Pump it or Leave it?
A Water Resources Evaluation in Sub-Saharan Africa

Jacqueline M. Fortin Flefil, Marios A. Galanis, Vladimir B. W. Kozlow

jackieff@stanford.edu, margalan@stanford.edu, vkozlow@stanford.edu

Stanford ENGINEERING

Motivation

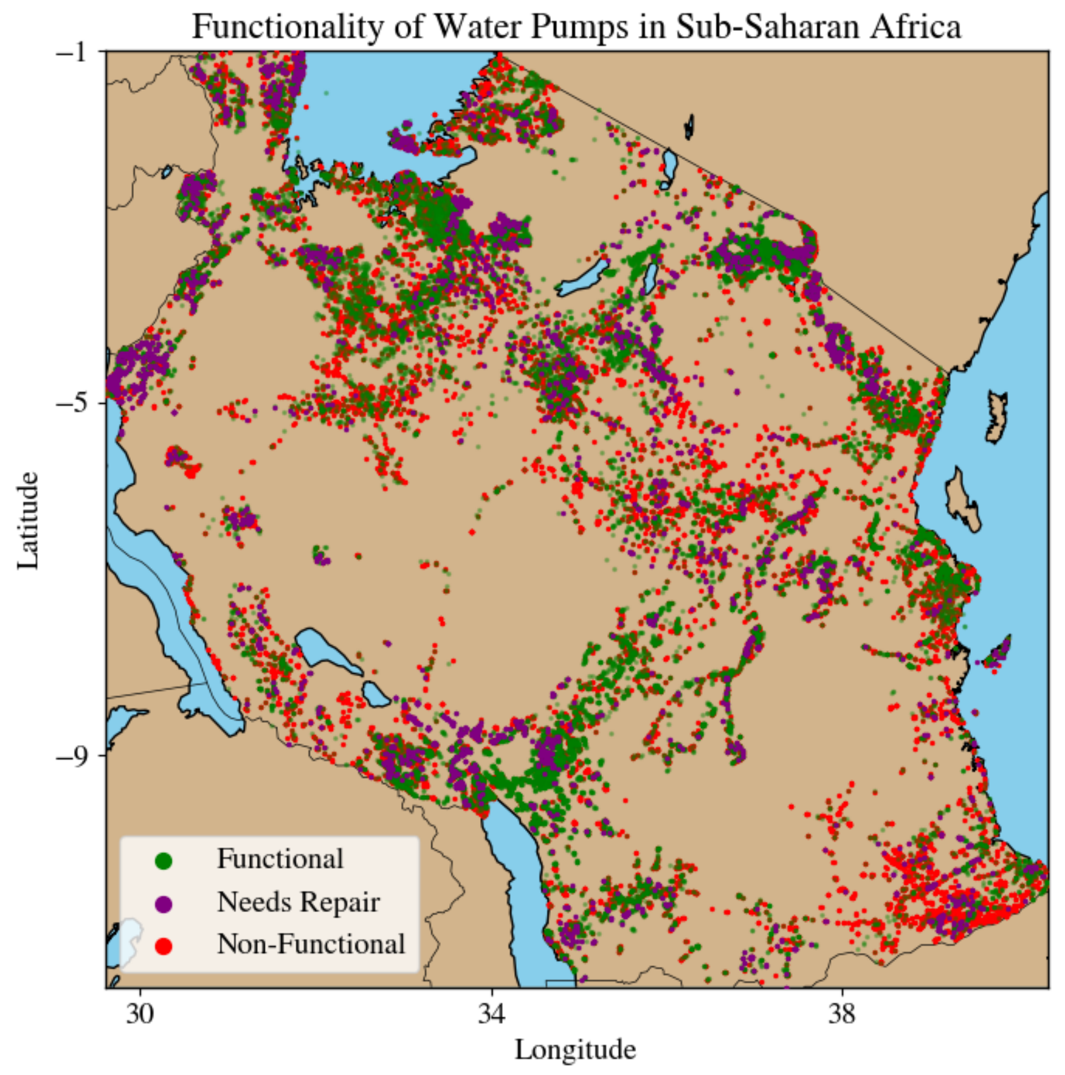
In Sub-Saharan Africa, failure rates of handpumps, the main source of water for millions of people, is almost 15% just one year after installation[1]. The goal of this study is to develop an algorithm that can predict the functionality of a handpump as well as the quantity and the quality of water it outputs based on a minimum of data collected on the field. Predicting those characteristic of a handpump at a given point in time can help shorten the time required for managing agencies to provide support and plan targeted maintenance operations of handpumps in remote areas.

