Process Safety and Environmental Protection 148 (2021) 1090–1097

Contents lists available at ScienceDirect

Process Safety and Environmental Protection

journal homepage: www.elsevier.com/locate/psep

Artificial intelligence based monitoring system for onsite septic systems failure

Niranjan Ravi [a,*], Daniel P. Johnson [b]

[a] Electrical and Computer Engineering, Purdue School of Engineering and Technology, Indianapolis, IN, USA
[b] Department of Geography, Indiana University Purdue University Indianapolis, Indianapolis, IN, USA

ARTICLE INFO

Article history:
Received 16 November 2020
Received in revised form 26 January 2021
Accepted 28 January 2021
Available online 17 February 2021

ABSTRACT

An onsite sewage system (OSS) is a complex system that takes advantage of nature's biological processes to remove harmful pathogens from wastewater and reintroduces clean water back into the natural water cycle. However, the failure of an OSS can contaminate drinking water, surface water and release hazardous pathogen and chemicals creating hazardous conditions within the local environment. This research aims to provide an artificial intelligence (AI) solution using machine learning (ML) algorithms such as logistic regression (LR), random forest classifier, and K-nearest neighbors (KNN) to understand and examine the underlying factors that could cause septic failures. Septic records from 1970 to 2019 were collected from five different counties across the State of Indiana and ML algorithms were applied to predict areas of possible septic failures. The algorithms demonstrated accuracy of approximately 80% in prediction of OSS failures. Such algorithms can assist state and county health departments in alerting homeowners of impending failures. Such an approach has the ability to not only prevent failures but also in the maintenance of a safe environment for communities reliant of OSSs. This approach was implemented and tested in the Indiana State Department of Health (ISDH).

1. Introduction

Since the early 1950s, septic systems have been in use in neighborhoods and rural and suburban developments. According to the U.S. Census Bureau (1999)(Onsite Wastewater Treatment Systems Manual), around 23 percent of the estimated 115 million homes are using onsite septic systems. Approximately, one-third of the US population which are not connected by centralized public sewers utilize septic systems to treat the wastewater onsite (University of California Cooperative Extension, n.d.). The number of septic systems is increasing and about one-half million septic systems are being installed each year (Septic System Status and Issues Working Paper, 2021). More than 800,000 onsite sewage disposal systems are currently located in the State of Indiana. Septic systems are being constructed by homeowners or local companies following the design guidelines laid out by state or federal health departments. These systems have a wide range of components and configurations and septic tanks utilizing a soil absorption system predominate (Septic Systems, 2021b). According to the design manual of onsite wastewater treatment and disposal system (Design Manual of Onsite Wastewater Treatment and Disposal System, 2021), inorder to prevent dense soils suffering from absorption issues deeper and wider trenches can be installed. Following the developed guidelines,the bottom of the trenches are covered with coarse aggregate before the drain tile is installed. The introduced aggregate provides a manufactured porous environment which enables the effluent from the OSS to flow in to the surrounding soil. From this description it is clear to see that the trench system also provides a storage medium for the effluent as infiltration rates into the surroundings are time dependent (i.e. soil which has been saturated by recent flooding).

The complexity of the soil plays a vital role in the effective functionality of OSSs. Soil characteristics widely vary between the geographic locations. Sandy soils, soil with low clay content, and loamy soils are ideal for effective functioning of septic systems. Massive soils, where soil particles are not arranged into real structural units and wet soils are not suitable for septic system. It has been estimated that only 32% of the total land area in the United States have soils suitable for traditional septic systems (U.S. EPA, 1980). Despite this observation, septic systems have been installed in many locations due to necessity and remote location, where it is difficult to connect to public sewer systems

When OSSs fail they can contaminate groundwater, and by default private wells. They can also add to sudden nutrient enrich-

* Corresponding author.
E-mail addresses: ravin@iu.edu (N. Ravi), dpjohnso@iupui.edu (D.P. Johnson).

https://doi.org/10.1016/j.psep.2021.01.049
0957-5820/

ment in lakes and catalyze the eutrophication process (Septic Systems, 2021e). Inappropriate design, poor maintenance, and poor soil parameters are the predominant cause of failure (Septic Systems, 2021f). Outdated design practices can lead to the inappropriate design of new systems. Failing OSSs are an onerous challenge for state, county and local health departments which catalog OSSs in their jurisdiction. The challenge is becoming increasingly difficult as suburban and rural development increases.

Motivating this research into the use of modern artificially intelligent (AI) techniques, is the need for an interactive monitoring system that will predict likely OSS failures. This approach has the potential to reduce the burden on health departments and can provide specificity on locations to perform field inspections. We provide several modeling options that can be relatively easy to follow by technical personnel at interested health departments.

2. Background

Traditionally, information on failing septic systems are collected by the field-based investigation of households. OSS failure locations are often characterized by lush vegetation, standing wastewater and dark soil where organic matter has accumulated. The associated high cost and variable reliability of these conventional techniques has encouraged health departments to seek alternatives. Although we do not use the technique in our research, one method receiving much attention recently is the use of unmanned aerial vehicles (UAVs) to acquire aerial photography/ground survey to locate OSS failures (Evans, 1982). This technique involves piloting an UAV to acquire aerial photographs and performing image analysis to identify failures. Such an intervention is followed by a site visit to validate findings. The initial research was carried out by Crouch (1979)(Evans, 1982) and personnel at two of the U.S. Environmental Protection Agency's field stations: the Environment Photographic Interpretation Center in Warrenton, Virginia and the Environmental Monitoring and Support Laboratory in Las Vegas, Nevada. According to studies by (Evans, 1982) the aerial surveys had interpretation accuracies ranging from 60% to 95%. There are some limitations associated with this technique. The primary limitation is that UAV sensors (Ravi et al., 2019) can identify only surface level failures. Also, the use of a UAV would likely be seasonal for many locations due to potential tree canopy. Furthermore, if a UAV is flown after a rainy day or during a monsoon season, photo analysis may not yield accurate classification. However, UAV imagery could provide another layer of data that would increase the accuracy of the models presented in this research.

