JOURNAL OF
EMERGING INVESTIGATORS

Article

Machine learning approach to detect renal calculi by studying the physical characteristics of urine

Semontii Mandal[1], Sudeshna Roy[2]

[1] Auxilium Convent School, Kolkata, West Bengal, India

[2] BP Poddar Institute of Management and Technology, Kolkata, West Bengal, India

SUMMARY

Renal calculi (also called kidney stones) are hard deposits made of minerals and salts that form inside the kidneys. A kidney stone is usually asymptomatic until it moves around within the kidney or passes into one of the ureters, after which one may experience a sharp pain similar to a muscle strain. The current methods of detecting kidney stones (like CT imaging) cannot be employed until the onset of the first symptoms. Our approach provides a solution to detect kidney stones before the appearance of any symptoms. Using existing data on the physical characteristics of urine, we trained a machine learning model using various classifiers – logistic regression, decision tree and Support Vector Machine – to detect kidney stones and calculate their accuracies. Using one-way ANOVA test to compare the accuracies yielded by each approach of these classifiers, we found that there was no statistically significant difference between using one classifier over another. We plotted a confusion matrix and calculated the F1 scores for each classifier, in order to evaluate the performance of the algorithms. All the classifiers reported an accuracy of >0.80 and an F1 score of >0.85. Thus, our findings suggest that it is possible to detect kidney stones before the onset of symptoms by analyzing the physical characteristics of urine using machine learning classifiers.

INTRODUCTION

Kidney stones are irregularly shaped solid masses or crystals that can be as small as a grain of sand, up to the size of a golf ball. Kidney stones can cause severe pain, but symptoms may not occur until the stone begins to move down the ureters (1). This severe pain is called renal colic. Kidney stones can affect any part of the urinary tract, from the kidneys to the bladder and can cause pain on one side of the back or abdomen (2). Even small stones can cause extreme pain.

Research shows that about 1 in 10 people get a kidney stone during their lifetime (3). Obesity, some medical conditions like recurrent urinary tract infections (UTIs) and gout, taking certain medications, and not drinking enough liquids can increase the risk of kidney stones (4). Urine has various wastes dissolved in it. When there is too much waste in too little liquid, crystals begin to form. The crystals attract other elements and join together to form a solid that will get larger unless it is passed out of the body with the urine. Since kidney stones continue to get bigger, it is important that they are detected early to avoid damage to the ureters. Stones that do not move may cause a back-up of urine in the kidney, ureter, the bladder, or the urethra (5). As kidney stone is not often diagnosed until a patient presents with symptoms, there lies a need to develop techniques for early detection.

According to a study conducted by the University of Toronto and Imperial College London, it may be anticipated that some of the physical characteristics of urine are highly correlated with the presence of calcium oxalate crystals in the kidney (6). The study involved analysis of 79 urine specimens. The physical characteristics of urine studied to detect the presence of crystals were specific gravity, pH, osmolarity (mOsm), conductivity (mMho), urea concentration (millimoles per liter), and calcium concentration (millimoles per liter) (**Table 1**).

Though there are various methods to detect kidney stones like computerized tomography (CT) imaging and blood tests, most cannot be employed until after the onset of symptoms (7). In certain cases, these methods are expensive as compared to a simple urine analysis. Analyzing the physical characteristics of urine through a Machine Learning model is thus an easier way to detect kidney stones.

We hypothesized that machine learning classifiers can detect the presence of kidney stones if supplied with data containing certain urine characteristics. We trained our machine learning model using the above described dataset containing 79 samples of urine (6). To be more accurate about the performance of our Machine Learning model, we used multiple Machine Learning classifiers and tested it using multiple cross validations. The classifiers used were logistic regression, decision tree, and support vector machine (SVM). Our study achieved greater than 0.85 accuracy and supported our hypothesis that Urine analysis through Machine Learning models is a promising technique for the early detection of calcium oxalate crystals before the onset of symptoms. As supported by our results, regular urine screenings can provide a cheaper way to diagnose calcium oxalate crystals early.

