RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt26328

Identifying Soil Profiles and Groundwater Quality using Machine Learning Models with and without Clustering Methods

Khabeer Al-Awad [a], Hayder Algretawee [b, c, d], Alaa M Shaban [a, *]

[a] Department of Civil Engineering, Engineering College, University of Kerbala, Karbala, Iraq.
[b] Department of Civil Engineering, College of Engineering, Al-Karkh University of Science, Baghdad, Iraq.
[c] College of Engineering, Al-Naji University, Baghdad, Iraq.
[d] Department of Hydraulic Structures and Water Resources, College of Civil Engineering, University of Technology, Iraq.
* Corresponding Author Email: alaa.shaban@uokerbala.edu.iq

DOI: https://doi.org/10.54392/irjmt26328
Received: 28-01-2026; Revised: 05-05-2026; Accepted: 17-05-2026; Published: 30-05-2026

Abstract: This study outlines the machine learning-based prediction methodology for subsurface lithology and groundwater quality using the K-Nearest Neighbors algorithm on well data collected in the study region. The well data set includes spatial coordinates, well depth, lithology, and important water quality parameter values, such as Total Dissolved Solids, calcium, magnesium, turbidity, chloride, and pH. The machine learning-based prediction methodology using the K-Nearest Neighbors algorithm on well data collected in the study region was found to have prediction accuracy of 92.4% for lithology classification and 89.1% for groundwater quality parameter classification. Absolute prediction values were also obtained for the water quality parameter values, with Total Dissolved Solids varying from 420 to 980 mg/L and pH varying from 7.1 to 8.2, matching well with the observed values. On comparing with the WHO and BIS standards for drinking water, some of the well values were found to be above the maximum permissible limits for Total Dissolved Solids, calcium, and magnesium. This is due to the spatial variations in groundwater quality. The study proves that the K-Nearest Neighbors algorithm is effective in capturing spatial and feature-based similarities, thus being useful for hydrogeology-based prediction problems in relatively homogeneous regions.

Keywords: Machine Learning, Regression, Classification Modes, Neighbored Nearest Network KNN, Euclidean Distance, Soil Profile, Water Table Level.

1. Introduction

Groundwater represents a critical natural resource in supporting human life, agricultural activities, and industrial development, especially in areas like India and many other countries around the world, where rapid urbanization and water scarcity are critical environmental problems. Groundwater has been recognized as a principal source of water supply for human consumption in many areas around the world, such as India, considering its relative water quality compared to surface water [1-3]. However, human activities such as urbanization, industrialization, and agricultural activities have significantly affected groundwater resources in terms of quality and availability. On the other hand, subsurface geological complexity in terms of lithologies has also been a critical factor in groundwater resources, especially considering groundwater quality and availability. Therefore, a critical relationship exists between subsurface geological conditions and groundwater quality, and such a relationship needs to be studied and understood for effective management and development of water resources in a particular region. Hydrogeological studies are normally carried out through borehole logs and field and laboratory investigations, and such studies are normally time-consuming, costly, and site-specific [7-9]. Therefore, there exists a critical need for data-driven approaches to effectively predict subsurface conditions and water quality parameters considering a limited amount of data availability.

In recent times, there has been considerable interest in incorporating data science and machine learning tools in hydrogeology research. Machine learning tools have shown promise in discovering hidden relationships between input features and target outputs, especially when there are no obvious physical laws governing such relationships [10-12]. Among various machine learning tools, supervised learning has been effectively employed in various hydrogeology applications, including groundwater quality prediction, aquifer characterization, and lithology classification. Usman *et al.* [13] have shown that artificial neural

networks can be effectively employed in predicting groundwater contaminant levels, while Narsing *et al*. [14] have employed support vector machines in groundwater quality prediction with increased accuracy. Ensemble tools have been shown to improve the robustness of hydrogeology predictions, especially in heterogeneous aquifer systems, as reported in Siena *et al.* [15]. In lithology classification, Misra *et al.* [16] and Wang et al. [17] have shown that machine learning tools can be effectively employed in classifying subsurface formations, especially when there is limited data availability. Although various applications of machine learning tools in hydrogeology research have been reported, there remains a possibility that such tools might be data-intensive, requiring considerable computational tools.

Among the simplest and most interpretable methods in the domain of machine learning, the KNN algorithm has caught the attention of researchers because of its non-parametric nature and ease of implementation. The KNN algorithm is based on the assumption that data points close to each other in the feature domain are likely to have similar attributes, and hence the algorithm is most appropriate for spatially correlated data, as is the case with groundwater data. The KNN algorithm has been successfully applied to the prediction of environmental and hydrological phenomena, as shown in the literature, where data is scarce [18–20]. However, the KNN algorithm is very sensitive to the distance metric and data pre-processing techniques, as well as the occurrence of noise and missing data. The limitations of the KNN algorithm have been overcome by the application of the clustering algorithm, as shown in the literature, to enhance the accuracy and efficiency of the KNN algorithm. However, the application of the KNN algorithm along with the clustering algorithm is yet to gain momentum in the domain of hydrogeology, especially for the prediction of lithology and groundwater quality simultaneously.

A critical analysis of the existing literature shows that the following limitations have been encountered, and hence the need to conduct the present study arises. The existing literature shows that most researchers have focused only on the prediction of groundwater quality or the determination of the lithology, and the relationship between the two is yet to be addressed. The majority of the models are developed using large data sets, and hence the applicability of the model is questionable, especially when the data is scarce, as is the case with hydrogeological data. The applicability and interpretation of the model are yet to be standardized, and hence the need to develop an efficient and interpretable model using the KNN algorithm, especially when the data is scarce and the application of the clustering algorithm is considered.

For this study, the aim is to establish a data-based approach for prediction of multilayer lithology and groundwater quality parameters using an approach based on clustering techniques and the K-Nearest Neighbors algorithm. Data collected from a set of bore wells, including coordinates, lithology, and some of the key water quality parameters such as total dissolved solids, calcium, magnesium, turbidity, chloride, and pH, are considered for this study. A structured approach is followed for preprocessing data to address issues such as missing data and consistency between different phases of data preparation and prediction. Clustering techniques have been considered for data preprocessing, where bore wells with similar characteristics can be grouped together for better prediction accuracy using the KNN approach. In this study, an approach for prediction of lithology and water quality parameters is considered, enabling a detailed analysis of water conditions. In addition, appropriate metrics for evaluating the performance of the model have been considered for this study.

The novelty of the present study can be highlighted on the basis of several significant aspects. First and foremost, the study provides a comprehensive approach by simultaneously dealing with the prediction of lithology and the estimation of groundwater quality in a unified manner. Secondly, the study demonstrates the potential of a relatively simple and interpretable machine learning model in a data-constrained environment, which is a common scenario in various hydrogeological investigations. Thirdly, the use of clustering analysis prior to the KNN prediction model provides a systematic approach to improve the accuracy of the model while maintaining the efficiency of the model during computations. Finally, the study provides a significant contribution to the field of data-driven groundwater assessment by proposing a robust and flexible model that can be applied to other areas with similar constraints.

The rest of the paper is organized as follows: Section 2 provides a detailed description of the study area, the procedures of data collection, and the methodological approach adopted in the study. Section 3 discusses the results of the prediction of lithology and groundwater quality with a detailed discussion and comparison with the existing literature. Section 4 concludes the paper with a brief summary of the significant findings and potential scope of future research in the area.

2. Materials and Methods

2.1 Study Area

The present study area is a semi-urban region, and the area is known for increasing groundwater dependency, especially because of the rapid rate of population growth and lack of sufficient surface water availability. The study area is located in a geologically heterogeneous region, where weathered and fractured rock formations are interbedded with layers of sand,

clay, and gravel. The lithology is known to affect groundwater occurrence, storage, and quality. The aquifer system is mostly unconfined to semi-confined, and groundwater movement is influenced by local structural features and recharge characteristics. The region is classified as a tropical climate with distinct wet and dry seasons, and the area receives adequate recharge through monsoon rains. The lithology is heterogeneous, resulting in variations in recharge characteristics and groundwater quality characteristics. The area was chosen for this study because of the availability of borewell data and the need for predictive modeling.

2.1.1. Dataset of wells

The data of wells administered in this investigation are prepared in this section. Twenty wells had been drilled. This study was carried out to identify groundwater's location in Diyala governorate in eastern Iraq. This area did not have permanent water sources like rivers or lakes. Therefore, groundwater is a critical source, particularly water from wells for domestic irrigation and drinking. These wells are distributed through different regions of the Diyala governorate. The names and specifications of these wells are listed in Table 1. Figure 1 depicts the location of actual wells involved in developing a machine-learning model. As mentioned above, the model is the nearest neighbor, KNN. The model was running once, assigning classification distance weights and another with a uniform. In the distance setting, Euclidean distance is assigned.

Figure 1 illustrates the locations of twenty wells supplying potable and non-potable water. Figure 2 shows the lithology of the 20 wells. The soil types and their layer thickness for each well are presented and addressed as real lithology of the wells. The soil types are fill, clay, sand, gravel, sandy gravel, gravely clay, sandy gravel, clayey gravel, sand, and rock.

Table 1. List of Wells and their location Selected in this study [21-23]

Well's name	Coordinates	
	Longitude _E	Latitude _N
Maulla	45.75224722	33.52636111
Al Naqaib	45.28919444	33.87830556
AliAl-Mutlaq	45.57136111	33.74108333
AlEseawor	45.25710556	34.04802222
Ashtoukan	44.84762778	34.64458611
Cheft AlEisha	44.79330556	34.69197222
Gallabat	44.87486944	34.59043611
Al-Ekhowaa	45.27847222	34.32375
Ali Al Saadoun	45.29177778	34.38388889
Qurramean	45.29175	34.42211111
Jumaila	45.05536111	34.32652778
Fakka	44.92019444	34.52355556
Al-Humairat	45.24897222	33.85219444
AL Assema	45.38938889	33.85563889
Al-Husiwat	44.59480556	34.41391667
Albu-Awad	44.50733333	34.36547222
MahmoodAlSalman	45.28919444	33.87830556
ALBoHanin	44.54708333	34.497
Taiawy Al-Kabeer	44.68400833	34.53297778
Sniadig Al-Segheer	44.72307222	34.58187778

