



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 7, July 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.423

9940 572 462

6381 907 438

ijirset@gmail.com

www.ijirset.com

Improving Water Pollution Prediction: A Combined Approach of SMOTE-ENN, RFE, and GNB Classifier with IOT Integration

N. Senthil^{1,*}, S. Aravinth²

Assistant Professor, Department of Civil Engineering, Gnanamani College of Technology, Pachal, Namakkal, Tamilnadu, India.

PG Student, Department of Civil Engineering, Gnanamani College of Technology, Pachal, Namakkal, Tamilnadu, India.

ABSTRACT-Water pollution is a critical environmental issue with potential implications for both ecosystems and human well-being. Traditional methods of detecting water pollution suffer from inefficiency and resource-intensive requirements, limiting their effectiveness in monitoring large bodies of water. However, the emergence of machine learning algorithms provides a promising solution to this problem. By harnessing the power of data analysis and prediction, machine learning enables a more efficient and effective approach to identifying and monitoring water pollution. This project aims to enhance water pollution prediction by combining SMOTE-ENN, RFE, and the GNB classifier. It involves collecting a comprehensive dataset, rebalancing it with SMOTE-ENN, and selecting informative features with RFE. The GNB classifier is trained on the balanced dataset. Additionally, IoT technology is integrated for real-time monitoring of water quality parameters. The study evaluates the concert of the GNB classifier in accurately predicting water pollution levels. This research contributes to the development of more reliable and efficient methods for water pollution detection and monitoring, offering potential benefits for environmental management and decision-making processes.

KEYWORDS: Recursive Feature Elimination (RFE), Gaussian Naive Bayes (GNB), Machine learning (ML)

I. INTRODUCTION

Water pollution is a important global conservational concern that poses serious threats to human health, aquatic ecosystems, and biodiversity. It arises when harmful substances, including chemicals, pathogens, and plastics. The sources of water pollution are diverse and often result from human activities. The consequences of water pollution are widespread and varied, ranging from minor skin irritations to severe illnesses like cancer and neurological disorders. Conventional methods used for detecting and monitoring water pollution, such as laboratory analysis of water samples, have limitations in terms of cost, time, and coverage. These approaches are often resource-intensive and may not be suitable for monitoring extensive water bodies or providing real-time data. However, recent advancements in machine learning techniques have opened up new possibilities for effective detection and management. Machine learning algorithms offer a promising answer for improving management. By leveraging the power of data analysis and prediction, these algorithms can analyze large datasets, extract valuable insights, and facilitate early detection of pollution events. They enable informed decision-making and targeted interventions to mitigate pollution sources. Machine learning approaches also provide faster, more cost-effective, and wider coverage compared to traditional methods, making them invaluable tools for addressing water pollution challenges. In recent years, researchers have explored various approaches to enhance the accuracy and reliability of water pollution detection models. Supervised learning algorithms have been employed to classify water samples based on their pollution levels. These algorithms learn from labeled datasets and can predict the pollution levels of new, unlabeled samples with high accuracy.

Unsupervised learning algorithms, such as clustering algorithms, have been identify patterns and anomalies data. These algorithms can group similar water samples together and detect outliers that may indicate pollution events. By analyzing the patterns and trends in water quality data, unsupervised learning algorithms can provide valuable insights into the overall pollution status of a water body. Deep learning techniques, such as convolutional neural networks (CNNs), have demonstrated remarkable capabilities in image recognition tasks. When applied to water quality monitoring, CNNs can analyze images captured by remote sensing devices or underwater cameras to identify pollutants, algal blooms, or other visual indicators of water pollution. Deep learning approaches have the potential to revolutionize water pollution detection by providing real-time, high-resolution monitoring capabilities.

