International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 5, Number 2, July 2026, *Page. 1641 - 1655*
Email : editorijhess@gmail.com

Design Of A Drinking Water Feasibility Identification System Using The Random Forest Algorithm And Principal Component Analysis (PCA)

Khusnul Karomah[1], Asrul Abdullah[2], Barry Ceasar Octariadi[3]
[1,2,3] Informatics Engineering Study Program, Faculty of Engineering and Computer Science, Universitas Muhammadiyah Pontianak

*Corresponding Author
Email : karomahkhusnul5@gmail.com

Abstract

Clean and safe drinking water is a basic need for society. Based on clean water statistical data from Statistic Indonesia (BPS), the average percentage of households in West Kalimantan Province with access to safe drinking water sources in urban areas was 89.21%, while in rural areas it was 77.76% in 2023. With a population growth that outpaces the available reserves of safe drinking water, there is a clear need to monitor and analyze water quality more carefully and on a data-driven basis. This research aims to design and implement a drinking water suitability classification system using the Random Forest method as the main classification algorithm, with Principal Component Analysis (PCA) serving as the data dimensionality reduction technique. The interactive visualization was built using Python and Streamlit, taking into account User Interface (UI) and User Experience (UX) aspects, which are divided into two access levels: Public Access (for the community) and Officer Access (for agencies). Water quality data is identified based on standards from the Ministry of Health (Permenkes No. 492/MENKES/PER/IV/2010), reduced in dimensionality using PCA, and subsequently classified for its suitability using Random Forest. The test results demonstrate that this system is capable of classifying the suitability status with an accuracy of 95.35%. to ensure the model does not experience overfitting, K-Fold Cross-Validation (5-Fold) testing was conducted, yielding an average accuracy of 94.94% and proving the model's stability. Furthermore, the User Acceptance Testing (UAT) results indicate an excellent level of user acceptance regarding the bulk upload feature and the monitoring of drinking water quality in a fast, accurate, and data-driven manner.

Keywords: Drinking Water Quality, Random Forest, PCA, Classification, User Acceptance Testing, UI/UX.

INTRODUCTION

Clean water is one of the most vital natural resources for human life. Its function extends beyond fulfilling basic needs such as drinking, cooking, and bathing, as it also plays a crucial role in various other aspects of life, including public health, agriculture, industry, and ecosystems. According to the Regulation of the Minister of Health of the Republic of Indonesia No. 492/MENKES/PER/IV/2010, drinking water quality must meet specific standards, including a pH value between 6.5 and 8.5, a maximum turbidity level of 5 NTU, and a free chlorine concentration ranging from 0.2 to 0.5 mg/L.

Based on water supply statistics from the Central Bureau of Statistics (BPS), the average percentage of households in West Kalimantan Province with access to adequate drinking water sources in urban areas reached 89.21%, while in rural areas it was 77.76% in 2023. Access to safe water is essential, particularly for preventing infectious diseases such as diarrhea, cholera, and typhoid. Insufficient access to clean water can also reduce agricultural productivity, potentially leading to food crises and affecting food supply. With the increasing population growth compared to the availability of safe drinking water reserves, this condition indicates the need for more precise and data-driven monitoring and analysis of water quality.

One of the main challenges in analyzing water quality for institutions such as regional water supply companies (PDAM) is the time-consuming nature of manual monitoring processes and the complexity of water quality testing parameters. To address this issue, a reliable Machine Learning algorithm for classification is required. The Random Forest algorithm is known for its high accuracy, robustness to outliers, and strong performance in handling classification data. However, before processing data with Random Forest, the high dimensionality of the dataset needs to be simplified without losing essential information. Therefore, Principal Component Analysis (PCA) is employed as a preprocessing technique to reduce data dimensionality and extract key features.

In addition to model accuracy, the effectiveness of a system largely depends on how users interact with it. A system requires an intuitive User Interface (UI) and an efficient User Experience (UX) to facilitate both individual self-checks by the general public and large-scale assessments by officials. Accordingly, this study develops a web-based application using Streamlit, focusing on user-friendliness. To ensure the model does not suffer from overfitting, evaluation is conducted using K-Fold Cross-Validation, while User Acceptance Testing (UAT) is performed to assess system acceptance by end users.

