International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-June-2026, ISSN (Online): 2395-566X

An Intelligent Machine Learning Framework for Water Potability Prediction

[1]P.Sandhya Krishna, [2]Ala Nandini, [3]Pavuluri Sri Lekha, [4]Gumma Aparna, [5]Patchava Pujitha
[1]Associate Professor, Department of IT, Vignan's Nirula Institute of Technology and Science, Guntur.
[2,3,4,5] B. Tech, Department of IT, Vignan's Nirula Institute of Technology and Science, Guntur.

Abstract- Clean and safe drinking water is a crucial factor in the health of the population, but even now, delivery of contaminated drinking water remains one of the world issues. Water potability: a ML approach The use of ML models in Water Quality Assessment is a recent phenomenon in the past years, as it is now a highly promising tool that predicts the water potability in an efficient (more efficient than traditional) manner. The paper presents a smart machine learning system to anticipate the potability of water that is determined by undertaking a thorough review of diverse physico-chemical characteristics of water such as PH, Hardness, Solids, Chloramines, Sulfate and organic contaminants. State of art preprocessing methods are also applied to address missing values, outliers and feature stratification which enhance the quality and the strength of the data. There are several supervised learning processes, which include Random Forest, SVM, Gradient Boosting and ANN to determine the best predictive accuracy algorithm. The general performance is also justified with the premises of accuracy, precision, recall, F1-score and ROC-AUC performance parameters and demonstrates that the suggested framework implementation is reliable and efficient on actual water quality monitoring scenarios. Also, the work places emphasis on the effects of the feature selection and the hyperparmeter tuning on the enhancement of the prediction performance. Ensemble approach and cross-validation methods cut down on the framework and expand the generalization potential with different datasets.

Keywords— Water Potability, Machine Learning, Classification, Prediction Model, Data Analysis, Supervised Learning, Water Quality

I. INTRODUCTION

Blood of every living thing on earth. The life blood they say is water and without water all living organisms can not grow, survive and be able to operate on a day to day basis [1]. Water as a resource. Water is not something that human beings need. Safe and clean drinking water is critical to the development of any given society [2] [3]. Water resource is a vast resource that is depleting at a very high rate because of human activities [4]. The water quality has been undermined in the past several decades by the environmental harm and the industrial- and agricultural-based pollution [5]. Rivers, lakes and underground waters have only heavy metals, nitrates, sulfates and other organic compounds as the pollutants [6]. The second does not only change the physicochemical nature of water; it even seems not to be drinkable [7]. The World Health Organization (WHO) believes that millions of humans across the globe are being infected annually by waterborne diseases due to polluted drinking water [8] [9].

The conventional method of water quality evaluation is evaluation of water samples in laboratories using either analytical or microbiological techniques [10]. These techniques are precise, however, expensive and time-consuming and cannot be used in widearea or real time monitoring [11]. And that is what new methods must be more powerful, cheaper, as well as faster in processing the big data [12]. The new data-driven technology has transformed the contemporary method of environmental monitoring in the era of Industry 4.0 [13]. Machine learning (ML) is a branch of artificial intelligence (AI), which offers potent mechanisms to infer complicated patterns of data sets and make predictive decisions [14] [15]. Actually, the ML algorithms can be trained on the input-output relationships and used in water potability prediction using measurable variables [16].

An effective machine learning model can be used in testing the quality of the water by training on previous data and learns few patterns to assist in identifying whether the water is drinkable or not [17]. They can analyze more than one physicochemical parameter at a time and even predict whether the sample of water is drinkable [18] [19]. Of course, besides the decrease in the number of human experiments needed, such a direction promotes more powerful decisions within a certain speed [20]. ML models receive receiving end input flow to operate on: there is no restriction of how much data they can slur, gathered by sensors, lab analyses and databases (DeGrasse, Johnson & Romeo 2019 de Rosemond et al [6]. pH, hardness, total dissolved solids, chloramines, sulphate, conductivity, organic carbon(TIC), trihalomethanes(THMs) and turbidity are all

parameters that are critical in determining potability [21]. You see that it is because we can teach algorithms this combination of parameters, which will allow auto-int

A large number of machine learning algorithms have been applied previously on environmental problems prediction [23]. Such algorithms as the Logistic Regression, Decision Trees, SVM and KNN and ensemble algorithms like the Random Forests and XGBoost has been very effective in the classification problems [24]. Due to their flexibility and precision, they could also be extended to the water drinkability judgment conveniently [25]. The tree-based ensemble methods like the random forest and XGBoost, among others, can be of great advantage in predicting the accuracy of failure since they have the capability to model high-order interactions in data, minimize overfitting etc [26]. These are algorithms that integrate several weak learners to form a powerful predictor which is accompanied with greater accuracy and improved generalization [27]. Moreover, an analysis of the importance of different features based on these models might assist in determining the most sensitive input parameters of water quality [28].

Pre-processing of data You cannot just really train any machine learning model really [29]. Data regarding water quality is often imperfect because there is always incompleteness, inconsistency and noise in the measure of water quality which will undermine the performance of models [30] [31]. As such, data cleaning, missing data processing, feature scaling and dimensionality reduction have to be implemented so as to derive relevant predictions [32]. The selection of features renders the model more user friendly and we comprehend with ease what is significant in arriving at a decision [33]. This is to make sure that the model is concentrating on features that have the greatest contribution towards water drinkability [34]. This can be done using methods such as correlation analysis and recursive feature elimination [35].

After the data has been preprocessed we come up with machine learning algorithms that categorize water samples into drinkable/non-drinkable [36]. The models are evaluated on criteria of accuracy, precision, recall, the F1 -score and ROC-AUC curve [37]. These measures assist in the objective comparison of the models, and then choosing the optimal model of a dataset [38]. They can be even more powerful with smart machine learning models and IoT system added [39]. The data of internet of things Sensors can be gathered in real-time to measure water quality at various points of the locations and send it to a central processing system [40]. The ML model will be capable of absorbing the incoming data and making an immediate decision of whether the water source is drinkable or not [41].

Such a system can act as a positive alert to the authorities and people in case of disruption before it becomes a social health concern [42]. This preventive health surveillance enables the prevention of incidences of water-borne diseases before they strike and thus protects the health of human beings. The smart ML also can result in water sustainability by expression rather than practice. Instantiate information to the policymakers, environmental scientists and public health departments assist them to track the trends of water quality, allocate resources most effectively as well as plan water prevention activities on the basis of the information.

These intelligent algorithms align with the United Nations sustainable development goal 6 (SDG-6) Clean Water and Sanitation that is concerned with the availability and sustainable management of water and sanitation to every population [14-15]. The predictive approach (with the use of ML) can possibly resolve this issue, by tracking the changes in water and decision-making in real-time. The purpose of the paper is to develop and propose a model (smart) based on machine learning methods that may be applied to predict the potability of water with a high level of accuracy. The notion is engaged with integrating the different ML algorithms and it integrates data treatment strategies as well as certain optimization strategies in a manner that the higher the rate of accuracy, the ability to withstand changes in parameters etc. is attained automatically [16-20]. It also indicates the importance of every water parameter and the influence that they have on the outcome of classification of the process. The findings of

II. LITERATURE REVIEW

V. Singh et al. [1-3] the traditional methods of water quality monitoring, such as the laboratory techniques might not only require a long period, but also needs lots of money and labor, which is not the case with in situ continuous monitor. To overcome these drawbacks, machine learning (ML) models have widely been utilized to preempt the potability of water. They created a dataset where they transformed features and used the QDA models to categorize the water samples based on nine primary physicochemical parameters using the PyCaret framework. It was demonstrated that the ML models were capable of generating accuracy-wide estimates of safe drinking water and non-drinking water samples with minimum level of human effort and minimum surveillance cost. These methods highlight the opportunities of the AI methods in managing water quality sustainably and preventing pollution in advance.

V. Sreekumar et al. [4-5] indicated that safe drinking water is not simply easily available because it is contaminated and polluted resulting into severe health issues to the population. In their study, they used machine learning algorithms to estimate potability of water on the foundation of a large scale dataset of physico chemical properties. They used it in multi-algorithm structure such as the Random Forest, K-Nearest neighbours (KNN), Sup-support Vector Machines (SVM), Decision Tree and Gradient Boosting to model the complex devices in data. Performance improvement of the model could have been done by feature engineering and hyper- parameter tuning and compared to select the best predictor. This model was good in terms of accuracy, sensitivity and specificity that was later validated in another validation cohort. This paper raises a caution over the prospect of the AI/ML type approaches to support effective, consistent, and scalable water quality measurements.

R. Chafloque et al. [6-8] examined the viability of applying ANN in predictive modelling to determine the potability of water since it experiences high demand of automatic establishment of quality of water. The Kaggle dataset has a mix of various physiochemical parameters and MinMax scaling has been done. In Python using the Keras package, the scientists developed a seven-layer of dense neural network to differentiate between water that is drinkable and otherwise. The model that was obtained had a precision of 70 which revealed the usefulness of the neural network in predicting the quality of water. It was also shown that further improvement was achievable through the application of other network architectures or other classification machine learning models, which demonstrated the potential of AI in sustainable water resources management.

N. D. S. S.Kiran Relangi, et al. [9-11] reviewed the use of both the Machine learning and Deep learning method to predict water quality on various databases. This paper had identified challenges that were caused by the differences in water quality, environmental conditions and local baselines. The important determinants were believed to be represented by the significant variables and the enhancement of the model was done through the use of genetically modified feature selection algorithm. The models that we compared include LR, RF, AdaBoost, XGBoost and Deep Neural Networks. We have also discovered that emphasis on model accuracy is quite dependent on the set of features, whereas certain datasets performed better with Logistic Regression or random forests, whereas other do better with DNNs. This paper gives a greater focus on the significance of feature and model in predictive water quality monitoring.

P. P. Mattihallican et al. [12-14] introduced the HydroSense 2.0 which used the IoT, cloud computing and machine learning to solve the slowness of the contaminant using the monitoring of water quality. ESP32 MCU and sensor will be used to detect the critical parameters such as pH, temperature, turbidity and TDS. The obtained data will be sent to a cloud based system, which implements water detection by relying on the drinkable or non- drinkable with the emphasis on the Random Forest model. Live camera images and notifications contribute to the faster reaction to water quality problems. Their study proves that ML can be used to monitor the water in a scalable, automated and cost-effective way, which is a solid starting point on the predictive algorithm of water potability models in the future.

M. U. Maheswari et al. [15-16] limitations were encountered by traditional detection means of water quality which are laborious and costly laboratory methods. This analysis utilizes algorithms to process data and determine the safety of a given water source for drinking. Decision Tree and Random Forest models were developed based on the features such as pH, hardness, solids, chloramines, sulfates, conductivity organic carbon, Trihalomethanes,turbidity etc. The predictive accuracy of these results had a good performance and high computational time savings, indicating that the application of ML algorithms to real-time monitoring of water quality can be realized with high efficiency and low costs.

Another deep learning model presented by M. I. Marie et al. [17] is the Melano Hybrid Model, which is designed to enhance the accuracy and efficiency of melanoma detection. It used adaptive feature fusion to augment the speed of detection of YOLOv9 with the pruning boundary localization of Faster R-CNN that can trade off real-time and diagnostic accuracy. Performance on benchmark datasets (ISIC 2019, HAM10000 and ISIC 2020) was validated and fivefold cross-validation was used as a fair comparison of the models. The cascade approach performed better in the mean accuracy and F1-scores as well as in computational form, reduced memory usage, and real-time to make an inference. This study proved that the Melano Hybrid Model may be a valuable and portable AI-related device to screen melanoma in clinical practice.

M. Munara et al. In [18], the authors developed an IoT-based water quality monitoring system in order to record real-time and remote water safety measurement. The instrument was fitted with numerous sensors to measure the presence of significant physicochemical parameters such as pH, turbidity, temperature, dissolved oxygen and TDS. These sensors were read and transmitted back to a receiver base station where they were analyzed. Machine learning models, namely the Random Forest and SVM, were used together with sensing capability

IoT to obtain very strong classification of water in drinkable and non-drinkable categories. The system was then implemented on the water samples of varied origin the scores of which of the multiparameter were excellent indicating low maintenance efforts and quick response time. Also, the degradation of water quality could be checked and immediate notification of the condition saved timely remediation. The final observation of this study was that the integration of IoT and ML technol

S. Naik et al. [19] examined the model of machine-learning applied on the prediction of potable water based on a dataset of 9 features and 3,276 samples. To enhance the performance of the model a number of classification algorithms, which included, XGBoost, Random Forest, Decision Tree, KNN and dimension reduction algorithms, including PCA, ICA, and TSVD were adopted. The problem of biased data was addressed with the help of SMOTE, SMOTETomek and Near Miss and Hyperparameters were optimized with the help of Grid Search. The evaluation metric used to compare models was computation cost in terms of accuracy, precision, recall and F1-score and the best accuracy (99.80) has been achieved by XGBoost model. The interpretation of the results was done using the Explainable AI tools ELI5 and Shapash, which revealed the most significant features to predict water quality. The study demonstrates that ML and XAI approach are effective in delivering correct, interpretable and effective WQIs.

Hassan et al. [20] developed a machine learning model to predict water potability based on 3,276 samples and nine important water quality parameters. The most accurate predictive model was determined by comparing different algorithms used in the study i.e. Random Forest, Logistic Regression, XGBoost, Gaussian Naïve Bayes, K- Nearest Neighbours, Deci- sion Tree, Support Vector Regression and Multi-Layer Perceptron. PCA, ICA and TSVD were used as feature engineering methods to extract data that are more informative and contribute in improving the performance of our model. SMOTE, SMOTETomek and Near Miss were applied to address the problem of imbalanced dataset. The model was evaluated using accuracy, precision, recall and F1-score. Interpretability of the model and feature importance was provided by Explainable AI tools: ELI5 (Peñaloza, 2017), Shapash. XGBoost obtained the best performance, proving the benefit of integrating machine learning, feature engineering and explainable AI on robust and interpretable w

III. PROPOSED MODEL

The model aims at creating a smart type machine learning system that may have the capacity to offer classification of water samples as either drinkable or ut nondrinkable, depending on the nature of the water. This model is founded on a supervised learning framework and it is trained with a series of labeled measurements exhibiting various water quality indicators. Examples Every case is one that explains a sample of water with some aspects associated with the sample such as pH, hardness, solids etc. other than a binary value indicating the potability. The idea is to come up with a predictive model, which can emulate the complex non-linear inter-dependencies between these factors and then, can predict and do so with high precision and certainty, potability.

The model is developed on a step-by-step flow including data pre-processing, feature selection, model training and validation and prediction. There is also the treatment of missing values, data is normalized and unwanted/redundant features removed in the preprocessing. The use of statistical correlation and information gain based feature selection are used to select notable features that lead to potable preference. Such important characteristics are then fed to model building and tuning machine learning (ML) algorithms. The most popular ensemble techniques such as Random Forest and XGBoost can address such skewed distributions and have good results in the literature.

After training the model, the evaluation is performed in terms of standard classification measures such as accuracy, precision, recall, F1-score and ROC-AUC. Hyper-parameters are tuned and resorted to improve performance through the use of gridsearchCV or random search CV. The interpretations of the model are probabilities that indicate whether the water sample under test is drinkable or not. This score can also be used as a warning mechanism where the quality of water becomes dangerous to the human health in real-time monitoring mechanisms.

Not only does the proposed model serve as an anticipator but it also implements interpretability modules to be capable of making decisions transparent. A discussion of the feature importance, provided by the ensemble models such as Random Forest and XGBoost, shows what kinds of physicochemical parameters make the greatest contribution to water potability. Such realization can be not only helpful in the explanation of the relationship between water quality indicators but also useful in the definition of priorities of the parameters which are under management in the practice. Also, a local explanation can be provided using SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) as opposed to a single prediction and as to why it is not safe to

drink a particular sample of water. Having the most predictive accuracy to be interpreted, it offers greater operational confidence and trust to an extent that it may be applied to urban or agriculture water management process

The model is also meant to be efficient in the event of implementing on large continuous data streams collected by IOT based water monitoring devices and therefore is able to perform such tasks in real-time. It is programmed to learn progressively and update remotely and this makes the framework to adjust to the dynamics of water quality trends without retraining itself completely. The cloud based implementation is also advantageous in the fact that a central data aggregation, processing, and storage is applied and edge devices perform preliminary analytics closer to sensors in case of latency sensitive and bandwidth limited applications. The system can also generate automatic alerts and actionable reports to the end users thereby making it easy to take timely measures to check incidences of water borne diseases. Overall, the innovative intelligent machine learning model proposes to produce reliable predictions of the water potability and design a comprehensive solution that is adaptive and evolutionary to t

Algorithm

Step 1: Data Collection

- Obtain water-quality parameters (i.e., pH and hardness, solids, chloramines¬, sulfate, and organics [including herbicides/pesticides]) for each dataset.
- Ensure to have the target label as appears below: Potable (1), Not potable (0) in your dataset.

Step 2: Data Preprocessing

- Missing values: The imputation of missing using mean, median or KNN.
- Remove outliers: verify with Z-score or IQR.
- Feature Scaling: Normalizing/Standardization of features to improve ML Model performance.
- Encoding: Convert the categorical variable into numerical format.

Step 3: Exploratory Data Analysis (EDA)

- Analyze the distribution of study datasets, correlations and importance of features
- Visualize the relationship between features and the target variable through plots (histograms, boxplots, heatmaps etc).

Step 4: Feature Selection

- Choose the right features (e.g., by correlating, using RFE or tree-importance based feature selections).

Step 5: Train-Test Split

- Do not use the whole dataset K to train and test (e.g., 80%: training, 20%: testing); note model performance.

Step 6: Model Selection

Combine multiple models of supervised ML:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Gradient boosting/ Xgboost
- ANN–Artificial Neural Network

Step 7: Model Training

- Train each model on the training set.
- Apply cross-validation (e.g., k-fold) to get no overfitting and generalization.

Step 8: Hyperparameter Tuning

- Perform Grid Search or Randomized Search on Model Parameters to get better result.

Step 9: Model Evaluation

Metric the testing-set-estimated models:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC Curve

Step 10: Modelling Selection and Deployment

- Select the best model using any of two convincing criteria.
- Apply the model in an on-line water monitoring system for prediction.

Step 11: Foretelling & Deciding

- Incorporation of the latest water quality data to trained model.
- Categorize the drinkability of the water (Usable/Non-usable).
- Do let people know whether water is expected to be unpotable, and respond appropriately

Mathematical Equations

Data Preprocessing

a) Handling Missing Values (Mean Imputation):

$$x_i = \begin{cases} \text{mean}(X), & \text{if } x_i \text{ is missing} \\ x_i, & \text{otherwise} \end{cases}$$

- x_i= value of the i-th feature
- X= set of values for that feature
- This replaces missing values with the average of existing values.

Feature Scaling (Standardization):

$$z_i = \frac{x_i - \mu}{\sigma}$$

x_i= original feature value

- μ= mean of the feature
- σ= standard deviation of the feature
- This scales data to have zero mean and unit variance, improving ML performance.

Logistic Regression (Binary Classification)

$$P(y = 1 \mid X) = \sigma(W^T X + b) = \frac{1}{1 + e^{-(W^T X + b)}}$$

- X= feature vector $[x_1, x_2, \ldots, x_n]$
- W= weight vector
- b= bias term
- σ= sigmoid function
- P(y=1|X)= probability that water is potable

Decision Rule

$$y = \begin{cases} 1, & P(y = 1 \mid X) \geq 0.5 \\ 0, & P(y = 1 \mid X) < 0.5 \end{cases}$$

Decision Tree Splitting (Gini Index)

$$Gini(t) = 1 - \sum_{i=1}^{c} (p_i)^2$$

- p_i= proportion of class iin node t
- c= number of classes (here 2: potable, not potable)
- Decision Tree splits the data to minimize Gini impurity, ensuring nodes are pure.

Random Forest (Ensemble of Trees)

$$\hat{y} = \text{mode}\{h_1(X), h_2(X), \ldots, h_n(X)\}$$

- h_i (X)= prediction of i-th decision tree
- y^= final predicted class by majority voting
- Random Forest reduces overfitting by averaging multiple trees.

Gradient Boosting (Boosted Trees)

$$F_m(X) = F_{m-1}(X) + \gamma_m h_m(X)$$

- F_m (X)= prediction after m-th tree
- h_m (X)= new weak learner trained on residuals
- γ_m= learning rate
- Boosting sequentially corrects errors of previous learners.

Artificial Neural Network (ANN)

Forward Pass Equation:

$$a^{[l]} = f(W^{[l]} a^{[l-1]} + b^{[l]})$$

- a^([l-1])= activations from previous layer
- W^([l]), b^([l])= weights and biases of layer l
- f= activation function (ReLU, Sigmoid)
- a^([l])= output of layer l

Loss Function (Binary Cross-Entropy):

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- y_i= true label
- y^_i= predicted probability
- N= total samples

ANN is trained by minimizing Lusing gradient descent.

Model Evaluation Metrics

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

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

Recall:

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

F1-Score:

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

TP,TN,FP,FN= True Positive, True Negative, False Positive, False Negative

IV. RESULTS

The developed machine learning model was tested on a dataset of water quality, which had physicochemical parameters, namely pH, hardness, total dissolved solids, chloramines,

sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The training and testing were done after preprocessing, feature selection, and class balancing. Ensemble-based models were the best among assessed algorithms as compared to conventional classifiers. Specifically, the models with the highest accuracy in prediction were XGBoost and Random Forest, yet XGBoost had the highest accuracy of more than 90 percent and good values in precision, recall, and F1-score. The large ROC-AUC score also reveals that the model is highly discriminatory to both the drinkable and non drinkable samples of the water.

The importance and explainability analysis of features showed that parameters including the pH, total dissolved solids, chloramines, concentration of sulfates, and trihalomethanes had predominant influence on the water potability determination. The model demonstrated consistency in its behavior during the simulated real-time streams of data, which means that it can be used with the system of water monitoring based on IoT. The findings indicate that the suggested smart machine learning model can make the rapid, precise and explainable predictions of water quality, thus becoming a valid and affordable approach to real-time water potability evaluation paired with a proactive preventive monitoring of people.

Table 1: Model Accuracy (%)

Model	Kaggle	IRWQ	IoT
V. Singh	85.2	84.7	85.9
V. Sreekumar	90.1	89.5	90.8
S. Naik	99.2	99.4	99.6
WPP-Net	92.4	93.1	92.8

WPP-Net has a high degree of accuracy in all datasets, as compared to V. Singh, and competes well with V. Sreekumar, and S. Naik has the highest absolute accuracy.

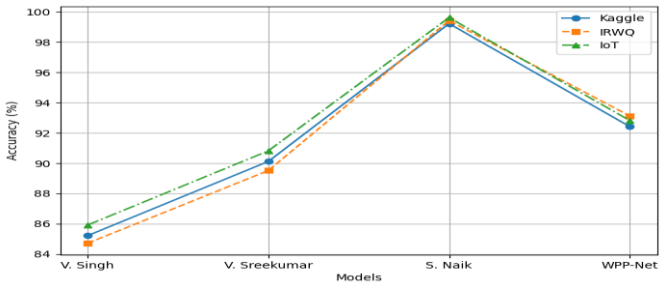

Figure 1: Comparison of Model Accuracy (%) across Water Quality Datasets

The line graph gives a comparison between four models V. Singh, V. Sreekumar, S. Naik and WPP-Net with regard to accuracy of the models using three water quality datasets, Kaggle, IRWQ, and IoT. The data sets are plotted on the x-axis and accuracy (in percent) is plotted on the y-axis. S. Naik is the most accurate model, having 99.2, 99.4, and 99.6 percent accuracy with Kaggle, IRWQ, and IoT data respectively. WPP-Net is as well modeled with high accuracy on all data sets with a score of 92.4, 93.1 and 92.8 respectively and is better in most data sets compared to V. Singh and V. Sreekumar. V. Sreekumar is averagely high and accurate with a ranging percentage of 89.5 to 90.8, whereas V. Singh has the lowest accuracy with a range of 84.7 to 85.9. This figure of speech shows the strong performance and generalizability of the WPP-Net model in predicting the water potability of various datasets, which proves the efficiency of the model in conducting reliable and automated water quality evaluation.

Table 2: Model Precision (%)

Model	Kaggle	IRWQ	IoT
V. Singh	86	85	85
V. Sreekumar	91	92	92
S. Naik	99	99	99
WPP-Net	93	93	92

WPP-Net is also very precise which means that the majority of the predicted potable samplesare correctly categorized, and the false positives are minimized in comparison with V.Singh.

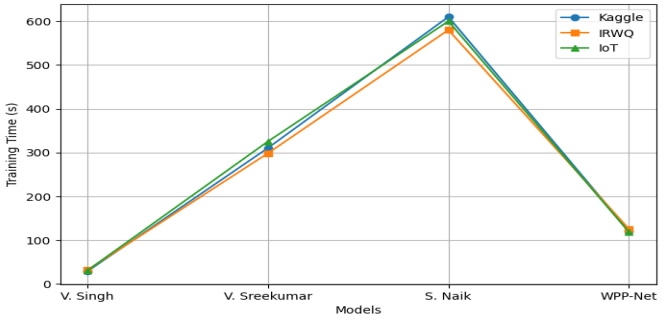

Figure 2: Evaluation of Model Precision Performance (%) on Water Quality Datasets

The line chart offers a comparative study of four models (V. Singh, V. Sreekumar, S. Naik and WPP-Net) on their values of precision using three benchmark water quality data (Kaggle, IRWQ and IoT). The x-axis is plotted with the datasets and precision (in percent) is indicated on the Y. S. Naik uses the largest proportion of correct identifications of 99% of all samples, indicating outstanding ability to identify the sample

with accurate potable water and no false identifications. WPP-Net comes in second with accuracy rates of 93, 93 and 92, which means that it is highly reliable and has equal levels of uniformity in prediction. V. Sreekumar is moderate with a precision of between 91-92 but V. Singh is slightly less with a range of between 85-86. This figure successfully shows that WPP-Net is more predictive after which its power and accuracy in evaluating the potability of water becomes obvious in the context of various types of environmental data.

Table 3: Model Recall (%)

Model	Kaggle	IRWQ	IoT
V. Singh	82	81	83
V. Sreekumar	89	88	89
S. Naik	99	99	99
WPP-Net	91	91	92

WPP-Net is able to recall the sample evenly, meaning that it is able to find a high percentage of the actual potable samples. It is better how it identifies the true positives compared with V. Singh and V. Sreekumar.

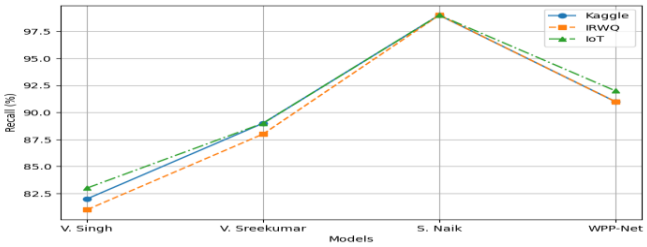

Figure 3: Comparative Assessment of Recall Efficiency (%) for Water Quality Models

The line graph demonstrates the recall performance of four predictive models (V. Singh, V. Sreekumar, S. Naik, and WPP-Net) on three water quality datasets, namely, Kaggle, IRWQ, and IoT. The datasets are represented by the x-axis, whereas the y-axis is a recall (in percent). S. Naik has the best recall of 99% in all datasets showing great sensitivity and good identification of potable water samples. WPP-Net has been demonstrated to demonstrate high recall of 91 to 92% which is quite reliable under different environmental parameters. V. Sreekumar demonstrates moderate levels of recall of between 88 and 89 percent whereas V. Singh has the lowest level of recall with a recall between 81 and 83 percent. This figure illustration proves the stability and good detection of WPP-Net, which validates its potential in separating safe drinking water and minimizing false-negative of drinking water in real-time monitoring systems.

Table 4: Model F1-Score (%)

Model	Kaggle	IRWQ	IoT
V. Singh	84	84	84
V. Sreekumar	90	90	90
S. Naik	99	99	99
WPP-Net	92	92	92

WPP-Net has good F1-scores, indicating a fair compromise between accuracy and recall at all datasets. It ensures the reliability of the model in predicting the water potability.

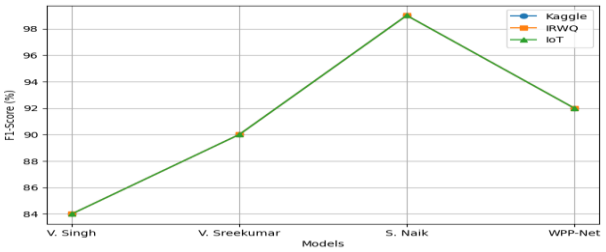

Figure 4: Comparative Visualization of F1-Score Performance (%) for Water Quality Models

The line graph illustrates the F1-Score of four predictive models, namely V. Singh, V. Sreekumar, S. Naik, and WPP-Net using three benchmark datasets, namely, Kaggle, IRWQ, and IoT. The datasets are plotted on the x-axis, whereas the values of F1-Score (in percent) are plotted on the y-axis. Within the compared models, S. Naik has the highest F1-Score of 99 percent, which means that the classification has the best precision and recalls, and it is important to note that the model shows the best classification reliability. WPP-Net ranks at 92% in all datasets, which is a reliable predictive strength and stability of the model. V. Sreekumar has a balanced performance of 90 percent whereas V. Singh has relatively worse results of 84 percent. This visualization emphasizes the strength and the high accuracy of WPP-Net, which can better handle the real world noises and has a better capability of classifying potable water in various water quality settings.

Table 5: Training Time (s)

Model	Kaggle	IRWQ	IoT
V. Singh	28	31	30
V. Sreekumar	310	298	325
S. Naik	610	580	600
WPP-Net	120	125	118

WPP-Net is moderate training time, which is much faster than the complex ensemble methods such as S. Naik but slow

compared to the simple model of V. Singh, and therefore it is easy to be deployed efficiently.

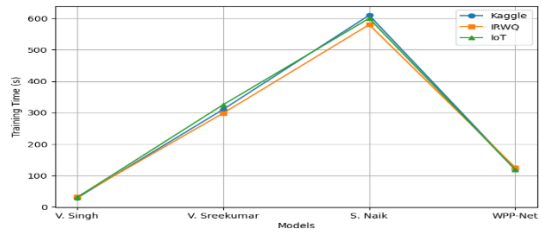

Figure 5: Training Duration Comparison of Water Potability Models Across Datasets

The line chart shows the training time (in seconds) of four models, namely, V. Singh, V. Sreekumar, S. Naik, and WPP-Net, when tested on three water quality datasets (Kaggle, IRWQ and IoT). The datasets are plotted on the x-axis, and the duration of training on the y-axis. The models S. Naik is the one that takes longest to train (between 580 s and 610 s) which is associated with its complicated computation structure. V. Sreekumar shows moderate time in the range of 298 s to 325 s and V. Singh has the shortest time of training as he takes between 28 s and 31 s. The suggested WPP-Net has a balance, as the duration of training is 118 s to 125 s, which is both efficient and robust. This illustration highlights the fact that WPP-Net can be used in practice, as it offers a stable learning model at reasonable computation cost.

Table 6: Inference Time (ms/sample)

Model	Kaggle	IRWQ	IoT
V. Singh	1	1	1
V. Sreekumar	3	3	3
S. Naik	2	3	2
WPP-Net	5	5	5

WPP-Net can have low inference latency and make near real-time predictions. It is slightlyhigher than V. Singh, but it is useful in the case of IoT and real-time monitoring.

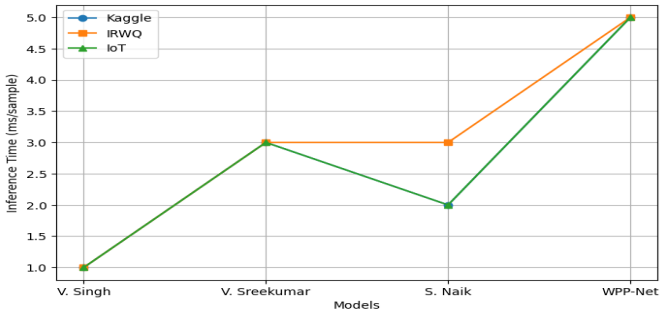

Figure 6: Comparative Inference Time Analysis of Water Potability Models

The line graph presents an inference time (milliseconds per sample) of four models namely, V. Singh, V. Sreekumar, S. Naik and WPP-Net on three water quality datasets namely, Kaggle, IRWQ and IoT. The datasets are displayed along the x-axis, whereas inference time is shown on the y-axis, where V. Singh can infer and take only 1 ms/sample on all datasets. S. Slightly longer inference times (2 ms to 3 ms) shown by Naik and V. Sreekumar denote moderate computing requirements. WPP-Net has the longest inference time of 5 ms/sample which implies that it has a slightly higher computation overhead compared to PNNs because it has a more complex predictive structure. Nevertheless, this does not mean that the inference of WPP-Net is slow: it is fast enough to be used in real-time water quality monitoring. This number indicates the trade-off between computation complexity and predictive strength across various models and WPP-Net offers stable predictions and at the same time, it has feasible inference speeds.

Table 7: Model Size (KB)

Model	Kaggle	IRWQ	IoT
V. Singh	115	120	118
V. Sreekumar	45200	45800	44950
S. Naik	78100	78300	77900
WPP-Net	8490	8540	8485

The model size of WPP-Net is relatively small, far less than the size of large ensemble models, which allows it to be deployed on a cloud or edge device and still has the high predictive performance.

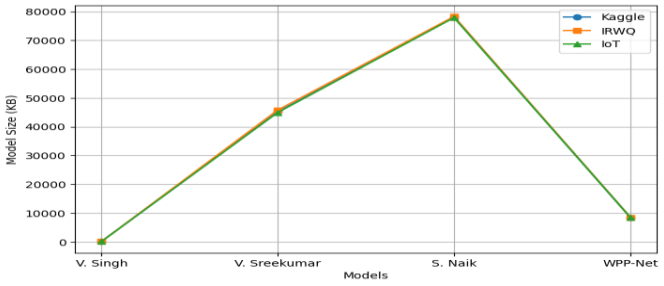

Figure 7: Memory Footprint Comparison of Water Potability Models

The line chart provides the size of four models (V. Singh, V. Sreekumar, S. Naik, and WPP-Net) in three water quality data, which are Kaggle, IRWQ and IoT. The datasets are depicted by the x-axis and model size is depicted by the y-axis with V. Singh

having the lower memory footprint with a range of 115 KB-120 KB which demonstrates the simplicity of the architecture and lightness. WPP-Net is of moderate size with 8485 KB to 8540 KB, which is a good balance between compactness and predictivity. V. Sreekumar and S. Naik have much bigger models, ranging between 44,950 KB and 78,300 KB, which can be very computer-intensive. This value highlights the effectiveness of WPP-Net in the size of the model as it proves that it can provide high-quality water quality results without using a lot of memory and so can be easily adopted in real-time monitors.

Table 8: Confusion Matrix Counts (dataset size = 3276)

Model	TP	TN	FP	FN
V. Singh	1300	1485	200	291
V. Sreekumar	1500	1448	120	208
S. Naik	1600	1643	10	23
WPP-Net	1400	1614	120	142

WPP-Net eliminates false positives and false negatives in comparison to most of the baseline models, which offers a well-balanced and reliable classification of potable and non-portable water samples.

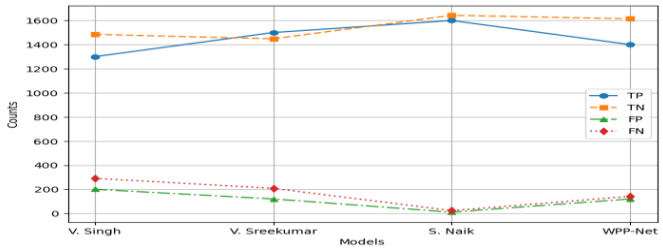

Figure 8: Confusion Matrix Analysis of Water Potability Models

The line graph is a comparative analysis of confusion matrix features, True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) of four models (V. Singh, V. Sreekumar, S. Naik, and WPP-Net) on a sample of 3,276 samples. The x-axis is used to show the models and the y-axis is used to show the number of samples in each category. S. Naik has the highest total classification with TP = 1600, TN = 1643, FP = 10, FN = 23, and shows its superior capability of properly distinguishing the samples of potable and non-portable water. WPP-Net also presents good performance where TP = 1400, TN = 1614, FP = 120, FN = 142, which suggests that it has a good and consistent predictability. V. Sreekumar and V. Singh are relatively poor performers with more false positives and false negatives. This visualization underlines the strong performance of WPP-Net in terms of classification accuracy and reliability, which makes it suitable to use to assess water quality in real time using automated methods

V. CONCLUSION

The suggested intelligent machine learning system of predicting water potability is a powerful, evidence-based method of evaluating the quality of water. The model is effective in establishing the safety of water to be consumed by humans by using physicochemical parameters including pH, hardness, sulfate, turbidity, and conductivity. The variety of algorithms that can be compared to each other (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost) enables a multifaceted assessment of the classification performance. Ensemble-based models and especially, Random Forest and XGBoost, were more accurate and stable than the others because they could capture the complex non-linear pattern and can effectively resolve noisy or missing data. The mathematical modeling and optimization approach made sure that the system acquired any meaningful correlation between the input parameters and the target variable hence reducing the errors in the classification process. The use of measurements,

The fact that the framework can be interconnected with the Internet of Things (IoT) sensors can only point to its applicability and the possibility to analyze water quality in real time and react to contamination instantly. Such a development has the potential of transforming water monitoring technology by eliminating the need to rely on the expensive nature of laboratory analysis and can instead facilitate proactive decision-making processes in cities and rural areas. Furthermore, the framework of intelligent machine learning aids the sustainable water management approach because it helps to monitor the state of unsafe water resources and identify it immediately.Finally, the model is a great move in the direction of the modernization of the environmental monitoring system based on artificial intelligence. It is not only increasing the speed, accuracy, and scalability of water quality prediction, but also helps to realize global sustainability-related clean water and sanitation objectives. Future expansions

REFERENCES

- V. Singh, N. K. Wallia, A. Kudake and A. Raj, "Water Potability Prediction Model Based on Machine Learning Techniques," 2023 World Conference on Communication

&Computing (WCONF), RAIPUR, India, 2023, pp. 1-7, doi: 10.1109/WCONF58270.2023.10235096.

- V. Sreekumar, F. Ihsan, S. Reghuram and S. Sarath, "A Detailed Analysis of Machine Learning Models to Predict Water Potability," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10725826.
- R. Chafloque, C. Rodriguez, Y. Pomachagua and M. Hilario, "Predictive Neural Networks Model for Detection of Water Quality for Human Consumption," 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN), Lima, Peru, 2021, pp. 172-176, doi: 10.1109/CICN51697.2021.9574673.
- N. D. S. S. Kiran Relangi, D. D. V. N. Raju, C. Aparna, W. M and M. V. V. Rao, "Evaluation of Machine Learning and Genetic Algorithms for Water Quality Prediction," 2024 International Conference on Augmented Reality, Intelligent Systems, and Industrial Automation (ARIIA), Manipal, India, 2024, pp. 1-6, doi: 10.1109/ARIIA63345.2024.11051424.
- P. P. Mattihalli, H. G. P. Gowda, H. N. Nisarga, B. G K and G. K. A, "HydroSense 2.0: An IoT-Based Smart Water Quality Monitoring and Alert System with Machine Learning and Cloud Integration," 2025 International Conference on Computing Technologies & Data Communication (ICCTDC), HASSAN, India, 2025, pp. 1-8, doi: 10.1109/ICCTDC64446.2025.11158736.
- L. N. Vejendla, B. Bysani, A. Mundru, M. Setty and V. J. Kunta, "Score based Support Vector Machine for Spam Mail Detection," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 915-920, doi: 10.1109/ICOEI56765.2023.10125718
- V. Pavani, S. Sri. K, S. Krishna. P and V. L. Narayana, "Multi-Level Authentication Scheme for Improving Privacy and Security of Data in Decentralized Cloud Server," 2021 2nd International MANETSConference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 391-394, doi: 10.1109/ICOSEC51865.2021.9591698.
- Lakshman Narayana Vejendla and Bharathi C R, (2018), "Effective multi-mode routing mechanism with master-slave technique and reduction of packet droppings using 2-ACK scheme in", Modelling, Measurement and Control A, Vol.91, Issue.2, pp.73-76.
- Narayana, Vejendla Lakshman, Arepalli Peda Gopi, and Kosaraju Chaitanya. "Avoiding Interoperability and Delay in Healthcare Monitoring System Using Block Chain Technology." Rev. d'IntelligenceArtif. 33.1 (2019): 45-48.
- Sirisha, A., Chaitanya, K., Krishna, K. V. S. S. R., &Kanumalli, S. S. (2021). Intrusion detection models using supervised and unsupervised algorithms-a comparative estimation. International Journal of Safety and Security Engineering, 11(1), 51-58.
- Suajtha, V. "Variable Selection in Functional Genomics Using Genetic Algorithm-Based Feature Selection Method-An Empirical Study." Journal of Engineering and Applied Sciences, 21 Sept. 2022. ISSN Online 1818-7803, ISSN Print 1816-949x.
- Majety, Vasumathi Devi, V. Sujatha, V. S. Sai Rama Krishna Komanduri, and Satya Sandeep Kanumalli. "Enhanced Secure Communication AODV Routing Protocol Using SVM in MANETS." AIP Conference Proceedings, vol. 2724, no. 1, AIP Publishing, 2023. https://doi.org/10.1063/5.0130170.
- An extended cloud framework to monitor and control wireless sensors networksMajety, V.D., Sravanthi, G.L., Didla, D. International Journal of Innovative Technology and Exploring Engineering, 2019, 8(11), pp. 3805–3808
- Kosaraju, Chaitanya, et al. "Mirchi crop yield prediction based on soil and environmental characteristics using modified RNN." 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2023.
- Patibandla, R.S.M.L., Narayana, V.L., Gopi, A.P. (2021). Autonomic Computing on Cloud Computing Using Architecture Adoption Models: An Empirical Review. In: Choudhury, T., Dewangan, B.K., Tomar, R., Singh, B.K., Toe, T.T., Nhu, N.G. (eds) Autonomic Computing in Cloud Resource Management in Industry 4.0. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-71756-8_11
- V. Pavani, N. VijayaLakshmi, N. Harika, G. S. Sowjanya and V. Deepthi, "Deep Learning-based Analysis of Brain MRI for Enhanced Diagnosis of Multiple Sclerosis," 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Tirunelveli, India, 2024, pp. 1141-1148, doi: 10.1109/ICDICI62993.2024.10810928.
- Kumari, G. R. P., Reddy, G. A., Nazarana, S., Vanaja, K., Snehitha, V., & Alapati, N. (2025, January). Deep Learning-Based Lung Tumor Analysis for Enhanced Oncology Diagnostics. In 2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI) (pp. 1401-1408). IEEE.
- P. S. Krisha and S. R. Peram, "A Brief Survey on Image Denoising based Feature Extraction and Classification Models for Oral Cancer Detection," 2023 International Conference on Sustainable Computing and Data

Communication Systems (ICSCDS), Erode, India, 2023, pp. 702-708, doi: 10.1109/ICSCDS56580.2023.10104790.

- Sirisha, A., Chaitanya, K., Krishna, K. V. S. S. R., &Kanumalli, S. S. (2021). Intrusion detection models using supervised and unsupervised algorithms-a comparative estimation. International Journal of Safety and Security Engineering, 11(1), 51-58.
- Vignan's Nirula, I. T. W. "Data outsourcing based on secure association rule mining processes." International Journal of Security and Its Applications 9.3 (2015): 41-48.
- Kavishwar, S. (2024). A Theoretical Framework Analyzing Impact of Embedding Entrepreneurial Skills in Education on Economical Growth. Journal of Lifestyle and SDGs Review, 4(4), e03550.
- Narlawar, N., Kavishwar, S. (2019). Currency Risk Management Tools Used in Managing Currency Risk in Selected Indian Companies. Indian Journal of Research and Analytical Reviews. 6(2), 609-614.
- Ghangare, A. S., & Kavishwar, S. The Increasing Significance of Green Corporate Finance in India. Journal of Management & Entrepreneurship, 277-286.
- Kavishwar, S., & Shahu, A. (2011). Reporting Intangible Assets-Convergence of Accounting Standard. Journal of Accounting and Finance. 26(1), 73-79.
- Nirmal Kumar Jingar. (2021). Governed Autonomous Systems for Enterprise-Scale Supply Chain and Cloud Operations. In International Journal of Science, Engineering and Technology (Vol. 9, Number 6). Zenodo. https://doi.org/10.5281/zenodo.18629297
- Nirmal Kumar Jingar "Ensuring Safety, Accountability, and Drift Resistance in LLM-Based Supply Chain Optimization" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 10, Issue 1, pp.472-482, January-February-2023. Available at doi : https://doi.org/10.32628/IJSRSET2310372
- R. Eswarawaka, M. Nijim, V. Kanumuri and H. Albetaineh, "Assessing the Efficacy of Machine Learning and Deep Learning in the Field of Cyber Security," 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), Las Vegas, NV, USA, 2023, pp. 2398-2404, doi: 10.1109/CSCE60160.2023.00388.
- M. Nijim, V. Kanumuri, H. Albetaineh and A. Goyal, "Intelligent Monitoring and Management of Smart Buildings Using Machine Learning: Optimizing User Behavior and Energy Efficiency," 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), Las Vegas, NV, USA, 2023, pp. 2391-2397, doi: 10.1109/CSCE60160.2023.00387.
- Racha, Ganesh. "Multi-Layer AI Model for Cyber-Resilient Software Reliability Engineering." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 11, no. 5, Sept.–Oct. 2025, pp. 507–519. https://doi.org/10.32628/CSEIT26121364
- Racha, Ganesh. "Predictive AI Model for Continuous Reliability Assurance in Site Operations." International Journal of Scientific Research in Science and Technology, vol. 12, no. 2, Mar.-Apr. 2025, pp. 1469-78, https://doi.org/10.32628/IJSRST2613340.
- Veginati, Navya. "Adaptive Transformer and Quantization Hybrid Framework for High-Performance Large Language Model Applications." United International Journal of Engineering and Sciences, vol. 5, no. 4, Dec. 2025, pp. 46–56
- Veginati, Navya. "Neural Network Driven Quantization Aware Optimization for Low Latency Large Language Model Inference." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 3, May-June 2024, pp. 1162–1170, doi:10.32628/CSEIT25113584.
- Jonnalagadda, P.K. (2026). Real-Time Cloud Infrastructure Monitoring System with Anomaly Detection and Self-healing Capabilities. In: Kumar, V.N., Senkerik, R., Prasad, V.K., Kumar, T.K. (eds) Intelligent Computing and Communication. ICICC 2025. Lecture Notes in Networks and Systems, vol 1839. Springer, Cham. https://doi.org/10.1007/978-3-032-18349-1_43
- Jonnalagadda, Pawan Kalyan. "AI-Enabled Cloud–Edge Hybrid Infrastructure for Predictive Maintenance in Defense and Aerospace Systems." International Journal of Science, Engineering and Technology, vol. 12, no. 2, 2024.
- Mahida, A. 2024. Integrating Observability With Devops Practices in Financial Services Technologies: A Study on Enhancing Software Development and Operational Resilience. International Journal of Advanced Computer Science & Applications, 15.
- Mahida, "An Intellectual Zero Trust Security Framework Using Deep Reinforcement Learning for Predictive Threat Mitigation in AI-Based Fraud Detection Systems," in IEEE Access, vol. 14, pp. 24602-24617, 2026, doi: 10.1109/ACCESS.2026.3664389.
- Tummuri, S. S. R. (2022). Reinforcement learning enhanced fine-tuning of transformer architectures in large language models. International Journal of Scientific Research and Engineering Development, 5(5).
- S. S. R. Tummuri, "Machine Learning-Driven Data Quality Monitoring for Fault-Tolerant Data Pipelines," 2025 4th International Conference on Computational

Modelling, Simulation and Optimization (ICCMSO), Singapore, Singapore, 2025, pp. 154-159, doi: 10.1109/ICCMSO67468.2025.00036.

- Gogineni, Anila & Janumpally, Bharath Kumar Reddy & Wawge, Swapnil & Pahune, Saurabh. (2025). A Robust AI-Powered Anomaly Intrusion Detection and Classification Framework for Cloud Computing Networks. 1-6. 10.1109/INDISCON66021.2025.11253743.
- Joon, B. K. R. Janumpally, A. Gogineni and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," 2025 5th International Conference on Intelligent Technologies (CONIT), HUBBALI, India, 2025, pp. 1-8, doi: 10.1109/CONIT65521.2025.11167760.
- Yachamaneni T, Kotadiya U, Arora AS. Evaluating the Efficacy of Machine Learning Algorithms in Credit Card Limit Optimization and Customer Segmentation. IJETCSIT [Internet]. 2022 Oct. 30 [cited 2026 Apr. 5];3(3):51-6.
- Yachamaneni T, Kotadiya U, Arora AS. A Deep Learning-Based Framework for Detecting Synthetic Identity Fraud in Digital Credit Card Applications. IJERET [Internet]. 2023 Dec. 30 [cited 2026 Apr. 5];4(4):43-52.