Predicting the functionality of water points in Tanzania using machine learning

1st Sibusiso Ziqubu
Johannesburg, South Africa
ziqubu.sibusiso@gmail.com

***Abstract*—Access to clean and reliable water sources is essential for community well-being and development. However, in many developing countries like Tanzania, the functionality of water points, such as wells and boreholes, is a significant concern. Tanzania, in particular, faces a crisis with nearly half of its 57 million people lacking improved access to safe water. This study employs machine learning models, specifically Random Forest and XGBoost, to predict water point functionality in Tanzania. It aims to investigate the impact of the choice of machine learning algorithm on predictive accuracy. The results indicate that Random Forest, especially when combined with SMOTE to address the class imbalance, outperforms XGBoost in predicting water point functionality with 70%. Confusion matrices and feature importance analysis highlight the effectiveness of Random Forest in correctly identifying positive instances and provide insights into influential features in water point functionality prediction. This study contributes to a deeper understanding of which models are most effective in predicting the status of water points In Tanzania.**

***Index Terms*—Machine learning, water points, XGBoost, random forest**

I. Introduction

Access to clean and reliable water sources is critical for the well-being and development of communities [7]. In many developing countries, including Tanzania, the functionality of water points, such as wells and boreholes, is a significant concern. Ensuring the sustainable operation of these water points is crucial for providing safe and consistent water access to communities. However, the challenge lies in identifying functional, nonfunctional, and repairable water points promptly to enable effective maintenance and resource allocation. In Tanzania, nearly half of its 57 million people lack improved access to safe and clean water, and the repercussions of this crisis are far-reaching [12]. The absence of reliable water sources also undermines agricultural productivity, hindering food security and economic growth. Moreover, the prevalence of waterborne diseases, such as cholera and diarrhea, is exacerbated due to contaminated water sources, leading to high mortality rates and increased healthcare expenditures [10]. Predicting the functionality of water points is pivotal for efficient resource allocation and maintenance planning [1]. By comparing the performance of different machine learning algorithms, this study contributes to a deeper understanding of which models are most effective in predicting the status of water points.

This study uses machine learning models to predict water point functionality in Tanzania. The study seeks to test whether the choice of machine learning algorithm, specifically Random Forest and XGBoost, impacts the accuracy of predicting water point functionality. It hypothesizes that Random Forest, especially when combined with techniques like SMOTE to address class imbalance, will exhibit superior predictive accuracy compared to XGBoost. The research aims to provide empirical evidence supporting or rejecting this hypothesis.

II. Background and related work

Previous studies have explored various approaches for water point functionality prediction and maintenance in different contexts. These studies highlight the significance of utilizing advanced technologies, particularly machine learning, to enhance decision-making processes related to water infrastructure management. Anomaly detection techniques, including machine learning algorithms, have been applied to identify abnormal patterns in water point data [9]. By detecting deviations from normal operational behavior, these methods can assist in predicting potential failures and the need for maintenance.

The study conducted by Arymurthy et al. [1] focused on predicting the status of water pumps using a data mining approach. The approach was applied to the data from the Tanzania Ministry of Water to predict the present and future statuses of water pumps in the region. The chosen data mining method was eXtreme Gradient Boosting (XGBoost) and Recursive Feature Elimination (RFE) was also employed to identify crucial data features, enhancing model accuracy. The study achieved its best accuracy using 27 input factors selected by RFE, coupled with XGBoost as the learning model. The resulting accuracy rate was 80.38%. The results from this study could help the government to improve inspection planning, maintenance strategies, and identification of factors that could harm water pumps. Ultimately, this contributes to the availability of clean water in Tanzania. Compared to manual inspection methods, the data mining approach proves to be cost-effective, quicker, and less time-consuming.

The study by Pathak and Shalini [8] addresses the imperative need for effective water pump maintenance by introducing a novel approach. The study employs a sequential attentive deep neural architecture known as TabNet to predict the water pump repair status in Tanzania. This model seamlessly integrates the strengths of both tree-based algorithms and neural networks, offering advantages such as end-to-end training, model interpretability, sparse feature selection, and efficient learning from tabular data. The study conclusively shows

that TabNet surpasses established gradient boosted tree-based models like XGBoost, LightGBM, and CatBoost in predictive accuracy. Furthermore, by training TabNet with focal loss instead of categorical cross-entropy loss, the performance is further elevated when dealing with imbalanced datasets. Overall, this study offers a robust solution to the critical challenge of water pump maintenance, contributing to improved water supply management and reduced water crisis impacts.