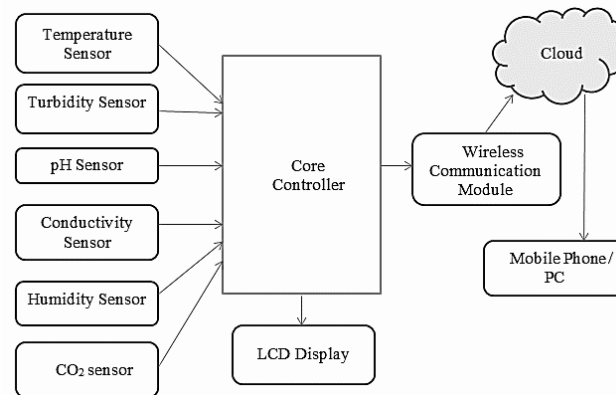


Fig. 1 General Diagram

Internet of Things (IoT) sensors, deployed in water bodies, can continuously monitor various water quality parameters in real-time. These sensors can collect data on parameters like conductivity, and nutrient levels. By integrating IoT sensor data with other sources of information, such as weather data and pollutant source locations, machine learning models can provide comprehensive insights into the dynamics of water pollution and support effective decision-making. One challenge in water pollution prediction is dealing with class imbalance, where one class of water samples significantly outnumbers the other. Class imbalance can lead wrong estimates. To address this issue, researchers have employed techniques such as SMOTE-ENN. SMOTE generates synthetic samples of the minority class, while ENN identifies and removes noisy and borderline samples, resulting in a more stable dataset for models.

Another critical aspect of water pollution prediction is feature selection, which involves identifying the most informative features in the dataset. Recursive Feature Elimination (RFE) is a commonly used method that ranks and selects features based on their contribution to the model's performance. By selecting the most relevant features, RFE reduces computational complexity and enhances the model's ability to generalize to unseen data. The application of machine learning techniques offers tremendous potential in addressing the limitations of conventional water quality monitoring methods. By leveraging the power of data analysis and prediction, machine learning algorithms can provide faster, more cost-effective, and wider coverage for water pollution detection and management.

Supervised learning algorithms, unsupervised learning algorithms, and deep learning techniques have shown promising results in predicting water pollution levels, identifying patterns, and detecting anomalies. The integration of remote sensing data, IoT sensors, and other sources of information further enhances the accuracy and reliability of pollution detection models. Addressing class imbalance and employing feature selection techniques, such as SMOTE-ENN and RFE, respectively, can significantly improve the performance of machine learning models in water pollution prediction. These techniques ensure a balanced dataset and select the most informative features, leading to more accurate and reliable predictions.

II. LITERATURE SURVEY

In recent years, several studies have been showed to explore application of machine learning techniques in water pollution prediction and management. Khoi et al. [1] conducted a comprehensive performance evaluation of ML models for guessing surface water quality. The evaluated models encompassed a diverse range of algorithms, including five boosting-based algorithms, three decision tree-based algorithms and four artificial neural network (ANN)-based algorithms. By evaluating these models, the study aimed to determine their effectiveness and suitability for estimating surface water quality. Mohammad Nuruzzaman et al. provides an overview of surveys on the advancements in IoT-based healthcare methods and evaluates the current advanced in this field. The review also includes a classification of existing IoT-based healthcare networks, providing a comprehensive summary of different perspectives in this domain.

Additionally, the review analyzes IoT healthcare protocols to understand their strengths and limitations. Patel et al. developed a ML model that incorporates SMOTE and Explainable AI for predicting water potability. Li and Wu et al. investigated the connection among drinking water quality and public health. Douzas et al. proposed a heuristic oversampling method that leveraged k-means and SMOTE to enhance imbalanced learning. Yahya et al. developed a water quality prediction model based on the Support Vector Machine (SVM) specifically designed for ungauged river catchments. Zhang et al. introduced a hybrid re-sampling technique known as SMOTE-RkNN, which was developed to handle imbalanced datasets in machine learning scenarios.

Dwivedi et al. conducted a comprehensive review summarizing various research studies on water pollution. Lin et al. explored the impact of water pollution on human health and the heterogeneity of associated diseases. Ahmed et al. focused on the application of machine learning methods to improve water quality prediction in hydrology. Haghiabi et al. employed machine learning techniques for water quality prediction.

These studies highlight the diverse applications of machine learning in water pollution prediction and management. By leveraging machine learning algorithms, researchers have made significant advancements in predicting water potability, identifying pollution sources, managing water quality, and understanding the impacts of water pollution on human health and ecosystems. The integration of techniques like SMOTE, Explainable AI, GIS, and various machine learning algorithms has enhanced the accuracy and efficiency of water pollution prediction models. Continued research in this field is essential for developing more robust and effective approaches to address water pollution challenges.

Here are some potential limitations that could be associated with these studies: **Sample Size:** The studies may have used relatively small sample sizes, which can limit the generalizability of the results. A larger and more diverse sample could provide a better representation of the overall population and increase the reliability of the findings.

