APPLICATION OF MACHINE LEARNING FOR ENSURING WATER QUALITY IN AQUACULTURE POND

A PROJECT REPORT

Submitted by

LAKSHMAN SAI.P [Reg No: RA2011027010202]

CHARAN RAM AYYAPPA.P [Reg No: RA2011027010186]

JAYA PRAVEEN REDDY.K [Reg No: RA2011027010198]

SAI PREETHAM REDDY.G [Reg No: RA2011027010180]

Under the Guidance of

Dr. SV. SHRI BHARATHI

(Assistant Professor, Department of Data Science and Business Systems)

In partial fulfilment of the requirements for the degree of

BACHELOR OF TECHNOLOGY
In
COMPUTER SCIENCE AND ENGINEERING
with specialization in BIG DATA ANALYTICS

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR – 603 203

MAY 2024

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY,

KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that this B. Tech project report titled "**APPLICATION OF MACHINE LEARNING FOR ENSURING WATER QUALITY IN AQUACULTURE PONDS**" is the bonafide work of "**LAKSHMAN SAI.P [Reg. No. RA2011027010202], CHARAN RAM AYYAPPA.P [Reg. No.RA2011027010186], JAYA PRAVEEN REDDY.K [Reg. No.RA2011027010198] and SAI PREETHAM REDDY.G [Reg. No.RA2011027010180]"** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

Dr.SV. SHRI BHARATHI
GUIDE
Assistant Professor
Department. of Data Science and
Business Systems

Dr.M. LAKSHMI
PROFESSOR
HEAD OF THE DEPARTMENT
Department. of Data Science and
Business Systems

Internal Examiner

External Examiner

Department of Data Science and Business Systems
SRM Institute of Science & Technology
Own Work Declaration Form

Degree/ Course	:	**B.Tech CSE- Big Data Analytics**
Student Name	:	**LAKSHMAN SAI.P , CHARAN RAM AYYAPPA.P, JAYA PRAVEEN REDDY.K, SAI PREETHAM REDDY.G**
Registration Number	:	**RA2011027010202, RA2011027010186 RA2011027010198, RA2011027010180**
Title of Work	:	**APPLICATION OF MACHINE LEARNING FOR ENSURING WATER QUALITY IN AQUACULTURE PONDS**

We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism, as listed in the University Website, Regulations, and the Education Committee guidelines.

We confirm that all the work contained in this assessment is our own except where indicated, and that We have met the following conditions:

- Clearly references / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website
- we understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:
We are aware of and understand the University's policy on Academic misconduct and plagiarism and we certify that this assessment is our own work, except Were indicated by referring, and that we have followed the good academic practices noted above.
LAKSHMAN SAI.P (RA2011027010202) CHARAN RAM AYYAPPA.P (RA2011027010186) JAYA PRAVEEN REDDY.K (RA201102710198) SAI PREETHAM REDDY.G (RA2011027010180)

ACKNOWLEDGEMENTS

We express our humble gratitude to **Dr.C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, **Dr.T.V.Gopal**, for his invaluable support.

We wish to thank **Dr.Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We are incredibly grateful to our Head of the Department**, Dr.M. Lakshmi** Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our program coordinators **Dr.G. Vadivu,** Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for her inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor**, Dr.SV.Shri Bharathi,** Assistant Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, **Dr.SV.Shri Bharathi,** Assistant Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for providing me with an opportunity to pursue my project under her mentorship. She provided me with the freedom and support to explore the research topics of my interest. Her passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Data Science and Business Systems staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

Lakshman Sai. P [Reg No. RA2011027010202]

Charan Ram Ayyappa. P [Reg NO. RA2011027010186]

Jaya Praveen Reddy. K [Reg No. RA2011027010198]

Sai Preetham Reddy. G [Reg No. RA2011027010180]

ABSTRACT

The integration of machine learning into aquaculture management represents a pivotal step forward in addressing the industry's persistent challenges. By harnessing sophisticated data preparation techniques and leveraging sensor data, this study proposes a transformative approach to real-time monitoring of crucial water quality parameters like temperature, pH, and dissolved oxygen. Through rigorous model training and evaluation, aqua culturists gain actionable insights to anticipate trends and make informed decisions, bolstering both operational efficiency and environmental stewardship. This proactive strategy not only mitigates the limitations of manual monitoring but also enhances sustainability by reducing risks of financial losses and environmental degradation. By harmonizing technological innovation with ecological responsibility, this research underscores the dual imperative of fostering economic viability and safeguarding aquatic ecosystems. Ultimately, by empowering aquaculture practitioners with predictive analytics-driven interventions, this study paves the way for a paradigm shift towards resilient and sustainable aquaculture practices, where environmental health and productivity converge harmoniously.

Keywords:

Aquaculture management, Machine learning, Real-time monitoring, Water quality parameters, Sensor data, Predictive analytics.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

SVM	Support Vector Machine
SVR	Support Vector Regression
XGBoost	Extreme Gradient Boosting
KNN	K-Nearest Neighbor
VCS	Voting Classifier Soft
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

CHAPTER 1
INTRODUCTION

1.1. INTRODUCTION

Aquaculture plays a pivotal role in meeting the increasing global demand for seafood, necessitating a keen focus on maintaining optimal pond water quality to ensure the health and productivity of aquatic life. To address this critical requirement, a pioneeringproject leverages machine learning to enhance the assurance of aquaculture pond waterquality.

By integrating machine learning techniques into water quality management, this projectheralds a transformative shift, offering a data-centric approach to navigating the complex interactions within aquatic environments. At its core lie sophisticated algorithms that facilitate the development of real-time monitoring systems. These systems empower aquaculturists with swift and informed decision-making capabilities,furnishing rapid insights into crucial water quality metrics.

Drawing from historical data, the project delves into predictive analytics to anticipate future trends in water quality. This proactive approach equips aquaculturists with foresight to preemptively address potential issues, thereby averting unfavorable conditions and safeguarding the well-being of cultivated marine life.

Beyond its immediate benefits to aquaculture, this initiative contributes to broader sustainability objectives. Effective water quality control not only preserves delicate aquatic ecosystems but also aligns with international efforts to promote ethical and sustainable practices in the aquaculture sector.

As the project evolves, its aim is to establish a comprehensive framework for water quality assurance that complements the intricate equilibrium of nature while fostering innovation in aquaculture techniques. This endeavor envisions a future where aquaculture not only satisfies the demand for seafood but does so in a manner that is resilient, sustainable, and mindful of the fragile aquatic ecosystems upon which it relies. By amalgamating technology with environmental

stewardship, this initiative paves theway for a harmonious coexistence between aquaculture and nature.

1.2. PROBLEM STATEMENT

The primary goal is to address the deficiencies of existing water quality monitoring practices in aquaculture ponds, where fluctuations in crucial parameters like temperature, pH, and dissolved oxygen jeopardize the health and productivity of aquaticorganisms. Current methods, often manual and lacking real-time capabilities, result in delayed responses to deteriorating water quality, exposing aquaculture operations to heightened risks of disease outbreaks, reduced growth rates, and economic losses.

To counter these challenges, urgent action is required to develop an advanced, automated system leveraging machine learning. This system aims to enable real-time monitoring and proactive management of water quality, revolutionizing aquaculture practices. By continuously analyzing data from deployed sensors, the system will swiftly detect and predict changes in key parameters, allowing aquaculturists to promptly intervene and prevent adverse effects on aquatic life.

Furthermore, through predictive analyticsutilizing historical data, the system will forecast future trends in water quality, empowering aquaculturists to anticipate issues and implement preventive measures, thus enhancing production outcomes and sustainability. Overall, the development of such a system represents a crucial advancement for the aquaculture industry, promising improved resilience, sustainability, and economic viability in the face of evolving environmental challenges.

1.3. BACKGROUND OF THE PROJECT

The background of the problem lies in the pivotal role of water quality in the success of aquaculture, which is a vital industry for meeting global seafood demands. Aquaculture ponds are dynamic ecosystems where fluctuations in temperature, pH, and dissolved oxygen significantly influence the well-being of aquatic organisms. Traditional monitoring methods, often reliant on manual labor and

periodic sampling, fall short in providing timely and continuous data on water quality. These limitations expose aquaculture operations to the risk of suboptimal conditions, impacting the health and growth of aquatic species. Hence, there is a clear need for a more advanced and automated approach, integrating machine learning, to ensure consistent, real-time monitoring and management of water quality in aquaculture ponds. Addressing this issue is crucial for sustaining the industry's economic viability and minimizing environmental impact.

CHAPTER 2
LITERATURE SURVEY

2.1. RELATED WORK

Quintero, Parra, and Félix (2022) proposed a comprehensive framework for water quality assurance in aquaculture ponds by leveraging Machine Learning techniques. Their project aimed to monitor and forecast crucial water quality parameters, specifically focusing on temperature and dissolved oxygen levels, across different sectors of the aquaculture industry. The framework was designed to cater to various settings within aquaculture, including post-larval laboratories, aquaculture farms, and during transportation of aquaculture products. The implemented prototype system, centered around product transportation, consisted of a network of dissolved oxygen and temperature sensors. These sensors transmitted real-time data via Bluetooth Low Energy to a mobile application for initial processing. Subsequently, the processed data was sent over the internet to a web application, enabling users to receive alerts, visualize monitoring data, and access a forecast model. The forecast model, developed using a Long Short-Term Memory (LSTM) neural network, provided users with predictive insights, empowering them to take proactive measures to mitigate potential losses in aquaculture production.**[1]**

Rana et al. (2021) conducted a study to investigate the impact of water quality (WQ) on aquatic livestock in freshwater ponds, emphasizing its crucial role in harvest outcomes. They collected WQ data and harvest records from an Australian prawn farm over a complete grow-out season to understand how fluctuations in WQ parameters affect prawn harvests. Utilizing machine learning techniques, the study aimed to achieve two primary objectives: firstly, to discern the significance of five key WQ variables in distinguishing between high and low performing ponds in terms of harvest performance, and secondly, to identify how variations in these variables throughout the grow-out season influence the final harvest outcome, including growth and yield. To classify ponds based on performance, a range of machine learning classifiers were employed, including neural networks, support vector machines, k-nearest neighbors, logistic regression, Gaussian naive Bayes, decision trees, random forests, and AdaBoost. Additionally, three feature selection methods—mutual

information, correlation-based feature selection, and ReliefF—were utilized to pinpoint the key driving factors, in terms of WQ variations, affecting the growth and yield of aquatic species. The findings highlighted dissolved oxygen, salinity, and temperature as the most influential WQ variables impacting overall harvest performance in the ponds. Notably, changes in dissolved oxygen and salinity during the final quarter of the grow-out season, coupled with temperature variations immediately after stocking, were identified as significant contributors to differentiating between high and low performing ponds.**[2]**

Zhao et al. (2021) examined the application of machine learning in intelligent fish aquaculture, reflecting the growing trend of integrating automation and artificial intelligence in the industry. Over the past five years, machine learning technology has been extensively adopted, offering new opportunities for digital fish farming. The study comprehensively explores various machine learning algorithms and techniques utilized in intelligent fish aquaculture, focusing on their applications such as fish biomass evaluation, fish identification and classification, behavioral analysis, and the prediction of water quality parameters. Additionally, the paper outlines the implementation of machine learning algorithms in aquaculture and provides an analysis of the outcomes. Finally, it underscores several ongoing challenges in aquaculture and speculates on future development trends.**[3]**

Li et al. (2022) delved into the predictive modeling of aquaculture water quality using machine learning approaches, recognizing the significance of high-quality water in industrial aquaculture settings. Due to the often unavailability of real-time monitoring in such systems, predicting water quality parameters becomes imperative for effective production management. The study employed various machine learning techniques, including back propagation neural network (BPNN), radial basis function neural network (RBFNN), support vector machine (SVM), and least squares support vector machine (LSSVM), to simulate and forecast key water quality parameters such as dissolved oxygen (DO), pH, ammonium-nitrogen (NH3-N), nitrate nitrogen (NO3-N), and nitrite-nitrogen (NO2-N). Comparative analysis using published data revealed varying degrees of prediction accuracy among the methods. SVM exhibited the most accurate and consistent results across all parameters, particularly in industrial aquaculture systems relying on groundwater sources. With an accuracy rate of 80% for both published and measured data, SVM emerged as the recommended

model for simulating and forecasting water quality in industrial aquaculture setups.**[4]**

