

## Article

# Evaluation of Electrical Impedance Spectra by Long Short-Term Memory to Estimate Nitrate Concentrations in Soil

Xiaohu Ma , Luca Bifano  and Gerhard Fischerauer 

Chair of Measurement and Control Systems, Faculty of Engineering Science, University of Bayreuth, 95440 Bayreuth, Germany

\* Correspondence: xiaohu.ma@uni-bayreuth.de

**Abstract:** Monitoring the nitrate concentration in soil is crucial to guide the use of nitrate-based fertilizers. This study presents the characteristics of an impedance sensor used to estimate the nitrate concentration in soil based on the sensitivity of the soil dielectric constant to ion conductivity and on electrical double layer effects at electrodes. The impedance of synthetic sandy soil samples with nitrate nitrogen concentrations ranging from 0 to 15 mg/L was measured at frequencies between 20 Hz and 5 kHz and noticeable conductance and susceptance effects were observed. Long short-term memory (LSTM), a variant of recurrent artificial neural networks (RNN), was investigated with respect to its suitability to extract nitrate concentrations from the measured impedance spectra and additional physical properties of the soils, such as mass density and water content. Both random forest and LSTM were tested as feature selection methods. Then, numerous LSTMs were trained to estimate the nitrate concentrations in the soils. To increase estimation accuracy, hyperparameters were optimized with Bayesian optimization. The resulting optimal regression model showed coefficients of determination between true and predicted nitrate concentrations as high as 0.95. Thus, it could be demonstrated that the system has the potential to monitor nitrate concentrations in soils in real time and in situ.

**Keywords:** nitrate concentration; electrical impedance spectroscopy; EIS; recurrent artificial neural network; RNN; long short-term memory; LSTM; random forest



**Citation:** Ma, X.; Bifano, L.; Fischerauer, G. Evaluation of Electrical Impedance Spectra by Long Short-Term Memory to Estimate Nitrate Concentrations in Soil. *Sensors* **2023**, *23*, 2172. <https://doi.org/10.3390/s23042172>

Academic Editors: Waheb Abdullah, AbdulRahman Alsewari and Mario De Oliveira

Received: 17 January 2023  
Revised: 6 February 2023  
Accepted: 10 February 2023  
Published: 15 February 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

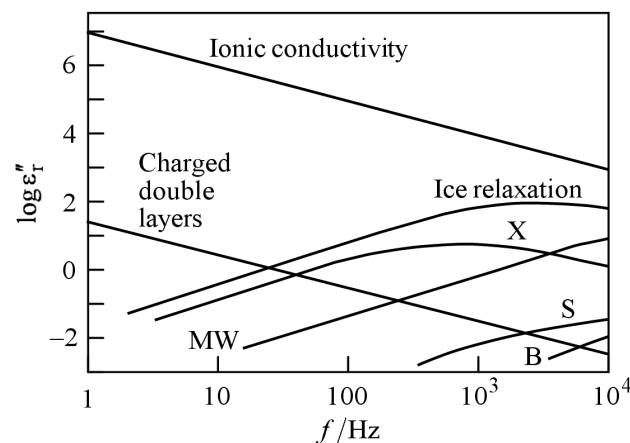
Nitrate nitrogen (nitrate-N) is one of the most important nutrients for securing the yield and quality of harvested products. However, excessive content of nitrate in soil leads to environmental problems. In Germany, nitrate is the main source of pollution for the groundwater. Nationwide more than a quarter of all near-surface groundwater fails to meet the nitrate target (50 mg/L) of the European Water Framework Directive [1]. Currently, nitrate concentrations in soils are measured by laboratory methods, such as flow injection analysis [2] and spectrophotometry [3]. Such methods require expensive equipment and a time-consuming collection and preparation of samples. This is why the methods are not used for continuous measurements, but only for sampling at larger intervals. To date, a reliable low-cost and in situ nitrate monitoring system is still not available and much needed [4,5].

A method that could eventually lead to such a monitoring system is electrical impedance spectroscopy (EIS). EIS is a powerful measurement technique used to measure the complex impedance of materials and systems over a range of frequencies. A small A-C signal stimulus is applied to a conductive electrode, thus obtaining a characteristic response that can be observed on the system surface. It has a high potential for fast, inexpensive, on-site measurements and is therefore used in many fields, such as material characterization [6], bioimpedance [7], battery diagnosis for state of health [8] and state of charge [9]. It is also very common for quantitative measurement of the ion concentrations in electrochemical

sensors [10,11]. Its potential applicability to the soil monitoring problem is justified by the fact that soil is a three-phase system, composed of the solid, liquid, and gaseous phases. This results in a dielectric constant, or relative permittivity, of between 2 and 14, depending on the soil. The liquid phase is an ion solution containing nutrients, such as nitrate, necessary for soil fertility and has a high static dielectric constant, which is strongly influenced by the ion concentration—the higher the ion concentration, the smaller the dielectric constant [12,13]. Water is a polar liquid with high dielectric relaxation and therefore has a high static dielectric constant of 81 at room temperature. With the increase in ion concentration, the ions orient the water molecules around them, thereby reducing the dielectric constant to values as low as 40 by a local high-field effect [13].

In an electric A–C field, energy is absorbed and stored in soil solution mixtures. The effect of energy storage is connected with the real part  $\epsilon'(\omega)$  of the permittivity, while the energy absorption is connected with the imaginary part  $\epsilon''(\omega)$  of the permittivity. Here,  $\omega$  denotes the angular frequency.

Figure 1 shows a schematic representation of dielectric losses in soil as a function of frequency [14]. Various effects contribute to the complex dielectric constant. It is obvious that the ionic conductivity and the double-layer effect play major roles in the imaginary part of the soil permittivity at lower frequencies. Both effects are strongly correlated to the ion concentration [15]. At the same time the double-layer effect leads to a frequency-dependence, which also influences the real part of the permittivity [15].



**Figure 1.** Contributions to the imaginary part  $\epsilon''_r = \epsilon'' / \epsilon_0$  of the relative permittivity of soil (after [14]). B: bound water relaxation; MW: Maxwell–Wagner effect; S: surface conductivity; X: crystal water relaxation.

When there is more than one ion species present in the soil, they all contribute to the impedance spectrum. The mixture problem is always a challenge in chemical sensing [16]. Here we do not deal with it, but assume that only nitrate ions, if any, are present in the soil. In the long run, one will want to remove this simplification, but in a first step it may be justified by the fact that by far the biggest soil problem is precisely nitrate ions and not any other ions [17].

