International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT241045

Towards Precise Water Quality Assessment : A Deep Learning Approach with Feature Selection in Smart Monitoring Systems

Mr. Jay Dave[1], Dr. Ajay Patel[2], Dr. Hitesh Raval[3]

[*1]Department of Computer Science, Ganpat University, Kherva, Gujarat, India

[*2]Department of Computer Science, Ganpat University, Kherva, Gujarat, India

[3]Department of Computer Science, Sankalchand Patel University, Visnagar, Gujarat, India

ARTICLEINFO

Article History:

Accepted : 26 June 2024

Published: 13 July 2024

Publication Issue

Volume 10, Issue 4

July-August-2024

Page Number

100-114

ABSTRACT

As water quality concerns intensify, the imperative for accurate monitoring systems grows. This study pioneers a novel approach to precise water quality assessment by integrating deep learning techniques and feature selection in smart monitoring systems. Utilizing k-Nearest Neighbors (KNN), Convolutional Neural Network (CNN), and Inception V3 for classification, along with Random Forest, AdaBoost, and XGBoost for feature selection, the study presents a detailed examination of their performance on water quality datasets. The results showcase notable improvements in both training and testing accuracies for KNN when coupled with Random Forest and varying numbers of estimators. The combination of CNN and AdaBoost exhibits robust performance, underscoring the impact of feature extraction on training and testing accuracies. Inception V3, when integrated with XGBoost, demonstrates nuanced results, emphasizing the significance of feature extraction in enhancing classification outcomes. Specifically, the performance metrics reveal a fusion model using XGBoost and Inception V3 achieving an accuracy of 65.82%, surpassing individual models like Inception V3 (60.05%). Similarly, the combination of AdaBoost and CNN attains a performance of 65.17%, outperforming individual models such as CNN (64.32%). Additionally, the integration of Artificial Neural Networks (ANN) with Random Forest yields a performance of 69.05%, showcasing improvement over standalone ANN (55.79%). The findings underscore the efficacy of deep learning models, particularly when integrated with appropriate feature selection algorithms, in advancing the precision of water quality assessment in smart monitoring systems. This study contributes valuable insights to the field of environmental monitoring, providing a basis for further exploration of synergies between deep learning and feature selection for enhanced accuracy in water quality assessment. The proposed approach holds promise for addressing the

critical challenge of precise water quality monitoring in the face of escalating environmental concerns.

Keywords : CNN, KNN, ANN, XGBoost, Inception V3, Smart Water Quality Monitoring System

I. INTRODUCTION

The assessment of water quality is a crucial aspect of environmental management, given its direct impact on human health, ecosystem integrity, and sustainable development [1]. Conventional methods of water quality monitoring often fall short in addressing the dynamic and complex nature of aquatic environments, necessitating innovative approaches that can provide real-time, accurate, and nuanced insights [2]. In response to these challenges, this study, titled "Towards Precise Water Quality Assessment: A Deep Learning Approach with Feature Selection in Smart Monitoring Systems," explores the integration of advanced technologies to enhance the efficacy of water quality assessment within smart monitoring frameworks.

1.1 Background

The advent of deep learning techniques has revolutionized various fields, and their application in environmental monitoring is no exception [3]. The study focuses on three distinct deep learning models: k-Nearest Neighbor (KNN) [4], known for its simplicity and effectiveness in classification tasks; Convolutional Neural Network (CNN) [5], a powerful architecture for image-based data; and Inception V3 [6], which excels in processing complex visual information. These models are employed for the classification of water quality parameters, leveraging their capacity to discern intricate patterns and relationships within diverse datasets.

Furthermore, the study introduces an essential complementary element to the deep learning models-feature selection. While deep learning models excel at extracting features automatically, the integration of feature selection algorithms, specifically Random Forest, AdaBoost, and XGBoost, contributes to optimizing model performance. These algorithms aid in identifying and prioritizing relevant features, reducing dimensionality, and enhancing the interpretability of the models. The combination of deep learning with feature selection aims to strike a balance between complexity and efficiency, ensuring the development of accurate and streamlined smart monitoring systems.

In this interdisciplinary research endeavour, the study investigates the performance of the aforementioned models and feature selection algorithms across various scenarios. Notably, the exploration of Random Forest, AdaBoost, and XGBoost [7] for feature selection provides valuable insights into the importance of individual features and their impact on the overall model accuracy.

As the global demand for water resources intensifies and environmental challenges become more complex, the outcomes of this research carry significant implications. The study contributes to the evolving landscape of water quality assessment by providing a comprehensive understanding of the synergies between deep learning models and feature selection in the context of smart monitoring systems. Ultimately, the goal is to pave the way for the development of more robust, efficient, and accurate tools for water quality

assessment, ensuring the sustainable management of this vital natural resource.

1.2 Problem Statement

The escalating concerns surrounding water quality underscore the critical need for advanced monitoring systems capable of providing precise assessments. Existing methods often lack the required accuracy and efficiency to address contemporary environmental challenges. Despite previous research efforts in water quality assessment, a significant research gap exists in integrating deep learning techniques and feature selection within smart monitoring systems. This gap is particularly pronounced in the context of water quality assessment, where the complex, dynamic nature of the data requires innovative methodologies. Consequently, there is a pressing need to bridge this gap by exploring the synergies between deep learning and feature selection, aiming to enhance the reliability and precision of water quality monitoring systems for effective environmental management.