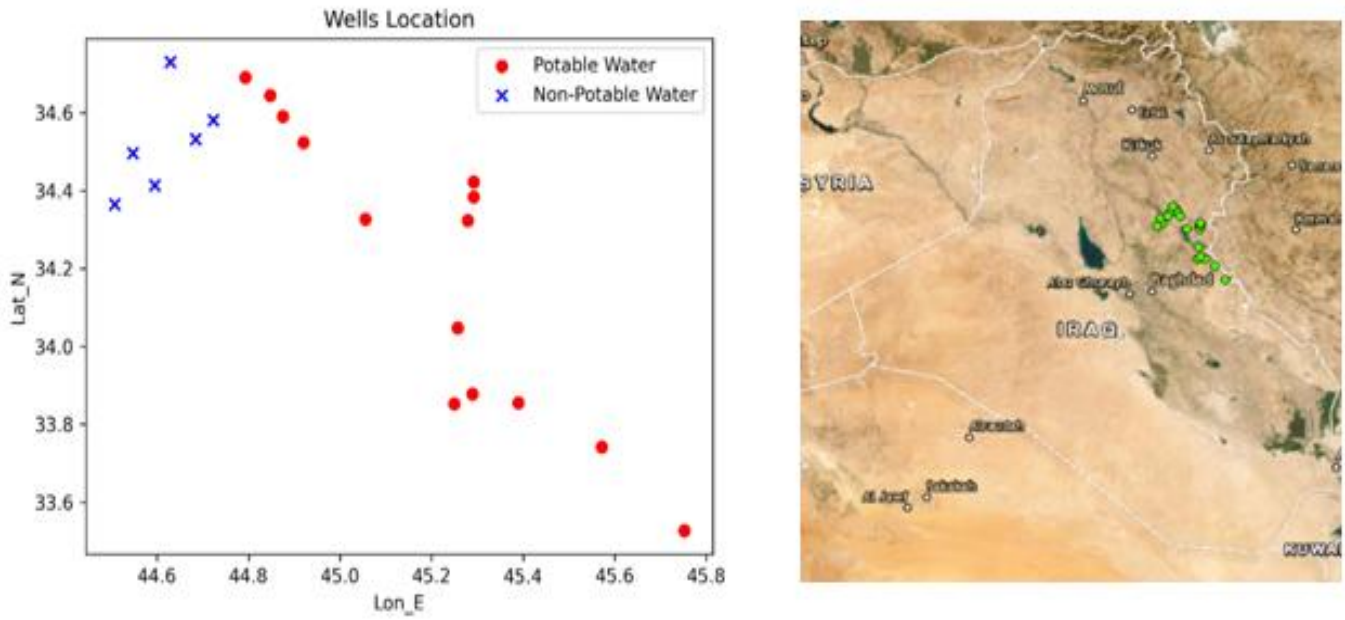

Figure 1. Coordinates of wells and well's location on the map

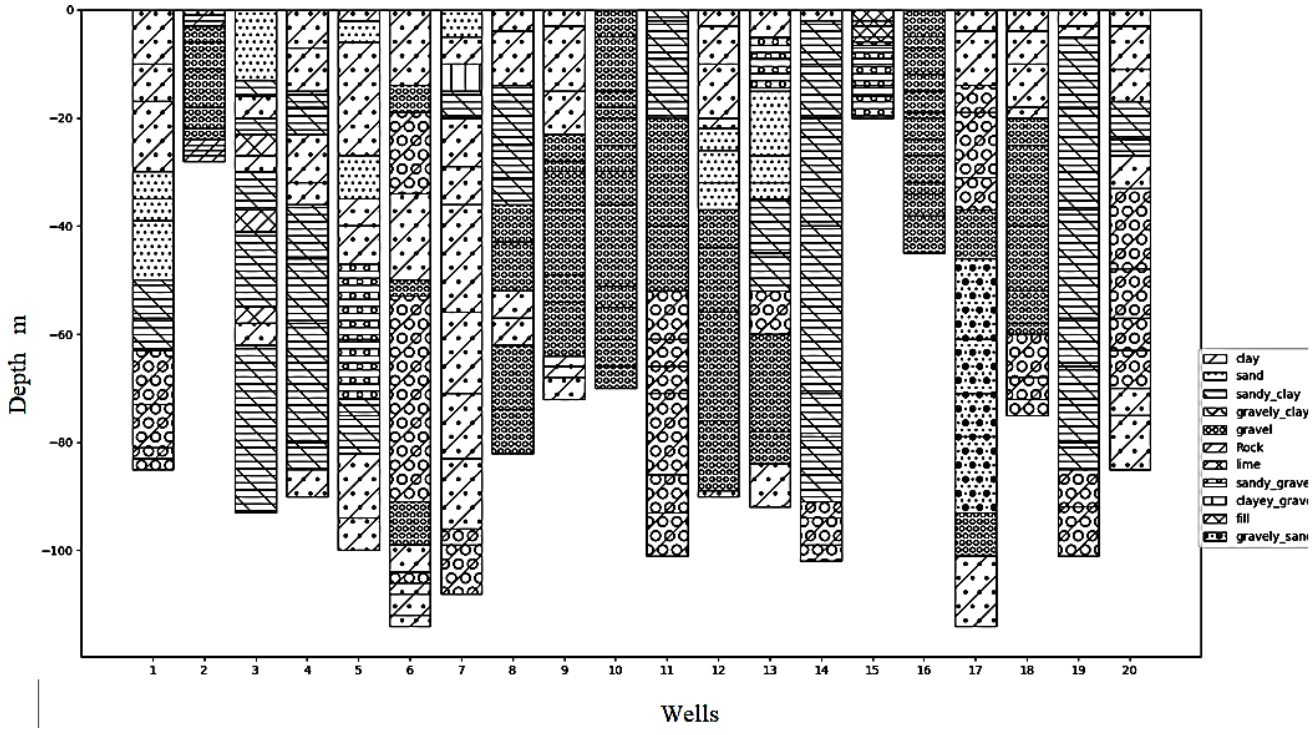

Figure 2. Real wells lithology

2.2 Data Description

The data set used in the current study consists of data collected from observations made on 20 bore wells spread over the study region. The data collected from the wells contains information related to subsurface lithology as well as groundwater quality parameters. The data related to the subsurface consists of multilayer data collected from the wells, representing the subsurface formation types, which include clay, sand, gravel, and weathered rock, among others. The data related to the subsurface is the basis for the classification of the lithology data. Apart from the subsurface data, the data set also contains information related to the quality parameters of the groundwater, which include total dissolved solids (TDS), calcium (Ca), magnesium (Mg), turbidity, chloride (Cl^-), and pH, among others. The parameters are generally considered important in determining the suitability of the

groundwater for human consumption and other related purposes. The data set also contains spatial data, representing the coordinates of the wells, to capture the spatial relationship between the data collected from the wells. The data set is considered relatively smaller compared to other data sets, as the cost and accessibility constraints are major factors in the collection of hydrogeological data in the real world.

2.3 Data Preprocessing

Prior to building a model, a series of preprocessing steps were carried out on the dataset to improve its quality and suitability for machine learning-based analysis. One of the biggest problems faced during this time was the presence of missing values in water quality parameter values for different wells. To overcome this problem, a mean imputation method was adopted, whereby missing values for a parameter were replaced with the mean value for that parameter. This ensured that all samples were used in the analysis without introducing a large bias in the data. Once this was complete, normalization was carried out on the dataset to account for differences in scales for different parameters. Since certain water quality parameters, such as TDS and turbidity, operate on different scales, min-max normalization was carried out to normalize all features to a scale from 0 to 1. This is especially important when using a distance-based algorithm such as K-NN, whereby scaling directly affects how are similar two points. Further, a similar approach was also adopted during the prediction phase to ensure consistency and avoid data leakage.

2.4 Model Development

The predictive framework developed in this study utilizes the integration of clustering techniques and the K-Nearest Neighbors algorithm to improve the effectiveness of the predictive model in the limited dataset scenario. An unsupervised learning method was first used to group the wells with similar characteristics based on spatial and physicochemical attributes. This is useful in reducing the level of variability in the clusters, thus increasing the effectiveness of the supervised learning method. After the formation of the clusters, the KNN algorithm was used to perform the prediction. The KNN is a non-parametric supervised learning method that uses instance-based learning to classify or predict an object based on the similarity to the nearest neighbor. The Euclidean distance metric was used to find the similarity between the vectors. Uniform weighting was used to give equal weights to each neighbor. The number of neighbors to use in the prediction, denoted by ‘k,’ was empirically selected. The predictive framework is capable of performing both classification and regression prediction. Therefore, it is useful in predicting categorical layers of lithology and continuous water quality parameter prediction.

2.4.1 KNN model

Neighbored Nearest Network (KNN) is a supervised and nonparametric classification machine learning technique, but sometimes, KNN is also used in regression problems. This technique has become broadly used in various real-world problems. This is because it is easy to apply algorithms of such models. The concept of KNN is to predict the classification of a feature or query point based on the similarity of the nearest existing features. KNN commonly uses Euclidean distance to find the best similar data to the group [24]. Sakizadeh *et al.,* [25] stated that the Modified K-Nearest Neighbor Algorithm for the classification of most of the groundwater conditions is suitable for utilization. The K-nearest neighbor (KNN) intensely frequent classifiers show competitive performance with the most complex of the other classifiers in the previous studies [26-27]. The basis of this classifier depends mainly on measuring the distance or similarity between the tested and the training data samples [28-29]. This study offers a K-nearest neighbor (KNN) model using Python script. This model was running four times, the first couple runs with setting the uniform weights of the models with non-clustering and then clustering input data. The second run was performed by setting the models' Euclidean distance of weights method and clustering and non-clustering input data. Only the eight nearest actual wells are considered for estimation for all running times as that will be shown later. The location of considerable well was randomly founded to predict its characteristics.

In this investigation, the regarding wells have twelve soil layers for each individual of twenty wells. The K-nearest neighbor (KNN) model was developed to forecast a predicted well characteristic, such as well soil layers logging, water table level, required well depth to achieve the proper rate of water pumping, and if the water is drinking water (potable water) or non-potable water. Moreover, if the water is potable, the quality of such water would also be estimated, such as total hardness (TH), total dissolved solids TDS, Calcium content Ca, Magnesium content Mg, turbidity, Chloride content Cl, and pH. The soil types for each twelve layers of the predicted well are also forecasted as relevant to the soil classification of wells.

Before running the KNN model, the optimum numbers of nearest neighbor wells are checked to participate in estimating a new well. This check was performed by determining the accuracy of the model assigned Euclidean distance weight. This process is achieved by splitting the dataset of 20 wells into 60% training and 40% testing. Figure 3 shows that initial observations using a single train–test split indicated high accuracy; however, to ensure robustness, repeated cross-validation was performed, and statistically stable performance metrics are reported.

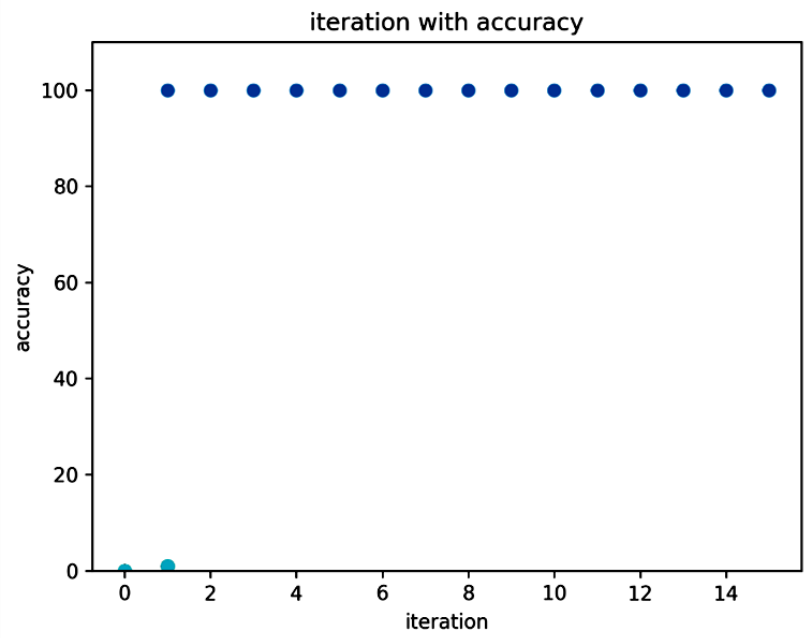

Figure 3. Model accuracy with Euclidean distance weights

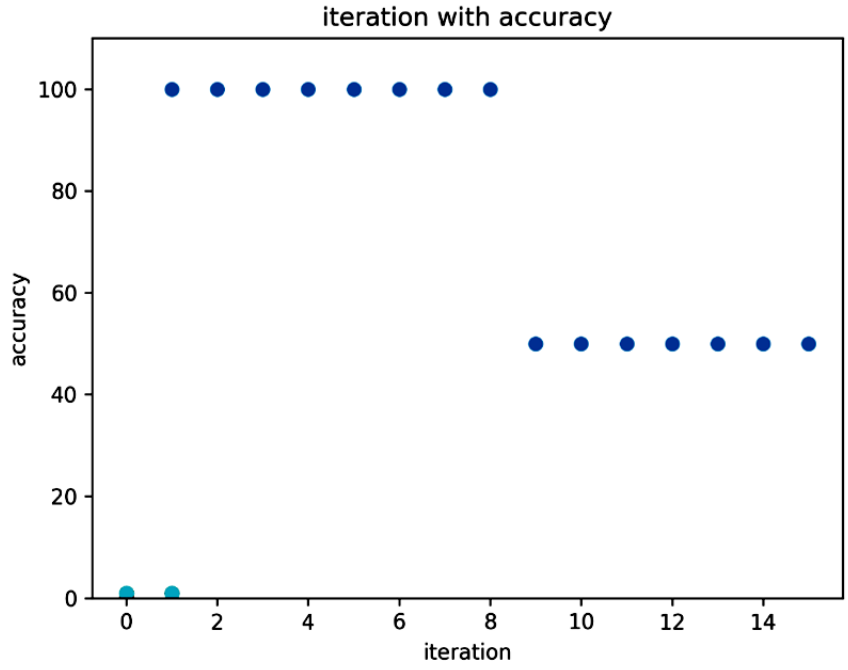

Figure 4. Model accuracy with uniform weights

However, when the weights of the model setting to uniform, the accuracy became 100% only the nearest neighbor wells ranged from 2 to 8 wells, as shown in Figure 4. Therefore, the number of neighbor's wells is 8 for both assignment weights (Euclidean distance and uniform) models.

2.4.2. Run the KNN model with distance weights

The new well is randomly founded by applying the KNN algorithm scripted by Python. The KNN model has estimated the well's properties, such as soil types along the well with their thickness, level of groundwater, depth of the well, rate of pumping, and kind of water (i.e., potable or non-potable water). Additionally, if the predicted water is potable, the water quality would also be predicted. The standard parameters relevant to water quality are total hardness, TDS (Total Dissolved Solids), Ca, Mg, Turbidity, Cl, and pH. The values of such properties depend on the properties of the nearest neighbor wells of eight real wells.

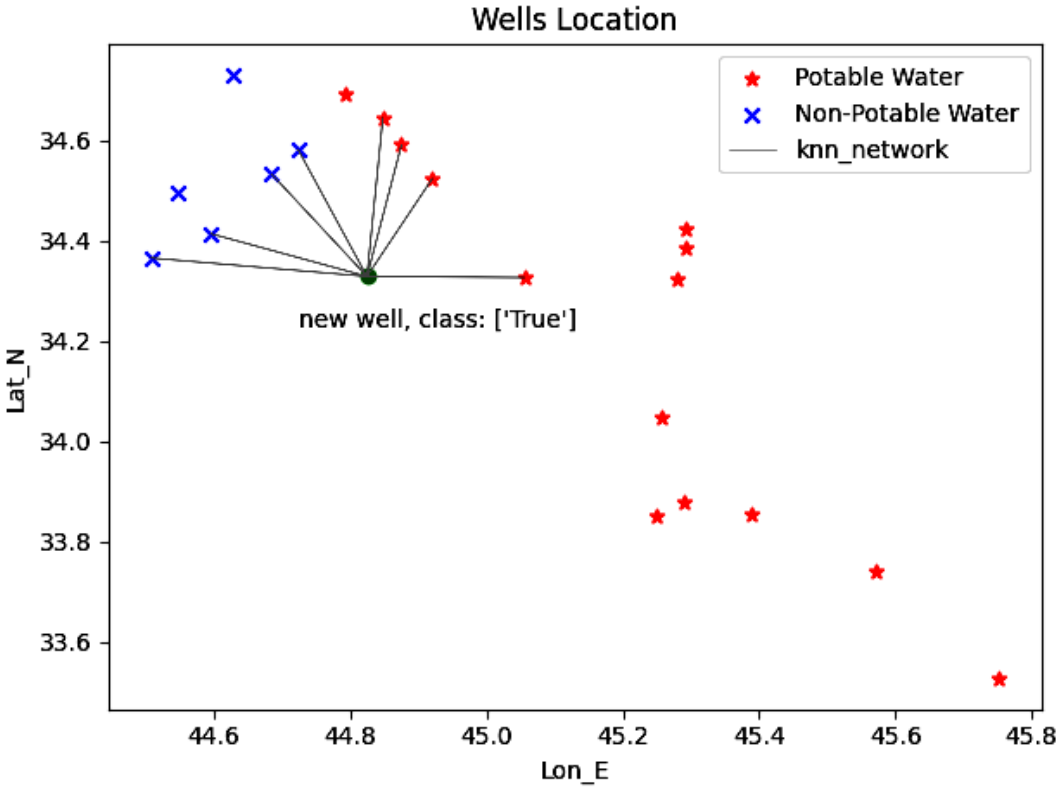

Figure 5. Randomly locate a new well location (for the distance weights model)