With this assumption, in principle, these relationships in Figure 1 allow one to estimate nitrate concentrations in soil indirectly from measured impedance spectra of soil samples. In practice, however, there is no one-to-one mapping between given physical causes and single features of an impedance spectrum. There exists a vast literature on physics-based models of impedance spectra, mostly relying on equivalent circuits [18–20]. The spectrum of a distributed system (with its uncountably infinite number of degrees of freedom) can only be approximated by a lumped-element circuit (with its finite number of degrees of freedom). The approximation may be made arbitrarily good by increasing the number of circuit elements. Alternatively, one can use elements with frequency-varying characteristics, such as Warburg impedances [21] or constant-phase elements [22]. These approaches

exacerbate the one-to-one mapping problem and usually do not allow for predicting the spectral response to physical changes in the system, such as geometry variations or analyte concentration changes. This remark applies even more to mere a-posteriori description methods, such as DRT [23,24] or to situations in which the entire spectrum cannot be measured for practical reasons (e.g., when measurement times are limited to perform a fast measurement such that small-frequency responses cannot be observed). However, even in the absence of a one-to-one relationship between a nitrate concentration and a single feature of a measured impedance spectrum, the desired information is somehow hidden in the spectrum and entangled with other pieces of information. EIS was used in [25] together with effective-media models to estimate nitrate concentrations in soils. The results demonstrate the potential of the approach. However, model-based methods rely on soil properties and ambient conditions, and the individual influences of water, nitrate, and soil composition still remain unclear. For this reason, the resulting error is less than satisfactory and one cannot state that the method has been fully established. In field environments, other than in laboratories, it is impossible to measure all relevant influence quantities or to isolate the measurement system from all the quantities. For reasons of effort, it is also impractical to describe all relevant effects. Such problems are best handled by machine learning approaches [26].

In this work, machine learning and neural networks are used to extract information from impedance spectra to explore how well they are able to untangle the information mess. This is consistent with the reasoning that we only want to gain information on the nitrate concentration in the soil, despite the many influence effects, but do not require physical cause-and-effect relationships (which, incidentally, are also not provided by the majority of parametrized models, such as equivalent circuits).

To be more concrete, we consider long short-term memory (LSTM), a variant of recurrent artificial neural networks (RNN) [27], to estimate nitrate concentrations in soils based on measured impedance data. LSTM RNNs are particularly useful to cope with long-term dependencies between different data points in time series (hence their name) [27,28]. They can detect patterns in a time series, such as a voice recording, because they not only consider the data values but also include the positions of the data points in the series in their prediction. There is no reason why the method should be restricted to time series and should not be applicable to arbitrary data sequences, such as impedance spectra. In fact, LSTM RNNs have been applied repeatedly to EIS data, e.g., to predict fuel-cell degradation [29] and estimate the state of healthy lithium-ion batteries [30].

Often, it is not clear a priori which measurement data features are best passed to a neural network to optimize its prediction accuracy. Before training the network (here: the LSTM), the most suitable feature subset must be identified by a feature selection algorithm. The resulting removal of redundant and irrelevant data facilitates the design and the training of the neural network without incurring much loss of information [31]. Redundant data may lead to overfitting and irrelevant data may cause long training times. In many studies [32–34] LSTM RNNs are combined with different feature selection algorithms. In [32], LSTM RNNs combined with random forest as feature selection methods (cf. below) are used to predict the battery degradation. In [33] an improved forward selection using light gradient boosting machines selected the optimal feature set and then LSTM RNNs served to predict photovoltaic power generation. In [34] the features are selected by Pearson's correlation method with maximum relevancy and minimum redundancy as criteria and then the selected features are passed to LSTM RNNs to predict the short-term energy consumption of customers.

All these feature selection methods treat the series data first as discrete input data and then estimate the importance of each feature. The dependency between data points in the series is not considered. The selected features are delivered as sequence of data to LSTM RNNs in which the dependencies between data points play an important role [27]. Therefore, the feature selection methods mentioned above may not be optimal for LSTM RNNs. In this work, a new feature selection method, combining sequential forward selection and

LSTM, is developed, which processes variable-length sequences of inputs to evaluate the relative importance of features in the measured data (electrical impedance spectra, soil properties, and ambient conditions). With selected features the hyperparameters of LSTM RNNs were tuned by Bayesian hyperparameter optimization and an optimized regression model was built to estimate the nitrate concentration in sandy soils.

To validate the performance of this new method and compare the evaluation result of nitrate concentration in synthetic soils with LSTM, feature selection and evaluation with random forest (RF) is applied for comparison, all in the context of processing and interpreting impedance spectra of soil samples.

RF is a known ensemble learning method that constructs a multitude of decision trees independently at training time. In particular, each RF decision tree is trained using a bootstrap set, which is randomly sampled from the input data with replacement (here the input data are impedance spectra and measured values of other quantities, such as soil density). In this way, about two thirds of the instances in the bootstrap set are unique and used to build a decision tree, and about one third are left out as out-of-bag (OOB) instances used to validate the decision tree [35]. The ensemble produces a forest of trees and outputs of all trees are aggregated to produce a final prediction. In classification problems, the final prediction is voted by the majority of trees. In regression, the output is the average of the individual tree predictions.

As all decision trees are built over a random extraction of the instances from the database and a random extraction of the variables, RF is also widely used to measure variable importance. A suitable metric for this is the mean decrease in impurity (MDI), also known as Gini importance, defined as the sum of all decreases in Gini impurity [36] due to a given variable when this variable is used to form a split in the RF. Another metric is the permutation importance which quantifies the variable importance based on the reduction in model accuracy when permuting the values of a variable included in a subset of variables that were split to grow the tree. Both metrics have been discussed and compared extensively [37–39]. The mechanism of MDI is biased, i.e., it tends to inflate the importance of continuous or correlated variables [37,39]. When the data set is too small, permutation importance is also unstable and may lead to a wrong ranking of variable importance [38]. Thus, the metric should be chosen carefully according to the measurement problem. It is not clear a priori which metric is most suitable for the selection of features from impedance spectra.

Another advantage of RF is the ability to measure variable association, which is an indication of correspondence between variables. Variable correspondence can increase the standard errors of a machine learning model [40]. It is an essential step to recognize the relationship between variables for establishing a robust and accurate machine learning model. RF is a tree-based algorithm and is able to use surrogate splits to measure the association between variables. Such a surrogate split is an alternative to the optimal split at a given node in a decision tree. It uses a similar or correlated variable and split criterion to replace the optimal split variable and split criterion [41].

To the best of the authors' knowledge, this work enhances the state of the art in the following respects:

1. The method of measuring nitrate concentration in soil based on the combination of EIS with LSTM networks is characterized for the first time.
2. The advantage of a new feature selection method over conventional methods, especially random forest, is demonstrated for the first time, the advantage being a higher coefficient of determination between the actual nitrate concentration and the concentration predicted by the LSTM network fed with the selected features.
3. The results presented improve the quantitative understanding of the potential of the described approach in the context of the low-cost, robust, fast, in situ, and soil property-independent measurement of nitrate concentrations in soils.