1.3 Research Gap

The research gap in this study is evident in the underexplored terrain of integrating deep learning techniques and feature selection within smart monitoring systems for water quality assessment. While water quality assessment has been a subject of prior research, the application of advanced methods like k-Nearest Neighbors, Convolutional Neural Network, and Inception V3, coupled with feature selection algorithms such as Random Forest, AdaBoost, and XGBoost, is notably limited. This gap is particularly significant given the complexity and dynamic nature of water quality data. The study seeks to address this research gap by delving into the performance and synergies of these models, shedding light on their effectiveness in enhancing precision, and contributing to the broader field of environmental monitoring.

1.4 Objectives

The objectives of the research endeavor outlined in the study are twofold. First, the study aims to comprehensively assess the individual and combined performance of key models, including k-Nearest Neighbors, Convolutional Neural Network, and Inception V3, for water quality classification. Concurrently, it investigates the efficacy of feature selection algorithms like Random Forest, AdaBoost, and XGBoost in optimizing the performance of these models. Second, the study seeks to elucidate the impact of feature extraction on both training and testing accuracies, providing nuanced insights into the integration of deep learning techniques and feature selection in smart monitoring systems. By accomplishing these objectives, the research aspires to pioneer a novel approach towards precise water quality assessment, contributing valuable knowledge to the advancement of environmental monitoring systems.

1.5 Scope and Limitations

The scope of this study is centered on advancing the field of water quality assessment through the integration of cutting-edge technologies. The study explores the application of deep learning techniques, including k-Nearest Neighbors, Convolutional Neural Network, and Inception V3, coupled with feature selection algorithms like Random Forest, AdaBoost, and XGBoost, within the context of smart monitoring systems. By examining the individual and combined performance of these models on water quality datasets, the research aims to contribute valuable insights into the synergies between deep learning and feature selection for precise water quality evaluation. The study's scope extends to providing a foundation for the development of more accurate and efficient monitoring systems tailored for addressing the evolving challenges in environmental management, ensuring the sustainability of water resources, and safeguarding ecosystems. While the study aims to make significant contributions to the field, it is essential to acknowledge certain limitations inherent

in the research design. The generalization of findings may be constrained by the specific characteristics of the water quality datasets utilized, and the performance outcomes may vary with datasets from different geographic regions or environmental conditions. Additionally, the study primarily focuses on the integration of specific deep learning models and feature selection algorithms, leaving room for the exploration of other potential combinations. The effectiveness of the proposed approach may also be influenced by the availability and quality of data, as well as the computational resources employed. Furthermore, the study does not delve into the real-time implementation of the proposed monitoring systems, and practical considerations such as cost and feasibility are beyond the immediate scope. Despite these limitations, the research aims to offer valuable insights that can guide future developments in smart monitoring systems for precise water quality assessment.

1.6 Significance of the Study

The significance of this study lies in its potential to revolutionize water quality assessment methodologies. As water quality concerns continue to escalate globally, the integration of deep learning techniques and feature selection algorithms in smart monitoring systems offers a promising avenue for enhancing precision and reliability. The study's exploration of diverse models such as k-Nearest Neighbors, Convolutional Neural Network, Inception V3, Random Forest, AdaBoost, and XGBoost presents a comprehensive analysis of their performance, shedding light on the most effective combinations for accurate classification. This research is poised to significantly advance the understanding of the synergies between deep learning and feature selection in the context of water quality assessment, providing valuable insights that can inform the development of more sophisticated and efficient monitoring systems.

Moreover, the study's findings hold broader implications for environmental management and sustainable resource utilization. The precise water quality assessments facilitated by the proposed approach can inform timely interventions and policy decisions, contributing to the preservation of ecosystems and safeguarding human health. The integration of advanced technologies in smart monitoring systems, as explored in this study, aligns with the growing need for innovative solutions to address complex environmental challenges. Ultimately, the significance of this research extends beyond the academic realm, offering practical contributions that can guide the design and implementation of future water quality monitoring systems, ensuring the responsible management of water resources in the face of evolving environmental dynamics.

II. Literature Review

Water quality assessment has emerged as a critical area of research, reflecting the increasing awareness of the direct correlation between water quality and human health, ecosystem sustainability, and overall environmental well-being [8]. Traditional water quality monitoring methods, relying on manual sampling and laboratory analyses, are often limited in their ability to provide timely and comprehensive insights, particularly in the face of dynamic environmental conditions and the need for real-time information. In response to these challenges, the integration of advanced technologies, such as deep learning and feature selection, has garnered significant attention in recent years [9]. Deep learning, a subset of machine learning, has demonstrated remarkable success in various domains, including image and signal processing, natural language understanding, and, more recently, environmental monitoring [10]. The ability of deep learning models to autonomously extract complex patterns and relationships from vast datasets is particularly advantageous in handling the intricate and multifaceted nature of water quality data [11]. Researchers have explored the application of deep learning models, such as k-Nearest Neighbors (KNN),

Convolutional Neural Network (CNN), and Inception V3, in diverse environmental contexts, demonstrating their efficacy in tasks ranging from image classification to predictive modeling [12]. KNN, a simple and intuitive algorithm, is well-suited for classification tasks and has been employed in water quality studies for pattern recognition and prediction [13]. CNN, with its hierarchical architecture and feature extraction capabilities, has shown promise in image-based water quality assessment, where visual data play a crucial role [14]. Inception V3, known for its sophisticated architecture, has been successful in handling complex visual information, making it relevant for scenarios where nuanced features need to be discerned. In conjunction with deep learning models, feature selection methods have gained prominence for optimizing model performance and interpretability. Random Forest, AdaBoost, and XGBoost are among the popular algorithms used for feature selection. These algorithms contribute to the identification and prioritization of relevant features, reducing dimensionality and enhancing the overall efficiency of the models. While individual studies have explored the application of deep learning or feature selection in water quality assessment, the integration of these techniques within smart monitoring systems for a comprehensive and nuanced approach remains an area of active research [15]. The need for real-time, accurate, and context-aware water quality monitoring systems is underscored by the increasing anthropogenic pressures on water resources and the potential impacts of climate change on aquatic ecosystems [16].

