Int.J.HIT.TRANSC:ECCN.Vol.9: Issue 1A (2023) Page 54-68 ISSN: 0973-6875

Available Online at www.hithaldia.in/locate/ECCN
All Rights Reserved

ORIGINAL CONTRIBUTION

A COMPARATIVE ANALYSIS OF MULTIPLE APPROACHES MACHINE LEARNING FOR PREDICTING AND ANALYSING URINE PH AMOUNT

[a]Prajna Bhunia , [b]Sirsendu Das Adhikary, [c]Supriya Maity, [d]Dipankar Dey, [e]Samiram Pal

[abcde]*Global Institute of Science & Technology, Haldia, Purba Midnapur-721657, West Bengal, India*

Email: sirsendu1979@gmail.com, supriyamaity1234@gmail.com, deydipankar2014@gmail.com, samiran.sip@gmail.com

ABSTRACT

Prenatal treatment includes clinical urine testing as a crucial element.Medical professionals now evaluate urine test strips using an operator-dependent, labor-intensive, and visually color-coded process that takes a long time. Procedures and methods: By using various treatment and resource recovery techniques, urine has the potential to offer numerous useful resources. Selecting which technique to utilize and what re- sources might be retrieved from human urine, we paid particular attention to pH because it was thought to be the most significant parameter. We made a distinction between fresh, hydrolyzed, and stabilized urine treat- ment methods. For optimum resource recovery, future studies should concentrate on a thorough economic and life-cycle assessment of the urine treatment process. It has been shown that ML and AI are beneficial in a variety of fields, particularly with the current explosion of data. Making quicker and more accurate judgments in terms of illness forecasts may be possible using this method. Machine learning algorithms are therefore increasingly being used in prediction applications. Because of its high degree of accuracy, ML has been adopted by clinical diagnostics as one of the main computational approaches and analytics for illness identification. In order to increase the consistency and quality of disease reporting, building a model can also help us visualize and analyze diseases. This article has investigated how to predict the average pH value of urine. Different ML algorithms, including Linear Regression, Support Vector Machine, Neural Network, Gaussian Process Regression, and Fine Tree, are used to learn and find meaningful patterns. There are sev- eral insightful discoveries in this article. The R^2 number is used to assess the accuracy of machine learning methods, including Fine Tree, Gaussian Process Regression, Neural Network, Support Vector Machine, and Linear Regression. According to recent research, the Linear Regression algorithm has the lowest RMSE value when compared to other algorithms and a high accuracy rate of 0.99997 for R^2. Nevertheless, the difficult and future research area for these studies will be to raise the accuracy rates of the machine learning algorithms.

KEY WORDS: Machine Learning, Linear Regression, Support Vector Machine, Neural Network, Gaussian Process Regression, Tree, Urine, pH

1. INTRODUCTION

The kidneys' waste product, urine, acts as a vehicle for the body to remove undesired substances from circulation. Water, salts, and kidney waste are all components of urine. The pH value used by experts to evaluate urine acidity can be affected by the balance of various chemicals. Urine is an easy material to collect and may be obtained in large quantities, making it useful for analysis. Urine also contains a number of biological and chemical components that can be utilised as biomarkers for different illnesses. Waste products from the blood are extracted by urine, including salts and ions (2.8% combined), ammonia (0.2%), creatinine, uric acid, and urea (2% of the total urine ejected). The remaining 95% of the pee that is evacuated is composed of water (Biswas et al. 2022).

**Corresponding Address: sirsendu1979@gmail.com* Doi: 10.5281/zenodo.14740344

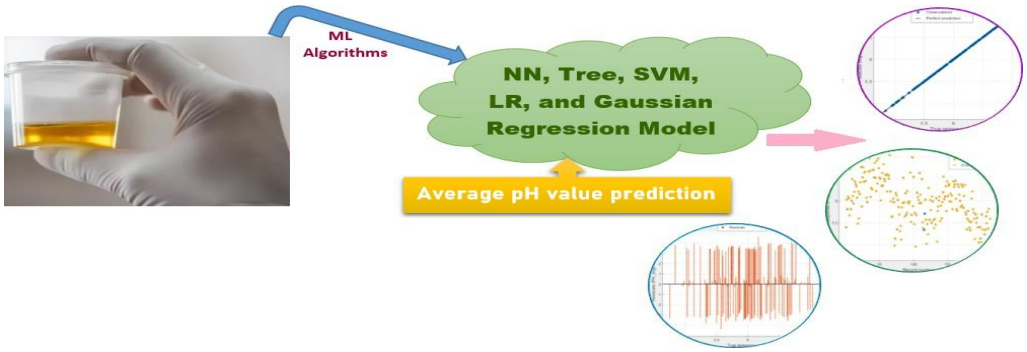

Figure 1: Graphical abstract

Urine analysis has a lot of potential in this regard because of its biological diversity and ability to be a practical and affordable health testing medium (Sheerin 2011). The average person excretes between 400 and 2000 mL of urine per day in $6-7$ urinations, therefore there are many samples available for collection and analysis. Utilizing urine analysis, one may assess the concentrations of typical urine components or look for unusual compounds that could point to a medical issue. There are four types of urine analysis tests: bacterial, chemical, physical, and microscopic. Commonly observed physical characteristics include colour, appearance, specific gravity, volume, and smell. Chemical elements measured include blood, urobilinogen, glucose, pH, ketones, protein, bilirubin, ascorbic acid, and nitrites. Urine can also contain bacteria, crystals, cells, and other materials when examined under a microscope. Urine analysis has been used for other diagnostic objectives, such as the diagnosis of pregnancy, dehydration, renal function, diabetes, and urinary tract infections (UTIs). Along with a number of biomolecules, urine often contains urea, salt, chloride, amino acids, sulphate, potassium, phosphate, and other trace chemicals. Certain medical disorders can be diagnosed by looking for abnormal amounts of these compounds, unusual urine chemical characteristics, or the presence of specific additional chemicals or molecules. Urine's specific gravity and pH can be used to track problems with the urinary system, such as dehydration and renal tubular acidosis (Massy et al. 2022).

A subset of artificial intelligence (AI), machine learning (ML) enables software applications to anticipate more accurately without explicit coding. Machine learning algorithms predict new output values by using historical data as input. The area of machine learning is vast and diverse, with a growing range of applica- tions. As a result, machine learning is now a crucial differentiator for many businesses. Machine learning employs supervised, unsupervised, and ensemble learning classifiers to forecast and assess a dataset's cor- rectness. ML algorithms may build a model from train data, which are little samples of data, to draw a conclusion or make a prediction (Chang et al. 2022). This research project looks into how machine learning techniques are applied in the medical field. Its main goals are to mimic specific human behaviours or mental processes and identify diseases based on a range of inputs.

