SSRG International Journal of Civil Engineering
ISSN: 2348-8352/ https://doi.org/10.14445/23488352/IJCE-V13I3P113

Volume 13 Issue 3, 189-203, March 2026
© 2026 Seventh Sense Research Group®

Original Article

Improving Embankment Settlement Predictions using Artificial Intelligence: Applications in Geotechnical Engineering

Youssef Elbalghiti[1], Mouna El Mkhalet[2], Nouzha Lamdouar[3]

[1,2,3]*Civil Engineering and Construction structure GCC laboratory, Mohammadia School of Engineers, Mohammed V University, Rabat, Morocco.*

[1]*Corresponding Author : y.elbalghiti@gmail.com*

Received: 10 December 2025 Revised: 12 January 2026 Accepted: 15 February 2026 Published: 23 March 2026

Abstract - *In road geotechnics, predicting embankment settlement is one of the most critical challenges, including management of completion in time, forecasting costs, and optimizing technical solutions at the design stage, especially when related to compressible soil types. Nonetheless, the nature of soil behaviour is such that heritage methods used for calculating and predicting soil settlement and deformation under load are often unreliable and tend to have only a limited capability with respect to the estimation of actual settlement. In this paper, a machine learning-based approach is introduced to enhance the prediction of deformation in structures on compressible soil. The proposed technique involves the application of an Artificial Neural Network (ANN) and density-based clustering and ordering (DBSCAN) to a real database collected during construction monitoring for high-speed train works in Morocco. DBSCAN was very effective for the organization and processing of databases, as well as a useful tool for identifying predominant spatial patterns of occupation and erroneous measurements caused by the heterogeneities and complexities found in compressible soils.*

Keywords - *Artificial Neural Networks, Cluster, Compressible soils, DBSCAN, Embankment, Prediction, Settlement.*

1. Introduction

Scientific and technological progress in the field of Artificial Intelligence (AI) technologies has been so fast that their bracing and strong impact on the engineering and construction industry is felt at practically every corner. Besides enhancing production, quality control, and predictive maintenance, AI can offer new ways to optimise the design and management of civil engineering projects.

One of the most attractive applications, AI can be a valuable tool to accurately predict complex soil deformations, particularly for large-scale projects where the management of deadlines and cost overruns plays a crucial role in successful completion.

The very same is in particular the case with complicated foundation works like slab foundations and fills on compressible soil. The predictability of these settlements and of virtually every type of deformation that is induced by the process is important as it leads to smarter construction planning, work scheduling, cost control, execution time management issues, and provision of solutions about reinforcement techniques needed for the stability-safety co-existence even during the intense phases in construction.

In this context, the use of predictive models with machine learning can allow very large amounts of data to be effectively incorporated and complex situations represented as much more immediate than in former times, when it was not easy to take into account using conventional methods.

This study thus is a step in the direction of increasing the precision for geotechnical diagnoses, considering that in such conditions, when soil support presents particular features which are hard to control, and when high overloads due to tall embankments occur. It also shows how such advanced techniques can be utilised to predict, expect settlement of structures with reflection in the future risk mitigation plan (preactive risk management), and optimize the design intervention process as well as civil engineering designs.

According to the procedures adopted for analyzing settlements of embankments on compressible soils, these methods can be classified into two classes: linear methods and nonlinear methods.

In linear methods, soil behaviour is assumed to be represented entirely by constant value elastic moduli, which reduces calculations significantly, simplifying the problem;

however, the accuracy of many predictions suffers, particularly in cases of significant deformation and varying load conditions. These methods are typically applicable only for soil displacements in moderate ranges, and the soil behavior can be modeled as a linear one.

In this paper, a statistical and machine learning-based model is proposed to predict the settlement of embankments on soft soils. While the use of machine learning has attracted much attention to geotechnical engineering applications as a result of the results obtained from previous studies, this work is distinctive for its focus on a particular challenge, which is data heterogeneity in compressible soils.

Most of the previous work is based on simple neural networks that directly generalize global regression models to raw data sets. Naturally, such methods fall short in many cases in representing real soils due to the influence of outlier noise and spatial variability on them.

A two-step hybrid approach is used in this work:

- Data cleaning and structuring by unsupervised learning (DBSCAN): This represents a novelty point. In contrast to most classical methods that assume homogeneous soil behavior, the DBSCAN algorithm is utilized in processing step 1 upstream to identify and separate specific soil behaviors (clusters) automatically.

Importantly, it enables us to successfully filter out ‘noise’ that would confound a standard model.

- Targeted predictive modelling: This initial segmentation allows the Artificial Neural Network (ANN) to learn from consistent and trusted data, leading to high-quality predictions.

In contrast with conventional, linear or nonlinear and generally simplified theory-based calculation methods, and traditional AI solutions that generalize all information equally, regard the specific features of such a non-homogeneous and complex system as is in-situ soils being present.

Indeed, nonlinear procedures consider the real geomechanical response of ground, which is often nonlinear, especially for saturated clay with heavy loading. They describe their response to loading using nonlinear behaviour laws. One example of such a law is the semi-logarithmic compressibility law, which is employed for the interpretation of oedometer tests. This connection considers that the soil deformation does not increase according to the amount of load applied, but follows a logarithmic curve, which is a function of the stress and deformation.

Oedometer models may be periodically applied to the study of soils loaded under heavy overburden, such as embankments, slab foundations, and others whose dimensions are much larger than the thickness of the deformable soil. These methods are based on the hypothesis that the geometric deformation of a soil in the vertical direction primarily results from uniformly applied loads, and thus make it possible to predict its thickening character by modelling this effect using laws of compression, in particular semi-logarithmic laws adapted for saturated and very compressible clay.

The total settlement St can be decomposed into a number of components, each corresponding to different phenomena [1]. The combined settlement is as follows :

$$St = Si + Sc + Sfl + Slat$$

With instantaneous settling according to the formula of Boussineq, Egorov, and Caquot Kerisel [2].

$$Si = \frac{\gamma h_r a^2}{E_u(a - a')}(r_h - \left(\frac{a}{a'}\right)^2 r'_h$$

a: Half width of the large base; a': Half width of the small base;

Eum: Average undrained Young’s modulus

rh and r’h are derived from Giroud’s chart for calculating settlement [3].

Consolidation settlement Sc is derived from oedometer settlement Soil [4]:

Normally consolidated soil

$$S_{oed} = C_C \frac{H_i}{1 - e_0} Log_{10} \frac{\sigma'_{vo} + \Delta\sigma_Z}{\sigma'_p}$$

Overconsolidated soil [5]:

$$\sigma'_{vo} + \Delta\sigma_Z < \sigma'_p$$

$$S_{oed} = C_S \frac{H_i}{1 - e_0} Log_{10} \frac{\sigma'_{vo} + \Delta\sigma_Z}{\sigma'_{v0}}$$

Cas

$$\sigma'_{vo} + \Delta\sigma_Z < \sigma'_p$$

$$S_{oed} = C_S \frac{H_i}{1 + e_0} Log_{10} \frac{\sigma'_p}{\sigma'_{v0}} + C_C(\frac{H_i}{1 + e_0}) Log_{10} \frac{\sigma'_{vo} + \Delta\sigma_Z}{\sigma'_p}$$

[6]

In the final one-dimensional settlement, correction is made for the effect of lateral deformation as per Skempton and Bjerum’s chart [7].

$$S_c = \mu S_{oed}$$

The finite element method simulations of elastic settlement could be carried out by means of specified software using a perfectly matched model for the elastic behaviour. Hence, the various calculation models can be used with numerical methods [8].