3. Septic systems

OSSs are the primary method of draining wastewater from houses or buildings not connected to public sewer systems. A properly designed system can work efficiently for decades with regular maintenance. The septic tank, distribution system, manhole and drain field/absorption field are the major components of OSSs. The distribution box is responsible for even distribution of wastewater. Septic tanks are constructed with concrete and are buried underground and typically located to the back or side of a building. They are responsible for digesting organic matter and solid materials from wastewater. These systems discharge the fluids from the septic tank into a series of pipes, chambers, or any special units designed to release the effluents into the soil (Septic Systems, 2021a). The tank collects the outgoing drainage from toilets, sinks, tubs, and laundry from the property. Inside the tank, grime and waste are separated from the water by gravity. The water is distributed to an outgoing grid of drain pipes where it is then released into the soil. Bacteria also play a predominant role in the smooth

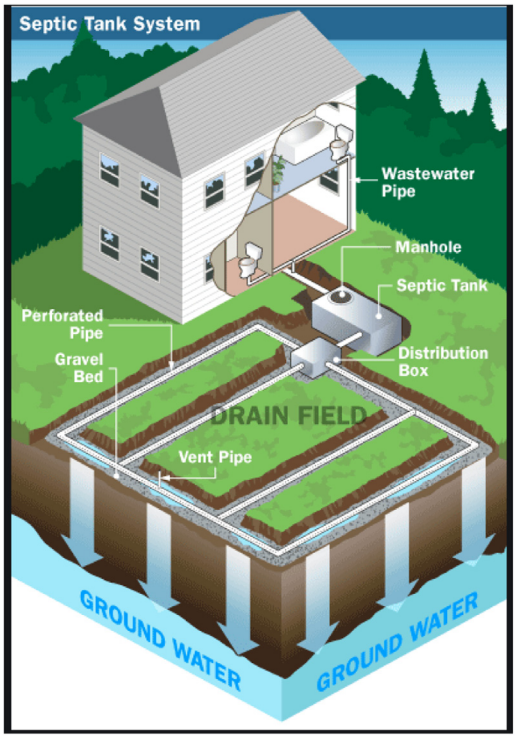

Fig. 1. Working of a septic system (Septic System Working, 2021).

and efficient running of an OSS. These exist both in the septic tank and drainage field and serve to break down organic materials in the discharge. This aids in the creation of a layer of sludge at the bottom of the tank while a layer of effluent floats at the top. Often a filter is attached in the outlet of the septic tank to prevent clogging aw effluents are discharged into the drain field. The drainage field provides a large area where bacteria can survive and holes in the drain field allow the treated water to reach underground (Fig. 1).

A properly installed septic system requires occasional pumping to remove the sludge and scum from the tank and to keep the system in working order. Unlike public sewer systems, septic systems cannot accommodate waste once a particular threshold is crossed. Numerous factors are used to determine the proper threshold level and when a system should be replaced. Excessive water usage and the disposal of non-biodegradable substances constitutes improper use of an OSS and can lead to clogging and total system failure.

Geographic factors such as the installation of a septic system in an area with inappropriate soils, excessive slope, or high water table can also contribute to improper function or failure. Design factors such as inappropriate choice of septic tank size can also lead to failure. For example, a commercial building would require a larger septic tank than a single household building. Any individual factor listed above or a combination of them could lead to a catastrophic failure impacting the household, neighboring systems and the environment.

4. Factors involved in designing an efficient system

While designing a new septic system for a house/commercial building, many factors are considered for its efficient functionality. There are various types of OSS such as a simple septic tank, conventional system, chamber system, mound systems, and drip distribution system available. Choosing an appropriate septic system varies based on the soil characteristics and dwelling type; whether it is a commercial building, residential unit or multiple housing units in a residential building. For example, a chamber

system is typically installed in areas of high water tables or areas with significant ground water saturation (Septic Systems, 2021c). A mound system distributes, effluents from the septic tanks into a pump chamber where it is discharged in prescribed doses. An adequate size of pump chamber is selected to prevent overflow. When a pump chamber fails, the result is a discharge of untreated effluents (Septic Systems, 2021d). In the case of multiple housing units, more than one septic tank is installed, usually providing a separate septic tank for each unit. The size of septic tank is dependent upon the number of bathrooms/bedrooms and expected water usage in the housing unit (Septic Tank Serving Multiple Housing Units, 2021). A dosing tank is usually installed in addition to septic tank which ensures the effluents are uniformly transmitted to the tile bed (About Dosing Tank, 2021). Dose size and dose volume are physical parameters associated with a dosing tank. If the drain absorption field square feet is overloaded with liquid, it can flood causing sewage to flow to the ground surface or create backups in soils and within the tanks themselves. As mentioned in section one, type of soil plays a significant role in the proper choice of a specific septic systems and its efficient functionality. The slope of the soil determines how much slope is required for the installation of septic lines to aid flow. Detailed understanding of these are essential in construction of an efficient OSS.