In this work, we have the ability to forecast the average quantity of pH present in urine by taking 212 samples of urine. The prediction was done using neural network (Li, Sengupta, and Hanigan 2019), tree (Zhang et al. 2023), support vector machine, gaussian process regression (Schulz, Speekenbrink, and Krause 2018), and linear regression. Finding a machine-learning model that can predict the pH amount of urine using experimental data collected from numerous places across India and readily

accessible synthetic data is a major goal of this study.

- This study provides a description and statistics of the data set used in this investigation.
- The proposed approach is then addressed in greater detail, along with the process of urine in the human body and the pH amount in urine.
- In the last section, a conclusion is drawn based on the best accuracy rate of prediction after an expla- nation of the experimental results and analysis utilizing comparison analysis.

2. Process of urine formation

There are various stages in which the kidneys create urine: filtration, reabsorption, secretion, and excretion. Nephrons in the kidneys—the renal tubule and renal corpuscle—produce urine. The first step takes place in the renal corpuscles and is called filtration. The Bowman's capsule, a densely packed network of capillaries encircling the glomerulus (Khan et al. 2022), is one component of the renal corpuscle. Filtration occurs when blood enters the capillaries at a locally high blood pressure. Plasma, water, dissolved waste products, and small proteins are transported into Bowman's capsule by the capillaries. These capillaries shouldn't be present in typical pee since they serve as a filtration barrier, keeping big proteins and blood cells in the blood while letting other filtrates pass. The gathered material, known as the renal filtrate, moves on to the following stage of the urine generation process by moving further into the nephron.

Reabsorption, the second stage of urine generation, takes place in the renal tubule. This stage aims to recover and reintroduce any beneficial solutes and nutrients into the bloodstream in order to preserve them. From the renal filtrate, the renal tubule gathers vital ions, glucose, amino acids, vitamins, and other minerals. These substances collect in the nephron's interstitial fluid, where they osmotically draw water out of the renal filtrate. Through capillaries, the water and other elements subsequently make their way back to the renal vein.

During simultaneous secretion and reabsorption, waste ions such as sodium, potassium, hydrogen, cal- cium, ammonia, creatinine, and other chemicals such as pharmaceuticals pass from the blood through adja- cent capillaries to the kidney's interstitial fluid and then into the renal filtrate in the renal tubules. The urine is now ready to be drained. The last renal filtrate exits the kidneys and is retained in the bladder for a brief period of time before being subsequently eliminated from the body. Urine's production process reveals that it contains a wide range of compounds that are connected to the body's functionality and overall health, both directly and indirectly. Blood filtration directly results in urine.

3. pH quantity in urine

How acidic or alkaline a person's urine is determined by its pH. When a patient exhibits symptoms that might be indicative of a kidney or urinary tract issue (Argyle and Baldwin 1988), doctors frequently test the pH of the urine and do additional diagnostic procedures. The pH of the urine should be between $4.5 - 8$. The typical urine pH range for most persons is $6 - 7.5$. One of the key elements influencing urine pH is the food a person eats. Diet, certain drugs, infections, diarrhea, high blood sugar, acid-base disorders, poor kidney functions, genes, vomiting, and endurance exercise can all affect the pH of urine. While diets heavy in fruits and vegetables can raise urine pH, diets high in protein from meat, fish, dairy, and grains can lower urine pH. Higher urine pHs either promote the solubility of the poison or increase the fraction of the poison that is ionised, improving urine elimination, in the case of weak acids (those with a dissociation constant, pKa, that is only slightly lower than the pH of the urinary physiologic solution). The main problems linked to high and low urine pH levels are kidney stones, obesity and excess body weight, metabolic syndrome, non-alcoholic fatty liver disease, bladder cancer, and advanced chronic kidney illnesses. After an overnight fast, a urine sample is given for a normal urinalysis. A fasting urine sample is what it is called. Typically, an early-morning sample is more alkaline. A spot urine sample is obtained at random during the day. It is well known that the

pH of urine fluctuates throughout the day, rising in the middle of the day and falling after each meal. 24-hour urine tests are performed to provide an accurate assessment of the average urine pH throughout the day. A dipstick or a pH meter that employs an electrode can be used to determine the pH of urine.

4. Methodology

Primarily, the initial phases of this study entail obtaining information and selecting relevant characteristics. After that, the relevant data is preprocessed into the required format. Next, two groups are created from the provided data: training datasets and testing datasets. The supplied data and procedures are then used to train the model. The accuracy of this model is determined using the testing data. The procedures of this inquiry are loaded by a number of modules, including data collection, attribute selection, pre-processing, data balancing, and disease prediction.

4.1 Data Collection

This article's dataset was acquired in order to predict the pH quantity of pee throughout the day. In order to construct an efficient pH prediction model, we train k models, one for each category of urine quality. Every day, a total of 212 data samples are obtained from the urine sample. These data samples were used in this study and were previously explained. Three times a day, human urine samples were taken to track variations in the urine's pH quantity over time. The data samples were used in this investigation to determine the regression value. This study starts with organising the dataset, which is then split into training and testing phases. Eighty percent of the dataset is used for testing, while the remaining twenty percent is used for training. This article considers all the relevant elements in addition to the variable quality rating. The variations in pH amount during the day are depicted in Fig. 1.

4.2 Dataset and Attributes

The properties of a dataset are its qualities, which are important to analyse and predict concerning our problem. The patient's time, nutrition, and other characteristics are taken into account while estimating the pH level of urine.

4.3 Pre-processing of Data

Data cleaning is the process of removing noise and missing values from a dataset in order to obtain exact and perfect results. We can use a couple standard techniques to fill in the gaps and noise. Next, we need



Figure 1: pH quantity of a day of human body

to adjust our dataset by considering its normalisation, smoothing, aggregation, and generalisation. One of the most crucial phases of data preprocessing is integration, which takes a number of things into account. Occasionally, the dataset is more intricate or challenging to comprehend. To achieve the optimal outcome in this instance, the dataset must be reduced to the necessary format.

4.4 Balancing of Data

To enhance machine learning algorithms' performance, the dataset must be balanced. Each output class (also known as the goal class) in a balanced dataset has an equal number of input samples. It is possible to balance the unbalanced dataset by taking into account two techniques, such as under- and over-sampling.

4.5 Prediction of Disease

Five distinct machine learning algorithms are used for categorization in this article. An examination of the algorithms' comparisons has been conducted. In conclusion, the ML algorithm presented in Fig. 2 provides the highest accuracy

rate for the prediction of heart disease and is the subject of this article.