2. Context of Study

2.1. Novelty of the Research

The originality of this study is the development and application of a hybrid approach associating Density-Based Spatial Clustering of Applications with Noise (DBSCAN) with artificial neural networks to estimate settlement in embankment on compressible soil. Although Artificial Neural Networks (ANN) have been used extensively in geotechnical engineering to model nonlinear soil behavior and settlement mechanisms, most of the available tools are described by a single global prediction model, which does not directly account for data heterogeneity or noise commonly observed in field monitoring sets [9, 10]. Density-based clustering methods such as DBSCAN have successfully been used for finding the underlying data structure and detecting noisy observations in complex datasets; however, how they interact with predictive models in geotechnical practice is still sparse [11, 12].

To the authors' knowledge, no similar study has been conducted on a national scale in Morocco using a hybrid artificial intelligence method that combines unsupervised clustering and predictive modeling to analyze embankment settlement. The proposed approach is further justified through real-scale monitoring data on a high-speed railway project, which shows its applicability in complex geotechnical scenarios. Cluster-based data structuring and neural network prediction are introduced as a closure to state-of-the-art geotechnical practice, showing, in addition to a more robustly valid and interpretable and context-adapted solution, a comparison with classical analysis methods and the above-quoted AI-based research [13] et al. [14].

2.2. Research Gap

While significant strides to advance the prediction of embankment settlement have been made with recent developments in artificial intelligence, there are some key gaps present within the geotechnical communities. Such AI models are in the form of global predictors, and they implicitly disregard spatial scatteredness and nonlinear response that characterize compressible soils with depth [10, 13]. Furthermore, field measurement data are typically noisy and plagued by local anomalies as well as abnormal soil responders; however, these features are seldom explicitly treated in the prediction model context ([15]). Most previous works mainly concentrate on improving the prediction accuracy via some supervised learning approaches, by neglecting unsupervised learning algorithms, which are able to detect the inherent settlement patterns and filter out noise or non-representative data. However, the incorporation of density-based clustering methods such as DBSCAN in predictive models in geotechnical engineering has not been established to a great extent [11] through [12]. As a result, there is no robust and interpretable data-driven framework considering the heterogeneity and uncertainty in embankment settlement prediction.

As can be seen in Table 1, in the majority of previous works, global models are used to predict without any particular clustering or noise treatment.

Table 1. Comparative analysis of existing studies and the proposed method

Study	Method	Clustering	Noise handling	Field validation
Goh (1995) [10]	ANN	No	No	Laboratory / numerical data
Das & Basudhar (2006) [13]	ANN	No	No	Laboratory data
Moayedi et al. (2019) [16]	SVM	K-means	No	Limited field data
Shahin et al. (2020) [17]	ANN	No	No	Case-specific field data
This study	DBSCAN + ANN	Yes (density-based)	Yes (explicit noise detection)	Real-scale monitoring data

In comparison, the proposed approach combines density-based clustering with neural network–based prediction and is compared with real-scale embankment monitoring data. The structure based on clustering increases not only the model reliability and interpretability, but is also more capable of representing both the natural physical variability and heterogeneous behaviors of compressible soil under embankment loading.

3. Related Works

3.1. State of the Art of Artificial Neural Networks

An Artificial Neural Network (ANN) is a computational simulation of the way in which the human brain processes information. It has a lot of applications in artificial intelligence and machine learning, such as, but not limited to: Classification, Regression, Voice or face recognition, Automatic translation, and so on [9].

A neural network consists of multiple layers of neurons that are interconnected.

According to Nielsen, the structure of the Neural Network is presented [18].

- Main Layers
- Input layer: raw data.

- Hidden layers: transform the data through learned patterns.
- Output layer: Final results after processing.

3.1.1. How a Neuron Works

Each neuron takes inputs, multiplies them by their respective weights, adds a bias, and then applies an activation function to the result.

3.1.2. Neuron Computation

$$z = \sum_{i=1}^{n} w_i\, x_i + b$$

Activation :
$a = \varphi(z)$
x_i: input values
w_i: weights
b: bias
φ: activation function

3.1.3. How the Network Learns

The goal is to adjust the weights so that the network makes accurate predictions.

3.1.4. Learning Steps

1. Forward propagation:
   - The input passes through the network to generate an output.
2. Loss computation:
   - The output obtained is compared to the expected result.
   - A loss function is used to evaluate the error
3. Backpropagation:
   - The error is propagated backward through the network.
   - The error is sent back through the network.
   - The outputs are adjusted using the gradient descent method in order to reduce the error

A simple network with:

- 3 input features,
- 1 hidden layer with 4 neurons,
- 1 output (e.g., binary classification)

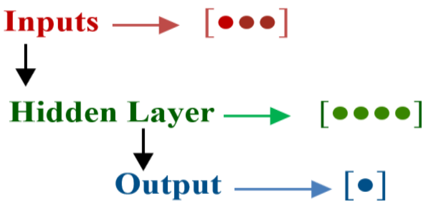

Fig. 1 Example of an Artificial Neural Network

In this part, the Backpropagation algorithm used in artificial neural networks was examined. The employed algorithm minimizes the gradient descent part. It scales the weights of each network individually according to how much it is contributing towards the error in the output.

Therefore, the neural network has L layers as follows:

- The first layer is equal to the 1st layer
- The final layer (layer L) is the output layer.

For each layer l, the definitions are as follows:
$W^{[l]}$: weight matrix
$b^{[l]}$: bias vector
$a^{[l]}$: activation
$z^{[l]} = W^{[l]}a^{[l-1]} + b^{[l]}$:
linear combination before activation

- Step 1: Forward Propagation
  For each training example x(i), compute:

  For each layer l=1 to L:

$$z^{[l]} = W^{[l]}a^{[l-1]} + b^{[l]}$$
$$a^{[l]} = \phi^{[l]}(z^{[l]})$$

Where ϕ is the activation function.

- Step 2: Compute the Loss

Compute the loss between the predicted output $\hat{y}$ and true output y

$$J = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\,(\hat{y}^{(i)}, y^{(i)})$$

For binary classification, the following applies :

$$\mathcal{L} = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$$ [19]

- Step 3: Backward Propagation

Goal: compute gradients of the cost function w.r.t. the weights and biases.

Compute error at output layer:

$$\delta^{[L]} = a^{[L]} - y$$

(assuming sigmoid + cross-entropy)

Backpropagate the error to previous layers: For l=L−1 down to 1:

$$\delta^{[l]} = \left(W^{[l+1]}\right)^{\top}\delta^{[l+1]}\circ\varphi'(z^{[l]})$$

Where: δ is the element-wise product (Hadamard product) and ϕ′ is the derivative of the activation function

- Step 4: Compute Gradients
  For each layer l: Gradient of the cost function with respect to W[l]:

$$\frac{\partial J}{\partial W^{[l]}} = \frac{1}{m}\delta^{[l]}\left(a^{[l-1]}\right)^{\top}$$

Gradient of the cost function with respect to b[l] :

$$\frac{\partial J}{\partial b^{[l]}} = \frac{1}{m}\sum_{i=1}^{m} \delta^{[l](i)}$$

- Step 5: Update Parameters (Gradient Descent)

Update the weights and biases

$$W^{[l]} := W^{[l]} - \alpha \frac{\partial J}{\partial W^{[l]}}$$

Update of b[l] :

$$b^{[l]} := b^{[l]} - \alpha \frac{\partial J}{\partial b^{[l]}}$$

where α is the learning rate.