RESEARCH METHODS

The research methodology in this study is carried out through several systematic stages, beginning with a literature review and problem identification related to the limitations of conventional methods in drinking water quality testing, which highlights the need for a faster and more efficient web-based system. The data used in this study was obtained from Perumda Air Minum Tirta Khatulistiwa (January–July 2024) with parameters including pH, turbidity, and free chlorine, which then underwent preprocessing using StandardScaler and dimensionality reduction using PCA to address scale differences and multicollinearity issues. Furthermore, the data was modeled using the Random Forest algorithm and validated using 5-Fold Cross Validation to ensure model stability and prevent overfitting. The system was then designed as a web application with two access levels (public and operator), and finally evaluated using User Acceptance Testing (UAT) to assess functionality, UI/UX, and overall system efficiency.

RESULTS AND DISCUSSION

Machine Learning Model Implementation Results

The implementation of artificial intelligence in this system integrates two main techniques: Principal Component Analysis (PCA) as a data dimensionality reduction method, and Random Forest as the primary classification algorithm. The Python code used to execute this model is designed with a structured workflow (pipeline), ensuring that data flows systematically from the feature extraction stage to the prediction stage.

The first step in the code execution involves defining the PCA algorithm. In this study, the *n_components* parameter is set to 2 (*n_components = 2*). This value is determined based on cumulative variance analysis, where the first two principal components are considered sufficient to represent most of the information (variance) from the original dataset, which includes higher-dimensional features (pH, turbidity, and chlorine). By reducing the data into two dimensions, PCA effectively addresses multicollinearity issues (high correlation between features) and reduces noise, while also enabling easier visualization of the data in a two-dimensional (2D) space.

After the original data matrix is transformed into a new feature space by PCA, the transformed data is then fed into the Random Forest algorithm for model training. In this study, the Random Forest algorithm is configured with the parameter *n_estimators = 300*. This parameter represents the number of decision trees constructed within the model's "forest." The selection of 300 trees is considered optimal, as it provides a balance between high accuracy and computational efficiency. With 300 decision trees performing majority voting for each data sample, the model becomes highly robust against outliers and is capable of forming complex and accurate decision boundaries.

Confusion Matrix Evaluation (Train/Test Split)

To evaluate how well the trained Random Forest model can recognize patterns in drinking water quality, a Train/Test Split method is applied. In this approach, 80% of the dataset is used as training data to build the model, while the remaining 20% serves as testing data that is "unseen" by the model, ensuring an objective evaluation of its predictive performance.

The prediction results on the testing data are then represented using a Confusion Matrix. This matrix is an evaluation table that compares actual labels (real-world water quality conditions) with predicted labels (model outputs). Based on the conducted evaluation, the following distribution is obtained:

1. **True Positive (TP):** This value indicates that 25 water samples are actually classified as “Safe” for consumption, and the model correctly predicts them as “Safe.” A high TP value suggests that the model is highly sensitive in identifying characteristics of clean and safe water.
2. **True Negative (TN):** This value indicates that 16 water samples are actually “Unsafe,” and the model correctly classifies them as “Unsafe.” Accurate detection of negative cases is crucial for water management authorities to prevent contaminated water from entering distribution systems.
3. **False Positive (FP):** Also known as a Type I Error, this occurs when 1 water sample that is actually “Unsafe” is incorrectly predicted as “Safe.” In the context of public health, FP is the most critical error, as it may allow contaminated water to be consumed. Reducing FP to only 1 is a significant achievement for this model.
4. **False Negative (FN):** Also known as a Type II Error, this occurs when 1 water sample that is actually “Safe” is incorrectly predicted as “Unsafe.” While this error does not pose a direct health risk, it may lead to operational inefficiencies, as clean water may be unnecessarily discarded or reprocessed.

From the confusion matrix values, mathematical calculations were performed to obtain four main evaluation metrics:

1. Accuracy: 95.35%.
   Accuracy measures the overall correctness of the model and is calculated using the formula:
   $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
   A value of 95.35% indicates that more than 95% of all tested samples were correctly classified into their respective categories.
2. Precision: 94.12%.
   Precision measures the reliability of the model’s positive predictions and is calculated as:
   $$\text{Precision} = \frac{TP}{TP + FP}$$
   A score of 94.12% means that when the model predicts a water sample as “Safe,” it is correct 94.12% of the time, thereby minimizing the risk of unsafe water being misclassified as safe.
3. Recall: 94.12%.
   Recall evaluates the model’s ability to identify all actual positive samples and is calculated as:
   $$\text{Recall} = \frac{TP}{TP + FN}$$
   A score of 94.12% indicates that the model is highly effective in capturing truly safe water samples, with very few being missed (false negatives)
4. F1-Score: 94.12%.
   The F1-Score is the harmonic mean of Precision and Recall. A value of 94.12% demonstrates that the model maintains a well-balanced and stable performance between correctly identifying safe water and avoiding misclassification of unsafe water.