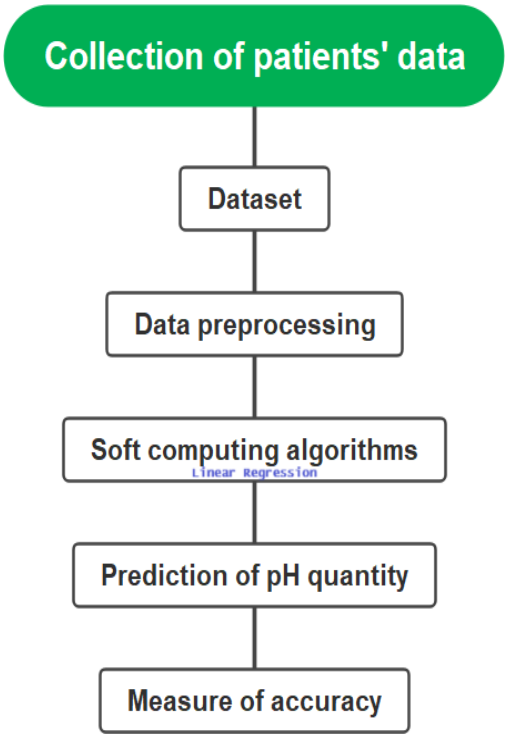

Figure 2: Architecture of prediction models using soft computing algorithm

5. Soft Computing Algorithms

The word ”soft computing” refers to a broad range of methods, strategies, and theoretical frameworks that share the objective of solving complex problems (like regression, pattern recognition, systems control, opti- misation, and prediction) using models and a variety of characteristics like fuzziness, missing data, and noise and non-linearity distribution. Furthermore, the data itself can be vectorial, scalar, etc., with a temporal or geographical component, and its sources can be either discrete or continuous variables. In this observation, five distinct algorithms are examined to determine which one is the most accurate.

5.1 Mechanism on Neural Network

A neural network’s cutting-edge perspective comes from the neurons that have been assembled into a single unit to construct it. A neural arrangement can be an artificial neural architecture designed to address artificial intelligence issues, or it can be a naturally occurring brain structure composed of several neurons (Basheer

and Hajmeer 2000). Because they function similarly to the human brain, they aid computers in recognising designs and solving common issues in the fields of computer-based intelligence, artificial intelligence, and deep learning. Neural networks are the basis of deep learning techniques; one subject of neural networks is artificial neural networks (ANNs). The human cerebrum served as the inspiration for both their name and appearance, which are based on the communication between actual neurons. The multi-input, multi- output systems that make up neural networks are constructed from artificial neurons. Because all neurons are connected to one another, each one affects the others. The main objective of a neural network is to translate input into a predefined output. Each element of the current data collection, as well as any potential relationships between the various data points, can be identified and highlighted by the organisation. In this approach, massive volumes of data can be mined by neural networks to find incredibly complicated examples. Fig. 3 displays the neural network’s schematic design.

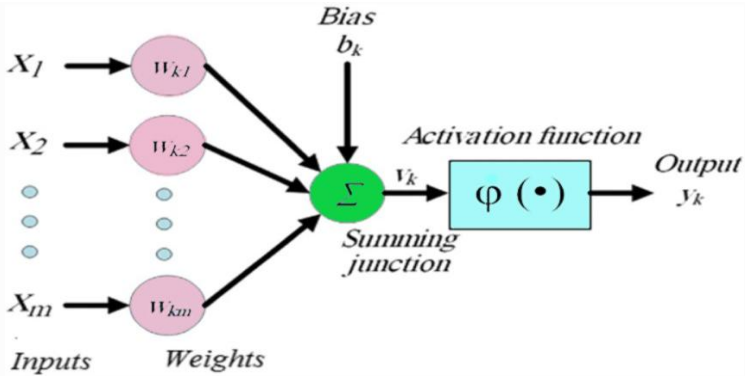

Figure 3: Schematic representation of Neural Network Model

5.2 Mechanism on Support Vector Regression

Vladimir Vapnik created the Support Vector Machine (SVM) in 1979. SVM stands for

Supervisory Type Machine Learning, which examines and identifies patterns in input data to perform regression analysis or classification. SVM is used in a wide range of applications, such as time series forecasting, face detection, handwriting recognition, digit identification, and cancer classification. Fig. 4 depicts all the components of this method.

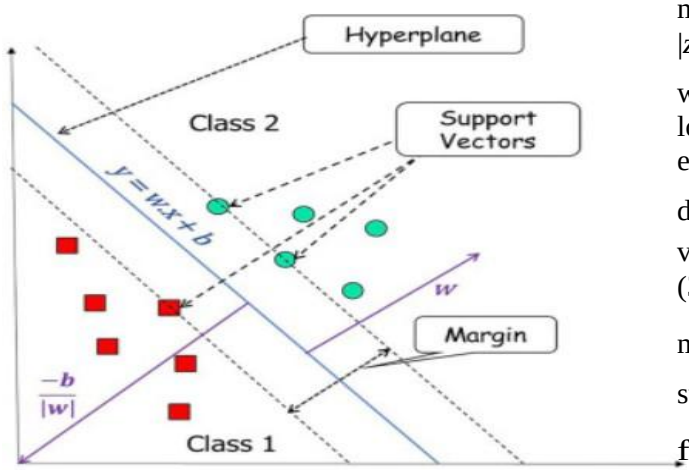

Figure 4: Basic components of SVM

Support vector regression (SVR) is the common name for SVM when it is used for regression. Fig. 5 displays the SVR network design. By employing SVR, users can freely define the maximum permitted error for the model, and the model will find the optimal line or hyperplane in higher dimensions to fit the data. SVR is one of the powerful algorithms that allows users to choose how forgiving they are of errors by using an acceptable error margin (ϵ) and modifying our tolerance for exceeding that acceptable error rate.

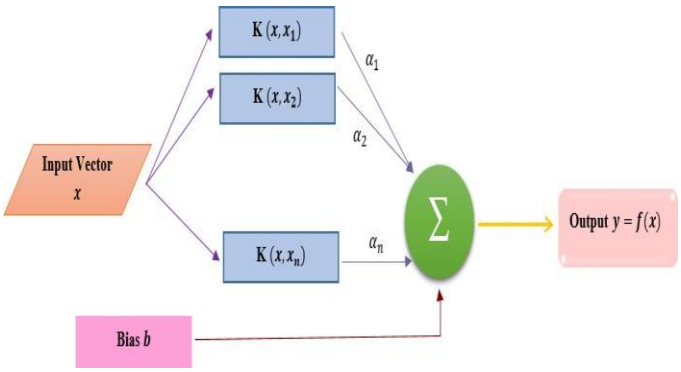

Figure 5: Basic network architecture of SVR

Let Eq. (1) represents the model for SVR,

$$f(x) = w^T x + b \qquad (1)$$

where w and b are the undermined parameters. The primitive optimization problem for SVR be formulated as Eq. (2)

$$\min_{1/2}\ w^2 + C\ {}^{n}_{i=1}\ (l_e(f(x_i - y_i) \text{ and } l_e(z) = 0 \text{ if } |z| \leq \in_i) \qquad (2)$$

where l denotes a margin of tolerance and the loss function used to assess the quality of the estimate, and C

denotes a regularisation parameter. The slack variables ξ_i and $\hat{\xi}$ can be introduced to write Eq. (3) as

$$\min_{w,b}\ 1/2\ \|w\|^2 + C\ {}^{n}_{i=1}\ (\xi_i + \hat{\xi}) \qquad (3)$$

subject to the following constraints

$$f(x_i) - y_i \leq +\xi_i,\ y_i - f(x_i) \leq +\hat{\xi}_i \xi \geq 0,\ \hat{\xi}_i 0,\ i = 1, 2, \cdots, N. \qquad (4)$$

The optimisation problem given above can be converted into a dual problem and solved by

$$f(x) = X \sum^{n}_{i=1} (\hat{\alpha}_i - \alpha_i)\kappa(x, x_i) + b \qquad (5)$$

$$0 \leq \alpha_i C,\ 0 \leq\leq C \qquad (6)$$

where $\kappa(x_i, x_i) = \Phi(x_i)T\phi(x_j)$ is the function of the kernel.