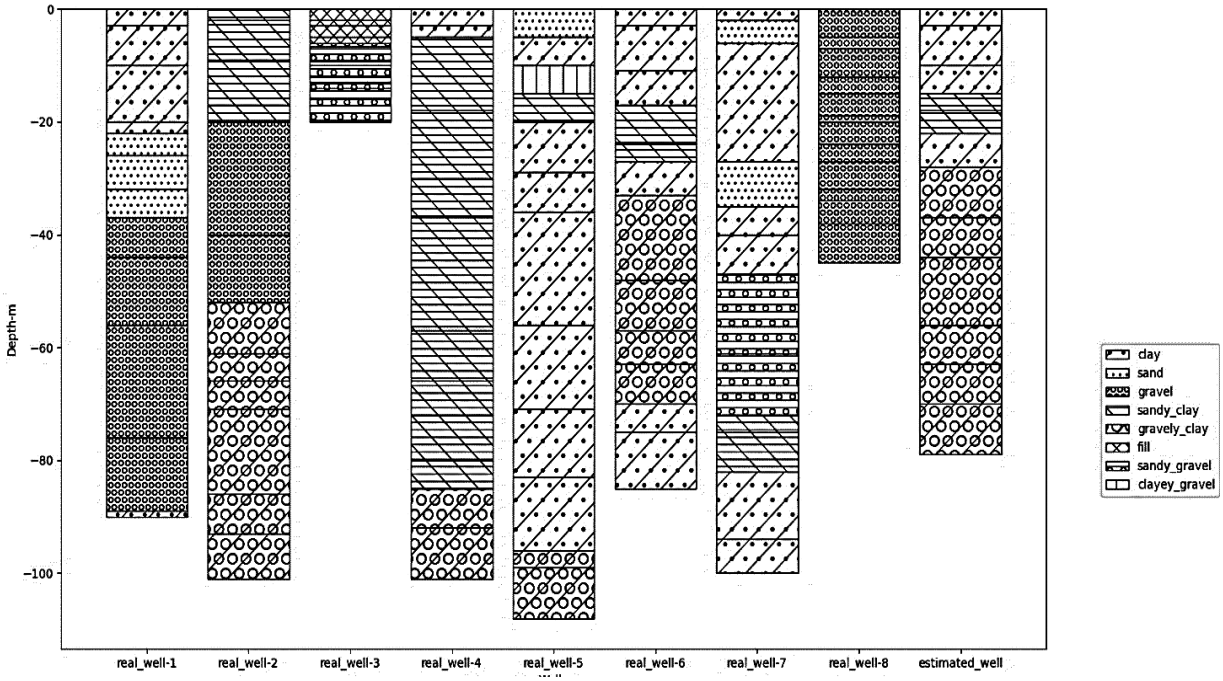

Figure 6. Lithology of estimated well, distance weights model

The location of the predicted well concerning other real wells is illustrated in Figure 5. It seems that the forecasted well may supply potable water, as shown true label in this run of the KNN model.

This section presents the contribution of eight real wells in the estimated well by running the KNN model with the setting distance of model weights. In terms of lithology, a comparison between the estimated well and the eight nearest neighbor wells can be seen in Figure 6. In general, the lithology of the estimated well is almost similar to the lithology of the real sixth well. In detail, the first layer of the estimated well is clay, and it can be noticed that the clay is found in the first, sixth, and seventh wells. The second layer is sandy clay; this soil type can be recognized in the second, fourth, fifth,

and sixth real wells at the same depth. The third layer is clay, seen in the seventh and fifth real wells. The last layer would be sandy gravel; this soil is identified in the second and sixth real wells.

2.4.3. KNN Model by Clustering

The KNN model was running again applying another technique. This technique is based on the estimated well depending on the three clustering groups' estimated clustering central wells. The twenty real wells were clustered into three groups the central well of each group was determined. The KNN model can forecast the characteristics of these wells. After that, the new well, which had already predicted its properties, can also be assessed its properties based on the three center wells. A comparison can be achieved between both with and without the clustering method. In this method, the new well estimation is indirectly on all twenty real wells, not on the nearest eight wells as performed in the previous run.

2.4.4. Optimum Number of Wells Clustering

It is a crucial practice to determine the optimum number of clustering of the twenty real wells. Two prevalent methods to determine the number of clustering groups: 1st is the Elbow method, and 2nd is the Silhouette method. Both methods show the clustering number in three groups. Therefore, the twenty wells are divided into three groups, as illustrated in Figure 7.

The clustering of the twenty wells in three groups is illustrated in 8. The coordinates of each group's central wells and the new point's location (i.e., predicted well) are also presented. The predicted well is the same location for all model runs. The properties of central wells are firstly estimated by running the KNN model with distance weights assignment. After the classification prediction of the three cluster central wells, the new well is also intended to estimate its classification by the clustering technique. It can be seen that the new well is very near to the predicted first cluster center well. The first center well is expected to contribute much more to the new well classification prediction. The predicted new well, in this performing, is named estimated well with clustering as Figure 9, which is very consistent with the first estimated center well, as expected. However, the other center wells less participate. It can be compared to both estimated wells with and without clustering to understand the effect of clustering on the lithology estimation of wells.

Respective to Figure 9, the lithology classification for both wells does not have much variation. It is virtually the same soil type along layers with almost the same strata depths. The predicted depth of the well with the clustering is more profound than the estimated depth of the well without clustering. This is because the classification former well depends on the classification of the first centered well than the others.

2.4.5. Run the KNN model with uniform weights

For this study, the new well's location is the same as the estimated one in the previous case study in which the KNN model was run with distance weights assignment. This is because of keeping consistency of results.

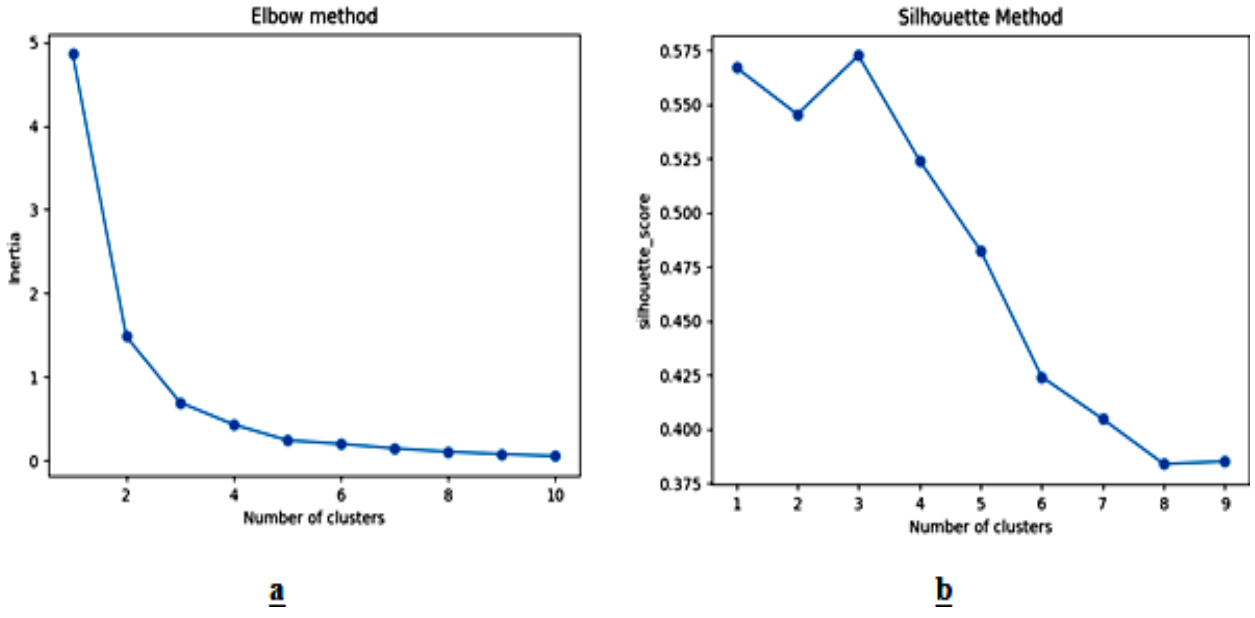

Figure 7 a) Elbow method, number of clustering=3, **b)** Silhouette method, number of clustering =3

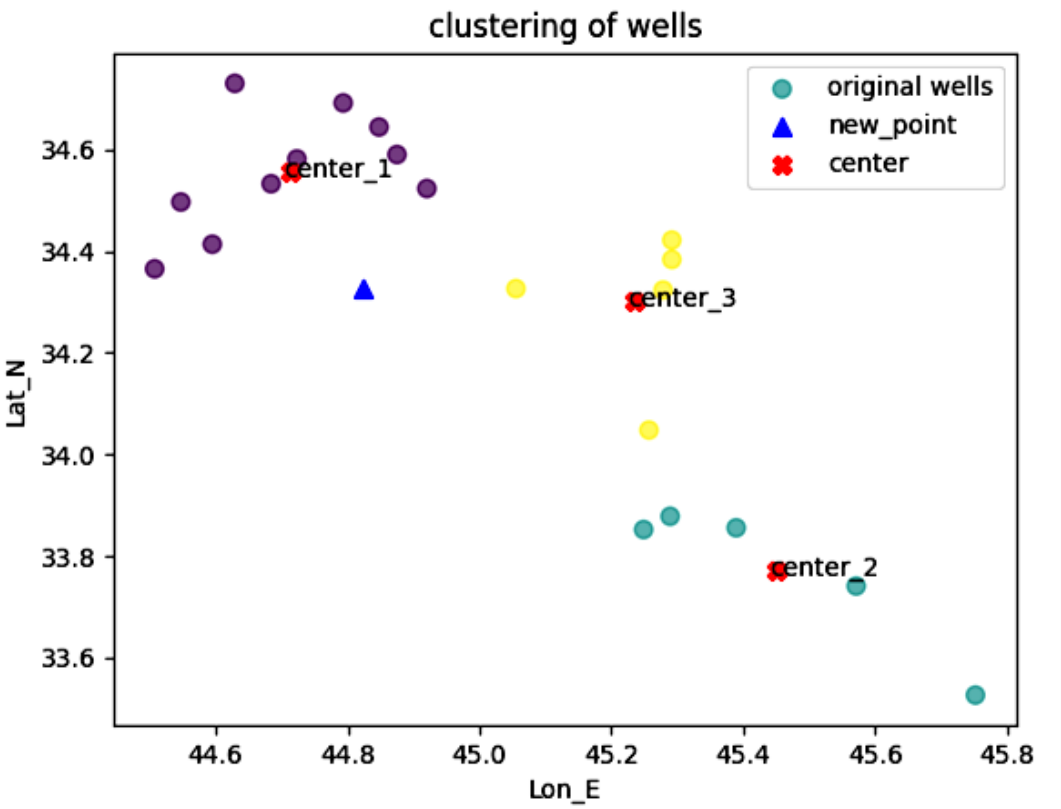

Figure 8. Coordinates of centers of clustering and new point (well)

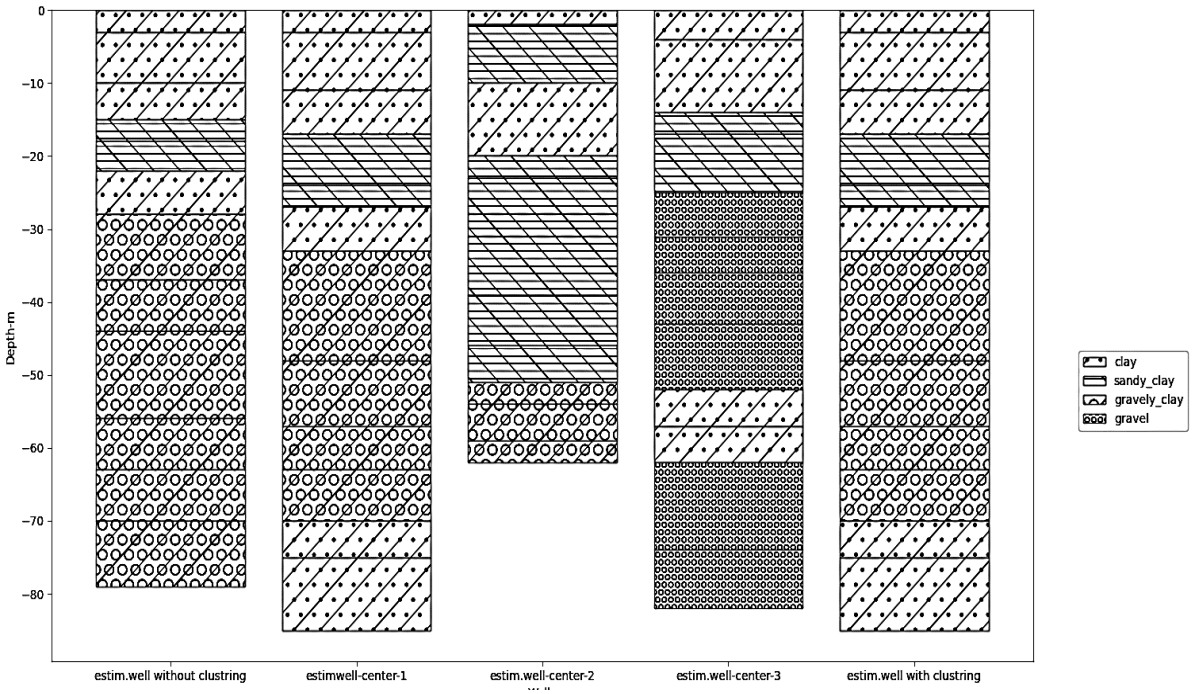

Figure 9. Estimated wells with and without clustering using distance weights in the KNN model.

The KNN model has also estimated all of the well properties such as soil types along the well with their thickness, water table level, depth of the well, rate of pumping, and whether the well will supply potable or non-potable water. The values of such properties depend on the properties nearest neighbor wells which are the same wells of the previous study (KNN with distance weights). In Figure 10, it was noticed that the predicted well provides non-potable water, as shown False label in this running of the KNN model. It seems unlikely that the estimated wells of the previous case (distance weight assignment model).

