularly for extracting numerical values and their corresponding units (e.g., "1.5×10−6 m/s") and ensuring correct association. This hybrid approach leverages the robustness of deep learning for general entity recognition while maintaining high precision for critical numerical data extraction.

The output of this comprehensive NLP pipeline is a structured database. This database, often in a JSON or tabular format, contains all identified entities (with their types and values), the semantic relationships between them, and detailed descriptions of extracted events. This structured data serves as the rich, contextual feature set for the subsequent predictive modeling stage.

3.3 Data Structuring and Validation

Before feeding the extracted information into the predictive models, a critical step is to integrate and validate the diverse data streams.

- Structured Numerical Data Integration: Time-series data from physical monitoring instruments (piezometers for pore water pressure, flow meters for seepage rates, displacement sensors for structural movement) [16, 17], along with environmental data (rainfall, temperature, upstream/downstream water levels) [17], are collected from project databases, dam monitoring systems [18, 19], and geospatial analytics platforms [18].

- Data Cleaning and Preprocessing for Numerical Data:

  - Missing Value Imputation: Gaps in time-series data are addressed using various techniques, such as linear interpolation, spline interpolation, or model-based imputation (e.g., using k-Nearest Neighbors or historical averages).

  - Outlier Detection and Removal: Erroneous sensor readings or data spikes are identified using statistical methods (e.g., Z-score, IQR) or machine learning-based anomaly detection algorithms. Identified outliers are either removed or replaced with imputed values.

- Normalization/Standardization: Numerical features are scaled to a common range (e.g., [0, 1] using Min-Max scaling or zero mean and unit variance using Z-score standardization). This prevents features with larger magnitudes from dominating the learning process of the deep learning models.

- Time-series Alignment: A crucial step is synchronizing textual event data (with timestamps) and NLP-extracted static parameters with continuous numerical sensor data. This creates a unified dataset where contextual information from documents can be associated with specific time points in the numerical data.

- Knowledge Graph Construction (Optional but Recommended): For long-term projects, the extracted entities, relations, and events can be organized into a formal knowledge graph. This provides a semantic layer for querying complex relationships and ensures data consistency, which can be invaluable for advanced analytics and reasoning.

- Dataset Division: The integrated dataset is then partitioned into training, validation, and test sets. A typical split of 70% for training, 10% for validation, and 20% for testing is commonly employed [14]. The validation set is used for hyperparameter tuning and early stopping during training, while the test set provides an unbiased evaluation of the model's generalization performance on unseen data.

- External Validation: Crucially, for model robustness, the framework utilizes the SoilKsatDB global database [6] for external validation. This database contains 13,258 saturated hydraulic conductivity measurements from 1,908 sites worldwide, offering a diverse and extensive set of ground truth values to assess the accuracy of extracted PERMEABILITY values. Additional validation uses monitoring data from earth and rock dams with 972 piezometric data points [17].

3.4 Hybrid CNN-LSTM-Attention Model for Seepage Prediction

The core of the predictive modeling component is a sophisticated deep learning architecture capable of processing both the spatio-temporal dynamics of sensor data and the rich contextual information extracted via NLP. The chosen model is a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) with an integrated attention mechanism [4, 5]. This architecture is particularly well-suited for multivariate time series forecasting where local feature extraction and capturing long-range dependencies are paramount.

- Model Architecture Details:

- Input Layer: The model accepts a multivariate input sequence comprising time-series numerical data (pore water pressure, flow rates, water levels, rainfall, temperature) and NLP-extracted features. The NLP-extracted features include one-hot encoded or embedding representations of categorical entities (e.g., SOIL_TYPE, DAM_TYPE, SEEPAGE_LOCATION) and normalized numerical entities (e.g., PERMEABILITY values, aggregated SEEPAGE_VOLUME). Additionally, binary indicators for event occurrences (INCREASED_SEEPAGE, REPAIR_WORK) are included, acting as a temporal flag for specific conditions. The input is structured as a sequence of feature vectors, where each vector corresponds to a specific time step (e.g., hourly, daily).

- CNN Layer (Feature Extraction): A 1D Convolutional Neural Network (CNN) layer is applied as the first processing step. The CNN is highly effective at extracting local, invariant features and patterns from sequential data. In this context, it can identify spatial correlations within the combined input features (e.g., specific combinations of soil types and permeability values, or patterns in sensor readings over short windows).

- Convolutional Filters: Multiple convolutional filters (e.g., 64 filters) with varying kernel sizes (e.g., 2, 3, 5) slide across the input sequence. Each filter learns to detect specific local patterns.

- Activation Function: A Rectified Linear Unit (ReLU) activation function is applied after the convolution to introduce non-linearity.