5.3 Mechanism on Tree

A tree-shaped technique called a decision tree is typically used to categorize and estimate targets. This approach divides a group into leaf neurons, inner neurons, and root neurons, which together create the branches of an overturned tree. These sections and nods are displayed and viewed as approximations of models in data mining (Amiri-Ramsheh et al. 2022). Major advantages of the DT method are its modest needs for data preparation and its high performance on big samples (Yang and Fong 2013).

The primary goal of DT model construction is realising a feature to be tested on one neuron before detaching from another neuron. Finding a

feature to test and the branch (Vanfretti and Arava 2020) is the process of splitting.

A split in a tree can lower the uncertainty in the dataset regarding the class by measuring the information gain. There are two parts to this measurement process. First, the dataset's Entropy (ENT) is determined. By using entropy, the projected information gain is calculated.

5.4 Mechanism on Linear Regression

Linear regression is among the most basic and widely applied Machine Learning algorithms. It's a statistical method for doing analysis that's predictive. Forecasts are generated using linear regression for continuous, real, or numerical data. The most accurate linear equation capturing the relationship between the explanatory factors and the dependent variable is found by an autoML tool using supervised machine learning. To do this, a line is fitted to the data using least squares. The line tries to bring down the residuals' squared sum. The residual is the difference between the actual value of the explanatory variable and the line. Finding the line of best fit requires iterative determination. The linear regression procedure, also referred to as linear regression, can demonstrate a linear relationship between a dependent variable (*y*) and one or more independent (*y*) variables. Finding out how the value of the dependent variable changes as a function of the value of the independent variable may be done using linear regression since it demonstrates a linear relationship. In the linear regression model, as shown in Fig. 6, the relationship between the variables is represented by a sloping straight line. A linear regression can be mathematically represented as Eq. (7)

$$y = a_0 + a_1 + \epsilon \qquad (7)$$

where, y = Dependent Variable (Target Variable), x = Independent Variable (predictor Variable), a_0 = intercept of the line (shows an additional degree of freedom), a_1 = Linear regression coefficient (scale factor to each input value), ϵ = random error. The values for x and y variables are training datasets for

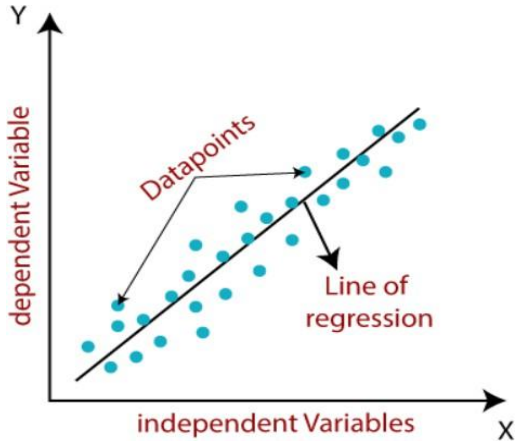

Figure 6: Graphical representation of Linear Regression Model

Linear Regression model representation. Eq. (8) is an example of a resulting linear regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots \qquad (8)$$

In the example above, y is the dependent variable, and x_1, x_2, and so on are the explanatory variables. The coefficients (b_1, b_2, and so on) explain the correlation of the explanatory variables with the dependent vari- able. The sign of the coefficients ±) designates whether the variable is positively or negatively correlated. b_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

A linear regression model can be used to forecast the value of a dependent variable and assess the pre- diction's accuracy. This is indicated by the p-value and R-squared values. The R-squared number indicates how much of the variance in the dependent variable can be explained by the explanatory variable, but the p-value indicates how reliable that explanation is. R-squared values range from 0 to 1. With a value of 0.8, the explanatory variable may explain 80% of the variation in the dependent variable's observed values. A forecast may be made perfectly if the number is 1, however this seldom happens in real life. If the explana- tory variable has a value of 0, it cannot predict the dependent variable in any way. It may be determined whether the

explanatory variable's impact on the dependent variable differs substantially from 0 by using a p-value.

5.5 Mechanism on Gaussian Progress Regression

Every finite subset of the collection of random variables in a Gaussian process has a multivariate normal distribution. That is to say, for an index set *X*, a real-valued stochastic process $f(x)$, $x \epsilon X$ is a Gaussian process if, for any subset $x = (x_1, \cdots, x_n) \epsilon X$, $f(x)$ has a joint Gaussian distribution. Then, its mean function *m* and covariance function *k* may be expressed as Eq (9) to represent it perfectly.

$$f \sim N(m(x), K(x, x')) \qquad (9)$$

where $K(x, x')$ is the covariance matrix with entries $K_{i,j} = k(x_i, x')$. Alternatively, it is possible to think of the Gaussian process as a multivariate Gaussian with an infinitely large number of random variables. Any real-valued function can be the mean function *m*, and it is frequently made to zero by deducting the mean from the data. Any legitimate Mercer's kernel, such as the exponential squared kernel already demonstrated, may be used as the kernel function *k*. A Gaussian process is frequently expressed as Eq (13) in this manner.

$$f(x) \sim GP(m(x), k(x, x')) \qquad (10)$$

To sample functions from the Gaussian process, one just needs to supply the covariance and mean func- tions. The covariance function k represents the joint variability of the random variables employed in the Gaussian process. The modelled covariance of each pair of inputs is given back. This covariance function is specified by the kernel function, which also proposes a distribution across functions. With this distribution, prior information can be set by choosing a specific kernel function. Sample function evaluations for a func- tion obtained from a Gaussian process can be performed at a finite, yet arbitrary, number of locations. The grey region and the black line in Fig 7 represent the standard deviation and the mean, respectively, while the 10 samples from the previous and posterior distributions are shown. We can see that the weighted average of the observable variables is a Gaussian process.