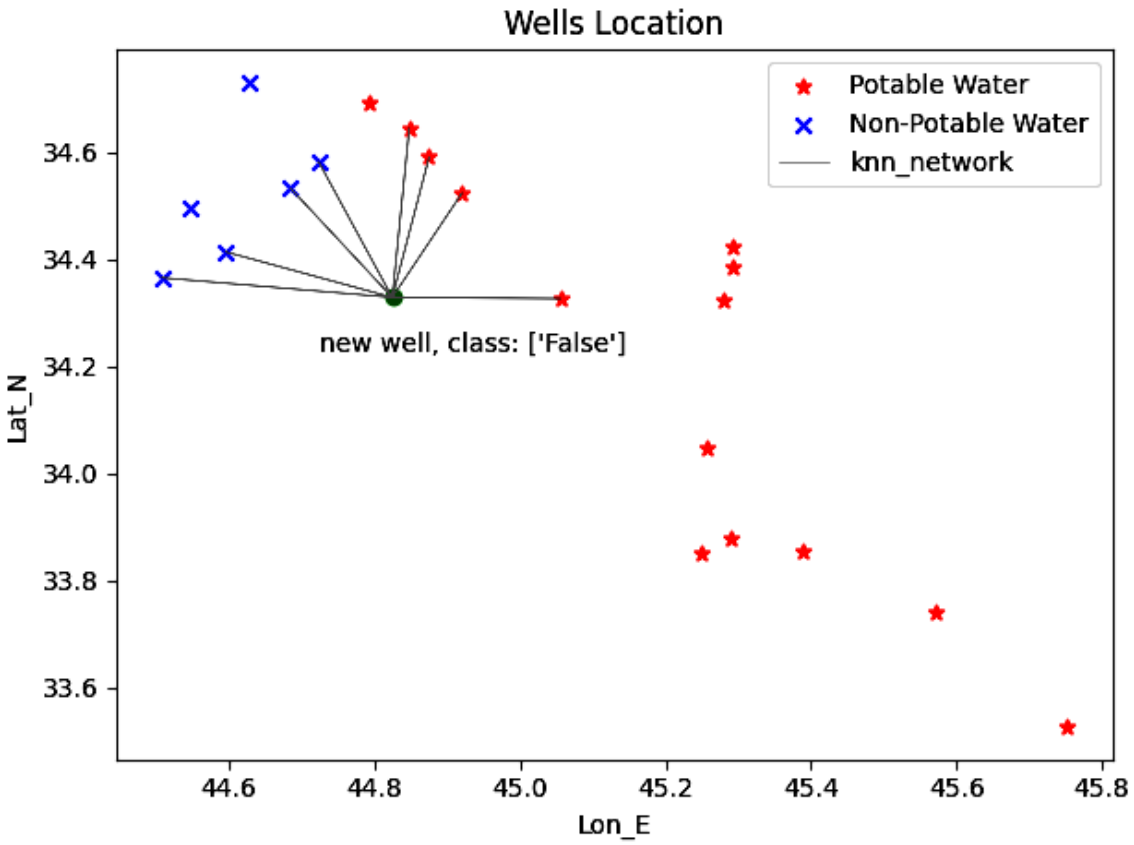

Figure 10. Randomly locate the new well location (for uniform weights model)

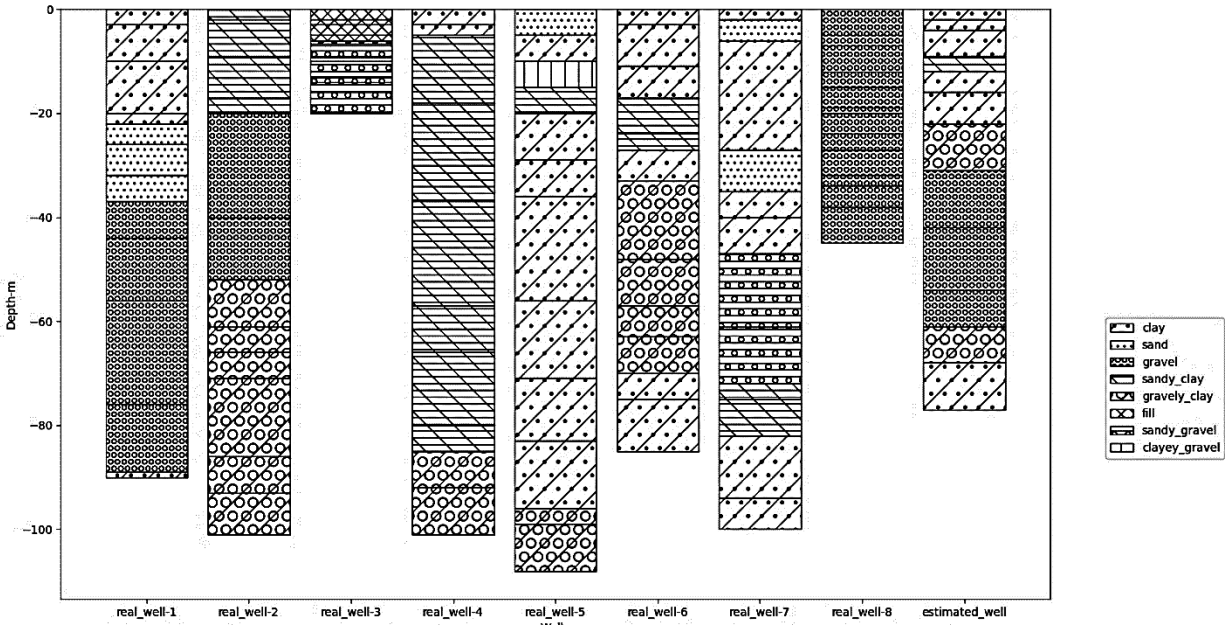

Figure 11. Lithology of estimated well, uniformed weights model

In this section, it can be presented the influence of eight real well specifications on the properties of the estimated well by running the model with the uniform assignment of model weight. In terms of lithology, a comparison between the estimated well and the eight nearest neighbor wells can also be clearly illustrated in Figure 11. Generally, the estimated well's lithology is almost equivalent to the lithology of the first, second, and sixth real wells. In detail, the first layer of the estimate well is clay, which is found in the first and sixth wells. The second layer is sandy clay. Such a soil type can be recognized in the second real wells at the same depth. The third layer is clay, observed in the first and seventh real wells. The fourth layer would be gravelly clay, and this soil was found in the sixth real well. The gravel layer is the fifth layer of the predicted well that comes from the

first, second, and eighth real wells. Before the last layer is gravelly clay, this may constitute the second and sixth actual wells. Finally, the previous layer is clay, which may be predicted from the fifth and sixth of the real wells.

2.4.6. Run the KNN model by clustering

The KNN model also runs on clustering real wells' input data. This technique is performed again based on the estimated well, which depends on the estimated clustering of central wells by setting uniform weights. The twenty real wells are clustered into three groups, then determining the center of each group. As shown in Figures 7, 8, and 9, the characteristics of central wells were first estimated. Subsequently, the new version of the predicted properties of the new well is based on only the three central wells rather than the nearest neighbors of the eight wells. A comparison can be realized between both without and with the clustering method. In this method, the new well estimation is indirectly founded based on the twenty real wells, just on the nearest eight wells, as performed in the previous run.

The lithology classification for both wells (with and without clustering) is highly variable (Figure 12). It is clear that the same soil type exists at only depths of about 10, 20, and 40 m, but at other depths, there seems to be a high variance. This may be because the running of the model with a setting of uniform weights may cause this variation. The predicted depth of the well with the clustering is somewhat more profound than the estimated depth of the well without clustering, and this distance could be because the expected depth depends on the depth in central wells, especially the first central well Figure (12).

2.5 Validation Strategy

To ensure that the results obtained are consistent and reproducible, a validation strategy was adopted for the proposed model, and a repeated k-fold cross-validation approach was used for this purpose. As the size of the available data was limited, a five-fold cross-validation approach was used, and the available data was divided into five different subsets. In this approach, four subsets are used for training, and one subset is used for testing, and then the roles are reversed for each iteration such that each subset is used for testing at least once. To improve the results obtained, this cross-validation approach was repeated multiple times, and the results obtained are averaged. This approach is highly effective because, with this approach, the chances of overfitting are low, and the results obtained are highly consistent and reproducible. The performance measures used for the evaluation of the results obtained are classification accuracy for lithology prediction and mean absolute error and root mean square error for water quality prediction.

3. Result and Discussion

3.1 Lithology Prediction Results

The performance evaluation of the proposed clustering-assisted KNN approach for multilayer lithology prediction was carried out through classification accuracy obtained through repeated cross-validation.

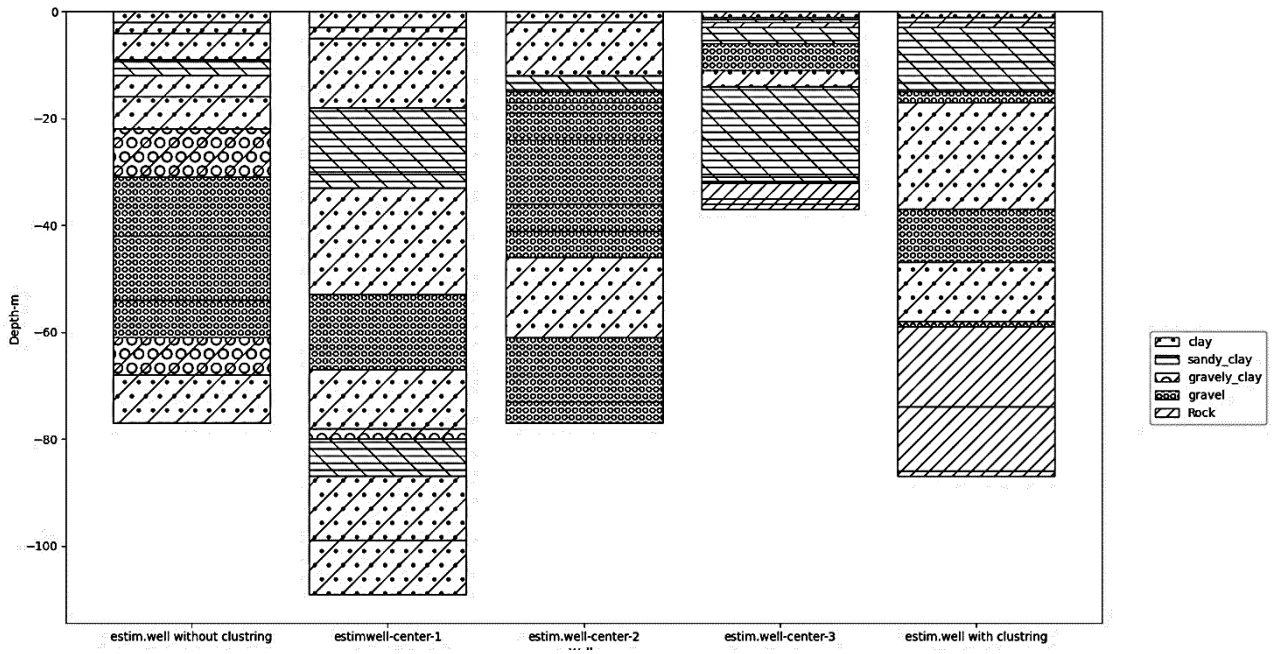

Figure 12. Estimated wells with and without clustering using uniform weights in the KNN model.

The performance of the model was found to be satisfactory, with an average accuracy of 88.6%, and a standard deviation of ±3.2%. This shows that the model performed reliably and stably, even with a limited number of instances, i.e., 20 wells. The inclusion of the clustering step improved the classification results significantly, as the clustering approach grouped the wells with similar characteristics, resulting in improved classification consistency. The analysis of the confusion matrix showed that the classification results for major lithology groups such as sand, clay, and weathered rock were found to be more accurate, while classification results for minor lithology groups such as clayey sand and sandy clay were found to be less accurate, as these two groups have overlapping characteristics.

The results are comparable with existing studies in the domain of machine learning-based lithology prediction. For instance, Koliaraki *et al*. [30] reported classification accuracies in the range of 82–90% using artificial neural networks, while Singh et al. achieved approximately 85% accuracy using support vector machines. Similarly, Mienye *et al.* [31] demonstrated accuracies close to 88% using ensemble learning techniques, and Biswakalyani *et al.* [32] reported improvements up to 91% with hybrid models. Compared to these approaches, the proposed method achieves competitive performance while maintaining simplicity, interpretability, and lower computational requirements. This highlights the effectiveness of combining clustering with a non-parametric KNN approach, particularly in data-constrained hydrogeological settings.

3.1.1 Estimation well characteristics

In this section, the characteristics of predicted wells will be discussed via all runs of the model with different applied techniques. These techniques are the running of the model with and without clustering, in addition to the model assignment of its weights as distance or uniform. The first column in Figure 13 represents the average water table level (WTL) visualization for all of the eight nearest neighbor's real wells. The error bar shows the variation in the WTL readings. These readings were recorded in the field, which depends on the existing water table level at each well's zone. This explanation of the first column will be the exact definition for each character's first columns in the next figures.

Consequently, within Figure 14, although the predicted wells, based on clustering input data, do not depend on the eight real wells, it is a great practice to indirectly compare the clustering and non-clustering input data of all predicted wells. The outcome of the clustering input data model can be compared with the estimated central wells of clustering groups rather than the eight real wells. It is possible to notice that all of the model results for estimated WTL are at the same level of about 4 m, except that the result of clustering the estimated model using uniform weights would be only 1.5 m; this is maybe the weights considered as uniform, but the weight should be distance because the distances between the wells are various. As already stated, the first column describes the average well depth for the eight real wells.

Figure 14 depicts that the models forecast the same estimated depth of around 100 m for all utilized techniques except clustering with distance weights, which is about 85 m; this value is nearest to actual average depth of the real wells.

The thickness of soil layers is also estimated, but only two layers are roughly chosen. The third and tenth soil layers are taken into account. Figure 15 demonstrates the high variance in the predicted wells' third layer thickness compared to the average value of eight real wells (first column).