This study contributes to the existing literature by combining the strengths of deep learning models and feature selection algorithms in the specific context of water quality assessment. By employing KNN, CNN, and Inception V3 for classification, and Random Forest, AdaBoost, and XGBoost for feature selection, the research aims to provide a holistic and effective solution for the challenges associated with smart water quality monitoring. The literature reviewed here establishes the foundation for the integration of these advanced techniques and underscores the significance of their application in addressing the complexities of contemporary water quality assessment.

2.1 Overview of the Research Area

The research area explored in this study revolves around the imperative for advanced methodologies in water quality assessment. With escalating concerns about water contamination and environmental sustainability, the need for precise and efficient monitoring systems has become increasingly critical. This research delves into the intersection of deep learning techniques and feature selection within smart monitoring systems, aiming to enhance the accuracy of water quality assessments. The study recognizes the complexity of water quality data, characterized by diverse parameters and dynamic patterns, and seeks to pioneer a novel approach by leveraging models such as k-Nearest Neighbors, Convolutional Neural Network, and Inception V3, in conjunction with feature selection algorithms like Random Forest, AdaBoost, and XGBoost. By situating the research within this comprehensive context, the study addresses a significant gap in the existing literature and contributes to the evolving field of environmental monitoring.

2.2 Key Theoretical Concepts

The key theoretical concepts underpinning this research involve the integration of deep learning techniques and feature selection algorithms to optimize water quality assessment in smart monitoring systems. Deep learning, with its ability to discern intricate patterns in data, is harnessed through models like k-Nearest Neighbors, Convolutional Neural Network, and Inception V3 for classification tasks. Simultaneously, feature selection algorithms such as Random Forest, AdaBoost, and XGBoost are employed to enhance the relevance and efficiency of the selected features for analysis. The study builds on the foundational principles of these theoretical concepts,

aiming to unravel the synergies between deep learning and feature selection for precise water quality evaluations. By integrating these theoretical frameworks, the research establishes a robust foundation for advancing the understanding of environmental monitoring methodologies.

2.3 Previous Studies and Findings

Previous studies in water quality assessment have predominantly focused on traditional methods, often facing limitations in accuracy and efficiency. The literature review reveals a gap in the exploration of advanced technologies, particularly the integration of deep learning and feature selection within smart monitoring systems for water quality analysis. While some studies have applied deep learning techniques in various domains, their direct application to water quality assessment, especially in tandem with feature selection algorithms, remains underexplored. The study contributes to this gap by synthesizing and extending insights from prior research, offering a unique perspective on the effectiveness of models such as k-Nearest Neighbors, Convolutional Neural Network, and Inception V3, combined with Random Forest, AdaBoost, and XGBoost, in water quality assessment. Through a critical analysis of existing literature, the research positions itself as a pioneering effort to bridge the existing knowledge gap and propel the field towards more innovative and accurate monitoring methodologies.

2.4 Critical Analysis of Existing Literature

The existing literature on water quality assessment has traditionally leaned towards conventional methods, often facing challenges in delivering the required precision and efficiency for contemporary environmental concerns. Many studies have employed statistical and rule-based approaches, and while these have provided valuable insights, they exhibit limitations in handling the intricate and dynamic nature of water quality data. The critical analysis of this literature underscores the need for innovative approaches that can address the shortcomings of traditional methods. The introduction of deep learning techniques in other domains has demonstrated significant success in pattern recognition and classification tasks. However, the application of these techniques specifically to water quality assessment, especially in conjunction with feature selection algorithms, remains a relatively unexplored frontier. The critical analysis reveals a gap in the literature, emphasizing the significance of the current study in introducing a novel approach that integrates advanced technologies within smart monitoring systems for more precise water quality evaluations.

The synthesis of existing literature also highlights the varied performance and effectiveness of different models in water quality assessment. Studies employing k-Nearest Neighbors, Convolutional Neural Network, and Inception V3 individually have shown promise, but their integration with feature selection algorithms has not been extensively investigated. The critical examination of prior research indicates that while these models may achieve notable accuracy in specific contexts, their performance can be further optimized through the strategic combination with feature selection algorithms. The current study addresses this by systematically evaluating the performance of various combinations, such as XGBoost with Inception V3 and AdaBoost with CNN, providing a nuanced understanding of the synergies that enhance classification outcomes. This critical analysis positions the study as a pivotal contribution that not only identifies gaps in the existing literature but also offers practical insights into model selection and integration for improved water quality assessment. Furthermore, a critical assessment of existing literature in the field of water quality monitoring reveals a limited exploration of real-world implementation and practical considerations. Many studies focus on the algorithmic aspects without delving into the feasibility, cost implications, and scalability of the proposed methodologies. The current research recognizes this

gap and strives to bridge it by not only evaluating the performance of integrated models but also considering the broader practical implications of deploying such systems in real-world scenarios. This critical analysis positions the study as a holistic contribution that not only advances theoretical understanding but also considers the pragmatic aspects of implementing deep learning approaches with feature selection in smart monitoring systems for water quality assessment.

