Ensemble learning paradigms for flow rate prediction boosting

Laurent Kouao Kouadio  
Central South University

Jianxin Liu  
Central South University

Serge Kouamelan Kouamelan  
University of Felix Houphouet-Boigny: Universite Felix Houphouet-Boigny

Rong Liu  (✉ liurongkaoyan@csu.edu.cn)  
Central South University

Research Article

Keywords: ensemble machine learning, electrical method, groundwater, flow, borehole.

Posted Date: February 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2520334/v1

License: ☺️ ));// This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Ensemble learning paradigms for flow rate prediction boosting

Kouao Laurent KOUADIO¹,²,³, Jianxin LIU¹,², Serge Kouamelan KOUAMELAN³, Rong LIU¹,²

¹ School of Geosciences and Info-physics
   Central South University
   Changsha, Hunan 410083, China
   lkouao@csu.edu.cn, ljx@csu.edu.cn, liurongkaoyan@csu.edu.cn,

³Hunan Key Laboratory of Nonferrous Resources and Geological Hazards Exploration
   Changsha, Hunan 410083, China

³UFR des Sciences de la Terre et des Ressources Minières
   Université Félix Houphouët-Boigny,
   Abidjan, 22 BP 582 Abidjan 22, Cote d’Ivoire
   kouamelan2006@yahoo.fr

Corresponding Author:

Rong LIU (liurongkaoyan@csu.edu.cn)

Submitted to

Water Resources Management

January 30, 2023
Abstract

In developing countries, climate change has considerably affected population welfare by increasing drinking water scarcity. Global organizations and governments have initiated many drinking water supply projects to fight against this issue. Most of these projects are led by geophysical companies in partnership with drilling ventures to locate drillings expected to give the recommended flow rate (FR). Known as cheap methods, electrical resistivity profiling (ERP) and vertical electrical sounding (VES) were the most preferred. Unfortunately, the project objective was not achieved due to numerous unsuccessful drillings, thereby creating a huge loss of investments. To reduce the repercussion of unsuccessful drillings, we introduced the ensemble machine learning (EML) paradigms composed of four base learners. The aim is to predict at least 80% of correct FR in the validation set before any drilling operations. Geo-electrical features were defined from the ERP and VES and combined with the collected boreholes data to compose the binary dataset ($FR \leq 1m^3/hr$ and $FR > 1 m^3/hr$ for unproductive and productive boreholes respectively). Then, the dataset is transformed before feeding to the EMLs. As a result, the benchmark and the pasting EMLs performed 85% of good predictions on the validation set whereas the extreme gradient boosting and the stacking performed 86% and 87% respectively. Finally, the correct prediction of FRs will reduce the losses in investment beneficial for funders and state governments, and geophysical and drilling ventures.

Keywords: ensemble machine learning, electrical method, groundwater, flow, borehole.
1. Introduction

Historically, groundwater constitutes the best source of potable water outside of human-made pollution (Faillat 1986; Biemi 1992; White et al. 2016; Anaba Onana et al. 2017). It has become a scarce commodity over past the years. As a result, the population welfare has considerably degraded due to the climate change effect (AMCOW 2008, 2011; Esnault et al. 2014; Singh et al. 2015; Jakeman et al. 2016). To solve this issue, many campaigns for drinking water supply (CDWS) were recommended by governments in collaboration with global organizations (Nomquphu 2005; AMCOW 2008; MHCI 2012; UNICEF 2017; United Nations 2019; Bayu et al. 2020). The geophysical with drilling ventures were the selected companies to supply villages and cities with potable water (Chinyem 2017). During the CDWS, electrical resistivity profiling (ERP) and vertical electrical sounding (VES) were the favored methods because they are cheap and fast to respect the project timeline (Mohamaden and Ehab 2017; Adagunodo et al. 2018; Olanrewaju and Abdulkadir 2020). However, the numerous unsuccessful drillings recorded during these projects have indicated that the aim goal was far from satisfactory. These are dry boreholes and unsustainable boreholes (i.e., out-of-service boreholes after a few years of usage). They also include the drilling with flow rate (FR) under the project required FR (RFR) (Soro et al. 2020). Indeed, the drilling operations are made after the geophysical survey and the objective is to find the best location expected to get at least the RFR. Thus, each unsuccessful drilling has required new geophysical investigation which increased the project budgets. As a consequence, a lot of investments from the governments and funders as well as the profits for geophysical and drilling companies have been lost (Bjornlund 2004; Xepapadeas and Koundouri 2004). To address this problem while using
ERP and VES, with a reasonable amount of data, we proposed an ensemble machine learning (EML) method composed of four base learners (BLs) algorithms: (i) the Logistic Regression (LR) (Cox, 1958, Geron, 2019; Raschka and Mirjalili, 2019), (iii) K-nearest neighbors (KNN) (Altman, 1992; Grus, 2015), (ii) Decision Tree (DT) (Ho, 1995; Wu et al., 2008; Nath & Levinson 2014, Breiman 2019, and Yariyan et al., 2020) and (iv) the pre-trained Support Vector Machines (pSVM) of Kouadio et al. (2022). The latter corresponds to the Gaussian radial basis function (RBF) kernel of SVMs (Vapnik 1998; Chen et al. 2006; List and Simon 2009; Chang and Lin 2011; Murphy 2012; Xu et al. 2014). The aim is to find the EML with good performance capable to achieve at least 80% of correct FR prediction on the validation set with a very low error useful for the future CDWS before the drilling operations.

Indeed, the EML is a machine learning (ML) paradigm that used multiple learners to build a set of hypotheses and combines them to solve the same issue (Pedregosa et al. 2011). Whilst, the ordinary ML approaches learn one hypothesis from the training data (Zhou 2007). The approach introduced in this workflow first entails using electrical features defined from the ERP and the VES by Kouadio et al., (2022) such as type, shape, magnitude, power, pseudo-fracturing index, and ohmic-area additional to the geology of the area to compose the geo-electrical features. Secondly, the observed FRs such as the unproductive boreholes (FR0) composed of the dry boreholes \( FR = 0 \ m^3/hr \), the unsustainable boreholes \( 0 < FR \leq 1 \ m^3/hr \) and the productive boreholes FR1 \( FR > 1 \ m^3/hr \) are categorized into two classes. They are combined with the geo-electrical features to generate the binary dataset. The latter is first trained separately with the BLs except for the pSVM. Secondly, the BLs with the optimal hyper-parameters are aggregated
with the pSVM to compose the base estimator of EMLs. The latter is trained in turn and evaluated on the validation set and test sets for performance generalization.

