Optimization of the SRIKANDI E-Government System Using XGBoost-Based Classification and One-Class SVM Anomaly DetectionType

Fitri Damaryanti[a,1], Aji Supriyanto[a,2,*]

[a] Magister of Information Technology, Universitas Stikubank , Jawa Tengah, Semarang, 50243, Indonesia
[1] rr.fitri0038@mhs.unisbank.ac.id, [2]ajisup@edu.unisbank.ac.id [*]
* corresponding author

ARTICLE INFO

ABSTRACT

Keywords
SRIKANDI
Classification
Archive Anomalies
XGBoost
One-Class SVM

Accurate and efficient digital archive management is a crucial component of Electronic-Based Government Systems (SPBE) in Indonesia. The Integrated Dynamic Archival Information System (SRIKANDI), widely used by government agencies, continues to face various challenges such as incomplete metadata, inconsistent classification, and difficulties in archive retrieval and retention scheduling. This study aims to optimize the SRIKANDI system by implementing machine learning algorithms XGBoost for document classification and One-Class SVM (OCSVM) for automatic anomaly detection in metadata. The methodology involves data preprocessing, feature selection, label generation, and the application of classification and anomaly detection models on archival data from the Meteorological, Climatological, and Geophysical Agency (BMKG), Central Java. The XGBoost model achieved a classification accuracy of 77%, showing strong performance in identifying "Destructible" archives but limited ability in detecting the "Permanent" category due to data imbalance. Meanwhile, the OCSVM model successfully identified 16 anomalous entries (9.14%) out of 175 archives, with key indicators including extreme item counts and illogical retention periods. The results demonstrate that integrating machine learning into digital archival systems significantly improves classification accuracy, operational efficiency, and metadata integrity. Furthermore, this approach supports proactive auditing and validation of archival metadata. The findings offer valuable insights for developing AI-powered archival classification and anomaly detection systems to enhance accountability, transparency, and data governance in the public sector.

1. Introduction

Records management plays a crucial role in modern governance to enhance transparency, accountability, and administrative efficiency. Archives serve not only as sources of historical information but also as critical tools for present-day decision-making and forecasting future possibilities. Such information possesses evidentiary and utilitarian value and functions as an organizational memory in administrative processes[1]. The ISO 15489 standard and Indonesian governmental regulations on archival management emphasize the importance of accurate records storage. However, practical implementation often faces challenges such as improper file location and neglected retention schedules. Effective records management is therefore essential to uphold public sector transparency and accountability [2].

Poor records management practices can lead to service inefficiencies, legal risks due to non-compliance, organizational memory loss, and impaired decision-making processes[3]. Numerous government institutions encounter challenges such as limited facilities and infrastructure for archival administration, resulting in inefficiencies[4]. Moreover, inadequate technological infrastructure inhibits the full potential of digital transformation efforts[5]. To address this, the Indonesian government has implemented an archival system known as *Sistem Informasi Kearsipan Dinamis Terintegrasi* (SRIKANDI), a standardized archival information system used across all government agencies—both central and regional—for managing dynamic electronic records[6]. This system also serves as a strategic component in advancing the maturity of Indonesia's Electronic-Based Government System (SPBE or e-Government), as mandated by Presidential Regulation No. 95 of 2018[7][8][9].

At the Meteorological, Climatological, and Geophysical Agency (BMKG) – Climatology Station Class I in Central Java, several issues have been identified with SRIKANDI, such as complex metadata structures, inconsistent classification labels, and file duplications. Similar problems were reported by Bhara Nurpasma Miawani (2024), who analyzed SRIKANDI implementation at the Ministry of Agriculture via the University of Indonesia digital library [https://lib.ui.ac.id/]. Out of 16 metadata requirements for record creation and classification, three critical elements were unmet—such as the absence of mandatory metadata fields, incomplete classification codes or creation dates, and reduced efficiency in search and automated classification functions. Dewi Yulianti (2024) also found inconsistencies between metadata creation dates and electronic signature timestamps, as well as the absence of automated metadata maintenance mechanisms, resulting in decreased archival integrity. Consequently, misdated records were not properly disposed of, and data discrepancies impaired system interoperability and search reliability.

Amid these issues, Artificial Intelligence (AI) technologies have emerged as vital solutions for the future[10]. Machine Learning (ML), a subset of AI, has shown great promise in automating archival document classification into standardized codes and document types[11][12][13]. ML models are also capable of detecting metadata anomalies—such as illogical dates, blank fields, duplication, and mismatches between document content and metadata—thereby improving efficiency, accuracy, and system integrity. The XGBoost algorithm, in particular, is highly effective in automating document classification using text and metadata features[14]. Some studies have proposed hybrid systems combining XGBoost for classification and One-Class Support Vector Machine (OCSVM) for anomaly detection, utilizing novel feature engineering techniques such as mapping classification codes to physical archive locations[15].

ML technologies also enable anomaly detection in e-Government datasets, highlighting the need for a balance between accuracy and data ethics. Models such as Isolation Forest, OCSVM, and deep learning architectures have been employed for this purpose[16]. ML approaches including SVM and ensemble models are capable of classifying archival documents based on content and metadata[17]. Open-set classification-based anomaly detection methods have been suggested to detect rare or unexpected document categories in systems like SRIKANDI[18]. XGBoost remains a preferred choice for classification due to its robustness in handling large, complex datasets with high performance[19]. Compared to SVM, XGBoost has proven more stable and reliable when applied to real-world, imperfect data, as it incorporates regularization to control model complexity and prevent overfitting[14][20].

Chen et al. (2024) utilized XGBoost for feature selection in safety management systems, demonstrating its utility in identifying system anomalies[11]. He & Chen (2025) further extended XGBoost in semi-supervised anomaly detection by integrating multiple anomaly scores[21]. Shilton et al. (2020) proposed an innovative SVM model that integrates multi-class classification and anomaly detection (OCSVM) into a unified framework[22], which aligns well with the needs of SRIKANDI to simultaneously classify document types and detect anomalous metadata entries. Studies comparing XGBoost and SVM show XGBoost achieving classification accuracy as high as 99%, outperforming SVM's 93.8% in network traffic anomaly detection tasks[23]. In the context of Industrial IoT anomaly detection, boosting models achieved the best AUC (0.992), surpassing

SVM [12]. In other studies on network and IoT device anomaly detection, XGBoost also achieved the highest accuracy (99.98%) and computational efficiency[13].

Based on the aforementioned background and literature, this study aims to identify types of metadata anomalies that affect search functionality, record retention, and system interoperability. Furthermore, the study applies and evaluates ML algorithms—specifically XGBoost and OCSVM—to perform document classification and metadata anomaly detection within the e-Gov SRIKANDI system at BMKG Climatology Station Class I, Central Java. The primary contribution of this research is the integration of XGBoost and OCSVM to enhance archival classification accuracy and automatically detect metadata anomalies, supporting improved efficiency and accountability in digital records governance under the e-Gov SRIKANDI initiative.