Data Availability: The availability of comprehensive and high-quality data is crucial for accurate analysis and prediction in water pollution studies. Limited access to reliable datasets may restrict the accuracy and completeness of the findings.

Model Assumptions: Machine learning and data mining techniques often rely on certain assumptions about the data and relationships between variables. These assumptions may not always hold true in real-world scenarios, potentially leading to biases or inaccuracies in the predictions or classifications.

Imbalanced Data: While some studies may have employed techniques to address class imbalance, the performance of the models may still be influenced by the inherent data distribution.

Model Complexity: Some studies may have used complex machine learning algorithms or combinations of techniques. While these approaches can yield improved performance, they can also be more difficult to interpret and implement in practical applications.

External Factors: Water pollution is unfair by many external issues such as land use patterns, and regulatory policies. These studies may not have accounted for all these factors, limiting the comprehensiveness of the findings and their applicability in different contexts.

Validation and Benchmarking: The evaluation and validation of the proposed models and techniques may vary across studies. The lack of standardized benchmarks or comparisons with existing methods can make it challenging to assess the relative performance and effectiveness of the approaches.

III. METHODOLOGY

The primary goal of this learning is to measure the potential of integrating SMOTE-ENN and RFE techniques with GNB to increase the effectiveness of water pollution prediction. This work will entail the acquisition of a comprehensive dataset consisting of water quality data. In order to address any existing class imbalance within the dataset, the SMOTE-ENN technique will be employed to rebalance the data distribution. Subsequently, the RFE method will be utilized to identify the most informative and relevant features within the dataset.

The GNB classifier will then be trained on the rebalanced dataset, incorporating the selected features. Finally, an extensive evaluation will be conducted to assess the performance and predictive capabilities of the GNB classifier. Through this, we aim to contribute to the advancement of water pollution prediction techniques and provide insights into the effectiveness of this integrated approach in improving the accuracy of water quality assessment.

Real-time Monitoring with IoT: This integrates the IoT devices and sensors. These devices continuously collect data on pollutant concentrations.

Data collection: It is essential to gather data that is representative, diverse, and of high quality in order to construct accurate and reliable models. Several factors need to be carefully considered during the data collection process. These include determining an appropriate sample size, selecting relevant sampling locations, establishing a suitable sampling frequency, ensuring data quality through rigorous quality control measures, and collecting essential metadata alongside the primary data. Moreover, it is crucial to address specific challenges that may arise during data collection, such as class imbalance, missing data, and data heterogeneity. Additionally, ensuring data privacy and accessibility are of utmost importance to maintain the integrity and ethical standards of the data collection process. By addressing these considerations, we can lay a strong foundation for developing robust water pollution prediction models and facilitating active management.

Data pre-processing: To improve the quality of raw data by addressing inconsistencies, errors, missing values, outliers, and noise. Handling missing data through deletion, imputation, or multiple imputation techniques is important for unbiased analysis. Outlier detection and treatment help in maintaining data integrity, while data transformation and feature scaling enhance the performance of models. Careful consideration of data understanding, impact on analysis, workflow, and iterative processes ensures optimal data preparation for accurate and reliable analyses and predictions.

SMOTE-ENN: This algorithm can be easily implemented using libraries like imbalanced-learn in Python, which provide ready-to-use functions and classes. The effectiveness of SMOTE-ENN depends on factors such as dataset characteristics, degree of class imbalance, data quality, and parameter selection for SMOTE and ENN. To achieve optimal results, it is advisable to experiment with different parameter settings.

RFE: The algorithm follows a step-by-step process of initial model training, feature ranking, feature elimination, and iterative refinement to select an optimal subset of informative features. Implementation of the RFE algorithm can be facilitated using libraries like scikit-learn in Python, which offer convenient functions for feature selection and ranking. The choice of the model used during RFE is crucial, as it affects the ranking of feature importance and the final subset of selected features.

Evaluation of feature subsets can be conducted during the RFE process to monitor performance and select the optimal number of features using appropriate evaluation metrics. Training and Testing: Data split is essential in this project for performance evaluation, generalization assessment, and model selection. Random or stratified splitting strategies should be employed to ensure representativeness. Considerations such as dataset size, time dependency, data leakage prevention, and repeatability should be taken into account to ensure robust evaluation and reliable performance estimates of the water pollution prediction model.