The above function in eq. 13 consists of the mean function m(x) and the covariance function $Covf(x, x')$

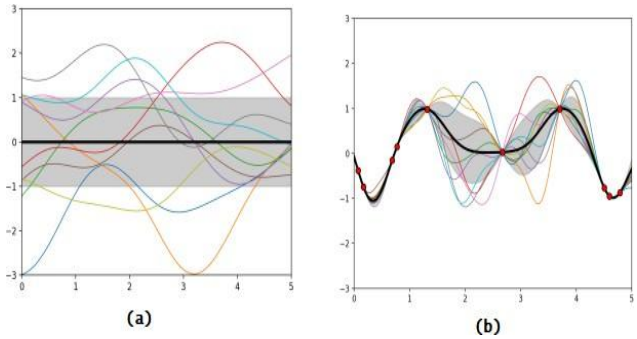

Figure 7: Example of Gaussian Process functions sample for the prior (a) and post (b) distributions

as follows:

$$m(x) = E[f(x)] \quad (11)$$

$$covf(x, x') = E[(f(x) - m(x)) \cdot (f(x') - m(x'))] \quad (12)$$

The purpose of this study is to figure out what the function f(x) means at the input point x. To make the computation easier, the mean value function m(x) is usually set to zero, and the covariance function of the square index is utilized as the covariance function, which is represented as:

$$m(x) = 0 \qquad (13)$$

$$covf(x, x) = \delta_{cov}^2 exp - (x-x)^2/2s^2 \qquad (14)$$

Where δ_{cov}^2 is the variance of the white Gaussian noise signal and *s* is the variance scale.

6. Results and Discussions

In soft computing, it is not necessary for consumers to have a detailed understanding of internal implementa- tion. According to this theory, the initial data sets could also be used as training examples to keep the models updated. The structure uses several techniques to create the final prediction model. Different algorithms use various theories to learn the models. Choosing the

optimum method or size fit to learn the models is difficult.

Five soft computing techniques were used in this study to configure models. To get a better outcome, the structure was modified to include the model that demonstrated the best-fitting performance.

6.1 Outcomes of Various Algorithms

This dataset has been analysed before the performance of taking machine learning techniques into consider- ation in this research is examined. Collecting urine samples from a human body three times a day to predict the average pH quantity of any person to identify whether there is any issue related to urine and the urinary system or not.

6.2 Evaluation Matrices

To evaluate the performance of various algorithms following evaluation matrices have been used.

- A statistical indicator known as the R-squared (R^2) value indicates the proportion of the dependent variable's variance that the independent variables in a regression model were able to predict.
- MSE is a statistical measure of the average squared difference between the predicted and actual values in a dataset. Outliers can affect the model's overall error since the MSE is susceptible to them.
- The root mean square error (RMSE) is equal to the square root of the mean squared error (MSE).
- Mean absolute error (MAE) is a statistical measure of the average absolute difference between the values that were predicted and the actual values.

6.3 Performance Analysis

In this article, various machine learning algorithms like Linear Regression (LR), Support Vector Machine (SVM), Fine Tree (FT), Gaussian Regression (GR), and Neural Network Regression (NNR) are studied broadly to predict the pH quantity. Through the response plot of each algorithm, the discrepancy between the expected and actual values i.e., error has been shown in Fig 8, 12, 16, 20, and 24. The comparison between the observations and the perfect prediction values of each model is shown in Fig 9, 13, 17, 21, and 25. The discrepancy between the fitted response values and the observed response is displayed in a residual plot. This residual plot of each model is shown in Fig 10, 14, 18, 22, and 26. RMSE, R^2, MSE, MAE, Prediction Speed, and Training Time, all for each algorithm are computed shown in Fig 11, 15, 19, 23, and 27. The accuracy rate of each algorithm has been measured, and the algorithm with the highest accuracy has been selected. Here the accuracy rate has been calculated through R^2. The statistical measure of fit known as R^2 shows how much of the variation in a dependent variable in a regression model is explained by the independent variable or variables.

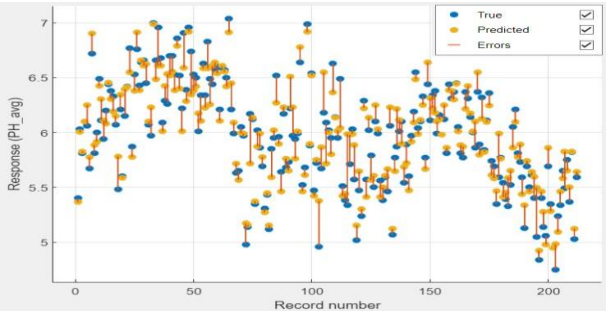

Figure 8: Response plot of Tree

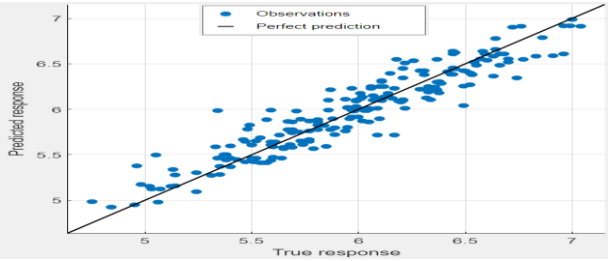

Figure 9: Validation prediction Vs actual value plot of Tree

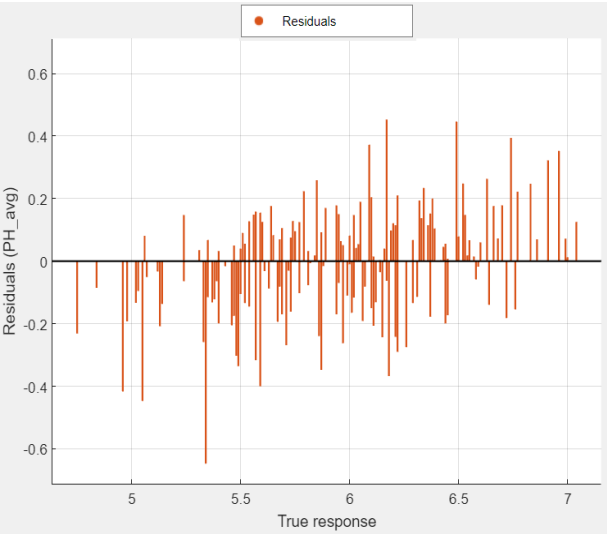

Figure 10: Validation residual plot of Tree

Model : Tree
Status: Trained

Training Results

RMSE (Validation)	0.16431
R-Squared (Validation)	0.89
MSE (Validation)	0.026998
MAE (Validation)	0.12759
Prediction speed	~3200 obs/sec
Training time	2.2256 sec