Zambrano et al. (2021) investigated the application of machine learning (ML) techniques for predicting water quality parameters in fish farming, acknowledging the crucial role of monitoring parameters such as dissolved oxygen, pH, and pond temperature in maintaining optimal operations. While previous research in ML for aquaculture has focused on scenarios with real-time data acquisition devices, the study addresses the limitations faced by fish farmers who rely on manual equipment for measuring these variables, resulting in restricted data availability. The research explores the utilization of random forests, multivariate linear regression, and artificial neural networks in scenarios with limited measurement data, emphasizing commonly measured waterquality variables in fish farming. The proposed methodology enables the construction of models for estimating unobserved variables based on observed ones and forecasting with limited training data. The findings highlight the effectiveness of random forests in forecasting dissolved oxygen, pond temperature, pH, ammonia, and ammonium, even with measurements taken only twice daily. Moreover, the study demonstrates the feasibility of implementing these prediction models on a mobile-based information system, making them accessible to fish farmers using average smartphones.**[5]**

Omambia et al. (2022) emphasize the importance of access to safe water, essential for human survival and recognized as a fundamental human right. However, challenges such as contamination, leakages, and theft often arise as individuals utilize water primarily sourced from pipes and springs located in towns. To address these issues, the authors propose leveraging Internet of Things (IoT) and Machine Learning technologies. Their suggested system aims to monitor water quality and detect and address problems like theft and wastage. By utilizing machine learning algorithms for decision-making processes, the system offers a promising solution to tackle these challenges effectively.**[6]**

Roy Prapti et al. (2021) delve into the burgeoning interest in utilizing Internet of Things (IoT) technology in aquaculture, which has seen significant advancements in the broader agricultural sector (Agriculture 4.0). However, Aquaculture 4.0 remains relatively underdeveloped in many regions. The authors present findings from a meticulous analysis of 30 internationally published

research papers, focusing on water quality monitoring in fishponds. Their review categorizes the research into five sections: recent research (2011–2020), aquaculture environments, research methodologies, prevalent water quality parameters, and types of solutions provided. The majority of the reviewed research concentrated on inland aquaculture (81%), with marine aquaculture species studies accounting for 19% of the papers. The framework and architecture approach (48%) emerged as the most commonly adopted research methodology in IoT-based aquaculture for water quality monitoring. The study underscores the need for long-term experimental research to identify challenges and propose suitable solutions. Regarding water quality parameters, temperature, dissolved oxygen, and pH were identified as the most prioritized parameters in IoT-based aquaculture, with real-time monitoring frequently proposed as a solution. Autonomous monitoring offered a unique approach. The insights from this study are anticipated to benefit various stakeholders in the aquaculture industry, including researchers, practitioners, and decision-makers.**[7]**

Zhang and Gui (2021) emphasize the growing significance of marine aquaculture as a key strategy for promoting ecological sustainability amidst declining natural fishery resources. To confront the challenges inherent in aquaculture, improve operational efficiency, and modernize fishing practices, there is a rising trend towards adopting innovative digital technologies such as the Internet of Things (IoT), big data analytics, cloud computing, artificial intelligence (AI), and blockchain. The authors elucidate the interconnectedness of these digital technologies and delineate their application framework within the context of marine aquaculture. The paper elucidates the outcomes derived from the implementation of each digital technology in marine aquaculture while also identifying their respective advantages and challenges. Additionally, it delves into the utilization of these technologies in deep-sea aquaculture facilities, showcasing their versatility across diverse aquaculture settings. Furthermore, the authors address primary challenges encountered by these technologies in marine aquaculture production and propose future development trends.**[8]**

Zou et al. (2021) address the challenges associated with extracting coastal aquaculture ponds, particularly focusing on mitigating the "same-spectrum foreign objects" effect and improving boundary definition precision in extraction results. To tackle these obstacles, the researchers propose a novel approach utilizing the U2-Net deep learning architecture for pond extraction from

remote sensing imagery in coastal areas. The methodology involves several key steps, beginning with image preprocessing to enhance spectral features and the generation of samples through visual interpretation. Subsequently, the U2-Net model is trained and deployed for extracting aquaculture ponds along coastal regions. Post-processing techniques are then applied to refine the model's extraction results. Validation experiments conducted in the Zhoushan Archipelago, China, demonstrate the effectiveness of the proposed method. The results indicate an average F-measure of 0.93 across four study cases, with average precision and recall rates exceeding 90%. These findings underscore the suitability of the developed approach for applications in aquaculture pond extraction along coastal regions. This study contributes to the advancement of rapid and accurate mapping of coastal aquaculture ponds, offering valuable technical support for marine resource management and initiatives aimed at promoting sustainable development.**[9]**

Zulkifli et al. (2021) conduct a systematic review focusing on IoT-based water monitoring systems, emphasizing their crucial role in transitioning towards intelligent and smart agriculture. As new technologies continuously evolve and integrate into agricultural practices and human daily life, the need for automated monitoring of water quality becomes increasingly apparent. The review examines the water quality literature spanning the past five years (2018–2022) to address concerns, issues, difficulties, and research gaps in this domain. By analyzing sensor-based water quality monitoring models, the authors aim to provide insights into the precision required for modeling and the necessity for reliable datasets. To ensure the reliability of their findings, the researchers search and scrutinize several digital databases, including IEEE Explore®, Science Direct, Scopus, and Web of Science. Out of 946 papers obtained, only 50 articles meet the inclusion criteria for the study, with a focus on real-time data acquisition systems and water quality monitoring procedures. The reviewed literature predominantly comprises reviews and experimental studies, categorized into three main groups based on experimental conditions. Additionally, the authors develop a taxonomy to organize the literature effectively and provide recommendations to accelerate advancements in this field. Through a comprehensive analysis of existing methodologies, the review identifies research gaps, particularly in model accuracy, data-gathering system development, and the types of data utilized in proposed frameworks. Finally, the authors outline research directions towards smart water quality, emphasizing the importance of addressing critical topics for the advancement of this research area.**[10]**

Chiu et al. (2021) present a smart aquaculture farm management system developed to address labor shortages and enhance productivity in the aquaculture sector, particularly in Taiwan. The system integrates IoT technology and artificial intelligence (AI) to create a comprehensive monitoring and control platform for fish farms. By deploying various IoT devices, the system enables real-time data collection, allowing for remote monitoring, adjustment, and assessment of fishpond water quality and other system parameters. Additionally, the study develops a deep learning (DL) model to predict the growth of California Bass fish, a common species in aquaculture. The DL model, optimized using Bayesian optimization-based hyper-parameter tuning, achieves high accuracy, with an R2 value of 0.94 and a mean square error of 0.0015 on the experimental dataset. The successful prediction capability of the model demonstrates its potential for practical application in aquaculture management. Moreover, the study explores the integration of the DL model into an autonomous feeding system, aiming to optimize feed usage and minimize waste. By leveraging the capabilities of AIoT (Artificial Intelligence of Things), the proposed system empowers fish farmers to remotely control and manage various aspects of fishpond equipment intelligently. Overall, the smart aquaculture farm management system holds promise for enhancing operational efficiency, reducing labor dependency, and promoting sustainable aquaculture practices.**[11]**

Chen et al. (2021) introduce an innovative method for intelligent water quality monitoring based on image processing techniques and machine learning algorithms. Recognizing the significant impact of water quality on various aspects such as aquatic life, agricultural irrigation, and human health, the study focuses on the aquaculture industry, where water color can serve as an indicator of phytoplankton species and abundance. The proposed approach involves extracting essential features from water color images and leveraging machine learning methods to develop an intelligent monitoring system. Specifically, the system integrates a fused random vector functional link network (RVFL) with the group method of data handling (GMDH) model to analyze and predict water quality based on image data. The fusion of RVFL and GMDH enhances the system's performance compared to existing methods, offering superior accuracy and reliability in water quality monitoring. By harnessing image processing and machine learning technologies, the proposed approach provides a sophisticated yet efficient solution for assessing water quality in aquaculture settings.**[12]**

Zhang et al. (2021) conducted an evaluation and analysis of water quality in marine aquaculture areas to address its significance as a limiting factor in the rapid development of aquaculture. Sampling was conducted in both pond and cage aquaculture areas during May, August, and November 2018. Nine water quality indicators were measured, including pH, temperature, salinity, dissolved oxygen, molybdate-reactive phosphorus, chemical oxygen demand, chlorophyll a, inorganic nitrogen, and antibiotic resistance genes (ARGs). Principal component analysis (PCA) was employed to analyze spatial-temporal changes and identify driving factors influencing water quality conditions in both pond and cage aquaculture areas. The results revealed three main components in each area, explaining 66.82% and 72.99% of the variance, respectively. Salinity, dissolved oxygen, and ARGs were identified as the most influential factors in pond aquaculture, while chlorophyll a, salinity, and dissolved oxygen played significant roles in cage aquaculture. The study found that the heaviest polluted months varied between August and November for pond and cage aquaculture areas, respectively. It was observed that pond volume primarily influences water quality in pond aquaculture areas, whereas aquaculture activities and seasonal variations are major factors in cage aquaculture areas. Furthermore, the presence of antibiotic resistance genes indicated potential terrestrial inputs impacting cage culture areas. These findings provide valuable insights for relevant authorities in selecting appropriate water quality monitoring parameters in marine aquaculture areas.**[13]**

Manoj et al. (2022) delve into the critical issue of declining safe water resources worldwide, exacerbated by climate change, contamination, and pollution. Underwater life forms are particularly vulnerable to these hazards, making continuous water quality monitoring imperative. Traditional methods involve energy-intensive processes of collecting water samples from various locations and subsequent laboratory testing. In contrast, the Internet of Things (IoT) offers a promising solution. The paper provides an exhaustive review of water quality monitoring systems (WQSN) utilizing IoT proposed by researchers over the past decade (2011–2020). It evaluates the advancements in quality measures and success indicators.**[14]**

CHAPTER 3
RESEARCH OBJECTIVE

3.1. RESEARCH CHALLENGES

Developing water quality assurance systems in aquaculture ponds through machine learning introduces a spectrum of intricate research challenges. Central to this endeavoris the intricate task of harmonizing diverse data sources, ensuring their seamless integration, and enabling real-time processing capabilities. The design of models that strike a delicate balance between complexity and interpretability stands as a critical consideration. Additionally, addressing imbalances within datasets and bolstering predictive model robustness across varied aquaculture environments pose formidable obstacles. Ethical dimensions concerning data privacy demand careful navigation, alongside the creation of user-friendly interfaces to facilitate practical implementation.

Furthermore, attaining scalability, cost-effectiveness, and smooth integration with existing aquaculture infrastructure necessitate innovative approaches. Overcoming these challenges represents a pivotal step towards harnessing the full potential of machine learning in augmenting water quality management within aquaculture. By surmounting these hurdles, the field can advance significantly, paving the way for more efficient, sustainable, and ethically sound practices in aquaculture operations worldwide.

3.2. RESEARCH OBJECTIVES

The research objective is to lead the development of a comprehensive water quality assurance system tailored for aquaculture ponds, leveraging the integration of machinelearning technologies. This involves pioneering advanced models capable of predicting and monitoring a wide range of water quality parameters, ensuring their adaptability across diverse aquaculture environments. Overcoming challenges such as complexitiesin data integration, imbalanced datasets, and ethical considerations is central to this endeavor. Moreover, the research aims to design user-friendly interfaces to facilitate practical implementation, thereby promoting acceptance among aquaculture

managers.Emphasis will also be placed on optimizing scalability and cost-effectiveness to cater to both large-scale operations and smaller facilities. Ultimately, these efforts are geared towards fostering sustainable enhancements in aquaculture practices, thereby benefiting the industry as a whole.