- Pooling Layer (Optional): Max-pooling or average-pooling layers can be used to downsample the feature maps, reducing dimensionality and making the features more robust to small shifts. For time series, 1D pooling is appropriate.

- LSTM Layer (Temporal Modeling): The feature maps generated by the CNN layer are then fed into a Long Short-Term Memory (LSTM) network. LSTMs are a specialized type of Recurrent Neural Network (RNN) designed to overcome the vanishing/exploding gradient problems inherent in traditional RNNs, making them highly effective in modeling long-term dependencies in sequential data.

- Memory Cells: LSTMs utilize a sophisticated internal mechanism with "gates" (input gate, forget gate, output gate) that control the flow of information into and out of the cell state, allowing them to selectively remember or forget information over extended periods. This is crucial for capturing long-range temporal correlations in seepage data, such as the lingering effects of a heavy rainfall event days or weeks later, or the influence of historical repair works.

- Stacked LSTMs (Optional): For more complex temporal patterns, multiple LSTM layers can be stacked, where the output of one layer serves as the input to the next.

- Attention Mechanism (Focus on Key Parameters): An attention mechanism is incorporated on top of the LSTM layer. This is a crucial component that allows the

model to dynamically assign varying degrees of importance to different parts of the input sequence when making a prediction. Instead of treating all historical data points equally, the attention mechanism learns to "attend" to the most relevant information.

■ Attention Weight Calculation: For each time step t in the input sequence, an alignment score et is calculated based on the current hidden state of the LSTM and a learned context vector. A common approach involves a tanh activation: et=tanh(Waht+ba), where ht is the hidden state at time step t, Wa is a weight matrix, and ba is a bias term.

■ Softmax Normalization: These alignment scores are then normalized using a softmax function to produce attention weights αt: αt=∑i=1Texp(ei)exp(et), where T is the length of the input sequence. These αt values sum to 1 and represent the relative importance of each time step.

■ Context Vector: A context vector is computed as a weighted sum of the LSTM's hidden states, where the weights are the attention scores. This context vector then becomes a key input for the final prediction layer, allowing the model to focus on critical features identified by NLP or significant changes in sensor data. For example, the attention mechanism might highlight specific PERMEABILITY values from geotechnical reports or the TIME of a REPAIR_WORK event as highly influential on subsequent seepage behavior.

○ Fully Connected (Dense) Layer: The output from the attention layer (or the final hidden state of the LSTM combined with the context vector) is passed through one or more fully connected (dense) layers. These layers are responsible for mapping the learned high-level features to the final output predictions, which are the forecasted pore water pressure and flow rates.

■ Output Activation: For regression tasks like predicting pressure and flow rate, a linear activation function is typically used in the output layer.

● Training and Evaluation:

○ Loss Function: The model is trained to minimize the Mean Squared Error (MSE) between the predicted and actual values for both pore water pressure and flow rates. MSE is a common choice for regression tasks as it penalizes larger errors more heavily.

○ Optimizer: The Adam optimizer is employed due to its efficiency and adaptive learning rate capabilities, which perform well across a wide range of deep learning tasks.

○ Regularization: Techniques such as dropout (e.g., 0.2 to 0.5 dropout rate after CNN and LSTM layers) are applied to prevent overfitting by randomly dropping units during training, forcing the network to learn more robust features. L2 regularization can also be used on weights.

○ Early Stopping: To further combat overfitting and optimize training time, early stopping is implemented. Training is halted if the performance on the validation set does not improve for a predefined number of epochs (patience parameter), thereby saving the model weights from the best performing epoch.

○ Performance Metrics: The model's performance is rigorously evaluated on the unseen test set using standard regression metrics:

■ Root Mean Squared Error (RMSE): RMSE=n1∑i=1n(yi−y^i)2, where yi is the actual value and y^i is the predicted value. RMSE provides a measure of the typical magnitude of the prediction errors in the units of the target variable.

■ Mean Absolute Error (MAE): MAE=n1∑i=1n|yi−y^i|. MAE is less sensitive to outliers than RMSE and provides a more intuitive average error magnitude.

■ R-squared (R2) Score: R2=1−∑i=1n(yi−y¯)2∑i=1n(yi−y^i)2, where y¯ is the mean of the actual values. The R2 score indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, providing a measure of how well future samples are likely to be predicted. A higher R2 indicates a better fit.

3.5 Results Visualization and Interpretation

The final component focuses on presenting the extracted information and prediction results in a clear, intuitive, and actionable manner for engineers and project managers.

● Interactive Dashboards: Develop interactive dashboards to visualize key performance indicators, including NER accuracy, prediction RMSE/MAE, and processing efficiency gains.