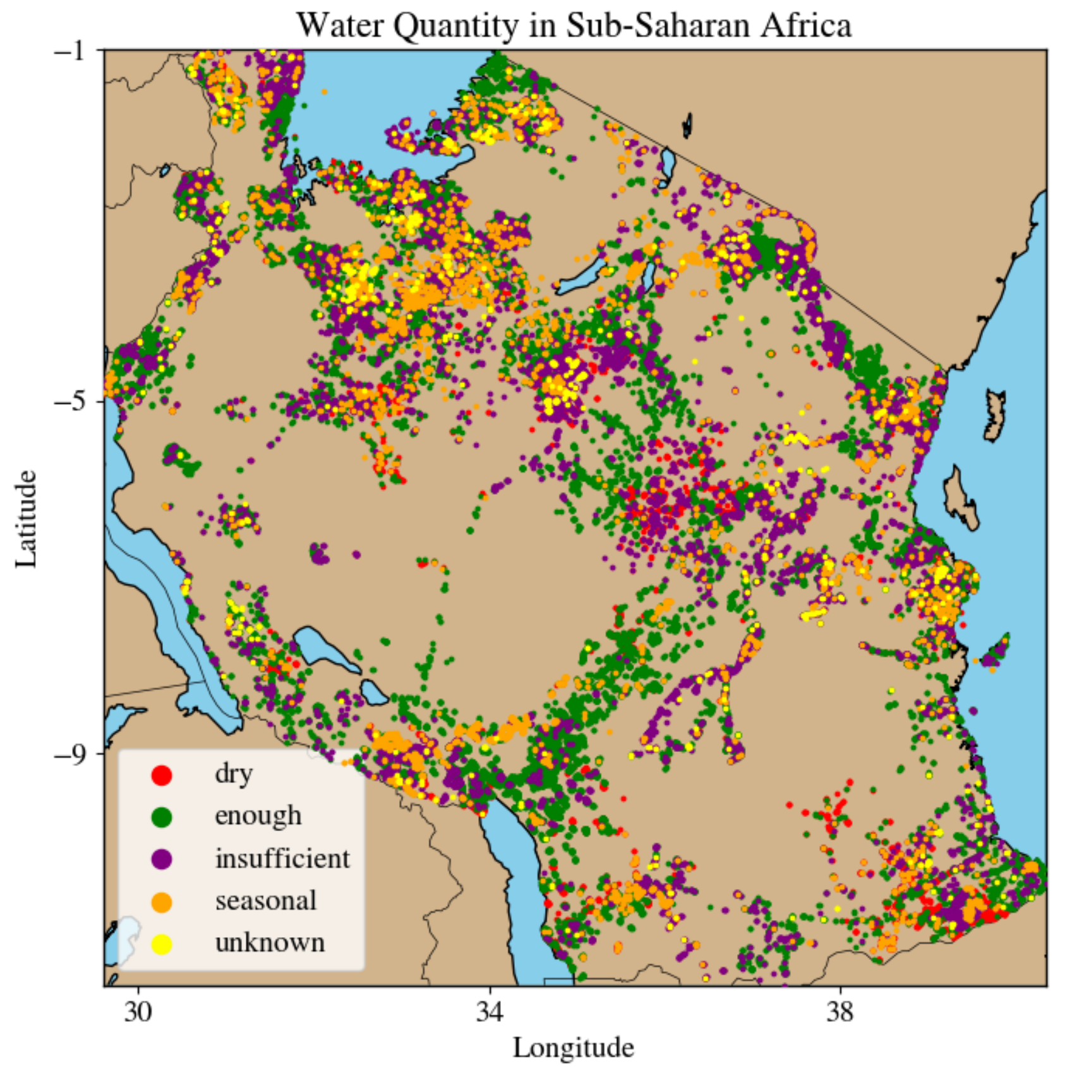
Preprocessing

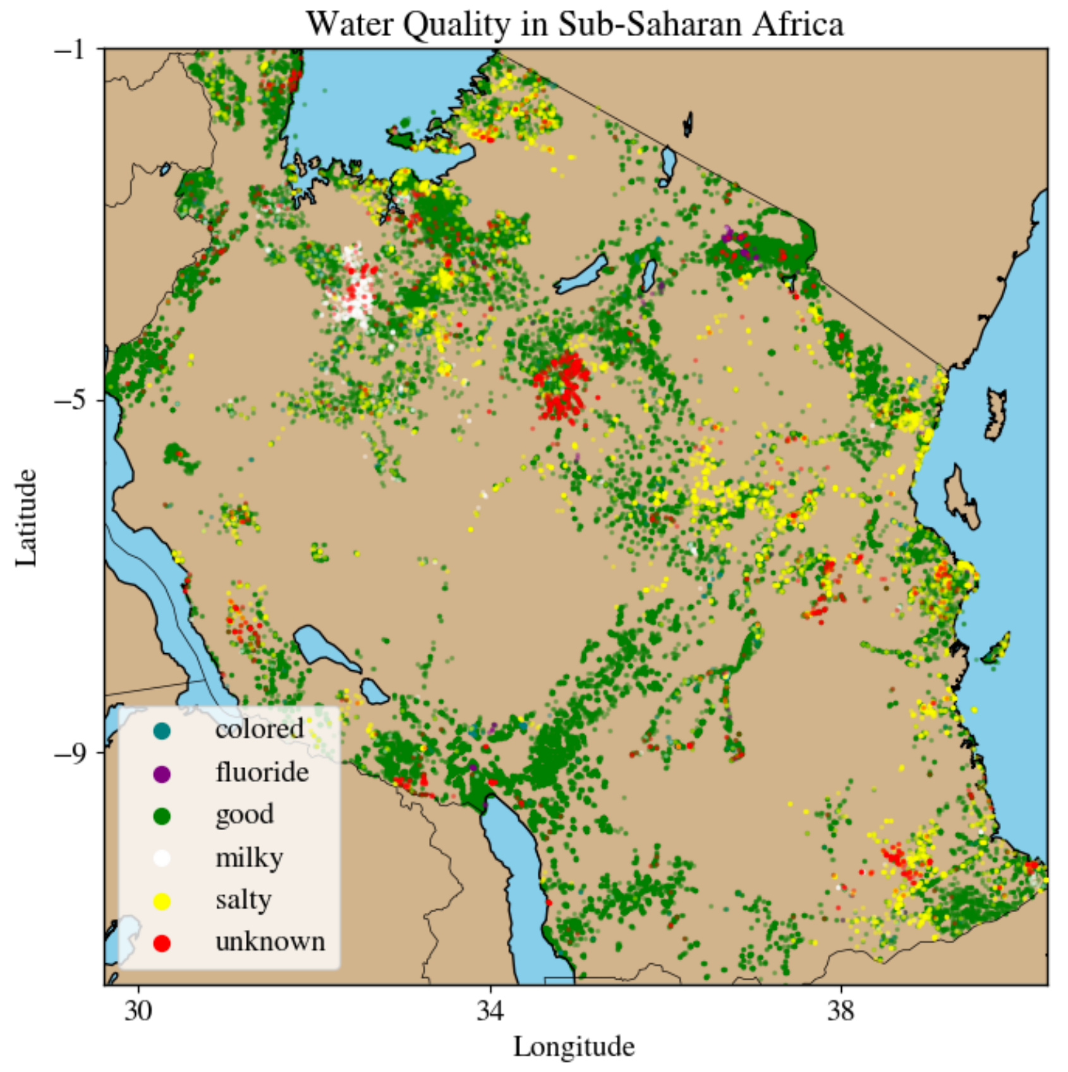
Only 24 of the 40 original features were used. The categorical features were transformed into binary feature using One Hot Encoding (OHE). Missing or incoherent feature values were replaced by the mean (numerical) or mode (categorical/binary) of this feature over the dataset. To deal with class imbalance, the Synthetic Minority Over Sampling Technique (SMOTE) was applied as described in [2].