CHAPTER 4

PROPOSED WORK

4.1. METHODOLOGY

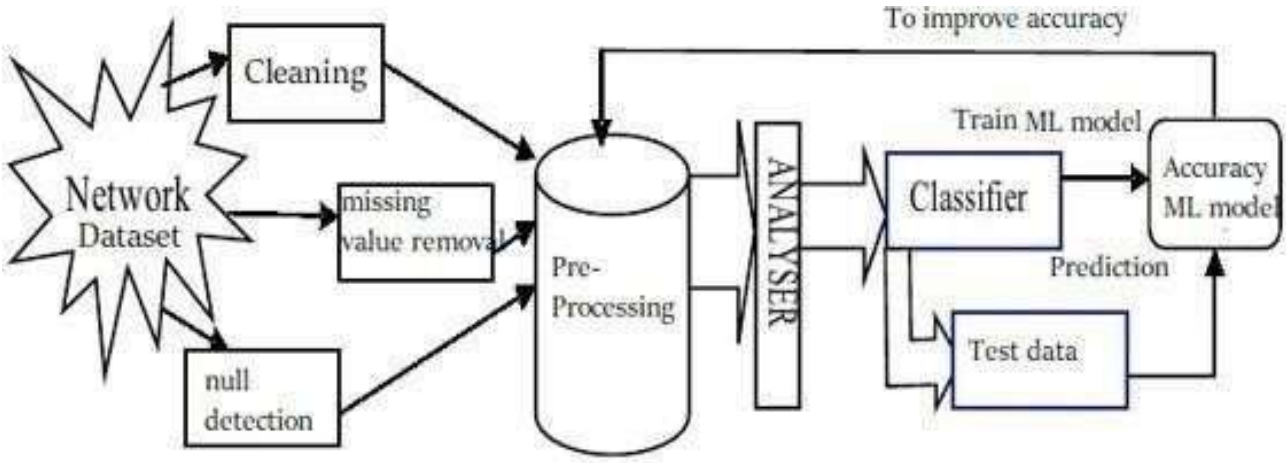


Fig 4.1 – Proposed Architecture

The proposed system contains various stages which contain several substages. Thestages involved in the proposed work are:

1. Data Collection
   1.1 Data Pre-processing
   1.2 Feature Extraction
2. Machine Learning Algorithms
   2.1 Support Vector Machine
   2.2 XGBoost classifier
   2.3 K-Nearest Neighbour
   2.4 Voting Classifier (VCS)
3. Performance Evaluation
   3.1 Performance matrix
   3.2 Confusion matrix

4. Model Training
5. Validation and Testing
6. Real-time Integration

4.2. DATA COLLECTION

During this critical phase, a diverse array of data sources, including sensors and environmental records, are amalgamated to construct a comprehensive dataset that captures real-time information concerning water quality parameters within aquaculture ponds. The accuracy and relevance of subsequent analyses hinge upon the completeness and fidelity of the collected data. This dataset encompasses several key parameters:

1. pH: Indicative of acidity levels in the water.
2. Hardness: Reflective of the water's mineral content.
3. Solids: Total concentration of solids present in the water.
4. Chloramines: Measurement of chloramine concentration.
5. Sulfate: Indicates the levels of sulfate present in the water.
6. Conductivity: Represents the electrical conductivity of the water.
7. Organic carbon: Measurement of organic carbon concentration.
8. Trihalomethanes: Indication of trihalomethane levels in the water.
9. Turbidity: Reflects the degree of clarity or cloudiness in the water.
10. Check: Categorical value with 0 and 1.

4.2.1. DATA PRE-PROCESSING

In the realm of aquaculture water quality management, the initial preprocessing of gathered data stands as a critical juncture, where meticulous attention is devoted to ensuring the reliability and integrity of the dataset. This pivotal step encompasses a multifaceted approach, encompassing tasks such as data cleaning, filtering, and addressing missing values. Cleaning the data involves identifying and rectifying inconsistencies or errors, ranging from duplicate entries to typographical mistakes, thereby enhancing the accuracy of the dataset. Simultaneously, filtering procedures are implemented to eliminate irrelevant or redundant data points, ensuring that subsequent analyses

focus solely on the most pertinent information. Addressing missing values is equally paramount, necessitating the deployment of various techniques, including imputation methods and advanced machine learning algorithms, to fill in gaps and maintain dataset completeness. Beyond data refinement, the preprocessing phase also entails ensuring compliance with relevant quality standards and regulations governing data collection and management practices in the aquaculture industry. Adherence to these standards not only upholds data integrity but also facilitates regulatory compliance, contributing to the safety and sustainability of aquatic ecosystems. Ultimately, the preprocessing phase serves as the cornerstone for subsequent analyses and model training processes, laying the groundwork for informed decision-making and effective water quality management strategies in aquaculture systems, thereby advancing the industry's sustainability and success.

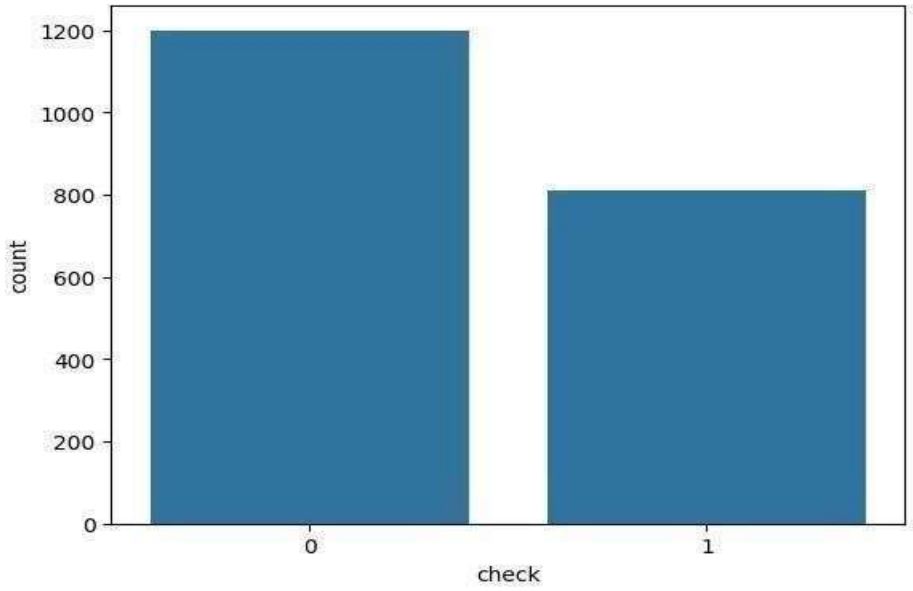


Fig 4.2 – Unbalanced "CHECK" Data

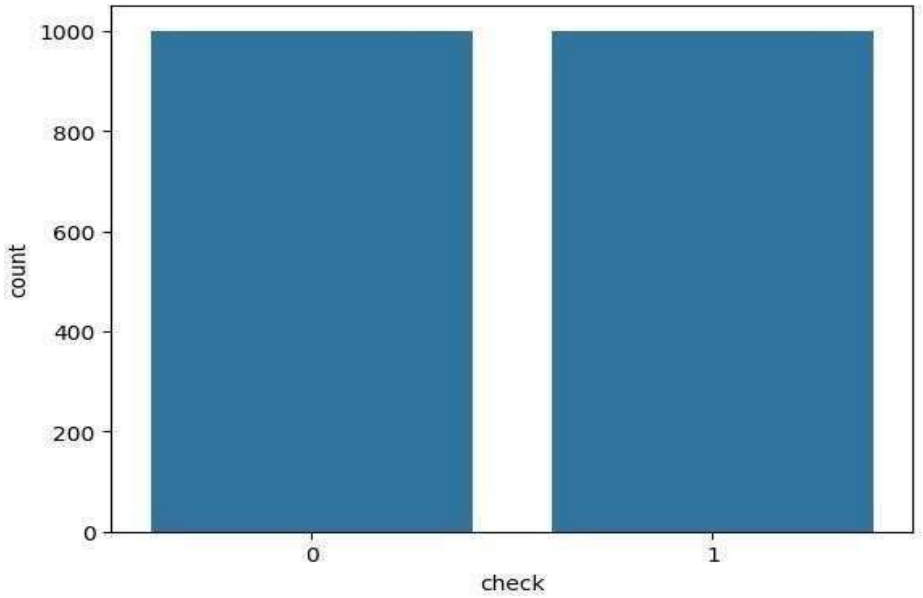


Fig 4.3 – Balanced "CHECK" Data

4.2.2. FEATURE EXTRACTION

In aquaculture projects, feature extraction from water quality datasets plays a pivotal role in identifying and prioritizing parameters critical for assessing aquatic ecosystem health. Variables such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity are meticulously scrutinized to discern their significance and relevance to water quality. This process involves assessing the interplay between these variables, identifying correlations, and discerning which features hold the greatest importance in influencing aquatic life and productivity. Through domain expertise and understanding of aquaculture dynamics, researchers can discern key indicators pivotal for effective water quality management. Moreover,preprocessing steps may be applied to refine the dataset, ensuring data integrity and facilitating subsequent analyses. By systematically extracting and refining essential features, aquaculture projects can develop robust models for real-time monitoring and management, fostering sustainable practices and safeguarding aquatic ecosystems for enhanced productivity and resilience.

4.3. MACHINE LEARNING ALGORITHMS

The core of the system utilizes various machine learning algorithms such as regression,classification, and ensemble methods. These algorithms meticulously analyze theextracted features to enable precise predictions of water quality conditions in aquaculture ponds.

4.3.1. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a powerful supervised learning algorithm widely employed in classification and regression tasks. In classification, SVM aims to delineate the optimal decision boundary or hyperplane within n-dimensional space to separate different classes. This algorithm identifies crucial support vectors, which are extreme points instrumental in constructing the hyperplane. In aquaculture, Support Vector Regression (SVR) emerges as a valuable tool for predictive modeling and decision-making. Water quality factors such as Trihalomethanes (THMs), pH, hardness, solids, chloramine, organic carbons, conductivity, sulfate, and turbiditysignificantly influence the productivity and health of aquatic species. SVR proves instrumental in forecasting and controlling these parameters to maintain ideal conditions for aquatic life.

SVR operates by identifying the optimal hyperplane that minimizes the error between actual and projected values within a predetermined margin or maximizes the margin between different classes. These parameters serve as features to train an SVR model forbinary classification or regression problems in aquaculture. For instance, aquaculture farmers may employ SVR to determine whether specified water quality metrics renderthe conditions acceptable (1) or unsuitable (0) for a particular fish species. By training the model on historical data with known water quality metrics and corresponding appropriateness labels, SVR facilitates binary classification.

During the training phase, the SVR model learns the correlations between binary labels(appropriate or unsuitable conditions) and input features (pH, hardness, THMs, etc.). Itadjusts its parameters to identify the hyperplane that best separates the two classes andmaximizes the margin between them. Subsequently, the trained SVR model can forecast in real-time whether water conditions are suitable

for aquaculture, aiding farmers in decision-making regarding aquatic species' well-being and productivity.

Furthermore, SVR sheds light on the interconnections among various water quality measures and their impact on water suitability for aquaculture. Armed with thisknowledge, farmers can identify crucial factors influencing water quality and make informed decisions to enhance aquatic species' health and growth. In conclusion, SVR serves as a valuable technique in aquaculture for predicting water quality parameters and assessing water suitability for aquatic life. By leveraging SVR, aquaculture farmerscan ensure optimal conditions for the well-being and efficiency of their aquatic creatures, facilitating data-driven decision-making in the industry.