A study conducted by Chowdavarapu and Manikandan [5] used various data mining models to predict the functionality of water pumps. The assessment of validation misclassification rates highlighted Random Forest as the most effective model for classifying water pumps accurately. The findings of the study revealed significant factors influencing pump functionality included longitude, region code, district code, source of water, GPS height, water quantity, extraction type, payment methods, and water-point type. The model demonstrated a sensitivity of 67.75% and a specificity of 92.72%, indicating its capability to identify functional and non-functional pumps. Overall, the study's results offer valuable information for water infrastructure management, emphasizing key factors affecting water pump functionality and providing a comprehensive understanding of maintenance patterns and lifespan based on geographical and demographic factors.

These research endeavors collectively emphasize the potential of machine learning in predicting the functionality of water points and enhancing water infrastructure management. By harnessing the power of data-driven insights, these studies contribute to the development of accurate and reliable predictive models that aid in identifying functional, nonfunctional, and repairable water points. The successful application of machine learning techniques in this context has the potential to significantly improve water supply reliability and contribute to the sustainable development of communities.

III. Research Methodology

This section provides an in-depth explanation of the research methodology, including the use of data, pre-processing, data cleaning, and the proposed method.

A. Data

The data used in this research is obtained from the Taarifa water points dashboard, which aggregates data from the Tanzania Ministry of Water [6]. The data consists of 59400 water points data with 40 features related to demographic information, administrative information, spatial information, and structural information. The water points are distributed into three classes, functional, functional but need repair, and non-functional. Fig. 1 shows the distribution of functionality of water points in Tanzania. Out of a total of 59 400 water points, 32 259 water points are functional, 22 824 are non-functional, and 4 317 need repairs, indicating a class imbalance. All three classes are important for predicting water point functionality in Tanzania.

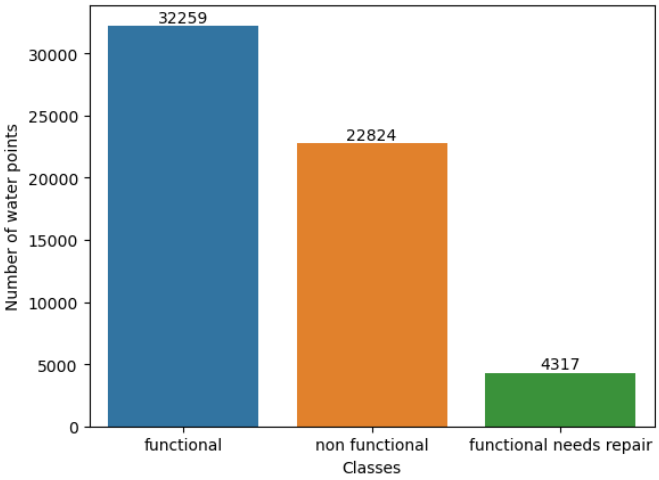

Fig. 1. Distribution of functionality of water points

B. Data pre-processing

The data contains some features with the same information, such as *quantity* and *quantity_group*. These features may lead to over-fitting, increase time cost in the learning process, and influence the accuracy rate. Therefore, in this study, we dropped features with the same information. There were many features with missing values, some important features like *scheme_name* having as much as 28 166 missing values out of 59 400 data points, the details can be seen in Table I. The missing values were imputed using the mean/median approach. The feature engineering process improved the *funder* and *installer* columns by reducing dimensionality and addressing spelling errors. A new categorical column was created to represent these values. Additionally, *construction_years*, which was an integer format, was converted into categorical values by grouping them into decades, as individual year values did not provide meaningful continuity for the model.

TABLE I
Features with missing values

Features	Number of missing values
scheme_name	28 166
scheme_management	3 877
installer	3655
funder	3635
public_meeting	3334
permit	3056
subvillage	371

C. Balanced data

Class imbalance refers to a situation where the distribution of classes in a dataset is uneven, with one class being significantly underrepresented compared to another class [3]. As seen in Fig. 1, the data used in this study is extremely imbalanced. In the context of classification tasks, when dealing with class imbalance, it is essential to address this issue to ensure that the

model learns to accurately predict the minority class as well. The technique used in this study to address class imbalance is SMOTE (Synthetic Minority Over-Sampling Technique) is a popular oversampling method. SMOTE works on the idea of nearest neighbors and create its synthetic data [3]. It is one of the most popular techniques for oversampling.

