Development of a Web-Based Application for Verifying the Authenticity of Digital Population Documents Using Extreme Gradient Boosting (XGBoost)

Andi Arma Gansa [1,a], Muh. Irsan S, [1,b] and Alvian Bastian [1,*,c]
[1] Department of Informatics and Computer Engineering, Politeknik Negeri Ujung Pandang, Makassar, Indonesia
[a] andiarmagansa@gmail.com, [b] muhirsan@poliupg.ac.id, [*,c] alvianbastian@poliupg.ac.id

***Abstract*— Along with the increasing use of digital population documents in Indonesia, the risk of document forgery has become a serious challenge that threatens data validity in various administrative processes. This research aims to develop and evaluate the performance of a website-based application capable of automatically verifying the authenticity of digital population documents. The proposed method integrates the Image Forensics approach with Machine Learning.The system extracts a series of features from the document image , including metadata analysis , Error Level Analysis (ELA) processed using a Convolutional Neural Network (CNN), and image statistical features. This feature set is then analyzed using the Extreme Gradient Boosting (XGBoost) algorithm to classify the document as "Authentic" or "Fake". The application is built with an e-Government architecture, using Laravel for the frontend and Flask as the API server for the analysis process. Performance testing results using a confusion matrix against 90 data samples show that the system successfully achieved an accuracy level of 93.33%. The average response time for one complete prediction cycle was recorded at 12.66 seconds. These results prove that the developed system is an effective and efficient solution for mitigating the risk of digital population document forgery.**

Keywords*— *XGBoost; Digital Population Documents; Document Verification; Image Forensics; ELA-CNN.

I. Introduction

The development of digital technology has brought a significant impact on document management transformation, encouraging the rapid use of digital documents to increase operational efficiency and simplify administrative processes in various sectors, including government. In Indonesia, this digitalization is driven by the e-Government policy to improve the efficiency and quality of public services [1]. Important examples include the legalization of electronic land certificates and the use of digital documents in population administration. However, the increased use of digital population documents such as Indonesian Family Cards (Kartu Keluarga/KK) and Indonesian Birth Certificates (Akta Kelahiran) faces a serious challenge in Indonesia, namely document forgery [2]. Document and letter forgery crimes, such as KTP (ID card) forgery using software to change identities in order to open bank accounts, are a growing problem, with the Indonesian National Police acting on an average of 7 cases per day in 2022 [3]. The conventional approach to verification relies heavily on manual forensic analysis, where analysts must use software like JPEGsnoop to check metadata and techniques such as Error Level Analysis (ELA) to visually identify image compression anomalies [4]. Previous research has also shown the effectiveness of ELA in detecting image file manipulation [5].

To address this challenge, this research proposes a modern approach that combines image forensics techniques with machine learning. This system will automatically extract a series of features to detect manipulation, including:

1. Metadata analysis to track software traces.
2. Statistical features to measure image property anomalies (e.g., Mean Intensity, Standard Deviation, Entropy, Edge Count, and FFT).
3. ELA analysis further processed using a Convolutional Neural Network (CNN) to obtain a manipulation probability score.

The results of this extraction are then quantified into tabular data and analyzed using the Extreme Gradient Boosting (XGBoost) algorithm, which is designed efficiently for tabular data, to classify document authenticity.

Based on the description above, the formulation of the research problems are:

1. How to develop a website-based application for verifying the authenticity of digital population documents.
2. How to measure the performance of a website-based application for verifying the authenticity of digital population documents using Extreme Gradient Boosting (XGBoost).

The objective of this research is to create a website-based application for the verification of digital population document authenticity and measure its performance using XGBoost. It is hoped that this research can be a means to minimize document forgery crimes. The scope of the research is limited to the verification of Indonesian Family Cards (Kartu Keluarga/KK) and Indonesian Birth Certificates (Akta Kelahiran) in JPG and PNG formats, using dummy data.

II. Research Methodology

A. Method

This research uses the Agile Software Development Life Cycle (SDLC) Method, which is flexible and iterative, with the following stages: Requirements, Design, Develop, Test, Deploy, and Review [6].