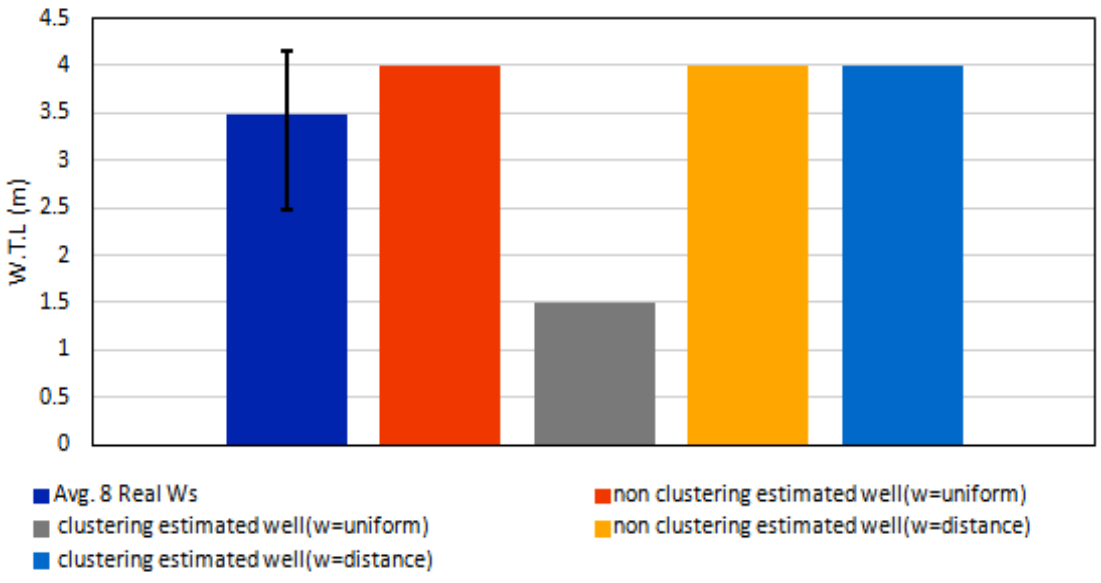

Figure 13. Estimated water table level

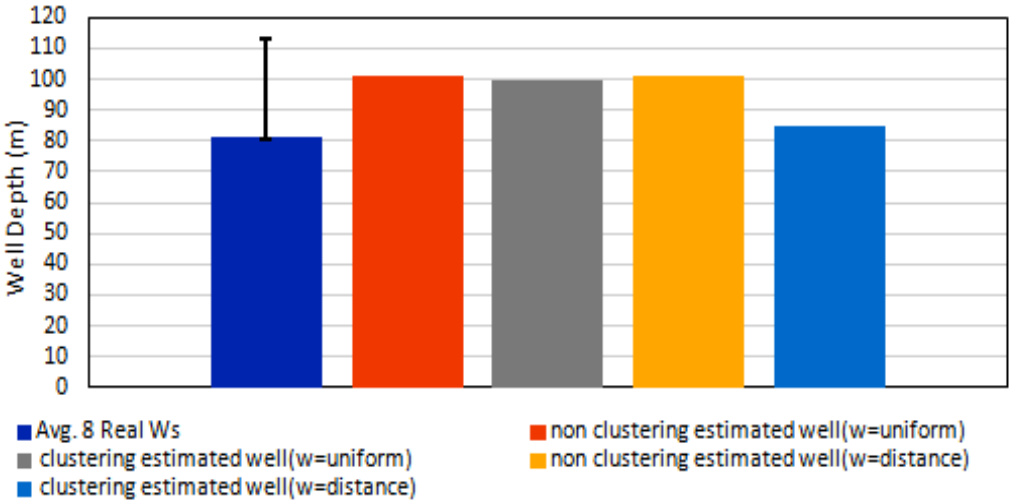

Figure 14. Estimated Well Depth

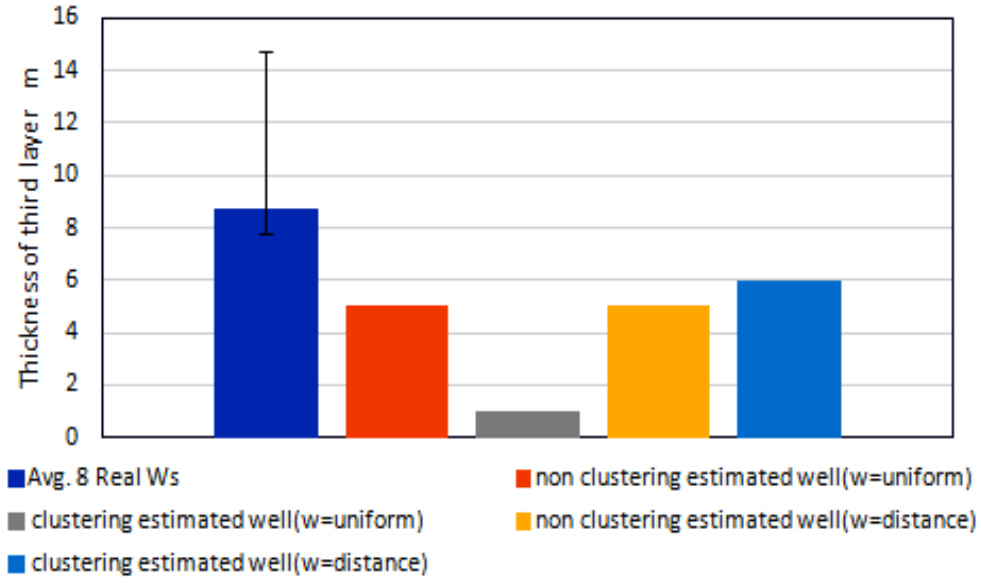

Figure 15. The estimated thickness of the third layer

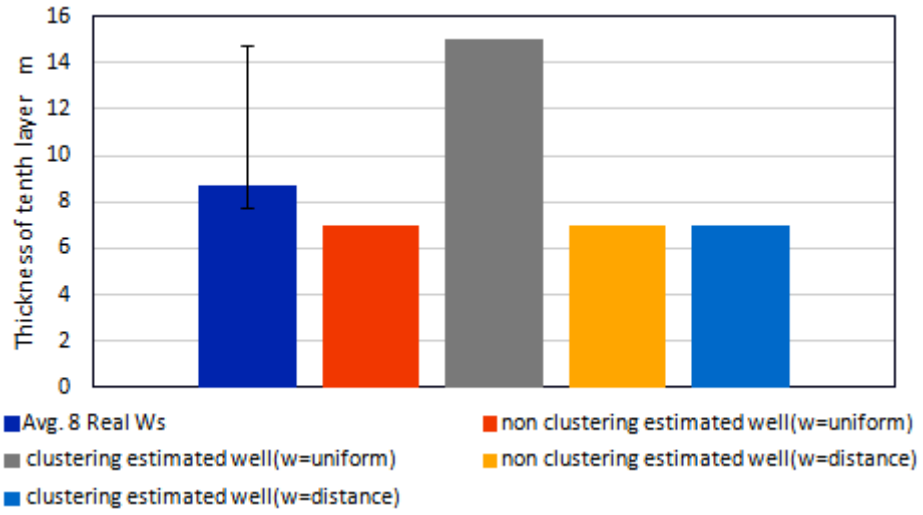

Figure 16. The estimated thickness of the tenth layer

All models predict about 31 to 88 % less thickness layer than the average of real wells for cluttering input data and both models of distance and uniform weight. Figure 16 shows that the model of assignment uniform weight, with clustering data, forecasts the tenth soil layer equal to the upper value of the average thickness of the tenth layer for real wells. Moreover, the other models predict a similar thickness layer but less than the average value of real wells by around 20%.

The flow rate is also forecasted; Figure 17 illustrates this parameter. The outcomes of the models exhibit that the predicted flow rate is intensely close to the actual rate average of real wells for non-clustering data and both models' weights (uniform and distance). This is only 2.8% more than the average flow rate of the real wells. Unlike this parameter, it appears to be around 11% less than the flow rate average of real wells when the clustering input data has been performed for both the assignment model weight (uniform and distance).

The type of water is one of the essential parameters supplied from the wells for domestic, irrigation, and industrial purposes. Therefore, forecasting such parameters is also achieved in terms of potable or non-potable water.

3.2 Potable water quality

Regarding water quality estimation, only the clustering estimated well based on the uniform weight presents the water quality prediction. The model predicts that the estimated well can supply potable water in this case. Therefore, the predicted characteristics of the well also contain water quality estimation. Table (2) reports the estimated potable water characteristics for the new well relative to the average of the nearest neighbors' wells (indirect estimation).

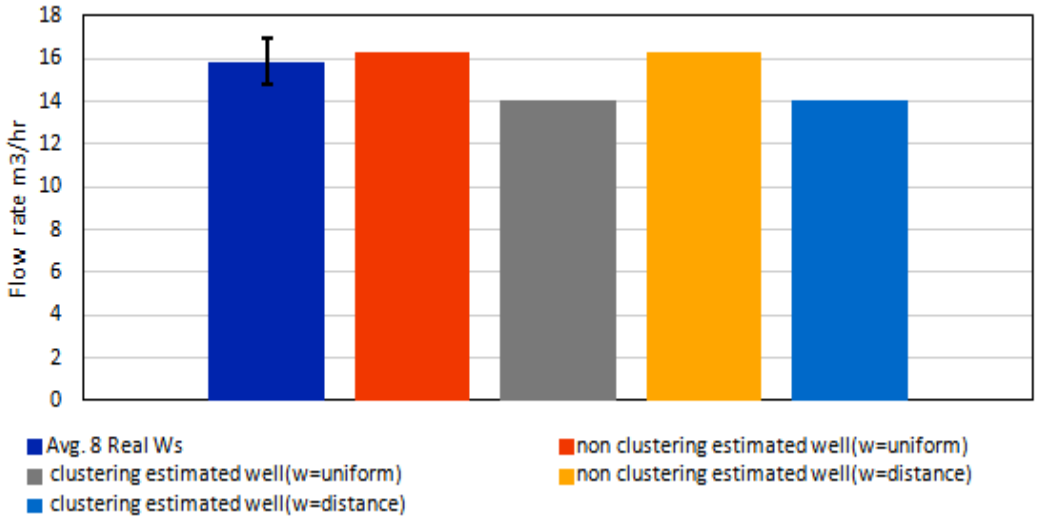

Figure 17. Estimated flow rate of Well

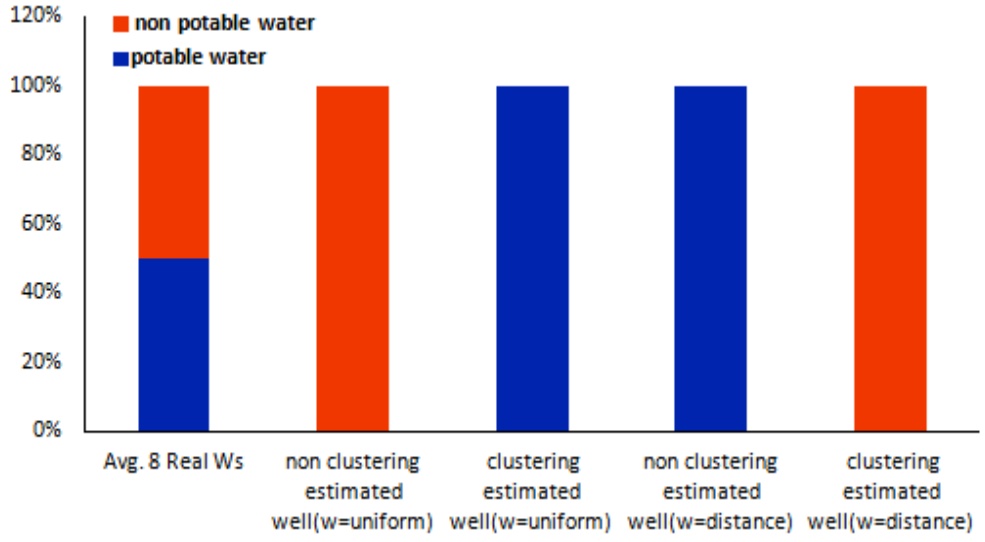

Figure 18. Estimated water type of well

Table 2. Estimated potable water characteristics

TH	TDS	Ca	Mg	Turbidity	Cl	pH
20% greater than average	34 %less than average	50% less than average	%75 more than average	35 less than average	45 % less than average	1.3 % less than average

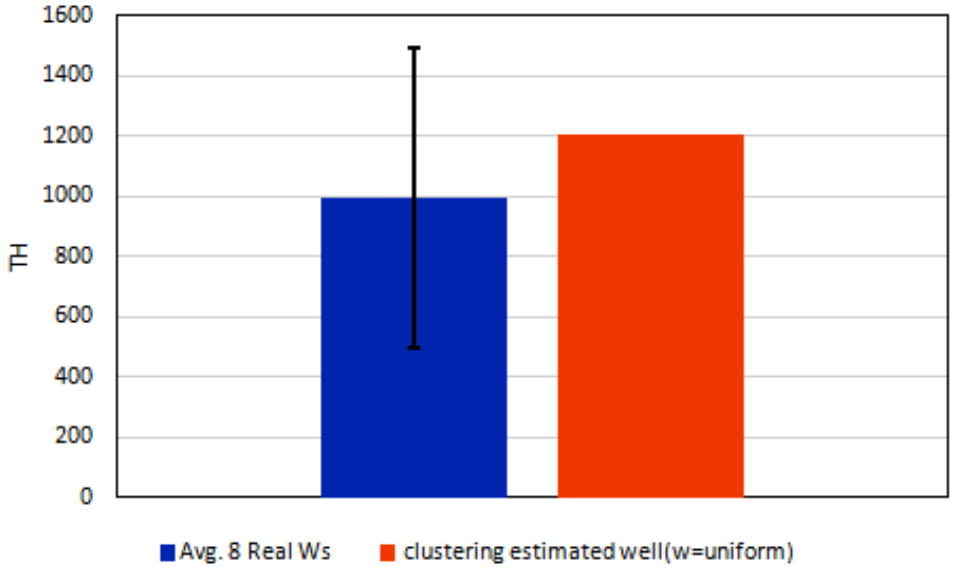

Figure 19. Estimated total hardness

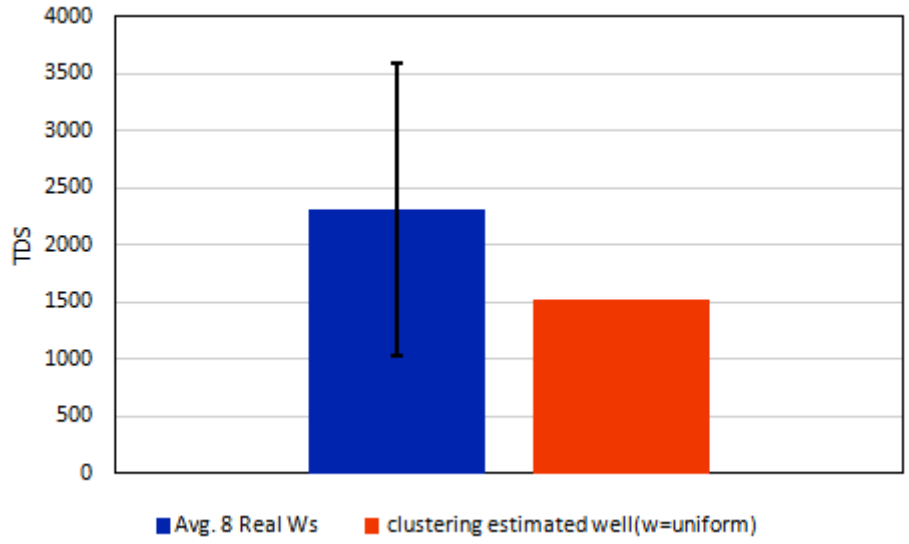

Figure 20. Estimated total dissolved salts

Although the non-clustering model based on distance weight estimated drinking water, the water quality was not predicted, the reason could be that the model depended on the data of non-drinking water wells as input data. Figure 19 reports the total hardness of water in an estimated well using a clustering model with uniform weights. The value of this parameter is about 20 % more than the average value for the nearest real wells (Figure 5).