D. Proposed method

1) XGBoost: XGBoost, introduced by Chen and Guestrin [4], is a popular machine learning algorithm that belongs to the gradient boosting family. It is built on the gradient boosting framework, which combines the predictions of multiple weak learners (usually decision trees) to create a strong ensemble model. Known for efficiency, flexibility, and high predictive performance.

2) Random Forest (RF): Random Forest is a popular ensemble learning algorithm used for classification and regression tasks [2]. It creates multiple decision trees during training and aggregates their predictions to make a final prediction. This ensemble approach helps to improve the overall performance and robustness of the model. It employs a technique called bagging which reduces model variance and improves generalization by training each decision tree on a bootstrap sample.

E. Hyperparameter tuning

In order to create the most effective classification model for the given dataset, it is essential to identify the ideal parameters. In this study, we have employed grid search as a method to find the optimal parameters. Grid search is particularly useful for automating the process of hyperparameter tuning and saving time and effort by testing various combinations in a structured manner Table II depicts how parameter tuning through grid search has led to notable increases in balanced accuracy, particularly for the XGBoost model. Prior to parameter tuning, the balanced accuracy for the XGBoost model stood at 59%, while after parameter tuning, it improved to 68%.

F. Performance evaluation metric

In this study, balanced accuracy and confusion matrix are used to evaluate the performance of each model. Balanced accuracy is a metric that considers both sensitivity and specificity, making it a more suitable choice for evaluating classifiers on imbalanced datasets [11]. A confusion matrix is a tabular representation used in the field of machine learning and statistics to evaluate the performance of a classification algorithm on a dataset with known ground truth labels.

G. Tools and libraries

This study utilizes Python, a programming language that has various libraries for data analysis and machine learning tasks. For data visualization, Matplotlib and Seaborn are employed, while Pandas and NumPy play a crucial role in data manipulation. Scikit-Learn provides various tools, including classifiers, preprocessing functions, and evaluation metrics. Addressing class imbalance in datasets is achieved through the imbalanced-Learn library, with SMOTE as one of the techniques applied. Hyperparameter tuning is facilitated by GridSearchCV, an integral component of Scikit-Learn. Additionally, MLxtend tools are used for machine learning model evaluation, providing functions for creating confusion matrices and assessing feature importance.

H. Experimental procedure

A stratified train/test split with an 80/20 ratio is employed to ensure representative training and testing datasets in a machine learning task. The training set is utilized for both model training and hyperparameter tuning using a grid search. Gridsearch is used to systematically explore various combinations of hyperparameters to find the optimal settings for the model. The procedure of the study is highlighted in Fig. 2 using the flowchart

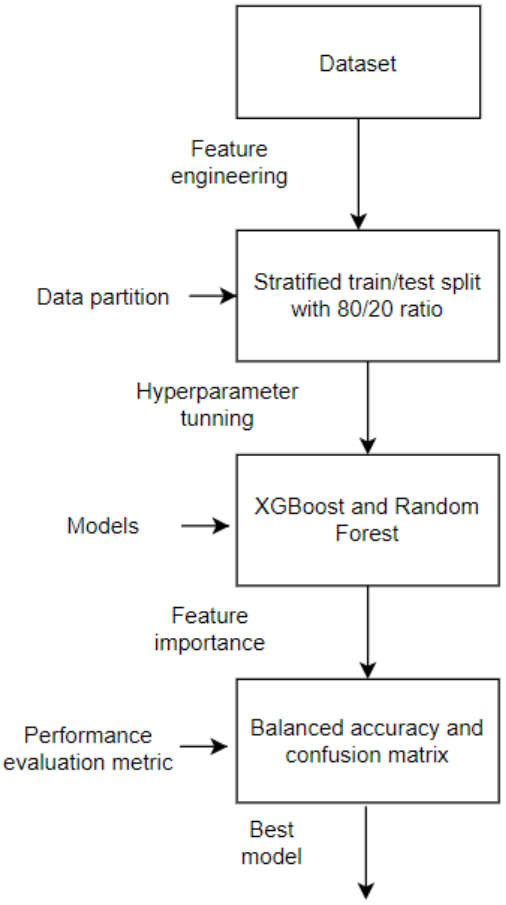

Fig. 2. Procedure of the study

IV. Results

A. Data exploration

Firstly, we conducted data exploration using the entire dataset to extract valuable insights, which are depicted in Fig. 3. This figure shows that many water points, despite having enough water, are non-functional. Dry water points have a

high correlation with non-functionality. Having enough water increases the likelihood of finding functional water points.