3.2. State of the Art of DBSCAN

DBSCAN is a density-based clustering algorithm that creates clusters of points in high-density regions and classifies as outliers the ones that are alone in low-density regions. DBSCAN, instead of manually choosing the number of clusters as in k-means Clustering, finds clusters and outliers by comparing the local density.

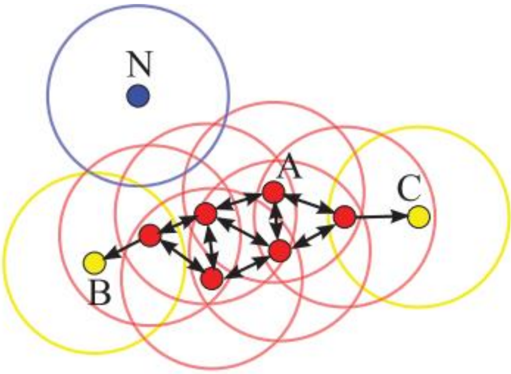

Fig. 2 Visualization of the DBSCAN clustering model [12]

3.2.1. Definitions

ε (epsilon): Radius to search neighbors.

MinPts: minimum number of points required to form a dense area

Core Point: a point with at least MinPts neighbouring points within the interval ε.

Directly Density-Reachable: A point q is considered directly accessible in terms of density from a point p if q is located at a distance less than ε from p and if p is a central point. [20]

Neighborhood of a given point,

$$\mathcal{N}_{\varepsilon}(\mathrm{p}) = \{\mathrm{q} \in \mathrm{D} \mid \mathrm{distance}(\mathrm{p},\mathrm{q}) \leq \varepsilon\}$$

Where D is the dataset.

Core-point condition

$$|\mathcal{N}_{\varepsilon}(p)| \geq \mathrm{MinPts}$$

Directly density-reachable

Point q is directly density-reachable from point p if:

$$q \in \mathcal{N}_{\varepsilon}(p) \quad \text{and} \quad |\mathcal{N}_{\varepsilon}(p)| \geq \mathrm{MinPts}$$

3.2.2. Density-Reachable

A point q is accessible in density from point p if the following condition is satisfied. There must be a chain of points p1, p2, …, pn with: p_1 = p, p_n = q, and for each pi+1, it must be directly accessible in density from pi.

3.2.3. Density-Connected

Two points p and q are said to be densely connected if there exists a point O from which both p and q are densely accessible [21].

3.2.4. Algorithms Steps

For each point p D, retrieve its ε-neighborhood Nε (p)

If $|\mathcal{N}_{\varepsilon}(p)| < \mathrm{MinPts}$, mark p as noise (or border point).

Else, create a new cluster and iteratively add all accessible points in terms of density from p [22].

3.3. Clustering Process

For each point p belonging to the data:

- Find all points within ε (the ε-neighborhood).
- If the ε-neighborhood contains at least MinPts, mark p as a core point and start a new cluster.
  1. Iteratively integrate all accessible points in terms of density from the central points to the cluster [22].
  2. Points that are not accessible or connected from any central point shall be considered noise.

3.4. Other Methods

RF aggregate predictions of decision trees for avoiding overfitting and enhancing generalization, and are very effective for capturing complex nonlinear associations. Zhang et al. have shown the strength and accuracy of RF for evaluating pile drivability, showing the adaptability to high-dimensional data and accurate predictions in several types of geotechnical conditions.

On the contrary, Gradient Boosting Machines (GBM), which include popular implementations like XGBoost, are working in a sequential manner to construct decision trees hint by hint, with each consecutive tree trying to correct the errors of its predecessors. It is this iteration that allows GBMs to reach the extraordinary levels of accuracy, as demonstrated by Zhou et al., who were able to apply XGBoost models to predict ground settlement due to tunneling. These two types of models are widely acknowledged for their capability to produce very accurate and stable predictions, even when dealing with noisy or incomplete data.

Apart from ensemble techniques, hybrid neuro-fuzzy systems provide an alternative field of interest that brings together the strengths of ANN and FLS. These architectures benefit from the capacity of ANNs to acquire complex patterns and nonlinear relationship from experiences, as well as from the ability of FL to model human reasoning and to handle inherent uncertainty and vagueness in geotechnical systems. Kurnaz and Kaya (2014) employed these hybrids with success in settlement prediction for shallow foundations on sandy soils.

Their work emphasises the practical relevance of these methods by showing that they not only yield good accuracy but also achievable interpretability of results, which is a significant advantage in engineering fields such as CFD, where insight into the underlying mechanisms through good interpretability is required. Nevertheless, even such advanced models powering these performing solutions are highly dependent on the quality and structure of input data. In geotechnical problems, compressible soil heterogeneity and outliers (noise) may undermine the efficiency of predictive models.

Thus, the works described primarily intend to apply these models directly to data that is typically considered pre-cleaned for this purpose. Such works complement our proposed methods, which deal with the challenge of handling heterogeneity and noise, generating a clustering risky population prior to building prediction models by DBSCAN clustering, followed by artificial neural networks for settlement prediction.

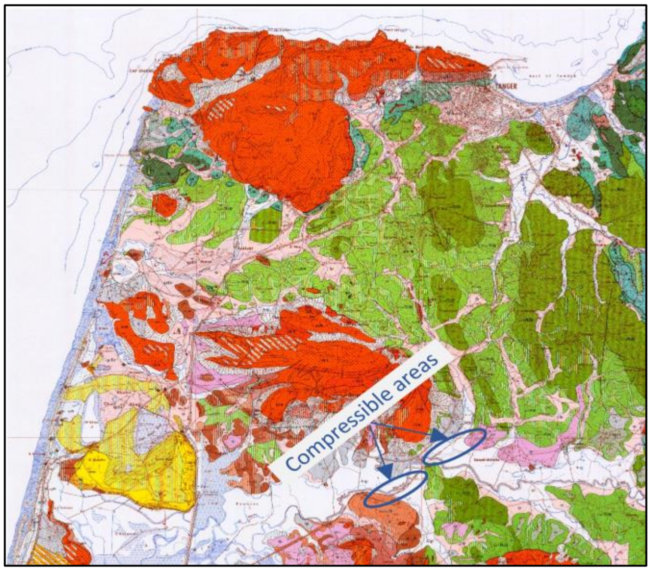

Fig. 3 Extract from the Tangier-Al Manzla geological map

4. Materials and Methods

4.1. Geotechnical Investigation

This settlement study focused on a number of embankment profiles of the Tangier-Kenitra high-speed rail line (LGV) in Morocco.

This line is located in Northern Morocco, where geomorphology is complex and one of the most affected by the problems related to deterioration of road and/or rail infrastructures caused by geotechnical matters [26].

The profiles studied intersected compressible valleys where the ground has a characteristic of high compressibility and a relatively high organic content.