2. Method

Based on the aforementioned background, this study adopts a quantitative exploratory analysis approach, utilizing Artificial Intelligence (AI) methods through Machine Learning (ML) models. Specifically, the research implements a hybrid technique by combining the Extreme Gradient Boosting (XGBoost) algorithm for classification and the One-Class Support Vector Machine (OCSVM) for anomaly detection. The overall methodological framework of the study is illustrated in Figure 1.

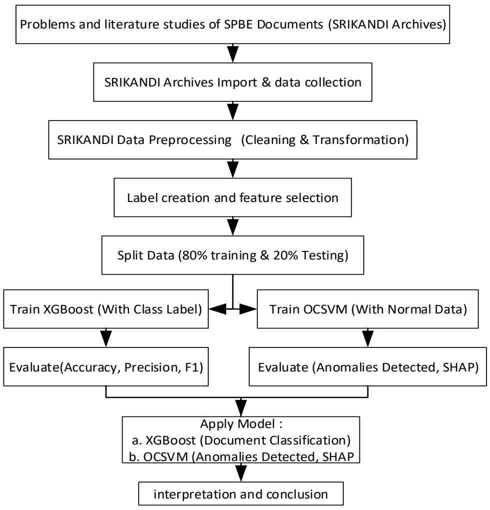

Problems and literature studies of SPBE Documents (SRIKANDI Archives)
SRIKANDI Archives Import & data collection
SRIKANDI Data Preprocessing (Cleaning & Transformation)
Label creation and feature selection
Split Data (80% training & 20% Testing)
Train XGBoost (With Class Label)
Train OCSVM (With Normal Data)
Evaluate(Accuracy, Precision, F1)
Evaluate (Anomalies Detected, SHAP)
Apply Model :
a. XGBoost (Document Classification)
b. OCSVM (Anomalies Detected, SHAP
interpretation and conclusion

Figure 1. Research Framework for Document Classification and Anomaly Detection in SRIKANDI Archives

2.1. Problem Statement and Literature Review

The archival documents stored in SRIKANDI, typically exported as Excel spreadsheets, exhibit several challenges: complex and inconsistent metadata structures, incomplete and ambiguous classification labels, and frequent data duplication. These issues hinder effective document categorization, impede efficient document retrieval and tracking, and complicate the detection of irregularities or anomalies (i.e., outlier detection).

To address these problems, this study employs the XGBoost algorithm to classify archival documents and the OCSVM algorithm to detect anomalies in the SRIKANDI system. In implementing the XGBoost classification model, the prediction formula is constructed by iteratively aggregating multiple decision trees, where each tree incrementally improves upon the errors of the previous one. This ensemble-based method enhances the accuracy and robustness of the classification process.

$$SVM\ Prediction = y\hat{}_i = \sum_{k=1}^{K} fk(x_i), fk\ \in F \qquad (2.1)$$

Next, the objective function of the XGBoost model is constructed, which combines the loss function and the regularization term, and is formulated as follows:

$$\mathrm{L}(\phi) = \sum_{i=1}^{n} \mathrm{l}(\mathrm{yi}, \mathrm{y}\hat{}_{i}) + \sum_{k=1}^{K} \Omega(f_k) \quad (2.2)$$

$$\Omega(\mathrm{f}) = \gamma\mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (2.3)$$

For binary classification tasks, XGBoost utilizes the logistic loss function, which is defined as follows:

$$\mathrm{l}(\mathrm{yi}, \mathrm{y\hat{}i}) = -[\mathrm{yi}\,log(\sigma(\mathrm{y\hat{}i})) + (1 - \mathrm{yi})log(1 - \sigma(\mathrm{y\hat{}i}))] \quad (2.4)$$

Where :

$$\sigma(\mathrm{y\hat{}i}) = \frac{1}{1 + e^{-y\hat{}i}}$$

Unlike classical Support Vector Machines (SVM), which aim to separate two distinct classes, the One-Class Support Vector Machine (OCSVM) focuses solely on learning from a single class (normal data) to identify deviations or outliers. It is specifically designed for unsupervised anomaly detection or novelty detection tasks[24]. The OCSVM model is particularly suitable for detecting abnormal or "novel" data points by learning the distribution of normal data—such as archival metadata entries—without requiring labeled anomaly data[15].

The objective function of the OCSVM is defined as follows:

$$= \min_{w,\xi\mathrm{i},\rho} \frac{1}{2} ||w||^2 + \frac{1}{vn} \sum_{i=1}^{n} \xi\mathrm{i} - \rho \quad (2.5)$$

After training, the decision function used to determine whether a data point is normal or anomalous in One-Class SVM is defined as follows:

$$\mathrm{f(x)} = \mathrm{sign}((\mathrm{w} \cdot \phi(\mathrm{x})) - \rho) \quad (2.6)$$

Since OCSVM is often used with kernel functions such as the Radial Basis Function (RBF) or Gaussian kernel its decision function is commonly expressed as follows:

$$\mathrm{f(x)} = \sum_{i=1}^{n} \alpha\mathrm{iK}(\mathrm{xi}, \mathrm{x}) - \rho \quad (2.7)$$

Here, αi\alpha_iαi represents the Lagrange multipliers (or coefficients) obtained through the optimization process during model training. These coefficients determine the influence of each support vector xix_ixi in the decision boundary formulation.

$$\mathrm{K}(\mathrm{xi}, \mathrm{x}) = \exp(-\gamma \parallel \mathrm{xi} - \mathrm{x} \parallel 2) \quad (2.8)$$

2.2. Data Collection

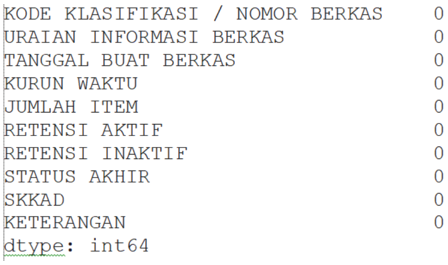
The dataset used in this study was obtained from the SRIKANDI archival system of the BMKG Climatology Station Class I Central Java, covering the period from August 2023 to May 2025. Data was extracted by importing Excel-format files from the SRIKANDI application interface. The imported data consists of two primary raw data files: (1) a file containing active document records and (2) a file containing classification reference structures. The imported data was structured into DataFrames, a two-dimensional tabular data structure used for digital storage, processing, and analysis of SRIKANDI archival records. The main archival DataFrame consists of 175 rows and 10 columns, each row representing an individual archival record. The classification reference DataFrame contains 385 rows and 5 columns, which define the hierarchical structure of field codes and sub-classifications. This classification reference is used as the foundation for labeling and validating document classification codes.