5. Research survey in OSS failures

A thorough research study was carried out to understand the reasons causing the failures in OSSs. The following research surveys has been carried out in various parts of the globe. According to a research report (Establishing Failure Indicators for Conventional On-Site Wastewater Treatment Systems, 2021), incorrect size of septic tank and drain field results in design failure of OSSs. Occupancy size of the household, frequency size of usage can contribute to system failure in primary stages. In another research (Beal et al., 2005), field surveys have been conducted and identified shallow water table, permeable soil types and hydrological and biogeochemical mechanisms are causing septic failures. The authors in this research (Forbis-Stokes, 2012), have discussed the reasons associated with drain pipe and drain field on how heavy rainfall and decreased evapotranspiration can result in clogging and leading to septic failures. Age of septic systems and irregular maintenance are also denoted as plausible reasons (Beal et al., 2005; Forbis-Stokes, 2012). In addition to usage of UAVs to predict septic failures as detailed in Section 2, another approach has been carried out to identify factors causing internal surface failures (Identification of Failing Septic Systems, 2021). Factors such as water bed area, biomat, number of bedrooms/bathrooms, drain field and pump chamber details associated with each septic record have been taken consideration. A statistical approach was carried out to analyze these records on a dataset size of 312 records to predict degree of septic saturation. But further analysis/understanding of importance of other factors were not discussed. This triggered a need to develop a more efficient and effective monitoring system to predict failures. Currently, in order to prevent septic failures, constant field inspections have been carried out health departments.

6. Machine learning

The term Machine learning was coined in the year 1959 by Arthur Samuel, a pioneer in the field of AI. Data is increasingly being captured on a variety of phenomenon all the time. These data contain valuable information but the volume of collection makes human interaction practically impossible. As a result machine learning (ML) has developed into an integral component of AI. ML can utilize a variety of practices such as using frequentist statistics to conduct analysis based on trends in data and historical relationships between the data records; thereby "learning" intricacies in the dataset. ML also evolved as an alternative to conventional engineering approaches to develop algorithmic solutions to various problems. ML has been applied in various fields such as information retrieval, game playing, bio-informatics, stock market predictions, seismic predictions, and weather forecasting. Currently, there are various ML applications used daily which many are unaware of, such as Netflix™, YouTube™, weather prediction or something as simple as detecting a spam email in your account. This technology has vast potential owing to recent advances in the fields of computing and engineering.

Once there is a necessary level of understanding to the phenomenon in question, the second phase involves developing a mathematical model to numerically capture the behavior of a failing septic system. Finally, an algorithm is developed and coded that closely mimics the behavior of the system in question; assuming it is an accurate representation of the problem (Simeone, 2018).

There are three major categories of machine learning.

- Supervised learning.
- Unsupervised learning.
- Reinforcement learning.

In supervised learning, the developed model directly informs the algorithm specifically of what to look for. Usually, this is entered into the algorithm by the user and is thereby "supervised". Unsupervised learning uses methods to look for random or systemic patterns within the data. Reinforcement learning is similar to the trial and error and is one of the more recent trends in ML.

ML has two major types:

- Classification problem.
- Regression problem.

The first step of ML problem would be to collect data from various sources; the so called dataset. Using these data, a model will be developed consisting of crucial features which have a direct impact on the unknown status of the OSS. These features are inturn referred to as independent variables and our target variable is denoted as the dependent variable. It's intuitive to understand, the dependent variable has a dependency upon the independent variable. An ML algorithm is developed based on this model. This algorithm would predict households with possible septic failures. For better understanding terms like a dependent variable, independent variables and features will be used in the following details. ML approach is chosen is justified by the case-to-case basis of its suitability and potential advantages. In this research, ML algorithms are programmed using python programming language, and algorithms deployed for this research are:

- Logistic regression (state of art technique).
- Random forest classifier.
- K-nearest neighbors.

7. Models

7.1. Logistic regression (LR)

Logistic regression is useful in many applications which have a binary response; in our case failed or not-failed OSS. This probabilistic linear classifier consists of a weight matrix w and a bias b. It helps a system to categorize dependent variable with help of

independent variables. LR is predominantly applied in classification problems. The classifier equation is

$$y = \sigma(w^T x + b),$$

where as y belongs to $\{-1, 1\}$ indicates the output class for the input x which was fed into the system. The weight matrix w is initialized to 1 and it is constantly updated by the model as it passes through the training samples. The equation of weight updates is

$$\theta_j = \theta_j(n-1) + \alpha_j v_j,$$

$$v_j = \sum_{i=1}^{m} (y^i - h_\theta(x^i))x_j^i,$$

where as α is the learning rate, θ is the augmented weight matrix, m denotes the training samples, h_θ is logistic function given by

$$h_\theta(x) = 1/(1 + e^{-}\theta^T x).$$

Logistic function serves as an input for cost function in which the average cost is minimized using gradient descendent method. Minimizing the average cost is necessary to calculate optimum weights which would help the logistic regression model in accurate prediction of dependent variable.