Fig. 4 Compressible zone, Mharhar Valley

They are, in fact, soft soils despite the height of the embankment and therefore suffer from a high settlement:

- Oued Mharhar embankment;
- R2288 embankment (PR 228+460 to 229+420): Rharifa Valley;
- R3053 embankment (305+020 to 306+530): Al Hamar embankment

The following data were used for this study:

- Coring: the taking of core samples of soil for testing in a laboratory.
- pressuremeter tests: to find the necessary limit and creep pressure, as well as other important parameters, such as pressuremeter modules, which are crucial for foundation design.
- CPT tests allowing for the in-situ characterisation of compressible soils;

Furthermore, the collected samples were subjected to laboratory testing, such as:

- Index tests: including grain size analysis, Atterberg limits, and methylene blue value for profile characterisation.

Soil mechanics tests are essentially re-compressibility tests with an oedometer, giving all the mechanical parameters of formations under load.

Table 2. Investigation program

Area	Situations	Programs	Laboratory tests
R2023 Mharhar river	PR201+760 to PR202+260	6 borehole logs, 22 CPT penetrometers 1 pressuremeter 4 Sissometers	8 identifications tests 6 oedometer tests 1 organic matter content
	PR202+260 to PR202+600		
	PR202+600 to PR202+851		
R2039 Mhahar Valley Studies	PR*203+520 to PR203+660	2 borehole logs 10 CPT 2 pressuremeter 2 Sissometers	12 identifications tests 15 oedometer 12 organic matter content
	PR203+690 to PR204+040		
	PR204+040 to PR204+190		
R2063 Seguedla	PR206+380 to PR206+560	3 borehole logs 5 CPT penetrometers 9 pressuremeter 1 Sissometers	5 identifications tests 2 oedometer tests 2 organic matter content
R2083	Du PR 208+080 au PR 208+310	2 borehole logs, 5 CPT penetrometers 4 pressuremeter 1 Sissometers	15identifications tests , 6 oedometer tests 2 organic matter content
	From PR208+350 to PR208+450		

**PR: Reference Point*

This dual screening approach with field and laboratory data led to a thorough understanding of the soil properties. On the basis of these findings, geotechnical computation models were developed in order to make an evaluation of embankment stability.

For the embankment instrumentation results, the field-measured values were taken from those recorded on profiles established with settlement balls for each zone.

4.2. Methodology

Three methods were employed in this research for the prediction of embankment settlement level on railway: classical computation, artificial intelligence modelling, and actual deformation measured in an experiment. These methods were used on a number of embankment sites for which instrument measurements were obtained during the consolidation process.

- The conventional approach: elastic methods, which employ the soil reaction modules obtained from the interpretation of laboratory and in situ tests
- Settlement prediction with artificial intelligence; this methodology adopts the DBSCAN algorithm for the partitioning of data

Experimental determination of the real deformation: a comparison will be made with the calculated results and traditional calculation methods, based on real settlement values measured in each one of the profiles.

These cross-validation and analysis techniques have been applied and used at several embankment sites in various locations with measured instrumentation data throughout the course of an embankment consolidation stage. A total of 86 profiles, in all compressible sections detected over the route, have been considered.

Conventional models, Traditional methods for predicting settlement are conventional geotechnical equations and predictions based on the geometry of embankments, applied load, and soil properties. Results of elastostatic measurements were tested for deformation behaviour described in terms of displacement.

Concurrently, an AI model for better predictions was constructed. Employing geotechnical parameters (density, cohesion), geometric properties, and loadings as input parameters, a NN was developed.

It should be pointed out that to the experimental database there corresponds the real deformation measurements performed by means of compaction balls settled in the embankment body. These measurements yielded the settlement value in practical conditions and were used as a benchmark for the accuracy of the proposed method.

In effect, instruments were installed at least under each embankment by means of extensometers, which were monitored weekly until settlement rates became stabilized.

Surface extensometers are geotechnical devices used to monitor the vertical displacement of a soil (settlement) prior to, during, and after the construction of an embankment.

A minimum of three instrumented profiles were spread throughout all of the compressible areas. The estimated loading heights are the maximum loads for each profile.

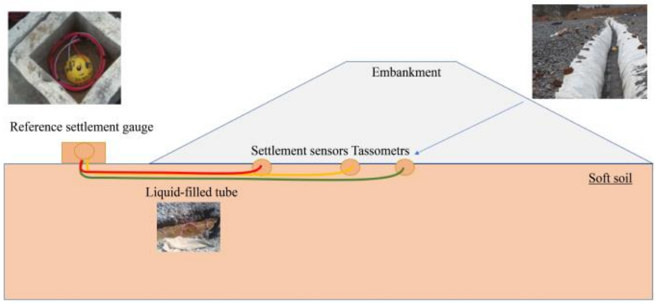

Fig. 5 Instrumentation profile of a railway embankment

The measurement scheme is an underground, differential hydraulic levelling system, and the level difference is measured between a stable reference ball outside of the affected area of embankment and a series of measuring balls buried in the ground inside the embankment.

Measurements are made at predetermined intervals (a function of the project progress and project-specific requirements) until settlement has stabilized or the ground has consolidated.

Finally, the database was split into training and validation data sets for calibration and testing of the AI-based model. The backpropagation method was employed for the training of the algorithm, while its performance was assessed by means of statistical measures, such as mean square error (MSE). Mutual comparisons in terms of graphs and statistics provided a trustworthy way to test the consistency of each method with respect to experimental data.

This integrated procedure makes it possible to critically appraise the adopted prediction method and identify the most appropriate model for predicting settlement in case of embankments founded on compressible soils.