K-Fold Cross-Validation Testing Results (Detecting Overfitting)

Although the Train/Test Split evaluation produced a very high accuracy metric (95.35%), in the discipline of Machine Learning, high accuracy from a single test does not necessarily guarantee that the model is reliable in real-world scenarios. There is a potential risk known as Overfitting, which is a condition where the model excessively "memorizes" patterns in the training data, including noise and specific anomalies within it, but loses its ability to generalize (recognize) new data.

International Journal Of Health, Engineering And Technology (IJHET) E-ISSN 2829 - 8683
Volume 5, Number 2, July 2026, Page. 1641 - 1655
Email : editorijhess@gmail.com

To ensure and validate that this classification model is capable of generalizing consistently, further testing was conducted using a cross-validation technique, namely K-Fold Cross-Validation. In this study, the parameter K was set to 5 (K = 5). This means the entire dataset was randomly divided into 5 equally sized folds. The model was then instructed to train in 5 consecutive iterations. In each iteration, 4 folds were used as training data, and the remaining fold was alternately used as validation data to measure the accuracy score.

The results of the five iterations are summarized in the table below:

Table 1. K-Fold Cross Validation Testing Results

Fold	Validation Accuracy Score	Evaluation Status
Fold 1	95.30%	Stable
Fold 2	93.50%	Stable
Fold 3	96.20%	Stable
Fold 4	94.80%	Stable
Fold 5	94.90%	Stable
Mean CV Score	94.94%	Proven Overfitting-Free Model

1. Fold 1: Produced a validation accuracy score of 95.30%, indicating that in the first data split, the model's performance was very close to the manual testing result (95.35%). The evaluation status is considered stable.
2. Fold 2: Produced a score of 93.50%. This slight decrease in the second iteration is common in Cross Validation, usually caused by the presence of more difficult samples to classify in that fold. However, this result is still above 90% and is considered stable.
3. Fold 3: Achieved the highest score of 96.20%. This demonstrates that the Random Forest algorithm was able to identify very clear decision patterns in the subset combination during this iteration.
4. Fold 4: Obtained a score of 94.80%, returning to the general performance balance of the model.
5. Fold 5: Concluded the testing with a score of 94.90%, again showing a stable evaluation status in the final fold.

After completing all five testing iterations, aggregation was performed to determine the central performance value of the model. The average calculation (Mean CV Score) from the 5-Fold Cross Validation resulted in a value of 94.94%.

Comparative analysis of this value is very important. The average accuracy score of 5-Fold (94.94%) shows a very small difference (only about 0.41%) compared to the initial Train/Test Split accuracy (95.35%). The closeness of these two values empirically proves that the Random Forest classification performance is consistent across all parts of the dataset, does not depend on a particular random split, and most importantly: safe and proven to be free from overfitting. This conclusion confirms that the designed system has high reliability and is ready to be implemented for water quality monitoring in real-world scenarios.

User Acceptance Testing (UAT) Results

After the Machine Learning model evaluation stage mathematically (through Train/Test Split and K-Fold Cross-Validation) proved that the Random Forest and PCA algorithms have accurate performance and are free from overfitting, the next testing phase focuses on the software engineering aspect and human interaction. This testing was conducted through User Acceptance Testing (UAT). UAT is a crucial final validation stage before a system is declared ready for real-world implementation. The main objective of UAT is to measure how well the user interface (UI) and user experience (UX) that have been developed can be accepted, understood, and easily operated by end-users, which in this case include the general public and water management officers.

Based on the recapitulation of testing results conducted with respondents, the system achieved an overall UAT acceptance index of 90%. This percentage places the application in the category of "Highly Feasible" or "Highly Accepted." This high achievement demonstrates that the complexity of artificial intelligence algorithms on the backend (dimensionality reduction and ensemble classification processes) has been successfully translated and hidden behind a user-friendly interface on the frontend.