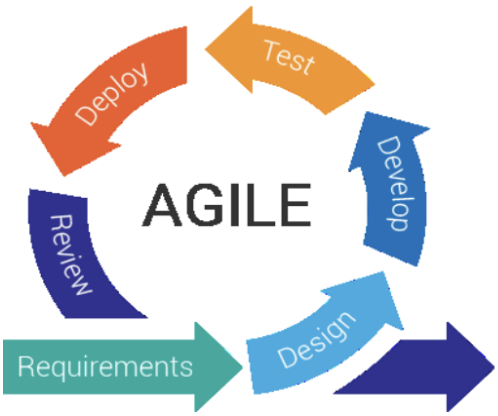


Figure 1 Metode Agile

1. Requirements: Involves a literature study, needs analysis, and the preparation of the product backlog. Data collection includes collecting dummy document images (Indonesian Birth Certificates (Akta Kelahiran) and Indonesian Family Cards (Kartu Keluarga)) that will be used to train and test the classification model.
2. Design: Encompasses creating the application architecture (Flowchart, Activity Diagram, Use Case Diagram) and sketches (wireframes) for the user interface (User and Admin).

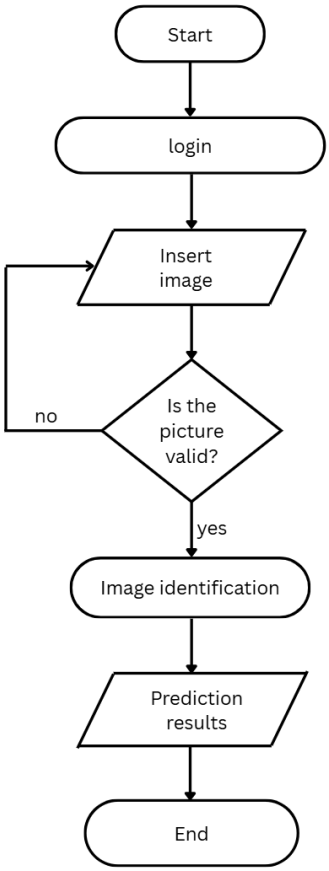


Figure 2 Flowchart

3. Development: This phase translates the design into functional program code, based on the tasks in the sprint backlog.
4. Testing: Performed continuously (Unit Testing, Integration Testing, and User Acceptance Testing) to ensure quality.
5. Deploy: The process of releasing the software (increment) to the testing environment for UAT (User Acceptance Testing).
6. Review: A formal evaluation session at the end of the cycle to obtain feedback and update the product backlog.

Prediction Model Architecture The model uses a hybrid approach (decoupled architecture) for authenticity verification. This architecture separates:

1. Web Application (Frontend): Built with Laravel (PHP) and hosted on cPanel.

2. API Server (Backend): Built with Flask (Python) and deployed on a Windows Virtual Private Server (VPS) to manage all heavy computational logic.

The identification process involves:

1. Metadata Analysis: The system tracks traces of software that may have been used to manipulate the documents, such as Exif data or file compression properties.
2. Digital Image Statistical Features: Extraction of statistical features to measure image property anomalies, including Mean Intensity, Standard Deviation (Std Dev), Signal-to-Noise Ratio (Snr), Entropy, Edge Count, Mean Fast Fourier Transform (FFT), and Std FFT.
3. ELA Analysis Processed by CNN: Error Level Analysis (ELA) is performed, and the output from this ELA is further processed using a Convolutional Neural Network (CNN) to obtain a manipulation probability score.
4. XGBoost Classification: The combined tabular data from the three extracted features (metadata, statistics, and ELA-CNN score) is fed into the Extreme Gradient Boosting (XGBoost) algorithm to generate the final prediction ("Authentic" or "Fake").