Each archival entry includes the following key attributes: *Classification Code*, *Information Description*, *Creation Date*, *Time Span*, *Item Count*, *Active Retention*, *Inactive Retention*, *Final Status*, *SKKAD*, and *Remarks*. The classification reference provides structured fields such as *Field Code*, *Main Code*, *Sub Code*, and *Subclassification Description*. A merged DataFrame was created by combining both data sources, resulting in a final dataset with 15 main attributes, which serves as the input for classification and anomaly detection tasks.

2.3. SRIKANDI Data Preprocessing

This phase involves data cleaning and transformation to prepare the dataset for machine learning analysis. First, the dataset was checked to ensure the absence of missing values in all critical columns. A zero count of missing entries confirms data completeness. Additional preprocessing steps included the removal of invalid characters and the standardization of date formats, which are essential to avoid errors during model training.

Next, the cleaned dataset underwent feature transformation, including categorical feature encoding, creation of new derived features such as date difference calculations, and validation of archival status fields. As a result, a final preprocessed DataFrame with 175 rows and 15 features was produced.



```
KODE KLASIFIKASI / NOMOR BERKAS    0
URAIAN INFORMASI BERKAS            0
TANGGAL BUAT BERKAS                0
KURUN WAKTU                        0
JUMLAH ITEM                        0
RETENSI AKTIF                      0
RETENSI INAKTIF                    0
STATUS AKHIR                       0
SKKAD                              0
KETERANGAN                         0
dtype: int64
```

Figure 2. SRIKANDI Archive Data Cleaning Process

As illustrated in Figure 2, the dataset contains no missing values, exhibits uniform data formatting, and has been successfully merged with the official classification reference. Additionally, 36 mismatched records were identified, these entries do not align with the official classification structure and thus serve as potential focal points for anomaly detection analysis in subsequent stages.

2.4. Feature Selection and Label Construction

Following the data cleaning stage, feature selection was conducted to support the construction of target labels for archival document classification using AI-based Machine Learning (ML) models. In document classification tasks, a label represents the type or classification category of the archival record.

In the context of SRIKANDI archival data processing, label creation is based on the integration of document attributes with the official classification reference. The selected features serve as inputs for training the classification model, while the label corresponds to the validated classification code assigned to each document. The structure of the input file (i.e., the feature set from SRIKANDI) consists of several columns as listed in Table 1.

Table 1. Column List of the SRIKANDI Data File (Feature Set)

Column	Description
KODE KLASIFIKASI / NOMOR BERKAS	Official classification code of the archive
URAIAN INFORMASI BERKAS	Description or title of the document content
TANGGAL BUAT BERKAS	Date the document was created
KURUN WAKTU	Validity period of the document
JUMLAH ITEM	Quantity of items or volumes in the archival entry
RETENSI AKTIF, RETENSI INAKTIF	Active retention period in years, Inactive retention period in years
STATUS AKHIR	Archival status (disposed, permanent)
SKKAD, KETERANGAN	Document reference related to retention schedule decisions
Kode Bidang, Kode Utama, Kode sub	Additional notes or comments, Extracted from classification structure, Extracted from classification structure
Sub Klasifikasi, Deskripsi	Name of classification based on reference, Indicator of whether the classification is valid or not

Based on the columns listed in Table 1, and the results of merging the active archive records with the official archival classification reference table, a label construction process was performed. The objective of this process is to assign a valid classification label to each archival record, which can subsequently be used for training a supervised machine learning classification model. The output of this merging and labeling process produces a structured dataset where each record is associated with a clearly defined classification code, enabling automated model learning. An example of the labeled data structure is presented in Table 2.

Table 2. Sample Labeled Archival Classification Data

Classification Code	Subclassification	Label
DB.00.00	Database Administration	Database Administration
DB.00.00	Database Management	Database Management
DB.00.01	Database System Operations	Database System Operations

The construction of the "label" column, as shown in Table 2, is the final result of merging the document classification code (e.g., DB.00.00) with its corresponding subclassification description. This combined field is then used as the target variable for the automated document classification task using XGBoost or Support Vector Machine (SVM) algorithms. The labeled dataset, once finalized, was exported and stored in an external file named Hasil_Label_Klasifikasi_Arsip.xlsx, which serves as the input for model training and evaluation in the classification pipeline.

2.5. Document Classification Modeling Using XGBoost

Following the labeling process, a document classification model was developed using the XGBoost algorithm. The first step involved splitting the labeled dataset into training (80%) and testing (20%) subsets to support supervised learning. Prior to training, an initial inspection of the dataset was conducted to identify the unique labels and determine whether any minority labels (i.e., low-frequency classes) should be filtered to ensure training stability. After splitting, the dataset was analyzed to determine the total number of training and test instances along with the feature dimensionality. The XGBoost model was then trained on the selected features to classify archival documents based on their metadata and classification attributes. The performance of the classification model was evaluated using standard classification metrics, including precision, recall, F1-score, and support.

2.6. Anomaly Detection Modeling Using OCSVM

An anomaly detection model was developed using the OCSVM algorithm to identify outliers within the archival metadata. Unlike supervised models, OCSVM does not require labeled anomalies. Instead, it learns the structure of normal data and flags any deviations from this learned distribution as potential anomalies. The modeling steps included: Reading the entire dataset to extract all relevant input features, Fitting the OCSVM model to the normal class, Performing anomaly detection and estimating pseudo-accuracy.

This process aims to identify records that may exhibit inconsistent metadata, such as logical errors, formatting issues, or structural mismatches. Such detection can be particularly useful for validating archival metadata, improving data integrity, and preventing misclassification or improper archival disposal. The performance of the anomaly model was reported based on the distribution of normal vs. anomalous data points, as well as a pseudo-accuracy metric derived from internal model evaluation.

2.7. Model Application, Evaluation, and Interpretation

This stage describes the results and interpretation of both the classification model (XGBoost) and the anomaly detection model (OCSVM). For classification, the results were analyzed based on the predicted archival status labels (e.g., active categories). For anomaly detection, the output was categorized into normal and anomalous groups. Model evaluations were conducted using: Confusion Matrix and SHAP (SHapley Additive exPlanations) summary plots for classification, Principal Component Analysis (PCA) visualizations for anomaly detection, Correlation analysis between retention features (active vs. inactive), and Boxplots for numeric feature distribution across anomaly categories.

Model interpretability was enhanced through: Feature Importance Graphs from XGBoost, Confusion Matrix Visualizations, SHAP summary plots to explain feature contributions, Anomaly distribution plots, PCA plots showing feature relationships and clustering patterns, Boxplots of item counts by anomaly class, Distribution plots for retention duration and major classification codes. All modeling and analysis procedures were implemented using the Python programming language on the Google Colab platform. The following libraries were used: Pandas, numpy for data manipulation, Xgboost for classification modeling, Scikit-learn (sklearn) for OCSVM modeling and evaluation, SHAP for model interpretability, and Matplotlib and seaborn for data visualization.