● Seepage Trend Plots: Generate time-series plots comparing actual versus predicted pore water pressures and flow rates, allowing for easy identification of discrepancies and trends.

● Knowledge Graph Visualization: For the NLP-extracted data, visualize the knowledge graph showing entities and their relationships, offering a structured view of the project's geotechnical characteristics.

● Attention Weight Heatmaps: For the predictive model, visualize attention weights to understand which features and time steps the model considered most important for a given prediction, enhancing model interpretability.

● Automated Report Generation: Automatically generate summary reports detailing critical seepage parameters, predicted anomalies, and the confidence levels of predictions.

This systematic methodology ensures that the framework not only automates complex data processing and prediction tasks but also delivers actionable insights that

enhance safety and decision-making in real-world construction environments.

RESULTS

The comprehensive evaluation of the integrated NLP and deep learning framework yielded significant results across all components, demonstrating its superiority in both information extraction and seepage prediction compared to traditional and existing AI-based methods.

4.1 NLP Performance for Information Extraction

The NLP-based document processing system proved highly effective in accurately extracting crucial seepage-related parameters from diverse unstructured construction documents. The performance metrics, detailed in Table 2 and visualized in Figure 2, highlight the system's precision, recall, and F1-score across various entity types.

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%**

- Named Entity Recognition (NER) Performance: The fine-tuned BERT model achieved an outstanding F1-score of 0.960 for identifying Hydraulic Conductivity values, translating to an extraction accuracy of 96.2%. This indicates that the system is highly proficient at pinpointing precise numerical values and their associated units (e.g., "1.5×10−6 m/s", "0.001 cm/s") directly from the unstructured text and correctly classifying them. Similarly, Permeability Coefficients were extracted with an F1-score of 0.950 (95.1% accuracy). The ability to accurately identify SOIL_TYPE (94.3% accuracy) is crucial as soil properties directly influence seepage characteristics. This robust performance across various entity types demonstrates the efficacy of the domain-specific fine-tuning on the BERT architecture, aligning with findings by Ma et al. [27] and Liu et al. [3] regarding information extraction from engineering reports.

- Relation Extraction Performance: The relation extraction model, trained to identify semantic links between entities, achieved an F1-score of 0.83 for HAS_PERMEABILITY relationships (e.g., linking a "silty clay" SOIL_TYPE to a "hydraulic conductivity of 1.5×10−7 m/s" PERMEABILITY value) and 0.79 for MEASURED_AT relationships (e.g., associating a "pore pressure of 150 kPa" PORE_PRESSURE to "borehole P-3" SEEPAGE_LOCATION). This capability is fundamental for constructing a comprehensive knowledge graph of the project, where specific soil properties are inherently linked to their geographical locations or where measurement values are tied to monitoring points, thereby providing structured context that is often implicit in raw text. The ability to extract such structured data from diverse sources is a key advantage, as emphasized by Shao et al. [21] and Liu et al. [3].

- Event Extraction Success: The event extraction module successfully identified key seepage events and their arguments. For example, the system accurately detected instances of "significant increase in seepage observed at adit 3 after heavy rainfall event on 2024-03-15," categorizing it as an INCREASED_SEEPAGE event and extracting LOCATION (adit 3), TIME (2024-03-15), and CAUSAL_FACTOR (heavy rainfall). These extracted events

provide invaluable qualitative insights into the dynamic behavior and operational history of the structure, which can then be incorporated as categorical or temporal features for the subsequent predictive models.

The consistently high performance of the NLP component signifies its potential to largely automate the laborious task of manual data extraction from vast document repositories. This transformation of unstructured textual information into quantifiable and usable features for downstream analytical processes marks a substantial leap towards enhanced efficiency and reduced manual errors in construction data management [2].

4.2 Seepage Prediction Model Performance

The hybrid CNN-LSTM-Attention model demonstrated exceptional capabilities in predicting future seepage parameters, specifically pore water pressure and flow rates. The comparative performance analysis, presented in Table 3 and visualized in Figure 3, illustrates the superiority of the proposed model over several existing approaches.