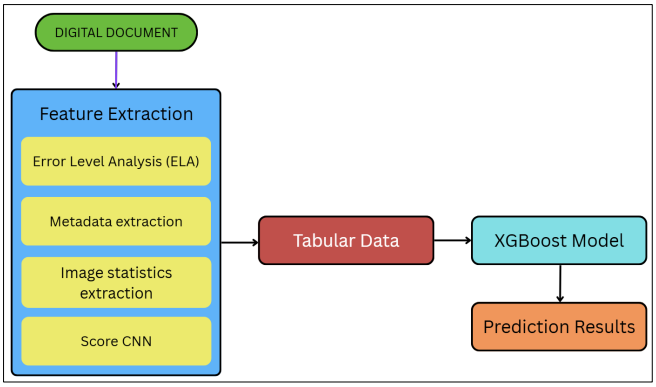


Figure 3. Identification Stages

Testing Steps

The testing steps include:

1. Functionality Testing: Uses the Black Box Testing method to verify that all application functions (e.g., document upload, verification, result display) run according to the design.
2. Model Performance Testing: Uses the Confusion Matrix (TP, TN, FP, FN) to calculate key metrics (Accuracy, Precision, and Recall).
3. Response Time Testing: Measures the total time required for the system to respond to a user request, calculated from when the document is identified until the API server returns the result.

B. Research Data

1. Error Level Analysis

The ELA implementation uses the basis of 3 standard color components: red, green, and blue (RGB). The parameters used in this research are a comparative image quality level of 60% and an error scale of 25 [7]. The value of each color component (R, G, B) at the pixel coordinates and of the image being analyzed (Input) is calculated with the color component and pixel at the same coordinate in the reconstructed image (Comp). This value is then multiplied by the scale (S) to obtain the component value of the same pixel in the ELA result (ELA) [8]. Mathematically, this process can be written as:

$$ELA_{x,y}R = (S|(Imput_{x,y}\ R - Comp_{x,y}\ R|))\%256 \quad (1)$$

$$ELA_{x,y}G = (S|(Imput_{x,y}\ G - Comp_{x,y}\ G|))\%256 \quad (2)$$

$$ELA_{x,y}B = (S|(Imput_{x,y}\ B - Comp_{x,y}\ B|))\%256 \quad (3)$$

The use of the scale creates color component values that are not directly proportional once the process is complete. The modulo 256 operation is necessary because of the limited range of values for JPEG and PNG image formats, which is 8-bit, having a maximum value of 255 [7]. An example image of an Indonesian Birth Certificate (Akta Kelahiran) document can be seen in the following figure.

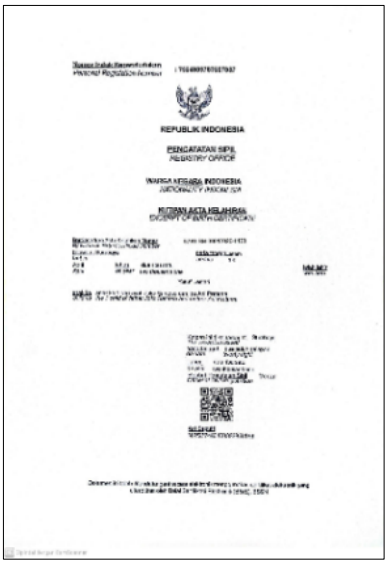


Figure 4. Document Sample

Next, the image processed using ELA with a comparative image quality level of 60% and an error scale of 25 can be seen in the following figure.

Figure 5. Document Sample After ELA

2. Metadata

Metadata plays an important role in document verification as it provides structured information that functions to explain, simplify searching, and facilitate the use and management of information. Details about the date and time of capture, the camera or cell phone used to take the picture, and other information are typically included as one of the additional details that complement and explain certain data for a photo or image [4]. An example of this can be seen in the following table.