## 2. Materials and Methods

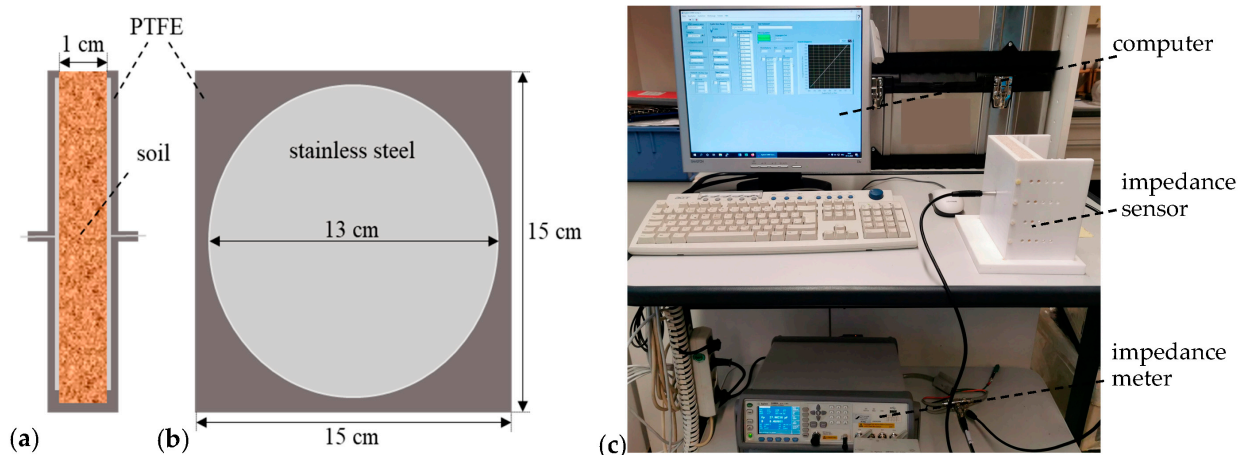
As material under test (MUT), dry quartz sand from Karl Inzelsberger Sand and Kaolinwerk in Creußen, Germany, prepared with six different nitrate-N concentrations, viz., 0, 2, 4, 6, 8, and 10 mg/kg, was used as synthetic soil. These nitrate-N concentrations are below the recommended levels of 20 to 40 mg/kg in soils depending on crop [42]. A total of 500 g of pure quartz sand was enriched in a beaker with sodium nitrate solutions prepared according to international standards [2]. After mixing and drying, the sand–nitrate mixture was examined at different water contents (the deionized water concentrations were between 0 and 150 g/kg). The water content here is the ratio of the mass of water and mass of dry mixture. Dry mixture mass is expressed by the mass of mixture dried (up to the constant weight) at 105 °C. In agriculture, the mass water content is not used frequently but a necessary step for the measurement of volumetric water content with gravimetric methods, as will be seen later. In the field, the volumetric water and nitrate-N contents offer more information about the ratio of pores filled with water and make it easier to calculate the quantity of water and nitrate in an area. Thus, the above-mentioned water and nitrate-N mass contents were converted to volumetric contents by:

$$\theta_W = \frac{m_W}{V} = \frac{m_S \cdot \omega_W}{V} = \rho_S \cdot \omega_W \quad (1)$$

$$\theta_N = \frac{m_N}{V} = \frac{m_S \cdot \omega_N}{V} = \rho_S \cdot \omega_N \quad (2)$$

where  $\theta_W$  and  $\theta_N$ , respectively, are the volumetric water and nitrate-N contents,  $\omega_W$  and  $\omega_N$  are the corresponding mass contents,  $m_W$ ,  $m_N$  and  $m_S$  are the masses of water, nitrate-N, and sand in the measurement volume  $V$ , and  $\rho_S$  is the density of dry sand in the measuring cell. Obviously, even when the mass content remains constant, the vol–metric content can vary.

To investigate various synthetic soil mixtures, the measuring cell shown in Figure 2 was used. As polytetrafluoroethylene (PTFE) is a high-temperature-resistant non-wetting material and stainless steel is stable to nitrate solutions, the measuring cell consists of two opposing square plates made of PTFE with a side length of 15 cm and two circular stainless-steel electrodes with a diameter of 13 cm centered in them, forming a plate capacitor with an electrode spacing of 1 cm. Both plates are removable so they can be easily cleaned after each measurement of the previous sample. The MUTs were filled into the space bit by bit between the electrodes until the material was flush with the top edge of the measuring cell. Care was taken to keep the bulk density as planned.

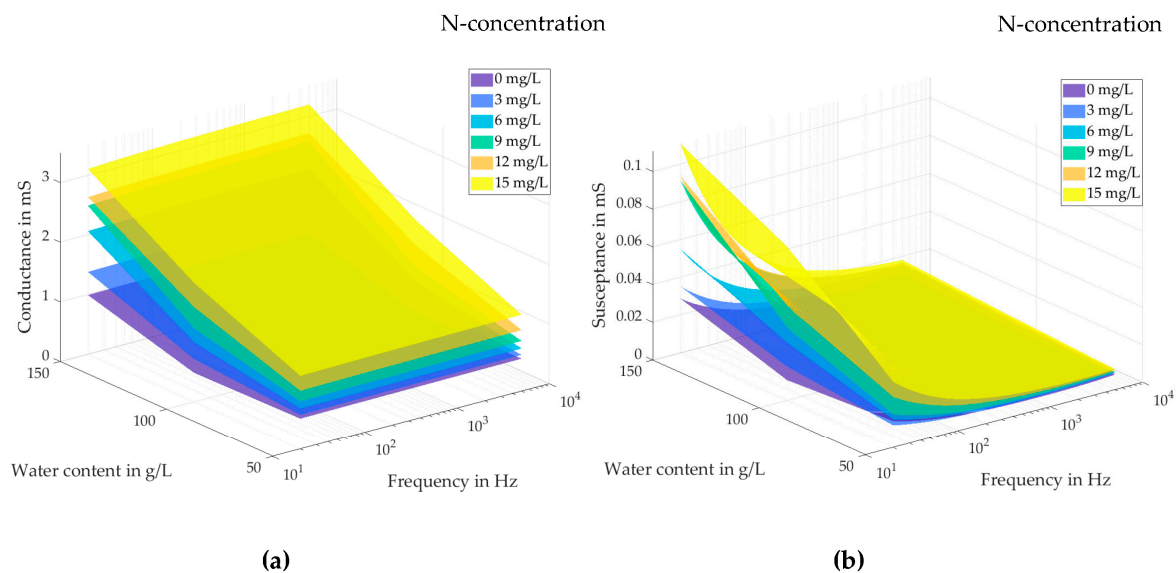


**Figure 2.** Parallel-plate impedance sensor. (a) Cross section. (b) View of the right plate from the left. (c) Photograph of the setup with impedance meter and computer.



The impedance  $\underline{Z} = |\underline{Z}|e^{j\varphi} = R + jX$  was measured over a frequency range from 20 Hz to 5 kHz using an Agilent E4980A impedance meter (98 measurement points per frequency sweep). For electrochemical systems, lower frequencies (as low as 1 mHz) would provide more information but would also take more time for the measurement. Therefore, as we are interested in fast measurements over larger areas (agricultural fields), we restricted the frequency to values above 20 Hz. At each frequency, both the apparent impedance  $|\underline{Z}|$  and the impedance phase  $\varphi$  were recorded and then evaluated in Matlab R2021a. Each MUT was repeatedly measured 20 times.

Figure 3a shows the measured frequency-dependent conductance  $G = \text{Re}\{\underline{Y}\} = \text{Re}\{1/\underline{Z}\}$  as a function of nitrate-N and water concentrations.  $G$  increases with increasing nitrate-N and water contents, the nitrate-N sensitivity being significantly greater than the moisture sensitivity. This demonstrates the potential for monitoring the nitrate-N concentration in soil with water as influence quantity, but also emphasizes the need for prior calibration.



**Figure 3.** Measured test cell conductance (a) and susceptance (b) as a function of the nitrate-N concentration (in mg/L) and of the water content (in g/L) in synthetic soil (quartz sand), and of frequency, reprinted from Ref. [43].