The testing was conducted using a scaled questionnaire method (Likert Scale) distributed to respondents after they performed hands-on testing of the application on both desktop and mobile devices. Based on the tabulated questionnaire data, the evaluation recap is presented in Table 2 below:

Table 2. Recapitulation of User Acceptance Testing (UAT) Results

No	**Testing Aspect**	**Evaluation Indicators**	**Approval Percentage (%)**	**Feasibility Category**
1	System Functionality	System's ability to process data input, automatic Excel-to-CSV conversion, bulk upload feature, and prediction algorithm speed without errors.	92%	Highly Feasible
2	User Interface (UI) Design	Responsiveness of layout on mobile and desktop devices, clarity of typography, and effectiveness of color indicators (green/red) in validation result boxes.	88%	Highly Feasible
3	User Experience (UX)	Level of efficiency and ease of navigation of the application, both for general users (public access) and in reducing operational work time for analysts (operator dashboard).	90%	Highly Feasible
Total	Overall Acceptance Index (Average)	Accumulation of all human-computer interaction testing aspects.	**90%**	**Highly Feasible / Accepted**

Based on Table 2 above, the Drinking Water Feasibility Identification System achieved an overall acceptance index of 90%, placing this software in the category of "Highly Feasible" or fundamentally very well accepted by users. This figure is not merely a quantitative value, but a representation of the successful integration between the artificial intelligence model (backend) and interface engineering (frontend). The following is an in-depth analysis of the three testing pillars:

Analysis of System Functionality Aspect (Score: 92%)

The functionality aspect achieved the highest score, namely 92%. This proves that the computational engine built behind the application operates optimally. Respondents gave high appreciation for the system's ability to handle data manipulation. In particular, the automatic conversion feature from Excel template (.xlsx) to Comma Separated Values (.CSV) format in the background process was considered highly essential. In addition, the Random Forest algorithm proved capable of processing bulk data uploads and executing hundreds of rows of water sample data without experiencing system crashes or excessive loading times. This technical success builds trust among operators that the application is robust for daily laboratory operations.

Analysis of User Interface / UI Aspect (Score: 88%)

In terms of visual and interface aspects, the system obtained an approval score of 88%. This very high score is driven by the implementation of Responsive Web Design (RWD) principles and conditional rendering. Respondents appreciated how the application layout dynamically adapts without breaking element proportions, whether accessed via large desktop screens or smaller mobile screens. Furthermore, the chosen color palette provides a strong psychological impact. The use of bright green indicators with check icons (☑) for compliant water, and red warning indicators with cross icons (✖) for non-compliant water, was considered highly communicative. This design eliminates ambiguity, allowing the general public to immediately understand water feasibility status at a glance.

International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 5, Number 2, July 2026, *Page. 1641 - 1655*
Email : editorijhess@gmail.com

Analysis of User Experience / UX Aspect (Score: 90%)

The user experience aspect measures how intuitive the application flow is in solving user problems. With a score of 90%, the dual access architecture (Role-Based Access) of this system successfully satisfies two different user demographics. On the General Access side, the learning curve is minimized close to zero; users only need to input three parameters (pH, Turbidity, Chlorine) to obtain instant results. On the Operator Access side, the system effectively addresses time inefficiency issues. The Operator Dashboard UX eliminates tedious manual cross-checking processes. Operators simply upload data, and the system immediately presents aggregated conclusions through the Prediction Table and Distribution Bar Chart. This significantly reduces reporting time, increases productivity, and proves that the implemented Machine Learning technology delivers positive impact. Based on the testing conclusion through the achievement of a 90% UAT index, it can be concluded that the Drinking Water Feasibility Identification System using the Random Forest Algorithm and Principal Component Analysis (PCA) Method can support the operational activities of PDAM institutions as well as facilitate independent water quality checking for the general public.