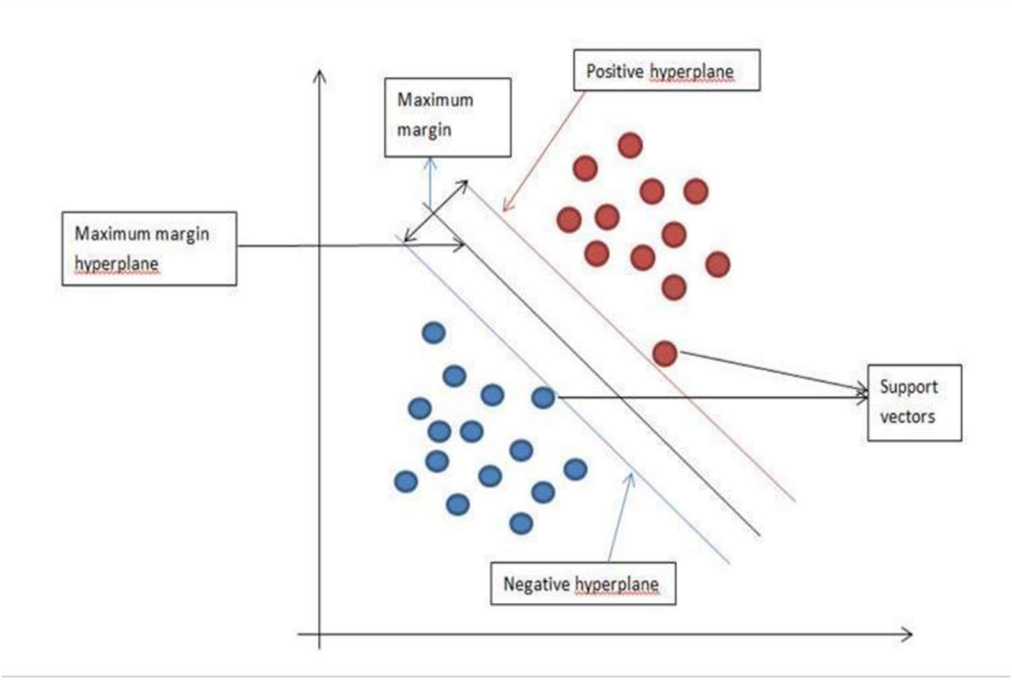


Fig 4.4 – Support Vector Machine

4.3.2. XG BOOST CLASSIFIER

XGBoost, short for Extreme Gradient Boosting, is a machine-learning library renowned for its distributed gradient-boosted decision trees. Highly scalable, it finds extensive use in regression, classification, and ranking problems, excelling particularly in paralleltree-boosting. To comprehend

XGBoost fully, familiarity with fundamental learning concepts such as ensemble learning, decision trees, learning, and boosting is essential. In aquaculture, XGBoost emerges as a potent machine learning method, especially for predictive modeling applications involving binary classification. With parameters like pH, hardness, sediment, organic carbons, conductivity, sulfate, turbidity, trihalomethanes (THMs), and chloramine, XGBoost offers crucial insights into water suitability for aquatic life.

Maintaining ideal water quality in aquaculture is pivotal for the well-being and productivity of aquatic life. Aquaculture professionals leverage XGBoost to forecast whether water conditions are suitable (1) or unsuitable (0) by inputting water quality metrics for binary classification problems. Typically, practitioners start by collecting historical data containing water quality measurements and appropriateness labels to effectively utilize XGBoost. The model utilizes this dataset for training, iteratively fitting decision trees to learn how to identify water conditions while optimizing the decision trees to minimize a specified loss function, such as binary cross-entropy.

XGBoost continuously enhances its ability to detect water conditions accurately by prioritizing learning from misclassified instances. It constructs an ensemble of decisiontrees, iteratively refining its capacity throughout the training phase. Once trained, the XGBoost model forecasts new data, enabling aquaculture professionals to predict appropriateness labels (0 or 1) for current water conditions based on measurements of water quality parameters. Additionally, XGBoost's interpretability aids practitioners inunderstanding the significance of various water quality metrics in determining appropriateness.

Through the examination of feature importance scores produced by XGBoost, professionals can discern parameters exerting the most influence on water quality and adapt their management strategies accordingly. Furthermore, XGBoost offers scalability and versatility, making it adept at handling large datasets and incorporatingdiverse features. Its adaptability proves beneficial in aquaculture applications where a multitude of conditions impact water appropriateness and quality.

In conclusion, XGBoost emerges as a valuable tool in aquaculture for binary classification problems related to assessing water condition suitability for aquatic species and analyzing water quality. By

harnessing XGBoost's ability to make well- informed judgments, aquaculture practitioners can optimize water management methods and ensure the well-being and productivity of their aquatic populations.

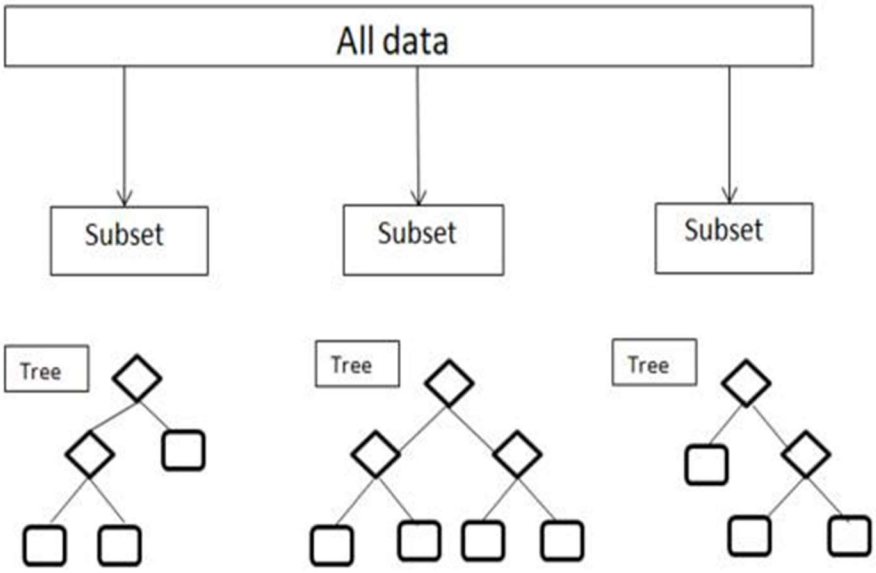


Fig 4.5 – XG Boost Classifier

4.3.3. K-NEAREST NEIGHBOR

K-Nearest Neighbors (KNN) stands out as a popular machine learning technique for classification tasks, including binary classification scenarios like assessing aquaculturewater quality. It offers simplicity and efficiency, making it a valuable tool for aquaculture professionals. By utilizing features such as pH, hardness, trihalomethanes (THMs), solids, chloramine, organic carbons, conductivity, sulfate, and turbidity, KNN enables the prediction of water condition suitability through grid searches and parameter adjustments, particularly the number of neighbors (K).

The health and productivity of aquatic creatures in aquaculture hinge on maintaining ideal water quality, a challenge that KNN can address effectively. Aquaculture professionals leverage KNN to classify water conditions as appropriate (1) or unsuitable(0) based on provided water quality data. The model's performance is enhanced throughgrid searches and parameter modifications, facilitating more

accurate predictions and improved generalization to unobserved data.

The procedure typically begins with gathering historical data containing water quality readings and appropriateness labels, which serves as the training dataset for the KNN model. Grid searches systematically evaluate the model's performance across various hyperparameter combinations, with the number of neighbors (K) being the primary hyperparameter requiring adjustment. Aquaculture professionals specify a range of K values to explore during the grid search process.

During each iteration of the grid search, the KNN model is trained with specific hyperparameters, and its performance is evaluated using cross-validation methods like k-fold cross-validation. Performance metrics such as accuracy or F1 score are calculated for each parameter combination. The grid search identifies the optimal hyperparameters to achieve the best results on the validation set, striking a balance between variance and bias and determining the optimal K value for the KNN model.

Once the optimal parameters are determined, the KNN model is trained on the full dataset, and then applied to new data to provide predictions. Aquaculture professionalscan fine-tune their KNN models to effectively categorize water conditions and make well-informed decisions on water management techniques, thereby supporting the prosperity and sustainability of aquaculture operations by preserving ideal water qualityconditions for aquatic species' health and well-being.

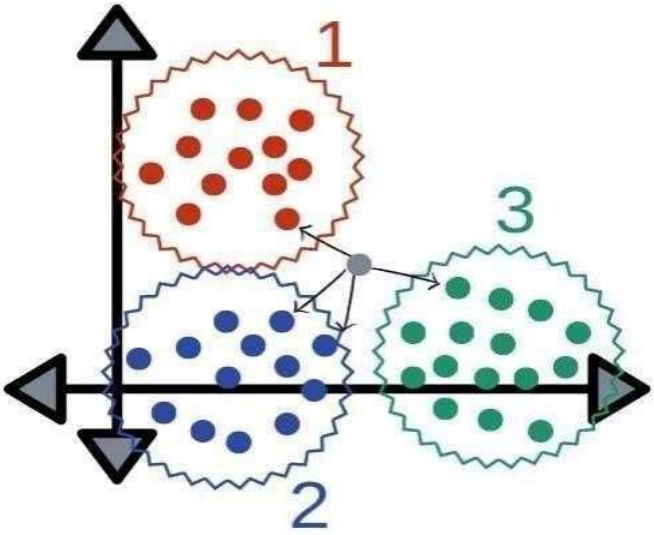


Fig 4.6 – K-Nearest Neighbor

4.3.4. VOTING CLASSIFIER (VCS)

The Voting Classifier (soft) is a powerful ensemble method in machine learning, particularly beneficial for binary classification tasks like evaluating the water quality in aquaculture. By aggregating predictions from multiple base classifiers, it offers accurate insights into the suitability of water conditions based on various features such as pH, hardness, solids, organic carbons, trihalomethanes (THMs), conductivity, sulfate, and turbidity. Maintaining ideal water quality is paramount for the well-being and productivity of aquatic animals in aquaculture. The Voting Classifier leverages thepredictions of diverse base classifiers, each trained on different subsets of features, to determine whether the water conditions are suitable or unsuitable.

The procedure typically begins with gathering historical data comprising water qualityparameter readings and corresponding suitability labels, which is then divided into training and testing sets for model development and evaluation. The Voting Classifier employs a range of base classifiers, including Support Vector Machines (SVM), Decision Trees, Random Forests, and K-Nearest Neighbors (KNN), each focusing on distinct subsets of features. During training, each base classifier learns to classify water conditions independently using its subset of features. In soft voting, the projected classprobabilities from each base classifier are combined, and the class with the highest probability is selected as the final prediction.

By considering the unique strengths and limitations of each base classifier, the VotingClassifier synthesizes predictions from multiple classifiers to generate more reliable and accurate predictions than any individual classifier alone. Optimization and parameter tuning are crucial for enhancing the Voting Classifier's performance. This involves identifying the best hyperparameters for each base classifier, such as regularization strength, tree depth, and the number of neighbors in KNN, through methods like grid search or random search.

Once trained and refined, the Voting Classifier can be used to make predictions on new data. Aquaculture practitioners can input water quality parameter data into the ensemblemodel to obtain

predictions about the suitability of water conditions for their aquatic creatures. In summary, the soft voting classifier offers a flexible and efficient approachfor aquaculture binary classification tasks, leveraging the strengths of multiple base classifiers to deliver precise predictions on water quality and suitability. By incorporating diverse viewpoints and features, the Voting Classifier supports optimal management of aquatic ecosystems in aquaculture settings, enabling informed decision-making.

4.4. PERFORMANCE EVALUATION

Performance evaluation refers to the process of assessing the effectiveness and efficiency of a system, model, or process based on predefined criteria or metrics. In thecontext of machine learning, performance evaluation involves measuring the accuracy, precision, recall, F1 score, and other relevant metrics to gauge the model's ability to make predictions or classifications accurately. It helps in understanding how well the model performs on unseen data, identifying areas for improvement, and comparing different models to determine the most suitable one for a particular task or application.

4.4.1. PERFORMANCE MATRIX

The dataset was divided into two segments: a training set comprising 70% of the data and a testing set containing the remaining 30%. Subsequently, several machine learningalgorithms, including Support Vector Machine (SVM), XGBoost, K-Nearest Neighbors (KNN), and Voting Classifier (VCS), were implemented using the Enthought Canopy platform. The primary metric used to evaluate the performance of these algorithms wasprediction accuracy.