Figure11: Model details of Tree

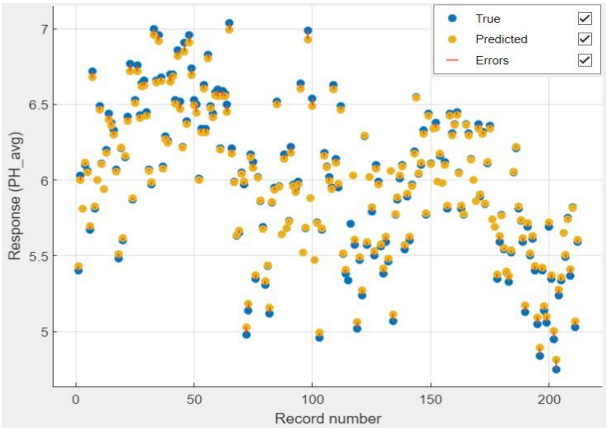

Figure 12: Response plot of Support Vector Machine

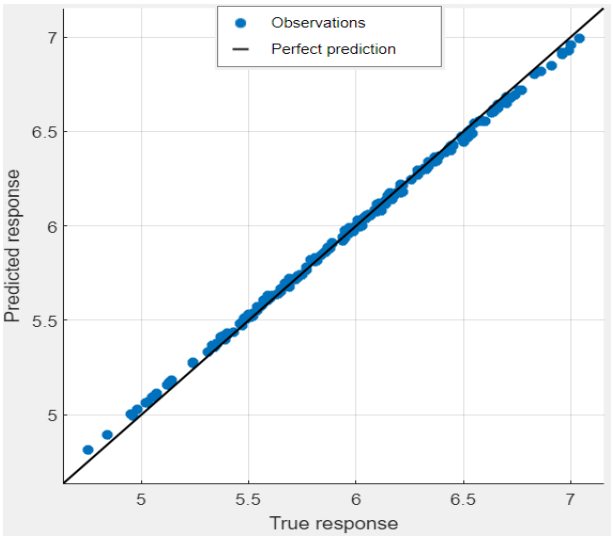

Figure 13: Validation prediction Vs. actual value plot of Support Vector Machine

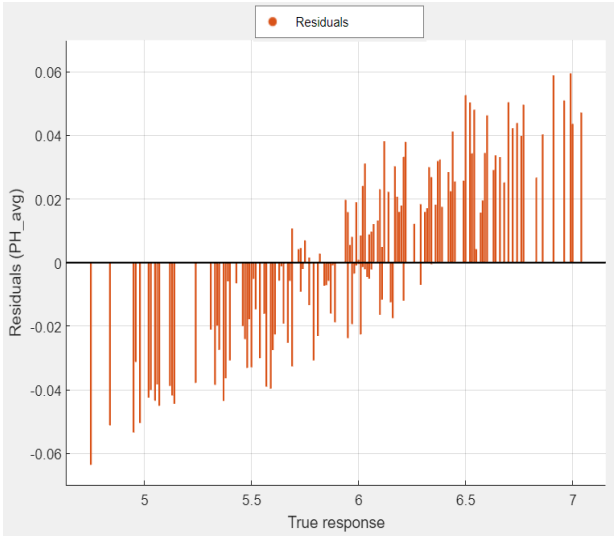

Figure 14: Validation residual plot of Support Vector Machine

Model : SVM
Status: Trained

Training Results

RMSE (Validation)	0.02576
R-Squared (Validation)	1.00
MSE (Validation)	0.00066355
MAE (Validation)	0.021138
Prediction speed	~12000 obs/sec
Training time	4.8543 sec

Figure 15: Model details of Support Vector Machine

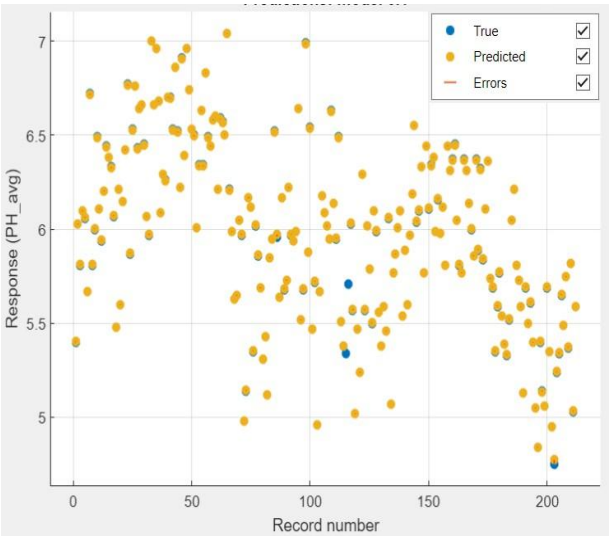

Figure 16: Response plot of Neural Network

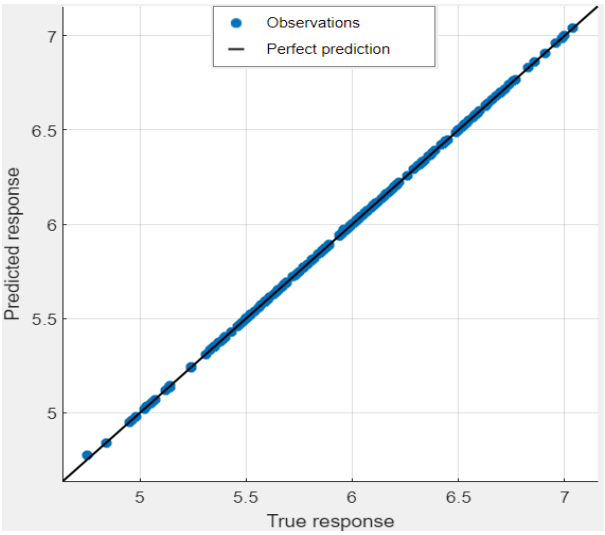

Figure 17: Validation prediction Vs actual value plot

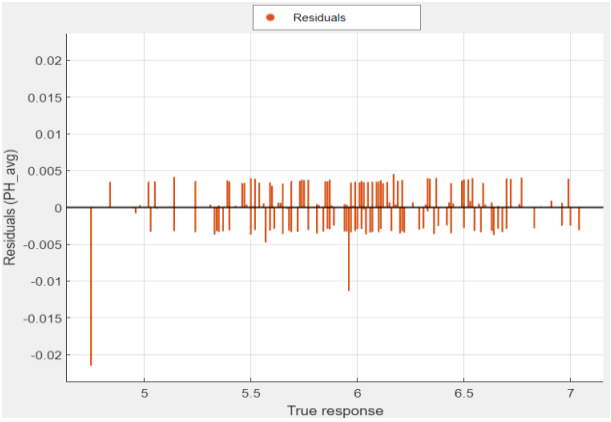

Figure 18: Validation residual plot of Neural Network

Model : Neural Network
Status: Trained

Training Results

RMSE (Validation)	0.0031853
R-Squared (Validation)	1.00
MSE (Validation)	1.0146e-05
MAE (Validation)	0.0024217
Prediction speed	~11000 obs/sec
Training time	6.0245 sec