The EML paradigms approach is implemented using data from an experiment area affected by climate change and also faced a significant drinking water scarcity to test its efficiency. Henceforth, the approach could be considered as an alternative opportunity to select the suitable drilling location expected to obtain the RFR before any drilling operations. Moreover, it will considerably minimize the numerous unsuccessful drilling and unsustainable boreholes profitable for geophysical and drilling ventures. Furthermore, it will reduce the loss of investments by global organizations and governments, and indirectly diminish the problem of water scarcity helpful for the population welfare.

2. Background

EML has already been used in diverse applications such as groundwater modeling (Zhou 2007; Rojas et al. 2008), groundwater head prediction (Li and Tsai 2009; Sahoo et al. 2017), groundwater potential modeling (Nguyen et al. 2020; Liao et al. 2023), and groundwater storage change prediction (Yariyan et al. 2020; Yin et al. 2021). In recent times, numerous studies in the literature have been conducted using the EML methods for prediction purposes such as drought forecasting (Li et al. 2020; Salim et al. 2023) using boosting algorithms (Extreme Gradient Boosting (XGB) and AdaBoost (Pedregosa et al. 2011; Tharwat 2018; Li et al. 2023)) and tree-based algorithms (Yariyan et al. 2020). XGB was also used for monitoring drought damage coupled with the Random-forest (Ho 1995; Breiman 2019; Buthelezi et al. 2020). Thus, EMLs try to find several possible solutions by weighing many data-driven individual paradigms (e.g., ML, Bayesian models (Barber 2012;
EMLs combined their results to increase the performance of the individual BLs models (Zounemat-kermani et al. 2019). In this workflow, we used a set of BLs as a base estimator of EMLs for FR prediction. Therefore, it is better to give a synopsis of each BLs for clarity in the following.

2.1. BL algorithms

Commonly, BLs should be accurate and diverse as possible to compose the base estimator of EMLs. They are combined to address a variety of real-world (Zhou 2007; Nath and Levinson 2014; Breiman 2019; Zounemat-kermani et al. 2021). However, the goal is not to aggregate multiple ML algorithms but to smartly select the optimal ones which at the same time obey the law of wisdom of crowds (Surowiecki 2005) and the theorem of “many could be better than all” (Zhou 2007). The author demonstrated that more BLs will not always lead to better performance. The selected BLs for this workflow are listed in Table 1 with their corresponding hyper-parameters. Except for pSVM, the other BLs are evaluated on the training set to get the optimal hyper-parameters before aggregating them as a base estimator of EMLs.
Table 1: BL hyper-parameters for fine-tuning. Each BL has extra parameters; however, the selected parameters are not exhaustive and are considered the most interesting ones for fine-tuning (Buitinck et al. 2013).

<table>
<thead>
<tr>
<th>BL</th>
<th>Hyper-parameters</th>
<th>Fine-tuning value range ((\exists \lambda \in \mathbb{N}))</th>
<th>Descriptions</th>
</tr>
</thead>
</table>
| LR  | \(p\) \{l: l \in \{l_1, l_2\}\} | Penalty norms (Geron 2019);  
   - \(l_1\): Manhattan norm noted \(\|\cdot\|_1\);  
     \[l_1 = \frac{1}{N} \sum_{i=1}^{n} |y_i - \hat{y}_i|\]  
   - \(l_2\): the Euclidean norm noted \(\|\cdot\|_2\);  
     \[l_2 = \frac{1}{N} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2\]  
   where \(N\) is the number of observations/rows in the dataset. The difference between actual \(y\) and predicted \(\hat{y}\) values take on every \(i\) value ranging from 1 to \(n\) (Pedregosa et al. 2011; Geron 2019) |
|     | \(C\) \{c: c = 10^4, \lambda \in \{-3, -2, \ldots, 3\}\} | Inverse regularization strength. Smaller is stronger regularization (VanderPlas 2016) |
|     | \(N\) \{n: n = \lambda, \lambda \in \{2, 3, \ldots, 10\}\} | Number of neighbors (Altman 1992; Grus 2015) |
| KNN | \(M\) \{m: \sum_{i=1}^{n} (x_i - y_i)^2\} | Distance metrics \(m\):  
   - Euclidian: \(d_{euc}(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2\)  
   - Manhattan: \(d_{man}(x, y) = \sum_{i=1}^{n} (x_i - y_i)\)  
   where \(n\) is a number of variables, \(x_i\), and \(y_i\) are the variables of vectors \(x\) and \(y\) respectively, in the two-dimensional vector space. i.e., \(x = (x_1, x_2, x_3, \ldots)\) and \(y = (y_1, y_2, y_3, \ldots)\). |
|     | \(p\) \{p: \sum_{i=1}^{n} (x_i - y_i)^2\} | Minkowski power parameters (Harrison 2019) |
| **DT** | $\mathcal{C}$ | $\{c : c \in \{\mathcal{G}, \mathcal{E}\}\}$ | Splitting function:  
- $\mathcal{G}$: Gini impurity: $\mathcal{G}_i = 1 - \sum_{k=1}^{n} P_{i,k}^2$  
- $\mathcal{E}$: Entropy $\mathcal{E}_i = 1 - \sum_{k=1}^{n} \frac{P_{i,k}}{P_{i,k} \neq 0 \log_2 (P_{i,k})}$  