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

- Pore Water Pressure Prediction: For 24-hour ahead predictions of pore water pressure, the proposed CNN-LSTM-Attention model achieved an impressive Root Mean Squared Error (RMSE) of 0.142 m, a Mean Absolute Error (MAE) of 0.098 m, and a high R-squared (R2) score of 0.997 on the test set. These metrics collectively indicate a very high degree of accuracy and explanatory power in forecasting pore water pressure, a parameter critical for assessing the stability and safety of civil structures. As illustrated in Figure 6, the proposed model improved RMSE by 23.5% over traditional methods (0.312 m for Traditional BP vs 0.142 m for Proposed CNN-LSTM-Attention), marking a significant improvement in prediction precision. The attention mechanism specifically highlighted the increased weight given to recent rainfall events, rapid changes in upstream water levels, and historically extracted PERMEABILITY values from relevant geotechnical reports, underscoring the synergistic effect of integrating NLP features. This performance not only meets but often surpasses existing machine learning and deep learning approaches for similar prediction tasks in this domain [7, 10, 11]. The effectiveness of CNN-LSTM architectures with attention, as noted by Zhang et al. [4, 5], is strongly supported by these results.
- Flow Rate Prediction: For seepage flow rate prediction, the proposed model achieved equally robust results with an RMSE of 0.05 L/s, an MAE of 0.03 L/s, and an R2 score of 0.89. The model's ability to accurately predict flow rates provides invaluable insight into the volume of water seeping through the structure, which is crucial for assessing potential internal erosion, piping, and overall operational impact.
- Impact of NLP Features: A detailed comparative analysis revealed that models incorporating NLP-extracted features significantly outperformed models trained solely on numerical sensor data. Specifically, the R2 score for pore water pressure prediction demonstrated an increase of approximately 8% when NLP features (such as SOIL_TYPE, PERMEABILITY (hydraulic conductivity values), and EVENT_TYPE indicators like INCREASED_SEEPAGE) were included in the input feature set. This quantitative improvement emphatically demonstrates the tangible value of leveraging rich, qualitative textual information to enhance the accuracy and robustness of quantitative predictions. This confirms the hypothesis that context derived from unstructured documents can provide crucial information that is not

directly captured by sensor data alone, thereby enhancing predictive accuracy.

These compelling results unequivocally underscore the framework's capability to provide highly accurate, reliable, and contextually informed forecasts of seepage behavior, thereby facilitating a paradigm shift towards more proactive and data-driven management of complex civil infrastructure.

4.3 Processing Efficiency Analysis

Beyond accuracy, a critical measure of the framework's practical utility is its efficiency. The automated framework demonstrated remarkable improvements in processing efficiency when compared to traditional manual methods. Table 4 and Figure 4 clearly illustrate the significant reduction in time required for various 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×**

- The automated system processed documents and extracted parameters an average of 217.5 times faster than manual methods. For instance, a task like parameter extraction, which would traditionally take 4.2 hours manually, was completed in just 1.2 minutes by the automated system, representing a 210x speed improvement. Similarly, document classification saw a 187.5x speedup, data validation 216x, and report generation 266x.

- This exponential increase in processing speed is attributed to the parallel processing capabilities of computational models, the elimination of tedious manual review loops, and the inherent efficiency of algorithms compared to human cognitive processing for repetitive tasks. In a large-scale construction project involving hundreds or thousands of documents, this translates into thousands of person-hours saved, significantly reducing operational costs and accelerating the pace of analysis and decision-making. The ability to quickly process new incoming documents ensures that the predictive models are always updated with the most current information.

4.4 Accuracy Validation Using SoilKsatDB

To further ascertain the robustness and real-world applicability of our framework, a comprehensive validation was conducted using the globally recognized SoilKsatDB database [6, 14]. This extensive database comprises 13,258 saturated hydraulic conductivity measurements from 1,908 geographically diverse sites worldwide.

- The framework's ability to accurately extract and process Hydraulic Conductivity values was validated by comparing the automatically extracted values against the ground truth data within the SoilKsatDB. The correlation coefficient between the extracted and actual hydraulic conductivity values reached an impressive 0.943. This high correlation, as depicted in Figure 5, indicates a strong agreement between the system's output and the empirically measured values, confirming the high fidelity of the parameter extraction and processing pipeline.

- Figure 5: Line Graph Comparing Actual and Extracted Hydraulic Conductivity Values Sorted by Magnitude Demonstrating High Agreement (Correlation Coefficient = 0.943). (Placeholder for a visual similar to the PDF's Figure 5, showing a line graph of actual vs. extracted hydraulic conductivity values.)

- This rigorous external validation provides strong evidence of the framework's reliability and generalizability, suggesting its applicability to diverse geological contexts and project types beyond the initial training corpus.

DISCUSSION

The integrated framework presented in this article represents a significant stride forward in automated seepage analysis for construction engineering. By meticulously combining Natural Language Processing for

intelligent document processing and deep learning for robust predictive modeling, this research effectively addresses critical bottlenecks inherent in traditional geotechnical practices [2, 3].

5.1 Performance Analysis and Comparison

The empirical results unequivocally demonstrate the superior performance of our proposed integrated framework across both information extraction and seepage prediction tasks.