Figure 19: Model details of Neural Network

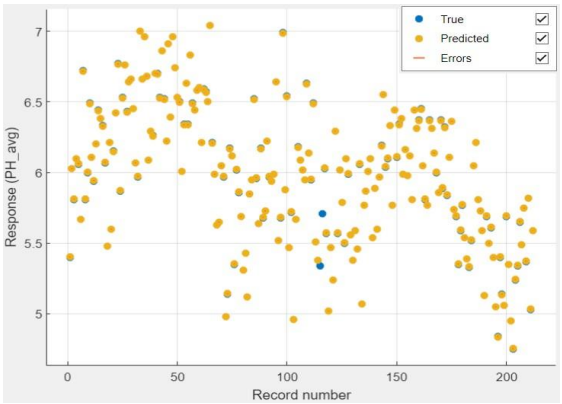

Figure 20: Response plot of Gaussian Process Regression

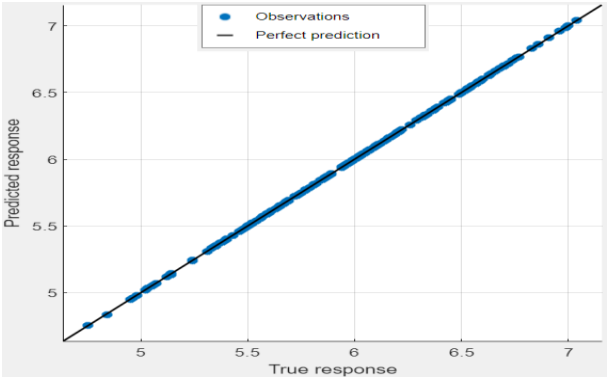

Figure 21: Validation prediction Vs actual value plot of Gaussian Process Regression

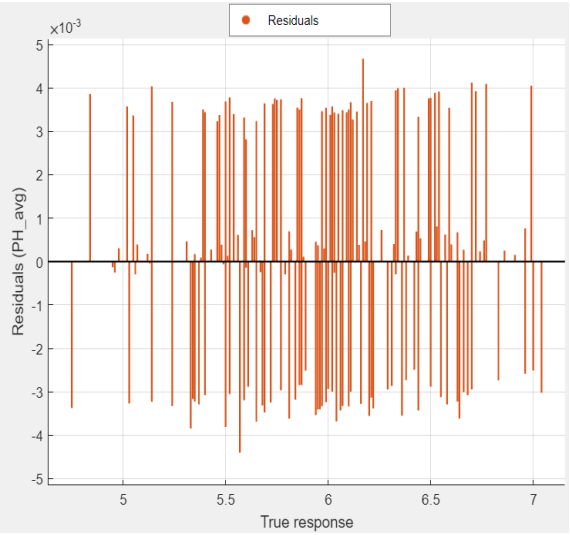

Figure 22: Validation residual plot of Gaussian Prosioncess Regression

Model : Gaussian Process Regression
Status: Trained

Training Results

RMSE (Validation)	0.0027304
R-Squared (Validation)	1.00
MSE (Validation)	7.4552e-06
MAE (Validation)	0.002303
Prediction speed	~13000 obs/sec
Training time	7.1909 sec

Figure 23: Model details of Gaussian Process Regression

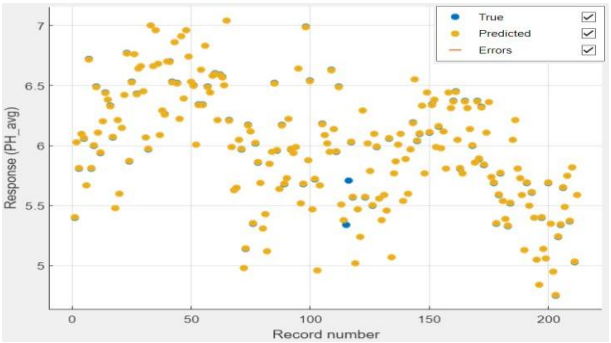

Figure 24: Response plot of Linear Regression

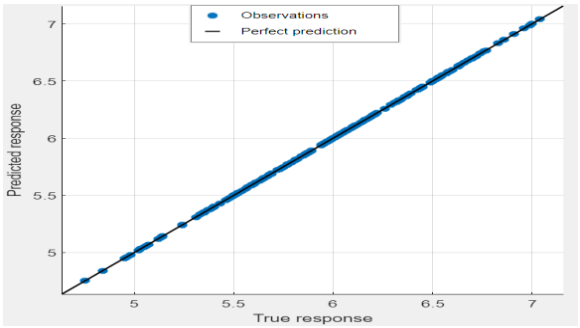

Figure 25: Validation prediction Vs actual value plot of Linear Regression

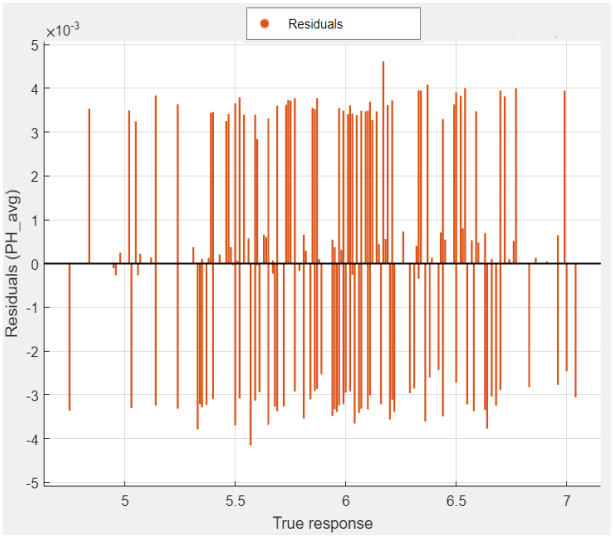

Figure 26: Validation residual plot of Linear Regression

Model : Linear Regression
Status: Trained

Training Results

RMSE (Validation)	0.002718
R-Squared (Validation)	1.00
MSE (Validation)	7.3876e-06
MAE (Validation)	0.0022881
Prediction speed	~5900 obs/sec
Training time	2.8559 sec

Figure 27: Model details of Linear Regression

In this evaluation, R^2 value has been computed for each algorithm and the algorithm with the highest R^2 value has been selected. For the case of the same R^2

value, the lowest RMSE value has been accepted for the best-fit model. As shown in Table 1, the Linear Regression algorithm gives us the best accuracy of prediction to compare with other algorithms. Accuracy (R^2) comparison of machine learning algorithms is shown in Fig 28.