Likewise, Figure 3b shows the measured frequency-dependent susceptance  $B = \text{Im}\{\underline{Y}\} = \text{Im}\{1/\underline{Z}\}$  as a function of nitrate-N and water concentrations. In the low-frequency region from 20 Hz to 1 kHz,  $B$  increases with increasing nitrate-N and water contents. This can be explained by the electrical double-layer effect, which is strongly related to the ion concentration [15]. It may provide a possibility to monitor the nitrate-N concentration at low frequencies.

The data in Figure 3 suggest that the designed impedance sensor can measure the nitrate-N concentration, possibly with prior calibration and possibly in the field (rather than the laboratory). As calibrations to take into account specific soil types and water content are likely to require too much time and effort, we looked for ways to remove the need for calibration. The reasoning is about the following: impedance spectra contain an abundance of signal features. This richness of information should allow one to estimate both analyte concentrations and influence quantity values (soil density, water content, etc.). The estimation must be automated, which calls for a machine learning algorithm.

Before the analysis using machine learning, it is an important step to prepare data, which involves reformatting data, making corrections to data and enriching data. The raw data sampled from our experiments consist of measured impedances  $\underline{Z}(f)$  (2520 spectra from 126 synthetic soils compositions, each spectrum comprising 98 discrete frequencies)

and physical properties of soils for each sand composition (density  $\rho_b$ , volumetric water content  $\theta_w$ ). The measured impedance was transformed into conductance  $G$  and susceptance  $B$  to provide a better view of the electrochemical changes in the measuring cell. As shown in Figure 3,  $G$  and  $B$  are functions of frequency and the derivatives of them may contain useful information. Therefore discrete approximations of  $dB/df$ ,  $dG/df$ ,  $d^2B/df^2$ , and  $d^2G/df^2$  as additional features are added to test whether they can facilitate the machine learning. As LSTM RNN process sequences of data, the features passed to it were  $G$  and  $B$  (vectors of length 98),  $\rho_b \cdot a$  and  $\theta_w \cdot a$  (where  $a = (1, 1, \dots, 1)^T$  is a size-98 column vector of 1s), and the derivatives just discussed (also padded to a length of 98). The padding of the scalars or shorter vectors to the length of the impedance spectrum (98) does not add information but is needed to pass data to all neurons of the RNN input layer. The features passed to RF were frequency-discrete. Before passing the features, they were centered and scaled to have zero mean and unity standard deviation.

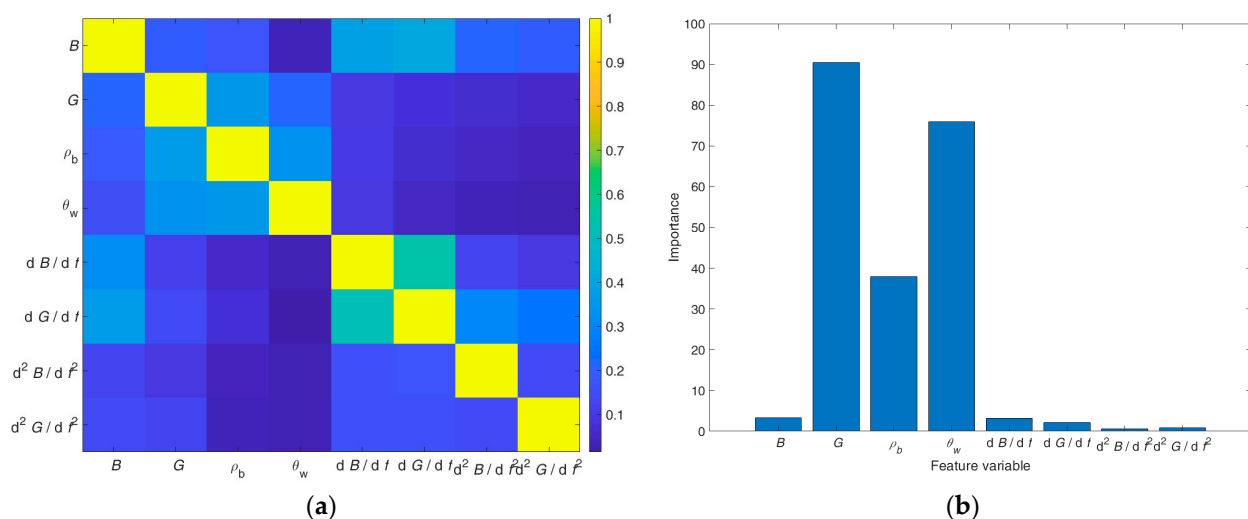
### 3. Results

#### 3.1. Feature Selection

In the following, feature selection by RF and by a newly developed method based on LSTM is evaluated with respect to the requirements for nitrate concentration measurement.

##### 3.1.1. Feature Selection by Random Forest

RF provides estimates both of the association between various features and of the feature importance. The association and importance of the eight features available in this work are respectively presented in Figure 4a,b. The importance was calculated using permutation importance as metric.



**Figure 4.** Feature association estimation (a) and importance estimation (b) using RF, reprinted from Ref. [44].

Figure 4a confirms that the association between distinct features is small, which indicates that the features are independent. The bar graph in Figure 4b reveals that the conductance  $G$  is the most important feature, followed by the volumetric water content  $\theta_w$  and the soil density  $\rho_b$ . Other features have but a negligible effect on the nitrate concentration estimation. Similar results have been obtained with the sequential feature selection method combined with a support vector machine.

##### 3.1.2. Feature Selection by Sequential Forward Selection (SFS) Based on LSTM

Since feature selection with RF does not consider the dependencies between different frequency points, it may not provide the optimal result when training LSTM RNN. We

therefore developed a new feature selection method combining SFS with LSTM to select proper features of impedance spectra and their derivatives.

SFS selects features sequentially based on a criterion to be minimized over all feasible feature subsets. In our case, the validation root mean square error (RMSE) of the LSTM network output was used as criterion. In SFS the features are sequentially added to an empty candidate set until the addition of further features no longer improves (i.e., decreases) the RMSE by more than a preselected minimum value, set to 0.1 mg/L in the given context.

The difficulty of using the RMSE of the LSTM output as a criterion for SFS is that the optimum hyperparameters of the LSTM network vary with the feature subsets. To solve this problem, Bayesian hyperparameter optimization was used to find out the optimum hyperparameters for the LSTM RNNs of all feature subsets. Bayesian optimization calculates the optimum hyperparameters in a specified range of values and uses a metric (in our case, the RMSE of the LSTM RNNs) to optimize. In this way, not only the optimum hyperparameters are determined, but also the RMSE for the SFS process. For every subset, an LSTM RNN with optimized hyperparameters and minimized RMSE was trained. The one feature that decreases the RMSE most at each step is retained in the further process.

The main structure of an LSTM RNN is shown in Figure 5. A dropout layer after each LSTM layer is used to avoid overfitting. The dense layers are to reduce the dimension of the output from previous layers to provide a smooth transition from the hidden layers to the output layer. The ReLU layer just before the output ensures a non-negative output.

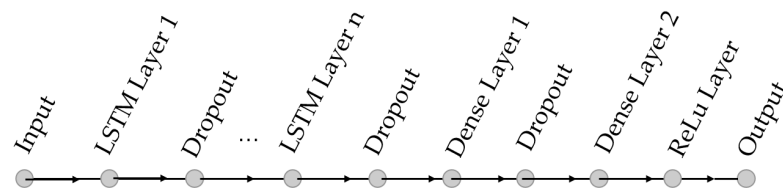


Figure 5. LSTM network structure.