7.2. Random forest classifier

The random forest (RF) algorithm is primarily chosen in many classification problems due to its ability to handle non-linear classification tasks and out-performs other state-of-art algorithms like linear and logistic regression. RF handles data imbalances in large datasets, which are a drawback in most real-world problems since many data points will not be recorded. The basic unit of RF is a binary tree constructed using recursive partitioning (RPART). RF creates multiple CART like trees in which the binary response recursively partition's the tree into many homogeneous nodes until it reaches an end/leaf node. Each decision tree is trained on a bootstrapped sample of training data. The output of RF is selected by the majority of votes from the trees (Nguyen et al., 2013). For example, consider a training set T where vector X of n features is an independent variable and the corresponding dependent variable is Y. As mentioned above, each of the decision trees would generate a result and a majority chosen by the algorithm. These decision trees are referred to as weak predictors h_k (Lahouar and Ben Hadj Slama, 2015) because each individual decision tree would predict one result based on its conditions and a cumulative value from a collection of decision trees are referred as strong predictors. The number of decision trees are denoted as *n*tree. Then,

$$Y = \frac{1}{ntree} \sum_{i=1}^{1ntree} h_i(X)$$

During the training phase, RF uses randomness to select a subset of input data to create the individual decision trees and is often represented as a collection of hundreds of thousands of trees (a forest). The selection of the number of trees is often a user-defined parameter. As a result of this representation, RF can handle high dimensional data by utilizing a large combination of trees. The application of randomness to split the data assists in reducing the correlation between individual trees and lowering any variance inflation factors. Furthermore, individual RF trees can be computed in parallel, aiding in the computational time of very large datasets (Fig. 2).

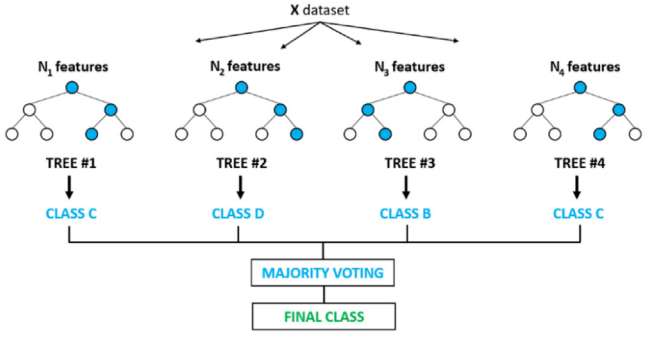

Fig. 2. Random forest with four decision trees (Random Forest, 2021).

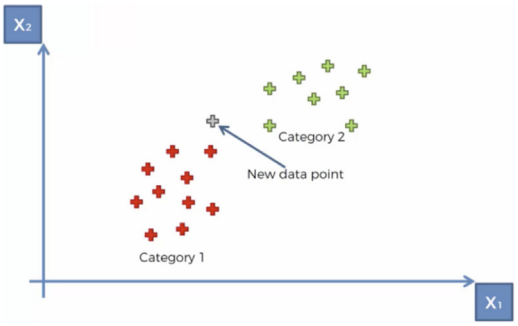

Fig. 3. K-nearest neighbors for binary classification problems (KNN, 2021).

7.3. K-nearest neighbors

K nearest neighbors (KNN) is one of the most applied supervised learning classification algorithms (Fix and Hodges, 1951). Owing to its simplicity, effectiveness, and intuitiveness KNN is used across many study domains. When a set of training samples and a query is entered into a KNN algorithm it proceeds by finding the closest point/neighbor to unknown dependent variable and assigns it a label. KNN is therefore an extension to the nearest neighbors (NN) rule (Gou et al., 2012). The dependent variable is labeled by a majority vote of its K-nearest neighbors present in the training set. Let

$$T = \sum_{i=1}^{1N} (x_i, y_i)$$

represent our training set, x_i is the vector of m dimensional feature space of independent variables and y_i represent the dependent variable. A query x' is our objective to find an unknown independent variable. This is calculated in two steps. In the first step, k similar targeted neighbors are selected for x'. In this new set T', the samples are arranged in the increasing order of the Euclidean distance between x' and x_i^{NN} (Fig. 3).

Euclidean distance is defined as:

$$d(x^i, x_i^{NN}) = \sqrt{(x^i, x_i^{NN})^T (x^i, x_i^{NN})}$$

The class label y' of our queried dependent variable is predicted by majority voting of its nearest neighbors by using the below formulae:

$$= argmax_y \sum_{x_i^{NN} y_i^{NN}} \delta(y = y_i^{NN})$$

where as, x_i^{NN} $y_i^{NN} \in T\prime$. In the above equation, y is the class label, y_i^{NN} is the class label for *i*th nearest neighbor among its K nearest

neighbors. The Dirac Delta function takes a value of one if y = y_i^{NN} and zero otherwise.

7.4. Determining model performance

ML algorithms are constructed by splitting the dataset into 70% for training and 30% for testing. The models are trained with the training data subset and its performance is evaluated with the testing subset. In binary classification problems, probability of a class variable (dependent variable) is mapped into either category 1 or 0, but before mapping, it should be compared to a threshold value/point usually 0.5. The values greater than 0.5 can be mapped into category 1 and values less than 0.5 are category 0. This threshold point is commonly referred to as a decision threshold. Before we analyze the performance of our models, the concepts of various performance metrics such as accuracy, precision, F1-score and recall, region under the curve, and area under the curve (ROC-AUC) has to be understood (Huang and Ling, 2005). The choice of performance metrics is typically performed on a case by case basis. Accuracy is the ratio of the number of correct predictions obtained to the total number of predictions made. This metric works well only when there is an equal distribution of data. Precision is the ratio of number of correct positive results to the total number of positive results that is predicted by the classifier. Recall is the ratio of correct positive results to the number of all relevant samples. F1-score is the harmonic mean between precision and recall. It indicates how precise the classifier performs. In-order to calculate F1, precision, and recall, ROC we employ a confusion matrix for better understanding.