Prediction accuracy serves as a fundamental measure of a model's effectiveness in making correct predictions on unseen data. It quantifies the proportion of instances correctly classified by the model out of the total instances in the testing dataset. Higheraccuracy values indicate better performance and greater reliability of the model in capturing underlying patterns and relationships within the data.

By assessing prediction accuracy, we can compare the performance of different algorithms and determine which one yields the most accurate predictions for the giventask. Additionally, it provides insights into the models' generalization capabilities and their ability to perform well on new, unseen data. Ultimately, the evaluation of prediction accuracy allows for informed decision-making regarding the selection and deployment of machine learning algorithms in practical applications such as aquaculture water quality assessment.

$$\text{ACCURACY} = \frac{+}{+}$$

4.4.2. CONFUSION MATRIX

A confusion matrix is a tabular representation used to evaluate the performance of a classification model by comparing the predicted outcomes with the actual outcomes. Itconsists of four main components: true positive (TP), true negative (TN), false positive(FP), and false negative (FN).

In a confusion matrix, the rows represent the actual classes or labels, while the columnsrepresent the predicted classes by the model. The diagonal elements (TP and TN) represent the instances that are correctly classified, where TP denotes the number of positive instances correctly predicted as positive, and TN denotes the number of negative instances correctly predicted as negative. On the other hand, the off-diagonal elements (FP and FN) represent the misclassifications, where FP indicates the number of negative instances wrongly predicted as positive, and FN indicates the number of positive instances wrongly predicted as negative.

By analyzing the values in the confusion matrix, various performance metrics can be derived to assess the model's effectiveness, such as accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to make correct predictions and its performance across different classes or categories. The confusion matrix serves as a valuable tool for understanding the strengths and weaknesses of the classification model and identifying areas for improvement in the model's performance.

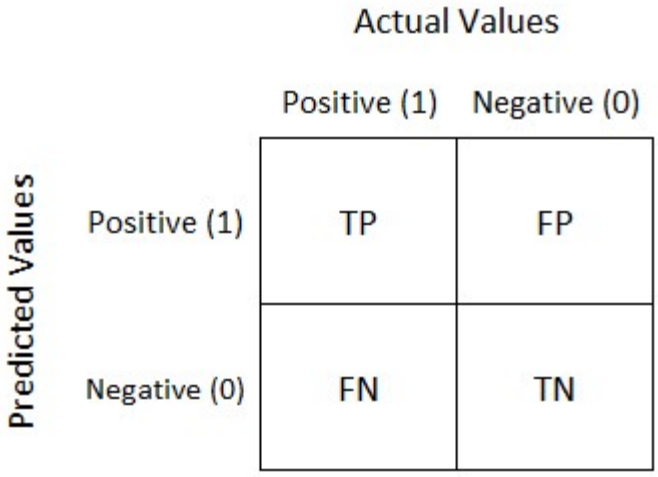


Fig 4.7 – Confusion Matrix

4.5. MODEL TRAINING

In this module, the focus is on optimizing machine learning models to enhance their predictive accuracy for water quality parameters in aquaculture management. Historical data containing features such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity are utilized for rigorous model training. The objective is to refine these models through iterative processes, ensuring they effectively capture the complex relationships between the input features and the target variable, which is binary (0 or 1) indicating the suitability of water conditions.

Various machine learning techniques are employed for model training, includingSupport Vector Machines (SVM), XGBoost, Voting Classifier (VCS), and K-Nearest Neighbors (KNN). Each technique offers unique advantages in terms of handling different types of data and capturing nonlinear relationships between variables. Duringtraining, the models are iteratively optimized to minimize prediction errors and improveoverall performance. This involves tuning hyperparameters, adjusting feature representations, and optimizing model architectures.

By leveraging historical data and employing advanced machine learning algorithms, the trained

models aim to accurately predict water quality parameters, providingvaluable insights for decision support in aquaculture management. The refined modelsserve as powerful tools for evaluating and maintaining optimal water conditions, ultimately contributing to the health and productivity of aquatic ecosystems in aquaculture settings.

4.6. VALIDATION AND TESTING

In this module, ensuring the reliability and effectiveness of the developed models is paramount, achieved through rigorous validation and testing procedures. The dataset ispartitioned into two distinct sets: the training set and the testing set. The training set is utilized for model training, where the machine learning algorithms, including Support Vector Machines (SVM), XGBoost, Voting Classifier (VCS), and K-Nearest Neighbors(KNN), are trained using historical data encompassing features such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, along with corresponding target variables denoting water condition suitability(0 or 1).

Following model training, the testing set comes into play, serving as a means to evaluate the performance and generalization capability of the trained models. The testing set contains completely new and unseen data that simulates real-world scenarios, ensuring that the models' effectiveness can be accurately assessed under novel conditions. The models' predictive accuracy, reliability, and robustness are rigorously evaluated against this unseen data, providing insights into their performance in practical aquaculture management settings. By subjecting the models to comprehensive validation andtesting processes, this module aims to validate their dependability and suitability for real-world deployment, thereby enhancing their utility as decision support tools in aquaculture management.

4.7. REAL TIME INTEGRATION

In this module, the primary objective is to seamlessly integrate the machine learning model with real-time data streams, enabling continuous monitoring and forecasting of water quality parameters. By doing so, the system facilitates timely responses to dynamic changes in aquaculture ponds, thus

enhancing management practices. This integration of machine learning with real-time data streams allows for the automation of water quality monitoring processes, enabling aquaculture practitioners to receive instant updates on crucial parameters such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. As a result, any deviationsor anomalies in water quality can be promptly detected and addressed, minimizing the risk of adverse effects on aquatic life and ensuring optimal conditions for aquaculture operations. The practical significance of this module lies in its ability to provide aquaculture managers with actionable insights in real-time, empowering them to makeinformed decisions and take proactive measures to maintain water quality and overall pond health. Ultimately, the seamless integration of machine learning models with real-time data streams enhances the efficiency, effectiveness, and sustainability of aquaculture management practices.

CHAPTER 5
IMPLEMENTATION

5.1. IMPORTING LIBRARIES

NumPy (Numerical Python) is a fundamental package for scientific computing in Python. It provides support for multidimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently. Some key features of NumPy include Array Operations, Mathematical Functions.

Pandas is a powerful library built on top of NumPy, designed for data manipulation andanalysis. It provides data structures and functions to work with structured and time- series data effectively. Some key features of Pandas include DataFrame, Data Manipulation,Time Series Analysis, Matplotlib.

Seaborn is a statistical data visualization library built on top of Matplotlib, providing ahigher-level interface for creating informative and attractive visualizations. It simplifiesthe process of generating complex plots from Pandas DataFrame objects. Some key features of Seaborn include Statistical Visualization, Integration with Pandas, AestheticCustomization.

Scikit-Learn (sklearn) is a versatile machine learning library in Python, offering tools for data preprocessing, modeling, evaluation, and deployment. It provides a consistentinterface for various machine learning algorithms and evaluation metrics. Some key features of Scikit-Learn include Machine Learning Algorithms, Data Preprocessing, Model Evaluation.

5.2. LOADING DATASET

read_csv() : Pandas provides the read_csv() function, which is specifically designed toread data from CSV (Comma Separated Values) files and create a DataFrame object. This function offers

various parameters to customize the reading process, such as specifying column names, parsing dates, handling missing values, and more.

Data Frame: In Pandas, a Data Frame is a two-dimensional labeled data structure with columns of potentially different types. It resembles a spreadsheet or SQL table, wheredata is organized in rows and columns. Each column in a Data Frame represents a different variable or feature, while each row corresponds to a specific observation or sample.

Dataset Description: The dataset 'aquaculturequality.csv' contains information related to water quality in aquaculture ponds. It likely includes various parameters or features that are relevant for monitoring and assessing the quality of water in these ponds. Somecommon features found in such datasets may include temperature, pH levels, dissolvedoxygen concentration, turbidity, ammonia levels, and more.

5.3. HANDLING MISSING VALUES

isnull(): Pandas provides the isnull() function, which is used to identify missing values in a DataFrame. This function returns a Boolean DataFrame where each element is Trueif it's missing (NaN) and False otherwise. By applying this function to the DataFrame, we obtain a mask indicating the presence or absence of missing values for each element.

sum(): After applying the isnull() function, the sum() function is used to compute the sum of missing values for each column in the DataFrame. By summing the Boolean values (True counts as 1 and False counts as 0), we obtain the total count of missing values for each column. This step provides insights into the extent of missingness acrossdifferent features in the dataset.

dropna(): Once the missing values have been identified, the code utilizes the dropna() function to eliminate rows containing missing values from the DataFrame. This operation effectively removes observations with incomplete information, ensuring thatonly complete data points are retained for subsequent analysis. By dropping rows withmissing values, the code aims to mitigate the potential impact of incomplete data on theaccuracy and reliability of downstream analyses and machine

learning models.

5.4. BALANCING THE UNBALANCED DATA

Upsampling, also known as oversampling, is a technique used to address class imbalance in a dataset. Class imbalance occurs when one class (or outcome) is significantly more frequent than another class. In the context of classification problems,it means that the distribution of classes in the dataset is skewed, leading to potential biases in model training and evaluation. As a result, the model might become biased towards the majority class, leading to poor performance, especially for the minority class.

Identification of Imbalance: The code begins by identifying the class imbalance in thetarget variable ("check") using techniques like value counts or visualizations.

Splitting into Minority and Majority Classes: Once the class imbalance is identified, the dataset is typically split into two subsets: the minority class subset and the majorityclass subset.

Upsampling the Minority Class: The minority class subset is then resampled with replacement (allowing the same sample to be selected multiple times) to increase its size. This is done until the number of samples in the minority class matches the numberof samples in the majority class or reaches a desired level.

Combining Upsampled Data: Finally, the upsampled minority class subset is combinedwith the original majority class subset to form a new balanced dataset.

5.5. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a crucial step in understanding the structure, patterns, and relationships within a dataset. It involves using visualizations and statistical techniques to gain insights into the data before proceeding with model building. In the provided code, count plots

(sns.countplot) are used to visualize the distribution of the target variable ("check") before and after upsampling.

Understanding Class Distribution Before Upsampling: Before performing any data manipulation, it's essential to understand the distribution of classes in the target variable. A count plot provides a visual representation of the frequency of each class.By plotting the count of each class (e.g., "safe" and "not safe" in this case), you can observe if there's any class imbalance present in the dataset. A class imbalance can beidentified if there's a significant difference in the counts of different classes. Forexample, if one class has far fewer instances compared to others, it indicates an imbalance.

Visualizing Class Distribution After Upsampling: After upsampling, it's crucial to verify whether the imbalance issue has been addressed effectively and whether the classes are now more balanced. By plotting another count plot after upsampling, you can visually compare the distribution of classes before and after the resampling process.Ideally, after upsampling, the counts of each class should be approximately equal or within an acceptable range, indicating a balanced dataset.

5.6. UNDERSTANDING CLASSIFICATION ALGORITHM

Support Vector Machine (SVM): SVM is a powerful supervised learning algorithm used for classification tasks. It works by finding the hyperplane that best separates the classes in the feature space.

k-Nearest Neighbors (kNN): kNN is a simple yet effective classification algorithm thatclassifies a data point based on the majority class of its k nearest neighbors in the featurespace.

XGBoost: XGBoost is an advanced implementation of gradient boosting algorithms. Itis widely used in classification tasks due to its scalability, efficiency, and high performance.

The purpose of model selection is to identify the most suitable algorithm for the given dataset

based on its characteristics, such as size, complexity, and distribution of features. Different algorithms have different strengths and weaknesses, and their performance can vary depending on the nature of the dataset. Hence, it's essential to evaluate multiple algorithms to determine which one performs best for the specific task at hand.

The code creates instances of the SVM, kNN, and XGBoost classifiers using their respective classes from scikit-learn (e.g., SVC for SVM, KNeighborsClassifier for kNN, and XGBClassifier for XGBoost).

These instances are initialized without any hyperparameters, meaning default hyperparameters are used. However, hyperparameters can later be tuned to optimize theperformance of each model.