where $P_{i,k}$ is the fraction of class $k$-instances among the training instances in the $i$-th node (Geron 2019). |
| **M_{dt}** | $\{ m_x : m_x = \lambda, m_x \in \{2,3,\ldots,20\}\}$ | $M_{dt}$: Maximum depth of trees. It determines the greatest number of terminal nodes in each child (Buitinck et al. 2013; Naghibi et al. 2020). |
| **Kernel** | Gaussian radial basis function (RBF) | RBF kernel: $K(x_i, x_j) = \exp \left(-\gamma \| \phi_l(x_i) - \phi_l(x_j) \|^2 \right)$ where $x_i$ ($x_i \in \mathbb{R}^{n \in \mathbb{N}}, i = 1, \ldots, m$) is the training vectors, and $m \in \mathbb{N}$ is the set of examples. The dot product $\phi_l(x_i)^T \phi_l(x_j)$ is called the kernel trick and is based on the original vectors $x_i$ and $x_j$. (Cristianini and Shawe-Taylor 2000; Mohri 2017; Kouadio et al. 2022) |
| **pSVM** | $C$ | $C = 2.0000$ | $C > 0$ is the regularization parameter. It controls the number of errors to tolerate in the training set (Murphy 2012). |
| | $\zeta$ | $\zeta = 0.0010$ | $\zeta_i$ is a slack variable for each instance. It measures is allowed to level of margin violation of the $i$-th instance (Vapnik and Cortes 1995; Chang and Lin 2011; Murphy 2012) |
| | $\gamma$ | $\gamma = 0.1250$ | $\gamma$ acts like a penalty hyper-parameter $C$. (Murphy 2012) |
2.2. EML paradigms

We selected four EML methods including the benchmark EML to compose our framework based on probabilistic learning principles (Kim et al. 2019). These are:

- **Benchmark EML** ($B_0$)

  $B_0$ is a meta-learner using the majority vote technique resulting from the combination of the individual output of BLs. The type of voting algorithm implemented for performance estimating is “hard” voting (Buitinck et al. 2011; Geron 2019).

- **EML Pasting** ($Past$)

  The pasting ($Past$) algorithm trains every predictor on random subsets of the training set (Zounemat-kermani et al. 2021). The sampling is performed with no replacement and the majority vote is used for pooling the individual output from bootstrap samples (Raschka and Mirjalili 2019). The $Past$ paradigm is selected in this workflow because we are dealing with a small dataset to avoid introducing a significant bias (Wu et al. 2008).

- **EML Boosting** ($XGB$)

  XGB utilizes precise approximations to generate the best prediction model (Friedman 2001). Moreover, the main advantage of XGB is its speed. It also creates a variety of training data sets from random sampling (Friedman 2001; Georganos et al. 2018; Zounemat-kermani et al. 2021).

- **EML Stacking** ($Stc$)
Stc is the last specific category of EML and produces strong learners from poorer ones (Buitinck et al. 2011, 2013; Pedregosa et al. 2011). It achieves better performance, using a higher-level model to integrate lower-level models, and finally increases the predictive power of the model (Zounemat-kermani et al. 2021). The blender (also known as the final estimator) algorithm aggregates the predictions of all predictors in an ensemble.

Table 2 records all the most interesting EML hyper-parameters for fine-tuning.
Table 2: EMLs hyper-parameters. *yes* indicates that the EML uses the specific parameter value as an argument. ∨: Logical OR.

<table>
<thead>
<tr>
<th>Hyper-parameters</th>
<th>Fine-tuning values range</th>
<th>$B_0$</th>
<th>Past</th>
<th>$XGB$</th>
<th>$Stc$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootstrap</td>
<td>False</td>
<td>yes</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Sampling performs without replacement</td>
</tr>
<tr>
<td>BLs</td>
<td>{ LR, KNN, DT, pSVM }</td>
<td>yes</td>
<td>yes</td>
<td>–</td>
<td>yes</td>
<td>Set of BL with optimal hyper-parameters</td>
</tr>
<tr>
<td>$n_e$</td>
<td>$n: n = 50 \times \lambda, \lambda \in {1, 2, \ldots, 10}$</td>
<td>–</td>
<td>yes</td>
<td>yes</td>
<td>–</td>
<td>The number of rounds/estimators of boosted trees.</td>
</tr>
<tr>
<td>$M_{dt}$</td>
<td>$m_x: m_x = \lambda, m_x \in {2, 3, \ldots, 20}$</td>
<td>–</td>
<td>–</td>
<td>yes</td>
<td>–</td>
<td>Refer to the DT hyper-parameters description in Table 1</td>
</tr>
<tr>
<td>$\eta$</td>
<td>$e: e = \lambda/100, \lambda \in {1, 2, \ldots, 20}$</td>
<td>–</td>
<td>–</td>
<td>yes</td>
<td>–</td>
<td>Learning rate for boosting trees (between 0 and 1). (Friedman 2001; Harrison 2019)</td>
</tr>
<tr>
<td>$B$</td>
<td>{ Lin, Tree, DART }</td>
<td>–</td>
<td>–</td>
<td>yes</td>
<td>–</td>
<td>$B$ is the booster. It could be linear (Lin), tree (Tree), and Dropouts with multiple Additive Regression Trees (DART)(Friedman 2001; Harrison 2019).</td>
</tr>
</tbody>
</table>
\[ \gamma \left\{ \begin{array}{l} n: n = \frac{\lambda}{2}, \\
\lambda \in \{0, 1, \ldots, 20\} \end{array} \right\} \]

- - yes -

\(\gamma\) control the pruning. Minimum loss reduction is needed to further split a leaf. Higher \(\gamma\) is more conservative (Friedman 2001; Harrison 2019).

<table>
<thead>
<tr>
<th>Blender</th>
<th>(LR \lor KNN \lor DT \lor pSVM)</th>
<th>-</th>
<th>-</th>
<th>-</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The blender takes predictions from each BL with optimal hyper-parameters and makes the final prediction (Geron 2019).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Data and Methods

3.1. Dataset

The data for the experiment area is available on zenodo (Kouadio et al. 2021). It was collected in the Bagoue region in Cote d’Ivoire (West Africa) during the CDWS projects (Presidential Emergency Program in 2012-2013 and the National Drinking Water Supply Program in 2014 (Kra et al. 2016; Mel et al. 2017; Mel 2018)). It is composed of 431 ERP, 407 VES, and 431 boreholes data. Moreover, the array configurations of the ERP and VES were Schlumberger with 200 m for the current electrodes distance and 20 m for the potential electrodes.