		Predicted	
		Negative	Positive
Actual	Negative	*True Negative*	*False Positive*
	Positive	*False Negative*	*True Positive*

7.5. Region operator characteristics (ROC)

Concepts of ROC-AUC is used in the situation of an imbalanced dataset, such as in our case number of failed systems is higher than working systems. A ROC curve is used to analyze the performance of binary classification algorithms. As seen in Figs. 5 and 6, the threshold point is dotted orange diagonal line which is 50%. But the threshold point is optimal only when number of false negatives equals false positives. Choosing a right threshold point is a trade off, so its determined by balancing false positive rate with false negative rate. ROC curve provides a solution by plotting true positive rates against false positive rates. This metric would provide the performance of our algorithms by averting the loss acquired from false negatives.

7.6. Key limitations of the present study

One of the key limitations of this study is the way in which data regarding failed septic systems was collected. The Indiana State Health Department relies on county-level agencies to report septic failures. Most counties have a unique way of collecting this information and all data between counties is not consistent. To minimize this effect we selected data that was available through Indiana's Network for Tracking of Onsite Sewage Systems (iTOSS) system which assists county agencies in reporting consistent data regarding failures.

Additionally, the data was collected for four (4) separate counties within the State. These data represented the most consistent reports of failed systems within the time-frame of the study and Fig. 5 illustrates the geographic dispersion of the counties studied.

8. Data mining and prepossessing technique

To successfully build an ML model to predict failure in OSS, it is mandatory to gather data about existing OSSs. As shown in Fig. 4, the first step is data mining which involves gathering enough data for modeling. Data mining was performed in two phases across the U.S. State of Indiana. As detailed in subsection F, the reason for two phases of mining data is because data scarcity and data bias were observed in phase one. Data bias is an error which indicates certain elements in the dataset have been represented more which would limit the ML model's performance. Most of the data points had blank features creating an imbalance in distribution. This imbalance causes high variability in ML model, yielding a poor prediction during testing. In order to overcome these challenges, second phase of data mining was carried out with help of iTOSS. Equal distribution of data was observed eliminating the problems faced in phase one. This newly retrieved dataset had many additional independent variables compared to dataset from previous phase.

In the initial phase, data from Marion County's (Indianapolis) septic records were collected. These records have been manually collected by various state officials over several decades. The earliest records date back to 1970 with the most recent to 2019. 11,527 records of septic systems and their associated information were present in the dataset. Each of the 11,527 records denotes an individual household/commercial building in Marion County. The details associated with each record include pump chamber size, system tank size, absorption field square feet, number of bedrooms in building, soil slope, soil type, and 62 other extraneous variables. In data cleaning phase, these extraneous values such as zip-code, State id, contractor name, contractor address, phone number, township name, address, description, comments were removed. Data prepossessing streamlined the data into integers/floats and categorized the data points using a technique called one-hot encoding. One-hot encoding is a technique where an integer coded variable is removed and a new binary column/variable is added for each integer value. For example, a few locations had a septic tank size of 750 gallons, others had 750.00 gallons and 7,50 gallons. All three values denote 750 gallons and the discrepancies were removed. For qualitative variables such as, soil type or soil fill any material, a two-step approach of encoding was adopted. For a variable such as soil type, all the unique values were located and each unique value was assigned an integer/float number. This ensures the complete dataset is in an unified quantitative format. After these prepossessing steps, the dataset size was 11,527 × 17 (rows × columns). One column denoted as Septic system status is chosen as dependent variable. This variable was further grouped into two categories such as repair and working systems and was chosen as the (Septic System status) dependent variable. The remaining 16 variables are declared as independent Table 1 and 2 . This dataset is referred as dataset one.

Since the dataset one was focused on a single county and it lacked a wide range of independent variables within Indiana, we sought other sources from across the state for a new dataset. iTOSS developed by the Indiana State Department of Health was utilized (Mettler and Atwood, 2010). iTOSS, a relatively recently developed database within an online platform provides web access for state personnel to analyze data, monitor reports, and perform real-time data retrieval. Utilization of iTOSS is more efficient than traditional data reporting since the format is streamlined and provides essential information rapidly. With the aid of iTOSS, a new dataset was collected composed of septic system records across the counties of Clark, LaPorte, Huntington, Spencer in the State of Indiana as shown in Fig. 5. The steps in Fig. 4, was carried out in the same manner in dataset two. The size of the retrieved dataset was 4332 × 17. The new dataset is referred as dataset two in further sections.

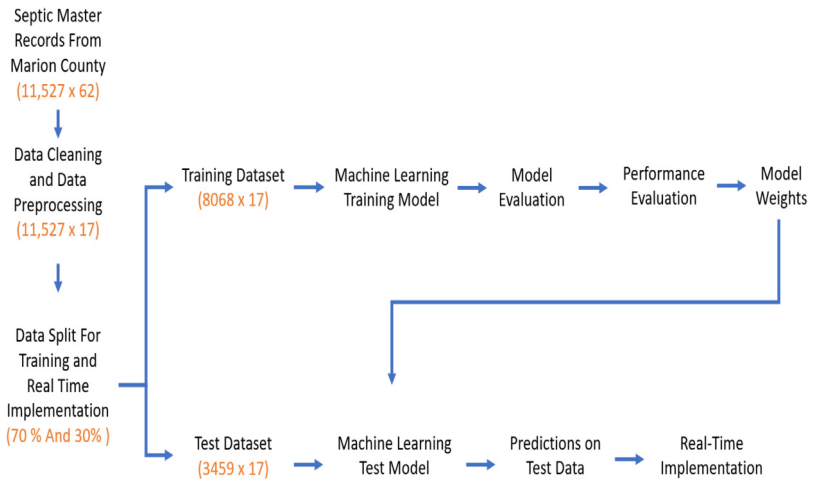

Fig. 4. Schematic layout of model development in dataset one.