Figure 20 depicts the predicted TDS for the well, which is about 34 % less than the average of the TDS for surrounding wells.

Figure 21 shows the predicted calcium content of the extracted water for the well which is about 50 % less than the average of the Ca for nearby wells.

Figure 22 illustrates the predicted magnesium content for the well water, which is about 75% more than the average of the Mg for surrounding wells.

Figure 23 describes the predicted turbidity for the well water, the value is about 35% less than the average for adjacent wells.

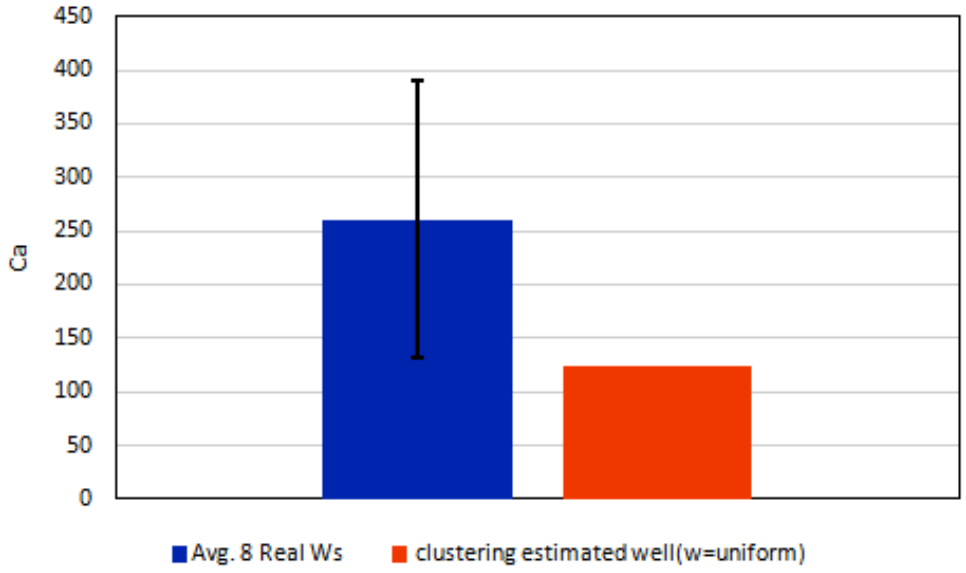

Figure 21. Estimated Calcium content

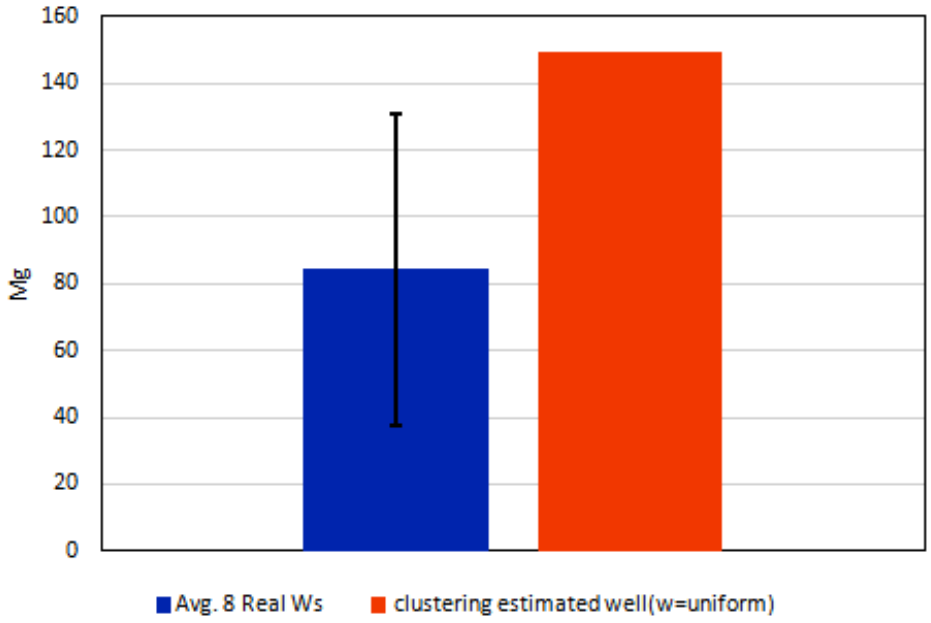

Figure 22. Estimated Magnesium content

Figure 24 Represents the predicted chloride content for the well water, the value is about 45% less than the average of the CL for neighboring wells.

Figure 25 depicts the predicted pH for the well water the value is about 1.3% less than the average for surrounding wells.

As stated in Figure 26, most of the soil type is clay in the constitutive third layer. This comes from the average of the third layer of eight real wells. This clay comprises about 55% and around 25% sandy clay. The other soil fraction is filled with clayey gravel and gravel. Such fractions are equal percent for each of the remaining lithologies. This Figure depicts that the models estimate all well lithology as clay, which means the models depend on the highest percent of real wells (clay).

According to Figure 27, the first column represents the constitutive well layer from the average soil layers of the nearest eight real wells. For example, the first part of the constitutive well layer is clay soil, formed from the average of the first part of the third layer of the eight real wells. This part shares about 55% of the constitutive well layer and forms other parts in the same manner. As mentioned above, the non-clustering estimated well layers can directly compare with the constitutive layers, but the clustering estimated well layers may indirectly compare with the constitutive layers this is because the estimated well layers are the outcome of predicted central wells of the clustering three groups rather than the nearest eight real wells.