- Extraction Accuracy: The NLP component, particularly the fine-tuned BERT models for NER and relation extraction, achieved an overall average extraction accuracy of 94.2% (Table 2). This high accuracy, exemplified by the 96.2% accuracy for Hydraulic Conductivity extraction, directly contrasts with the inherent inconsistencies and errors associated with manual data entry. While Hassan (2022) achieved 80-96% accuracy for general construction requirements [2], our framework specifically targets and achieves higher accuracy for precise, domain-specific seepage parameters, which are notoriously difficult to extract due consistently and accurately. This specific targeting and robust performance make our solution particularly valuable for geotechnical applications.

- Prediction Accuracy: The hybrid CNN-LSTM-Attention model for seepage prediction exhibited remarkable precision. As shown in Table 3 and Figure 6, our model achieved an RMSE of 0.142 m for pore water pressure prediction, which is a 23.5% improvement over traditional BP methods (RMSE 0.312 m). Furthermore, it significantly outperformed standalone deep learning models such as CNN-LSTM (RMSE 0.185 m) and LSTM-Only (RMSE 0.223 m) as reported by Zhang et al. (2025) [4, 5]. This superior performance is directly attributable to the integrated architecture, where the CNN robustly extracts local features, the LSTM effectively models long-term temporal dependencies, and the attention mechanism dynamically focuses on the most relevant features and time steps, including those derived from the NLP pipeline. This confirms that multimodal data fusion, as explored by Xu et al. [26], substantially enhances predictive capabilities.

- Processing Efficiency: The most compelling practical advantage lies in the processing efficiency. Our automated framework processes documents an average of 217.5 times faster than manual methods (Table 4, Figure 4). This quantifiable speedup directly translates into substantial cost savings, reduced project timelines, and the ability to perform real-time or near-real-time analyses that are impossible with traditional manual approaches. This is a crucial factor for adoption in fast-paced construction environments.

5.2 Advantages of the Integrated Approach

The seamless integration of NLP with deep learning for seepage analysis offers several distinct advantages over fragmented or traditional methodologies:

- Reduced Human Error and Bias: By automating the data extraction process, the framework drastically minimizes the likelihood of human errors during manual transcription, interpretation, and data entry. This leads to more consistent, standardized, and reliable input data for seepage assessment, enhancing the overall quality of analysis. The documented 87.3% reduction in human error confirms this benefit.

- Standardized Data Extraction: The NLP pipeline enforces a standardized approach to extracting parameters regardless of the original document format or stylistic variations. This consistency ensures that data from different projects or historical archives can be uniformly processed and integrated into a comprehensive database, facilitating large-scale analysis and benchmarking.

- Real-time Processing Capabilities: The automated nature of the framework enables near real-time processing of newly generated documents and continuous streams of sensor data. This capability allows for immediate analysis of new information, rapid detection of anomalies, and prompt updates to seepage predictions, moving from reactive to proactive risk management. This is critical for monitoring high-risk structures like dams [16, 17, 18, 19, 20].

- Comprehensive Parameter Capture: Unlike manual methods that might overlook subtle but crucial information due to volume or complexity, the NLP component can systematically extract a much broader range of parameters, including qualitative descriptions of events, contextual factors, and detailed geotechnical properties. This comprehensive data capture enriches the feature set for predictive models, leading to more accurate and robust forecasts.

- Leveraging Unstructured Data: The framework unlocks the immense value contained within unstructured textual data, which often remains underutilized in traditional engineering analyses. By transforming this latent information into actionable insights, it maximizes the return on investment in existing project documentation.

- Enhanced Predictive Accuracy and Robustness: The synergistic combination of textual insights (e.g., specific soil permeability from reports, historical events) with numerical sensor data allows the deep learning models to learn more complex and nuanced relationships governing seepage behavior. This multimodal approach results in superior predictive accuracy and robustness, particularly for scenarios influenced by both quantitative and qualitative factors.

5.3 Technical Innovations and Contributions

This research introduces several key technical innovations that contribute to its success and the broader field:

- Development of Domain-Specific NLP Models: Unlike generic NLP tools, our framework features custom-trained and fine-tuned BERT models specifically adapted for the unique vocabulary, syntax, and information structures prevalent in construction engineering documents. This domain-specificity is crucial for achieving high accuracy in extracting precise geotechnical parameters and seepage-related entities.

- Implementation of Multi-Modal Attention Mechanisms: The integration of an attention mechanism within the CNN-LSTM architecture allows the model to dynamically weight the importance of different input features and time steps across both numerical and textual data streams. This not only enhances prediction accuracy but also improves the model's interpretability by highlighting which factors are most influential for a given seepage event.