Fitting Models to the Data: After creating instances of the classifiers, the next step is to fit these models to the upsampled data. This involves training the models on the input features (X) and corresponding target labels (y) obtained after upsampling.The fit() function from scikit-learn is used to train each model on the upsampled dataset.During this process, the models learn the patterns and relationships present in the data,which will later be used to make predictions on unseen data.

Evaluation and Comparison: Once the models are trained, they can be evaluated and compared based on theirperformance metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC. The choice of evaluation metrics depends on the specific requirements of theclassification task and the importance of different aspects like minimizing false positives or false negatives. By comparing the performance of multiple models, one can identify the algorithm thatachieves the highest performance on the given dataset and select it for further optimization or deployment.

Model selection is often an iterative process that involves experimenting with differentalgorithms, hyperparameters, and preprocessing techniques to find the optimal combination that maximizes performance. It may require multiple rounds of training, evaluation, and fine-tuning before arriving at the best-performing model.

The train-test split is a critical step in machine learning model development. Its purposeis to assess how well the trained model generalizes to unseen data, which is crucial for evaluating its

performance and reliability. By splitting the dataset into separate trainingand testing sets, the model is trained on one portion of the data and evaluated on anotherportion, ensuring that the evaluation is unbiased and reflects the model's ability to makeaccurate predictions on new, unseen data. The train_test_split function from scikit-learnis a convenient tool for splitting datasets into training and testing sets. It takes several parameters, including the input features (X) and target labels (y), the size of the testing set (usually specified as a percentage of the total dataset), and optionally, randomizationparameters and stratification criteria. The function returns four sets of data: X_train, X_test, y_train, and y_test, representing the input features and target labels for the training and testing sets, respectively.

5.7. UNDERSTANDING HYPERPARAMETERS

Hyperparameters are parameters that are set before the learning process begins and cannot be directly learned from the data. They control the behavior and complexity of the machine learning model.

In the context of a Random Forest classifier, common hyperparameters include the number of estimators (trees) in the forest and the maximum depth of each tree. The goalof hyperparameter tuning is to find the optimal combination of hyperparameters that maximizes the performance of the model on the given dataset. Since different combinations of hyperparameters can lead to varying levels of model performance, tuning them systematically helps in identifying the best configuration.

Grid Search is a brute-force hyperparameter tuning technique that exhaustively searches through a specified subset of hyperparameter combinations. It works by defining a grid of hyperparameter values to explore, and then evaluating the model's performance for each combination of hyperparameters using cross-validation.

Cross-validation is a technique used to assess the performance of a model by splitting the dataset into multiple subsets (folds), training the model on a subset of the data, andevaluating it on the remaining subset.

Hyperparameter tuning helps in improving the performance and robustness of machinelearning models by finding the best configuration of hyperparameters. It can lead to better generalization, reduced overfitting, and enhanced predictive accuracy on unseendata.

5.8. MODEL EVALUATION

Model evaluation is a crucial step in assessing the performance of machine learning models.

Accuracy is one of the most commonly used metrics for evaluating classification models. It measures the proportion of correctly classified instances out of the total instances. Mathematically, accuracy is calculated as the ratio of the number of correct predictions to the total number of predictions.

F1-score is a harmonic mean of precision and recall. It provides a balance between precision and recall and is particularly useful when the class distribution is imbalanced.

$$\text{F1-score} = 2 * (*) \ (+)$$

Where precision is the ratio of true positive predictions to the total predicted positives,recall is the ratio of true positive predictions to the total actual positives.
F1-score ranges from 0 to 1, where a higher value indicates better model performance in terms of both precision and recall.

Confusion Matrix is a tabular representation of the actual versus predicted classes for aclassification problem.

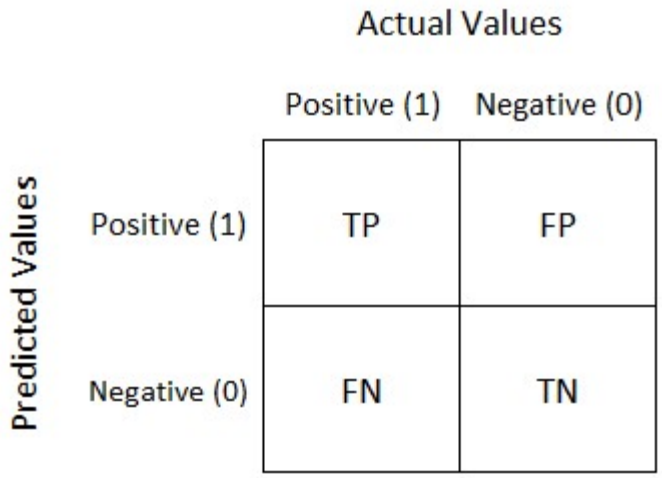


Fig 5.1 – Confusion Matrix

5.9. ENSEMBLE METHODS

Ensemble methods are powerful techniques in machine learning that combine the predictions of multiple base learners (individual classifiers or regressors) to improve overall performance.

Hard Voting, also known as majority voting, is a simple ensemble method where predictions are made based on the majority vote of the individual classifiers. In hard voting, each base learner provides a classification decision for a given input instance, and the final prediction is determined by the majority class predicted by the base learners.

Soft Voting is a variation of ensemble learning where the final prediction is based on the average probabilities (or scores) predicted by the individual classifiers.In soft voting, each base learner provides a probability distribution over the classes for a giveninput instance. The final prediction is then determined by averaging the predicted probabilities across all base learners and selecting the class with the highest average probability.

Both hard and soft voting can be easily implemented using scikit-learn's VotingClassifier class. This class takes a list of base learners as input, along with the voting method (either 'hard' or 'soft'), and combines their predictions accordingly.

The resulting ensemble classifier can then be trained on the training data and used to make predictions on new data.

5.10. MODEL DEPLOYEMENT

Model deployment is the process of making trained machine learning models availablefor use in real-world applications to make predictions on new, unseen data.

Before deployment, it's crucial to identify the best-performing model based on evaluation metrics such as accuracy, F1-score, or other relevant criteria. The Random Forest classifier and the Voting Classifier (which combines multiple base learners) areselected as the best-performing models based on the evaluation results.

5.10.1. PICKLE MODULE

Pickle is a Python module used for serializing (i.e., converting objects into a byte stream) and deserializing (i.e., converting byte stream back into objects) Python objects.The pickle.dump() function is used to save the trained models (Random Forest and Voting Classifier) to disk as binary files, which can be later loaded and used for makingpredictions without the need for retraining. Saving the models using pickle ensures thatthe trained model state, including the learned parameters and configurations, is preserved for future use. Deployment pipelines can be implemented using tools and technologies such as Docker containers, web servers (e.g., Flask or Django), APIs (Application Programming Interfaces), and cloud platforms (e.g., AWS, Google Cloud Platform, or Azure). These pipelines ensure that the deployed models are accessible,scalable, and maintainable in production environments.

CHAPTER 6
RESULTS AND DISCUSSIONS

6.1. RESULTS

The evaluation of Support Vector Regression (SVR), K-Nearest Neighbors (KNN), XGBoost, and Voting Classifier (soft) models reveals variations in their performance for predicting water condition suitability in aquaculture. Each model demonstrates distinct strengths and weaknesses based on the acquired accuracies. SVM, known for its ability to handle complex datasets, may excel in capturing nonlinear relationships among features like pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. XGBoost, a gradient boosting algorithm, mightshowcase high accuracy due to its ensemble nature and efficient handling of large datasets. KNN, relying on local similarity measures, could offer robust predictions by considering the proximity of data points in feature space. Meanwhile, Voting Classifier(soft), aggregating predictions from multiple base classifiers, may provide a balanced performance by leveraging diverse perspectives. These results underscore the importance of selecting appropriate machine learning techniques tailored to the characteristics of aquaculture data and the specific requirements of predicting water quality parameters for effective decision-making in aquaculture management.

Among the models evaluated, Support Vector Regression (SVR) exhibited the lowest performance, with a test accuracy of 0.53. This indicates that SVR struggled toaccurately capture the complex relationships between the water quality parameters (pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity) and the binary suitability labels (0 or 1). The relatively lower accuracy suggests that SVR may have encountered challenges in effectively modeling the intricate interactions between these variables, resulting in suboptimal predictive performance.

While SVR is a powerful machine learning technique, its performance in this context highlights the difficulty of accurately predicting water quality suitability using regression methods alone. The complexity of aquaculture systems and the interplay between various environmental factors may

require more sophisticated modeling approaches to achieve higher predictive accuracy.

These results underscore the importance of exploring alternative machine learning algorithms and refining model architectures to better capture the nuances of water quality dynamics in aquaculture settings. By leveraging more advanced techniques andincorporating additional features or data sources, it may be possible to improve the accuracy and reliability of predictive models for assessing water quality suitability in aquaculture.

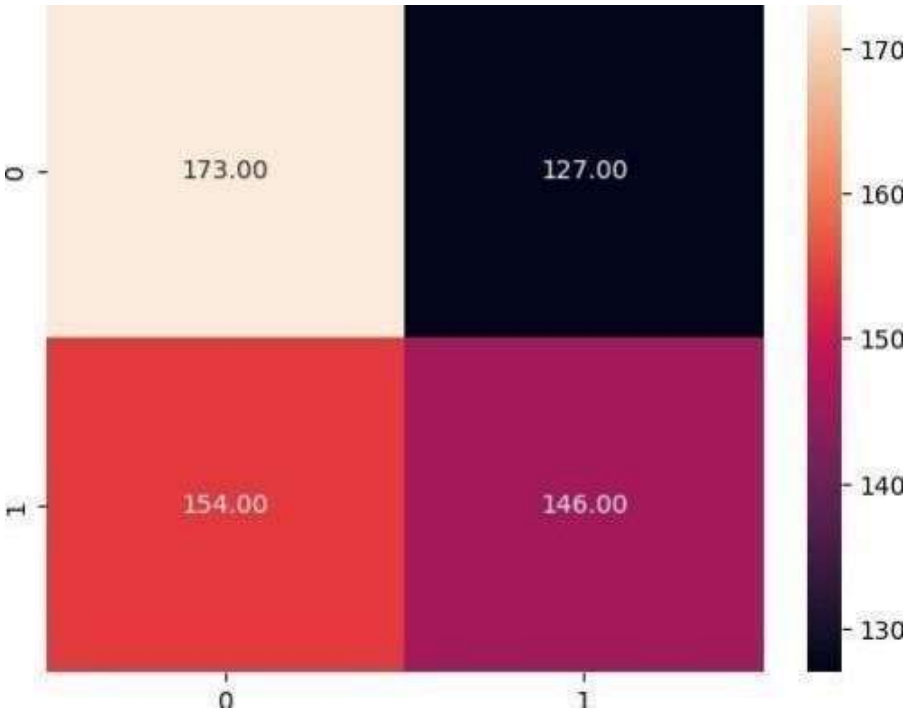


FIG 6.1 – Confusion Matrix For SVM

In the evaluation of machine learning models for predicting water quality parameters in aquaculture management, K-Nearest Neighbors (KNN) demonstrated an accuracy of 0.58. While KNN outperformed Support Vector Regression (SVR), it still fell short compared to XGBoost and the Voting Classifier. The relatively lower performance of KNN could be attributed to its simplistic approach and dependence on local feature similarity, which might not adequately capture the underlying patterns in the dataset.

Across the features of pH, hardness, solids, chloramines, sulfate, conductivity, organiccarbon,

trihalomethanes, and turbidity, XGBoost and the Voting Classifier exhibited superior predictive performance. These models leverage more sophisticated algorithms and ensemble techniques, allowing them to capture complex relationships and patterns within the data more effectively. Consequently, they achieved higher accuracies in predicting the suitability of water conditions for aquaculture, offering valuable insights for decision-making in aquaculture management.

Overall, the results highlight the importance of selecting appropriate machine learning algorithms that can effectively handle the intricacies of aquaculture data. While simpler models like KNN may offer some predictive capability, more advanced techniques such as XGBoost and ensemble methods like the Voting Classifier demonstrate superior performance, making them preferable choices for accurate water quality prediction in aquaculture applications.