Dataset

This study is based on the **Taarifa dataset** which contains information about 59,400 handpumps located in rural Tanzania. Each handpump has 40 features attached to it, most of which are categorical features, the rest being numerical. Three categorical features of the dataset were identified as possible indicators of the sustainability of a handpump: functionality of the handpump, quantity of water delivered, and quality of water delivered. Those three features were predicted separately in the study.







Methods

Models were optimized using a grid search with CV to fine tune hyperparameters. Final results were obtained using 5-fold CV with a 75%-25% train-test split. The voting ensemble method was used to optimize our final results. Algorithms were evaluated and optimized based on the F1 score.

Multinomial logistic Regression:

Multinomial logistic regression was performed and optimized with L2 regularization and coordinate gradient descent.

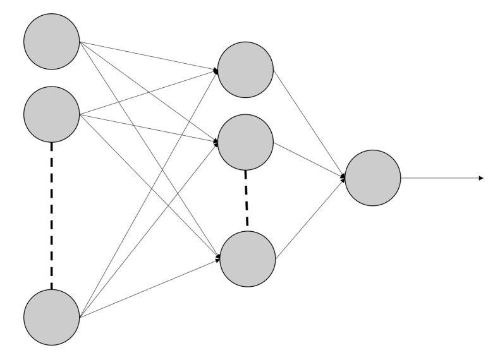
MLSE cost function with L2 regularization

$$J(\theta) = \|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2$$

Random Forest:

Hyperparameters that were optimized are: number of trees, max depth, max number of features per split, minimum number of samples by leaf, minimum number of sample by split.

Neural Network:



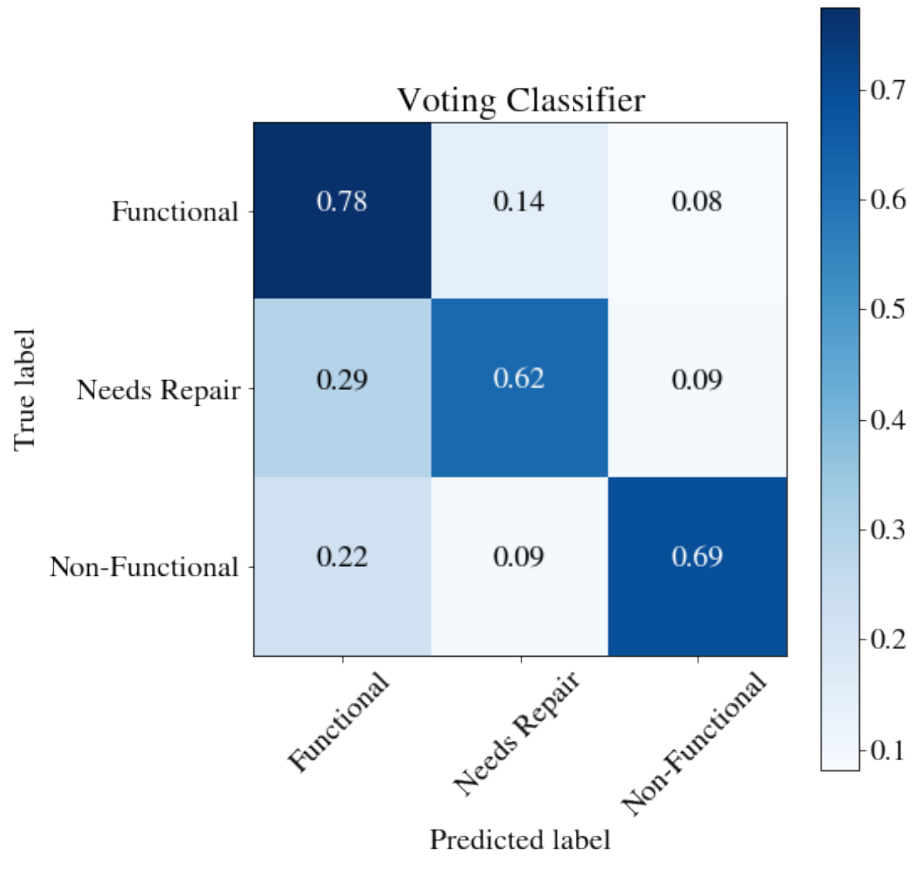
Sigmoid function:

$$g(z) = \frac{1}{1+e^{-z}}$$

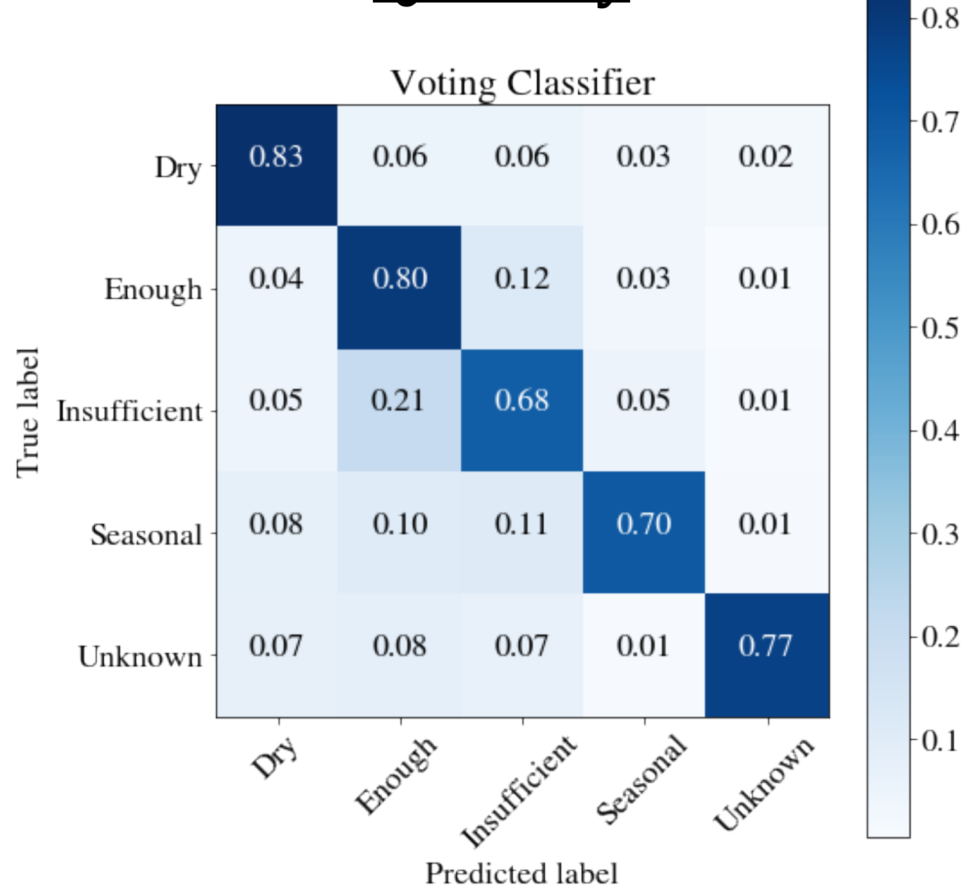
The Sigmoid function was used as the activation function. Different architectures were tested and optimized for each prediction.