Table 1 Metadata Sample

Fitur	Asli	Palsu
Size	345422	619609
YResolution	0/0	0/0
XResolution	0/0	0/0
Software	IJG Library	Adobe Photoshop
Length	APP2	APP14
Marker	472	33
Assessment	*Class* 1	*Class* 1

3. Digital Image Statistical Features

Statistical features in image classification refer to the extraction of numerical information that describes the statistical properties of the pixel intensity distribution within an image. These statistical features provide a numerical representation of the variation, spread, and distribution of pixel brightness within the image, which can be used for classification or pattern recognition purposes [9].

a. Mean Intensity: Measures the average brightness level of the image based on pixel intensity

$$Mean\ intensity\ = \frac{1}{N} + \sum_{i=1}^{N} I(i) \quad (4)$$

b. Standard Deviation (Std Dev): Measures the degree of variation or difference in contrast within the image.

$$Standar\ deviasi = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(I(i) - \mu)^2} \quad (5)$$

c. Signal-to-Noise Ratio (Snr): The ratio between the signal and the noise. A higher Snr indicates better image quality.

$$SNR = \frac{\mu}{\sigma} \quad (6)$$

d. Entropy: A measure of complexity or randomness in the image. A high entropy value indicates the image has many variations in color and texture, while a low value indicates the image is more uniform.

$$Entropy = -\sum_{1=0}^{255} p(i) log_2 p(i) \quad (7)$$

e. Mean FFT: The average value of the image's Fourier transform, which describes the frequency distribution within the image.

$$Mean\ FFT = \frac{1}{N}\sum_{i=1}^{N} |FFT(I(i)| \quad (8)$$

f. Std FFT: Indicates the variation in the frequency domain.

$$Std\ FFT = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\left|FFT\left(I(i)\right)\right| - \mu FFT)^2} \quad (9)$$

g. Edge Count: Used to determine the number of contours or object boundaries based on the results of the edge detection algorithm.

where is the number of pixels in the image, and is the intensity value of the -th pixel. is the number of intensity levels, while is the probability of the -th intensity level occurring. The results can be seen in the following table.

Table 2 Image Statistics Sample

Features	Authentic	Fake
Mean intensity	241.79	241.77
Std dev	49.78	49.48
Snr	4.86	4.89
Entropy	2.20	2.34
Edge count	136322	136294
Mean fft	1061.66	1054.45
Std fft	11313.89	11310.95

4. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is an advanced algorithm of the Neural Network. The CNN algorithm yields the most significant results in digital image recognition because CNN is implemented based on the image recognition system in the human visual cortex[10]. AlexNet is one of the CNN architectures, representing a new breakthrough in deep learning by implementing ConvsNet combined with the Dropout Regularization technique, utilizing the Rectified Linear Unit (ReLU) as the activation function, and data augmentation. AlexNet is designed to perform classification with 1000 categories. The AlexNet architecture consists of 5 convolution layers, 3 pooling layers, 2 dropout layers, and 3 fully connected layers [11]. The AlexNet architecture can be seen in the following Figure.

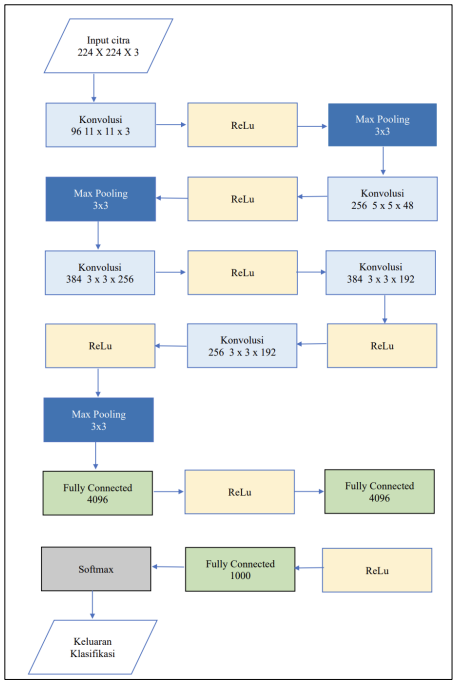


Figure 6. AlexNet Architecture

5. Extreme Gradient Boosting (XGBoost)

The XGBoost method is an advancement of gradient boosting proposed by Dr. Tianqi Chen from the University of Washington in 2014. Gradient boosting is an algorithm that can find optimal solutions for various problems, especially in regression, classification, and ranking. The basic concept of this algorithm is to iteratively adjust the learning parameters to lower the loss function (a model evaluation mechanism) [12]. XGBoost uses a more regularized model to build the regression tree structure, which can provide better performance and reduce model complexity to prevent overfitting. The final prediction result from XGBoost is the sum of the predictions from each regression tree. This decision tree-based algorithm performs well on data with categorical features and is less affected by data with imbalanced classes.