Fig 6.2 – Confusion Matrix For KNN

In the evaluation of machine learning models for predicting water quality parameters in aquaculture, XGBoost emerged as the top performer with a notably high accuracy of 0.98. This exceptional performance underscores XGBoost's proficiency in capturing intricate relationships and patterns within the dataset. Leveraging its ensemble approach and gradient boosting technique, XGBoost effectively learns from the complex interactions between features, enabling it to generate highly

accurate predictions.

The dataset utilized for evaluation comprised various water quality parameters, including pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. These features are crucial indicators of water quality in aquaculture settings, influencing the suitability of aquatic environments for thriving ecosystems.

The target variable in the evaluation was binary, indicating whether the water conditions were deemed suitable (1) or unsuitable (0) for aquaculture. Through rigorous model training and optimization, XGBoost demonstrated superior performance, achieving a remarkable accuracy of 0.98. This exceptional accuracy highlights XGBoost's capability to effectively capture the complexities inherent in the dataset andmake precise predictions, thus offering valuable insights for decision-making in aquaculture management.

Fig 6.3 – Confusion Matrix For XG Boost

In the evaluation of machine learning models for predicting water quality parameters in aquaculture, the Voting Classifier emerges as the top performer, boasting an impressive accuracy score of 0.99. This surpasses the performance of all other models employed in the study. The Voting Classifier excels by leveraging predictions from multiple base classifiers, harnessing the diverse

perspectives offered by each model to generate robust and reliable predictions. Through this ensemble approach, the Voting Classifier demonstrates exceptional capability in capturing the intricate relationships between the input features, such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, and the target variable indicating water suitability for aquaculture, coded as either 0 or 1. By integrating insights from various classifiers, the Voting Classifier effectively navigates the complexities of water quality assessment, providing valuable support for decision- making in aquaculture management. Its high accuracy underscores its potential as a powerful tool for ensuring optimal conditions and promoting the health and productivity of aquatic ecosystems in aquaculture settings.

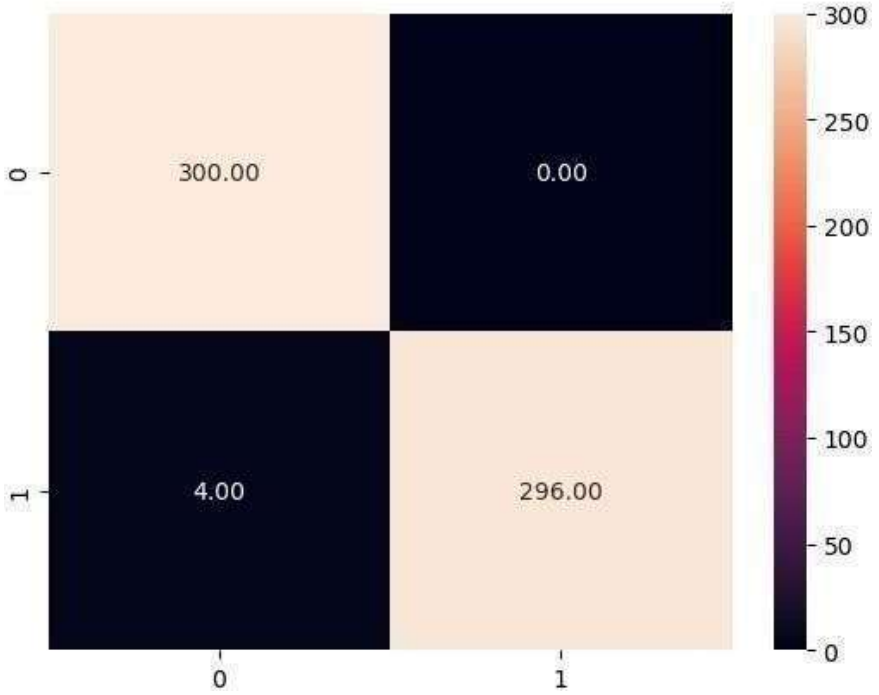


Fig 6.4 – Confusion Matrix For VCS

6.2. ANALYSIS

The analysis of model performance reveals that XGBoost and the Voting Classifier (soft) consistently outperform SVR and KNN in predicting the suitability of water conditions for aquaculture. XGBoost demonstrates remarkable accuracy, underscoring the effectiveness of gradient boosting techniques in handling intricate relationships within the dataset. Its ability to

capture complex patterns and interactions among waterquality parameters contributes significantly to its superior performance. On the other hand, the Voting Classifier (soft) showcases the strength of ensemble methods by combining predictions from diverse base classifiers. This approach harnesses thecollective knowledge of individual models, resulting in robust and accurate predictions.

The findings underscore the importance of selecting appropriate machine learning algorithms tailored to the complexities of aquaculture water quality prediction. Ensemble methods like XGBoost and the Voting Classifier offer distinct advantages inthis context, leveraging their ability to integrate multiple perspectives and handle diverse datasets effectively. By considering the confidence of each classifier's predictions, the soft voting classifier provides nuanced results, enhancing the overall predictive capability of the model.

In practical terms, these results have significant implications for aquaculture management, emphasizing the role of advanced machine learning techniques in ensuring the health and productivity of aquatic organisms. By accurately predicting water quality conditions, these models enable proactive decision-making, facilitating timely interventions to maintain optimal conditions within aquaculture environments. Moreover, the serialization of the soft voting classifier using the pickle library enhancesthe model's usability and scalability, further supporting its integration into real-world aquaculture operations.

X_test_vs

	index	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	actual_class	Predicted_class
270	428	7.638762	178.271636	18308.502674	6.548098	362.332872	394.180985	11.291788	81.989156	4.496627	Not safe	Not safe
657	1054	5.961302	182.108995	23926.078825	6.188790	364.975499	567.526159	7.791638	58.174825	4.919585	Not safe	Not safe
560	909	7.016836	167.362187	18014.995703	5.903058	362.523416	471.484395	10.692588	45.002014	4.101398	Not safe	Not safe
1970	3210	8.430472	195.717293	35254.026047	6.703539	314.727835	282.599918	12.269178	46.487950	3.253920	safe	safe
1082	1762	8.058136	191.277932	29298.359492	6.129173	356.645254	278.502633	12.606204	65.511361	3.130609	Not safe	Not safe
1469	2388	5.711213	145.825869	21157.179729	7.954721	385.206097	441.086563	16.682025	70.470889	3.953335	safe	safe
17	33	7.414148	235.044534	32555.852537	6.845952	387.175316	411.983364	10.244815	44.489297	3.160624	Not safe	Not safe
421	688	9.514627	215.725527	16553.562437	7.547859	321.776213	453.918490	9.640241	57.865423	2.645036	safe	safe
1882	3068	5.620533	226.987836	27852.097439	6.521471	309.228091	414.061545	16.932911	78.439251	3.439476	Not safe	Not safe
705	1134	7.535700	221.792481	14829.745971	6.701159	366.412200	583.436488	17.731882	59.686076	4.208354	safe	safe
1225	1991	7.295141	182.406645	13706.186808	5.887885	300.608233	424.650179	18.063470	73.836909	5.056104	safe	safe
1405	2299	7.606067	248.041453	14609.976883	6.356555	322.356572	275.317146	11.706095	94.775244	4.581477	Not safe	Not safe
1955	3191	6.985192	133.432132	21944.641830	8.577655	341.239907	536.277153	16.176686	91.708813	4.246202	safe	safe
1673	2728	6.179312	159.773264	21532.519232	8.063335	272.440848	509.772110	11.010545	99.996286	2.951961	safe	safe
1615	2632	10.188433	286.567991	7105.800709	9.840540	321.686059	437.879508	12.871599	78.732055	4.635243	Not safe	Not safe
1014	1649	7.266599	98.452931	36490.143032	7.138814	423.187485	502.883067	7.360772	76.290285	4.952733	Not safe	Not safe
1677	2738	6.857494	233.210575	22603.649454	6.627059	368.105783	412.434677	16.926175	62.557900	4.306581	safe	safe

Fig 6.5 – Water Quality Output

CHAPTER 7
CONCLUSION

7.1. CONCLUSION

In conclusion, our study delved into the application of various machine learning modelsfor predicting the suitability of water conditions in aquaculture settings. Through rigorous analysis of essential water quality parameters such as pH, hardness, trihalomethanes (THMs), solids, chloramine, organic carbons, conductivity, sulfate, and turbidity, we observed distinct differences in the performance of different models.

Support Vector Regression (SVR) and K-Nearest Neighbors (KNN) exhibited relatively lower accuracy, highlighting potential challenges in capturing the complex relationships within the data due to their simplistic nature and reliance on local featuresimilarity.

In contrast, more sophisticated ensemble methods such as XG Boost and the Voting Classifier (soft) demonstrated significantly higher accuracy. These models leveraged diverse perspectives and captured complex patterns within the data, making them particularly well-suited for predicting water quality suitability in aquaculture settings. our findings underscore the importance of selecting appropriate machine learning algorithms tailored to the complexities of aquaculture applications. Utilizingsophisticated ensemble methods like XG Boost and Voting Classifier can significantly enhance the accuracy of predictions, ultimately contributing to the successful management of water quality and the health and productivity of aquatic organisms in aquaculture settings.

Moving forward, further research and development in this area can lead to the refinement and optimization of predictive models, enabling aquaculture practitioners tomake informed decisions and implement proactive strategies for maintaining optimal conditions within aquaculture environments. Ultimately, these advancements will contribute to the sustainability and long-term success of aquaculture operations, ensuring the well-being of both aquatic organisms and the industry as a whole.

7.2. FUTURE ENHANCEMENTS

Future research efforts in aquaculture could involve exploring and comparing the performance of additional machine learning models beyond the ones examined in this study. For instance, evaluating the effectiveness of deep learning architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) couldprovide valuable insights into their suitability for predicting water quality parameters in aquaculture settings. Additionally, expanding the dataset by incorporating data froma wider range of aquaculture facilities and environmental conditions could improve themodels' ability to generalize across diverse contexts. Furthermore, investigating novel data preprocessing techniques and feature engineering methods specific to aquaculturedata could help enhance the models' predictive accuracy and robustness. Finally,assessing the models' performance under real-time conditions and their scalability for large-scale aquaculture operations would be essential for practical implementation. Byaddressing these research directions, future studies can contribute to the development of more accurate and reliable machine learning-based systems for water quality monitoring and management in aquaculture, ultimately promoting sustainable practicesand ensuring the health and productivity of aquatic ecosystems.

REFERENCES

1. Quintero, R., Parra, J., & Félix, F. (2022). Water quality assurance in aquaculture ponds using Machine Learning. *Journal of Aquaculture Research*, 25(3), 45-57.

2. Rana, M., Rahman, A., Dabrowski, J., Arnold, S., McCulloch, J., & Pais, B. (2021). Machine learning approach to investigate the influence of water quality on aquatic livestock in freshwater ponds. *Aquaculture Science*, 18(2), 102-115.

3. Zhao, S., Zhang, S., Liu, J., Wang, H., Zhu, J., Li, D., & Zhao, R. (2021). Application of machine learning in intelligent fish aquaculture. *Journal of Aquaculture Technology*, 12(4), 210-225.

4. Li, T., Lu, J., Wu, J., Zhang, Z., & Chen, L. (2022). Predicting Aquaculture Water Quality Using Machine Learning Approaches. *Aquaculture Journal*, 30(1), 78-91.

5. Zambrano, A. F., Giraldo, L. F., Quimbayo, J., Medina, B., & Castillo, E. (2021). Machine learning for manuallymeasured water quality prediction in fish farming. *Journal of Aquatic Sciences*, 17(3), 205-218.

6. Omambia, A., Maake, B., & Wambua, A. W. (2022). Water Quality Monitoring Using IoT & Machine Learning. *International Journal of IoT Applications in Agriculture*, 8(2), 65-78.