- Features

Features are composed of categorical parameters such as shape, type, and the geology of the area. They also contained numerical parameters such as the pseudo-fracturing index, power, magnitude, and ohmic-area (Kouadio et al. 2022). The categorical features included the text labels indicating the category of the selected conductive zone for drilling operations (also called an anomaly). For instance, the shape categorical feature is composed of four categories of anomalies: W, U, V, K, and C. Likewise the feature type is composed of four labels such as EC (wide conductive zone), NP (narrow conductive zone), CB2P (conductive zone of two distinct resistive plans) and CP (conductive zone with stretched resistive plans). The final categorical feature is the geology of the area (geol) where the drilling operations are carried out. It gathered the volcano-sedimentary schists (VOLCS), the migmatite-gneiss (MIGG), the geosynclinal-granites (GEOG), and the granites (GRAN). Figure 1 shows the distributions of the features.
Supposing the raw matrix $X$ composed of $n$-defined parameters ($n=7$) with $m$-samples (selected anomalies, $m=431$), the predictor $X$ is given as

$$X = \{x_{ij} | 1 \leq i \leq m; \ 1 \leq j \leq n\} \quad (1)$$

Thus, the categorical features such as shape, type, and geol are vectorized ($vec$) (transformed to an integer vector $x$) to compose the $(m \times 3)$-submatrix $X_{cat}$ defined as

$$x^{(shape)}: \{W, U, V, K, C\} \xrightarrow{vec} \{1, 2, 3, 4, 5\}$$
$$x^{(type)}: \{EC, CB2P, NC, CP\} \xrightarrow{vec} \{1, 2, 3, 4\}$$
$$x^{(geol)}: \{VOLCS, MIGG, GEOG, GRAN\} \xrightarrow{vec} \{1, 2, 3, 4\}$$

$$\rightarrow X_{cat} \quad (2)$$
The target corresponds to the borehole data. Prior, it was globally separated into three classes: the dry boreholes FR0 \((FR = 0)\), the unsustainable boreholes \(FR1(0 < FR \leq 1 \text{ m}^3/\text{hr})\) and the productive boreholes \((FR > 1 \text{ m}^3/\text{hr})\). However, for this workflow, the problem is simplified into binary classes by considering the dry and unsustainable boreholes as unsuccessful drillings encoded to \(\{0\}\) (FR0). Whilst, the productive boreholes (successful drillings) are encoded to \(\{1\}\) (FR1) (see Fig. 1). Thus, the target \(y\) is given as

\[
y : \{\leq 1, > 1 \text{ m}^3/\text{hr} \} \overset{\text{vec}}{\longrightarrow} \{0, 1\} \tag{3}
\]

\(y \ {0:FR \leq 1 \text{ m}^3/\text{hr}}\) size represents 49.65% of the dataset and is almost equal to \(y \ {1:FR > 1 \text{ m}^3/\text{hr}}\) (50.35%) i.e., there is only a 50% chance to achieve the project objective \((RFR > 1 \text{ m}^3/\text{hr})\). This ratio validates the critical issue of potable water in that region. In the following, we denoted \(\mathcal{D} \in \mathbb{R}^{m \times (n+1)}\), the dataset composed of vectorized predictor \(X\) and the target \(y\).

### 3.2. Processing of \(\mathcal{D}\)

First, the rate of 5.57% missing data in the VES (see Fig.1) was imputed using the median value (Geron 2019; Moroney 2020). Thus, the imputed \(431 \times 7\)-matrix \(X\) of categorical features \(X_{\text{cat}}\) and numerical features \(X_{\text{num}}\) was processed separately. \(X_{\text{num}}\) is standardized whereas \(X_{\text{cat}}\) is one-hot-encoded (Pedregosa et al. 2011). The one-hot-encoding of \(X_{\text{cat}}\) generates a transitory matrix \(\Lambda\) of \(431 \times 15\) dimension (Buitinck et al. 2013; VanderPlas 2016; Harrison 2019). In addition, the dummy features (number equals...
and the redundant dummy vectors (number equals 2) are eliminated. Thus, Λ holds a new dimension of \(431 \times 10\). Then, the concatenation of Λ from the one-hot encoding process with the standardized \(X_{num}\) yields a transformed matrix \(X\) of \(431 \times 14\) dimension (Raschka and Mirjalili 2019). Therefore, the new dimension of \(\mathcal{D}\) becomes \(431 \times 15\) (\(y\) is added) for stratifying and sampling.

Secondly, because \(\mathcal{D}\) is a small and imbalanced dataset (with an unequal class), the stratification technique is used to preserve in each fold the equal proportions of class labels (Buitinck et al. 2013). This is relevant to guarantee the consistency of each fold based on the class proportion in the training and test sets (Buitinck et al. 2013; Harrison 2019). In this workflow, we put aside 20% of \(\mathcal{D}\) for testing and the 80% remaining for pieces of training as

\[
\mathcal{D}: (X|y) \rightarrow \begin{cases} (X_{train}|y_{train}) \equiv 80\% X|y \\ (X_{test}|y_{test}) \equiv 20\% X|y \end{cases}
\] (4)

Furthermore, to ascertain that the test set is consistent by keeping the relative proportion of the classes, the number of folds \(k\) is given as

\[
k = \left[ \frac{80\% m_{X_{train}}}{20\% m_{X_{test}}} \right] = 4
\] (5)

3.3. Parameters search and performance evaluation

The optimization randomized search technique (Bergstra and Bengio 2012) is used to explore the BL and EML hyper-parameters recorded in Table 1 and Table 2. Furthermore, the learning curve is used to address the overfitting and underfitting issues (Moroney 2020). It detects the level of bias or variance of each model to deduce whether
the collection of more data is necessary to address the unsuccessful drilling problem. Finally, the model with an acceptable bias and variance (commonly with low variance and low bias) with a high score on the whole CV results, is selected with a superscript (+). It should be used to evaluate the $X_{test}$ for error generalization. Moreover, EML performances are evaluated for error estimation using the following metrics:

- **Accuracy**

  It is the ratio of the sum of true positive ($TP$) and true negative ($TN$) over the sum of all terms (VanderPlas 2016):

  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$  \hspace{1cm} (6)

  where $FN$ and $FP$ are the numbers of false negatives and false positives respectively.