3. RESULTS AND DISCUSSION

3.1. Archival Document Classification Results

In performing archival document classification, the process involves three primary stages: label construction, XGBoost model training, and classification result analysis. In the first stage, classification labels were constructed based on the subclassification field of each archival record. Initially, 83 unique labels were identified. However, to maintain proportionality and avoid data sparsity during model training, only the 7 dominant labels (those with ≥ 5 entries) were selected for modeling. The resulting filtered dataset comprised 71 entries, which were then split into training (80%) and testing (20%) subsets. In the second stage, the XGBoost algorithm was used to train the classification model. The model was built to predict the final archival status, specifically distinguishing between "*Musnah*" (Destroyed) and "*Permanen*" (Permanent) categories.

After splitting the data and training the model, a feature importance analysis was performed. This step generates scores indicating the contribution of each feature to the classification outcome, as computed by the XGBoost model. Feature importance analysis serves multiple purposes: Assessing the relative contribution of features, Supporting feature selection, Enhancing model interpretability, Providing a basis for model refinement.

The feature importance plot for the archival classification model is presented in Figure 3.

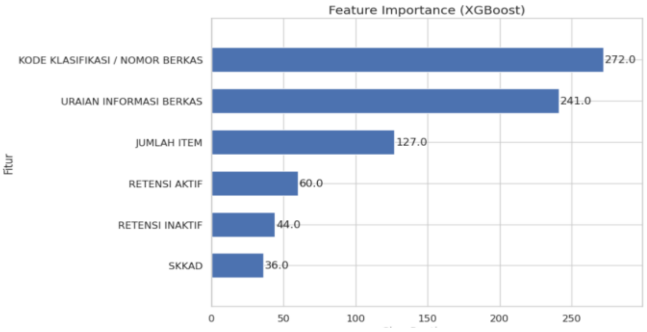


Figure 3. Feature Importance Plot (XGBoost) for Archival Document Classification

As shown in Figure 3, the feature "Classification Code / File Number" emerged as the most influential variable in the XGBoost model, with an importance score of 272, followed by "Document Information Summary" with a score of 241. These two features reflect the content semantics and administrative grouping of archival documents, which substantially impact the model's classification decisions, particularly in distinguishing between permanent archives and those eligible for disposal. The third most important feature was "Number of Items", with a score of 127, indicating that the volume or quantity of documents also serves as a significant indicator in the classification process. Other features include: "Active Retention" (score: 60), "Inactive Retention" (score: 44), and "SKKAD" (Official Decree on Archival Classification) with a score of 36.

Although these features contributed lower importance scores, they still played a relevant role in the model's predictions. Retention periods are closely related to the document's lifecycle—covering active and passive phases, while SKKAD serves as a legal or administrative guideline influencing the final archival status.

These results demonstrate that the XGBoost model effectively captures the underlying classification patterns by emphasizing the classification structure and informational description as core determinants. This insight is not only valuable for interpreting model behavior, but also forms the foundation for developing automated recommendation systems and validation mechanisms for digital archival classification platforms such as SRIKANDI. Based on the classification model evaluation, the performance metrics are summarized in Table 3 below.

Table 3. Performance Metrics of the Classification Model Using XGBoost Algorithm

Class Label	Precision	Recall	F1-Score	Data Summary (Support)
"*Musnah*"	0.84	0.90	0.87	29
"*Permanen*"	0.25	0.17	0.20	6
Accuracy	-	-	0.77	35
Macro Avg	0.54	0.53	0.53	35
Weighted Avg	0.74	0.77	0.75	35

Analysis of Table 3 Results :

1. Class-wise Evaluation
   a. "*Musnah*" (Destroyed) Class:
      - Precision (0.84): Among all instances predicted as "*Musnah*", 84% were correctly classified.
      - Recall (0.90): Of all actual "*Musnah*" archives, 90% were successfully identified by the model.
      - F1-Score (0.87): Represents the harmonic mean of precision and recall, indicating a well-balanced and high-performing classification.
      - Support (29): The number of actual samples labeled "*Musnah*" was 29, making it the dominant class in the dataset.

      These results indicate that the model performs very well in identifying archival documents that should be classified as destroyed.

   b. "*Permanen*" (Permanent) Class:
      - Precision (0.25): Only 25% of the predictions for the "*Permanen*" class were correct.
      - Recall (0.17): Of the actual "*Permanen*" documents, only 17% were correctly predicted.
      - F1-Score (0.20): The low precision and recall result in a low F1-score.
      - Support (6): The number of actual records labeled "*Permanen*" was only 6, indicating a severe class imbalance**.**

      These results show that the model is less effective in identifying documents belonging to the "*Permanen*" class**,** most likely due to the small number of training samples for that category.

2. Overall Model Evaluation
   - Accuracy (0.77): The model correctly classified 77% of the total data.
   - Macro Average (F1 = 0.53): Represents the unweighted mean of F1-scores across all classes, regardless of class distribution. Reflects suboptimal performance due to the poor classification of the "*Permanen*" class.
   - Weighted Average (F1 = 0.75): Represents the average F1-score weighted by the number of samples in each class. Provides a more realistic view of model performance, as it is heavily influenced by the dominant "*Musnah*" class.

Based on the results shown in Table 3**,** the XGBoost classification model demonstrates strong predictive performance for the "*Musnah*" class**,** but significantly underperforms on the "*Permanen*" class**.** This discrepancy is likely caused by data imbalance**,** where the minority class (*Permanen*) lacks sufficient examples for the model to learn meaningful patterns. The classification outcomes are further visualized in the confusion matrix**,** as shown in Figure 4**.**

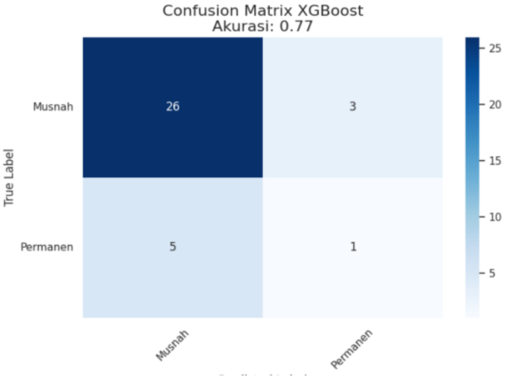


Figure 4. Confusion Matrix of Archival Classification Model Predictions

Based on the results of the confusion matrix**,** the XGBoost model successfully classified 26 archival records from the "*Musnah*" (Destroyed) class correctly as true positives**.** However, it misclassified 3 instances of the "*Musnah*" class as "*Permanen*" (Permanent)**,** which are considered false negatives**.** Conversely, out of 6 archival records that actually belonged to the "*Permanen*" class, only 1 record was correctly predicted **(**true positive**)**, while the remaining 5 records were incorrectly classified as "*Musnah*" (false positives). The model achieved an overall accuracy of 77%, which indicates a reasonably good performance given the imbalanced class distribution. Nonetheless, the performance for the minority class ("*Permanen*") remains suboptimal, as reflected by the low recall score for that class. This suggests that the model exhibits a bias toward the majority class ("*Musnah*"), a common issue in classification tasks involving imbalanced label distributions.