System Visualization Documentation

General Access Interface (Water Quality Checking)

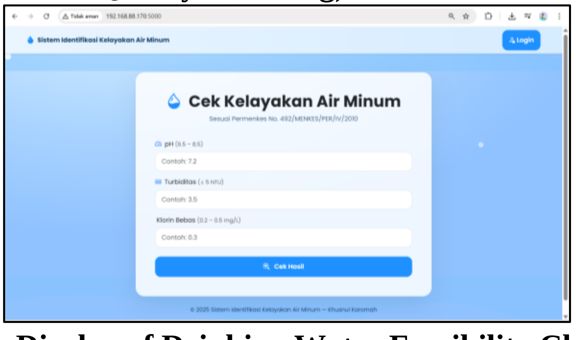

Figure 1. Desktop Version Display of Drinking Water Feasibility Check Application (General Access)

Figure 2. Landscape Mobile Screen Version Display of Drinking Water Feasibility Check Application (General Access)

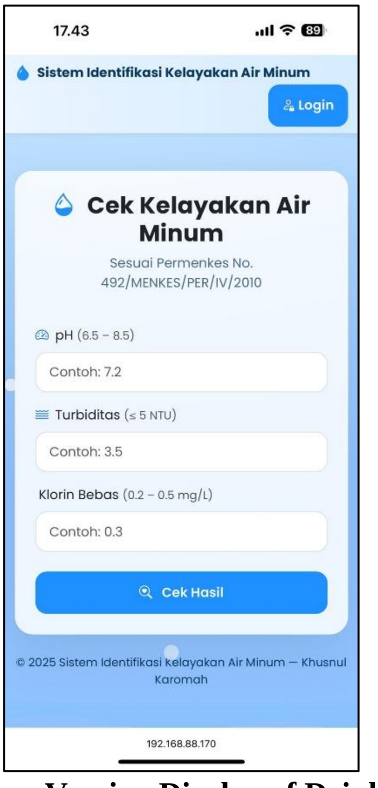

Figure 3. Portrait Mobile Screen Version Display of Drinking Water Feasibility Check Application (General Access)

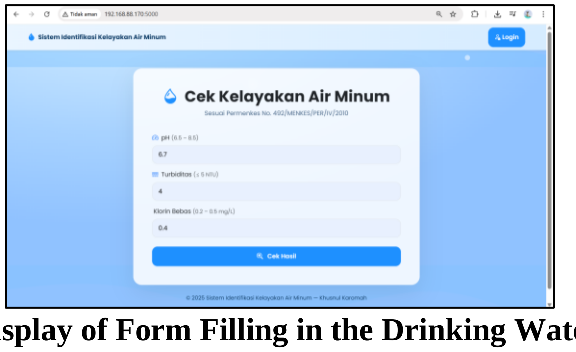

Figure 4. Example Display of Form Filling in the Drinking Water Feasibility Check Application (General Access) Desktop Version

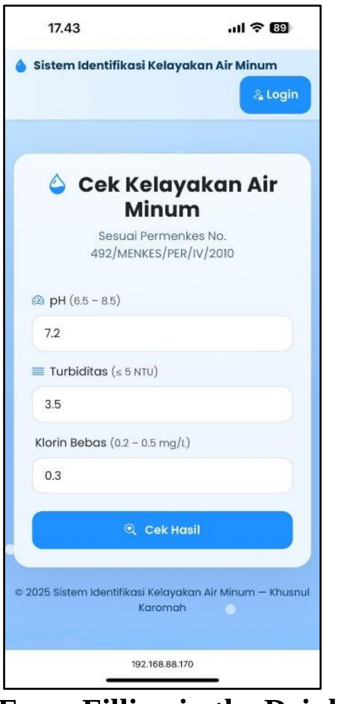

Figure 5. Example Display of Form Filling in the Drinking Water Feasibility Check Application (General Access) Portrait Mobile Screen Version

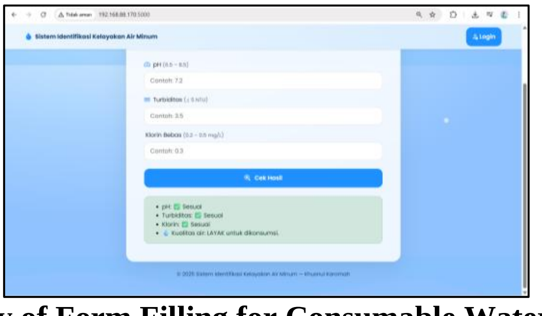

Figure 6. Example Display of Form Filling for Consumable Water Quality in the Drinking Water Feasibility Check Application (General Access) Desktop Version

Figure 7. Example Display of Form Filling in the Drinking Water Feasibility Check Application (General Access) Portrait Mobile Screen Version