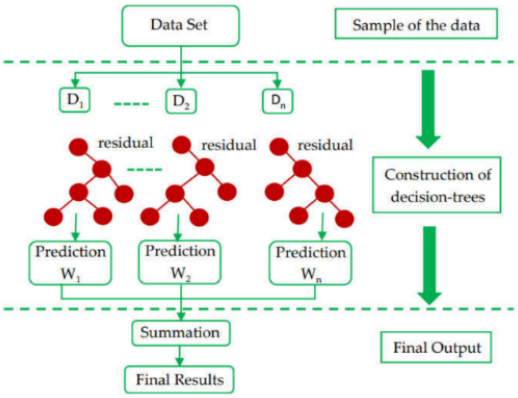


Figure 7. XGBoost Architecture

III. Results and Discussion

The research used a total of 571 document images (Indonesian Birth Certificates (Akta Kelahiran) and Indonesian Family Cards (Kartu Keluarga)), consisting of 273 authentic images and 298 fake images.

1. Model Training

Model training was conducted in stages:

a. CNN (for Indonesian Birth Certificates): Used 2,004 data (1,604 training, 400 validation). It achieved a Final Loss of 0.0812 and a Final Accuracy of 96.25%.

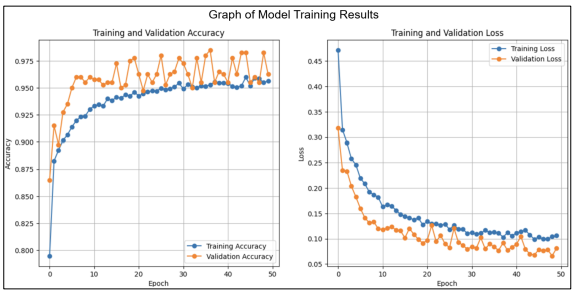


Figure 8. CNN Training Results for Birth Certificates

b. CNN (for Indonesian Family Cards): Used 2,103 data (1,683 training, 420 validation). It achieved a Final Loss of 0.2405 and a Final Accuracy of 91.19%.

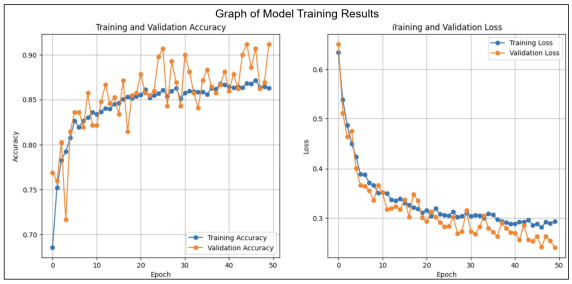


Figure 9. CNN Training Results for Family Cards

c. XGBoost: Training used 2,425 data samples (1,940 training, 485 validation). On the validation data, the model showed an accuracy of 95.05% (461 correct predictions), with 225 authentic and 236 fake documents correctly identified.

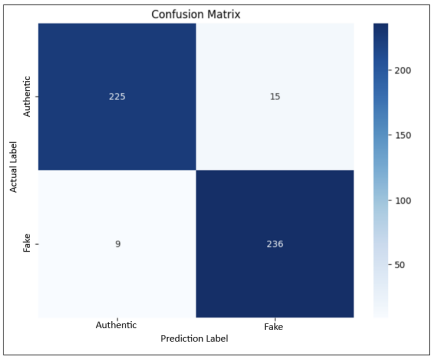


Figure 10. XGBoost Training Results

2. Model Performance Testing

Final performance testing was conducted using 90 test data samples (45 authentic and 45 fake), which consisted of various sources (Camscanner, standard scanner, mobile camera) and modifications (Adobe Photoshop, Picsay Pro).