- **Precision and Recall (PR)**

  The PR is defined as the percentage of the positive FR that was correctly classified while the recall is the percentage of positive FR correctly classified (Powers 2007). Both are expressed as

  $$\text{PR} \triangleq \begin{cases} 
  \text{precision} = \frac{TP}{TP + FP} \\
  \text{recall} = \frac{TP}{TP + FN} 
  \end{cases}$$  \hspace{1cm} (7)

- **Confusion Matrix (CM) and Binary Predictor Error (BPE)**

  CM metric counts the number of times the instances of each FR category are confused with the different FR classes (Nath and Levinson 2014; VanderPlas 2016; Geron 2019). The BPE is another representation of the CM for ease of interpretation. Instead of plotting
the true FRs against the predicted FRs, it plots the true FRs against the number of predicted FRs by topping the error instances of FN and FP over the TP and TN (Harrison 2019).

4. Results

The results are structured into two parts. The first part is the model selection which consists to find robust models capable to generalize to the unseen data through the learning curves. The second part entails predicting the FR on the test data separately for global error estimation.

4.1. Models selection

Figure 2(a) shows the different performances on the BL learning curves after fine-tuning the BL hyper-parameters (see Table 1) across the stratified four-fold CVs. Overall, the training scores of $BLs^+$ are higher than the validation score everywhere. The latter is improving when increasing the training samples and also depends on the model complexity. For instance, from the validation curve of $LR^+$, the optimal tradeoff between the bias and the variance is almost found around 100 samples as shown by its convergence line. However, typically after 100 training samples, the training and the validation curves are already close to each other. This behavior is also visible in $KNN^+$ but less pronounced with a higher bias at the beginning (less than 50 samples) before it stabilizes around 80%. In addition, the fact that the convergence line is closest to the training and validation curves indicates that collecting more data will not help to improve the fit. Conversely, the learning curve of $DT^+$ shows a higher variance indicated by the gap observed between the
validation and training curves which explains a real overfitting case. Thus, using the DT in particular, it is clear that gathering more data for the more complicated model could be an alternative solution to have a good fit with an improved score. Conversely, pSVM globally shows a good fit, and its validation starts increasing as its training curve decreases, and both curves never cross the convergence line. However, it seems not a good idea to improve the convergence score by providing more training data.

Figure 2(b) indicates the EMLs model performances on the learning curves after aggregating the $BLs^+$ to compose the base estimator. The model bias-variance problem is diagnosed for varying numbers of training samples. Contrary to the $BLs^+$, the $EMLs^+$ performances have a bit of improvement. The filled part surrounding the training curve is obtained using the mean training score $\pm$ the standard deviation of the training scores (Buitinck et al. 2011). The same calculus is performed for the validation curves. It indicates the range of the model stability during the training phase and is useful for generalization purposes (VanderPlas 2016).

Overall, for training samples less than 100, the difference between training and validation accuracy is larger, and the $EMLs^+$ seriously are overfitting (higher variance). As the training size increases, the validation curves become closer to the convergence line. With $B_0^+$ for instance, the training and validation curves are almost close to the convergence line. Therefore, collecting more data for $B_0^+$ score improvement seems not helpful though, as adding more data may result in noise (Raschka and Mirjalili 2019). In addition, considering both $B_0^+$ and $Past^+$, the validation and the training scores converge to a value that is quite low ($\approx 87\%$) compared to $XGB^+$ and $Stc^+$ ($\approx 89\%$). However, the latter learning curves ($XGB^+$ and $Stc^+$) show a large gap between training and
validation curves, similar to $Stc^+$. Both curves expect that adding more training samples will most likely increase the generalization to have improved convergence scores.

Based on the aforementioned analyses, $XGB^+$ seems a good paradigm to better generalize to the unseen data than other $EMLs^+$ although all are expected to do the same. The next section could validate this assumption through the error analyses after the prediction on the test data.
**a)**

\[
\begin{align*}
    LR^+ & \quad C = 1.0, \quad p = 12 \\
    KNN^+ & \quad N = 9.0, \quad \sigma = \text{d}_{\text{man}} \\
    DT^+ & \quad e = e, \quad M_d = 7.0 \\
    \text{pSVM} & \quad \text{Kernel} = "RBF", \quad C = 2.000, \quad \zeta = 0.01, \quad \gamma = 0.125 \\
\end{align*}
\]

**Diagram:**

- Score vs. training size for each model.
- Training Curve (blue), Validation Curve (red), Convergence Line (dashed black).
- Baseline Score (circles).

**b)**

\[
\begin{align*}
    B0^+ & \quad BL = \{LR^+ + KNN^+, DT^+, pSVM\} \\
    \text{Past}^+ & \quad \text{Bootstrap} = \text{False}, BL = \text{pSVM}, \quad \eta_{\text{e}} = 150.0 \\
    XGB^+ & \quad M_d = 7.0, \quad \eta = 0.01, \quad \gamma = 0.5, \quad B = \text{gbtree} \\
    \text{Stc}^+ & \quad BL = \{LR^+ + KNN^+ + DT^+ + pSVM\}, \quad \text{Blender} = LR^+ \\
\end{align*}
\]

**Diagram:**

- Score vs. training size for each model.
- Training Curve (blue), Validation Curve (red), Convergence Line (dashed black).
- Baseline Score (circles).
Figure 2: Learning curves. a) $BLs^+$ learning curves. b) $EMLs^+$ learning curves. $XGB^+$ and $Stc^+$ offer a better tradeoff between the variance and bias than other models.
4.2. Models evaluation

The evaluation of $EMLs^+$ consists to validate whether the selected paradigms can properly be generalized to the unseen data after applying them to the test set.