The hyperparameters listed in Table 1 were optimized inside predefined intervals. The feature selection results with SFS based on LSTM, the optimal hyperparameters for each step are presented in Table 2. The whole feature selection time for step 1 through step 4, shown in the table, took 32.8 h on a workstation with a 3.7 GHz quad core and a memory of 64 GB. In the first step, the susceptance  $B$  decreases the RMSE most and thus is the most important feature for a successful LSTM estimator. Adding the volumetric water content  $\theta_w$  decreases the RMSE more than the remaining features. Similarly, the third and fourth most important features are the conductance  $G$  and the soil density  $\rho_b$ . Adding other features had no appreciable effect on the target criterion. In particular, the first and second derivatives of  $G$  and  $B$  are not needed to estimate nitrate concentrations in soils.

Table 1. Hyperparameters to be optimized using Bayesian optimization with predefined parameter ranges.

No. of Hyperparameter	Description	Range	Type
1	Number of LSTM layers, $N$	1 ... 3	Integer
2	Numbers of neurons in each hidden layer	$(1 \dots 100) \times N$	Integer $\times N$
3	miniBatchSize <sup>1</sup>	2 ... 20	Integer
4	initialLearnRate <sup>1</sup>	$10^{-3} \dots 10^{-1}$	Real

<sup>1</sup> Nomenclature of MATLAB R2021a Deep Learning Toolbox.



**Table 2.** Results of feature selection by SFS with the validation RMSE of the LSTM output as criterion to be minimized.

Step	Features	Validation RMSE, in mg/L	Optimum Value of Hyperparameter No. (cf. Table 1)			
			1	2	3	4
1	$B$	2.1920	2	(25,75)	18	0.005
2	$B$ and $\theta_w$	1.2364	2	(60,12)	16	0.007
3	$B$ , $\theta_w$ , and $G$	0.6873	2	(89,99)	12	0.001
4	$B$ , $\theta_w$ , $G$ , and $\rho_b$	0.1431	2	(77,45)	16	0.001

### 3.1.3. Comparison

One notices that the feature selection results from RF on the one hand and SFS based on LSTM on the other hand are not compatible. The most important feature according to the latter, the susceptance  $B$ , had little or no effect according to the former. Conversely, the most important feature by RF, the conductance  $G$  (cf. Figure 4), is only ranked third by SFS. Both selection methods assessed the volumetric water content as crucial and soil density as less important. The reason for the differences in feature ranking and the consequences for the nitrate concentration measurement accuracy are discussed below.

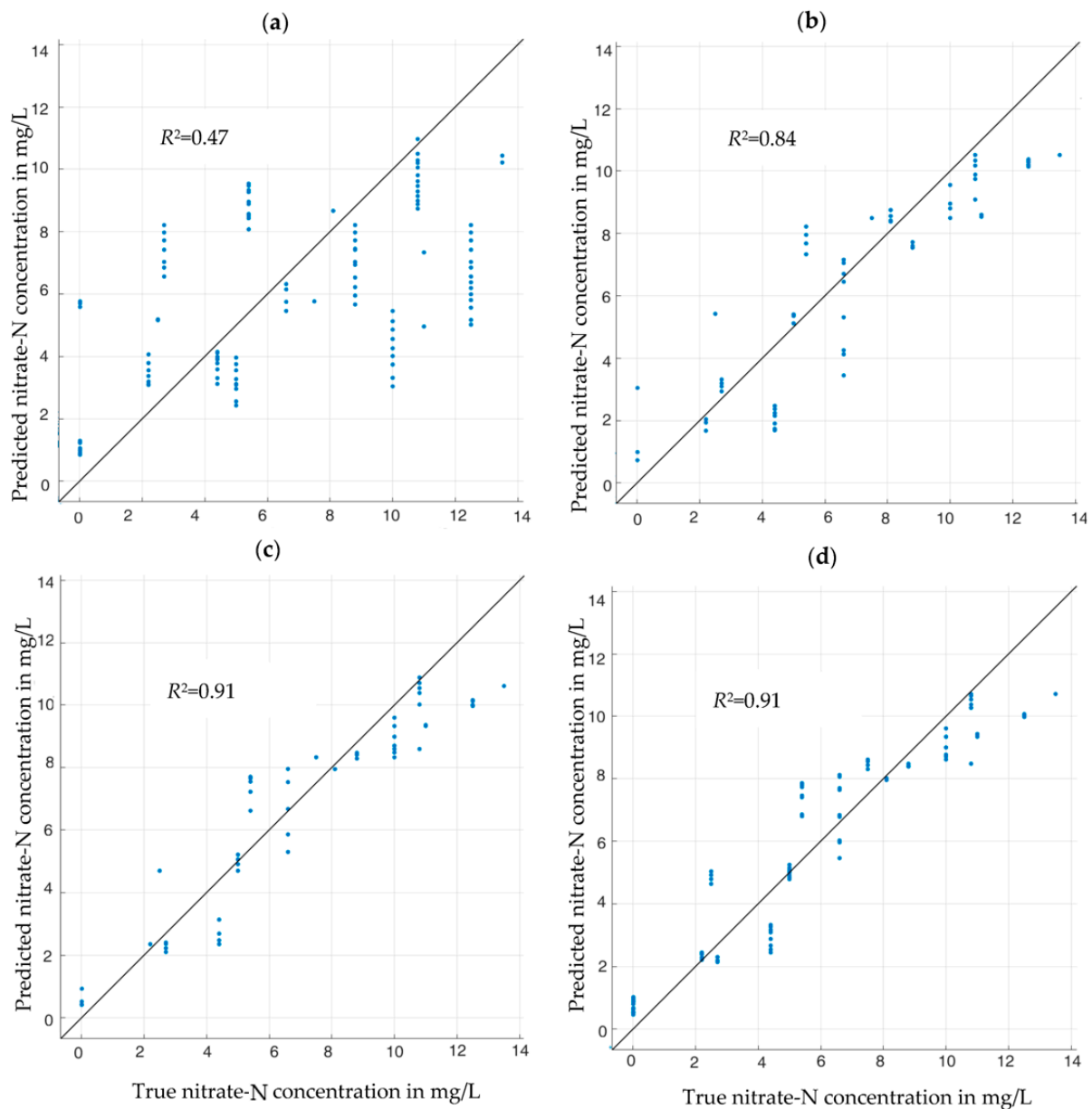
### 3.2. Experimental Results for Nitrate Concentration Measurement

In order to quantify the influence of the features on the estimation accuracy, the feature selection results were validated by the regression models from their own algorithms. The regression models built from RF take an input vector containing the measured quantities at one specific frequency as well as the numerical value of said frequency. When building regression models from RF and LSTM RNNs, our impedance measurements provided 2520 spectra. They were split into training and validation data in the ratio of 80:20. After the training, the models were tested using 360 spectra unused up to this point. Both algorithms took less than 10 min to train regression models and the inference times for both were around one second. The quality of the prediction results was judged by the coefficient of determination between true and estimated nitrate-N concentration, defined as:

$$R^2 = 1 - \frac{\sum_i (Y_{test-i} - Y_{pre-i})^2}{\sum_i (Y_{test-i} - \overline{Y_{test}})^2} \quad (3)$$

Here,  $i$  is the case index,  $Y_{test-i}$  and  $Y_{pre-i}$  are the actual and predicted nitrate-N concentrations of the test data, respectively.  $\overline{Y_{test}}$  is the mean of the actual concentration.