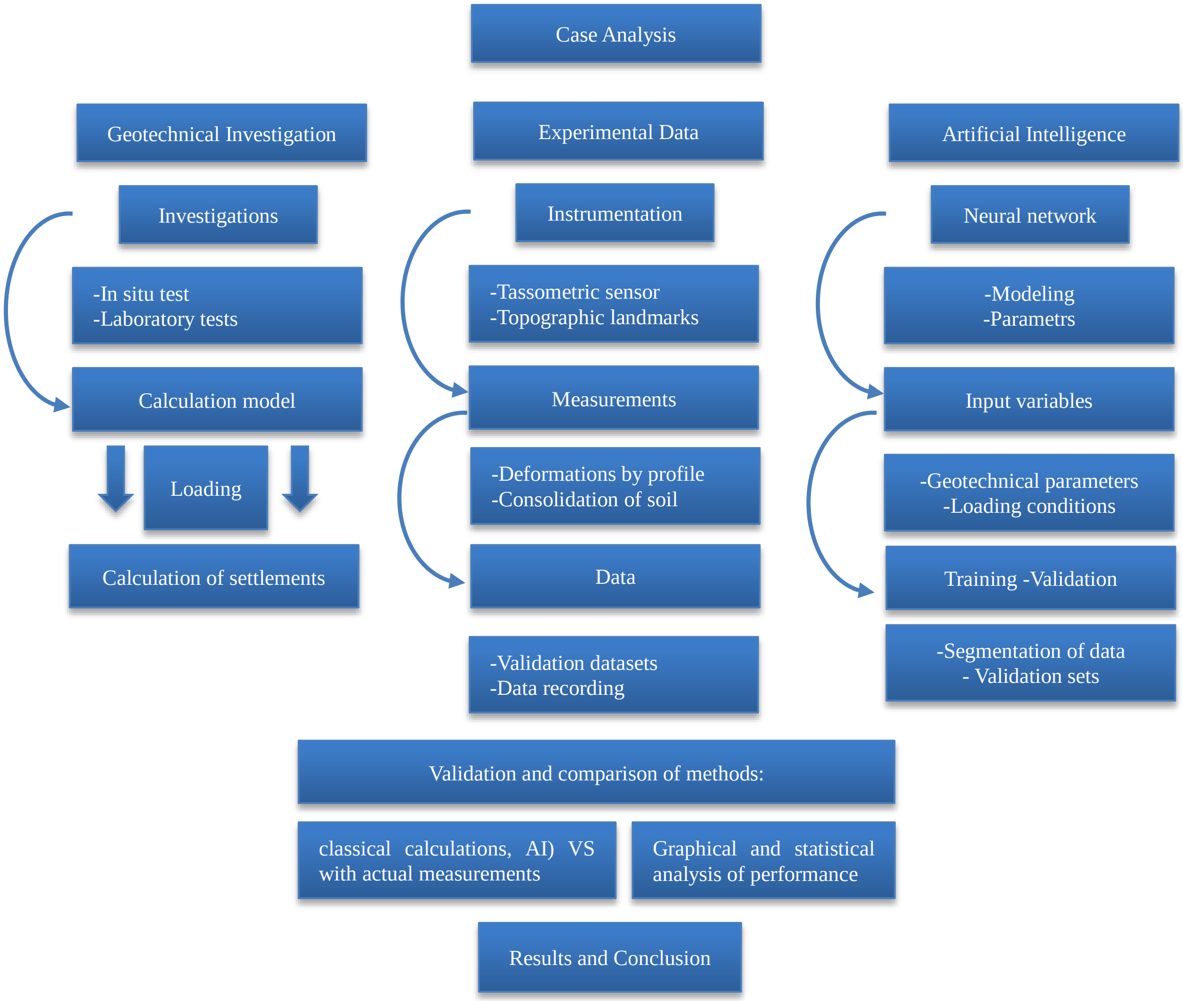

Fig. 6 Methodology flowchart

4.3. Results of Investigations

Overall, analysis of the results of investigations carried out along the geotechnical profile of the compressible areas under study reveals the following lithological sequence:

- Surface brown clays (1)
- Vasardes clays (2)
- Sandy silts (3)
- Transition sandy silts (5)
- Silty-sandy clays (6)
- Alluvial deposits with a sandy matrix (7)
- Pelitic substratum, weathered at the top into pelitic clay (8)

The table below summarizes the results of in situ and laboratory tests:

Table 3. Geotechnical characteristics of the formations

Formation	qc MPa	Pl* MPa	E MPa	γ kg/m^3	IP (%)	<80µm (%)
1-Brownish clay	1,5 à 1,9	-	-	1942	28-31	98
2-Clayey silt	0,5 à 1,4	-	-	1838	28	-
3-Alluvial lens	6 à 8	-	-	-	-	-
4-Sandy silt	0,4 à 1,5	-	-	-		
5-Sandy clay	1 à 1,4	-	-	-	-	
6-Sandy silty clay	1,2 à 2,1	-	-	2050	28	98
7-Alluvium with sandy matrix	3	1,8	19	-	-	-
8-Pelitic clays	>4	,95-1,17	10,9-11,2	1955	29	
9-Lower alluvium	3,1	-	-			
10-Blackish silt	0,8	-	-	1815	26	86
11-Brownish silty clay	1,0 à 1,2	-	-	1924	21	98
12-Greyish silt	0,9 à 1,0	-	-	1946	24	88
13-Grey pelite	> 8	,35-4,2	19,4-542	-	-	-

5. Results

5.1. Application of DBSCAN

5.1.1. Processing Workflow

The figure below illustrates the processing workflow used for the application of the DBSCAN algorithm; Table 3 presents the Data of this study, and Figure 8 shows clusters.

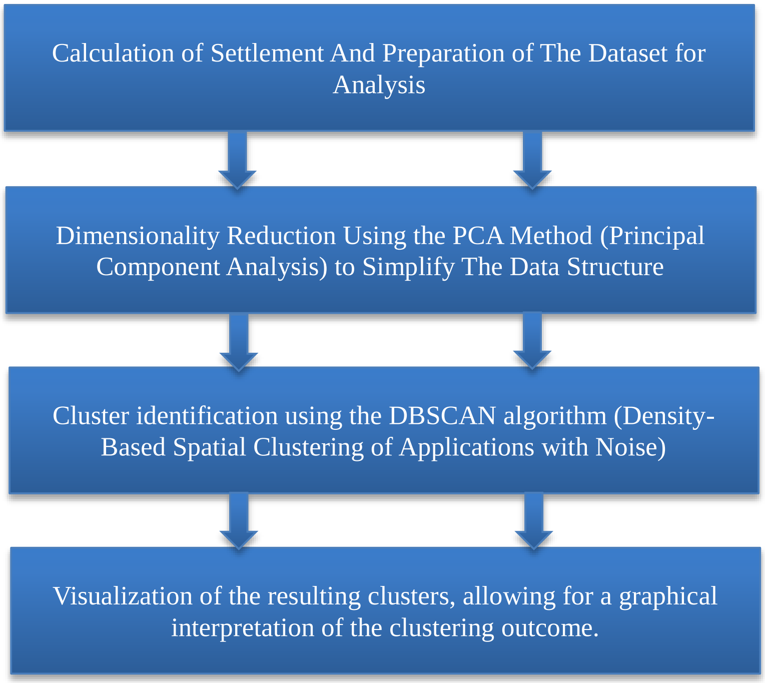
Calculation of Settlement And Preparation of The Dataset for Analysis
Dimensionality Reduction Using the PCA Method (Principal Component Analysis) to Simplify The Data Structure
Cluster identification using the DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise)
Visualization of the resulting clusters, allowing for a graphical interpretation of the clustering outcome.

Fig. 7 Processing workflow

Table 4. Data from this study

Samples	Th (m)	γ (kN/m^3)	σ'p (Kpa)	e0	Cc	Cs
Ech1	6,0	19,0	258	1,028	0,334	0,047
Ech2	6,0	19,0	258	1,028	0,334	0,047
Ech3	6,0	19,0	258	1,028	0,334	0,047
Ech4	6,0	19,0	258	1,028	0,334	0,047
Ech5	6,0	19,0	258	1,028	0,334	0,047
Ech6	8,0	20,0	235	1,11	0,375	0,052
Ech7	8,0	20,0	235	1,11	0,375	0,052
Ech8	8,0	20,0	235	1,11	0,375	0,052
Ech9	8,0	20,0	235	1,11	0,375	0,052
Ech10	1,5	18,0	253	0,703	0,243	0,087
Ech11	6,0	18,0	298	0,723	0,255	0,075
Ech12	6,0	18,0	298	0,723	0,255	0,075
Ech13	6,8	18,0	147	1,051	0,318	0,085
Ech14	8,0	18,0	174	0,572	0,179	0,055
Ech15	8,0	18,0	174	0,572	0,179	0,055
Ech16	8,0	18,0	174	0,572	0,179	0,055
Ech17	8,0	18,0	174	0,572	0,179	0,055
Ech18	4,0	19,6	253	1,028	0,334	0,047
Ech19	4,0	19,6	253	1,028	0,334	0,047
Ech20	4,0	19,6	253	1,028	0,334	0,047
Ech21	4,0	19,6	253	1,028	0,334	0,047
Ech22	5,6	18,3	312	0,94	0,231	0,072
Ech23	5,6	18,3	312	0,94	0,231	0,072
Ech24	5,6	18,3	312	0,94	0,231	0,072
Ech25	5,6	18,3	312	0,94	0,231	0,072
Ech26	5,7	18,8	254	0,918	0,316	0,112
Ech27	5,7	18,8	254	0,918	0,316	0,112
Ech28	3,8	19,8	265	0,741	0,26	0,088
Ech29	3,8	19,8	265	0,741	0,26	0,088
Ech30	3,2	18,0	222	0,794	0,328	0,089
Ech31	4,0	18,0	255	0,763	0,296	0,083
Ech32	4,0	20,0	300	0,99	0,305	0,097
Ech33	5,5	20,0	300	0,99	0,305	0,097
Ech34	8,5	18,1	136	0,73	0,24	0,07
Ech35	8,5	18,1	136	0,73	0,24	0,07