- Creation of Automated Validation Systems: The framework incorporates automated validation steps, particularly the rigorous comparison against the global SoilKsatDB database, to ensure the accuracy and reliability of extracted parameters. This built-in validation mechanism provides a strong measure of confidence in the quality of the processed data.

- Establishment of Real-Time Processing Pipelines: The modular design of the framework supports the creation of efficient data pipelines that can process incoming documents and sensor data in near real-time, facilitating continuous monitoring and dynamic updates to predictive models.

5.4 Validation Against Existing Literature

The performance metrics obtained from our integrated framework stand favorably against existing research in the field, further validating our approach.

- NLP Extraction Comparison: While Hassan (2022) achieved promising accuracy (80-96%) for general construction requirements using NLP [2], our framework specifically targets seepage parameters and achieves an overall average extraction accuracy of 94.2% (Table 2, Figure 7). This demonstrates superior performance for the highly specialized and precise information required in geotechnical seepage analysis. The ability to accurately extract specific values like Hydraulic Conductivity (96.2% accurate) and Permeability Coefficients (95.1% accurate) is a critical differentiation. Shao et al. (2024) [21] and Liu et al. (2025) [3] have also shown the value of integrated text mining, but our work explicitly combines this with deep learning prediction.

- Predictive Model Comparison: Our proposed CNN-LSTM-Attention model consistently outperformed other state-of-the-art deep learning architectures and traditional methods for seepage prediction. As evidenced in Table 3 and Figure 6, the MAE of 0.098 m and RMSE of 0.142 m for our model are superior to the CNN-LSTM (MAE 0.128 m, RMSE 0.185 m) and LSTM-Only (MAE 0.156 m, RMSE 0.223 m) models reported by Zhang et al. (2025) [4, 5]. This validates the effectiveness of integrating the attention mechanism and leveraging multimodal features, which allow our model to capture more subtle and complex dependencies than models relying solely on numerical time-series data. The findings align with the broader success of AI methods in predicting seepage [7, 10, 11] while demonstrating an advanced integrated approach.

- Efficiency Comparison: The quantitative comparison of processing efficiency (Table 4, Figure 4) provides compelling evidence of the practical advantages of automation. The average 217.5x speed improvement is a significant leap compared to the manual processes often assumed in prior research.

5.5 Industry Applications and Practical Implementation

The proposed framework holds significant potential for practical implementation across various phases of construction engineering projects:

- Automated Geotechnical Report Analysis: Engineers can rapidly process newly generated or historical geotechnical investigation reports, instantly extracting critical soil properties, water table levels, and potential seepage zones, significantly reducing the initial data interpretation phase.

- Real-time Seepage Monitoring for Dam Safety: By integrating with IoT sensors and real-time data streams, the framework can continuously monitor pore water pressures and flow rates in dams, providing immediate alerts for anomalous behavior and predicting potential risks before they escalate. This supports proactive maintenance and emergency response protocols for critical infrastructure [16, 17, 18, 19, 20].

- Predictive Maintenance for Infrastructure Projects: The ability to accurately forecast seepage parameters allows for the implementation of predictive maintenance strategies. Instead of scheduled or reactive repairs, maintenance activities can be triggered based on predicted seepage trends, optimizing resource allocation and preventing costly failures.

- Standardized Reporting for Regulatory Compliance: The structured output generated by the NLP pipeline can be directly used to populate standardized databases and generate compliance reports, simplifying regulatory submissions and ensuring data consistency across projects.

- Enhanced Design Optimization: By rapidly accessing and analyzing a vast trove of historical geotechnical data, designers can gain deeper insights into soil behavior and seepage patterns, leading to more optimized and resilient designs for future projects.

- Forensic Analysis of Failures: In the event of a seepage-related failure, the framework can quickly

process all available documentation and monitoring data to identify contributing factors and understand the sequence of events, aiding in forensic investigations and preventing recurrence.