4.2.1. FR scores on $X_{test} | y_{test}$

Table 3 gives the FR prediction scores using the four $EMLs^+$. 

Table 3: $X_{test}$ FR prediction Scores

<table>
<thead>
<tr>
<th>Metrics</th>
<th>$B_0^+$</th>
<th>$Past^+$</th>
<th>$XGB^+$</th>
<th>$Stc^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.850574</td>
<td>0.873563</td>
<td>0.896552</td>
<td>0.908045</td>
</tr>
<tr>
<td>Recall</td>
<td>0.772727</td>
<td>0.863636</td>
<td>0.886364</td>
<td>0.909091</td>
</tr>
<tr>
<td>Precision</td>
<td>0.851479</td>
<td>0.873678</td>
<td>0.906977</td>
<td>0.909091</td>
</tr>
</tbody>
</table>

Overall, the $EMLs^+$ predict a score >84% which corresponds to the minimum prediction threshold performed on the CV accuracy scores based on their convergence scores (see Fig. 2b). Therefore $EMLs^+$ can be considered able to properly generalize their performance to the unseen data. However, the optimal paradigm performance based on the evaluation of the metric is visible with $Stc^+$ ($\approx 91\%$) following by the $XGB^+$ ($\approx 90\%$), $Past^+$ ($\approx 87\%$) and $\approx 85\%$ for $B_0^+$. The lower recall score observed with $B_0^+$ ($\approx 77\%$), indicates that the model is unable to well demarcate the TP by confusing FR0 to FR1 (i.e., increase of the number of FN, see Eq.7). However, the precision and recall scores are almost close with $Past^+$ and perfectly equal with $Stc^+$ (same number of FP and FN). Moreover, $Past^+$ makes many mistakes (more FP) indicated by the precision score whereas $XGB^+$ improves its performance in the precision score ($\approx 91\%$) by predicting less
Furthermore, the confusion matrix can better assert each $EML^+$ error statement aforementioned.

- **4.2.2. CM and BPE metrics**

The CM metrics of Figure 3(a) indicate a bit more improvement using $EML^+$ paradigms in error estimation. Indeed, the best-fit error is found using the $B_0^+$ ($\approx 7\%$).

The latter is sensitive to FR0 i.e., $B_0^+$ can properly demarcate the FR0 composed of dry and unsustainable boreholes but is unable to well distinguish the productive FR1 class ($\approx 77\%$ success). Thus, the use of $B_0^+$ should not be considered the best paradigm for productive FR forecasting. However, compared to $Past^+$, $XGB^+$ is most preferred because it correctly classifies two more instances (one more instance for FR0 and FR1) thus reducing its prediction errors to 9% on FR0 and 11% on FR1. This is supported by the BPE (Fig. 3b). For instance, considering the FR prediction of $B_0^+$ in Figure 3(b), the top of the left bar is $FR = FR0$ while the model predicts $FR = FR1$ (i.e., FN). In addition, at the bottom of the right bar is the $FR = FR1$ whereas $B_0^+$ wrongly predicts an FR $=$FR0 (FP). Moreover, compared to $Past^+$ paradigm, the preference of $XGB^+$ can also be illustrated using both BPEs graphs in Figure 3(b). It is demonstrated by the higher number of FN and FP on $Past^+$ (topped on FR0 and FR1) compared to $XGB^+$. Furthermore, $Stc^+$ also performs better with fewer instances confused and a high prediction rate of 91%. This paradigm seems most interesting in CDWS to select the best anomaly (conductive zone) expected to give the RFR.
Finally, one can note that $B_0^+$ is more specific to FR0 while $Stc^+$ is the best EML for FR1 prediction and is capable of correctly demarcating FR0 and FR1 with a maximum error of 9%.
Figure 3: Confusion matrix (CM) and binary prediction error (BPE). a) $EML^+\text{CM}$ metrics representation. b) $EMLs^+\text{BPE}$ diagram.
4.2.3. PR tradeoff and cumulative gains plot (CGP)

Recall that $B_0^+$ is a benchmark EML and does not output the probability so it could not be evaluated with the following metrics.

Overall, in Figure 4(a), when lowering the threshold, the recall increases as the precision is reduced. This is illustrated when moving the average precision dash line (VanderPlas 2016; Bengfort and Bilbro 2019; Geron 2019). Moreover, the precision starts to fall around 87%, 84%, and 86% recall for $Past^+$, $XGB^+$, and $Stc^+$ respectively. Therefore, a precision/recall tradeoff can be selected just before 80%. However, decreasing the recall to increase the precision seems not a good idea. Indeed, the result will increase the rate of FN since predicting FR0 should be considered unsuccessful drilling and will lead to a new geophysical survey to achieve the project objective.

Figure 4(b) shows the CGP of $Past^+$, $XGB^+$ and $Stc^+$. It plots the TP rate (sensitivity) along the y-axis labeled as “gain” against the support rate (fraction of positive predictions) along the x-axis labeled as the percentage of the sample (Harrison 2019; Moroney 2020). The idea behind these plots is to sort all classifications by predicting the FR probabilities. For instance, if the first 10% of the predictions have 30% of positive samples of $EML^+$, we should plot a point from (0, 0) to (0.1, 0.3) and continue the process through all the samples. Considering the CGP of $Past^+$, if we want to demarcate 90% of FR0 or FR1 (sensitivity), we need to trace 0.9 on the y-axis to the right until we hit that curve. The x-axis at that point will show the total samples of FR0 or FR1 that are needed as support to get to a 90% rate of FR. Thus, from 0 to 50% on $Past^+$, $XGB^+$ and $Stc^+$, the total percentage of the sample has almost the same distribution of FR0 and FR1 (both lines are almost colinear).
In contrast, from 50% to 90% percentage of the samples, there is a slight difference in both classes FR0 and FR1. Indeed, the CGP plot for FR0 of Past$^+$ is above the FR1 which indicates the capability of the model to properly demarcate the FR0 class rather than FR1. For instance, 70% of the percentage of the samples yield a gain of 90% FR1 vs 95% of FR0 (Fig. 4b). As demonstrated by the confusion matrix metrics in Figure 3(a), the sensitivity of FR0 class of the Past$^+$ model is once again validated. Likewise, at the same 70% support, the XGB$^+$ shows a slight improvement in the FR1 demarcation by predicting the same gain of 100%. Moreover, the gains of FR0 and FR1 from Stc$^+$ are almost similar from 50% to 90% percentage of the samples. Furthermore, the CGP of FR0 and FR1 of XGB$^+$ and Stc$^+$ are far away from the baseline compared to the Past$^+$ which shows that both models could be considered as preference models among the four EMLs$^+$ designed in this workflow for FR forecasting.
Figure 4: Precision-recall (PR) curve and Cumulative Gain Plot (CGP). a) EML PR curves. b) EML CGP
To conclude, based on the previous analyses, although the EMLs have good abilities for generalization, XGB seems the best EML paradigms, followed by Stc. Past and \(B_0\) takes the third and fourth positions respectively for boosting FR.