To further interpret the model's decision-making process, SHapley Additive exPlanations (SHAP) analysis was conducted. The SHAP summary plot helps identify which features most strongly influence the model’s predictions. The most impactful features included: Number of Items, Active Retention, Subclassification, and Time Span. The SHAP beeswarm plot visualization reveals that extremely high or low values of the "Number of Items" feature had a strong correlation with the predicted archival status. These findings suggest that document volume and retention-related metadata play a key role in the model’s classification decisions. The detailed SHAP summary plot illustrating feature influence is presented in Figure 5.

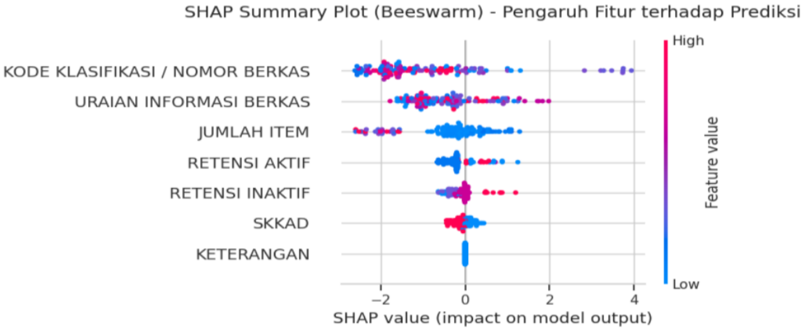


Figure 5. SHAP Summary Plot of Influential Features in Archival Classification

3.2. Results of Metadata Anomaly Detection in Archival Records

The metadata anomaly detection process involved three main stages: developing the anomaly detection model, identifying anomalous records, and performing statistical comparisons. The detection model was built using the OCSVM algorithm, trained on the merged archival dataset comprising 175 entries. The objective of the anomaly detection was to identify metadata entries that deviate from typical patterns, such as: Extremely high or low values in Number of Items, Illogical values in Active or Inactive Retention Periods, Creation Dates that are inconsistent with the stated Time Span. The OCSVM model detected 159 records (90.86%) as normal, and 16 records (9.14%) as anomalies. A statistical summary comparing normal and anomalous data is presented in Table 4:

Table 4. Summary Statistics: Normal vs. Anomalous Records

Category	Number of Items (Mean)	Active Retention	Inactive Retention
Normal	16.09	2.13	2.52
Anomaly	72.56	3.75	2.44

Interpretation of Table 4:

a. Number of Items: Anomalous records had an average of 72.56 items, significantly higher than normal records (16.09), indicating abnormal volume or document aggregation.
b. Active Retention: Anomalies showed longer retention periods (3.75 years vs. 2.13), suggesting potential misclassification, data entry errors, or policy inconsistencies.
c. Inactive Retention: Slightly lower in anomalies (2.44 vs. 2.52); while the difference is minor, it still indicates management inconsistencies.

These findings suggest that anomalous records statistically differ, especially in item count and active retention. Anomalies were 4.5 times more voluminous than normal entries, possibly due to: Misclassification (inconsistent classification codes), Incorrect metadata inputs, Structural deviations from institutional standards. These anomalies may serve as inputs for metadata audits, manual validation by archivists, or as a basis for enhanced business rules in digital systems like SRIKANDI.

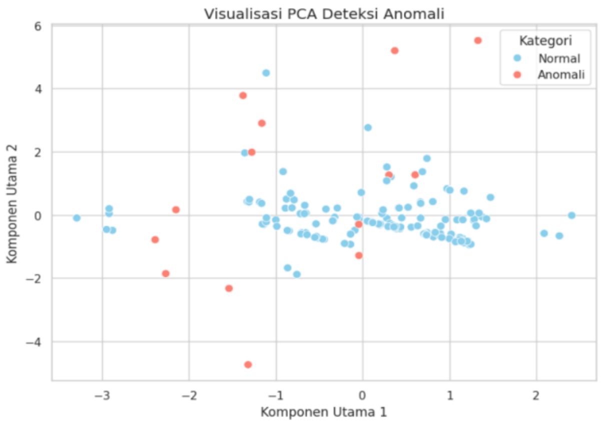


Figure 6. PCA Visualization of Anomaly Detection

The PCA (Principal Component Analysis) visualization in Figure 6 shows a clear spatial distinction:

- Blue points represent normal data, tightly clustered around the origin (0, 0).
- Red points represent anomalous entries, scattered more broadly across the PCA space.

This separation indicates that the OCSVM model was able **to** effectively distinguish deviations from the learned "normal" pattern, validating its pseudo-accuracy rate of 90.86% in identifying 16 outliers among 175 archival entries.

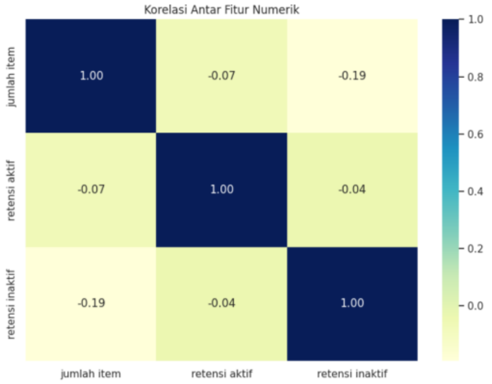


Figure 7. Heatmap of Pearson Correlation Among Numeric Features

A heatmap analysis was conducted to examine linear correlations among key numerical features: *Number of Items*, *Active Retention*, and *Inactive Retention*.

a. Correlation between Number of Items and Active Retention**:** -0.07 (no linear relationship).
b. Correlation between Number of Items and Inactive Retention**:** -0.19 (weak negative correlation).
c. Correlation between Active and Inactive Retention**:** -0.04 (minimal correlation).

These results suggest that the three features are relatively independent, implying low multicollinearity**.** This is beneficial for machine learning models such as XGBoost and OCSVM, as independent features often carry distinct and complementary information**,** improving model accuracy and interpretability.

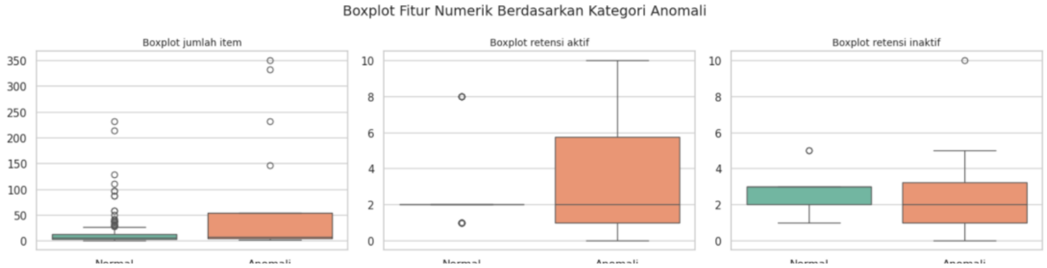


Figure 8. Boxplot of Numeric Features by Anomaly Category

The boxplot in Figure 8 compares the distributions of the three numerical features between normal and anomalous groups.

a. Boxplot: Number of Items. Anomalous entries exhibit a much wider distribution and higher item counts, with several outliers exceeding 300 items. Normal entries are more centered and consistent, suggesting extreme item count as a strong anomaly indicator.

b. Boxplot: Active Retention. Anomalies have higher medians and wider IQRs, reflecting unusual retention durations. Normal records are tightly distributed around a 2-year standard, indicating that anomalies deviate from institutional policy.

c. Boxplot: Inactive Retention. Although less pronounced, anomalies show greater variance and more outliers. While medians are similar across both categories, the inconsistent spread further supports the anomaly classification.