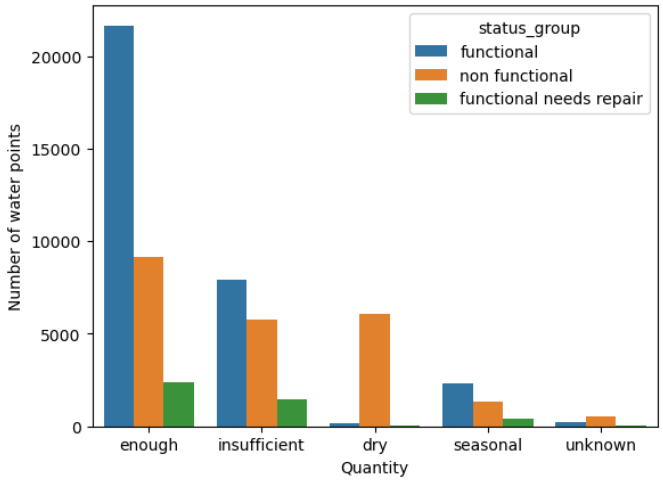

Fig. 3. Number of water points by water quantity

B. Balanced accuracy

In this study, we employed balanced accuracy on the test set to select the most effective model for predicting water point functionality in Tanzania. The results in Table II indicate that the Random Forest model, when coupled with SMOTE to address class imbalance, achieved the test balanced accuracy at 70%, surpassing XGBoost at 69%. Consequently, based on these findings, Random Forest is deemed the superior model. Notably, XGBoost exhibited improvements in performance after addressing the class imbalance issue using SMOTE, as evidenced by Table II.

TABLE II
PERFORMANCE EVALUATION METRIC: BALANCED ACCURACY

Model	Balanced accuracy test set	Tuning methods
RF	0.678	
XGBoost	0.587	
RF	0.707	Grid search
XGBoost	0.677	Grid search
RF	0.701	SMOTE
XGBoost	0.694	SMOTE

C. Confusion matrix

Confusion matrices, depicted in Fig. 4 and 5, serve as tools to assess the performance of each classification algorithm. They provide a summary of model predictions compared to actual outcomes in our dataset. Notably, the confusion matrix for the Random Forest model in Fig. 4 displays the highest True Positive (TP) value when compared to XGBoost. This observation highlights the Random Forest's effectiveness in correctly identifying positive instances, indicating its strong performance.