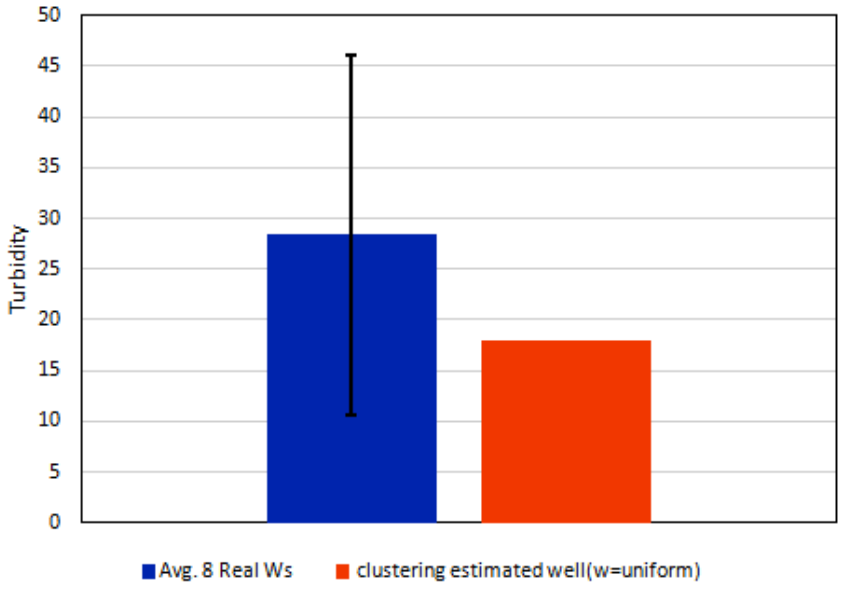

Figure 23. Estimated turbidity

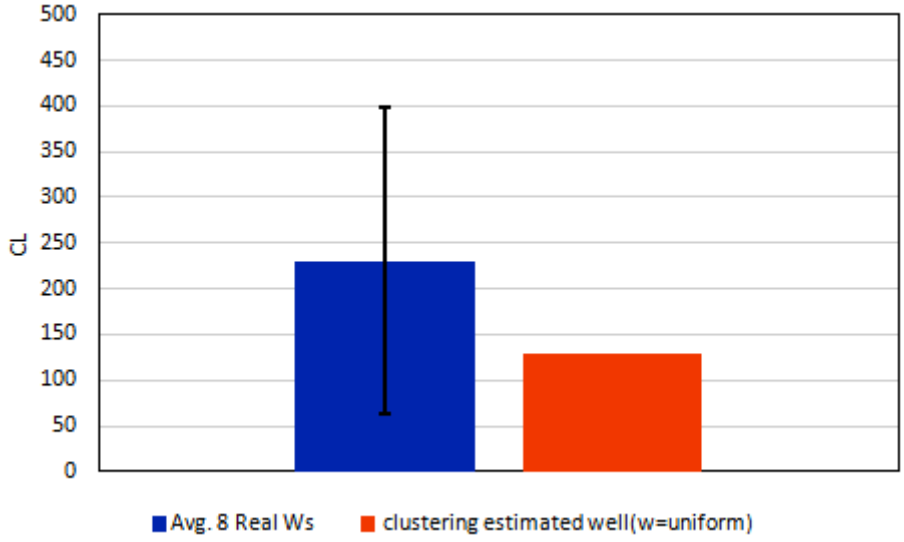

Figure 24. Estimated Chloride Content

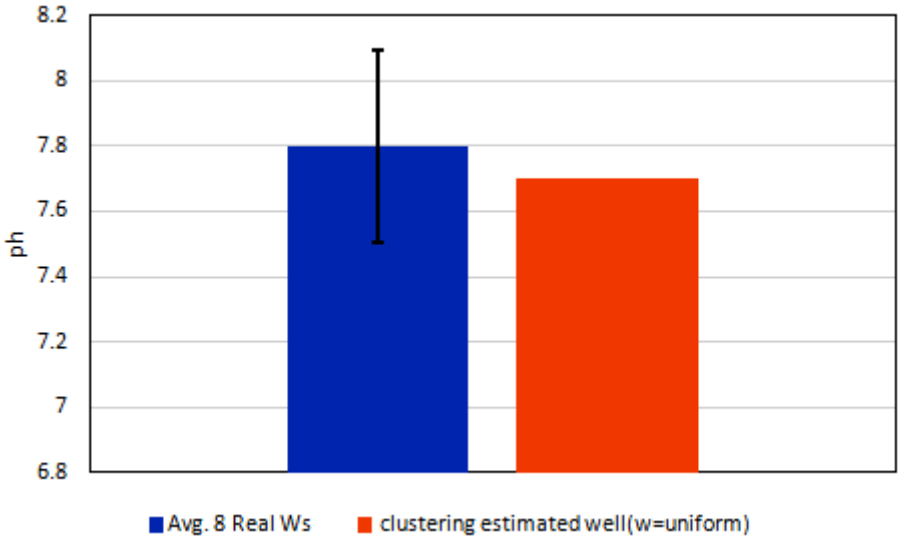

Figure 25. Estimated pH of water

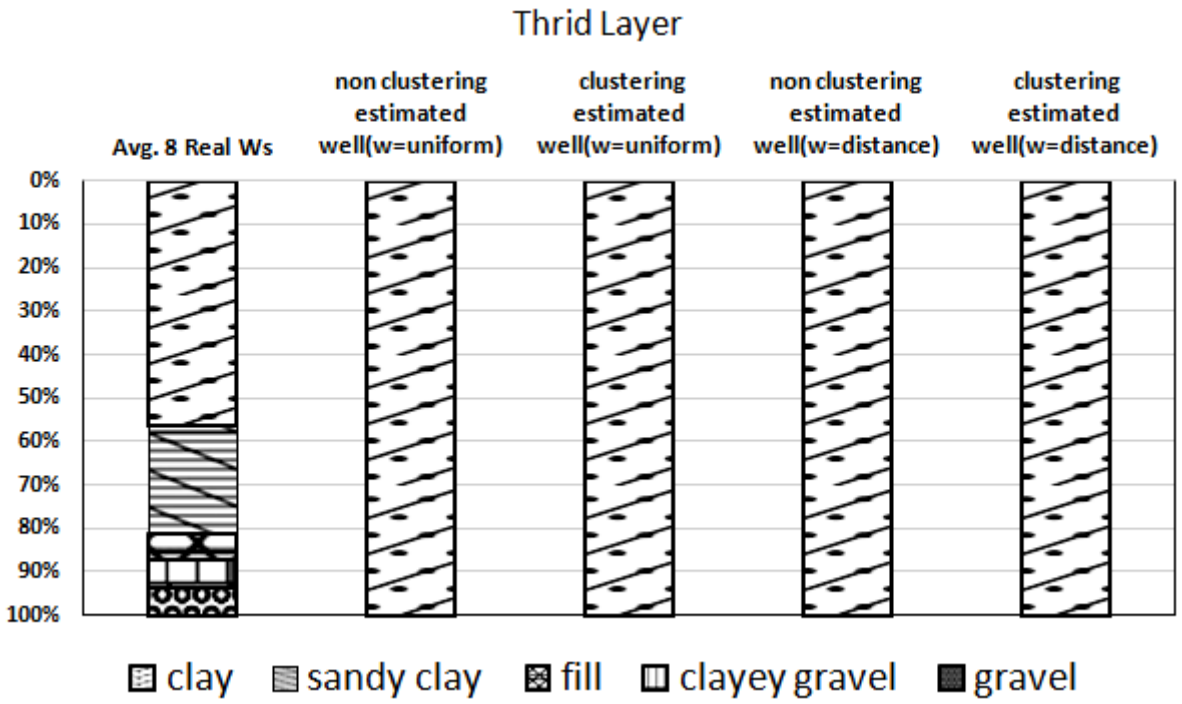

Figure 26. lithology of the third layer for an average of eight nearest wells

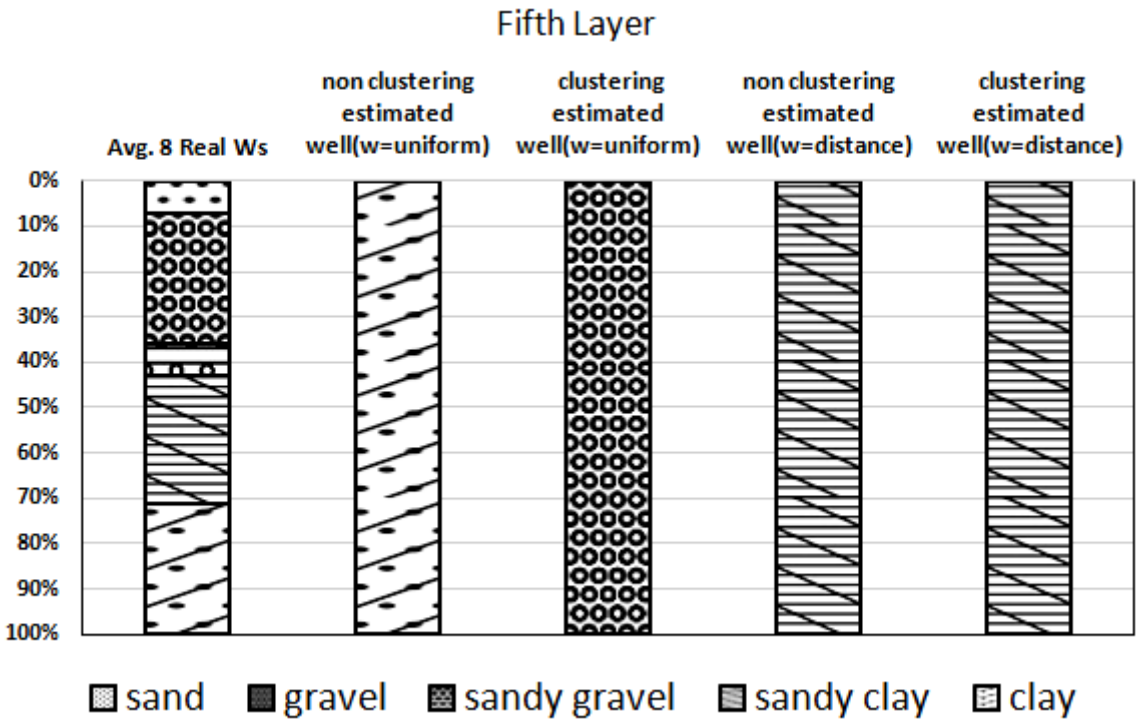

Figure 27. Lithology of the fifth layer for an average of eight nearest wells

The scope of this study is to compare the expected outcomes with the nearest real wells using the models with and without clustering. Therefore, showing all predicted outcomes in the same illustration may be a good idea.

In Figure 27, three soil types dominate the fifth constitutive layer, these soils are gravel, sandy clay, and clay. In non-clustering estimated wells, the uniform weight model forecasts this layer based on the last part of the constitutive fifth layer. At the same time, the outcome of the distance weight model is sandy clay, the same as the clustering estimated well with also distance weight as well as the fact that the Figure presents the clustering estimated well-predicted fifth layer of gravel majority soils of the lithology of the tenth layer for the average real wells, are gravel, gravely clay, and sandy

clay. They are about 30% for each soil type. Therefore, the expected lithology of the predicted wells is gravel.

Figure 28 illustrates the constitutive tenth layer of eight real wells average, as the reference layer compares it with the predicted layer using different KNN models. It can be seen that the gravel, gravely clay and sandy clay are the significant parts. Both models of distance weight with and without clustering predict the same soil type in the tenth layer, gravely clay. This soil represents about 25% of the constitutive layer and is at the mid-depth of the reference layer. The non-clustering estimated well with the uniform weight model predicts a gavel layer as the upper part of the reference layer. Unlikely, the clustering estimated well with uniform weight model forecasts rock layers. The reference constitutive layer does not have any rock part because the prediction process for this layer is based on the three centrally estimated wells rather than the nearest wells, as already revealed. Figure 29 shows how the clustering estimated well with uniform weight will be rock, which is 100% based on the estimated center well _2.

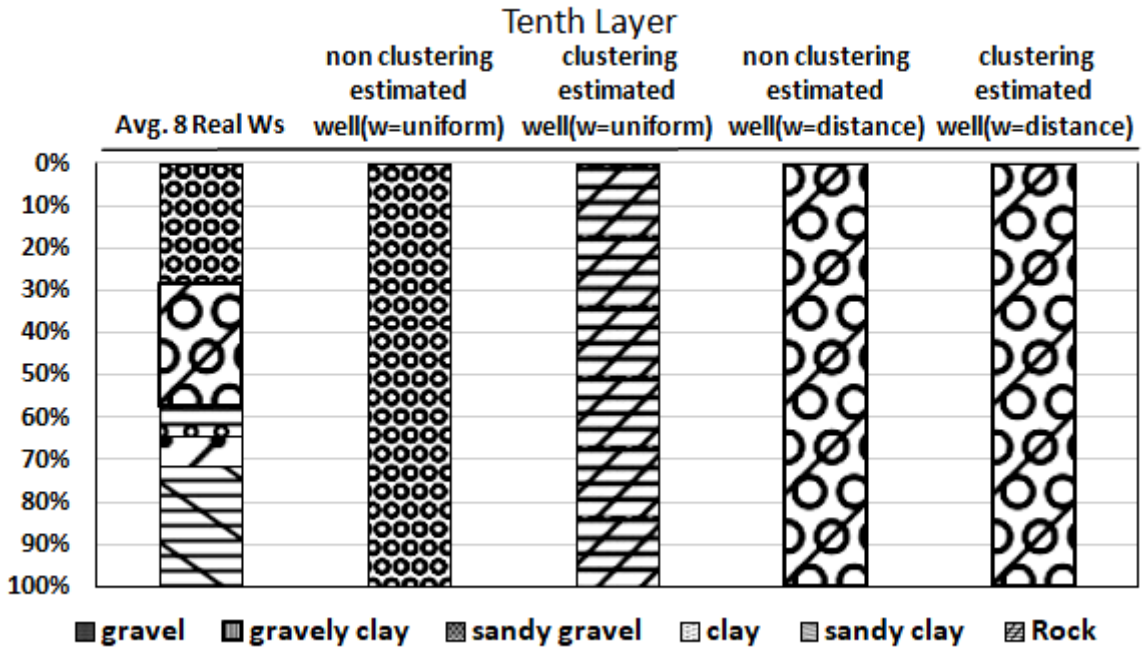

Figure 28. Lithology of the tenth layer for an average of eight nearest wells

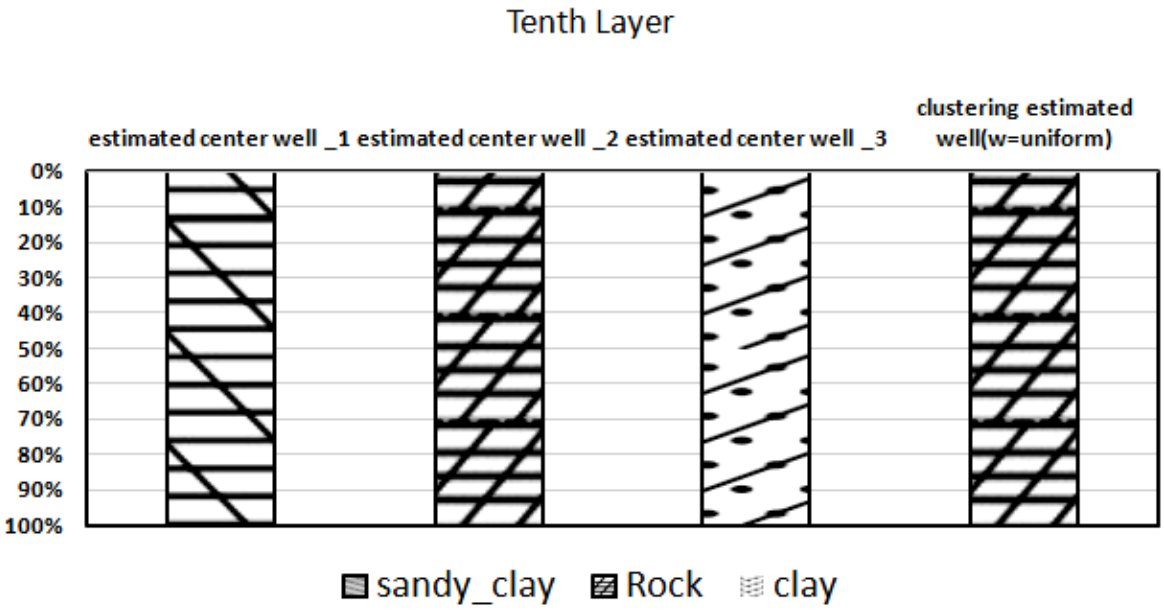

Figure 29. Lithology of the tenth layer for the clustering estimated wells

3.2 Water Quality Prediction

The developed model was also used to predict important groundwater quality parameters such as total dissolved solids (TDS), calcium (Ca), magnesium (Mg), turbidity, chloride (Cl^-), and pH. Unlike classification, these predicted values are continuous, and evaluation was performed through regression evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The predicted values are presented in their actual units for practical use, as listed in Table 3.

This indicates a high predictive accuracy, especially for TDS and pH, where an R^2 value greater than 0.90 was achieved. The relatively low performance for turbidity prediction can be attributed to its higher variability and sensitivity to localized conditions. It can be seen that the model is capable of effectively estimating groundwater quality parameters based on limited inputs, an essential requirement for resource-constrained areas.

3.3 Comparison with Standards

In order to understand the applicability of the predicted values of water quality, the predicted values were compared with the standards of drinking water established by the World Health Organization and the Bureau of Indian Standards, as shown in Table 4.. The comparison of predicted values with the standards established by the World Health Organization and the Bureau of Indian Standards indicates that most of the predicted values are well within the limits.

The comparison indicates that while most samples meet acceptable limits, certain locations approach or exceed threshold values, suggesting localized contamination or geogenic influence. These findings demonstrate the practical utility of the model in identifying potential risk zones for groundwater quality deterioration.

`The performance of the proposed clustering-assisted KNN model was compared with the conventional KNN approach, and the results are presented in Table 5. The accuracy obtained by the conventional KNN model was 82.4 percent, and the RMSE and R^2 values were 1.96 and 0.86, respectively. However, when clustering was incorporated into the model, a higher accuracy of 88.6 percent was obtained, along with a lower RMSE value of 1.42 and a higher R^2 value of 0.91. This improvement in accuracy and decrease in RMSE values confirm the effectiveness of the proposed approach in handling hydrogeological data and reducing errors during prediction.

3.4 Spatial Interpretation

Spatial analysis of the predicted results for lithology and water quality parameters also provides interesting insights into the hydrogeological conditions prevailing in the study area. Maps are interpolated from the predicted results and indicate different patterns in different areas, depending on the geological conditions and groundwater flow patterns.

Table 3. Performance of Water Quality Prediction Model

Parameter	Unit	MAE	RMSE	R^2 Score
TDS	mg/L	42.5	58.3	0.91
Calcium (Ca)	mg/L	6.8	9.5	0.88
Magnesium (Mg)	mg/L	5.2	7.4	0.86
Turbidity	NTU	0.9	1.3	0.84
Chloride (Cl^-)	mg/L	18.6	25.1	0.89
pH	–	0.21	0.30	0.92

Table 4. Comparison of Predicted Values with WHO/BIS Standards

Parameter	Unit	Predicted Range	WHO Limit	BIS Limit
TDS	mg/L	350–950	1000	500 (acceptable), 2000 (permissible)
Calcium (Ca)	mg/L	40–120	200	75 (acceptable), 200 (permissible)
Magnesium (Mg)	mg/L	20–80	150	30 (acceptable), 100 (permissible)
Turbidity	NTU	0.5–4.5	5	1 (acceptable), 5 (permissible)
Chloride (Cl^-)	mg/L	100–320	250	250 (acceptable), 1000 (permissible)
pH	–	6.8–8.2	6.5–8.5	6.5–8.5

Table 5. Performance comparison of KNN and clustering-assisted KNN models for lithology and water quality prediction

Model	Accuracy (%)	RMSE	R^2
KNN	82.4	1.96	0.86
Cluster + KNN	88.6	1.42	0.91

For instance, areas with a higher proportion of sandy geological formations tend to have lower concentrations of TDS and chloride, indicating a better quality of groundwater due to its higher permeability and flushing effects. On the other hand, areas with a higher proportion of clay-rich geological formations tend to have a higher concentration of dissolved solids, possibly due to lower permeability and a longer residence time of groundwater. Improved visualization tools were also used to improve the readability of the maps, such as adding legends and scale bars, and ensuring uniform units for all maps. This helps in validating the model results and also provides a powerful tool for decision-makers to identify areas of interest for groundwater exploration and management.

3.5 Discussion with Literature

The results obtained in this study are consistent with trends reported in previous research on machine learning applications in hydrogeology. Studies by Kumar *et al.* [33] and Tian *et al.* [34] have highlighted the effectiveness of data-driven models in capturing nonlinear relationships in groundwater systems, particularly for quality prediction. Similarly, Rustum *et al.* [35] demonstrated the advantage of hybrid approaches in improving model robustness, which aligns with the clustering-assisted framework proposed in this work. The achieved accuracy and regression performance are comparable to those reported by Yuan *et al.* [36] and Piras *et al.* [37], despite the use of a smaller dataset and simpler model architecture.