RESULTS

We hypothesized that machine learning classifiers can accurately detect kidney stones through urine analysis if they are previously trained on similar data containing the physical characteristics of urine. The urine analysis dataset consists of six features (the physical characteristics of urine) and the diagnosis, all of which were used to train our classifiers, logistic regression, decision tree and support vector machine (SVM). We preprocessed our dataset to account for any missing values or to find any outliers. We found that our dataset did not have many outliers and hence the missing values in each column could be filled using the mean value of that respective column (**Table 1)**. Since the values of various

https://doi.org/10.59720/23-066

Specific Gravity	pH	Osmolarity (mOsm)	Conductivity (mMho)	Urea Concentration (mM/L)	Calcium Concentration (mM/L)	Crystals
1.01	5.97	343	13.4	126	2.31	0
1.02	5.68	876	35.8	308	4.49	0
1.021	5.94	774	27.9	325	6.96	1
1.024	5.77	698	19.5	354	13	1
1.022	6.21	442	20.6	398	4.18	1
1.011	6.13	364	10.9	159	3.1	1
1.025	5.53	907	28.4	448	1.27	0
1.024	5.6	866	29.5	360	5.54	1
1.007	5.35	283	9.9	147	1.47	0
1.011	5.21	450	17.9	161	1.53	0

Table 1. Dataset preview. Each column shows specific characteristics of urine that were studied, and their respective values obtained during analysis. In the 'Crystals' column, 0 signifies negative (i.e. no crystal found) while 1 signifies positive (i.e. crystals are present). Dataset obtained from the laboratory of James S. Elliot, M.D. of the Urology Section, Veterans Administration Medical Center, Palo Alto and the Division of Urology, Stanford University School of Medicine, Stanford was used to train our Machine Learning Classifiers (6).

columns of the physical characteristics dataset were very distinct in their range of magnitudes, we used feature scaling to bring these values within a similar range for all columns. After our dataset was ready, we split it into training data (90%) and testing data (10%). We trained all classifiers using the same training dataset. Then, our classifiers were tested on both training and testing data, and their accuracies were calculated (**Table 2**).

For further performance analysis of our model, we plotted a confusion matrix, displaying the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each classifiers' prediction (**Figure 1**). We used the outputs of the confusion matrices to determine our model's performance. Values 0 – 4 in the confusion matrices are indicative of a scale, that is, the minimum and maximum values of each cell in the matrices are 0 and 4, respectively. For example, in the logistic regression confusion matrix, the number of FPs obtained was 0 while 4 TNs were obtained.

After plotting the confusion matrices and determining the TP, FP, TN and FN values, we used the classification report visualizer to calculate the evaluation metrics — precision, recall and F1 score — and determine the performance of our classifiers (**Table 3**). As shown in the table, the logistic regression classifier achieved a precision of 1.0, recall of 0.75 and F1 score of 0.857 while the decision tree both achieved precision of 0.8, recall of 1.0 and F1 scores of 0.889. Since all our models performed well on the accuracy test as well as the precision, recall, and F1 Scores, we concluded that these models are capable of successfully detecting kidney stones. Comparing the three classifiers using a one-way ANOVA, we found that there was no statistically significant difference between the outcomes of detecting the presence of kidney stones while using any specific classifier ($p > 0.05$). Thus, our machine learning models support our hypothesis and suggest that these can successfully detect kidney stones through urine analysis.

DISCUSSION

Our primary aim was to determine the efficacy of machine learning classifiers in detecting kidney stones through urine analysis. In our analysis, all three classifiers—logistic regression, decision tree, and support vector machine—demonstrated commendable performance in the task of kidney stone detection through urine analysis. The logistic regression classifier exhibited a precision of 1.0, a recall of 0.75, and an F1 score of 0.857. In contrast, the decision tree and support vector machine (SVM) classifiers achieved a precision of 0.8, a recall of 1.0, and an F1 score of 0.889. We passed the data through multiple classifiers to compare the accuracy and performance of each of them. Since all models could predict the presence of calcium oxalate crystals in the Kidney satisfactorily, we concluded that their performance was nearly the same.

Classifier	Accuracy Score for Training Data	Accuracy Score for Testing Data
Logistic Regression	0.845	0.875
Decision Tree	1.0	0.875
Support Vector Machine (SVM)	1.0	0.875

Table 2. Accuracy scores obtained using various classifiers both for training and testing Data. The first 'Classifier' column lists the classifiers used while the second and third columns show the calculated values obtained while working with each classifier, both for training as well as Testing Data.

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. The TP, FP, TN, and FP values obtained allow us to determine the Precision, Recall and F1 values of our model, which, along with the Accuracy, give us the overall performance of our classifiers. Based on these metrics, we analyzed our model's overall performance. Despite training our model on a small amount of data, we achieved greater than 80% accuracy. These promising results suggest that training on a larger dataset could further improve the model.

The detection of Kidney Stones based on regular urine screening can be performed early, saving patients from tremendous pain and cost burden (8). As stated earlier, the main advantage to our method of detection over other methods like CT imaging is that it can be employed without the appearance of symptoms by analyzing the urine during a regular health checkup. Even if a person shows no symptoms, our method, if incorporated as a part of regular health checkups, can be used to detect kidney stones without requiring a separate test. Additionally, since kidney stones continue to grow once they are formed, they can descend down the ureters and block them. This can also cause bleeding and prevent urine from leaving the body. If they are detected early when they are smaller in size, they can be passed or removed without damaging the ureters, typically with mild medication. The longer a stone stays in the kidney, the larger it gets and the condition typically requires a medical intervention and even surgery if the stone formed is greater than six mm in diameter (9). Thus, this shows that the earlier kidney stones are detected, the safer their removal is.