Results

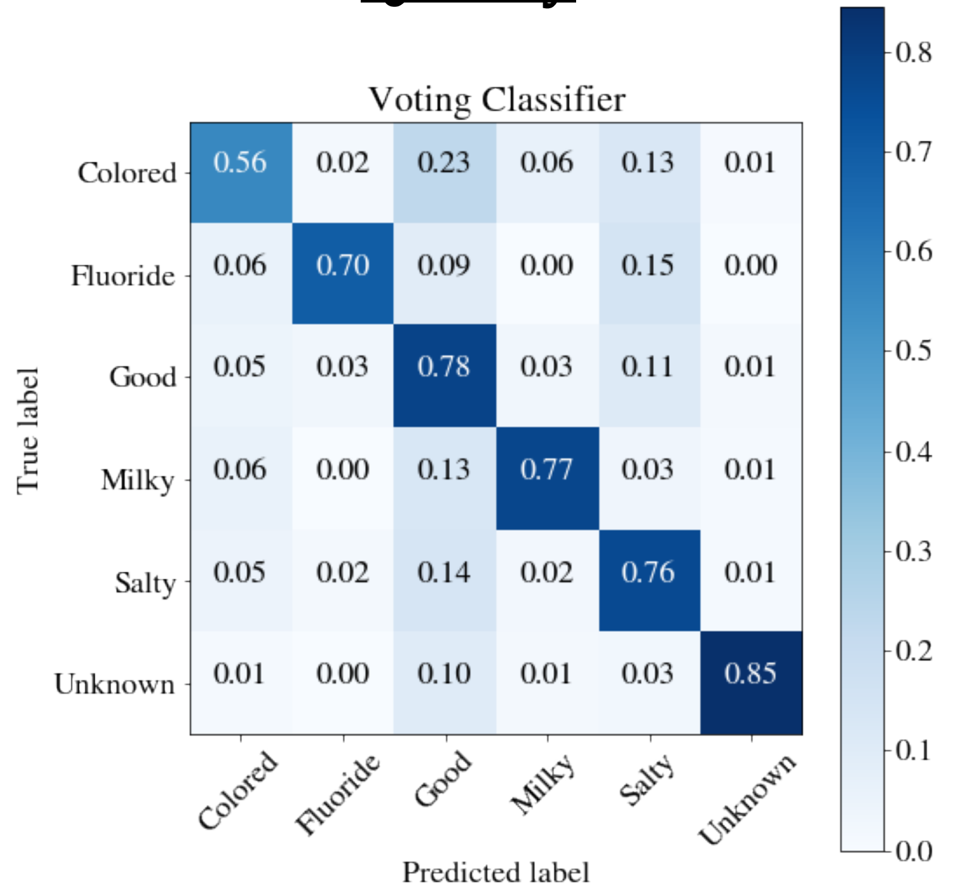
Functionality



Quantity



Quality



	F1-Score (Micro)					
Method	**Functionality**		**Quantity**		**Quality**	
LR	65.2%	64.3%	68.6%	57.0%	77.0%	59.8%
RF	86.2%	76.8%	91.9%	78.9%	97.9%	87.4%
NN	74.0%	70.4%	79.6%	66.4%	91.0%	73.3%
Voting	79.5%	73.5%	93.7%	77.0%	93.3%	78.3%

Discussion

The micro-average F1 score provided a good evaluation of the tested algorithms overall in terms of number of good predictions. However, it did not provide information on the distribution of our good predictions over the different classes. The class imbalance – still present in our test set - made it hard to get high overall F1 scores, so we relied visually on confusion matrices. We were able to achieve high accuracy at predicting the most represented categories, but usually still did not have great accuracy for the less common categories, despite the SMOTE process. The Random Forest algorithm achieved the best F1 scores overall, but was outperformed by the voting ensemble in terms of accuracy distribution.

Future Work

Future work could include looking at different models such as Convolutional Neural Networks or Support Vector Machine. Looking at differences in predictions between countries or regions would help us see how robust the algorithm is and how dependent on local conditions the sustainability of handpumps is. Finally, adapting the model to predict when a pump would fail would make it more applicable on the field.

References

[1] Rural Water Supply Network (RWSN) (2009) 'Handpump data 2009: selected countries in Sub-Saharan Africa' [online] <http://www.rural-water-supply.net/en/resources/details/203> [accessed 11/20/2018].

[2] Fisher, M. B., Shields, K. F., Chan, T. U., Christenson, E., Cronk, R. D., Leker, H., Samani, D., Apoya, P., Lutz, A., ... Bartram, J. (2015). "Understanding handpump sustainability: Determinants of rural water source functionality in the Greater Afram Plains region of Ghana". Water resources research, 51(10), 8431-8449.