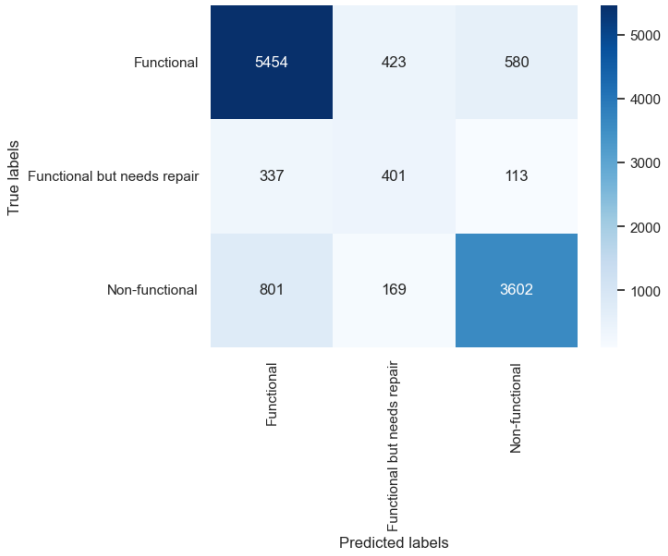

Fig. 4. Confusion matrix for Random Forest with SMOTE

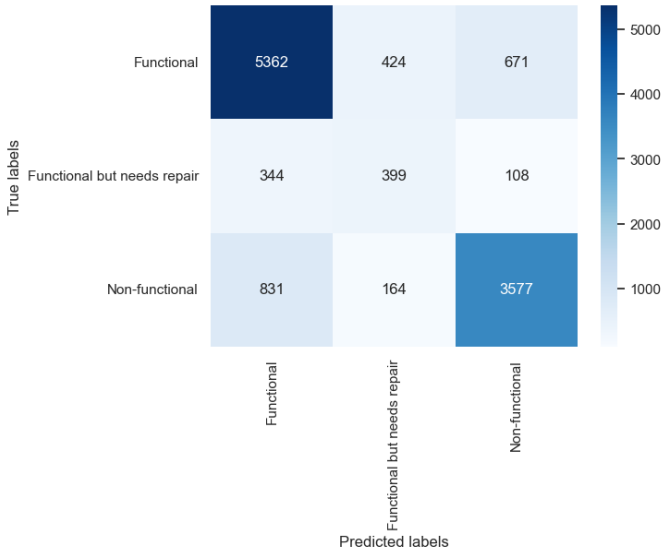

Fig. 5. Confusion matrix for XGBoost with SMOTE

D. Feature importance

Utilizing Random Forest for model building helps identify influential features within the dataset. In the context of water point status prediction, the importance of specific features can be determined. According to Fig. 6, features *public_meeting*, *permit* and *water_quality* have less importance in our model, whereas *longitude*, *quality* and *latitude* have higher importance.

V. DISCUSSION

The importance of clean and reliable water sources for community well-being and development cannot be overstated. In many developing countries, including Tanzania, the challenge of maintaining functional water points, such as wells and

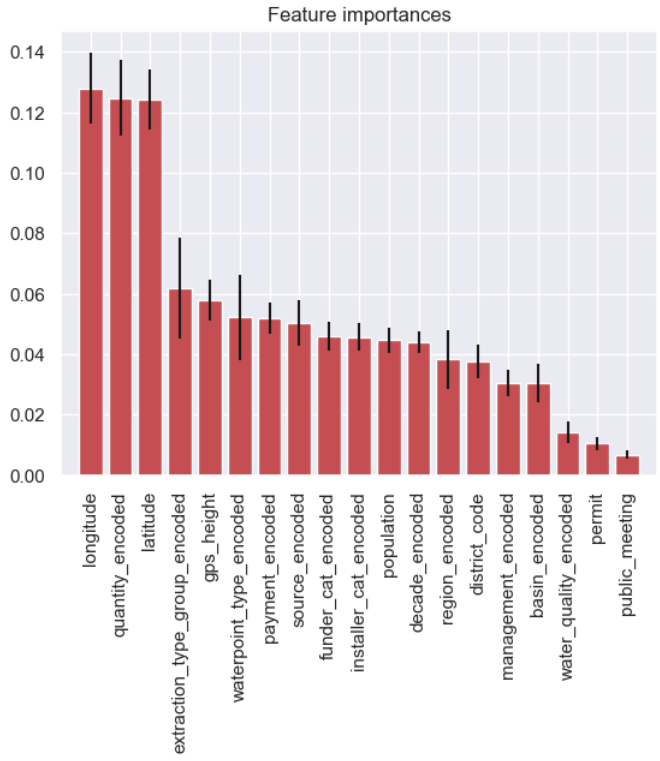

Fig. 6. Feature importance, showing 19 features that influence classification by RF

boreholes, presents a significant obstacle to ensuring consistent access to clean water. The repercussions of this challenge extend far beyond water supply issues, impacting agriculture, food security, and healthcare. Leveraging machine learning techniques to predict the functionality of water points offers a promising solution to address these challenges.

In the context of this study, we have reviewed previous research that explored various approaches to water point functionality prediction and maintenance. Notably, these studies highlight the significance of machine learning algorithms, in enhancing decision-making processes related to water infrastructure management. Techniques such as anomaly detection, data mining, and deep learning have been employed to identify abnormal patterns in water point data, predict potential failures, and optimize maintenance efforts. One of the studies we reviewed, conducted by Arymurthy et al. [1], employed eXtreme Gradient Boosting (XGBoost) as the machine learning model to predict the status of water pumps in Tanzania. They achieved a high accuracy rate of 80.38%, highlighting the effectiveness of machine learning in this context. The results have significant implications for government planning, maintenance strategies, and the availability of clean water in Tanzania. Another study by Pathak and Shalini [8] introduced TabNet, a sequential attentive deep neural architecture, to predict water pump repair status in Tanzania. TabNet's performance surpassed traditional gradient boosted tree-based models like XGBoost, LightGBM, and CatBoost, demonstrating the potential of deep learning approaches for water infrastructure management. Chowdavarapu and Manikandan [5] study focused on data mining models and highlighted Random Forest as an effective classifier for water pump functionality. This research revealed key factors influencing pump functionality and provided valuable insights into maintenance patterns and lifespan based on geographical and demographic factors.