MATERIALS AND METHODS

The dataset used for our analysis consisted of six physical characteristics of urine and 79 urine samples, with the data being obtained from the laboratory of James S. Elliot M.D. of

https://doi.org/10.59720/23-066

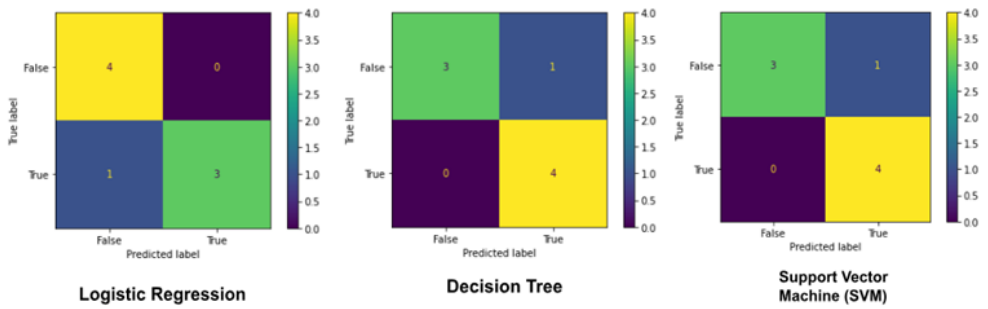

Figure 1. The confusion matrices plotted for the three classifiers. In our model's predictions, we categorized the results into four distinct scenarios based on their true and predicted labels. When the true label was marked as true and our model also predicted it as true, it was classified as a true positive (TP), representing instances where our model accurately identified the condition as positive. Conversely, when the true label was true, but our model predicted false, these instances were classified as false negatives (FN), indicating that our model incorrectly identified the condition as negative. On the other hand, when the true label was false, but our model predicted true, these cases were categorized as false positives (FP), signifying that our model mistakenly identified the condition as positive. Lastly, when both the true label and the model's prediction were false, these cases were labeled as true negatives (TN), illustrating instances where our model correctly identified the condition as negative. 0 – 4 in the Confusion Matrices depict the minimum and maximum values obtained in the matrices respectively.

the Urology Section, Veterans Administration Medical Center, Palo Alto and the Division of Urology, Stanford University School of Medicine, Stanford (6). The physical characteristics of urine studied in the dataset were: specific gravity, pH, osmolarity, conductivity, urea concentration, and calcium concentration. The specific gravity of urine is the density of urine relative to water. The pH of urine is the negative logarithm of the hydrogen ion concentration in urine. The osmolarity of urine is the measure of solute concentration, defined as the number of osmoles (Osm) of solute per liter (L) of solution (osmol/L or Osm/L) and is proportional to the concentration of molecules in solution. The conductivity of urine (in mMho) is the measure of a solution's ability to conduct electric current and is proportional to the concentration of charged ions in solution. The urea concentration in urine (in millimoles per liter) is the concentration of urea in urine. The calcium concentration in urine (in millimoles per liter) is the concentration of calcium ions in urine.

For training, we decided to split our data in a 9:1 ratio (90% training data and 10% test data) to train our models with as much data available given the small nature of the dataset. For building each classification model, we have used the scikit learn library. We employed the following methods to build our Machine Learning model and train it with data from our dataset.

Classifier	Precision	Recall	F1 Score
Logistic Regression	1.0	0.75	0.857
Decision Tree	0.8	1.0	0.889
Support Vector Machine (SVM)	0.8	1.0	0.889

Table 3. Summary of the precision, recall and F1 scores obtained for each classifier. The first column lists the three classifiers used while the second and third columns show the Precision and Recall scores obtained for each classifier, which in turn are used to calculate the F1 score shown in the fourth column. On comparing the three classifiers using a one-way ANOVA, we found that there was no statistically significant difference between the outcomes of detecting the presence of kidney stones while using any specific classifier ($p > 0.05$).

Data Preprocessing

Our urine analysis dataset consisting of 79 samples was preprocessed for filling any missing values. We programmed our model to fill in any missing values using the mean value of respective columns. We chose mean and not median because our dataset does not contain many outliers and the distribution is not skewed. We then used feature scaling to bring the values of different columns within a similar range.

Logistic Regression

Logistic regression is a statistical model that uses the logistic function to map y as a sigmoid function of x (10). We have used sklearn.linear_model.LogisticRegression class from the Scikit Learn library to build this classifier.