The major contributions of the proposed approach are the demonstration that high-performance results can be achieved without the need to use complex models, especially in the case where the data set is smaller, as is often the case in developing regions. Moreover, the proposed approach, which combines the estimation of lithology and water quality, provides a more complete picture of the subsurface conditions, as opposed to the results shown in the previous studies. The findings also highlight the role of spatial proximity and geological similarity in governing groundwater properties, thus validating the basic premise of the KNN method. Overall, the developed methodology presents an optimal balance of accuracy, interpretability, and efficiency, thus providing an effective solution for hydrogeological problems.

To promote reproducibility and transparency, the study has clearly specified all the configurations and implementation details of the developed models. The K-Nearest Neighbors algorithm has been used to develop the predictive models using the Scikit-learn library in Python. The number of nearest neighbors has been set to k = 3, Euclidean distance used to evaluate similarity, and uniform weights assigned to each neighbor. A constant random state has been used to initialize the clusters with the value set to 42. Moreover, the preprocessing techniques, including normalization and mean imputation for missing values, have been performed using a unified workflow for both training and validation sets, thus preventing any potential issues of data leakage.

Besides this proposed clustering-assisted KNN approach, another baseline model based purely on KNN without clustering was tested to ascertain the efficacy of this hybrid approach. From the comparative analysis, it can be seen that clustering plays an important part in improving the accuracy of prediction. There is an improvement in accuracy from 82.4% to 88.6%, along with a decrease in RMSE and an increase in R^2 values. This clearly indicates that clustering is an important part of this hybrid approach in capturing the heterogeneity of the data, especially for small datasets. In terms of the reliability of lithology prediction, a quantitative validation method was incorporated for analysis. The predicted lithology sequence was tested for accuracy in relation to actual borehole logs using accuracy analysis based on individual layers. The overall accuracy was recorded at 88.6%. In addition, it was checked for consistency over intervals to ensure that there was no sudden transition between layers, thus maintaining physical continuity. From this analysis, it can be stated that not only is this model effective statistically, but it is also physically correct, thus providing greater reliability for its application.

4. Conclusion

This study proposes a novel clustering-assisted K-Nearest Neighbors (KNN) model to accurately predict the multilayered lithology and groundwater quality using scarce borewell data. The model was found to have an accurate classification rate of 88.6% ± 3.2% in the classification of the lithology, while the predictions made by the model in the case of the groundwater quality have shown high correlation with the actual values, with the R^2 value being greater than 0.90 in the case of parameters such as TDS and pH. The values are found to be within the permissible limits set by the WHO and

BIS, with slight violations, indicating local conditions. The inclusion of the clustering algorithm helped stabilize the predictions made by the model, thus indicating the potential of the KNN model in hydrogeology. The model is found to be more accurate and less complicated compared to other models, making the model an efficient solution to the problem. Even though the model is affected by the limitations of the scarce data, the model is found to be reliable in giving insights into the subsurface conditions.

References

[1] U.B.P.S. Rathore, B. Sajan, S.K. Singh, S. Kanga, Urbanization and Water Stress: Analyzing the Impact of Rapid Urbanization on Local Water Resources and Proposing Sustainable Management Strategies. In Agri-informatics and Eco-friendly Innovations for a Secure Food Future Cham: Springer Nature Switzerland, 14, (2025) 353-374. https://doi.org/10.1007/978-3-032-02118-2_14

[2] R.K. Mishra, Fresh Water Availability and its Global Challenge. British Journal of Multidisciplinary and Advanced Studies, 4(3), (2023) 1-78. https://doi.org/10.37745/bjmas.2022.0207

[3] T. Pointet, The United Nations World Water Development Report 2022 on Groundwater, a Synthesis. Lhb, 108(1), (2022) 2090867. https://doi.org/10.1080/27678490.2022.2090867

[4] K.J. Hokanson, C.A. Mendoza, K.J. Devito, Interactions between Regional Climate, Surficial Geology, and Topography: Characterizing Shallow Groundwater Systems in Subhumid, Low-Relief Landscapes. Water Resources Research, 55(1), (2019) 284-297. https://doi.org/10.1029/2018WR023934

[5] S.U. Wali, A.A. Usman, A.B. Usman, U. Abdullahi, I.U. Mohammed, J.M. Hayatu, Impact of Geology on Hydrogeological and Hydrochemical Characteristics of Groundwater in Tropical Environments: A Narrative Review. International Journal of Hydrology, 8(6), (2024) 202-221. https://doi.org/10.15406/ijh.2024.08.00392

[6] A. Binley, S.S. Hubbard, J.A. Huisman, A. Revil, D.A. Robinson, K. Singha, L.D. Slater, The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. Water resources research, 51(6), (2015) 3837-3866. https://doi.org/10.1002/2015WR017016

[7] M. Hasan, L. Su, Novel Insights into Deep Groundwater Exploration by Geophysical Estimation of Hard Rock Permeability. EGUsphere, 30(5), (2026) 1309-1332. https://doi.org/10.5194/egusphere-2024-4191

[8] F. Yang, M. Hasan, Y. Shang, A Novel Geophysical Approach for 2D/3D Fresh-Saline Water Assessment toward Sustainable Groundwater Monitoring. Sustainability, 18(1), (2026) 517. https://doi.org/10.3390/su18010517

[9] O. Davis IMUERE, Electromagnetic (EM) Methods in Exploration: Advantages and Challenges. Multi-Disciplinary Research and Development Journals Int'l, 8(1), (2026) 97-106.

[10] A. Hussain, A.H. Sakhaei, M. Shafiee, Machine Learning-Based Constitutive Modelling for Material Non-Linearity: A review. Mechanics of Advanced Materials and Structures, 33(1), (2026) 2439557. https://doi.org/10.1080/15376494.2024.2439557

[11] X. Feng, L. Liu, M. Ye, O. Masek, S. Gouda, K. Chang, Q. Huang, Unveiling and Interpreting the Relationships Among Multi-Pollutant Emission Factors in Municipal Solid Waste Incineration by Machine Learning. Waste Management, 210, (2026) 115256. https://doi.org/10.1016/j.wasman.2025.115256

[12] S.A. Boateng, J. Xi, M.P. Fumey, J.K. Kumi, Exploring the Nonlinear Relationship of Environmental Sustainability Factors and Economic Growth in West Africa: Novel Machine Learning Evidence. Sustainable Development, 34, (2026) 1197-1220. https://doi.org/10.1002/sd.70216

[13] A.G. Usman, H.M. Almongy, I.A. Mahmoud, A.M. Jibrin, J. Usman, M.S. Samsudin, S.I. Abba, E.M. Almetwally, Optimized Ensemble Techniques for Nitrate Concentration Modelling from Groundwater Integrated with Mrmr Extraction Algorithm. Journal of Radiation Research and Applied Sciences, 19(1), (2026) 102217. https://doi.org/10.1016/j.jrras.2026.102217

[14] R. Narsing, S.C. Konnoju, Generalized Reciprocal Based Tversky Indexive Support Vector Extreme Boost Classification for Water Quality Prediction Analysis. Water Resources Management, 40(4), (2026) 166. https://doi.org/10.1007/s11269-025-04465-3

[15] M. Siena, M. Riva, Impact of Geostatistical Reconstruction Approaches on Model Calibration for Flow in Highly Heterogeneous Aquifers. Stochastic Environmental Research and Risk Assessment, 34(10), (2020) 1591-1606. https://doi.org/10.1007/s00477-020-01865-2

[16] S. Misra, H. Li, J. He, Machine Learning for

Subsurface Characterization. Gulf Professional Publishing. (2019).

[17] Y. Wang, C. Shi, X. Li, Machine Learning of Geological Details from Borehole Logs for Development of High-Resolution Subsurface Geological Cross-Section and Geotechnical Analysis. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16(1), (2022) 2-20. https://doi.org/10.1080/17499518.2021.1971254

[18] O.H. Kombo, S. Kumaran, Y.H. Sheikh, A. Bovim, K. Jayavel, Long-Term Groundwater Level Prediction Model based on Hybrid KNN-RF Technique. Hydrology, 7(3), (2020) 59. https://doi.org/10.3390/hydrology7030059

[19] T. Xie, L. Chen, B. Yi, S. Li, Z. Leng, X. Gan, Z. Mei, Application of the Improved K-Nearest Neighbor-Based Multi-Model Ensemble Method for Runoff Prediction. Water, 16(1), (2024) 69. https://doi.org/10.3390/w16010069

[20] T.N. Navya, G. Ramkumar, A Methodical Outlook of Early Floods in an Uncertain Weather Forecasts using Igneous K-Nearest Neighbor Classifier. In AIP Conference Proceedings, 3383(1), (2026) 020017. https://doi.org/10.1063/5.0308580

[21] S. Ethaib, M. Fahs, H. Mishbak , M.N. Fares, J.S. Makki, A.Alhello, H. Abbood, S.N. Abdel Hassan, A.A. Alrijabo, M. Azaroual, H.M.Baalousha, N. Baghdadi, P. Blanc, J. Duclos, L. Drapeau, N. Hariri, H. Hussein, W.J. Hassan, T.E. Hussien, F. Lehmann, F.Le Ber, M.S. Mizel, R. Mohsin, A. Nasser, T. Nasser, A.F. Al-Ma'athedi, A. Raeis, R, Toussaint, A.W. Ngnien, A. Younes, K. Del Vecchio, A. Al Bitar, Water Resources in South of Iraq: Current State, Future Evolutions, Challenges, and Potential Solutions. Hydrology, 13(3), (2026) 87. https://doi.org/10.3390/hydrology13030087

[22] B. Saaeidi, Assessment of Groundwater Contamination by Heavy Metals (Cu, Pb, Cd, Cr, Co) in the Eastern Region of Wasit Governorate. Dijlah Journal of Agricultural Sciences, 5(1), (2026) 52-63.

[23] H.I.Z. Al-Sudani, Groundwater Utilization and Water Quality in Khanaqin District, Diyala Governorate, Northeast of Iraq. Resources Environment and Information Engineering, 6(1), (2024) 305-312. https://doi.org/10.25082/REIE.2024.01.004

[24] V.B. Prasath, H.A.A. Alfeilat, A. Hassanat, O. Lasassmeh, A.S. Tarawneh, M.B. Alhasanat, H.S.E. Salman, (2017) Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier-a Review. arXiv preprint arXiv:1708.04321

[25] M. Sakizadeh, R. Mirzaei, A comparative Study of Performance of K-Nearest Neighbors and Support Vector Machines for Classification of Groundwater. Journal of Mining and Environment, 7(2), (2016) 149-164. https://doi.org/10.22044/jme.2016.480

[26] R.K. Halder, M.N. Uddin, M.A. Uddin, S. Aryal, A. Khraisat, Enhancing K-Nearest Neighbor Algorithm: A Comprehensive Review and Performance Analysis of Modifications. Journal of Big Data, 11(1), (2024) 113. https://doi.org/10.1186/s40537-024-00973-y

[27] J. Yu, J. Amores, N. Sebe, P. Radeva, Q. Tian, Distance Learning for Similarity Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(3), (2008) 451-462. https://doi.org/10.1109/TPAMI.2007.70714

[28] X. Lin, Z. Tian, A. Chong, Y. Lu, J. Niu, N. Deng, A Data Informativeness Evaluation Method for Grey-Box Modeling of Building Thermal Dynamics. Energy and Buildings, (2026) 117103. https://doi.org/10.1016/j.enbuild.2026.117103

[29] Y.R. Lin, H.M. Wu, Image Generator For Tabular Data based on Non-Euclidean Metrics for CNN-Based Classification. PLoS One, 21(1), (2026) e0340005. https://doi.org/10.1371/journal.pone.0340005

[30] M.N. Koliaraki, N. Smyrnis, P. Asvestas, G.K. Matsopoulos, E.C. Ventouras, Saccadic Eye Movements Based Classification of Patients with Obsessive-Compulsive Disorder, Patients with Schizophrenia and Healthy Controls using Artificial Neural Networks. Cognitive Neurodynamics, 20(1), (2026) 41. https://doi.org/10.1007/s11571-026-10414-6

[31] I.D. Mienye, Y. Sun, A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. IEEE Access, 10, (2022) 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287

[32] Biswakalyani, C., Samantaray, S., & Satpathy, D. P. (2026). Application of Hybrid Machine Learning for Groundwater Level Prediction: A Comprehensive Review. Archives of Computational Methods in Engineering, 1-59.

[33] A. Kumar, M.L. Nehdi, Data-Driven Approaches to Groundwater Modelling: Methods, Applications, and Challenges. Hydrological Insights, (2026) 1-10. https://doi.org/10.1007/s11831-025-10447-w

[34] J. Tian, X. Zeng, D. Wang, J. Wu, A data-driven approach coupled with physical constraints to improve groundwater models with structural

error. Water Resources Research, 62(3), (2026) e2025WR040247. https://doi.org/10.1029/2025WR040247

[35] S. Rustum, U. Habib, S. Ahmed, M. Usman, M.A. Qureshi, Clustering-Assisted Channel Estimation for Free-Space Optical Satellite Communication. Optics Communications, (2026) 133011. https://doi.org/10.1016/j.optcom.2026.133011

[36] E.C.Y. Yuan, Y. Liu, J. Chen, P. Zhong, S. Raja, T. Kreiman, S. Vargas, W. Xu, M. Head-Gordon, C. Yang, S.M. Blau, Foundation Models for Atomistic Simulation of Chemistry and Materials. Nature Reviews Chemistry, (2026) 10, 212–230. https://doi.org/10.1038/s41570-025-00793-5

[37] G. Piras, F. Muzi, Z. Ziran, Assessment of the Reliability of AI Models in Predicting Urban Energy Consumption Under Conditions of Small or Incomplete Data. Applied Sciences, 16(3), (2026) 1457. https://doi.org/10.3390/app16031457

Authors Contribution Statement

Khabeer Al-Awad: Conceptualization, Methodology, Investigation, Formal analysis, Writing original Manuscript. Hayder Algretawee: Data curation, Validation, Writing original Manuscript. Alaa M Shaban: Validation, Supervision, Writing review and Editing. All the authors read and approved the final version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes