Water Quality Analysis using Machine Learning

K Rajesh[†] , K Tharun Kumar Reddy[†] , Karthik Kemidi[†]
Chaitanya Bharathi Institute of Technology[†]
Department of Artificial Intelligence and Machine Learning[†]
Email: karthikkemidi2004@gmail.com*

Abstract

Water quality prediction was generated for predicting if the water is safe to drink or not. This experiment was also conducted to compare the machine learning model performance between Decision Tree, Random Forest, XGBoost, KNN, SVM, Gaussian Naive Bayes to determine the most suitable technique for predicting Water Quality.Water is the most crucial resource of life and it is necessary for the survival of all living creatures including human beings. The survival of business and agriculture depends on freshwater. An essential step in managing freshwater assets is the evaluation of the quality of the water. Before using water for anything, including drinking, chemical spraying (pesticides, etc.), or animal hydration, it is crucial to assess its purity. The ecosystem and the general public's health are directly impacted by water quality. Therefore, analysing and predicting water quality is necessary for both environmental and human protection. Machine learning can be used to analyse and predict the water quality based on the parameters like PH value, turbidity, hardness, conductivity, dissolved solids in water and other parameters. In this work, the water quality is predicted by giving the concentration of various parameters as input to machine learning algorithms and the water is classified as safe or unsafe for the usage of domestic purposes

***Keywords*—Decision Tree, Random Forest, KNN, SVM, XGBoost, Gaussian Naive Bayes, Performance Metrics,**

Introduction

The scientific field of machine learning, it is investigated how computers learn via experience. Since the capacity to learn is the fundamental quality of an entity regarded as intelligent in the broadest meaning of the word, the words "Machine Learning" and "Artificial Intelligence" are frequently used synonymously in the minds of scientists. Building adaptable, experience-based computer systems is the goal of machine learning. It is now possible to discover a solution to this problem because of the development of machine learning methods. We have developed a technique that uses data mining to identify whether the water is portable or not. The enormous amount of data related to water quality can be mined for hidden knowledge. As a result, it now has a more significant role in the study. This research aims to develop a system that can predict water quality more precisely.

Literature Survey

[1] One comprehensive model analyzed water quality parameters using machine learning algorithms to predict conditions with high accuracy and reliability. This study highlighted the importance of data preprocessing and feature selection in improving prediction outcomes.[2] Another study compared multiple machine learning techniques to identify the most effective approach for water quality prediction. It was found that methods such as random forests and support vector machines demonstrated superior predictive performance due to their ability to handle non-linear relationships among water quality parameters.

[3] Real-time monitoring systems integrated with machine learning algorithms have been developed to provide continuous water quality information. These systems enable prompt responses to potential health risks by detecting contaminants and other changes in water composition in real time.

[4] Hybrid models, which combine different machine learning approaches, have shown improved performance in predicting specific water quality indicators such as salinity. These hybrid methods are particularly effective in capturing complex interactions between variables that traditional models may miss.

[5] Some studies focus on analyzing and predicting water quality using datasets from field sensors and public records. By applying exploratory data analysis

techniques, researchers have identified key patterns in the data, enabling them to predict long-term trends and seasonal variations in water quality.[6] Recently, there has been a rise in research leveraging machine learning to assess and predict water potability, training models to evaluate water's suitability for drinking based on a variety of quality metrics.

[7] Studies have evaluated machine learning algorithms for sustainable monitoring of drinking water quality, emphasizing these models' potential to make water management more efficient. This research often discusses the implications for sustainability and public health.

[8] Comparative analyses of various machine learning algorithms on datasets with statistically imputed missing values have deepened the understanding of each model's strengths and limitations, highlighting how data preprocessing affects model accuracy.

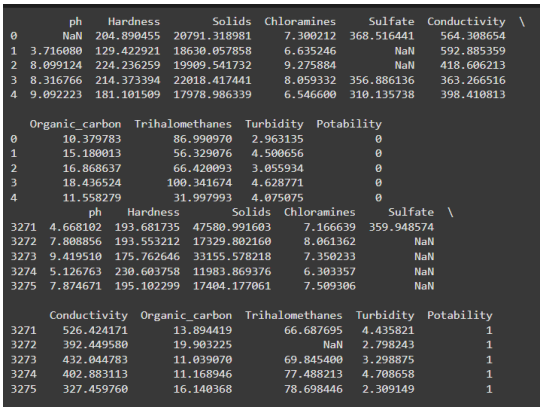
[9] The integration of machine learning with comprehensive weighting techniques has been explored to enhance prediction reliability by combining the outputs of multiple models. This method reduces errors from individual models, yielding more robust predictions.[10] Several studies have aimed to predict water potability by classifying water samples based on various quality indicators. This research is particularly valuable for public health monitoring.[11] Recent research has focused on optimizing machine learning models for water quality prediction by refining data preprocessing and tuning model parameters. These tailored models improve prediction accuracy significantly, especially when adapted to local environmental conditions and specific water quality attributes.

Methodologies

The machine learning model is used to detect whether the water is potable or non-potable. Import relevant libraries to test and train our data set and required to install some packages related to nature-inspired algorithms. Split the data as training data set and testing dataset they should bein the ratio 80:20 respectively and perform the Model Selection. The Support Vector Classifier (SVC), Decision Tree, GaussianNB, Random Forest and XGBoost are these different classifiersthat are taken into consideration.

The dataset consists of observations of water quality for 3276 different sources of water:

pH - The water's pH (0 to 14). According to EPA recommendations, tap water's pH should range between (6.5 and 8.5). The pH level is a crucial factor in determining the acid-base nature of water. Additionally, it shows if the water is either alkaline or acidic. The



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	\
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	

	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	10.379783	86.990970	2.963135	0
1	15.180013	56.329076	4.500656	0
2	16.868637	66.420093	3.055934	0
3	18.436524	100.341674	4.628771	0
4	11.558279	31.997993	4.075075	0

	ph	Hardness	Solids	Chloramines	Sulfate	\
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	
3272	7.808856	193.553212	17329.802160	8.061362	NaN	
3273	9.419510	175.762646	33155.578218	7.350233	NaN	
3274	5.126763	230.603758	11983.869376	6.303357	NaN	
3275	7.874671	195.102299	17404.177061	7.509306	NaN	

	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
3271	526.424171	13.894419	66.687695	4.435821	1
3272	392.449580	19.903225	NaN	2.798243	1
3273	432.044783	11.039070	69.845400	3.298875	1
3274	402.883113	11.168946	77.488213	4.708658	1
3275	327.459760	16.140368	78.698446	2.309149	1

Figure 1

present investigation's range fell between 6.52 to 6.83, which is within WHO guidelines.

Hardness - It is the amount of soap that may dissolve in one litre of water. Salts made of calcium and magnesium are the major causes of hardness. How long water is exposed to a hardness-producing substance influences how hard the water is while it is in the raw state. The ability of water to form soap due to calcium and magnesium precipitation was the original definition of hardness.

Total Dissolved Solids (TDS) - Water can dissolve a wide variety of chemicals and certain organic minerals or salts, including sodium, calcium, iron, zinc, bicarbonate ions, chloride ions, magnesium, and sulphates. These minerals affected the water's appearance and gave it foul smells. This is an important consideration while using water. A highTDS rating indicates that the water contains a lot of minerals. For drinking purposes, the maximum and desired TDS limits are 500 mg/l and 100 mg/l, respectively.

Sulfates - These are the organic substances that are found naturally in minerals, soil, and rocks. They are present in the air, groundwater, plants, and food in the area. Sulfate is mostly utilized for business purposes in the chemical sector. Around 2,700 mg/L of sulphate can be found in seawater. While certain places have significantly higher levels (1000 mg/L), most freshwater sources have values between 3 and 30 mg/L.

Conductivity - Pure water is great insulation of electrical current. By raising the ion concentration, the liquid's electric conductivity has been enhanced. The quantity of dissolved particles in the liquid often determines its conductivity. Electrical Conductivity measures how well they carry electricity through their ionic mechanism (EC). WHO recommendations state that the EC value shouldn't be higher than 400 S/cm.

Chloramines - The two primary disinfectants used

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               3276 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          3276 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3276 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

Figure 2: data description

```
Missing values after imputation:
 ph                 0
Hardness           0
Solids             0
Chloramines        0
Sulfate            0
Conductivity       0
Organic_carbon     0
Trihalomethanes    0
Turbidity          0
Potability         0
dtype: int64

Summary Statistics:
                ph     Hardness        Solids  Chloramines     Sulfate  \
count  3276.000000  3276.000000   3276.000000  3276.000000  3276.000000
mean      7.080795   196.369496  22014.092526     7.122277   333.775777
std       1.469956    32.879761   8768.570828     1.583085    36.142612
min       0.000000    47.432000    320.942611     0.352000   129.000000
25%       6.277673   176.850538  15666.690297     6.127421   317.094638
50%       7.080795   196.967627  20927.833607     7.130299   333.775777
75%       7.870050   216.667456  27332.762127     8.114887   350.385756
max      14.000000   323.124000  61227.196008    13.127000   481.030642

       Conductivity  Organic_carbon  Trihalomethanes    Turbidity   Potability
count   3276.000000     3276.000000      3276.000000  3276.000000  3276.000000
mean     426.205111       14.284970        66.396293     3.966786     0.390110
std       80.824064        3.308162        15.769881     0.780382     0.487849
min      181.483754        2.200000         0.738000     1.450000     0.000000
25%      365.734414       12.065801        56.647656     3.439711     0.000000
50%      421.884968       14.218338        66.396293     3.955028     0.000000
75%      481.792304       16.557652        76.666609     4.500320     1.000000
max      753.342620       28.300000       124.000000     6.739000     1.000000
```

Figure 3: data pre-processing

in public water systems are chloride and chlorine. Ammonia is used in combination with chlorine to clean potable water. Drinking water can include up to 4 mg/L of chlorine, which is regarded as a safe quantity.

Potability - It is a metric for determining whether water is fit for human consumption. Unpotable equals zero (0), while potable is one (1).

A. Data Pre-processing

The data quality must be improved at the processing stage of the data analysis process. The Water quality index has been determined in this phase using the important dataset parameters. The act of converting collected data into something an algorithm for machine learning can use is known as data preparation. The most important and first stage in building an algorithm for machine learning is this one. Remove all instances where the value is 0. (zero). Zero is not a possible value. Therefore, this instance is terminated. The process of deciding on feature subsets, which decreases the dimension of the data and helps to work more quickly, involves removing irrelevant characteristics and instances.

B. Correlation Matrix

By Visualizing the correlation of all characteristics using a thermal foot map function. But you can see from the heat map below that there is no correlation between any characteristic; this means that we cannot reduce the dimension.

*C. **Training and Testing of Data***

In machine learning, the model is instructed to perform a variety of tasks using a training set of data. The model is trained using certain features from the training set. Therefore, the prototype contains these structures. Words or word clusters are taken from tweets for sentiment analysis. They build connections, understand concepts, come to judgments, and assess

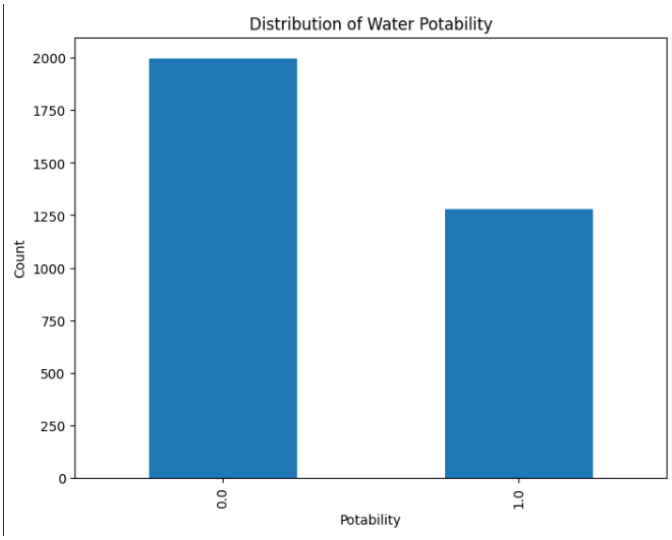


Figure 4: Water Potability Distribution

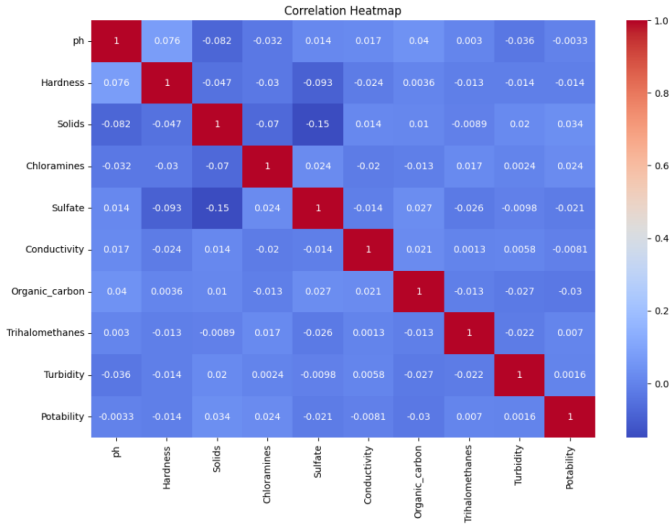


Figure 5: correlation matrix

their level of confidence using the training data. The quality and quantity of the Machine Learning training data, we use determines how well our data project performs, just as much as the algorithms they do. As a result, provided the training set is correctly labelled, the model will be able to learn about the features.

D. Decision Tree

The decision tree is a Machine Learning algorithm, it is mostly focused on classification-related issues. The decision tree has a structured classifier in which the nodes within display the components of a particular dataset. Decisionnodes and leaf nodes are both types of nodes found in decision trees.

E. Support Vector Machine

The SVM is an algorithm which is used in machine learning to categorize the task. It is frequently used for classification problems. SVM separates the data into two classes by mapping the data points to a high-dimensional space and then locating the best hyperplane.

F. Random Forest Classifier

The popular learning algorithm Random Forest is a part of the supervised learning methodology. It may be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating many classifiers to address difficult issues and enhance model performance. Random Forest, as the name implies, is a classifier that uses several decision trees on different subsets of the provided dataset and averages them to increase the dataset's prediction accuracy. Instead, then depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of most predictions.

G. XGBoost

Extreme Gradient Boosting is a framework that can run on multiple languages. It is popular supervised learning which works on large datasets. It is implemented on top of the gradient boost. The way the XGBoost algorithm is designed to work uses the parallelization concept. It uses sequentially generated shallow decision trees and a highly scalable training method to minimize overfitting to deliver accurate results.

H. KNN

K-Nearest Neighbors (KNN) is a simple, non-parametric machine learning algorithm used for classification and regression tasks. It works by finding the "k" closest data points (neighbors) to a given input and making predictions based on the majority class or average value of these neighbors. KNN is easy to implement but can be computationally intensive, especially with large datasets, since it requires calculating distances between points.

I. Performance Metrics

Accuracy - Accuracy is measured as the total count of actual predictions to the available predictions and it is multiplied by 100.

Precision - The ratio of actual positives to the total available positives is known as precision.

Recall - It mainly focuses on type-2 errors the ratio of true positives to false negatives is called recall.

Fl-score - The harmonic mean performance metric parameters precision with recall known as f1-score.

Results and Discussion

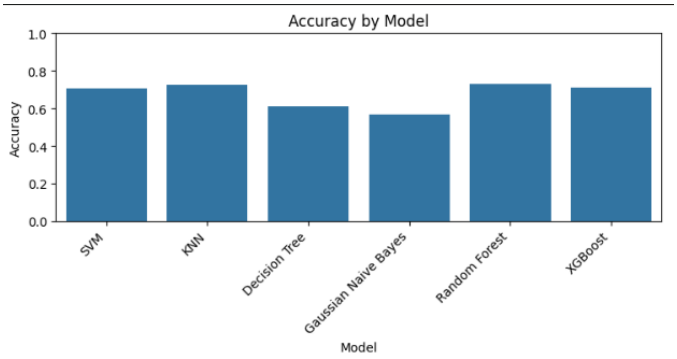


Figure 6: Accuracy

- The above graph shows the accuracy of six different machine learning models: SVM, KNN, Decision Tree, Gaussian Naive Bayes, Random Forest, and XGBoost. The accuracy is fairly consistent with SVM, KNN, Random Forest, and XGBoost achieving higher values compared to Decision Tree.

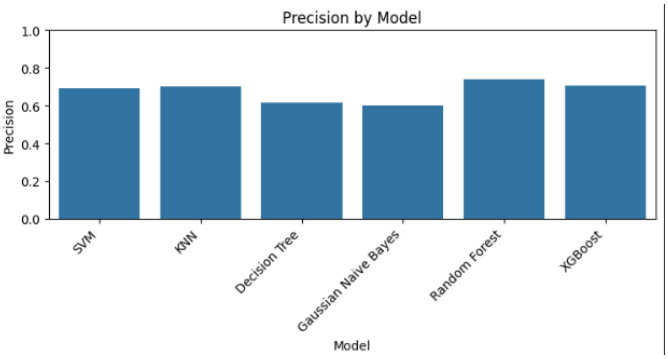


Figure 7: Precision

- This chart displays the precision for each model. Precision is highest for the Random Forest model, followed closely by KNN and XGBoost, while Gaussian Naive Bayes has slightly lower precision. Higher precision means fewer false positives for these models.

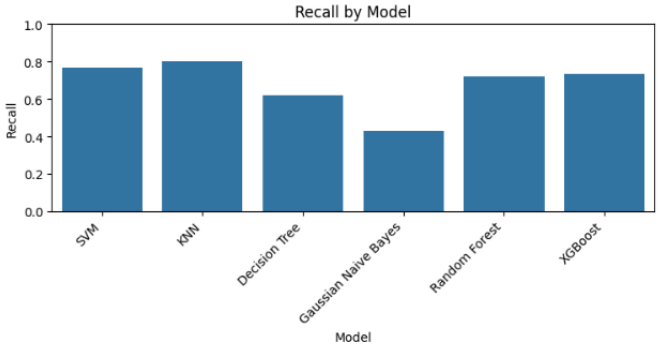


Figure 8: Recall

- This graph illustrates recall values, where SVM and KNN show the highest recall, meaning these models are better at identifying true positives. Gaussian Naive Bayes has the lowest recall, indicating it misses more true positives compared to the others.

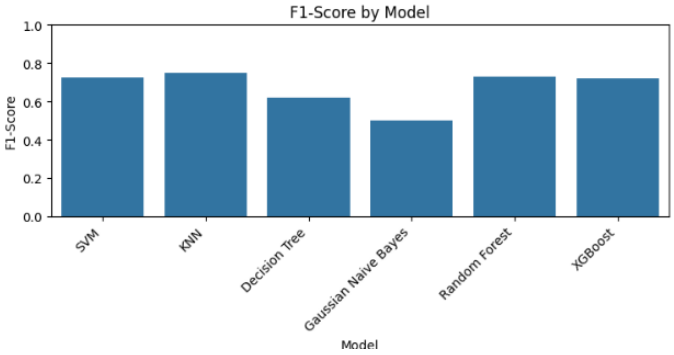


Figure 9: F1-score

- The F1-score, which is the harmonic mean of precision and recall, is shown here. KNN has the highest F1-score, indicating a good balance between precision and recall. Gaussian Naive Bayes has the lowest F1-score, suggesting it is less effective in balancing precision and recall.

In conclusion, the Random Forest Classifier emerged as the most robust model, achieving an optimal balance between accuracy, precision, recall, and F1-score. This model's superior balance indicates it is well-suited for applications that require consistent accuracy across multiple evaluation metrics.Random Forest Classifier worked the best to train the model, giving us an Accuracy (Balanced with precision recall) of around 70 percent.

```
Random Forest Results:
Accuracy: 0.6784
Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      0.86      0.77       412
         1.0       0.61      0.38      0.47       244

    accuracy                           0.68       656
   macro avg       0.65      0.62      0.62       656
weighted avg       0.67      0.68      0.66       656

Confusion Matrix:
[[353  59]
 [152  92]]
```

Figure 10: Best Model

Conclusions and Future Work

Future cities would benefit from real-time monitoring and evaluation of water quality due to the advancement of machine learning techniques. This work presented the results of our most recent literature analysis and comparative recent studies on the assessment of water quality using big data analytics and machine learning models and methods. Finally, it offers a few insights into theproblems, demands, and needs of future studies. Environmental protection greatly benefits from the modelling and forecasting of water quality. The algorithm implemented in this work improves the performance of water quality classifiers. We previously examined the performance metrics of machine learning algorithms, and we found that by utilizing Hyperparameter Tuning along with Random Forest Classifier, we delivered a better improvement in the execution of different performance metrics of the models using Hyperparameter Tuning. We have got better improvement in performance metrics. This strategy may be applied and improved for automated water quality monitoring.

References

[1] Khan, Y., & See, C. S. (2016, April). Predicting and analysing water quality using machine learning: a comprehensive model. In *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-6). IEEE.

[2] Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology, 578*, 124084.

[3] Vergina, S. A., Kayalvizhi, S., Bhavadharini, R., & Kalpana Devi, S. (2020). A real time water quality monitoring using machine learning algorithm. *Eur. J. Mol. Clin. Med,* 7(8), 2035-2041.

[4] Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T. (2020). River water salinity prediction using hybrid machine learning models. *Water, 12*(10), 2951.

[5] Kuthe, A., Bhake, C., Bhoyar, V., Yenurkar, A., Khandekar, V., & Gawale, K. (2022). Water quality analysis using machine learning. *International Journal for Research in Applied Science and Engineering Technology, 10*(12), 581-585.

[6] Akshay, R., Tarun, G., Kiran, P. U., Devi, K. D., & Vidhyalakshmi, M. (2022, December). Water-Quality-Analysis using Machine Learning. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 13-18). IEEE.

[7] Kaddoura, S. (2022). Evaluation of machine learning algorithm on drinking water quality for better sustainability. *Sustainability, 14*(18), 11478.

[8] Poudel, D., Shrestha, D., Bhattarai, S., & Ghimire, A. (2022). Comparison of machine learning algorithms in statistically imputed water potability dataset. *Journal of Innovations in Engineering Education, 5*(1), 38-46.

[9] Wang, X., Li, Y., Qiao, Q., Tavares, A., & Liang, Y. (2023). Water quality prediction based on machine learning and comprehensive weighting methods. *Entropy, 25*(8), 1186.

[10] Patel, S., Shah, K., Vaghela, S., Aglodiya, M., & Bhattad, R. (2023). Water Potability Prediction Using Machine Learning.

[11] Brindha, D., Puli, V., NVSS, B. K. S., Mittakandala, V. S., & Nanneboina, G. D. (2023, February). Water quality analysis and prediction using machine learning. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 175-180). IEEE.