-Th: Thickness (m) -initial void ratio:e_0
-Unit weight (kN/m3) -Preconsolidation pressure:σ'p
-Compression index: Cc -Swelling Index: Cs

5.1.2. Results of DBSCAN Clustering Analysis on Settlement Data

Clustering techniques were used to analyse embankment settlement data after an initial treatment of dimensionality reduction/feature scaling, in order to find underlying structures and separate behavior patterns.

The DBSCAN algorithm was implemented with the following settings:

- Neighborhood radius (varepsilon) =0.50
- MinPts =5 as the minimum no of points.

These settings were chosen to improve the demarcation of dense and effective noise separation.

Therefore, the clustering results are shown in Figure 8.
A subset of the features is projected and scaled on the first two reduced components.

Apart from a subset of points labelled as noise, there are five (5).

5.1.3. Characteristics of Identified Clusters

- Clusters 1 (red), 2 (yellow), and Cluster 3 (green): those are the three clusters that dominate the picture.

They are spread quite broadly in the chosen attribute space. The dense cores of these groups are represented by the central points (circles), whereas the peripheral ones, represented by squares, are spread-out configurations which cannot be defined as high-density areas.

As an instance, cluster 3(green) spreads significantly on 'feature 1' while having high density.

- Clusters 4 (blue) and 5 (purple): based on the structure of these clusters, it appears that they have a higher density. They are in two different places.

For instance, cluster 4 is positioned in the area where 'characteristic 1' values are high and 'characteristic 2' values are low, and cluster 5 is placed in the region of low 'characteristics 1 & characteristics 2'.

So this compactness and the quodistant position mean that there are special conditions that induce more homogeneous settlement behavior in this data set.

It is also mentioned that many points were labeled as drawn from noise (black crosses in Figure 8).

These points are not uniformly distributed and may result from the isolation of these sources and/or from their location in regions with a very low source density.

DBSCAN does an excellent job, as not one of these is found in a class (hence, keeping consistency among the clusters found).

5.2. Application of Neural Network

5.2.1. Architecture

- Network type: patternnet (neural network for classification)
- Number of hidden layers: 2
- Neurons per hidden layer: 10 neurons in layer 1, 10 neurons in layer 2
- Activation functions:
    - Layer 1: ‘tansig’ (hyperbolic tangent)
    - Layer 2: ‘tansig.’
- Output layer: implicit softmax (default in patternnet)

- Performance Function: The error function used is cross-entropy, which is suitable for multi-class classification
- Data Splitting
    - 70% for training
    - 15% for validation
    - 15% for testing

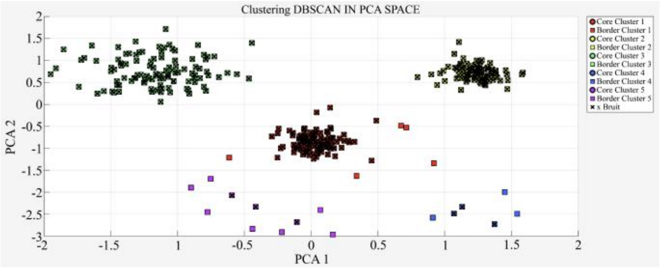

Fig. 8 Clustering results

5.2.2. Accuracy (Exactitude)

Accuracy is the ratio between the number of correct predictions and the total number of predictions made [27].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

5.2.3. Precision

Accuracy reflects the percentage of true positive predictions among all cases classified as positive by the model [24].

$$\text{Precision} = \frac{TP}{TP + FP}$$

5.2.4. Recall

This is the number of true positives, i.e., samples correctly identified as belonging to the positive class among all samples that actually belong to that class.

$$\text{Recall} = \frac{TP}{TP + FN}$$

5.2.5. F1-Score

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score represents the harmonic mean of precision and recall, making it an important indicator for achieving a balance between these two measures in classification tasks.

5.2.6. Confusion Matrix

The confusion matrix is a 2×2 table used for binary classification that displays the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). True Positive (TP) refers to a sample belonging to the positive class being classified correctly.

- True Negative (TN): a sample from the negative class that is correctly classified.
- False Positive (FP): a sample from the negative class that is incorrectly classified as belonging to the positive class.
- False Negative (FN): a sample from the positive class that is incorrectly classified as belonging to the negative class

5.3. Other Default Key Parameters

- Training algorithm: ‘trainlm’ (Levenberg-Marquardt), fast for small to medium-sized networks (default in patternnet)
- Maximum number of epochs: typically, 1000
- Stopping criterion: convergence is based on validation performance or the maximum number of cycles reached
- Weight initialization: random initialization is used at the start

6. Discussion

6.1. Results

The identification of five distinct groups and a ‘noise’ category makes it possible to highlight and differentiate the settlement behaviour of embankments for each group.

This confirms that there is no single, homogeneous behaviour, but rather a set of different profiles controlled by several combinations of geotechnical parameters depending on the loading conditions.

Thus, each group potentially represents a particular and unique settlement ‘behaviour’.

In this context, the use of the DBSCAN algorithm is particularly effective.

Its ability to define classes is a considerable asset. This is particularly true for class 3, which does not have a spherical shape that is easily identifiable by conventional methods.

In addition, DBSCAN excels at explicitly identifying noise points that can distort interpretation.

These noise points normally correspond to atypical measurements, experimental errors, or exceptional conditions at the site that do not correspond to any of the dominant settlement models. Isolating them prevents them from biasing the characterisation of the main clusters.

These results contribute directly to the development of predictive models. Thus, to predict actual settlement, instead of attempting to construct a single, comprehensive regression model, a more nuanced approach is now very beneficial for better controlling complex soil settlement. It is therefore relevant to explore:

- For each cluster: develop and train separate regression models. A specifically calibrated model will be more accurate in predicting settlements corresponding to that same profile.
- Integration of cluster membership: the identifier of each cluster could be integrated as an additional categorical feature in the predictive model to adapt to the different behaviours identified.
- In-depth analysis of atypical data: detailed examination of points classified as noise. By verifying the parameters that are less characteristic of the original input of these points, with the aim of determining the causes of these non-standard behaviours. Are these data anomalies or unique geotechnical behaviour?

In this study, DBSCAN was applied to classify the soil layers, resulting in four main clusters and one noise group:

- Cluster 1 → silty sands
- Cluster 2 → silts, silty sands, and sandy clays
- Cluster 3 → clay and silty clay
- Cluster 4 → Clay and silt
- Cluster -5 → Noise. These are outliers identified as noise by DBSCAN.