The next section will demonstrate whether there is any influence of the geology of the area on the FR prediction using the EML paradigms.

- **4.2.4. Relationship between the geology and the FRs prediction**

Figure 5 gives us a visual depiction to understand the possible relationship between the geology of the area and the EMLs FR predictions. The different structures are oriented in SW to NW direction which corresponds to the direction of mineralization of the area (Kouamélan 1996). This direction of mineralization was coerced by many fractures that occurred during the tectonic activities in that region (Tagini 1971; Yace 2002). Based on the test samples, the FR prediction using the \(B_0\) and Past seems to have no relationship to the geology of the area. Indeed, all the misclassified labels are found everywhere in different geological structures although the wrong prediction in MIGG with Past is a bit higher compared to the one in \(B_0\). Opposite \(B_0\) and Past, XGB and Stc predictions are a bit sensitive to the geology of the area. Most of the wrong predictions of XGB are found in the northern part of the study area dominated by the GRAN. The behavior of XGB seems obvious according to the geological information of the survey area. Indeed, during the CDWS, the higher rate of unsuccessful drillings and unsustainable boreholes is found in the northern part, especially on the granites (~36% over 55% on GRAN).
(Kouadio et al. 2022). This part of the region is mostly dominated by the fracture which
does not contain any groundwater for long-term exploitation. This is validated by the 100%
rate of wrong predictions performed on the GRAN using $XGB^+$ and $Stc^+$. Opposite to
$XGB^+$, the misclassified labels using $Stc^+$ do not mainly focus on the northern part of the
country, it is almost found everywhere only in the GRAN. At this stage, it is difficult to
conclude that geological information has a great influence on FR prediction. Indeed, two
reasons can explain this fact: First, the imbalanced proportions of geological data in $D$ (Fig.
1). Secondly, the GRAN itself is known as an unproductive structure especially the wide
fracture found on that structure (Lasm 2000). Based on the Baseyian computation of geo-
electrical features in Kouadio et al. (2022), the authors demonstrate that wide-fractures in
that area have a 6.07% chance to obtain groundwater contrary to the narrow-fracture
(82.35%). This confirms the geological complexity of the survey area (Gnamba et al. 2014).
Nonetheless, during a future CDWS, it could be useful to pay more attention to the granites
before proposing the drilling operations to avoid unsuccessful drillings.
Figure 5: Relationship between FR prediction and geological structures of the experiment area. All predictions were performed on the test data in the experiment area and were projected based on the geographical coordinates.
5. Discussion

The popular implementation of EML in groundwater mapping has paid more attention in the last decade of years (Sahoo et al. 2017; Zounemat-kermani et al. 2019; Nguyen et al. 2020; Yariyan et al. 2020; Li et al. 2023) with the use of pasting, bagging, and boosting methods (e.g. XGB, DT) mostly considered for storage change predictions (Naghibi et al. 2020; Yin et al. 2021). However, the prediction of FRs using geo-electrical features seems novel in the literature. The first work was published using the SVMs as predictors with a global score on validation equal to 70% of correct prediction (Kouadio et al. 2022). At a glance, this rate is satisfactory in terms of minimizing unsuccessful drillings.

Moreover, due to climate change, numerous organizations such as (AMCOW 2008, UNECA 2009, and UNICEF 2017) have investigated a lot in CDWS for population welfare (Bayu et al. 2020). If we consider the ratio of unproductive boreholes (49.65% of FR0; see Fig. 1) in the experiment area, we can notice the huge loss the organization and governments face against water scarcity. The approximative 1/2 rate of unsuccessful drilling raises doubt about the use of traditional methods like the shape, type, and geology of the area to specify the right place before drilling operations. However, although the score from the SVMs seems a good alternative for minimizing the unsuccessful drilling rate, the improvement of that score always remains a relief for investors and governors as well as the geophysical and drilling companies. Indeed, the loss of one unsuccessful drilling is estimated at around 24 810 $US (including the pumping test) and 1 654 $US per survey (CIEH 1993; Sombo 2012; Mobio 2018). This is a considerable expense especially when the CDWS covers at least 2000 villages (Xepapadeas and Koundouri 2004; MHCI 2012;
Mel et al. 2017). Therefore, predicting FR with accuracy can highly help to supply more villages thereby increasing the number of boreholes for the population. This should be a great relief for populations that daily faced water scarcity.