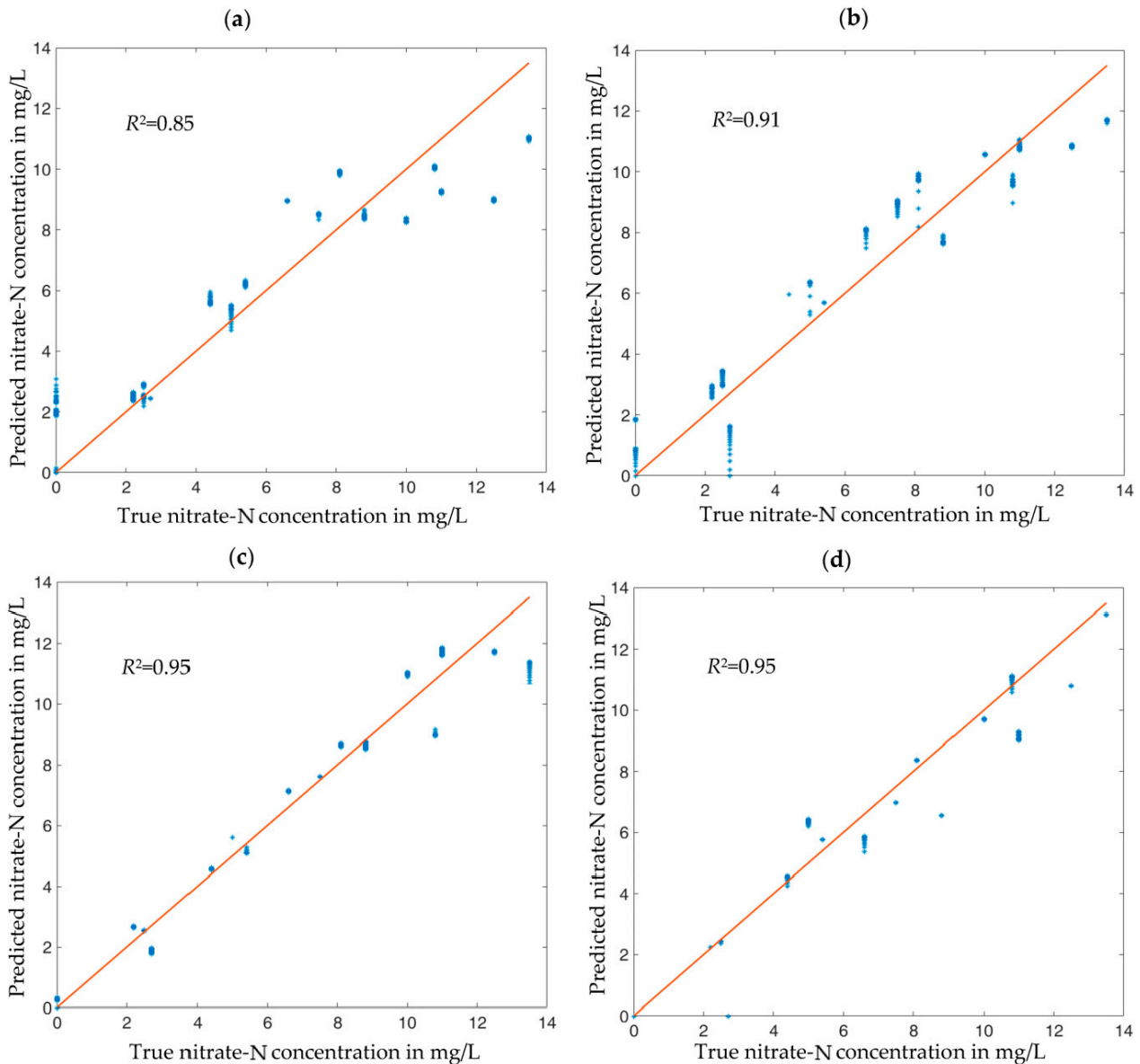
Figures 6 and 7 illustrate the quality of the nitrate-N concentration estimation by the various neural networks with different feature subsets. Using but one feature (the RF-favored conductance spectrum  $G$  or the SFS-favored susceptance spectrum  $B$ , respectively), one obtains results between rather unconvincing, Figure 6a, and quite good, Figure 7a. Adding other features step by step, as suggested by Figure 4b and Table 2, respectively, improved the estimation accuracy monotonically. In all cases, the LSTM estimator based on SFS performed better than the RF estimator, although the differences tend to become smaller, the more features are passed to the estimators. The optimum feature combination turned out to be  $B$ ,  $\theta_w$ , and  $G$  (in other words: the full complex impedance spectrum  $\underline{Z}$  enhanced by the additional information  $\theta_w$ ). This feature set led to an excellent measurement result with  $R^2 = 0.95$ , Figure 7c, which could not be improved further by adding more features, such as soil density  $\rho_b$ , Figure 7d. The additional features might prove helpful for the generalization of this approach to other soils, which was not part of this work.



**Figure 6.** Nitrate-N concentrations predicted by RF-built regression models from measured impedance spectra of synthetic soil samples plus selected other physical variables. (a) Results based on the conductance spectrum  $G$  as feature. (b) Results with features  $G$  and  $\theta_w$ . (c) Results with features  $G$ ,  $\theta_w$ , and  $\rho_b$ . (d) Results with features  $G$ ,  $\theta_w$ ,  $\rho_b$ , and  $B$ . Blue dots: estimation results for each of the 360 validation data sets. Straight lines: goal (predicted value = actual value). The  $R^2$  values shown in each chart are the coefficients of determination.

It is obvious that the RF-selected features are not the optimal choice for the LSTM RNN. In fact, a regression LSTM RNN trained with the top three RF-favored features ( $G$ ,  $\theta_w$ , and  $\rho_b$ ) resulted in a coefficient of dermination of only  $R^2 = 0.82$  instead of  $R^2 = 0.95$ . Both feature selection methods work well with regression models based on their own algorithms, but cannot be used interchangeably. This implies that a feature selection method based on point data, such as RF, may not be optimal for regression models based on vector-valued input data with dependencies between the vector elements (time or frequency series). With respect to our impedances, such a behavior is already noticeable in Figure 3. There is an unambiguous relationship between the measured conductance  $G$  and the nitrate concentration at all frequencies, Figure 3a. In contrast, the susceptance  $B$  depends less

and less on nitrate concentration with increasing frequency (and in practice, measurement errors would prevent one from determining nitrate concentration from high-frequency susceptance). However, the LSTM method takes the whole frequency spectrum as input, and as  $B(f)$  depends more strongly on  $f$  than  $G(f)$  and thus contains more information as sequence input, LSTM can estimate the nitrate concentration better from  $B$  than from  $G$  and better than the RF-based regression model.