6.2. Cluster 1: Silty Sands

Samples in this group have moderate pre-consolidation pressure and a compression index (cc) of less than 0.02. These parameters indicate that the soils are moderately compressible. They are therefore compact and relatively stable soils, with calculated settlements varying according to thickness.

6.3. Cluster 2: Silts, Silty Sands, and Sandy Clays

This group has a higher density with a compression index between 0.2 and 0.3; these are therefore relatively compressible soils that are sensitive to water. In this group, settlement is quite significant under embankment loads.

6.4. Cluster 3: Clay and Silty Clay

These samples are characterised by high pre-consolidation pressure, low density, and a compression index greater than 0.3.

These are, therefore, soils with significant potential for settlement under embankment loads.

6.5. Group 4: Clay and Silt

These soils have a lower density, are softer, and constitute the most critical layers in terms of settlement.

They are very sensitive to water and prone to significant settlement.

6.6. Group 5: Noise

DBSCAN excluded this group because these samples do not form a homogeneous group.

Based on these results, it appears that training neural networks made it possible to define cluster classes and obtain excellent results with an overall accuracy of 60.87%.

It should also be noted that the models also demonstrated solid performance according to the mean square error (MSE) measure (Figure 9).

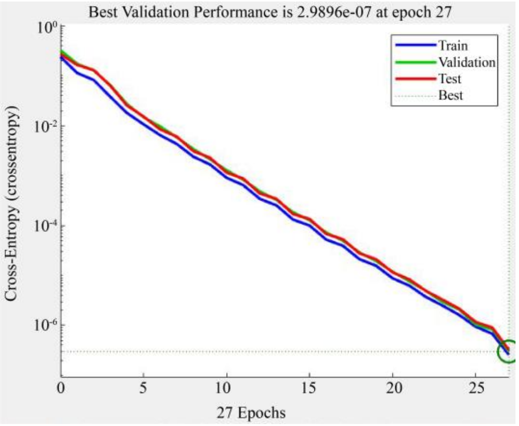

Fig. 2 Best Validation Performance

6.7. Validation

The graph below illustrates the differences between measured and actual settlements before treatment (Figure 10).

The results are as follows:

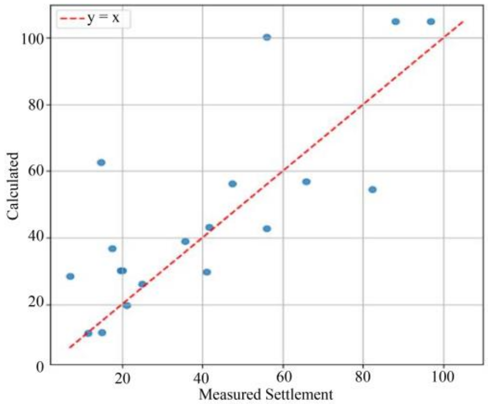

Fig. 10 Calculated settlement vs Measured settlement

- The values of MAE (Mean Absolute Error) = 17.08 cm and RMSE (Root Mean Square Error) = 21.52 cm. These statistics proved that an average large error and some very poor descriptions (RMSE > MAE).
- Large scatter: the model is not very well-calibrated and has large errors.
- This strong scatter is especially visible for average to highest settlements (> 50 cm), where some, albeit few, very significant overestimations can be observed.

If training the model, the following results are for reference:

After training the model, the results obtained are significantly improved (Figure 11):

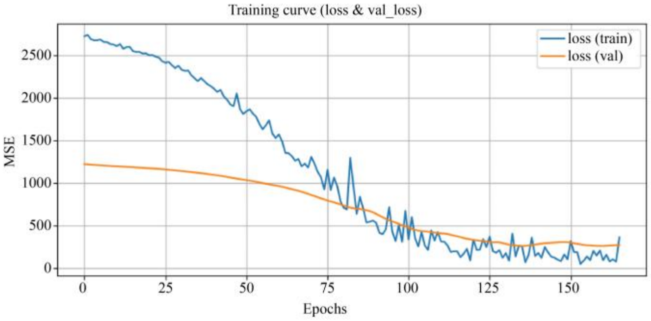

Fig. 11 Training results

- MSE: 12.593
- MAE: 3.351
- BIAS: 3.351 cm

These values demonstrate improved prediction with RMSE coefficients of 3.55 cm instead of 21 cm.

As areas for improvement, we can reduce the complexity of the networks or test models such as RandomForest and XGBoost, given the amount of data available.

Last, but not least, this paper offers a notable methodological contribution by designing a hybrid model that combines narrow bandwidth Density-Based Clustering (DBSCAN) with Artificial Neural Networks (ANNs) for embankment settlement prediction. This is an identical first attempt of such a novel approach in the domain of geotechnical civil engineering in Morocco. Through the use of DBSCAN, data heterogeneity is handled, and one can find homogenous data clusters from a set of measurements, while also being able to handle outliers and measurement noise. This feature contributes to the reliability and strength of the predictions.

Besides, the hybrid DBSCAN-ANN can achieve stronger prediction robustness and interpretability. It is able to account for the nonlinear behavior and spatial variability of compressible soils in a more interpretable form compared to traditional global ANN models. This is of utmost significance also in intricate geotechnical problems. The validity of the method is also supported by validation using field data from a high-speed railway project at real scale. The validation underlines the applicability of the framework to difficult geotechnical situations and demystifies its potential use in practice.

Furthermore, the proposed framework is an improvement over traditional algorithms. In contrast to the general analytical methods or global ANN models, it deals with data noise and local variability in a way that is context-specific and data-constrained, suitably adjusted for the particularities of each study area.

The study is not without limitations despite its strength. One limitation is the reliance on the parameters of DBSCAN, namely epsilon (ε) and minPts. The clustering quality is highly dependent on these parameters, and one therefore has to calibrate them empirically for best results.

Another problem of the hybrid DBSCAN-ANN method is that it has higher time complexity compared with ANN models, especially when processing a large-scale dataset. This higher computational cost may be prohibitive in both time and resources.

In addition, the verification of the method was limited only to one high-speed railway construction work, and hence it might be questionable for practical applications on different embankments or other soil profiles without relevant corrections. Finally, it should be noted that the validity of the neural network depends on data-driven restrictions and is closely related to the amount and quality of available monitoring information, which may affect predictions as a whole.

7. Conclusion

The work presented in this paper shows the use of AI techniques for improving embankment settlement estimation, especially when involving compressible soils with nonlinear behavior. Through an integration of conventional geotechnical methods and machine learning technologies, this work made a two-fold approach possible through the application clustering by DBSCAN to describe soil performance and neural networks to predict settlement.

DBSCAN successfully detected diversified, unique settlement patterns and recognized 5 clusters and ‘outlier’ points.

These results corroborate that the response of soils under an embankment is not consistent, but rather depends on different permutations of geotechnical factors and loading scenarios. Furthermore, this capacity to judge non-usual behaviours and do not push them into existing groups is a significant benefit towards the reliability of geotechnical results. It can be observed that the promise of development with improved performance, with an indexing of 60.87% overall accuracy for neural networks trained using cluster classification. While these findings imply relatively strong settlement type classification, they also present potential for more complex regression models, adjusted to each cluster separately.