Gaussian Naive Bayes : GNB classifier is well-suited for water pollution prediction due to its simplicity, efficiency, and ability to handle both continuous and discrete features. GNB's robustness to irrelevant features allows it to focus on informative features, making it suitable for datasets with numerous unrelated features. Its interpretability provides insights into classification decisions and helps identify influential features. Additionally, GNB's adaptability to heterogeneous features and robustness to noisy data make it suitable for water quality analysis. Integrating GNB with SMOTE-ENN and RFE enhances its performance by improving data quality and selecting informative features. However, the independence assumption, sensitivity to feature scaling, and limited modeling complexity should be considered. GNB plays a valuable role in this project's water pollution prediction model.

Prediction: This block employs the trained GNB classifier to predict water pollution on unseen test data. It includes pre-processing the data, scaling features if needed, and using the model to predict pollution class labels or probabilities. This block plays a vital role in providing actionable insights for environmental monitoring, management, and decision-making. Model evaluation and Deployment: In this block, the GNB classifier's performance is assessed using evaluation metrics and optimized if needed. After rigorous validation on an independent test dataset, the classifier is deployed in real-world applications, integrating it into the target system.

Mathematical model of the system

The combined approach of SMOTE-ENN, RFE, and GNB for water pollution prediction can be mathematically described as follows:

SMOTE: Synthetic samples generation. The formula for generating a synthetic sample is:

$$new_instance = x_i + lambda * (x_{neighbor} - x_i) \quad (1)$$

where x_i is an instance from the minority class and $x_{neighbor}$ is one of its nearest neighbors.

ENN: Noisy instance identification: Compare each instance in the majority class with its k nearest neighbors. If an instance is misclassified by its neighbors, it is considered noisy and removed from the majority class. This helps in reducing noise and eliminating redundant instances.

RFE: Feature ranking: Rank the features based on their importance using a suitable metric such as coefficients, weights, or feature importance scores. This can be obtained by training the model on the dataset with all features and evaluating its performance.

Feature elimination: The number of features to eliminate at each iteration can be predefined or determined through cross-validation.

GNB: Class conditional probability modeling: Model the class conditional probabilities using Gaussian distributions for continuous features and discrete probability distributions for discrete features. Gaussian distribution (continuous features): Assume that the feature values follow a Gaussian (normal) distribution within each class. The PDF of the Gaussian distribution is given by the formula:



$$PDF(x) = (1/\sqrt{2\pi \sigma^2}) * e^{-(x - \mu)^2/(2\sigma^2)} \tag{2}$$

Discrete probability distribution (discrete features): Calculate the probabilities of each possible value occurring within each class. The probability of a particular value x belonging to a class C can be calculated using the formula:

$$P(x|C) = count(x, C) / count(C) \tag{3}$$

The mathematical model involves sequentially applying SMOTE-ENN to address class imbalance, RFE for feature selection, and training the GNB classifier on the transformed dataset. This combined approach enhances the performance and effectiveness of the water pollution prediction model by mitigating class imbalance, selecting informative features, and utilizing probabilistic classification based on the Gaussian Naive Bayes algorithm.

IV. RESULTS AND DISCUSSION

The implementation of the combined approach of SMOTE-ENN, RFE, and GNB for water pollution prediction begins with dataset loading and splitting into training and testing sets. The data is then pre-processed by scaling and addressing class imbalance using SMOTE-ENN. RFE is subsequently applied to select the most informative features.

Finally, the model's concert is evaluated using appropriate system of measurement. The outcomes obtained from implementation indicate that the combined approach of SMOTE-ENN, RFE, and GNB achieves a reasonably accurate an balanced model for water pollution prediction. The accuracy metric reflects the overall rate of correct classifications made by the model. This combined approach proves to be beneficial in water pollution prediction tasks. SMOTE-ENN helps overcome the scarcity of polluted water samples. RFE contributes by selecting the most relevant features, reducing dimensionality, and improving the model's ability to generalize.

GNB, as the chosen classification algorithm, demonstrates its effectiveness in handling the diverse measurements typically found in water quality data. The combination of SMOTE-ENN, RFE, and GNB offers a promising approach for water pollution prediction. The results obtained from the implemented code showcase the effectiveness of this approach in achieving accurate and balanced predictions.