Figure 8. Example Display of Form Filling for Non-Consumable Water Quality in the Drinking Water Feasibility Check Application (General Access) Desktop Version

Figure 9. Example Display of Form Filling for Non-Consumable Water Quality in the Drinking Water Feasibility Check Application (General Access) Portrait Mobile Screen Version

Authentication Interface (Officer Login System)

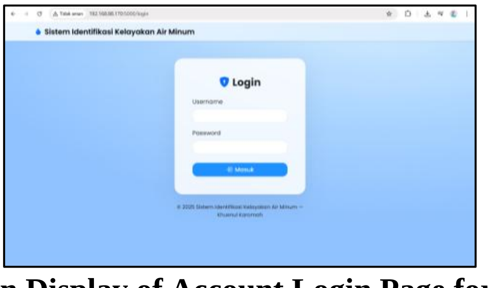

Figure 10. Desktop Version Display of Account Login Page for Drinking Water Quality Monitoring Officers

Figure 11. Mobile Screen Version Display of Account Login Page for Drinking Water Quality Monitoring Officers

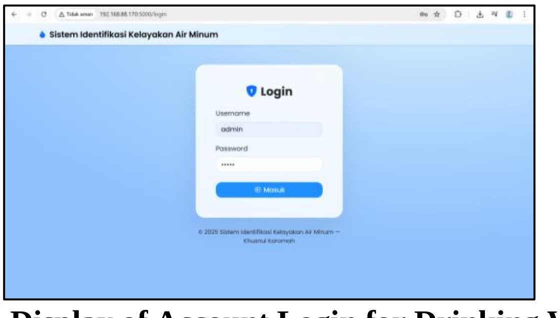

Figure 12. Desktop Version Display of Account Login for Drinking Water Quality Monitoring Officers

Figure 13. Landscape Mobile Screen Version Display of Account Login for Drinking Water Quality Monitoring Officers

International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 5, Number 2, July 2026, *Page. 1641 - 1655*
Email : editorijhess@gmail.com

Figure 14. Mobile Screen Version Display of Account Login for Drinking Water Quality Monitoring Officers

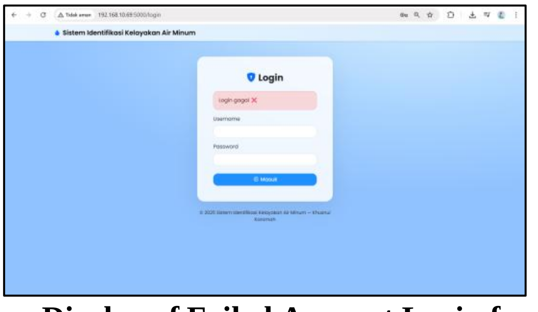

Figure 15. Desktop Version Display of Failed Account Login for Drinking Water Quality Monitoring Officers

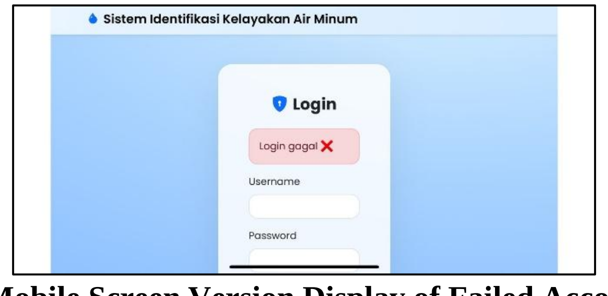

Figure 16. Landscape Mobile Screen Version Display of Failed Account Login for Drinking Water Quality Monitoring Officers

Figure 17. Mobile Screen Version Display of Failed Account Login for Drinking Water Quality Monitoring Officers

Confusion Matrix Evaluation (Train/Test Split)

This sub-chapter presents the documentation of User Interface (UI) and User Experience (UX) design within the Officer Dashboard environment. This interface is the core of the application's operations, specifically designed to accommodate high-level computational needs, bulk data processing, and to present analytical visualizations that facilitate decision-making processes for water quality management institutions.