Overall, Number of Items and Active Retention are the most discriminative features, consistent with previous SHAP and feature importance analyses. These variables should be prioritized in data quality checks and anomaly mitigation strategies.

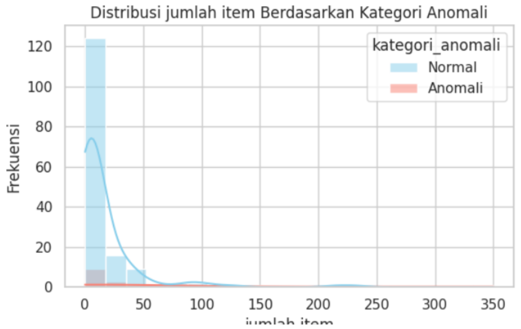


Figure 9. Distribution of Number of Items by Anomaly Category

The histogram and KDE curve in Figure 9 show:

a. Normal records cluster around 10–20 items, with most entries falling below 50.

b. Anomalies exhibit a wide and irregular distribution, with some entries exceeding 350 items.

This sharp contrast supports the hypothesis that abnormally high document volume is a key anomaly signal, potentially caused by: Data entry mistakes, Improper archival merging, Non-compliance with classification rules. This insight reinforces the importance of "Number of Items" as a critical variable in anomaly detection and supports its integration into AI-assisted data review processes.

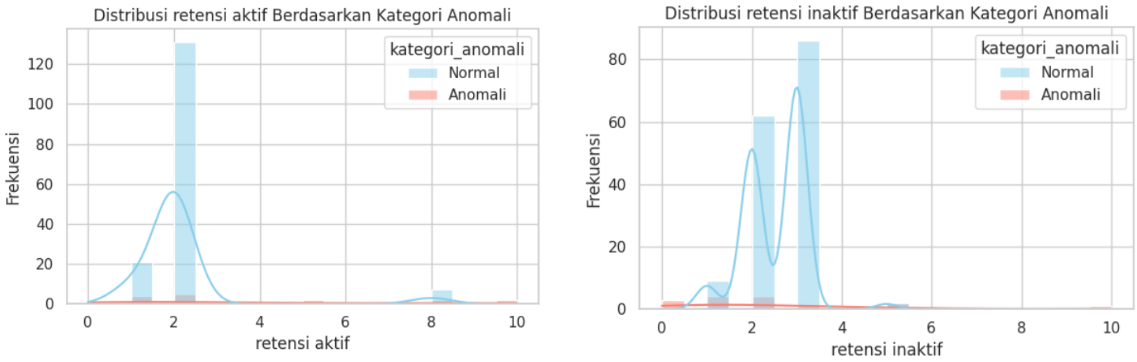


Figure 10. Distribution of Retention Periods by Anomaly Category

Caption Figure 10:

a. Active Retention (Figure 10, Left): Normal records are strongly centered around 2 years, suggesting policy compliance. Anomalies show broad dispersion, ranging from 1 to 9 years, indicating irregularities in data input or special cases.

b. Inactive Retention (Figure 10, Right): Normal entries cluster between 2–4 years with two distinct peaks, reflecting standard retention policies. Anomalous entries, although fewer, span extreme values (near zero up to 10 years), implying possible metadata inconsistencies.

These findings confirm that both retention fields, especially active retention, are reliable indicators for identifying metadata anomalies. They are particularly useful for triggering manual review processes or automating policy compliance checks.

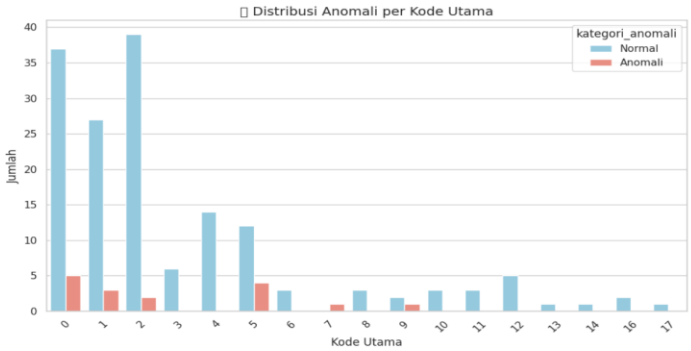


Figure 11. Anomaly Distribution by Major Classification Code

The histogram in Figure 11 shows the anomaly frequency by "Kode Utama" (Main Classification Code)**:**

a. Codes 0, 1, and 2 had the highest archival volumes and also showed the most frequent anomalies**,** particularly:
   - Code 0: 5 anomalies detected despite being a common administrative category,
   - Code 1 and 2**:** 3 and 2 anomalies respectively, suggesting higher risk in high-volume classifications.

b. Other codes like 4 and 5 also showed anomalies, despite lower total volume, likely due to document complexity or metadata inconsistency.

c. Codes 10–17 had lower frequency but still showed occasional anomalies (e.g., Codes 10 and 12), emphasizing that low frequency does not imply zero risk**.**

Implications**:**

1. High-volume classifications should be prioritized for automated validation and user training, as they carry a higher anomaly burden.
2. A machine learning–driven early warning system could use these insights to flag high-risk classifications, support manual audits, and guide metadata input improvements**.**

This analysis underscores the importance of a data-driven auditing approach in modern archival management systems to enhance accuracy, efficiency, and institutional accountability**.**

Table 4. Statistical Summary by Archival Anomaly Category

Anomaly Category	Number of Records	Average Number of Items	Average Active Retention	Average Inactive Retention
Anomalous	16	72.56	3.75	2.44
Normal	159	16.09	2.13	2.52

Descriptive Analysis of Table 4: Normal vs. Anomalous Archival Records. Table 4 provides essential descriptive insights into the characteristics of normal and anomalous archival records as identified by the OCSVM anomaly detection model. Three key numerical features are compared: *Number of Items*, *Active Retention*, and *Inactive Retention*, with the following interpretations:

a. Record Count. The majority of the dataset is classified as normal, comprising 159 records (90.8%), which aligns with the fundamental assumption of the OCSVM algorithm, that anomalies represent minority cases. The anomalous class contains 16 records (9.2%), which supports the model's high pseudo-accuracy score of 90.86%.

b. Number of Items. The mean number of items in anomalous records is 72.56, significantly higher than that of normal records (16.09). This marked difference suggests that high document volume per archival entry is a strong anomaly indicator. Possible causes include data duplication, redundant information, or bulk entry errors. This feature emerges as a primary variable in distinguishing irregular records.

c. Active Retention. Anomalous entries also exhibit longer active retention periods, with a mean of 3.75 years compared to 2.13 years for normal records. This may indicate that such records are retained for extended durations, possibly due to misclassification, incorrect retention scheduling, or non-standard archival policies.

d. Inactive Retention. The average inactive retention values are relatively similar between groups, 2.52 years for normal records and 2.44 years for anomalies. This suggests that the key differentiating variables are Number of Items and Active Retention, while Inactive Retention appears less decisive in identifying anomalies.