Table 1
Independent variables utilized in the OSS failure model.

First issue date
Premises type
Number of bedrooms
Type of septic system
Soil compaction ratio
Soil loading rate
Soil slope
Soil type
Soil fill any material
Septic tank size
Pump chamber size
Absorption field square feet
Depth of trench
Glacial till depth
High water table
Maximum trench depth

Table 2
Dependent variables utilized in the OSS failure model.

System in repair
Replaced system
New installed systems

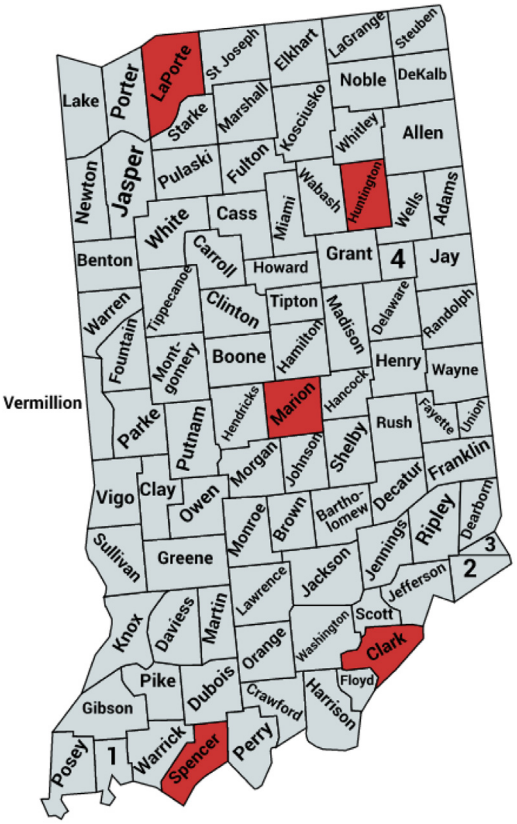

Fig. 5. Counties highlighted in State of Indiana.

At the end of data mining phase, we have two datasets. Third step involves splitting the dataset into testing and training components. 70% for developing and training ML models and 30% for real time implementation for later usage. Individual ML models were developed for each of datasets, one and two by considering the independent and dependent variables which are described in detail in Section 8. The performance evaluation and performance metrics of the models are discussed in Section 9.

9. Results

The models are trained in training data and its performance is evaluated in testing data. For the Marion county dataset, three ML models were built using logistic regression, RF, and KNN. Below is the table on the comparison of their performance metrics.

Model	Accuracy	Precision	Recall
KNN	79%	80%	81%
Logistic regression	80%	76%	81%
Random forest	80%	77%	80%

Based on the above observations from the models, there is a non-linearity between accuracy, prediction, and recall. This could be primarily due to an imbalanced dataset and many un-encountered points. Further analysis of ROC was carried out (Fig. 6).

The ROC curve (blue line) lies close to the threshold point and it shows that the developed models were not able to classify all the data points and RF classifier was able to achieve highest accuracy of 62.1%.

To achieve better performance and accurate results, we utilized the dataset from second phase of data mining. A dataset containing diverse data points across 4 major counties. The performance metrics yielded by RF classifier for both models are tabulated below.

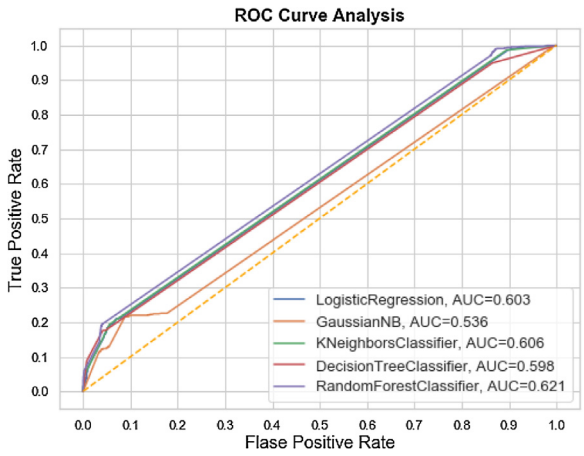

Fig. 6. ROC curve in dataset one.

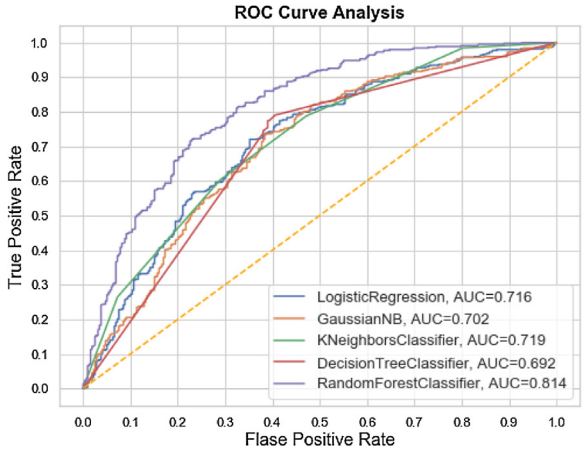

Fig. 7. ROC curve in dataset two.