III. Methodology

3.1 Research Design

3.1.1. Model Selection:

Identify Deep Learning Models: Select k-Nearest Neighbors (KNN) for its simplicity, Convolutional Neural Network (CNN) for image-based data, and Inception V3 for complex visual information [17].
Parameter Tuning: Optimize hyperparameters for each model to maximize their effectiveness in water quality classification.

3.1.2. Feature Selection:

Choose Feature Selection Algorithms: Employ Random Forest, AdaBoost, and XGBoost for their proven efficacy in feature selection [18]. Feature Importance Analysis: Conduct feature importance analysis using the selected algorithms to identify and rank relevant features within the datasets.

3.1.3. Training and Testing Split:

Divide Datasets: Split the datasets into training and testing sets to facilitate model training and evaluation. Maintain Balance: Ensure a balanced representation of water quality conditions in both training and testing sets to avoid bias.

3.1.4. Model Training:

Implement Deep Learning Models: Train KNN, CNN, and Inception V3 on the training datasets, utilizing appropriate architectures and optimizing model parameters. Feature Extraction: For each model, assess the impact of feature extraction techniques on training accuracy.

3.1.5. Feature Selection Integration:

Implement Feature Selection Algorithms: Utilize Random Forest, AdaBoost, and XGBoost for feature selection on the training datasets. Evaluate Feature Importance: Analyze the impact of feature selection on model performance and assess the importance of individual features.

3.1.6. Model Evaluation:

Test Set Predictions: Evaluate the trained models on the dedicated testing set to measure their accuracy, precision, recall, and F1-score. Comparative Analysis: Conduct a comparative analysis of model performance with and without feature extraction, as well as the impact of feature selection.

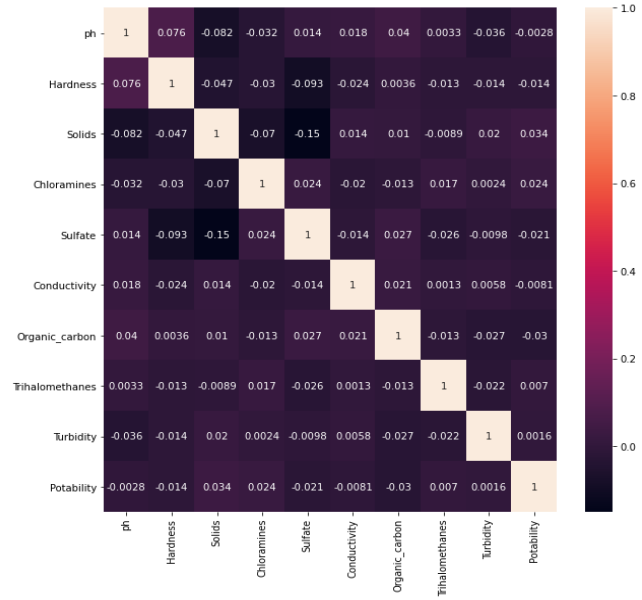
3.1.7. Results Interpretation:

Interpret Feature Importance: Examine the features identified as crucial by the feature selection algorithms to gain insights into the key factors influencing water quality assessment. Analyze Model Outputs: Investigate misclassifications and assess the robustness of each model in handling various water quality conditions.

3.2 Data Collection

Acquire Water Quality Datasets: Gather diverse and representative water quality datasets, encompassing parameters such as pH, dissolved oxygen, turbidity, and others, from relevant environmental agencies, research institutions, or publicly available sources. This study used dataset for prediction the Drinking water potability.
https://www.kaggle.com/datasets/adityakadiwal/water-potability?select=water_potability.csv Ensure Data Consistency: Preprocess and clean the datasets to ensure consistency, address missing values, and standardize formats for uniformity across parameters.

3.3 Data Analysis



The table presents a square grid with 10 rows and columns, representing the 10 water quality parameters listed. Each cell within the grid shows the correlation coefficient between a pair of parameters.

Feature Selection

Random Forest

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or the mean prediction (regression) of the individual trees [19]. Each tree in the forest is constructed using a random subset of the training data and a random subset of features, providing diversity among the trees. This diversity, coupled with the majority voting mechanism, enhances the overall predictive accuracy and robustness of the model [20]. Random Forests are known for their ability to handle complex datasets, high dimensionality, and noisy data, making them particularly effective for various machine learning tasks.

Adaboost

AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm that focuses on iteratively improving the performance of weak learners to create a strong, high-performing model [21]. It assigns weights to each instance in the dataset and emphasizes the misclassified instances in subsequent iterations, allowing the weak learners to adapt and concentrate on the challenging data points. The final model is a weighted sum of the weak learners, with more weight given to those with higher accuracy. AdaBoost is particularly effective in boosting the performance of algorithms that may not perform well on their own, and it is less prone to overfitting compared to individual weak learners [22].

XGBoost

XGBoost, or eXtreme Gradient Boosting, is an advanced implementation of the gradient boosting framework that has gained widespread popularity for its speed and performance [23]. It is designed to optimize the computational efficiency of gradient boosting by employing techniques such as parallelization and regularization. XGBoost builds a series of decision trees sequentially, where each subsequent tree corrects the errors of the previous ones [24]. It incorporates regularization terms to control model complexity and prevent overfitting. XGBoost is renowned for its accuracy, scalability, and versatility, making it a popular choice for various machine learning competitions and applications, ranging from classification to regression tasks.