Implications and Statistical Test Results. The results suggest that a high number of items is a primary indicator of anomalies and should be incorporated as a key variable in predictive models or automated validation processes for digital archival systems. Likewise, excessively long active retention periods may signal misclassification, such as archives designated for disposal ("*Musnah*") that are incorrectly retained as active for extended durations.

As a recommendation, archival records that exhibit both a high number of items and extended active retention periods should be flagged for manual audit to verify their validity and to determine whether they reflect genuine archival needs or are the result of metadata entry errors. To statistically validate the distinction between normal and anomalous categories, independent samples t-tests and Mann-Whitney U tests were conducted on the numerical features. The resulting p-values were extremely small**,** particularly for Number of Items and Active Retention**,** indicating that the differences between the two groups are statistically significant**.**

These statistical findings support and strengthen the earlier anomaly detection results, both numerically and visually, confirming that anomalous records exhibit distinct patterns in key metadata fields. The combination of machine learning–based detection and inferential statistical testing provides a robust foundation for developing data-driven auditing frameworks in digital recordkeeping platforms such as SRIKANDI**.**

Table 5. Statistical Test Results for Numerical Features (Anomalous vs. Normal Records)

Feature	Statistical Test	Test Statistic	p-value	Significance Interpretation
Number of Items	t-test	t = 3.55	0.00063	Significant (p < 0.01)
	Mann-Whitney U	U = 525.0	0.00003	Significant (p < 0.01)
Active Retention	t-test	t = 2.60	0.0102	Significant (p < 0.05)
	Mann-Whitney U	U = 800.5	0.0059	Significant (p < 0.01)
Inactive Retention	t-test	t = -0.25	0.803	Not significant (p > 0.05)
	Mann-Whitney U	U = 1211.5	0.6906	Not significant (p > 0.05)

A p-value less than 0.05 indicates a statistically significant difference between the anomalous and normal groups for the corresponding feature. These results confirm that both Number of Items and Active Retention differ significantly between normal and anomalous records, thus supporting the patterns identified through OCSVM anomaly detection and accompanying visual analyses.

In contrast, the Inactive Retention feature does not exhibit a statistically significant difference between the two groups. This suggests that, unlike the other two features, it plays a less decisive role in distinguishing anomalous metadata entries within the archival system.

3.3. Interpretation and Discussion

The classification results indicate that the XGBoost algorithm is capable of accurately classifying archive status, particularly for the majority class. However, the imbalance between the "*Musnah*" (to be destroyed) and "Permanen" (permanent) classes suggests a need for mitigation strategies such as oversampling techniques (e.g., SMOTE) or class weight adjustment to improve performance for the minority class. The anomaly detection model effectively identifies irregular metadata patterns, such as archives with a high number of items but short time spans, inconsistencies between creation dates and archival periods, and active retention durations that deviate from archival policies. These findings align with prior studies by Bhara Nurpasma (2024) and Dewi Yulianti (2024)**,** which reported structural weaknesses in the metadata of the SRIKANDI electronic archive system. Hence, the application of machine learning proves to be both relevant and strategic in enhancing data integrity, improving classification accuracy**,** and supporting the proactive management of digital records**.**

The implications of this study demonstrate that the integration of ML approaches, namely XGBoost and One-Class SVM, can contribute significantly to the automatic detection of input errors and potential data corruption**.** Furthermore, these methods improve the efficiency of metadata-based document classification**,** while supporting the interoperability and validity of electronic records management systems. Compared to previous research using the same algorithm, Rao's (2024)[23] study yielded better accuracy. This is due to the different case studies and the larger number of variables. Furthermore, a limitation of this study is the limited data available, as the case studies ranged from 2022 to 2025. Therefore, more data is needed.

4. Conclusion

The results of this study demonstrate that the XGBoost classification model exhibits reasonably good performance in predicting the final status of archival records (i.e., *Musnah* vs. *Permanen*), achieving an accuracy of 77%, with the highest F1-score recorded for the "Musnah" category. However, performance on the "*Permanen*" class remains low, likely due to class imbalance in the dataset. Feature importance and SHAP analysis reveal that the most influential metadata attributes for classification are Number of Items**,** Active Retention Period**,** and Sub-Classification**.** This highlights the critical role of descriptive and temporal metadata attributes in determining archival retention status. The OCSVM anomaly detection model successfully identified 16 anomalous records (9.14%) out of 175 archival entries. These anomalies were typically characterized by abnormally high item counts, inconsistencies between retention duration and archival time span, and illogical or missing metadata entries—such as “No Document Available.” Furthermore, anomalies involving mismatches between creation dates and Digital Signatures (TTE) point to potential reliability issues in search, retention, and interoperability within the SRIKANDI archival system. The integration of XGBoost and One-Class SVM provides a robust framework for the systematic evaluation and validation of digital government archives, offering scalable support for improving data governance quality. The limitation of this research is that the data used is still limited, namely the SRIKANDI documents for 2022-2025, so the data for training and testing is still inadequate.

Based on the research findings, it is recommended to enhance archivist capacity through training on consistent metadata entry, implement data balancing techniques such as SMOTE to improve classification accuracy, and regularly apply anomaly detection models to support internal audits. Additionally, future development of the SRIKANDI system should integrate machine learning modules for metadata validation and classification recommendations. Lastly, an AI-based metadata cleansing system is needed to automatically detect and correct anomalies before documents are permanently stored in the digital archive.

References

[1] A. A. Musaddad, M. Niswah, K. Prasetyo, and S. Hardjati, “Implementasi Manajemen Kearsipan Di Sektor Publik,” *Jurnal Governansi*, vol. 6, no. 2, pp. 133–143, 2020, doi: 10.30997/jgs.v6i2.2843.

[2] M. Farrell, “Accountability as a mechanism and a virtue in Irish public sector recordkeeping,” *Records Management Journal*, vol. 34, no. 2–3, pp. 190–204, 2024, doi: 10.1108/RMJ-09-2023-0051.

[3] R. Touray, “A Review of Records Management in Organisations,” *OALib*, vol. 08, no. 12, pp. 1–23, 2021, doi: 10.4236/oalib.1108107.