Performance metrics	Model one	Model two
Accuracy	80%	72%
Precision	77%	73%
Recall	80%	72%

The corresponding ROC-AUC Curve was also analyzed in this case (Fig. 7). We could visualize the performance highly improved and achieved an accuracy of about 81.4%. This is significantly better and the linearity in performance metrics shows that the model was not biased and works efficiently. The results were of the model are called weights. With help of a python library called a pickle, the weights were saved for real-time testing

10. Discussion and implementation

As mentioned in Section 8, 30% of our dataset two was reserved for real-time implementation. A test script was developed which takes the above dataset as an input and applies the weights of our model to it. This would attach a new column to the dataset called 'Results'. This column would have binary values such as 'Possible Failure' or 'Good System'. This would help state officials to be extra cautious about locations that have a result column as 'Possible Failure'. They can alert local county officials corresponding to the location and they can perform additional field inspections in those places.

Further analysis was carried out to understand the importance of variables. Feature importance is a technique that assigns each of the independent variables a score, which denotes how efficient they are in determining the dependent variable. This can be further

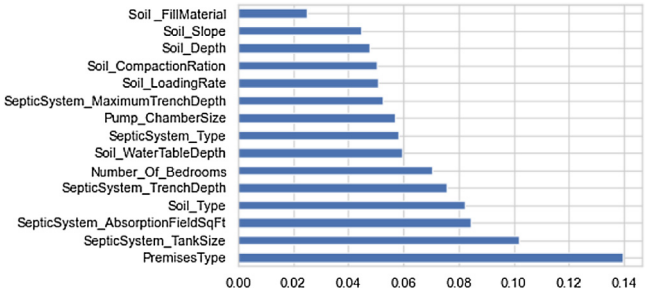

Fig. 8. Feature importance in dataset one.

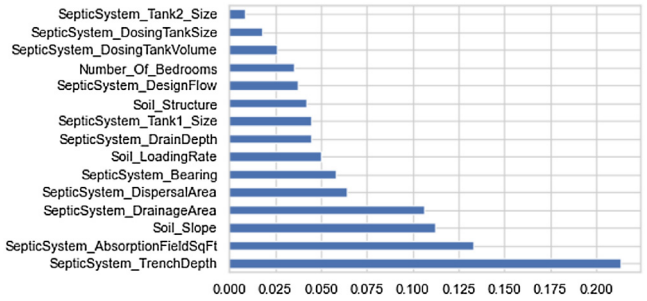

Fig. 9. Feature importance in dataset two.

explained as quantifying how accuracy will be affected if the feature/variable is removed. The first/largest 15 features with highest importance are obtained and shown in Figs. 8 and 9 . The y axis represents the feature names and x axis represents their impact in prediction. The factors with higher value indicates how useful they perform in actual prediction in terms of accuracy.

By visualizing the Figs. 8 and 9, we can identify the difference in feature importance between the datasets. The primary reason could be the size of datasets, dataset one was almost three times the size of dataset two. This causes varying distribution of independent variables in terms of mean, standard deviation and variance. Secondary reason is that, dataset one belongs to a particular singular geographic location (Marion county's) septic records. But the dataset two belongs to a wide range of geographic locations. Varying geographic locations can contribute various factors such as water table level, soil slope, type of soil and solid moisture content. Features from dataset one such as number of bedrooms, septic tank size, premises type have been identified as causes for septic failures in this research (Fix and Hodges, 1951). Soil type, soil slope features from dataset one also coincides with research findings with another research (Gou et al., 2012). On the other hand, highlighted features from dataset two such as drain depth, drainage area can cause clogging and result in septic failures as denoted by this research (KNN, 2021; Huang and Ling, 2005). In addition to that, there are few features which play a crucial role in determining septic failures in both datasets. They are soil slope, trench depth, soil loading rate, absorption field square feet, septic tank size and number of bedrooms. The findings of the research identify the critical parameters associated with septic system failures.

11. Conclusion

A brief overview of parameters involved in construction followed by working of an efficient septic system and causes for septic failures were discussed. Existing approaches of using UAVs and statistical approach to detect failure regions and their drawbacks to identify septic failures were discussed and need to develop a new monitoring system was examined. An AI based monitoring system was developed with the help of ML models we were able to iden-

tify underlying factors which play a crucial role in septic systems failure and identify locations where possible failures could occur. This research investigates the underlying factor as shown in Figs. 8 and 9, causing the septic system failures and how they align with other research findings in this sector. Septic records used in this research were collected across a range of counties in the State of Indiana which are geographically unique. Exploring various counties helped in understanding the diverse techniques employed in constructing a septic system. This approach is cost effective and could be expanded to a wide range of state regions. The above factors can be taken into consideration by State governments or local authorities to repair the existing systems or install superior systems in the respective households. This approach would help in ensuring and developing a safe and healthy environment.

Conflict of interest

This research was conducted with funding from Indiana State Department of Health and their direct supervision.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

We would like to show our gratitude to Mr. Mike Mettler, Mr. Mike Sutton, and Indiana State Officials for their excellent guidance and support throughout the work.