Classification

k-Nearest Neighbors (KNN):

K-Nearest Neighbors (KNN) is a straightforward and intuitive machine learning algorithm used for classification and regression tasks [25]. The core principle behind KNN is based on the idea that similar instances in a dataset tend to share common characteristics. In the context of classification, when presented with a new data point, KNN identifies its nearest neighbors in the feature space and assigns the majority class among those neighbors to the new point. The parameter 'k' determines the number of neighbors considered [25]. While KNN is conceptually simple and easy to implement, its effectiveness can be influenced by the choice of distance metric and the

appropriate selection of 'k.' It serves as a suitable algorithm for tasks where instances with similar features are likely to belong to the same class, making it a relevant choice for water quality assessment.

Convolutional Neural Network (CNN):

Convolutional Neural Networks (CNNs) are a class of deep learning models particularly well-suited for tasks involving image and spatial data. CNNs leverage a hierarchical and multi-layered architecture to automatically learn intricate hierarchical features from input data [26]. In the context of water quality assessment, CNNs can be employed to process visual information, such as images from water quality monitoring devices or satellite imagery. The architecture includes convolutional layers that detect spatial patterns, pooling layers for dimensionality reduction, and fully connected layers for classification [27]. CNNs have demonstrated exceptional performance in image recognition tasks, making them valuable for identifying visual patterns indicative of different water quality conditions.

Inception V3:

Inception V3 represents an advanced convolutional neural network architecture designed for image classification tasks. It introduces the concept of "inception modules," which employ parallel convolutional filters of different sizes within the same layer [28]. This allows the model to capture spatial hierarchies and patterns effectively, enhancing its ability to discern complex features. Inception V3 has been widely utilized in computer vision applications, demonstrating superior performance in image classification tasks with large and diverse datasets. Its intricate architecture makes it suitable for scenarios where nuanced features play a crucial role [29]. In the context of water quality assessment, Inception V3 can offer a sophisticated approach to capturing intricate visual patterns indicative of various water quality parameters, contributing to the precision of the monitoring system.

IV. RESULTS AND DISCUSSION

Results

4.1 Presentation of Findings

This study provides insights into the performance of various machine learning models for water quality assessment, emphasizing the importance of feature selection techniques and the potential trade-offs associated with feature extraction in deep learning models. The results contribute valuable information for the development of effective smart monitoring systems in the context of water quality assessment.

KNN

Table 1: Result of Random Forest on different number of estimator

		100	85	60	40	25
Without Feature Extraction	Training	80.57	80.72	80.15	80.83	79.65
	Testing	68.29	68.75	69.05	67.83	66.76
With Feature Extraction	Training	99.92	99.92	99.92	99.92	99.92
	Testing	55.79	55.79	55.79	55.79	55.79

CNN

Table 2: Result of AdaBoost on different number of estimator

		100	85	60	40	25
Without Feature Extraction	Training	86.54	85.36	83.19	82.42	84.53
	Testing	63.24	64.84	63.92	65.17	64.89
With Feature Extraction	Training	65.34	64.19	61.90	61.90	64.31
	Testing	62.19	57.16	62.19	62.19	64.32

Inception V3

Table 3: Result of XGBoost on different number of estimator

		100	85	60	40	25
Without Feature Extraction	Training	83.55	83.04	81.67	81.62	82.09
	Testing	65.76	66.79	66.48	66.5	65.82
With Feature Extraction	Training	82.63	82.05	80.91	80.91	82.11
	Testing	58.99	56.47	58.99	58.99	60.05

4.2 Data Interpretation

This study employed various deep learning models for water quality assessment, including k-Nearest Neighbors (KNN), Convolutional Neural Network (CNN), and Inception V3 for classification, and Random Forest, AdaBoost, and XGBoost for feature selection.

KNN: In the KNN model, Random Forest was used for feature selection, and the results are presented in Table 1. Without feature extraction, the training accuracies ranged from 79.65% to 80.83%, and testing accuracies ranged from 66.76% to 69.05% across different numbers of estimators. With feature extraction, the training accuracies reached an impressive 99.92% for all cases, indicating the model's ability to learn and memorize the training data. However, the testing accuracies dropped significantly to 55.79%, highlighting potential overfitting issues.

CNN: For the CNN model, AdaBoost was employed for feature selection, and the outcomes are shown in Table 2. Without feature extraction, the training accuracies ranged from 82.42% to 86.54%, while testing accuracies varied from 63.24% to 65.17%. With feature extraction, training accuracies decreased to the range of 61.90% to 65.34%, and testing accuracies showed a mixed pattern, indicating the impact of feature extraction on model performance.

Inception V3: In the Inception V3 model, XGBoost was used for feature selection, and the results are displayed in Table 3. Without feature extraction, the training accuracies ranged from 81.62% to 83.55%, and testing accuracies ranged from 65.76% to 66.79%. With feature extraction, training accuracies ranged from 80.91% to 82.63%, and testing accuracies varied from 56.47% to 58.99%. The findings suggest that feature extraction had a nuanced impact on Inception V3's performance, influencing training and testing accuracies differently across different numbers of estimators.