Figure 18. Desktop Version Display of Drinking Water Feasibility Identification Application for Officers

Figure 19. Landscape Mobile Screen Version Display of Drinking Water Feasibility Identification Application for Officers

Figure 20. Mobile Screen Version Display of Drinking Water Feasibility Identification Application for Officers

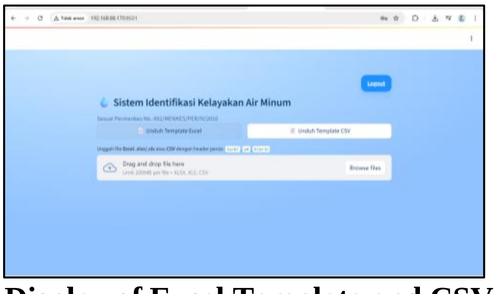

Figure 21. Desktop Version Display of Excel Template and CSV Template Download in the Application

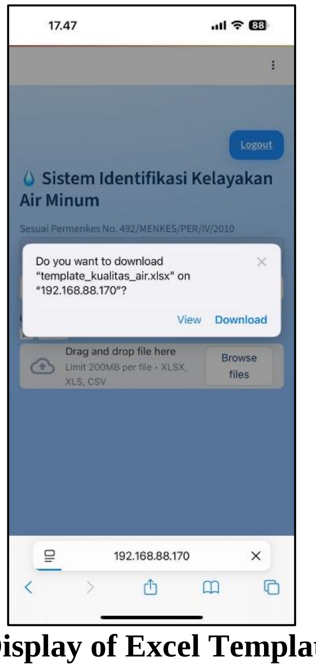

Figure 22. Mobile Screen Version Display of Excel Template and CSV Template Download in the Application

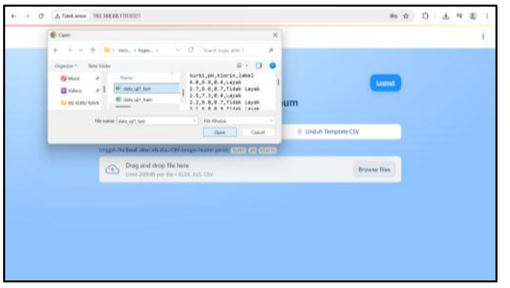

Figure 23. Desktop Version Display of Excel and CSV Dataset Upload in the Application

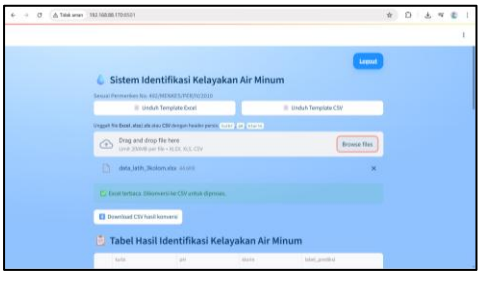

Figure 24. Desktop Version Display of the Application During Drinking Water Feasibility Identification from Uploaded Dataset

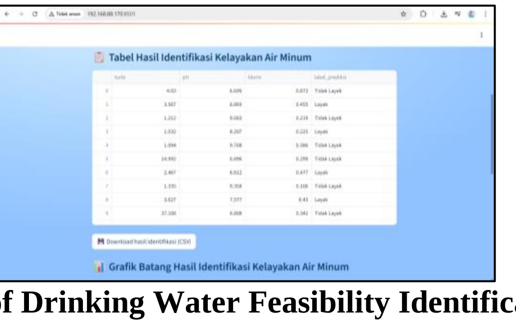

Figure 25. Display of Drinking Water Feasibility Identification Result Table

CONCLUSIONS

1. The drinking water feasibility classification system was successfully developed by implementing PCA specifically to reduce the dimensionality of features (pH, turbidity, and chlorine), while the Random Forest algorithm works effectively as the main classification model.
2. The performance of the Random Forest algorithm is highly satisfactory, achieving an accuracy of 95.35%. Model validation testing using K-Fold Cross-Validation (K=5) produced an average score of 94.94%, proving that the model is stable and free from overfitting conditions.
3. The program interface (UI) and user experience (UX) have proven to be highly flexible as they are designed with two layers of authentication. The public can perform quick one-way checks using color-based visual indicators (green/red), while the Officer Dashboard simplifies bulk reporting through automatic Excel/CSV format conversion. User Acceptance Testing (UAT) shows that the user acceptance level reached 90%.

REFERENCES

A. Abdullah, M. R. Pratama, and S. Hidayat, "Penerapan Machine Learning Untuk Deteksi Dini Kualitas Air Permukaan di Perkotaan," *Jurnal Teknologi dan Sistem Cerdas*, vol. 4, no. 2, pp. 102-115, 2023.