References

About Dosing Tank, 2021. https://www.onsiteinstaller.com. (Accessed 2 January 2020).
Beal, C.D., Gardner, E.A., Menzies, N.W., 2005. Process, performance, and pollution potential: a review of septic tank-soil absorption systems. Soil Res. 43 (7), 781–802.
Design Manual of Onsite Wastewater Treatment and Disposal System, 2021. https://www.epa.gov/. (Accessed 16 September 2020).
Establishing Failure Indicators for Conventional On-Site Wastewater Treatment Systems, 2021. https://ir.canterbury.ac.nz/bitstream/handle/10092/13692/Master. (Accessed 14 January 2020).
Evans, B.M., 1982. Aerial photographic analysis of septic system performance. Photogramm. Eng. Remote Sens. 48 (11), 17091712.
Fix, E., Hodges, J.L., 1951. Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Technique Report No. 4. U.S. Air Force School of Aviation Medicine, Randolf Field, Texas, pp. 238–247.
Forbis-Stokes, A., 2012. Modeling Onsite Wastewater Treatment Systems in the Dickinson Bayou Watershed (Doctoral dissertation). Texas A&M University.
Gou, J., Lan, D., Zhang, Y., Xiong, T., 2012. A new distance-weighted k-nearest neighbor classifier. J. Inf. Comput. Sci. 9 (6), 1429–1436.
Huang, J., Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. 17 (3), 299–310, http://dx.doi.org/10.1109/TKDE.2005.50.
Identification of Failing Septic Systems, 2021. https://www.hrwc.org/wp-content/uploads/HRWC20Septic20System20ID20Report20Final20v1.pdf. (Accessed 3 January 2020).
KNN, 2021. K Nearest Neighbor Algorithm in Python, . (Accessed 17 September 2020) https://towardsdatascience.com/k-nearest-neighbor-python-2fccc47d2a55.
Lahouar, A., Ben Hadj Slama, J., 2015. Random forests model for one day ahead load forecasting. In: The Sixth International Renewable Energy Congress (IREC2015), Sousse, pp. 1–6, http://dx.doi.org/10.1109/IREC.2015.7110975.
Mettler, M., Atwood, C., 2010. Indiana's Network for Tracking of Onsite Sewage Systems (iTOSS): Indiana State Department of Health's Data System for Managing Onsite Sewage System Data. Proceedings of the Water Environment Federation (11), 5568–5578, http://dx.doi.org/10.2175/193864710798193680.
Nguyen, C., Wang, Y., Nguyen, H., 2013. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomed. Sci. Eng. 6, 551–560, http://dx.doi.org/10.4236/jbise.2013.65070.
Random Forest, 2021. Applying Random Forest Classification, . (Accessed 16 September 2020) https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57.
Ravi, N., Chitanvis, R., El-Sharkawy, M., 2019. Applications of drones using wireless sensor networks. In: 2019 IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, USA, pp. 513–518, http://dx.doi.org/10.1109/NAECON46414.2019.9057846.
Septic System Status and Issues Working Paper, Metropolitan North Georgia Water Planning District, 2021. https://gowa.wildapricot.org. (Accessed 16 September 2020).
Septic System Working, 2021. https://www.vdh.virginia.gov/environmental-health/onsite-sewage-water-services-updated/how-systems-work/. (Accessed 16 September 2020).
Septic Systems, 2021a. How Does a Septic Tank Work?, . (Accessed 15 September 2020) https://www.familyhandyman.com/project/how-a-septic-tank-works/.
Septic Systems, 2021b. Pros and Cons of a Septic Tank System, . (Accessed 15 September 2020) https://www.mrrooter.com/oneida/about-us/blog/2017/march/pros-and-cons-of-a-septic-tank-system/.
Septic Systems, 2021c. The 6 Septic Systems You Must Know, . (Accessed 1 January 2020) https://buildwithabang.com/the-lowdown-topics/6-septic-system-types.
Septic Systems, 2021d. Types of Septic System, . (Accessed 1 January 2020) https://www.epa.gov/septic/types-septic-systems.
Septic Systems, 2021e. What to Do If Your Septic System Fails, . (Accessed 12 September 2020) https://www.epa.gov/septic.
Septic Systems, 2021f. Why Septic Systems Fail, . (Accessed 12 September 2020) https://thepinkplumber.com/news/why-septic-tanks-fail.
Septic Tank Serving Multiple Housing Units, 2021. http://ecp-inc.com/2017/09/07/can-a-septic-tank-serve-multiple-housing-units/. (Accessed 3 January 2020).
Simeone, O., 2018. A very brief introduction to machine learning with applications to communication systems. IEEE Trans. Cogn. Commun. Netw. 4 (4), 648–664, http://dx.doi.org/10.1109/TCCN.2018.2881442.
U.S. EPA, 1980. Design Manual: Onsite Wastewater Treatment and Disposal Systems. EPA 625/1-80-012. U.S. EPA, Washington, DC.
University of California Cooperative Extension, Calaveras County, n.d. Septic tanks: the real poop. University of California Cooperative Extension, Calaveras County, San Andreas, CA. http://cecalaveras.ucdavis.edu/realp.htm.
Onsite Wastewater Treatment Systems Manual. (Accessed 12 November 2020).

Niranjan Ravi has completed his Master's degree from Indiana University Purdue University, Indianapolis with his research in embedded systems and IoT. Currently, he is pursuing his Ph.D. in field of machine learning, a modern approach to solve various real world problems. His research interests are embedded systems, Internet of Things and UAV systems.

Dr. Daniel Johnson is an associate professor in the Indiana University – Purdue University, Department of Geography. Dr. Johnson's research interests are modeling complex human–environmental interactions particularly with a focus on environmental health and spatial epidemiological applications. His primary work focuses on vulnerability and its intersections with natural and man-made hazards.