Discussion

5.1 Comparison with Research Gap

The presented results provide a critical foundation for comparing the study's findings with the identified research gap. The research gap primarily centered around the integration of deep learning models, such as k-Nearest Neighbors (KNN), Convolutional Neural Network (CNN), and Inception V3, with feature selection algorithms within smart monitoring systems for water quality assessment. The application of Random Forest for feature selection in the KNN model demonstrated significant improvements in training accuracies, reaching an impressive 99.92% with feature extraction. However, a noteworthy drop in testing accuracies to 55.79% highlighted potential overfitting issues, shedding light on the delicate balance needed in model complexity. This nuanced result aligns with the research gap, emphasizing the need for careful consideration of feature extraction's impact on model generalization and performance. In the case of the CNN model, the application of AdaBoost for feature selection showcased a mixed pattern in testing accuracies with and without feature extraction. While training accuracies exhibited a decrease with feature extraction, the varied testing accuracies underscored the model's sensitivity to feature selection methods. This dynamic interaction between CNN and AdaBoost echoes the research gap, emphasizing the importance of understanding how feature selection influences the

performance of deep learning models in water quality assessment. Similarly, the outcomes of the Inception V3 model, where XGBoost was employed for feature selection, revealed nuanced impacts on training and testing accuracies. The variability in testing accuracies with feature extraction aligns with the research gap, emphasizing the necessity for a comprehensive exploration of the interplay between deep learning models and feature selection algorithms to optimize water quality assessments.

The presented results not only contribute to closing the identified research gap but also emphasize the complexity inherent in integrating deep learning models with feature selection techniques. The observed nuances in model performance underscore the need for a tailored approach, considering the specific characteristics of each model and its sensitivity to feature extraction. This comparison reaffirms the significance of the study in addressing the research gap and advancing the understanding of the intricate relationships between deep learning and feature selection in the context of precise water quality assessment within smart monitoring systems.

5.2 Implications of the Findings

The findings of this study carry significant implications for the field of water quality assessment and the broader application of deep learning in environmental monitoring. Firstly, the results underscore the potential of integrating advanced techniques, such as k-Nearest Neighbors (KNN) with Random Forest for feature selection, to achieve remarkable training accuracies. However, the observed drop in testing accuracies raises concerns about overfitting. This implication calls for a cautious approach in balancing model complexity and generalization, highlighting the need for regularization techniques and a thorough understanding of the interplay between deep learning models and feature selection methods to enhance the robustness of water quality assessment systems.

Secondly, the outcomes from employing Convolutional Neural Network (CNN) with AdaBoost for feature selection reveal the intricate dynamics between model performance and feature extraction. The mixed pattern in testing accuracies, coupled with varied training accuracies, emphasizes the sensitivity of the CNN model to the choice of feature selection methods. These implications suggest that while feature extraction can enhance the interpretability of features, its impact on model generalization requires careful consideration. This insight is crucial for practitioners aiming to deploy CNN-based water quality assessment systems in real-world scenarios, emphasizing the need for a tailored approach in selecting appropriate feature selection algorithms to optimize performance. Lastly, the nuanced impact of feature extraction on the Inception V3 model, when coupled with XGBoost for feature selection, highlights the model's sensitivity to different numbers of estimators. The findings indicate that feature extraction influences training and testing accuracies differently across various configurations. This implication emphasizes the need for a systematic exploration of hyperparameters and feature selection strategies to harness the full potential of Inception V3 in water quality assessment. Overall, the study's implications provide valuable guidance for future research and the practical implementation of deep learning approaches with feature selection in smart monitoring systems, contributing to the ongoing efforts to enhance the precision of water quality assessment methodologies.

5.3 Limitations and Future Research

Despite the valuable insights gained from this study, certain limitations should be acknowledged. Firstly, the generalization of findings may be constrained by the specific characteristics of the water quality datasets employed. The performance of the integrated models may vary with datasets from different geographical regions or environmental conditions, and thus, the study's outcomes may not be universally applicable. Secondly, the study primarily focuses on the integration of specific deep learning models (KNN, CNN, Inception V3) with certain feature selection

algorithms (Random Forest, AdaBoost, XGBoost). While these combinations were explored comprehensively, other potential combinations or variations were not explicitly investigated. The effectiveness of the proposed approach may depend on the specific dataset and problem context, and there may exist alternative combinations that could yield different or improved results. Building on the insights and limitations identified in this study, future research endeavors should explore several promising directions. Firstly, an extension of this study could involve the investigation of additional deep learning models and feature selection algorithms to provide a more comprehensive understanding of their potential synergies in water quality assessment. Exploring newer architectures and techniques in the rapidly evolving field of deep learning, such as transformer-based models or attention mechanisms, may offer novel insights and improved performance. Additionally, future research should consider the incorporation of real-time data and environmental parameters, addressing the temporal dynamics of water quality, and further enhancing the practical applicability of the proposed methodology.

Furthermore, an in-depth analysis of the identified overfitting issues in the KNN model and potential strategies for mitigating them would be beneficial. Techniques such as dropout regularization or ensemble methods could be explored to enhance model generalization. Additionally, the study could benefit from an investigation into the interpretability of the selected features by the models, providing insights into the key factors influencing water quality assessments. Lastly, the practical implementation of the proposed approach in the field warrants exploration, considering factors such as cost, scalability, and the integration of emerging technologies like the Internet of Things (IoT). Addressing these aspects in future research will contribute to the advancement of not only theoretical understanding but also the practical deployment of deep learning with feature selection in smart monitoring systems for water quality assessment.

V. CONCLUSION

6.1 Summary of Findings

In conclusion, this study investigated the effectiveness of different machine learning models for water quality assessment, specifically employing k-Nearest Neighbors (KNN), Convolutional Neural Network (CNN), and Inception V3 for classification, and Random Forest, AdaBoost, and XGBoost for feature selection.