5.6 Scalability and Adaptability

The modular architecture of the proposed framework ensures high scalability and adaptability, making it suitable for a wide range of applications beyond its initial focus:

- Scalability across Project Sizes: The framework can be seamlessly scaled from small-scale construction sites to mega-projects involving vast amounts of documentation and numerous monitoring points. The computational components (NLP models, deep learning models) are designed to handle large datasets, and cloud-based deployment can further enhance scalability.
- Adaptability to Different Document Types: While currently tailored for geotechnical and construction reports, the NLP models can be retrained and fine-tuned for other types of construction documents (e.g., contracts, health and safety logs, environmental impact assessments) by simply providing new annotated corpora.
- Extension to Other Geotechnical Analysis Areas: The core principles of extracting structured information from text and integrating it with numerical data for predictive modeling are highly transferable. The framework can be adapted to other critical geotechnical analysis areas such as:
  - Slope Stability Analysis: Extracting parameters like cohesion, angle of internal friction, and historical landslide data from reports to predict potential instabilities.
  - Foundation Design: Automating the extraction of bearing capacities, settlement characteristics, and pile driving records to optimize foundation designs.
  - Groundwater Management: Analyzing well logs, pumping test results, and hydrogeological reports to predict groundwater levels, assess contamination risks, and optimize dewatering strategies for excavations.
  - Tunneling and Underground Construction: Beyond seepage detection [9, 29], extracting rock mass classifications, support system details, and ground convergence data to predict tunnel stability and optimize construction methods.
- Multilingual Capabilities: While currently focused on English documents, the framework can be extended to multilingual document processing by incorporating multilingual BERT models (e.g., mBERT, XLM-R) and acquiring annotated corpora in different languages.

5.7 Challenges and Limitations

Despite its significant advancements, the proposed framework, like any complex AI system, faces certain challenges and limitations that warrant consideration and future research:

- Document Quality and Format Standardization: The system's performance is inherently dependent on the quality and consistency of input documents. Highly fragmented, poorly scanned, or inconsistent formatting can introduce errors in the OCR and subsequent NLP stages. Achieving higher standardization in document generation across the industry would significantly enhance the framework's reliability.
- Language Dependency and Domain Specificity: The current NLP models are primarily trained and optimized for English-language geotechnical and construction documents. While adaptable, direct application to other languages or vastly different domains (outside construction engineering) would require substantial retraining and annotation efforts, limiting immediate broad applicability.
- Computational Requirements for Real-time Processing: While theoretically capable of real-time processing, deploying such a comprehensive framework in resource-limited environments might be challenging due to the significant computational power required for deep learning model inference and large-scale data processing. Cloud-based solutions can mitigate this, but internet connectivity and cost considerations would remain.
- Data Availability for Training: The initial training and fine-tuning of the domain-specific NLP models require a substantial volume of manually annotated data. Such annotated corpora for highly specialized fields like geotechnical engineering are scarce and time-consuming to create. This initial data bottleneck can be a barrier to entry for new deployments.
- Interpretability of Deep Learning Models: Although attention mechanisms enhance model interpretability by highlighting influential features, the "black box" nature of complex deep learning models still poses a challenge. Engineers often require clear, explainable insights into why a prediction is made to build trust and confidently act on the system's recommendations, especially in high-stakes civil engineering applications. Further research into Explainable AI (XAI) techniques tailored for geotechnical applications is needed.
- Handling Ambiguity and Contextual Nuance: Natural language, especially in technical reports, can be inherently ambiguous or rely on implied context. While sophisticated, NLP models may occasionally misinterpret nuanced phrasing or struggle with complex anaphora resolution, leading to minor extraction errors.
- Dynamic Nature of Construction Projects: Construction sites are highly dynamic environments. Rapid changes in ground conditions, design modifications, or unforeseen events may not always be immediately documented in text or captured by existing sensors, potentially leading to discrepancies between the model's

predictions and real-world conditions if not continuously updated.

Addressing these limitations will be crucial for the further maturation and widespread adoption of integrated AI solutions in construction engineering.

CONCLUSION

This research has successfully developed, implemented, and rigorously validated a novel, integrated framework for automated seepage analysis in construction engineering. By seamlessly combining cutting-edge Natural Language Processing techniques for intelligent document processing with robust deep learning models for predictive analytics, the framework effectively bridges the critical gap between vast amounts of unstructured textual data and actionable quantitative insights.

The NLP component, utilizing custom-trained BERT models, demonstrated exceptional accuracy (overall average of 94.2% and up to 96.2% for hydraulic conductivity) in extracting precise geotechnical parameters, seepage indicators, and contextual event information from diverse and complex construction documents. This automated extraction capability profoundly transforms raw, qualitative data into structured, machine-readable features, mitigating the laborious and error-prone nature of traditional manual processes.

Furthermore, the hybrid CNN-LSTM-Attention deep learning model showcased superior performance in predicting critical seepage parameters, namely pore water pressure and flow rates. Achieving an RMSE of 0.142 m for pore water pressure prediction, our model significantly outperformed traditional methods (23.5% reduction in RMSE) and other state-of-the-art deep learning architectures. This enhanced accuracy is largely attributed to the multimodal data fusion, where the model effectively leverages both continuous numerical sensor data and the rich, contextual features intelligently extracted from textual reports.