A. Prihandoko, *Memahami Konsep dan Implementasi Machine Learning*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.

A. Izzuddin and M. Rizal Wahyudi, "Pengenalan Pola Daun Untuk Membedakan Tanaman Padi dan Gulma Menggunakan Metode Principal Components Analysis (PCA) dan Extreme Learning Machine." *Jurnal Teknik Elektro ITN*, 2020.

A. S. Ritonga and I. Muhandhis, "Teknik Data Mining Untuk Mengklasifikasikan Data Ulasan Destinasi Wisata Menggunakan Reduksi Data Principal Component Analysis (PCA)." *Jurnal Data Mining dan Sistem Informasi*, vol. 3, no. 1, 2021.

A. N. Kusnawi, *Belajar Mudah Dan Singkat Machine Learning*. Yogyakarta: CV Andi Offset, 2024.

A. Wijaya and B. Santoso, "Pengujian User Acceptance Testing (UAT) Terhadap Aplikasi Pemantauan Lingkungan Hidup Berbasis Mobile," *Jurnal Sistem Informasi Bisnis*, vol. 13, no. 2, pp. 134-142, 2023.

B. Pusat Statistik, "Statistik Air Bersih (Water Supply Statistics) Indonesia Tahun 2023," *Badan Pusat Statistik Jakarta*, 2023.

B. Farnham, S. Tokyo, and F. Sebastopol, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Beijing: O'Reilly Media, 2021.

B. C. Octariadi, D. W. Santoso, and M. I. Hakim, "Evaluasi User Experience pada Sistem Pakar Layanan Publik Berbasis Web Menggunakan System Usability Scale," *Jurnal Rekayasa Perangkat Lunak dan Informatika*, vol. 5, no. 1, pp. 45-52, 2022.

D. Novaliendry, *Deep Learning Untuk Pemula Jilid 1*. Grobogan: CV. Sarnu Untung, 2023.

D. T. E. Kurniawan, *User Interface dan User Experience: Prinsip dan Desain*. Surabaya: PT. Revka Petra Media, 2021.

F. A. Budianto, "Komparasi Algoritma Klasifikasi Machine Learning pada Data Numerik Lingkungan," *Jurnal Ilmiah Komputasi*, vol. 20, no. 3, pp. 201-210, 2021.

H. Yuliansyah, *Pendeteksian Overfitting pada Machine Learning*. Bandung: Informatika, 2022.

Jubilee Enterprise, *Python Untuk Programmer Pemula*. Jakarta: PT. Elex Media Komputindo, 2019.

Kementerian Kesehatan Republik Indonesia, "Peraturan Menteri Kesehatan Republik Indonesia Nomor 492/MENKES/PER/IV/2010 Tentang Persyaratan Kualitas Air Minum," *Kemenkes RI*, 2010.

M. A. Wibowo, "Panduan Praktis User Acceptance Test (UAT) dalam Siklus Pengembangan Perangkat Lunak," *Jurnal Informatika Terapan*, vol. 6, no. 2, 2020.

M. Septiani and Z. Abidin, "Pengenalan Pola Batik Lampung Menggunakan Metode Principal Component Analysis," *Jurnal Informatika dan Rekayasa Perangkat Lunak (JATIKA)*, vol. 2, no. 4, pp. 552–558, 2021.

R. Kurniawan, *Statistika Multivariat: Analisis Komponen Utama (PCA)*. Jakarta: Kencana, 2021.

R. Pratama, T. Wahyono, and A. Rahman, "Klasifikasi Kualitas Air Menggunakan Random Forest dan Validasi K-Fold," *Indonesian Journal of Computer Science*, vol. 12, no. 1, pp. 110-120, 2024.

S. A. Saputra and D. S. Maulana, "Optimalisasi Evaluasi Model Klasifikasi Menggunakan K-Fold Cross Validation," *Jurnal Ilmu Komputer dan Teknologi Informasi*, vol. 8, no. 1, 2023.

Z. Jayidan, A. Mutoi Siregar, S. Faisal, and H. Hikmayanti, "Improving Heart Disease Prediction Accuracy Using Principal Component Analysis (PCA) In Machine Learning Algorithms," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 3, pp. 821–830, 2024.