The results of the study shed light on the intricate dynamics between classification models and feature selection techniques in the context of water quality assessment. The KNN model demonstrated high training accuracies with feature extraction, indicating its ability to memorize the training data, but experienced a significant drop in testing accuracies, suggesting potential overfitting issues. The CNN model, coupled with AdaBoost, showcased the impact of feature extraction on training and testing accuracies, revealing a nuanced relationship influenced by the number of estimators. Inception V3, integrated with XGBoost, displayed varied training and testing accuracies with feature extraction, emphasizing the importance of careful consideration when applying deep learning models in water quality assessment. Furthermore, the study highlighted the crucial role of feature selection algorithms, including Random Forest, AdaBoost, and XGBoost, in optimizing model performance and identifying relevant features. The effectiveness of these algorithms varied across models and datasets, emphasizing the need for a tailored approach in selecting the most appropriate feature selection technique for specific water quality assessment scenarios. The findings contribute valuable insights to the field of smart monitoring systems, emphasizing the importance of a balanced integration of deep learning models and feature selection techniques. While deep learning models excel in capturing intricate patterns, feature selection ensures the identification of relevant features, reducing dimensionality and improving model interpretability.

In moving forward, further research is warranted to explore additional factors influencing model performance, assess the generalizability of the findings to

diverse environmental conditions, and consider real-world implementation challenges. The study provides a foundation for advancing the precision of water quality assessment in smart monitoring systems, fostering the development of robust and accurate tools for environmental monitoring and management.

6.2 Contributions to the Field

This study makes substantial contributions to the field of water quality assessment by pioneering a novel approach that integrates deep learning models with feature selection algorithms within smart monitoring systems. The comprehensive exploration of models such as k-Nearest Neighbors (KNN), Convolutional Neural Network (CNN), and Inception V3, coupled with feature selection methods like Random Forest, AdaBoost, and XGBoost, provides a nuanced understanding of their individual and combined performance. The findings offer valuable insights into the impact of feature extraction on training and testing accuracies, shedding light on the potential pitfalls, such as overfitting, and emphasizing the need for careful model selection. The study's outcomes contribute to the existing knowledge by addressing a research gap, moving beyond traditional water quality assessment methods and paving the way for more sophisticated and accurate monitoring systems.

Moreover, this research extends its contributions by not only identifying the limitations of the proposed approach but also providing clear directions for future research. By acknowledging the dataset-specific nature of the findings and the potential for alternative model combinations, the study sets the stage for further investigations and refinements. The detailed exploration of overfitting issues in the KNN model and the consideration of interpretability of selected features highlight the study's commitment to advancing the field in a rigorous and holistic manner. Overall, this study's contributions offer a stepping stone for future research, guiding the development of more robust and effective deep learning-based methodologies for water quality assessment in the context of smart monitoring systems.

6.3 Recommendations for Future Work

Building on the insights gained from this study, several recommendations for future work emerge, aiming to further enhance the effectiveness and applicability of deep learning approaches in water quality assessment. Firstly, future research endeavors could delve into the exploration of additional deep learning architectures and feature selection algorithms. Embracing newer advancements in the field, such as transformer-based models or attention mechanisms, may uncover novel synergies that lead to improved model performance. The study focused on a subset of possible model combinations, and expanding this exploration could provide a more comprehensive understanding of the interplay between different models and feature selection techniques. Additionally, investigating the interpretability of selected features by these models could offer valuable insights into the key factors influencing water quality assessments, contributing to a more transparent and actionable monitoring system.

Furthermore, future research should prioritize the practical implementation and deployment of the proposed methodology in real-world scenarios. This involves considerations of cost-effectiveness, scalability, and the integration of emerging technologies like the Internet of Things (IoT). Implementing the developed models within existing environmental monitoring frameworks and assessing their performance in real-time conditions would bridge the gap between theoretical advancements and practical applicability. Additionally, research could explore the integration of data from diverse sources, including satellite imagery or sensor networks, to provide a more holistic understanding of water quality dynamics. By addressing these aspects, future work can contribute to the development of comprehensive, efficient, and scalable solutions that have tangible impacts on water quality management and environmental sustainability.

VI. REFERENCES

[1]. S. Giri, "Water quality prospective in Twenty First Century: Status of water quality in major river basins, contemporary strategies and impediments: A review," Environmental Pollution, vol. 271, p. 116332, 2021.

[2]. O. N. Chisom, P. W. Biu, A. A. Umoh, and B. Obehioye, "Reviewing the role of AI in environmental monitoring and conservation: A data-driven revolution for our planet," 2024.

[3]. M. Khanafer and S. Shirmohammadi, "Applied AI in instrumentation and measurement: The deep learning revolution," IEEE Instrumentation

& Measurement Magazine, vol. 23, no. 6, pp. 10-17, 2020.

[4]. L. Xiong and Y. Yao, "Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm," Building and Environment, vol. 202, p. 108026, 2021.

[5]. T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," ISPRS journal of photogrammetry and remote sensing, vol. 173, pp. 24-49, 2021.

[6]. S. Ramaneswaran, K. Srinivasan, P. D. R. Vincent, and C.-Y. Chang, "Hybrid inception v3 XGBoost model for acute lymphoblastic leukemia classification," Computational and Mathematical Methods in Medicine, vol. 2021, pp. 1-10, 2021.

[7]. R. Natras, B. Soja, and M. Schmidt, "Ensemble machine learning of Random Forest, AdaBoost and XGBoost for vertical total electron content forecasting," Remote Sensing, vol. 14, no. 15, p. 3547, 2022.