Table 1: Accuracy (R^2) comparison of algorithms

Algorithms	R^2	RMSE
Support Vector Machine	0.99721	0.02576
Tree	0.8864	0.16431
Neural Network	0.99996	0.003185
Gaussian Process Regression	0.99997	0.0027304
Linear Regression	0.99997	0.002718

7. Conclusion

Here, we have a model that can forecast the pH of urine, which has normal values between pH 4.6 − 8.0. Therefore, we must create a machine learning model that can be more accurate and help forecast the pH level in urine with less expense and uncertainty. It could be the primary technique used to assess the condition of the urinary system. This article focuses on urine pH prediction based on the accuracy rate through R^2 in the study that intended to build a prediction technique for the quantity of average pH of daily pee. The general linear regression model, the support vector regression model, the Gaussian process regression model, the tree model, and the neural network model have all been evaluated, and the results show that all of the model types showed prediction values with errors. In line with this idea, the statistics of the given algorithms are applied to calculate the R^2 value and evaluate the statistics across machine learning methods. When the performance of the five methods is evaluated, it is discovered that the Linear Regression algorithm is chosen due to its high R^2 value (0.99997) and low RMSE value (0.002718). These estimated values suggest the highest accuracy of this algorithm. Model precision was always satisfactory for predicting the average pH value of urine after accounting for the variation of data. Even though the model's predictions were more accurate. These findings suggest that machine learning algorithms have the ability to learn about the likelihood of illnesses. It may be possible to develop this kind of research to diagnose additional diseases. For a more thorough investigation, we may further examine the data's prior history and integrate other machine-learning approaches. This study may also have further uses in the future, including the ability to forecast many diseases such as diabetes, cardiovascular disease, tumors, breast cancer, and other diseases.

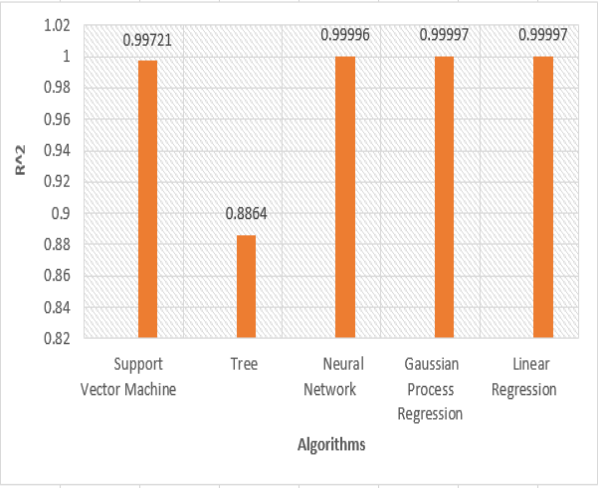

Figure 28: Accuracy (R^2) comparison of machine learning algorithms by bar diagram

REFERENCES

[1] Amiri-Ramsheh, Behnam et al. (2022). "Modeling of wax disappearance temperature (WDT) using soft computing approaches: Tree-based models and hybrid models". In: Journal of Petroleum Science and Engineering 208, p. 109774. ISSN: 0920-4105. DOI: https://doi.org/10.1016/j.petrol. 2021 . 109774. URL: https://www.sciencedirect.com/science/article/pii/ S0920410521013954.

[2] Argyle, J.L. and R.L. Baldwin (1988). "Modeling of Rumen Water Kinetics and

Effects of Rumen pH Changes". In: *Journal of Dairy Science* 71.5, pp. 1178–1188. ISSN: 0022-0302. DOI: https://doi.org/10.3168/jds.S0022-0302(88)79672-1. URL: https://www.sciencedirect.com/science/article/pii/S0022030288796721.

[3] Basheer, I.A and M Hajmeer (2000). "Artificial neural networks: fundamentals, computing, design, and application". In: *Journal of Microbiological Methods* 43.1. Neural Computting in Micrbiology, pp. 3–31. ISSN: 0167-7012. DOI: https://doi.org/10.1016/S0167-7012(00)00201-3. URL: https://www.sciencedirect.com /science/article/pii/ S0167701200002013.

[4] Biswas, Souvik et al. (2022). "Machine learning based urinary pH sensing using polyaniline deposited pa-per device and integration of smart web app interface: Theory to application". In: Biosensors and Bio- electronics 211, p. 114332. ISSN: 0956-5663. DOI: https:// doi. org/ 10 . 1016 / j. bios. 2022 . 114332. URL: https://www.sciencedirect.com/science/article/pii/ S0956566322003724.

[5] Chang, Victor et al. (2022). "An artificial intelligence model for heart disease detection using machine learning algorithms". In: Healthcare Analytics 2, p. 100016. ISSN: 2772-4425. DOI: https://doi.org/10.1016/j.health.2022.100016. URL: https://www.sciencedirect.com/ science/article/pii/S2772442522000016.

[6] Khan, Washim et al. (2022). "Quantitative analysis of primaquine and its metabolites in human urine us- ing liquid chromatography coupled with tandem mass spectrometry". In: Journal of Chromatography B 1213, p. 123517. ISSN: 1570-0232. DOI: https://doi.org/10.1016 / j . jchromb . 2022 . 123517. URL: https://www.sciencedirect.com/science/article/pii/ S1570023222004226.

[7] Li, Meng M., Srijan Sengupta, and Mark D. Hanigan (2019). "Using artificial neural networks to predict pH, ammonia, and volatile fatty acid concentrations in the rumen". In: Journal of Dairy Science 102.10, pp. 8850–8861. ISSN: 0022-0302. DOI: https://doi.org/10.3168/jds.2018-15964. URL: https://www.sciencedirect.com/science/article/pii/S0022030219306617.

[8] Massy, Ziad A. et al. (2022). "Machine Learning-Based Urine Peptidome Analysis to Predict and Under-stand Mechanisms of Progression to Kidney Failure". In: *Kidney International Reports*. ISSN: 2468- 0249. DOI: https://doi.org/10.1016/j.ekir.2022.11.023. URL: https://www.sciencedirect.com/science/article/pii/S2468024922018964.

[9] Schulz, Eric, Maarten Speekenbrink, and Andreas Krause (2018). "A tutorial on Gaussian process regres- sion: Modelling, exploring, and exploiting functions". In: *Journal of Mathematical Psychology* 85, pp. 1–16. ISSN: 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2018.03.001. URL:https://www.sciencedirect. com/ science/ article/pii/ S0022249617302158.

[10] Sheerin, Neil S. (2011). "Urinary tract infection". In: Medicine 39.7. Renal: Part 2 of 3, pp. 384–389. ISSN: 1357-3039. DOI: https://doi.org/10.1016/j.mpmed.2011.04.003. URL: https: //www.sciencedirect.com/ science/article/pii/S1357303911000946.

[11] Vanfretti, Luigi and Narasimham Arava (Dec. 2020). "Decision tree-based classification of multiple oper- ating conditions for power system voltage stability assessment". In: International Journal of Electrical Power Energy Systems 123, p. 106251. DOI: 10.1016/j.ijepes.2020.106251.

[12] Yang, Hang and Simon Fong (2013). "Incremental optimization mechanism for constructing a decision tree in data stream mining". In: Mathematical problems in engineering 2013.

[13] Zhang, Yi et al. (2023). "Tree-based machine learning model for visualizing complex relationships between biochar properties and anaerobic digestion". In: Bioresource Technology 374, p. 128746. ISSN: 0960- 8524. DOI: https://doi.org/10 .1016/j.biortech.2023.128746. URL: https:

//www.sciencedirect. com/science/article/ pii/S0960852423001724.