[4] H. P. Nur Soulthoni and M. Itasari, “The Implementation of Electronic-Based Archiving to Accelerate Government Digitalization in Indonesia,” *Indonesian Journal of Innovation and Applied Sciences (IJIAS)*, vol. 5, no. 1, pp. 49–57, 2025, doi: 10.47540/ijias.v5i1.1735.

[5] E. U. Janah, P. F. Hidayah, R. N. Adysti, and R. H. Arjuna, “Peran Digitalisasi Arsip untuk Meningkatkan Efektivitas Manajemen Dokumen Arsip di Dinas Arsip Kota Semarang,” *Tsaqofah*, vol. 5, no. 1, pp. 474–484, 2024, doi: 10.58578/tsaqofah.v5i1.4517.

[6] M. D. Harahap and F. Trimurni, “Kualitas Sistem Informasi Kearsipan Dinamis Terintegrasi (Srikandi) dalam Pelayanan Administrasi di Dinas Perpustakaan dan Arsip Kabupaten Deli Serdang,” *Jurnal Penelitian Inovatif*, vol. 5, no. 1, pp. 295–312, 2025, doi: 10.54082/jupin.984.

[7] A. Supriyanto and K. Mustofa, “E-gov readiness assessment to determine E-government maturity phase,” *Proceeding - 2016 2nd International Conference on Science in Information Technology, ICSITech 2016: Information Science for Green Society and Environment*, pp. 270–275, 2017, doi: 10.1109/ICSITech.2016.7852646.

[8] A. Supriyanto, A. Jananto, J. A. Razaq, B. Hartono, and F. Damaryanti, “Alignment of KAMI Index with Global Security Standards in Information Security Risk Maturity Evaluation,” *Cybernetics and Information Technologies*, vol. 25, no. 2, pp. 173–192, 2025, doi: 10.2478/cait-2025-0018.

[9] M. Suepa, “Pengimplementasian Sistem Informasi Kearsipan Dinamis Terintegrasi (SRIKANDI) Pada Dinas Perpustakaan dan Arsip Daerah Kabupaten Pandeglang,” pp. 304–314, 2024, [Online]. Available:

https://conference.ut.ac.id/index.php/semnasip/article/download/3094/1299/8039?TSPD_101_R0=088cbb75e0ab2000b589ab1714292dc9adb5555617e71884101b96b50053e12ab619a4f27036fa6a08149700701430001869109d3ce7d88453a84033dc73310c3ec9c42657b8eaf30d76db28b5eef3c94eda

[10] E. T. Suharmanto and A. Supriyanto, "Assessment Of IDW And ANN On Daily Rainfall Data Imputation in Semarang Central Java," *Sinkron*, vol. 9, no. 1, pp. 382–394, 2025, doi: 10.33395/sinkron.v9i1.14452.

[11] Z. Chen, Z. W. Li, J. Huang, S. Z. Liu, and H. X. Long, "An effective method for anomaly detection in industrial Internet of Things using XGBoost and LSTM," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-74822-6.

[12] M. Aly and M. H. Behiry, "Enhancing anomaly detection in IoT-driven factories using Logistic Boosting, Random Forest, and SVM: A comparative machine learning approach," *Scientific Reports*, vol. 15, no. 1, pp. 1–17, 2025, doi: 10.1038/s41598-025-08436-x.

[13] M. Balega, W. Farag, X. W. Wu, S. Ezekiel, and Z. Good, "Enhancing IoT Security: Optimizing Anomaly Detection through Machine Learning," *Electronics (Switzerland)*, vol. 13, no. 11, 2024, doi: 10.3390/electronics13112148.

[14] S. Suangli, F. Fahmi, and E. M. Zamzami, "Performance Analysis of Support Vector Machine and Xgboost Classifier Algorithms in Predicting Data Heart Disease," *Proceedings - ICT 2023 - 29th International Conference on Telecommunications: Next-Generation Telecommunications for Digital Inclusion and Universal Access*, pp. 1–6, 2023, doi: 10.1109/ICT60153.2023.10374048.

[15] K. Yang, S. Kpotufe, and N. Feamster, "An Efficient One-Class SVM for Anomaly Detection in the Internet of Things," pp. 1–23, 2021, [Online]. Available: http://arxiv.org/abs/2104.11146

[16] S. K. Devineni, S. Kathiriya, and A. Shende, "Machine Learning-Powered Anomaly Detection: Enhancing Data Security and Integrity," *Journal of Artificial Intelligence & Cloud Computing*, vol. 2023, no. May 2023, pp. 1–9, 2023, doi: 10.47363/jaicc/2023(2)184.

[17] S. Brokensha, E. Kotze, and B. Senekal, "Machine learning for document classification in an archive of the National Afrikaans Literary Museum and Research Centre," *Sasa Journal*, vol. 56, no. 1, pp. 134–147, 2021.

[18] L. Bergman and Y. Hoshen, "Classification-Based Anomaly Detection for General Data," *8th International Conference on Learning Representations, ICLR 2020*, pp. 1–12, 2020.

[19] P. R. Satriawan, G. M. Ferdinand, I. N. P. S. Natha, I. G. A. P. S. D. Sastrawan, N. W. Marti, and N. P. N. P. Dewi, "Evaluasi Dan Perbandingan Algoritma KlasifikasiDalam Analisis Penggunaan Lahan DenganTeknologi Remote Sensing: Sebuah Kajian Sistematik," *INSERT: Information System and Emerging Technology Journal.*, vol. 5, no. 2, pp. 97–109, 2024.

[20] A. Mehdary, A. Chehri, A. Jakimi, and R. Saadane, "Hyperparameter Optimization with Genetic Algorithms and XGBoost: A Step Forward in Smart Grid Fraud Detection," *Sensors*, vol. 24, no. 4, 2024, doi: 10.3390/s24041230.

[21] S. He and Y. Chen, "A high-performance extreme gradient boosting outlier detection framework for integrating the outputs of diverse anomaly detectors for detecting mineralization-related geochemical anomalies," *Journal of Geochemical Exploration*, vol. 273, no. February, p. 107741, 2025, doi: 10.1016/j.gexplo.2025.107741.

[22] A. Shilton, S. Rajasegarar, and M. Palaniswami, "Multiclass anomaly detector: The cs++ support vector machine," *Journal of Machine Learning Research*, vol. 21, pp. 1–39, 2020.

[23] G. B. N. Rao, N. Kumar, A. Kumar, and P. Raj, "Efficient Intelligent Network Intrusion Detection for SDN Using XGBoost," *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*, pp. 1–9, 2024, doi: 10.1109/ICCCNT61001.2024.10723841.

[24] N. Pinon, R. Trombetta, and C. Lartizien, "One-Class SVM on siamese neural network latent space for Unsupervised Anomaly Detection on brain MRI White Matter Hyperintensities," *Proceedings of Machine Learning Research*, vol. 227, pp. 1783–1797, 2023.