**Figure 7.** Nitrate-N concentrations predicted by LSTM-built regression models from measured impedance spectra of synthetic soil samples plus selected other physical variables. (a) Results based on the susceptance spectrum  $B$  as feature. (b) Results with features  $B$  and  $\theta_w$ . (c) Results with features  $B$ ,  $\theta_w$ , and  $\rho_b$ . (d) Results with features  $B$ ,  $\theta_w$ ,  $\rho_b$ , and  $G$ . Blue dots: estimation results for each of the 360 validation data sets. Straight lines: goal (predicted value = actual value). The  $R^2$  values shown in each chart are the coefficients of determination between true and estimated concentrations.

### 3.3. Literature Comparison

As discussed in the introduction, the result in [25] using EIS with effective-media models is less satisfying because of the dependency on soil conditions. In addition, the measuring range of nitrate-N is from 14 mg/L to 70 mg/L which can hardly meet the

field requirements. Another attempt was made to monitor the nitrate concentration in soils in situ without pretreatment using spectrophotometry in [45]. This system is strongly dependent on the water content in the soil and is not able to measure the nitrate over large area. Moreover, the measuring range of nitrate-N used in [45] is from 0 to 500 ppm, which is far beyond the field requirements. Through comparison, our work shows the best possibilities to overcome these problems. We hope to develop a system based on it that can be mounted on a tractor and allow measurements across large agriculture areas.

#### 4. Conclusions

Based on synthetic soil samples, we have demonstrated that machine learning methods can extract nitrate-N concentrations in soil from measured impedance spectra enhanced by selected additional physical variables with errors that are compatible with field requirements, and the demonstrated measurement accuracy was far below the recommended nitrate-N concentration in soils.

It was also shown that proper feature selection yields objective figures of merit for the impact of different features. Based on this, one can design RNNs that estimate the nitrate concentration in soils robustly and accurately against the influence of water content and soil density. Feature selection by SFS based on LSTM improved the nitrate concentration estimation accuracy of an LSTM RNN to  $R^2 = 0.95$ , which is better than the values obtained by more common feature selection. This result was obtained when the complete complex impedance spectrum and the water content of the soil were passed as input data to the neural network. Adding the soil density as additional input data did not improve the prediction quality any further. Therefore, in practice, soil density is not expected to influence field measurements significantly.

Of course, in the field one will be faced with many types of soil (not just sandy as in our experiments) and with additional influence quantities. Our further goal is to conduct more experiments to expand the measuring range and include more soil types, such as potting soils, brown soils, and clay mixtures. In this way, we hope to implement a method based on machine learning that is capable of measuring nitrate-N concentrations virtually independently of soil properties and types. True field measurements cannot rely on bulky and expensive laboratory equipment but will require application-specific printed-circuit-board impedance meters.

**Author Contributions:** Conceptualization, X.M. and G.F.; Methodology, X.M.; Software, X.M.; Validation, X.M. and G.F.; Formal analysis, X.M., L.B. and G.F.; Investigation, X.M.; Resources, X.M. and L.B.; Data curation, X.M.; Writing—original draft preparation, X.M.; Writing—review and editing, G.F.; Visualization, X.M.; Supervision, G.F.; Project administration, G.F.; Funding acquisition, L.B. and G.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Volkswagen Foundation (Grant no. Az. 9A220). Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—491183248. Funded by the Open Access Publishing Fund of the University of Bayreuth.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are openly available at [https://doi.org/10.15495/do\\_ubt\\_2067](https://doi.org/10.15495/do_ubt_2067), accessed on 5 February 2023, provided by the authors.

**Conflicts of Interest:** The authors declare no conflict of interest.

#### References

1. Jakobs, I.; Grimm, F.; Keppner, L.; Hilliges, F. *Nitratbericht 2020: Gemeinsamer Bericht der Bundesministerien für Umwelt, Naturschutz und nukleare Sicherheit sowie für Ernährung und Landwirtschaft* (= Nitrate Report 2020: Joint Report of the Federal Ministries for [ . . . ]); German Ministry for Food and Agriculture and German Ministry for the Environment, Nature Conservation and Nuclear Safety: Bonn, Germany, 2020. (In German)