The new approach developed in this workflow using the EMLs seems to respond to this objective. Indeed, the EMLs have improved the FR prediction score to 85% on the validation set compared to the pre-trained SVMs (~77% on the validation score globally). For instance, the 7% enhancement for around 2000 villages is estimated to be 1,736,700 $US profit and seems satisfactory. Figure 6 validates this enhancement using $EMLs^+$ compared to the pre-trained SVMs on the four kernels (Linear (Lin), polynomial (Poly), RBF, and sigmoid (Sigm)). Typically, the pre-trained SVMs show a good generalization with a maximum error of 16% and a minimum error of 12% (see Fig. 6a). However, the best error is found using the $Stc^+ (\approx 9\%)$ (Fig. 6b) although $Past^+$, $XGB^+$ also gives excellent results with a maximum error of 14% on FR1 and 9% on FR0 (which is the best minimal error). In contrast, $B_0^+$ can be considered the worst paradigm compared to the pre-trained SVMs when predicting the FR1 class.
Figure 6: Comparison of pre-trained SVMs and EML paradigm performances using the CM metric applied on the test set for FRs predictions. a) pre-trained SVMs; Lin, Poly, RBF, and Sigm are the linear, polynomial, Gaussian radial basis function and sigmoid kernels.
respectively. The model with a red frame corresponds to the pSVM used for EMLs aggregating. b) EMLs paradigms. The paradigms perform improved FR prediction scores (except for the B$_0^+$) with fewer errors compared to the pre-trained SVMs.
6. Conclusions

The approach proposed in this research study using the ensemble machine learning (EMLs) paradigms has predicted and improved the FR rate to a score greater than 80% on the validation test. The four base learners (BLs) were trained by fiddling with their hyper-parameters and cross-validated on four folds to get the optimal parameters. Thus, the optimal BLs were aggregated to compose the four categories of EMLs: the Benchmark EMLs ($B_0$), the Pasting (Past), the Extreme Gradient Boosting (XGB), and the Stacking (Stc) paradigms. Compared to the pre-trained SVM ($\sim77\%$) results, the EML scores have improved ($\geq7\%$) to reach 85\% for $B_0$ and Past, 86\% and 87\% for XGB and Stc respectively. This is satisfactory to reduce the repercussion of unsuccessful drillings thereby minimizing the huge loss of investments. Although the maximum performances of 90\% and 91\% on the test set were achieved using the XGB and Stc respectively, it is recommended to ascertain the relevance of the technique by testing it in another survey area with additional training samples.

7. Acknowledgments

The authors gratefully acknowledge Ibrahim Berthe and Kone Saramatou of the Water Resources Department, and Diarrassouba Mada of Cote d’Ivoire National Office for Drinking Water (ONEP) for their collaboration and support during the PPU and PNAEP programs. They also appreciated the creation of the monitoring framework of exchange by Sombo Boko Celestin and Gahe Emile of the School of Earth Science and Mineral
Resources (STRM) during the second part of the project. Finally, the authors express their gratitude to Ndeutala Selma Iita to make the manuscript more readable.

**Ethical Approval**

The authors LKK, JL, SKK and RL approve the following publications policies:

- Manuscripts submitted to this journal should not have been published, or simultaneously submitted elsewhere.
- Submitted articles should largely contain previously unreported material. The overlap of contents between related papers should be the minimum, normally confined to the introductory/review sections.
- Authors should provide information and preprints on any papers closely related to the submitted paper, which are not yet in the public domain (in review, in press). If deemed necessary to the review process, these will be copied to the reviewers.
- Authors should consult the Journal’s “Authors Instructions” for detailed stylistic guidelines, during preparation of their manuscripts. Editors may return manuscripts that are grossly inconsistent with these guidelines.
- Where the submitted manuscript is multi-authored, one individual will need to be the corresponding author. It will be assumed that all the authors have been involved in the work, have approved the manuscript, and agree to its submission.
- Manuscripts will be sent to at least two reviewers. Reviewers are requested to treat the manuscript confidentially. They may choose to identify themselves, or remain anonymous.
- The Editor’s judgement is final with regard to suitability for publication.
• The Publisher carries responsibility for typesetting, page layout and figure sizing, for all accepted papers.

• Authors are responsible for reading and correcting page proofs of their articles. Proof corrections are normally restricted to typesetting and printing errors. Major changes cannot be undertaken at the proof stage.

Consent to Participate

All authors whose names appear on the submission:

1) made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work;

2) drafted the work or revised it critically for important intellectual content;

3) approved the version to be published; and

4) agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Consent to Publish
The authors LKK, JL, SKK and RL agreed with the content and that all gave explicit consent to submit and that they obtained consent from the responsible authorities at the institute/organization where the work has been carried out.

**Authors Contributions**

- Conceptualization, Methodology, Writing - original draft preparation: Laurent Kouao Kouadio;
- Formal analysis and investigation: Serge Kouamelan Kouamelan;
- review and editing, Resources: Rong Liu;
- Supervision: Jianxin Liu.

**Funding**

No funding was received to assist with the preparation of this manuscript.

**Competing Interests**

The authors (LKK, JL, SKK, RL) have no competing interests to declare that are relevant to the content of this article.

**Availability of data and materials**

The data for this research study is available in zenodo (https://zenodo.org/record/5560937). The computer codes for reproducing the workflow
are hosted on GitHub(https://github.com/WEgeophysics/watex). The documentation (watex.readthedocs.io) gives a step-by-step guide for a non-specialized user.

References


Bayu T, Kim H, Oki T (2020) Water governance contribution to water and Sanitation


https://doi.org/10.1080/10106049.2020.1852612


https://doi.org/10.1109/TNN.2006.875973


https://doi.org/10.1002/2013WR014792


Lasm T (2000) Hydrogéologie des réservoirs fracturés de socle : analyses statistique et géostatistique de la fracturation et des propriétés hydrauliques ; application à la région des montagnes de Côte d’Ivoire (domaine archéen). Université de Poitiers

https://doi.org/10.1007/s11269-022-03410-y

https://doi.org/10.1029/2008WR007488

https://doi.org/10.1007/s00703-019-00675-8


https://doi.org/10.1007/s12065-020-00447-z


Mel EACT, Adou DL, Ouattara S (2017) Le programme présidentiel d’urgence (PPU) et


Powers DMW (2007) Evaluation: From Precision, Recall and F-Factor to ROC. Adelaide


Sombo PA (2012) Application des methodes de resistivites electriques dans la determination et la caracterisation des aquiferes de socle en Cote d’Ivoire. Cas des departements de Sikensi et de Tiassale (Sud de la Cote d’Ivoire). Universite Felix Houphouet Boigny


https://doi.org/10.1029/2003WR002158


Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- highlights.docx