The results of our study are consistent with these findings, emphasizing the potential of machine learning in predicting water point functionality. We conducted data exploration to gain valuable insights, which showed that many water points with an adequate water supply are non-functional, highlighting the need for timely maintenance. Employing balanced accuracy, we compared the Random Forest and XGBoost models. Notably, the Random Forest model, when combined with SMOTE to address the class imbalance, achieved the test balanced accuracy of 70%, surpassing XGBoost. The confusion matrices we used to assess the performance of classification algorithms confirmed the effectiveness of the Random Forest model in correctly identifying positive instances, further supporting its superiority. Additionally, feature importance analysis using Random Forest revealed influential features in water point status prediction, which can guide decision-making in water infrastructure management.

In conclusion, this study aligns with previous research in demonstrating the potential of machine learning in predicting water point functionality. The application of advanced technologies in water infrastructure management contributes significantly to the availability of clean water, the sustainable development of communities, and the mitigation of water crisis impacts. By leveraging data-driven insights, we can continue to enhance water supply reliability and address the critical challenges associated with water infrastructure management. This work underlines the importance of further research and practical applications of machine learning in water resource management in Tanzania and similar regions.

VI. Conclusion

In regions like Tanzania, where access to clean and reliable water sources is a critical concern, the application of machine learning techniques for water point functionality prediction holds significant promise. This study focused on comparing Random Forest and XGBoost models, with a specific emphasis on addressing class imbalance using SMOTE. The results demonstrate the effectiveness of Random Forest in achieving superior predictive accuracy, making it the preferred model for water point functionality prediction. The use of confusion matrices further substantiates the efficacy of the Random Forest model in correctly identifying positive instances, which is crucial in ensuring the reliable functionality of water points. Additionally, feature importance analysis sheds light on the critical factors that influence water point status, facilitating more informed decision-making in water infrastructure management. The findings of this study contribute to the sustainable development of communities in Tanzania by enhancing

water accessibility and quality. By leveraging machine learning to predict water point functionality, resources can be optimally allocated for maintenance, minimizing downtime, and ensuring consistent access to clean drinking water.

Opportunities for expanding upon the findings of this study exist, as the current research only offers a comparison of a restricted set of machine learning algorithms. Subsequent investigations could involve the evaluation of a broader spectrum of machine learning models trained on datasets sourced from the Tanzania Ministry of Water.

REFERENCES

[1] Aniati Murni Arymurthy et al. "Predicting the status of water pumps using data mining approach". In: *2016 International Workshop on Big Data and Information Security (IWBIS)*. IEEE. 2016, pp. 57–64.

[2] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *Test* 25 (2016), pp. 197–227.

[3] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[4] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[5] Indra Kiran Chowdavarapu and VD Manikandan. "Data Mining the Water Pumps: Determining the functionality of Water Pumps in Tanzania using SAS Enterprise Miner". In: SAS South Central User Group Forum. 2016.

[6] DrivenData. *Pump it Up: Data Mining the Water Table*. 2014. URL: https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/ (visited on 08/14/2023).

[7] Tula M Ngasala et al. "Analysis of water security and source preferences in rural Tanzania". In: *Journal of Water, Sanitation and Hygiene for Development* 8.3 (2018), pp. 439–448.

[8] Karan Pathak and L Shalini. "Pump It Up: Predict Water Pump Status using Attentive Tabular Learning". In: *arXiv preprint arXiv:2304.03969* (2023).

[9] Kai Qian et al. "Deep learning based anomaly detection in water distribution systems". In: *2020 IEEE International Conference on Networking, Sensing and Control (ICNSC)*. IEEE. 2020, pp. 1–6.

[10] R Rainey and M Weinger. "The role of water, sanitation and hygiene (WASH) in healthcare settings to reduce transmission of antimicrobial resistance". In: *AMR Control [Online Edition]* (2016), pp. 65–8.

[11] Alaa Tharwat. "Classification assessment methods". In: *Applied computing and informatics* 17.1 (2020), pp. 168–192.

[12] Mattana Wongsirikajorn et al. "High salinity in drinking water creating pathways towards chronic poverty: A case study of coastal communities in Tanzania". In: *Ambio* (2023), pp. 1–15.