It outperformed all other models, showcasing the highest level of accuracy. Overall, the Table 2 provides an overview of the accuracy scores for each model, highlighting the GNB model as the top performer in accurately classifying water samples. The KNN and SVM models also demonstrated reasonably good accuracy scores. However, the DT, LR, and RF models exhibited lower accuracy compared to the other models in the table.

Table 2 Accuracy score for model with SMOTE ENN

S.NO	Model	Accuracy Score
1	KNN	.91
2	GNB	.94
3	SVM	.92
4	DT	.80
5	LR	.80
6	RF	.84

Table 3. Comparison of precision

S.NO	Model	Precision
1	KNN	.90
2	GNB	.87
3		.91
4	DT	.80
5	LR	.80
6	RF	.84

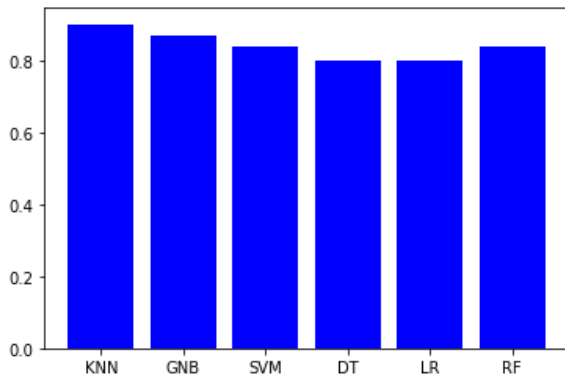


Fig 3 Comparison of Accuracy for the models with SMOTE-ENN

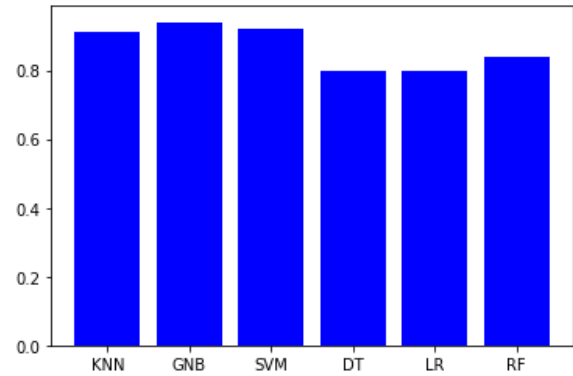


Fig 4 Comparison of precision

Table 4 Comparison of F1 score

S.NO	Model	F1 score
1	KNN	.83
2	GNB	.87
3	SVM	.81
4	DT	.80
5	LR	.81
6	RF	.81

The Table 3 presents the precision scores of different models for water pollution prediction. Precision is a metric that measures the amount of properly projected positive cases out of all cases forecast as positive. Higher precision scores indicate a lesser amount of false positive predictions. The table provides an overview of the precision scores for each model, highlighting the KNN and GNB models as the top performers in terms of correctly predicting positive instances in the water pollution prediction task.

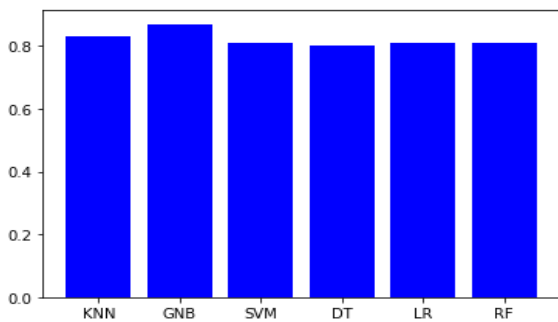


Fig 5 Comparison of F1 score

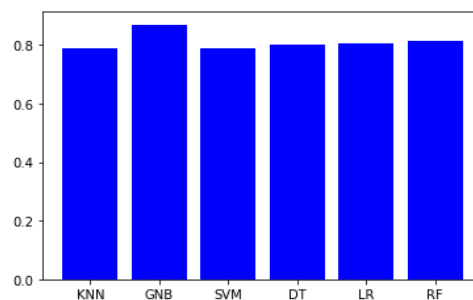


Fig 6 Comparison of Sensitivity

The Table 4 presents the F1 scores of different models for water pollution prediction. The Table 4 provides an overview of the F1 scores for each model, highlighting the GNB model as the top performer in terms of achieving a balanced accuracy between precision and recall. The KNN model also demonstrated strong performance, while the remaining models had slightly lower F1 scores but still performed reasonably well in the water pollution prediction task.