Decision Tree

A decision tree is an algorithm which splits the features in a dataset (root node) into multiple nodes (which may be split again until a terminal 'leaf node' is reached) to solve classification problems. For our Decision Tree classifier, we used the 'entropy' criterion and not 'gini'. Computationally, entropy is more complex since it makes use of logarithms and consequently, the calculation of the Gini Index is faster. However, the results obtained using the entropy criterion are slightly better in terms of accuracy (11). We used sklearn.tree. DecisionTreeClassifier class from the Scikit Learn library to build the Decision Tree.

Support Vector Machine (SVM)

The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a two-dimensional plane (12). We used gamma='scale' and not 'auto' for building our SVM model, as better performance was obtained with the former. We used sklearn.svm.SVC class from the Scikit Learn library to build our SVM.

https://doi.org/10.59720/23-066

Confusion Matrices

We also generated Confusion Matrices for further evaluation of our model's performance. Plotted using TP, FP, TN, and FN values, a confusion matrix is an N X N matrix for evaluating the performance of a classification model. This matrix was plotted using the Python Matplotlib library, and the Precision, Recall, and F1 Scores were calculated using the Scikit Learn libraries built-in functions.

Precision determines what proportion of positive identifications was actually correct (13).

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall determines what proportion of actual positives was identified correctly (13).

$$\text{Recall} = \frac{TP}{TP+FN}$$

Using the above obtained values, the F1 score was calculated as follows:

$$\text{F1 Score} = \frac{2*Precision*Recall}{Precision+Recall}$$

One-way ANOVA

We then implemented one-way ANOVA to compare the results of each model and calculated the p-value for each. Also referred to as one factor ANOVA, it is a parametric test used to test for a statistically significant difference of an outcome between 3 or more groups (14). The basic idea behind a one-way ANOVA is to take independent random samples from each group, then compute the sample means for each group. The variation of sample means among the groups is then compared to the variation within the groups and a test statistic is used to determine whether the means of the groups are all equal or not (15). Since we obtained a p-value greater than 0.05, there was no statistically significant difference between the outcomes of detecting the presence of kidney stones while using any specific classifier.

Received: March 31, 2023
Accepted: August 7, 2023
Published: March 06, 2024

REFERENCES

1. "Kidney Stones." *Mayo Clinic*. www.mayoclinic.org/diseases-conditions/kidney-stones/symptoms-causes/syc-20355755.
2. Alelign T, Petros B. "Kidney Stone Disease: An Update on Current Concepts." *Adv Urol.* 2018 Feb 4;2018:3068365. https://doi.org/10.1155/2018/3068365.
3. "Kidney Stones." *Cleveland Clinic*. my.clevelandclinic.org/health/diseases/15604-kidney-stones.
4. Cloutier, J., Villa, L., Traxer, O. et al. "Give me your stone, I will tell you who you are!". *World J Urol* 33, 157–169 (2015). https://doi.org/10.1007/s00345-014-1444-9.
5. "Kidney Stones." *National Institute of Diabetes and Digestive and Kidney Diseases*. www.niddk.nih.gov/health-information/urologic-diseases/kidney-stones.
6. Andrews, D.F. and A.M. Herzberg. "Physical Characteristics of Urines With and Without Crystals." Springer Series in Statistics, 1985, pp.249-50, https://doi.org/10.1007/978-1-4612-5098-2_45.
7. "Kidney Stones in Adults." *UpToDate*. www.uptodate.com/contents/kidney-stones-in-adults-beyond-the-basics.
8. Brisbane W, Bailey MR, Sorensen MD. "An overview of kidney stone imaging techniques." *Nat Rev Urol.* 2016 Nov;13(11):654-662. https://doi.org/10.1038/nrurol.2016.154.
9. "Surgery for Kidney Stones." *Washington University School of Medicine in St. Louis* urology.wustl.edu/patient-care/kidney-stones/surgery-for-kidney-stones/
10. "What is Logistic Regression?" *IBM.* www.ibm.com/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables.
11. "Decision Trees: Gini vs Entropy." *QuantDare*. quantdare.com/decision-trees-gini-vs-entropy/.
12. "Support Vector Machine Algorithm." *GeeksforGeeks*. www.geeksforgeeks.org/support-vector-machine-algorithm/.
13. "Classification: Precision and Recall." *Google Developers*. developers.google.com/machine-learning/crash-course/classification/precision-and-recall.
14. "One-way ANOVA with Python." *Python for Data Science*. www.pythonfordatascience.org/anova-python/
15. "ANOVA Test with Python." *Towards Data Science*. towardsdatascience.com/anova-test-with-python-cfbf4013328b