2. EN ISO 13395; 1996-07: Water Quality–Determination of Nitrite Nitrogen and Nitrate Nitrogen and the Sum of Both by Flow Analysis (CFA and FIA) and Spectrometric Detection. CEN: Brussels, Belgium, 1996.
3. Narayana, B.; Sunil, K. A spectrophotometric method for the determination of nitrite and nitrate. *Eurasian J. Anal. Chem.* **2009**, *4*, 204–215.
4. Subhash, C.; Malik, A.; Zargar, M.Y.; Bhat, M.A. Nitrate Pollution: A menace to human, soil, water and Plant. *Univ. J. Environ. Res. Technol.* **2011**, *1*, 22–32.
5. Ilsemann, J.; Goeb, S.; Bachmann, J. How many soil samples are necessary to obtain a reliable estimate of mean nitrate concentration in an agricultural field? *J. Plant Nutrition Soil Sci.* **2001**, *164*, 585–590. [[CrossRef](#)]
6. Bifano, L.; Fischerauer, A.; Liedtke, A.; Fischerauer, G. Characterization of sand and sand-binder systems from the foundry industry with electrical impedance spectroscopy. *J. Sens. Sens. Syst.* **2021**, *10*, 43–51. [[CrossRef](#)]
7. Xu, Y.; Li, C.; Jiang, Y.; Guo, M.; Yang, Y.; Yang, Y.; Yu, Y. Electrical Impedance Spectroscopic Detection of E. *Coils with Machine Learning. J. Electr. Soc.* **2020**, *167*, 047508. [[CrossRef](#)]
8. Bercibar, M.; Gandiaga, I.; Villarreal, I.; Omar, N.; Van Mierlo, J.; Van den Bossche, P. Critical review of state of health estimation methods of Li-ion batteries for real applications. *Renew. Sustain. Energy Rev.* **2016**, *56*, 572–587. [[CrossRef](#)]
9. Itagaki, M.; Honda, K.; Hoshi, Y.; Shitanda, I. In-situ EIS to determine impedance spectra of lithium-ion rechargeable batteries during charge and discharge cycle. *J. Electroanal. Chem.* **2015**, *737*, 78–84. [[CrossRef](#)]
10. De Beer, D.J.; Joubert, T.-H. Validation of Low-Cost Impedance Analyzer via Nitrate Detection. *Sensors* **2021**, *21*, 6695. [[CrossRef](#)] [[PubMed](#)]
11. Grossi, M.; Parolin, C.; Vitali, B.; Riccò, B. Electrical Impedance Spectroscopy (EIS) characterization of saline solutions with a low-cost portable measurement system. *Eng. Sci. Technol.* **2019**, *22*, 102–108. [[CrossRef](#)]
12. Kanoun, O.; Tetyuev, A.; Tränkler, H. Boden feuchtemessung mittels Impedanzspektroskopie (= Soil moisture measurement with impedance spectroscopy). *TM–Tech. Mess.* **2004**, *71*, 475–485. (In German) [[CrossRef](#)]
13. Gavish, N.; Promislow, K. Dependence of the dielectric constant of electrolyte solutions on ionic concentration: A micro field approach. *Phys. Rev. E.* **2016**, *94*, 012611. [[CrossRef](#)] [[PubMed](#)]
14. Hasted, J.B. *Aqueous Dielectrics*; Chapman and Hall: London, UK, 1973; pp. 238–252.
15. Ishai, P.B.; Talary, M.S.; Caduff, A.; Levy, E.; Feldman, Y. Electrode polarization in dielectric measurements: A review. *Meas. Sci. Technol.* **2013**, *24*, 102001. [[CrossRef](#)]
16. Kumaragamage, D.; Warren, J.; Spiers, G. Soil Chemistry. In *Digging into Canadian Soils: An Introduction to Soil Science*; Krzic, M., Walley, F.L., Diochon, A., Paré, M.C., Farrell, R.E., Eds.; Canadian Soc. Soil Sci.: Pinawa, MB, Canada, 2021; pp. 150–190.
17. Rockström, J.; Steffen, W.; Noone, K.; Persson, Å.; Chapin, F.S.; Lambin, E.F.; Lenton, T.M.; Scheffer, M.; Folke, C.; Schellnhuber, H.J.; et al. A safe operating space for humanity. *Nature* **2009**, *461*, 472–475. [[CrossRef](#)] [[PubMed](#)]
18. Zoltán, L.; Tamás, K. A generalized model of the equivalent circuits in the electrochemical impedance spectroscopy. *Electrochim. Acta* **2020**, *363*, 137199. [[CrossRef](#)]
19. Schönleber, M.; Klotz, D.; Ivers-Tiffée, E. A Method for Improving the Robustness of linear Kramers-Kronig Validity Tests. *Electrochim. Acta* **2014**, *131*, 20–27. [[CrossRef](#)]
20. Gao, P.; Zhang, C.; Wen, G. Equivalent circuit model analysis on electrochemical impedance spectroscopy of lithium metal batteries. *J. Pow. Sourc.* **2015**, *294*, 67–74. [[CrossRef](#)]
21. Taylor, S.R.; Gileadi, E. Physical Interpretation of the Warburg Impedance. *Corrosion* **1995**, *51*, 664–671. [[CrossRef](#)]
22. Brug, G.J.; van den Eeden, A.L.G.; Sluyters-Rehbach, M.; Sluyters, J.H. The analysis of electrode impedances complicated by the presence of a constant phase element. *J. Electroanal. Chem.* **1984**, *176*, 275–295. [[CrossRef](#)]
23. Nohl, M.; Raut, G.; Wolf, S.E.; Duyster, T.; Dittrich, L.; Vinke, I.C.; Eichel, R.A.; De Haart, L.G.J. Analysis of the DRT as Evaluation of EIS Data Analysis. *ECS-Trans.* **2021**, *103*, 1403. [[CrossRef](#)]
24. Xia, J.; Wang, C.; Wang, X.; Bi, L.; Zhang, Y. A perspective on DRT applications for the analysis of solid oxide cell electrodes. *Electrochim. Acta* **2020**, *349*, 136328. [[CrossRef](#)]
25. Pandey, G.; Kumar, R.; Weber, R.J. Real time detection of soil moisture and nitrates using on-board in situ impedance spectroscopy. In Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, 13–16 October 2013; pp. 1081–1086. [[CrossRef](#)]
26. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. *Sensors* **2018**, *18*, 2674. [[CrossRef](#)]
27. Hochreiter, S.; Schmidhuber, J. LSTM can solve hard long time lag problems. In Proceedings of the 10th Advances in Neural Information Processing Systems, Denver, CO, USA, 2–5 December 1996; pp. 473–479.
28. Houdt, G.V.; Mosquera, C.; Nápoles, G. A Review on the long short-term memory model. *Artificial Intell. Rev.* **2020**, *54*, 5929–5955. [[CrossRef](#)]
29. Caponetto, R.; Guarnera, N.; Matera, F.; Privitera, E.; Xibilia, M.G. Application of Electrochemical Impedance Spectroscopy for prediction of Fuel Celldegradation by LSTM neural networks. In Proceedings of the 2021 29th Mediterranean Conference on Control and Automation (MED), PUGLIA, Italy, 22–25 June 2021; pp. 1064–1069. [[CrossRef](#)]
30. Li, D.; Yang, D.; Li, L.; Wang, L.; Wang, K. Electrochemical Impedance Spectroscopy Based on the State of Health Estimation for Lithium-Ion Batteries. *Energies* **2022**, *15*, 6665. [[CrossRef](#)]



31. Kratsios, A.; Hyndman, C. NEU: A Meta-Algorithm for Universal UAP-Invariant Feature Representation. *J. Mach. Learn. Res.* **2021**, *22*, 4102–4152. [[CrossRef](#)]
32. Sin, S.; Cho, S.; Lee, P.; Abbas, M.; Lee, S.; Kim, J. Data-driven prediction of battery degradation using EIS-based robust features. In Proceedings of the 2022 IEEE Energy Conversion Congress and Exposition (ECCE), Detroit, MI, USA, 9–13 October 2022; pp. 1–5. [[CrossRef](#)]
33. Tao, C.; Lu, J.; Lang, J.; Peng, X.; Cheng, K.; Duan, S. Short-Term Forecasting of Photovoltaic Generation Based on Feature Selection and Bias Compensation-LSTM Network. *Energies* **2021**, *14*, 3086. [[CrossRef](#)]
34. Sun, P.; Jiang, C.; Wang, X.; Yang, X. Short-Term Building Load Forecast Based on a Data-Mining Feature Selection and LSTM-RNN Method. *IEEE Trans. Electr. Electr. Eng.* **2020**, *15*, 1002–1010. [[CrossRef](#)]
35. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
36. Menze, B.H.; Kelm, B.M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F.A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform.* **2009**, *10*, 213. [[CrossRef](#)]
37. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [[CrossRef](#)]
38. Luz Calle, M.; Urrea, V. Letter to the Editor: Stability of Random Forest importance measures. *Brief. Bioinform.* **2011**, *12*, 86–89. [[CrossRef](#)]
39. Nicodemus, K. Letter to the Editor: On the Stability and ranking of predictors from random forest variable importance. *Brief. Bioinform.* **2011**, *12*, 369–373. [[CrossRef](#)] [[PubMed](#)]
40. Daoud, J.I. Multicollinearity and Regression Analysis. *J. Phys. Conf. Ser.* **2017**, *949*, 012009. [[CrossRef](#)]
41. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth: Belmont, CA, USA, 1984; Chapter 5; pp. 140–149.
42. Bagshaw, D.; Moody, J.; Pattison, T. *Soil health for vegetable production in Australia*; Queensland Government: Queensland, Australia, 2010; p. 27.
43. Ma, X.; Bifano, L.; Oehme, M.; Fischerauer, G. *Validation of the Suitability of Electrical Impedance Spectroscopy for the In-Situ Monitoring of Nitrate in Sandy Soil*. ITG-Fachbericht 303 (= Proc. 21. GMA/ITG-Fachtagung Sensoren und Messsysteme, Nuremberg, Germany, May 10–11, 2022); VDE Verlag: Berlin, Germany, 2022; pp. 352–355.
44. Ma, X.; Fischerauer, G. Investigation of long short-term memory artificial neural networks as estimators of nitrate concentrations in soils from measured electrical impedance. In Proceedings of the 2022 International Workshop on Impedance Spectroscopy (IWIS), Chemnitz, Germany, 27–30 September 2022; pp. 117–120. [[CrossRef](#)]
45. Yeshno, E.; Arnon, S.; Dahan, O. Real-time monitoring of nitrate in soils as a key for optimization of agricultural productivity and prevention of groundwater pollution. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 3997–4010. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.