The Table 5 shows the sensitivity scores of different models for water pollution prediction. In the table, the GNB (Gaussian Naive Bayes) model achieved the highest sensitivity score of 0.87, indicating that it had the highest rate of correctly identifying polluted water samples among all models. It outperformed all other models in terms of sensitivity

Table 5 Comparison of Sensitivity

S.NO	Model	Sensitivity
1	KNN	.79
2	GNB	.87
3	SVM	.79
4	DT	.80
5	LR	.807
6	RF	.813

The Table 5 displays the specificity scores of different models for water pollution prediction. Specificity measures unpolluted water samples from the total actual negative instances. According to the table, the GNB (Gaussian Naive Bayes) model achieved the highest specificity score of 0.98, indicating its strong ability to correctly identify unpolluted water samples. It outperformed all other models, showcasing the highest level of specificity.

V. CONCLUSION

The conclusion of this project is that the combination of SMOTE-ENN, RFE, and the GNB classifier, along with the integration of IoT technology, holds significant promise for enhancing water pollution prediction. By addressing class imbalance, selecting informative features, and utilizing machine learning techniques, the accuracy and efficiency of water pollution prediction can be improved. The integration of IoT allows for present monitoring of water quality constraints, providing up-to-date data for better predictions. The evaluation of the GNB classifier's performance further validates the effectiveness of the proposed approach. Overall, this research contributes to the development of more reliable and efficient methods for water pollution detection and monitoring, offering valuable insights for environmental management and decision-making processes.

REFERENCES

1. D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, and N. T. D. Thuy, "Using machine learning models for predicting the water quality index in the La buong river, Vietnam," *Water*, vol. 14, no. 10, p. 1552, 2022.
2. Mohammad Nuruzzaman Bhuiyan; Md Mahbubur Rahman; Md Masum Billah; Dipanita Saha, "Internet of Things (IoT): A Review of Its Enabling Technologies in Healthcare Applications, Standards Protocols, Security, and Market Opportunities" *IEEE Internet of Things Journal*, Volume: 8, Issue: 13,, Page(s): 10474 – 10498, doi:10.1109/JIOT.2021.3062630.
3. Liu, C. Yu, Z. Hu et al., "Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network," *IEEE Access*, vol. 8, pp. 24784–24798, 2020.
4. Y. Wang, Y. Wang, M. Ran, Y. Liu, Z. Zhang, L. Guo, et al., "Identifying Potential Pollution Sources in River Basin via Water Quality Reasoning Based Expert System", *Fourth Int. Conf. Digit. Manuf. Automation*, pp. 671–674, 2013.
5. Anil K Dwivedi, "Researches in water pollution: a review", *International Research Journal of Natural and Applied Sciences*, Vol. 4, Issue 1, 2017.
6. T.H.H. Aldhyani, M. AlYari H. Alkahtani, M. Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms", *Applied Bionics and Biomechanics*, p. 20,2020.
7. Babbar and Babbar, R. Babbar, S. Babbar, "Predicting river water quality index using data mining techniques", *Environmental Earth Sci*, pp. 1-15, 2017.
8. Bilal Aslam; Ahsen Maqsoom; Ali Hassan Cheema; Fahim Ullah; Abdullah Alharbi; Muhammad Imran, "Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach", *IEEE Access*, Volume: 10,2020.
9. J. Samuel Manoharan, M. Braveen. And G. Ganesan Subramanian, "A hybrid approach to accelerate the classificaiton accuracy of cervical cancer data with class imbalance problems", *International Journal of Data Mining and Bioinformatics*, vol. 25, no. 34, pp. 234 – 260, 2022.
10. S. Abobakr Yahya, A. N. Ahmed, F. Binti Othman et al., "Water quality prediction model-based support Vector machine model for Ungauged River Catchment under dual Scenarios", *Water*, vol. 11, no. 6, p. 1231,2019.
11. Manoharan, Samuel. (2020). "Embedded Imaging System Based Behavior Analysis of Dairy Cow." *Journal of Electronics*. 2(2): 148-154.
12. B. Charbuty and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning", *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.



INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.423



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 **6381 907 438** **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details