

mgr inż. Piotr Wyborski¹⁾
ORCID: 0000-0003-1831-8300

Assessment of the impact of soil characteristics on crack formation and their prediction using machine learning

Ocena wpływu cech gruntu na powstawanie pęknięć i ich prognozowanie z wykorzystaniem uczenia maszynowego

DOI: 10.15199/33.2025.09.21

Abstract. The aim of the article was to develop a model identifying soil characteristics and environmental factors that determine its susceptibility to cracking. Additionally, a predictive algorithm was developed to forecast the intensity of soil cracking. The random forest method was used, based on 261 cases. The most significant variables influencing crack intensity were found to be soil moisture, specific density, and soil sample size. The developed algorithm demonstrated high predictive accuracy.

Keywords: soil cracking; soil desiccation; random forest; machine learning in geotechnics.

Streszczenie. W artykule zaprezentowano model identyfikujący cechy gruntu oraz właściwości środowiskowe decydujące o jego podatności na pękanie. Dodatkowo opracowano algorytm predykcyjny służący do prognozowania intensywności pęknięć gruntu. Wykorzystano metodę lasu losowego na bazie 261 przypadków. Najistotniejszymi zmiennymi wpływającymi na intensywność pęknięć okazały się: wilgotność gruntu, gęstość właściwa i wymiary próbki. Opracowany algorytm charakteryzuje się bardzo dużą dokładnością.

Słowa kluczowe: pęknięcia gruntu; wysychanie gruntu; las losowy; uczenie maszynowe w geotechnice.

Soil cracking caused by drying can significantly affect its mechanical, hydrogeological, physicochemical, and thermal properties [1]. From the point of view of soil mechanics, their presence may reduce strength, increase compressibility and accelerate consolidation. As a result, they negatively impact earth structures and foundations, reducing their load-bearing capacity and stability. Therefore, it is crucial to recognize soil characteristics and conditions that promote cracking, allowing for engineering measures to mitigate this phenomenon and its consequences. Table 1 presents the impact of soil cracking on various areas of civil engineering.

Soil is a three-phase material with a complex structure, whose fracture process depends on many different factors. The most frequently mentioned factors in the literature include soil moisture, mineralogical composition, soil density, and temperature. Expansive soils, whose volume changes significantly due to moisture fluctuations, are of particular importance in this context. Despite numerous studies, it has not been clearly established which soil characteristics determine its susceptibility to cracking. This paper attempts to develop a model that identifies soil characteristics and environmental properties that determine its susceptibility to cracking. A predictive algorithm for predicting crack intensity based on input data is also proposed. This interdisciplinary approach combines geotechnical engineering with data analysis, enabling a more comprehensive understanding of the problem and improved prediction accuracy.

Pęknięcia gruntu spowodowane wysychaniem mogą mieć duży wpływ na jego właściwości mechaniczne, hydrogeologiczne, fizykochemiczne i termiczne [1].

Z punktu widzenia mechaniki gruntów ich obecność zmniejsza wytrzymałość, zwiększa ścisłość i przyspiesza konsolidację. W efekcie pęknięcia gruntu negatywnie oddziałują na budowle ziemne i posadowione konstrukcje, zmniejszając ich nośność i stateczność. Kluczowa jest zatem ocena cech gruntu i warunków sprzyjających pękaniu, co umożliwia podejmowanie działań inżynierskich ograniczających to zjawisko i jego skutki.

Tabela 1 przedstawia wpływ pęknięcia gruntu na różne obszary inżynierii lądowej. Grunt jest trójfazowym materiałem o złożonej strukturze, którego proces pęknięcia zależy od wielu różnych czynników. W literaturze najczęściej wymieniana jest skład mineralogiczny, wilgotność oraz gęstość gruntu i temperaturę. Szczególne znaczenie w tym kontekście mają grunty ekspansywne, których objętość znacznie się zmienia pod wpływem wahań wilgotności. Pomimo wielu badań nie ustalono jednoznacznie, które cechy gruntu decydują o jego podatności na pękanie. W artykule podjęto próbę opracowania modelu identyfikującego cechy gruntu oraz właściwości środowiskowe, które decydują o jego podatności na pękanie. Zaproponowano również algorytm predykcyjny prognozujący intensywność pęknięć na podstawie danych wejściowych. Podejście ma charakter interdyscyplinarny, ponieważ łączy geotechnikę z analizą danych, umożliwiając pełniejsze ujęcie problemu i poprawę trafności prognoz.

¹⁾ Politechnika Wrocławska, Wydział Budownictwa Lądowego i Wodnego, piotr.wyborski@pwr.edu.pl

Research method

The study involved developing a database containing geotechnical soil properties, environmental conditions during drying, and a parameter describing the intensity of cracking in a given moisture regime. Based on the collected data and the application of machine learning methods, key factors influencing crack initiation and development were identified.

There are numerous parameters used to describe soil cracking, among which one of the most commonly applied is the Crack Intensity Factor (CIF). In the present study, CIF was selected as the indicator of cracking potential, and it is defined as follows:

$$CIF = \frac{A_c}{A_t} \cdot 100\% \quad (1)$$

where: A_c – total crack area; A_t – the total surface area of the soil specimen.

Since the value of CIF depends solely on the ratio of crack surface area to the total soil surface area, it provides a simple and unambiguous indicator for assessing the soil's susceptibility to cracking.

The analysis of crack formation is usually carried out by photographing the surface of the soil specimen during drying. The images are taken at regular time intervals or when the specimen reaches a specified moisture content.

In this study, three different cohesive soil materials were analyzed, differing, among other aspects, in grain size distribution and mineral composition, which influence their physical and mechanical properties. The materials investigated included a clayey silt and the same material amended with 15% and 35% montmorillonite by weight. The basic properties of the tested soils are summarized in Table 2. The soil was dried, pulver-

Table 1. Impact of soil cracking due to desiccation on various areas of civil engineering. Modified from [2]

Tabela 1. Wpływ pęknięcia gruntu wskutek wysychania na różne obszary inżynierii lądowej. Opracowano na podstawie [2]

Area / Obszar zagadnień	Consequences of desiccation soil cracking / Skutki pęknięć gruntu
Slope / Skarpy	cracks in the upper parts of slopes may initiate landslides that pose a threat to structures founded within their reach / pęknięcia w górnych partiach zboczy mogą inicjować osuwiska, zagrażające konstrukcjom posadowionym w ich zasięgu
Embankment / Wały	cracks can create privileged filtration paths in the embankment body, leading to hydraulic breakthrough and caverns / pęknięcia mogą tworzyć uprzywilejowane drogi filtracji w korpusie wału, prowadząc do przebiccia hydraulicznego i kawern
Dam / Zapory	cracks promote infiltration, increase hydraulic conductivity and reduce the mechanical properties of the soil; they can also act as a slip surface / pęknięcia sprzyjają infiltracji, zwiększają przewodność hydrauliczną i zmniejszają parametry mechaniczne gruntu; mogą też stanowić powierzchnię poślizgu
Foundation / Fundamenty	cracks can cause bearing capacity and settlement problems / pęknięcia mogą powodować problemy z nośnością i osiadaniem
Pipeline / Rurociągi	pipelines placed in soils susceptible to shrinkage are exposed to excessive settlement and, as a result, failure / rurociągi posadowione w gruntach podatnych na skurcz narażone są na nadmierne osiadanie i w efekcie awarie
Road subgrade / Podbudowy dróg	expansive soils prone to cracking contribute to road substructure failure / grunty ekspansywne, podatne na pęknięcia, przyczyniają się do awarii podbudowy dróg

Metoda badań

Badanie polegało na opracowaniu bazy danych zawierającej właściwości geotechniczne gruntu, warunki środowiskowe podczas wysychania oraz parametr opisujący intensywność pęknięcia w danym reżimie wilgotnościowym. Na podstawie zgromadzonych danych oraz zastosowania metod uczenia maszynowego zidentyfikowano kluczowe czynniki wpływające na inicjację i rozwój pęknięć.

Istnieje wiele parametrów służących do opisu pęknięcia gruntu. Jednym z najczęściej stosowanych jest współczynnik intensywności pęknięć – CIF (*Crack Intensity Factor*), który w badaniu wybrano jako wskaźnik potencjału pęknięcia. Jest on określony wzorem (1):

$$CIF = \frac{A_c}{A_t} \cdot 100\% \quad (1)$$

gdzie: A_c – powierzchnia pęknięć; A_t – całkowita powierzchnia próbki gruntu.

Wartość CIF zależy wyłącznie od stosunku powierzchni pęknięć do całkowitej powierzchni gruntu, dlatego też stanowi on prosty i jednoznaczny wskaźnik oceny podatności gruntu na pęknięcie. Analiza powstawania pęknięć polega zazwyczaj na fotografowaniu powierzchni próbki gruntu podczas wysychania. Zdjęcia wykonuje się w regularnych odstępach czasu lub gdy próbka gruntu osiągnie określoną wilgotność.

W ramach przeprowadzonego badania przeanalizowano trzy różne spoiste materiały gruntowe, zróżnicowane m.in. pod względem uziarnienia i składu mineralnego, co wpływa na ich właściwości fizyczne i mechaniczne. Przedmiotem analizy był pył oraz ten sam materiał z dodatkiem montmorylonitu w ilości 15% i 35% wagowo. Podstawowe właściwości badanych gruntów zestawiono w tabeli 2.

Table 2. Basic properties of the tested soils

Tabela 2. Podstawowe właściwości badanych gruntów

Soil type ¹⁾ / Rodzaj gruntu ¹⁾	Specific gravity ²⁾ / Gęstość właściwa ²⁾ ρ_s [g/cm ³]	Sand content ³⁾ / Frakcja piaszkowa ³⁾ [%]	Silt content ³⁾ / Frakcja pyłowa ³⁾ [%]	Clay content ³⁾ / Frakcja ilowa ³⁾ [%]	Plastic limit ⁴⁾ / Granica plastyczności ⁴⁾ w_p [%]	Liquid limit ⁴⁾ / Granica płynności ⁴⁾ w_L [%]	USCS classification / Klasyfikacja USCS
Clay / Il	2.69	5.8	55.3	40.9	49.3	83.4	CH
Silty clay / Il pylasty	2.71	7.6	69.7	22.7	31.0	53.4	CH
Clayey silt / Pył ilasty	2.72	9	82	9	17.2	31.7	CL

¹⁾ sklasyfikowano wg PN-EN ISO 14688-1, PN-EN ISO 14688-2; ²⁾ oznaczono wg PN EN ISO 17892-3;

³⁾ określono wg PN-EN ISO 17892-4; ⁴⁾ wyznaczono zgodnie z PN-EN ISO 17892-12.



ized, and moistened with distilled water to the target moisture content, added in portions and thoroughly mixed after each addition. Once a homogeneous consistency was achieved, the soil was placed in a circular mold with a diameter of 15 cm, carefully compacted to fill the form. The specimens were weighed, sealed to prevent evaporation, and left for three days at room temperature to allow sedimentation and the removal of air from the pores. They were then dried in a laboratory oven. Every 30 minutes, photographs of the surface were taken and the mass was measured to determine the moisture content. For clay and clayey silt, two test series were performed. The ranges of moisture content within which the analysis of the tested soils was conducted are presented in Table 3. The photographs were catalogued and processed using the *ImageJ* software for crack identification. The workflow of image editing and analysis is illustrated in Photo. This procedure enabled accurate determination of the CIF coefficient in accordance with equation (1).

Based on the presented methodology, 58 records were generated and entered into the database. To increase diversity and broaden the scope of analysis, the dataset was supplemented with literature data [1, 3÷9]. The basic soil parameters obtained from the literature are summarized in Table 4. These parameters were determined in accordance with various testing standards adopted in the cited publications. The referenced sources represent key studies documenting the relationship between surface crack intensity, expressed by the CIF factor, and soil moisture content. The final database comprises 261 cases containing information on soil properties, specimen dimensions, and drying conditions. The database includes the following information:

- specific gravity ρ_s [g/cm³];
- sand content f_p [%];
- silt content f_π [%];
- clay content f_i [%];

Table 3. Moisture content ranges used in the desiccation tests and crack analysis

Tabela 3. Zakres wilgotności zastosowany w badaniach wysychania i analizie spękań

Soil type / Rodzaj gruntu	Initial soil moisture / Początkowa wilgotność gruntu [%]	Mean final soil moisture / Średnia końcowa wilgotność gruntu [%]
Clay / Ił	151,6 ¹⁾ , 143,4 ²⁾	73,8 ¹⁾ , 88,6 ²⁾
Silty clay / Ił pylasty	100,0 ¹⁾	17,9 ¹⁾
Clayey silt / Pył ilasty	62,0 ¹⁾ , 59,7 ²⁾	3,3 ¹⁾ , 3,1 ²⁾

¹⁾ the first series of tests / pierwsza seria badań;

²⁾ the second series of tests / druga seria badań

Grunt wysuszono, rozdrobniono i nawilżono wodą destylowaną do założonej wilgotności, dodając ją porcjami i mieszając po każdej dawce. Po uzyskaniu jednolitej konsystencji grunt umieszczono w okrągłej formie o średnicy 15 cm i dokładnie ją wypełniono. Próbkę zważono, zabezpieczono przed wysychaniem i pozostawiono na trzy dni w temperaturze pokojowej, aby umożliwić sedymentację i usunięcie powietrza z porów. Następnie suszono je w suszarce laboratoryjnej. Co 30 min wykonywano zdjęcia powierzchni oraz mierzono masę w celu określenia wilgotności. W przypadku iłu oraz pyłu ilastego wykonano po dwie serie badań. Zakresy wilgotności, w których prowadzi-

zono analizę badanych gruntów, przedstawiono w tabeli 3. Zdjęcia skatalogowano i przetworzono w programie *ImageJ* w celu identyfikacji pęknięć. Przebieg edycji i analizy obrazu przedstawiono na fotografii. Procedura ta umożliwiła dokładne wyznaczenie współczynnika CIF zgodnie z wzorem (1).

Na podstawie przedstawionej metody opracowano 58 rekordów, które wprowadzono do bazy danych. W celu zwiększenia różnorodności i poszerzenia perspektywy analizy, katalog uzupełniono o dane literaturowe [1, 3÷9]. Podstawowe parametry gruntów zaczerpnięte z literatury zestawiono w tabeli 4. Parametry zostały określone zgodnie z różnymi normami badawczymi, przyjętymi w cytowanych publikacjach. Źródła

te stanowią kluczowe opracowania dokumentujące zależność intensywności spękań powierzchniowych, opisaną współczynnikiem CIF, od wilgotności gruntu. Finalna baza danych obejmuje 261 przypadków zawierających informacje o cechach gruntu, wymiarach próbek i warunkach suszenia. Informacje ujęte w bazie danych:

- gęstość właściwa ρ_s [g/cm³];
- udział frakcji piaskowej f_p [%];
- udział frakcji pyłowej f_π [%];
- udział frakcji ilowej f_i [%];

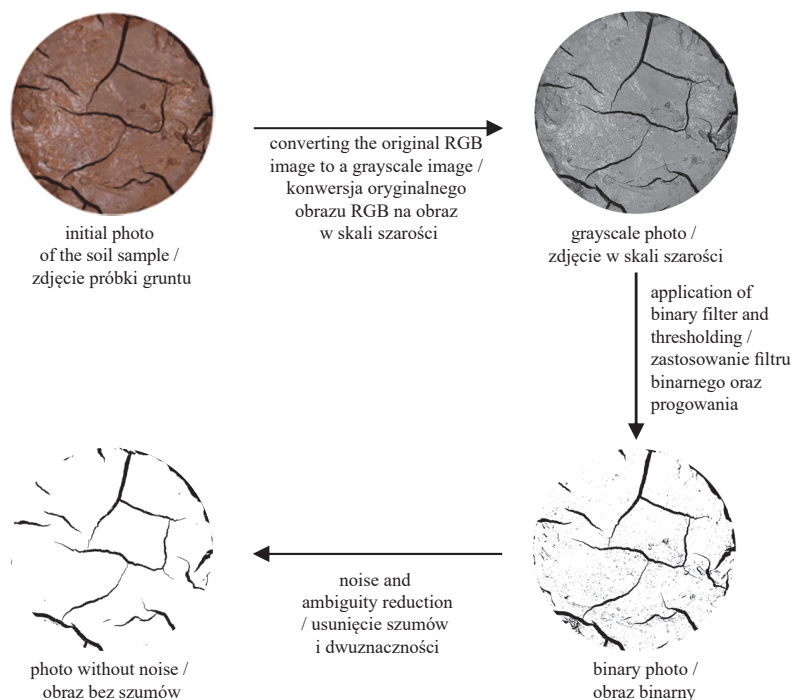


Image processing procedure for calculating the Crack Intensity Factor (CIF)
Procedura obróbki zdjęć w celu obliczenia współczynnika CIF

Table 4. Basic parameters of soils sourced from the literature

Tabela 4. Podstawowe parametry gruntów zaczerpnięte z literatury

Reference / Źródło	Specific gravity / Gęstość właściwa ρ_s [g/cm ³]	Sand content / Frakcja piaskowa [%]	Silt content / Frakcja pyłowa [%]	Clay content / Frakcja ilowa [%]	Plastic limit / Granica plastyczności w_p [%]	Liquid limit / Granica płynności w_L [%]	USCS classification / Klasyfikacja USCS
[1]	2.73	2	76	22	20	37	CL
[3]	–	–	–	–	–	48	–
[4]	2.73	2	76	22	18.9	36.7	CL
	2.7	4	67	29	22.7	34.5	CL
	2.61	2	48	50	32.9	52.6	CL
[5]	2.67	–	–	84	40	75	CH
[6]	–	24	34	34	24.8	40.4	–
[7]	2.71	–	–	–	29.1	76.2	CH
[8]	2.79	2	19	79	40	77	CH
[9]	2.74	3	55	42	–	57	CH
[10] ⁵⁾	–	3.36	29.2	67.44	–	–	–
[11] ⁵⁾	2.75	6	63	31	26.1	52.6	CH

⁵⁾ the data were used to test the algorithm / dane służyły do testowania algorytmu

- liquid limit w_L [%];
- plastic limit w_p [%];
- shrinkage limit w_s [%];
- plasticity limit I_p [%];
- USCS classification [–];
- montmorillonite content [%];
- specimen height h [cm];
- specimen shape;
- surface area A [cm²];
- drying temperature t [°C];
- relative humidity of air RH [%];
- sample moisture [%];
- crack intensity factor CIF [%].

To evaluate the influence of soil properties on crack formation, the random forest method was applied, which is a machine learning technique based on the structure of decision trees. A decision tree is a predictive model built on a sequence of logical “if-then” conditions that lead to a decision or prediction. Each such model partitions the data into smaller groups based on the values of input variables, referred to as predictor variables (in this case, soil properties and drying conditions). The random forest constructs multiple trees and then aggregates their outcomes into a single ensemble model, thereby improving the accuracy and stability of the prediction. This technique is applicable to both classification (e.g., assigning categories) and regression (e.g., predicting numerical values). Its objective is to build a model that, based on the values of input features (predictors), predicts the value of the response variable, in this case, the CIF factor.

The random forest model is based on recursive partitioning of the predictor variable space, that is, repeatedly splitting the input data into smaller subsets according to specified logical conditions. As a result, partitions are created to which predicted values of the response variable are assigned. During

- granica płynności w_L [%];
- granica plastyczności w_p [%];
- granica skurczalności w_s [%];
- indeks plastyczności I_p [%];
- klasyfikacja USCS [–];
- udział montmorylonitu w części ilowej [%];
- wysokość formy h [cm];
- kształt formy;
- pole powierzchni próbki A [cm²];
- temperatura suszenia t [°C];
- względna wilgotność powietrza RH [%];
- wilgotność próbki [%];
- wskaźnik intensywności pęknięć CIF [%].

Do oceny wpływu cech gruntu na powstawanie pęknięć zastosowano las losowy (random forest), czyli metodę uczenia maszynowego, bazującą na strukturze drzew decyzyjnych. Drzewo decyzyjne to model predykcyjny wykorzystujący ciąg warunków logicznych typu „if-then”, który prowadzi do decyzji lub prognozy. Każdy taki model dzieli dane na mniejsze grupy (partycje) na podstawie wartości zmiennych wejściowych, zwanych zmiennymi predykcyjnymi (w tym przypadku cechy gruntu i warunki suszenia). Las losowy buduje wiele drzew, a następnie łączy ich wyniki w jeden model zbiorczy, zwiększając dokładność i stabilność prognozy. Technika ta znajduje zastosowanie zarówno w klasyfikacji (np. przypisanie kategorii), jak i regresji, np. przewidywanie wartości liczbowej. Jej celem jest stworzenie modelu, który na podstawie wartości cech wejściowych (predyktorów) przewiduje wartość zmiennej odpowiedzi – w tym przypadku współczynnika CIF.

Model lasu losowego bazuje na rekurencyjnym podziale przestrzeni zmiennych predykcyjnych, czyli wielokrotnym dzieleniu danych wejściowych na mniejsze podzbiory wg określonych warunków logicznych. W efekcie tworzone





model construction, the bootstrap mechanism is applied, i.e., sampling with replacement, to generate multiple training sets. Each tree is trained on a different data sample, and at each node split a limited subset of predictor variables is randomly selected, which substantially reduces correlation among the trees. This lowers the variance of the model and helps to prevent overfitting. In this study, the set of predictors included soil properties and drying conditions contained in the database. The predicted variable was the CIF factor, which describes the susceptibility of soil to cracking. The number of trees in the random forest was set to 100, a value commonly adopted in the literature. The algorithm used in this study was implemented by the author in the *R* environment using the *randomForest* package for building the random forest model, and the *caret* package for hyperparameter tuning and cross-validation.

Model quality assessment

This chapter presents the evaluation of the random forest model quality. The assessment was based on the out-of-bag (OOB) error, which is typical for validation in random forest algorithms. In addition, hyperparameters were tuned to improve the predictive performance. The OOB error is a measure of model quality, calculated based on observations not included in the construction of particular decision trees. This results from the use of bootstrapping – each tree was assigned a random training sample with replacement, containing on average 62.92% of the data. The remaining 37.08% formed the OOB sample, which was not used in training the given tree. For each such observation, the prediction was based exclusively on the trees that had not seen it. An incorrect prediction was treated as an error, and the final OOB error value corresponded to the proportion of misclassifications within the training set.

In the analyzed case, the OOB error was used to assess the adequacy of the number of trees and its impact on the stability and quality of predictions. Figure 1 presents the variation of the OOB error with respect to the number of trees. The error value stabilizes at approximately 30 trees, oscillating around a relatively constant level. This indicates that further increasing the number of trees does not significantly improve the model accuracy. At the same time, a larger number of trees did not substantially affect the computation time, which allowed the original configuration to be maintained without loss of efficiency.

Despite applying an adequate number of trees, the model quality was not fully satisfactory. Therefore, model tuning was performed by constructing a grid of values for the hyperparameter defining the number of randomly selected predictors at each tree split. Each hyperparameter configuration was evaluated using tenfold cross-validation.

szą partycje, którym przypisuje się przewidywane wartości zmiennej odpowiedzi. Podczas budowy modelu stosuje się mechanizm bootstrapowania, czyli losowania ze zwracaniem w celu stworzenia wielu zbiorów treningowych. Każde drzewo uczy się na innej próbce danych, a przy każdym podziale węzła wybierany jest losowo ograniczony zestaw zmiennych predykcyjnych, co znacznie zmniejsza korelację między drzewami. To redukuje wariancję modelu i przeciwdziała przeuczeniu. W omawianym badaniu zestaw predyktorów obejmował cechy gruntu i warunki suszenia zawarte w bazie danych. Zmienną prognozowaną był współczynnik CIF. Liczbę drzew w lesie losowym określono jako 100 (jest to wartość stosowana w literaturze). Zastosowany algorytm przygotowano w środowisku *R* z zastosowaniem pakietu *randomForest*, służącego do budowy modelu lasu losowego oraz pakietu *caret*, przeznaczonego do strojenia hiperparametrów i przeprowadzania krosvalidacji.

Ocena jakości modelu

Podstawą oceny jakości modelu lasu losowego był błąd out-of-bag (OOB), typowy dla walidacji w algorytmach lasu losowego. Dodatkowo dostrojono hiperparametry w celu poprawy skuteczności predykcji. Błąd OOB to miara jakości modelu, wyznaczana na podstawie obserwacji nieuwzględnionych przy budowie konkretnych drzew decyzyjnych. Wynika to z zastosowania bootstrapowania – każdemu drzewu przypisywano losową próbkę treningową z powtórzeniami, zawierającą średnio 62,92% danych. Pozostałe 37,08% danych stanowiła próbka OOB, niewykorzystywana w uczeniu danego drzewa. W przypadku każdej takiej obserwacji predykcja opierała się wyłącznie na drzewach, które jej nie widziały. Niepoprawna predykcja traktowana była jako błąd, a ostateczna wartość błędu OOB, to odsetek błędnych predykcji w zbiorze treningowym.

W analizowanym przypadku błąd OOB posłużył do oceny adekwatności liczby drzew oraz jej wpływu na stabilność i jakość prognoz. Rysunek 1 przedstawia zmianę błędu OOB względem liczby drzew. Już przy ok. trzydziestu drzewach wartość błędu stabilizuje się, oscylując wokół poziomu względnie stałego. Oznacza to, że dalsze zwiększanie liczby drzew nie poprawia istotnie dokładności modelu. Jednocześnie większa liczba drzew nie wpływała istotnie na czas obliczeń, co umożliwiło utrzymanie pierwotnej konfiguracji bez utraty wydajności.

Mimo zastosowania odpowiedniej liczby drzew, jakość modelu nie była w pełni satysfakcjonująca, dlatego też przeprowadzono strojenie modelu, tworząc siatkę wartości hiperparametru, określającego liczbę losowo wybieranych predyktorów przy podziale drzewa. Każdą konfigurację hiperparametru oceniano za pomocą dziesięciokrotnej walidacji krzyżowej (krosvalidacji), każdo-

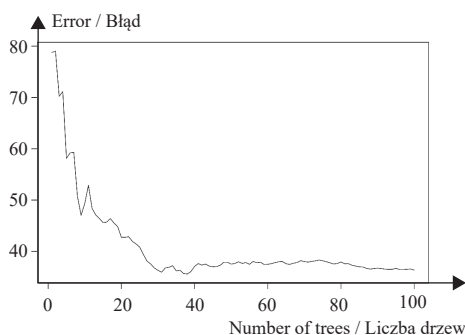


Fig. 1. Random forest error depending on the number of trees

Rys. 1. Błąd lasu losowego w zależności od liczby drzew

idation, where the data were repeatedly divided into training and test sets. The hyperparameter value that provided the highest predictive accuracy was selected.

Analysis of results

Figure 2 presents the variable importance based on the trained random forest model. The soil moisture content exerts the strongest influence on crack intensity, while other important factors include soil specific gravity, specimen height and surface area, drying temperature, and sand content. The least important variables are the USCS classification, specimen shape, and plastic limit.

Variable importance was assessed using the $\%IncMSE$ index, which represents the percentage increase in mean squared error after random permutation of a given variable's values. A high value of this index for a variable indicates that random permutations of that variable lead to a significant deterioration in predictive accuracy. For this reason, the $\%IncMSE$ index is used to evaluate variable importance in random forest models.

razowo dzieląc dane na zbiory uczące i testowe. Wybrano taką wartość hiperparametru, która zapewniała największą trafność predykcji.

Analiza wyników

Rysunek 2 przedstawia istotność zmiennych na podstawie wytrenowanego modelu lasu losowego. Najsilniej na intensywność pęknięć wpływa wilgotność gruntu, ale istotne są także gęstość właściwa gruntu, wysokość i pole powierzchni próbki, temperatura suszenia oraz zawartość frakcji piaszczystej. Zmienne najmniej istotne to klasyfikacja USCS, kształt formy oraz granica plastyczności.

Istotność zmiennych oceniano za pomocą wskaźnika $\%IncMSE$, określającego procentowy wzrost średniego błędu kwadratowego po losowej permutacji wartości danej zmiennej. Duża wartość tego wskaźnika w przypadku danej zmiennej oznacza, że losowe permutacje tej zmiennej prowadzą do istotnego pogorszenia jakości predykcji. Z tego względu wskaźnik ten służy do oceny istotności zmiennych w modelu lasów losowych.

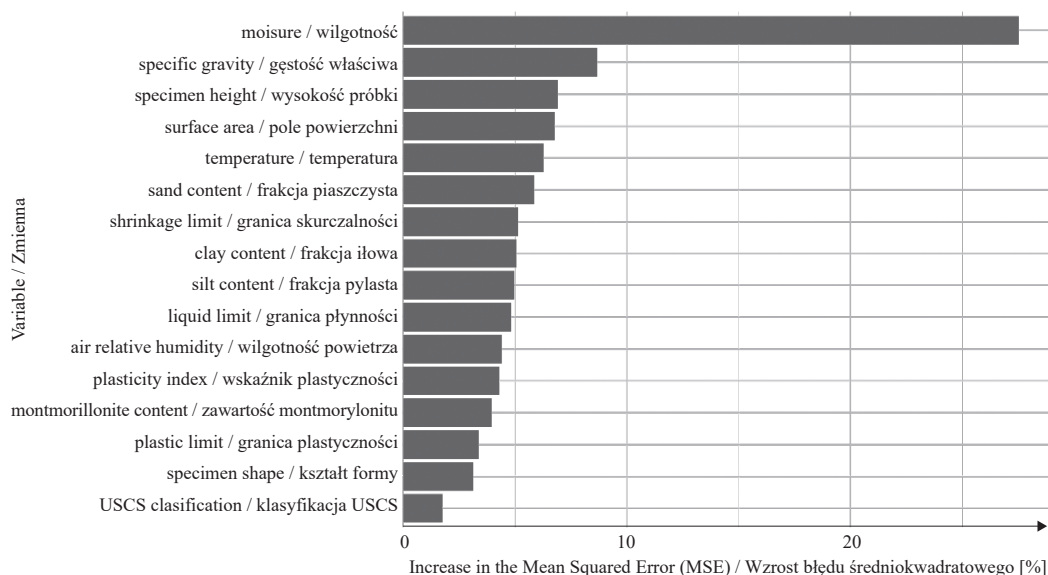


Fig. 2. Influence of variables on prediction in the random forest model

Rys. 2. Wpływ zmiennych na predykcję w modelu lasu losowego

From the comparison of predicted and observed CIF values after tuning the random forest model, the tuned model demonstrates relatively good fit. The mean absolute deviation (MAD) is 0.318, and the mean squared deviation (MSD) is 0.166. The best fit was obtained using polynomial regression ($R^2 = 0.993$) and linear regression ($R^2 = 0.992$). The low values of dispersion and variability indicate the high quality of the model. Due to the high accuracy of the model, an algorithm was developed on its basis to predict the value of the CIF using the variables employed in training the random forest. The algorithm allows for the input of only selected variables, while leaving the remaining ones without assigned values. To evaluate the predictive performance, the predicted and actual CIF values were compared

Z porównania prognozowanych i obserwowanych wartości CIF po dostrojeniu modelu lasu losowego, który wówczas charakteryzuje się relatywnie dobrym dopasowaniem, wynika, że średnie odchylenie bezwzględne (MAD) wynosi 0,318, a średnie odchylenie kwadratowe (MSD) 0,166. Najlepsze dopasowanie otrzymano za pomocą regresji wielomianowej ($R^2 = 0,993$) oraz regresji liniowej ($R^2 = 0,992$). Małe wartości rozproszenia i zmienności świadczą o bardzo dobrej jakości i dokładności modelu, dlatego też na jego podstawie opracowano algorytm prognozujący wartość współczynnika intensywności pęknięć z wykorzystaniem zmiennych, które posłużyły do wytrenowania lasu losowego. Algorytm umożliwia uzupełnienie jedynie wybranych zmiennych, dopuszczając pozostawienie pozostałych bez przypisanych wartości. W celu oceny jakości predykcji porównano pro-





for 35 cases from the training database (serial numbers in the range 1–261) and eight cases outside of it (300–307). Sample results of the prediction quality analysis are presented in Table 5, comparing experimental and predicted values together with the relative error. Prediction results with a relative error lower than 10% were considered of good quality, while predictions with an error in the range of 10–20% were regarded as acceptable. Predictions with errors exceeding this range were classified as unsatisfactory. The percentage distribution of results categorized as good, acceptable, and unsatisfactory for all data used in CIF prediction by the algorithm is presented in Table 6. Predictions of CIF based on the data included in the random forest training are characterized by very good quality. More than 74% of the 35 predictions can be considered good. Acceptable and unsatisfactory predictions account for 14.3% and 11.3% of all results produced by the algorithm, respectively. In the case of data not included in the random forest training (the test set), the prediction quality was lower – 25% of results were classified as good, 50% as acceptable, and 25% as unsatisfactory.

Conclusions

The random forest represents an effective tool for predicting variable values in complex nonlinear systems, such as processes related to soil cracking induced by drying. In the present study, the high predictive quality of the random forest was largely the result of appropriate model parameter

Table 5. Results of the quality analysis of the algorithm used for predicting the Crack Intensity Factor (CIF)

Tabela 5. Wyniki analizy jakości algorytmu służącego do predykcji wskaźnika intensywności pęknięć (CIF)

No. / Lp.	Reference / Źródło danych	Moisture / Wilgotność [%]	CIF [%]	Prediction CIF / Predykcja CIF [%]	Difference / Różnica	Error / Błąd [%]
5	[5]	35.0	15.0	15.5	0.5	3.0
11	[4]	1.5	14.1	13.3	0.8	6.1
33	[6]	18.9	2.4	3.0	0.6	27.3
47	[1]	12.0	18.8	18.3	0.5	2.6
55	[7]	34.1	6.7	6.8	0.1	0.8
70	[8]	8.5	14.1	14.0	0.1	0.4
126		14.5	13.8	14.2	0.4	3.0
142		14.3	14.0	14.3	0.3	2.0
162	[9]	6.0	22.2	20.9	1.3	5.8
195	[3]	43.5	16.0	16.6	0.6	3.8
208	Own research / Badania własne	27.0	19.1	18.0	1.2	5.7
222		122.9	36.4	32.1	4.3	11.9
236		75.8	26.5	26.5	0.0	0.0
248		0.1	23.5	24.4	0.9	3.8
259		104.0	36.6	36.7	0.1	0.4
303 ⁶⁾	[10]	35	6	5.6	0.4	5.8
306 ⁶⁾	[11]	10	24	18.1	5.9	24.6
307 ⁶⁾	[11]	10	25	20.1	4.9	19.6

⁶⁾ records from the test set / rekordy pochodzące ze zbioru testowego

Table 6. Percentage distribution of CIF prediction accuracy

Tabela 6. Procentowy rozkład trafności predykcji CIF

Reference / Źródło	Number of predictions / Liczba predykcji	Prediction CIF / Predykcje CIF [%]		
		good / dobre	acceptable / dopuszczalne	unsatisfactory / niezadowolające
Training set / Zbiór treningowy	35	74,3	14,3	11,3
Test set / Zbiór testowy	8	25	50	25

dopuszczalne oraz niezadowolające można zaklasyfikować odpowiednio 14,3% i 11,3% wszystkich wyników przewidywanych przez algorytm. W przypadku danych nieobjętych trenowaniem lasu losowego (zbioru testowego), jakość predykcji była niższa – 25% wyników zaklasyfikowano jako dobre, 50% jako zadowolające, a 25% jako niedopuszczalne.

Wnioski

Las losowy stanowi skuteczne narzędzie do prognozowania wartości zmiennych w złożonych układach nieliniowych, takich jak procesy związane z pękaniem gruntów w wyniku wysychania. W opisanym badaniu wysoka jakość predykcji lasu losowego była w dużej mierze efektem odpowiedniego

gnozowane i rzeczywiste wartości CIF dotyczące 35 przypadków z bazy treningowej (liczby porządkowe z zakresu 1–261) oraz osiem spoza niej (300–307). Przykładowe wyniki analizy jakości przedstawiono w tabeli 5, porównując wartości eksperymentalne i przewidywane wraz z błędem względnym. Za dobrą jakość predykcji uznano wyniki z błędem względnym mniejszym niż 10%, a jako dopuszczalne dopasowanie uznano predykcje, których błąd wynosił 10–20%. Powyżej tego zakresu prognozy CIF uznano za niezadowolające. Procentowy rozkład wyników zaklasyfikowanych jako poprawne, mieszczące się w granicach dopuszczalnego błędu oraz niezadowolające w przypadku wszystkich danych wykorzystanych do predykcji CIF za pomocą algorytmu, przedstawiono w tabeli 6. Predykcje CIF na podstawie danych objętych trenowaniem lasu losowego charakteryzują się bardzo dobrą jakością. Ponad 74% predykcji spośród 35 można uznać za dobre. Jako predykcje



tuning. The application of the random forest made it possible to identify which variables play the most significant role in the crack development process. Among the most important are soil moisture, soil specific gravity, specimen height and surface area, drying temperature, and sand content. Variables of minor importance include the USCS classification, specimen shape, and plastic limit. These results not only confirm the usefulness of the random forest algorithm in geotechnical studies, but also indicate potential directions for further optimization of the feature set and the development of tools supporting the prediction of nonlinear phenomena in geotechnics.

Based on the above analysis, it can be concluded that the developed algorithm constitutes a useful tool for predicting the intensity of soil cracking caused by drying. The CIF prediction results obtained for data not included in the random forest training process indicate the need for further expansion of the training dataset. With the advancement of research on soil cracking mechanisms, it becomes reasonable to extend the dataset used for model training and validation, which may significantly improve its predictive capability. The elevated relative errors observed in some cases may result from missing input data. When the value of a variable is not available, the algorithm replaces it with the median of the entire dataset. Although this enables further computations, it may reduce prediction accuracy, particularly for features with a strong influence on the outcome. Therefore, it is essential to provide as complete a dataset as possible, especially for variables that are critical to prediction.

Received: 17.04.2025

Revised: 04.06.2025

Published: 19.09.2025

dostrojenia parametrów modelu. Zastosowanie lasu losowego pozwoliło na określenie, które zmienne odgrywają najważniejszą rolę w procesie rozwoju pęknięć. Wśród najistotniejszych można wymienić: wilgotność gruntu, gęstość właściwą gruntu, wysokość oraz pole powierzchni badanej próbki, temperaturę suszenia i zawartość frakcji piaszczystej. Do mało istotnych zmiennych można zaliczyć klasyfikację USCS, kształt próbki oraz granicę plastyczności. Wyniki te nie tylko potwierdzają przydatność algorytmu lasu losowego w badaniach geotechnicznych, lecz także wskazują potencjalne kierunki dalszej optymalizacji zbioru cech i rozwoju narzędzi wspomagających prognozowanie zjawisk nieliniowych w geotechnice.

Na podstawie powyższej analizy można stwierdzić, że opracowany algorytm stanowi użyteczne narzędzie do prognozowania intensywności pęknięć gruntu spowodowanych jego wysychaniem. Wyniki predykcji CIF przeprowadzone dla danych nieuwzględnionych w procesie trenowania modelu lasu losowego wskazują na potrzebę dalszej rozbudowy zbioru danych uczących. Wraz z rozwojem badań nad mechanizmami pęknięcia gruntów zasadne staje się rozszerzanie zbioru danych wykorzystywanych do trenowania i walidacji modelu, co może istotnie poprawić jego zdolności predykcyjne. Podwyższone błędy względne w niektórych przypadkach mogą wynikać z brakujących danych wejściowych. W przypadku, gdy wartość zmiennej nie jest dostępna, algorytm zastępuje ją medianą z całego zbioru. Umożliwia to dalsze obliczenia, ale może zmniejszyć dokładność prognoz, szczególnie cech silnie wpływających na wynik. Z tego powodu należy zapewnić możliwie pełny zestaw danych, przede wszystkim zmiennych kluczowych dla predykcji.

Artykuł wpłynął do redakcji: 17.04.2025 r.

Otrzymano poprawiony po recenzjach: 04.06.2025 r.

Opublikowano: 19.09.2025 r.

Literature

- [1] Tang CS, Shi B, Cui YJ, Liu C, Gu K. Desiccation cracking behavior of polypropylene fiber-reinforced clayey soil. *Canadian Geotechnical Journal*. 2012;49; <https://doi.org/10.1139/t2012-067>.
- [2] Tang CS, Zhu C, Cheng Q, Zeng H, Xu JJ, Tian BG, Shi B. Desiccation cracking of soils: A review of investigation approaches, underlying mechanisms, and influencing factors. Elsevier B.V. 2021; <https://doi.org/10.1016/j.earscirev.2021.103586>.
- [3] Trabelsi H, Jamei M, Zenzri H, Olivella S. Crack patterns in clayey soils: Experiments and modeling. *Int J Numer Anal Methods Geomech*. 2012;36; <https://doi.org/10.1002/nag.1060>.
- [4] Tang CS, Shi B, Liu C, Zhao L, Wang B. Influencing factors of geometrical structure of surface shrinkage cracks in clayey soils. *Eng Geol*. 2008;101; <https://doi.org/10.1016/j.enggeo.2008.05.005>.
- [5] Cui YJ, Tang CS, Tang AM. Investigation of soil desiccation cracking using an environmental chamber. *Rivista Italiana di Geotecnica*, 2014;48.
- [6] Li JH, Lu Z, Guo LB, Zhang LM. Experimental study on soil-water characteristic curve for silty clay with desiccation cracks. *Eng Geol*. 2017;218; <https://doi.org/10.1016/j.enggeo.2017.01.004>.
- [7] Tang CS, Zhu C, Leng T, Shi B, Cheng Q, Zeng H. Three-dimensional characterization of desiccation cracking behavior of compacted clayey soil using X-ray computed tomography. *Eng Geol*. 2019;255; <https://doi.org/10.1016/j.enggeo.2019.04.014>.
- [8] Tang CS, Cui YJ, Tang AM, Shi B, Liu C. Quantification and Characterization of Temperature Effect on Desiccation Crack Network in Soil. 2013; doi:10.1061/9780784412787.076.
- [9] Tollenaar RN, van Paassen LA, Jommi C. Observations on the desiccation and cracking of clay layers. *Eng Geol*. 2017;230; <https://doi.org/10.1061/9780784412787.076>.
- [10] Lomeling D, Kenyi MC, Lodiong MA, Kenyi MS, Silvestro GM, Yieb JLL. Characterizing Dessication Cracking of a Remolded Clay (Eutric Vertisol) Using the Fractal Dimension Approach. *Open Journal of Soil Science*. 2016;06; <https://doi.org/10.4236/ojss.2016.64008>.
- [11] Luo Z, Wang S, Ou Q, Hamka MQ, Li X, Xu C, Ding X. Multi-effect investigation on desiccation crack evolution and mechanical behavior of swelling clay. *Acta Geotech*. 2023;18; <https://doi.org/10.1007/s11440-023-01910-8>.