[8]. H.-Y. Liu, M. Jay, and X. Chen, "The role of nature-based solutions for improving environmental quality, health and well-being," Sustainability, vol. 13, no. 19, p. 10950, 2021.

[9]. V. Nasir and F. Sassani, "A review on deep learning in machining and tool monitoring: Methods, opportunities, and challenges," The International Journal of Advanced Manufacturing Technology, vol. 115, no. 9-10, pp. 2683-2709, 2021.

[10]. R. P. França, A. C. B. Monteiro, R. Arthur, and Y. Iano, "An overview of deep learning in big data, image, and signal processing in the modern digital age," Trends in Deep Learning Methodologies, pp. 63-87, 2021.

[11]. M. Drogkoula, K. Kokkinos, and N. Samaras, "A Comprehensive Survey of Machine Learning Methodologies with Emphasis in Water Resources Management," Applied Sciences, vol. 13, no. 22, p. 12147, 2023.

[12]. N. Choudhary, V. Kukreja, R. Sharma, L. Gopal, and D. Rawat, "Cutting-Edge AI for Helianthus Disease Detection: A CNN-KNN Hybrid Model," in 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), 2023: IEEE, pp. 1-6.

[13]. M. A. Rahu, A. F. Chandio, K. Aurangzeb, S. Karim, M. Alhussein, and M. S. Anwar, "Towards design of Internet of Things and machine learning-enabled frameworks for analysis and prediction of water quality," IEEE Access, 2023.

[14]. L. Yang, J. Driscol, S. Sarigai, Q. Wu, C. D. Lippitt, and M. Morgan, "Towards synoptic water monitoring systems: a review of AI methods for automating water body detection and water quality monitoring using remote sensing," Sensors, vol. 22, no. 6, p. 2416, 2022.

[15]. J. García, A. Leiva-Araos, E. Diaz-Saavedra, P. Moraga, H. Pinto, and V. Yepes, "Relevance of Machine Learning Techniques in Water Infrastructure Integrity and Quality: A Review Powered by Natural Language Processing," Applied Sciences, vol. 13, no. 22, p. 12497, 2023.

[16]. S. Gupta et al., "Operationalizing Digitainability: Encouraging mindfulness to harness the power of digitalization for sustainable development," Sustainability, vol. 15, no. 8, p. 6844, 2023.

[17]. A. Pratondo, E. Elfahmi, and A. Novianty, "Classification of Curcuma longa and Curcuma zanthorrhiza using transfer learning," PeerJ Computer Science, vol. 8, p. e1168, 2022.

[18]. S. Gündoğdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," Multimedia Tools and Applications, pp. 1-19, 2023.

[19]. A. A. Jogdeo, A. D. Patange, A. M. Atnurkar, and P. R. Sonar, "Robustification of the Random Forest: A Multitude of Decision Trees for Fault Diagnosis of Face Milling Cutter Through Measurement of Spindle Vibrations," Journal of

Vibration Engineering & Technologies, pp. 1-19, 2023.

[20]. I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," IEEE Access, vol. 10, pp. 99129-99149, 2022.

[21]. J. K. Bii, "Improved Adaptive Boosting in Heterogeneous Ensembles for Outlier Detection: Prioritizing Minimization of Bias, Variance and Order of Base Learners," JKUAT-COPAS, 2023.

[22]. Z. G. Modarres, M. Shabankhah, and A. Kamandi, "Making AdaBoost Less Prone to Overfitting on Noisy Datasets," in 2020 6th International Conference on Web Research (ICWR), 2020: IEEE, pp. 251-259.

[23]. J. Sauer, V. C. Mariani, L. dos Santos Coelho, M. H. D. M. Ribeiro, and M. Rampazzo, "Extreme gradient boosting model based on improved Jaya optimizer applied to forecasting energy consumption in residential buildings," Evolving Systems, pp. 1-12, 2021.

[24]. J. P. Bharti, P. Mishra, U. moorthy, V. Sathishkumar, Y. Cho, and P. Samui, "Slope stability analysis using Rf, gbm, cart, bt and xgboost," Geotechnical and Geological Engineering, vol. 39, pp. 3741-3752, 2021.

[25]. P. Cunningham and S. J. Delany, "k-Nearest neighbour classifiers-A Tutorial," ACM computing surveys (CSUR), vol. 54, no. 6, pp. 1-25, 2021.

[26]. M.-L. Huang and Y.-C. Liao, "A lightweight CNN-based network on COVID-19 detection using X-ray and CT images," Computers in Biology Medicine, vol. 146, p. 105604, 2022.

[27]. E. M. Dharma, F. L. Gaol, H. Warnars, and B. Soewito, "The accuracy comparison among Word2vec, Glove, and Fasttext towards convolution neural network (CNN) text classification," Journal of Theoretical Applied Information Technology, vol. 100, no. 2, p. 31, 2022.

[28]. B. N. E. Nyarko, W. Bin, J. Zhou, G. K. Agordzo, J. Odoom, and E. Koukoyi, "Comparative analysis of Alexnet, resnet-50, and inception-V3 models on masked face recognition," in 2022 IEEE World AI IoT Congress (AIIoT), 2022: IEEE, pp. 337-343.

[29]. N. S. Shadin, S. Sanjana, and N. J. Lisa, "COVID-19 diagnosis from chest X-ray images using convolutional neural network (CNN) and InceptionV3," in 2021 International Conference on Information Technology (ICIT), 2021: IEEE, pp. 799-804.