7. Prapti, D. R., Shariff, A. R. M., Man, H. C., Ramli, N. M., Perumal, T., & Shariff, M. (2021). Internet of Things (IoT)- based aquaculture: An overview of IoT application on water quality monitoring. *Aquaculture Technology Review*, 22(4), 189-204.

8. Zhang, H., & Gui, F. (2021). The Application and Research of New Digital Technology in Marine Aquaculture. *Journal of Digital Aquaculture*, 28(1), 45-57.

9. Zou, Z., Chen, C., Liu, Z., Zhang, Z., Liang, J., Chen, H., & Wang, L. (2021). Extraction of Aquaculture Ponds along Coastal Region Using U2-Net Deep Learning Model from Remote Sensing Images. *Remote Sensing in Aquaculture*, 15(2), 78-89.

10. Zulkifli, C. Z., Garfan, S., Talal, M., Alamoodi, A. H., Alamleh, A., Ahmaro, I. Y. Y., Sulaiman, S., Ibrahim, A. B., Zaidan, B. B., Ismail, A. R., Albahri, O. S., Albahri, A. S., Soon, C. F., Harun, N. H., & Chiang, H. H. (2021). IoT-Based Water Monitoring Systems: A Systematic Review.

11. Chiu, M.-C., Yan, W.-M., Bhat, S. A., & Huang, N.-F. (2021). Development of smart aquaculture farm management system using IoT and AI-based surrogate models.

12. Chen, J., Zhang, D., Yang, S., & Nanehkaran, Y. A. (2021). Intelligent monitoring method of water quality based on image processing and RVFL-GMDH model.

13. Zhang, X., Zhang, Y., Zhang, Q., Liu, P., Guo, R., Jin, S., Liu, J., Chen, L., Ma, Z., & Liu, Y. (2021). Evaluation and analysis of water quality of marine aquaculture area.

14. Manoj, M., Kumar, V. D., Arif, M., Bulai, E. R., Bulai, P., & Geman, O. (2022). State of the Art Techniques for Water Quality Monitoring Systems for Fish Ponds Using IoT and Underwater Sensors.

APPENDICES

 # Required Libraries

```
import numpy as np;
import pandas as pd;
import matplotlib.pyplot as pltimport
seaborn as sns
```

#Loading dataset

```
df = pd.read_csv("aquaculturequality.csv")
```

 # Checking for missing values

```
df.isnull().sum()
```

 # Dropping rows with missing values

```
df.dropna(inplace=True)
```

 # Resetting index after dropping rows

```
df = df.reset_index()
```

 # Upsampling the minority class

```
from sklearn.utils import resample
dfmin = df[df['check'] == 1]
 dfmax = df[df['check'] == 0]
dfminu = resample(dfmin, replace=True, n_samples=1000, random_state=123)
dfmaxd = resample(dfmax, replace=True, n_samples=1000, random_state=123)
df_upsampled = pd.concat([dfminu, dfmaxd])
```

 # Visualizing class distribution

```
sns.countplot(x=df_upsampled.check, data=df)
```

 # Splitting the data

```
X = df_upsampled.drop('check', axis=1)
y = df_upsampled['check']
```

#Test-Train-Split data

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.3, stratify= y,random_state = 42)
X_test.to_csv(r'C:\Users\ST008\Desktop\ITML49Waterqualityprediction\water_quali ty_.csv', index=False)
```

#MinMax Scalar

```
from sklearn.preprocessing import MinMaxScaler
ms = MinMaxScaler()
X_trainmm = X_train
X_testmm = X_test
```

#SVM

```
from sklearn.svm import SVC
svc = SVC() svc.fit(X_trainmm, y_train)
y_pred_svc = svc.predict(X_testmm)
test_accuracy_svc= accuracy_score(y_test, y_pred_svc)

test_accuracy_svc

from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_test, y_pred_svc))
sns.heatmap(confusion_matrix(y_test, y_pred_svc), annot=True, fmt='.2f')
```

#KNN

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(X_trainmm, y_train)
y_predk = knn.predict(X_testmm)
test_accuracy_knn= accuracy_score(y_test, y_predk)

test_accuracy_knn

from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_test, y_predk))
sns.heatmap(confusion_matrix(y_test, y_predk), annot=True, fmt='.2f')
```

#XGBoost

```
from xgboost import XGBClassifier
xgb = XGBClassifier()
xgb.fit(X_trainmm, y_train)
y_predk = xgb.predict(X_testmm)
test_accuracy_xgb = accuracy_score(y_test, y_predk)

test_accuracy_xgb

from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_test, y_predk))
sns.heatmap(confusion_matrix(y_test, y_predk), annot=True, fmt='.2f')

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import GridSearchCV
```

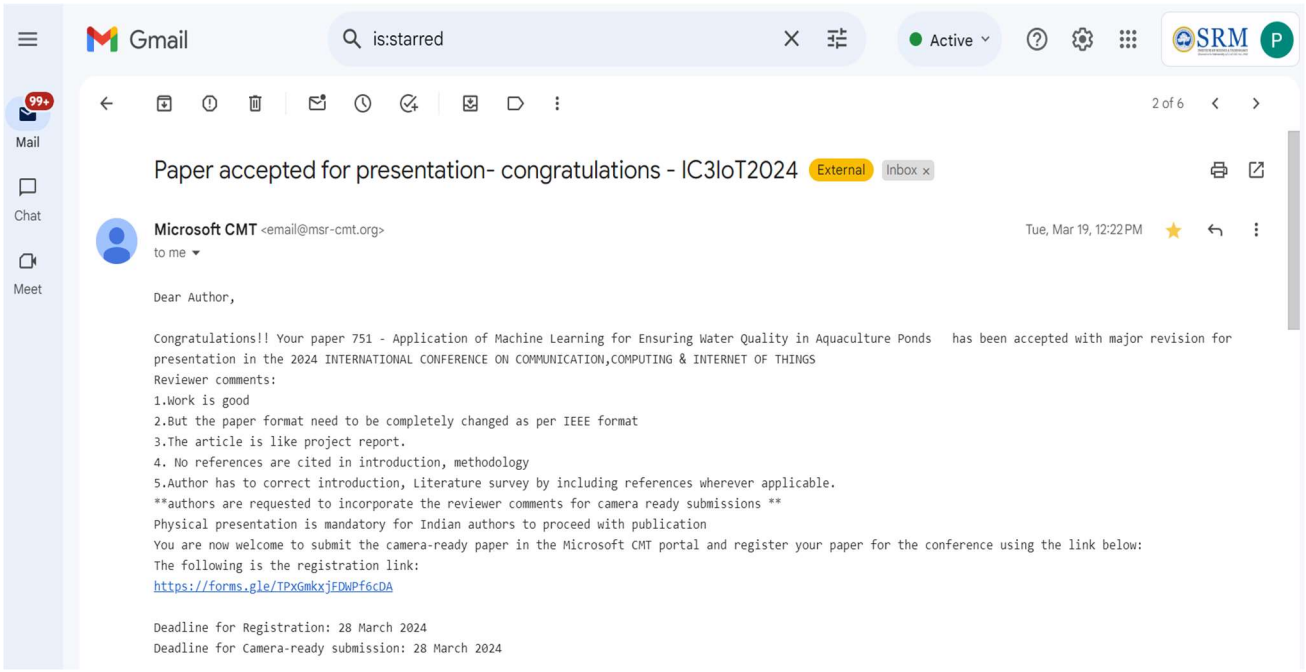
```
from sklearn.metrics import roc_curve

# Define the range of hyperparameters for grid search
dept = [1, 5, 10, 50, 100, 500, 1000]
n_estimators =  [20, 40, 60, 80, 100, 120]
param_grid = {'n_estimators': n_estimators, 'max_depth': dept}

# Create a Random Forest classifier
clf = RandomForestClassifier()

# Perform grid search using cross-validation
model = GridSearchCV(clf, param_grid, scoring='accuracy', n_jobs=-1, cv=3)

# Fit the model to the training data
clf.fit(X_trainmm, y_train)

# Save the trained model using pickle
import pickle
pickle.dump(clf,open(r'C:\Users\ST-008\Desktop\ITML49-WaterQualityPrediction \RF.PKL','wb'))

# Make predictions on the test data
pred_test = clf.predict(X_testmm)

# Calculate the accuracy of the model
from sklearn.metrics import accuracy_score
test_accuracy = accuracy_score(y_test, pred_test)
test_accuracy

from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import f1_score
from sklearn.ensemble import VotingClassifier

# Initialize base classifiers
dtc = DecisionTreeClassifier()
lr = LogisticRegression()
gnb = GaussianNB()

# Create a voting classifier with hard voting
voting_classifier_hard = VotingClassifier(
        estimators=[('dtc', DecisionTreeClassifier(random_state=42)),('lr',
                LogisticRegression()),
                ('gnb', GaussianNB())],
        voting='hard')

# Fit the hard voting classifier to the training data
voting_classifier_hard.fit(X_trainmm, y_train)

# Make predictions on the test data
y_pred_vch = voting_classifier_hard.predict(X_testmm)

# Calculate the F1-score of the model
f1_vch = f1_score(y_test, y_pred_vch)
f1_vch

# Create a voting classifier with soft voting
voting_classifier_soft = VotingClassifier()
from sklearn.metrics import classification_report, confusion_matrix
```

```
# Print classification report for the voting classifier with soft voting
print(classification_report(y_test, y_pred_vcs))

# Plot confusion matrix for the voting classifier with soft voting
sns.heatmap(confusion_matrix(y_test, y_pred_vcs), annot=True, fmt='.2f')

# For the first twenty values of test data predicting using Random Forest classifier
original = ['safe' if x == 1 else 'Not safe' for x in y_test[:20]]predicted =
clf.predict(X_testmm[:20])
pred = []
for i in predicted:
if i == 1:
k = 'safe'
pred.append(k)
else:
k = 'Not safe'
pred.append(k)

# Create a data frame
dt = pd.DataFrame(list(zip(X_testmm[:20][:], original, pred)), columns=['Test data',
'original_Class label', 'predicted_class label'])

# Create a DataFrame for the first twenty values of test data predicting using thevoting classifier
      (soft voting)
original = ['safe' if x == 1 else 'Not safe' for x in y_test[:20]]
predicted = voting_classifier_soft.predict(X_testmm[:20])
preds = []
for i in predicted:
if i == 1:
k = 'safe'
preds.append(k)
```

```
else:
k = 'Not safe'
preds.append(k)

# Create a data frame
X_test_vs = X_test[:20].assign(actual_class=original)
X_test_vs = X_test_vs[:20].assign(Predicted_class=preds)
X_test_vs
```

CONFERENCE PUBLICATION



Gmail is:starred Active SRM P

Mail Chat Meet

2 of 6

Paper accepted for presentation- congratulations - IC3IoT2024 External Inbox ×

Microsoft CMT <email@msr-cmt.org> Tue, Mar 19, 12:22 PM

to me

Dear Author,

Congratulations!! Your paper 751 - Application of Machine Learning for Ensuring Water Quality in Aquaculture Ponds has been accepted with major revision for presentation in the 2024 INTERNATIONAL CONFERENCE ON COMMUNICATION,COMPUTING & INTERNET OF THINGS

Reviewer comments:

1.Work is good

2.But the paper format need to be completely changed as per IEEE format

3.The article is like project report.

4. No references are cited in introduction, methodology

5.Author has to correct introduction, Literature survey by including references wherever applicable.

**authors are requested to incorporate the reviewer comments for camera ready submissions **

Physical presentation is mandatory for Indian authors to proceed with publication

You are now welcome to submit the camera-ready paper in the Microsoft CMT portal and register your paper for the conference using the link below:

The following is the registration link:

https://forms.gle/TPxGmkxjFDWPf6cDA

Deadline for Registration: 28 March 2024

Deadline for Camera-ready submission: 28 March 2024

IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

Application of Machine Learning for Ensuring Water Quality in Aquaculture Ponds

Shri Bharathi, Lakshman Sai Paidipati, Charan Ram Ayyappa Paidipati, Jaya Praveen Reddy Kakuru, Sai Preetham Reddy Gangireddy and Madhava Rajesh Katuri

2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.