In sum, relatively complex geotechnical scenarios are better understood by this combined approach than the conventional method with a single global model in comparison perspective. It is a significant step forward in the development of more precise predictive methods and in the design/control optimisation for embankment works built on compressible soil.

In the future, the development of cluster-specific regression models and considering in detail 'noise' points will extend these areas, displaying heterogeneous behaviour. In summary, in this study, a novel hybrid model is first introduced, which effectively combines unsupervised density-based clustering (DBSCAN) and ANN for the purpose of forecasting embankment settlement in geotechnical civil engineering applications in Morocco. By proposing a novel methodology, the work gets around data heterogeneity issues and is robust in terms of predictions, as well as being more interpretable than classical approaches.

The validation of this model with full-scale field data from a high-speed railway project demonstrates its practical utility and applicability in challenging geotechnical conditions. Although the promising findings are indicative, limitations of parameter dependence, computational cost, and specificity using a limited validation dataset are also recognized in their study.

Future work is very promising, and it can further improve the applicability of this framework based on other trucks, embankment conditions, and soil types, taking advantage of the integration of advanced AI technologies and automatic parameter optimization methods. Through these lines of investigation, this study can significantly benefit the field in general and serve as introductory work to the development of safer and more efficient solutions for embankment settlement prediction, safety, and performance improvement in Civil Engineering projects.

References

[1] Roger Frank, "Shallow Foundations," *Engineering Techniques*, 1998. [CrossRef] [Google Scholar] [Publisher Link]

[2] Robert D. Holtz, and William D. Kovacs, *Introduction to Geotechnical Engineering*, Presses Inter Polytechnique, 1991. [Google Scholar] [Publisher Link]

[3] Jean Pierre Giroud, "Soil Mechanics: Tables for Foundation Calculation, " *Dunod*, 1973. [Google Scholar]

[4] Gerard Philipponnat, and Bertrand Hubert, *Foundations and Earthworks*, Eyrolles, 2016. [Google Scholar] [Publisher Link]

[5] Serge Leroueil, Jean-Pierre Magnan, and François Tavenas, *Backfill on Soft Clays*, Lavoisier, Paris Cedex, France, 1985. [Google Scholar]

[6] NF P94-261, Justification of Geotechnical Work - National Application Standards for the Implementation of Eurocode 7 - Shallow Foundations, Afnor Editions, 2013. [Online]. Available: https://www.boutique.afnor.org/en-gb/standard/nf-p94261/justification-of-geotechnical-work-national-application-standards-for-the-i/fa174738/41472

[7] Jean-Pierre Magnan, and Philippe Mestat, "Soil Behavior Laws and Modeling," *Techniques Ingénieur*, 1997. [CrossRef] [Google Scholar] [Publisher Link]

[8] Fatima Dadouche-Zeroual, "Stability of Embankment Slopes on Soft Soils," *Proceedings of the 25th AUGC meeting*, Bordeaux, pp. 1-12, 2007. [Google Scholar] [Publisher Link]

[9] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., Springer New York, NY, pp. 1-19, 2006. [Google Scholar] [Publisher Link]

[10] Anthony T.C. Goh, "Back-Propagation Neural Networks for Modeling Complex Systems," *Artificial Intelligence in Engineering*, vol. 9, no. 3, pp. 143-151, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[11] Martin Ester et al., "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, pp. 226-231, 1996. [Google Scholar] [Publisher Link]

[12] Erich Schubert et al., "DBSCAN Revisited, Revisited: Why and How You Should (still) use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1-21, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[13] Sarat Kumar Das, and Prabir Kumar Basudhar, "Undrained Lateral Load Capacity of Piles in Clay using Artificial Neural Networks," *Computers and Geotechnics*, vol. 33, no. 8, pp. 454-459, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[14] John H. Schmertmann, "*Guidelines for Cone Penetration Test: Performance and Design*," Technical Report, United States Department of Transportation, Federal Highway Administration, 1978. [Google Scholar] [Publisher Link]

[15] Kok-Kwang Phoon, and Fred H. Kulhawy, "Characterization of Geotechnical Variability," *Canadian Geotechnical Journal*, vol. 36, no. 4, pp. 612-624, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[16] Hossein Moayedi, Hoang Nguyen, and Ahmad Safuan A. Rashid, “Novel Metaheuristic Classification Approach in Developing Mathematical Model-based Solutions Predicting Failure in Shallow Footing,” *Engineering with Computers*, vol. 37, no. 1, pp. 223-230, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Mohamed A. Shahin, Holger R. Maier, and Mark B. Jaksa, “Predicting Settlement of Shallow Foundations using Neural Networks,” *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 128, no. 9, pp. 785-793, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[18] Michael Nielsen, *Neural Networks and Deep Learning*, Determination Press, 2015. [Google Scholar] [Publisher Link]

[19] Sharmin Afrose et al., “Evaluation of Static Vulnerability Detection Tools with Java Cryptographic API Benchmarks,” *IEEE Transactions on Software Engineering*, vol. 49, no. 2, pp. 485-497, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Charles Swenson, “Using Machine Deep Learning AI to Improve Forecasting of Tax Payments for Corporations,” *Forecasting*, vol. 6, no. 4, pp. 968-984, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Jingxiao Yu et al., “Analysis of Spatiotemporal Characteristics of Microseismic Monitoring Data in Deep Mining based on ST-DBSCAN Clustering Algorithm,” *Processes*, vol. 13, no. 8, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[22] Darren J. Edwards, “A Functional Contextual, Observer-Centric, Quantum Mechanical, and Neuro-Symbolic Approach to Solving the Alignment Problem of Artificial General Intelligence: Safe AI Through Intersecting Computational Psychological Neuroscience and LLM Architecture for Emergent Theory of Mind,” *Frontiers in Computational Neuroscience*, vol. 18, pp. 1-45, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[23] Wengang Zhang et al., “Assessment of Pile Drivability using Random Forest Regression and Multivariate Adaptive Regression Splines,” *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, vol. 15, no. 1, pp. 27-40, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Xiangzhen Zhou, Chuang Zhao, and Xuecheng Bian, “Prediction of Maximum Ground Surface Settlement Induced by Shield Tunneling using XGboost Algorithm with Golden-Sine Seagull Optimization,” *Computers and Geotechnics*, vol. 154, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Talas Fikret Kurnaz, and Yilmaz Kaya, “The Performance Comparison of the Soft Computing Methods on the Prediction of Soil Compaction Parameters,” *Arabian Journal of Geosciences*, vol. 13, no. 4, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] H. Harmouzi et al., “Geomorphological and Geological Analysis of Akchour Landslide (Rif, Morocco),” *Geo-Eco-Trop Review*, vol. 1, no. 42, 19-31, 2018. [Google Scholar] [Publisher Link]

[27] Achraf Berrajaa, Mostafa Merras, and Issam Berrajaa, “Advanced CNN based on Genetic Algorithm to Automated Femoral Neck Fracture Classification,” *Signal, Image and Video Processing*, vol. 18, no. 6-7, pp. 5229-5238, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Fascicle No. 62 - Title V, - Technical Rules for the Design and Calculation of Foundations of Civil Engineering Structures, Ministry of Equipment, Housing and Transport, 1993. [Online]. Available: https://piles.cerema.fr/IMG/pdf/fascicule_62_titre_v_cctg_1993_regles_de_conception_et_calcul_des_fondations_des_ouvrages_gc_cle7a4325.pdf