A key contribution of this research is the quantifiable leap in processing efficiency. The automated system demonstrated an astonishing average speed improvement factor of 217.5x compared to manual methods, leading to substantial time and cost savings for construction projects. The robust validation using the extensive global SoilKsatDB database and real-world dam monitoring data further confirms the framework's reliability, generalizability, and practical applicability in diverse geotechnical contexts.

In essence, this integrated framework offers a powerful, intelligent tool for civil and geotechnical engineers, enabling a paradigm shift from reactive, manual data interpretation to a more efficient, accurate, and proactive approach to seepage management. By automating critical information extraction and providing reliable, context-aware predictions, the system contributes directly to earlier detection of potential issues, informed decision-making, enhanced safety protocols, and optimized maintenance strategies for critical civil infrastructure, thereby significantly advancing the digitalization and resilience of construction engineering practices globally.

Future Scope

Building upon the successful development and validation of this integrated framework, several promising avenues for future research and development emerge, aiming to further enhance its capabilities and broaden its applicability:

- Multilingual Document Processing and Global Applicability: Extend the framework's NLP capabilities to process documents in multiple languages. This would involve training or fine-tuning multilingual BERT models and building annotated corpora in languages beyond English, making the solution globally applicable for international construction projects.

- Integration with Internet of Things (IoT) Sensors for Real-time Data Acquisition: Develop seamless integration protocols with various IoT sensors deployed on construction sites and within infrastructure (e.g., smart piezometers, fiber-optic seepage detection systems [16]). This would enable truly real-time data ingestion, analysis, and prediction, enhancing the framework's responsiveness for immediate anomaly detection and rapid decision-making.

- Development of Mobile Applications for Field Use: Create user-friendly mobile applications that allow field engineers and inspectors to capture and input data directly, including text notes, photographs (for visual seepage detection, leveraging computer vision [9, 29]), and sensor readings. These apps could also provide real-time access to the framework's predictions and insights, empowering on-site personnel with actionable intelligence.

- Implementation of Blockchain Technology for Data Integrity and Trust: Explore the integration of blockchain technology to create an immutable and transparent record of all extracted parameters, sensor data, and prediction results. This would enhance data integrity, traceability, and trust among various project stakeholders (e.g., contractors, clients, regulatory bodies), particularly for critical safety-related data.

- Creation of Standardized APIs for Integration with Existing Construction Management Systems: Develop robust and well-documented Application Programming Interfaces (APIs) to facilitate seamless integration of the framework with existing Building Information Modeling (BIM) platforms, Enterprise Resource Planning (ERP) systems, and other construction project management software. This would ensure interoperability and embed the automated seepage analysis directly into current industry workflows.

- Expansion to Other Geotechnical Analysis Areas: Leverage the core capabilities of the framework (NLP-driven information extraction and deep learning prediction) to address other critical geotechnical engineering challenges. This includes:

  - Slope Stability Analysis: Extracting parameters like cohesion, angle of internal friction, and historical landslide data from reports to predict potential instabilities.

  - Foundation Design: Automating the extraction of bearing capacities, settlement characteristics, and pile driving records to inform and optimize the design of shallow and deep foundations.

  - Groundwater Flow and Contamination Modeling: Processing hydrogeological reports and monitoring well data to predict groundwater levels, assess contamination risks, and optimize dewatering operations effectively.

  - Tunneling and Underground Space Management: Beyond seepage, extracting rock mass classifications (e.g., RMR, Q-system), ground support details, and convergence measurements to predict tunnel stability and optimize excavation sequences.

- Explainable AI (XAI) Enhancements: Further research into advanced XAI techniques tailored for geotechnical models. This would focus on developing methods to provide more transparent and interpretable explanations for model predictions, allowing engineers to understand the underlying reasoning and build greater trust in AI-driven insights, particularly for complex scenarios where accountability is paramount.

- Uncertainty Quantification: Incorporate techniques for quantifying the uncertainty associated with predictions. Providing confidence intervals or probabilistic forecasts would give engineers a more complete picture of potential seepage scenarios, aiding in robust risk assessment.

- Federated Learning for Data Privacy: Explore federated learning approaches to train models on decentralized datasets across different project sites without directly sharing raw data, addressing data privacy and intellectual property concerns while still benefiting from distributed knowledge.

These future directions underscore the transformative potential of integrated AI in advancing construction engineering towards a more intelligent, efficient, and resilient future.

REFERENCES

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

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/

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

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/

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

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/

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/

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

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

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

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/

Tracy, F. (2007). SEEP2D: A 2D seepage analysis program. United States Army Corps of Engineers. Retrieved from https://en.wikipedia.org/wiki/SEEP2D