Table 3 Performance Testing

	Predicted Authentic	Predicted Fake	Total
Actually Authentic	40 (TP)	5 (FN)	45
Actually Fake	1 (FP)	44 (TN)	45

$$Accuracy = \frac{TP+TN}{Total} = \frac{40+44}{90} x\ 100\% = 93.33\% \quad (10)$$

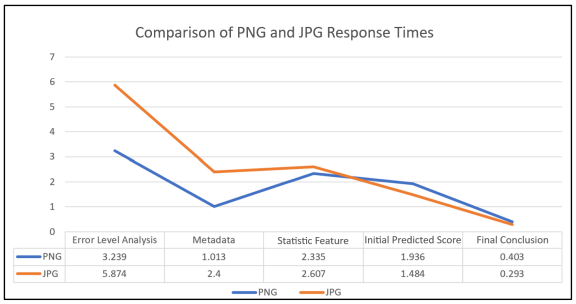
$$Precision = \frac{TP}{TP+FP} = \frac{40}{40+1} x\ 100\% = 97,56\% \quad (11)$$

$$Recall = \frac{TP}{TP+FN} = \frac{40}{40+5} x100\% = 88,89\% \quad (12)$$

The results indicate an Accuracy of 93.33% and a Precision of 97.56%. The very high precision validates the reliability of the model's positive predictions. Nevertheless, there were 5 cases of False Negative (authentic documents detected as fake), which represents an area for improvement.

3. Response Time Testing

This testing evaluates the system's speed in processing the main analysis features. The average time required for one complete image analysis process is 10.80 seconds.



	Error Level Analysis	Metadata	Statistic Feature	Initial Predicted Score	Final Conclusion
PNG	3.239	1.013	2.335	1.936	0.403
JPG	5.874	2.4	2.607	1.484	0.293

Figure 11. Comparison Chart of PNG and JPG

a. The Error Level Analysis (ELA) feature has the longest response time, averaging 4.56 seconds.

b. The processing time for the JPG format (12.658 seconds total) is much longer than the PNG format (8.926 seconds total), with ELA for JPG being almost twice as slow as ELA for PNG.

c. Other features, such as Metadata (1.71 seconds), Initial Prediction Score (1.71 seconds), and Final Conclusion (0.35 seconds), show good speed. The relatively long time is attributed to the complex and pixel-by-pixel intensive image processing required by the ELA feature, especially for JPG files that undergo lossy compression.

4. Functional Testing Analysis

Black Box Testing conducted by external developers showed that all 35 main functions (User Management, Core Image Analysis Functions, and Administration Functions) were successfully completed with a "Successful" status. Functionally, the system is solid and ready for use.

Figure 12. Website Display

IV. Conclusion

1. The website-based application for verifying the authenticity of digital population documents has been successfully developed by implementing a decoupled architecture that separates the frontend (Laravel) and the analysis server (Flask/VPS). The system is functional, capable of running the entire process flow from authentication to displaying the results of image forensic analysis and final classification.

2. The application's performance shows very effective and efficient results. The Extreme Gradient Boosting (XGBoost) classification model achieved an accuracy level of 93.33% against 90 test data. Although there is time variation in the initial ELA analysis stage, the time required for one complete end-to-end prediction process is 10.80 seconds, providing a step-by-step overview to the user.

3. In light of the findings presented in this study, future research is encouraged to employ larger-scale and more heterogeneous datasets to improve the robustness and generalizability of the proposed model. Moreover, a comprehensive comparative analysis involving a range of machine learning algorithms is highly recommended to rigorously assess model feasibility and performance. Such efforts are essential to identify the most effective methodological framework for advancing the development of a web-based application aimed at verifying the authenticity of digital population

documents using the Extreme Gradient Boosting (XGBoost) algorithm.

Acknowledgement

We would like to express our deepest gratitude to Department of Informatics and Computer Engineering at the State Polytechnic of Ujung Pandang for their valuable support.

References

[1] Kementerian Kominfo/Direktorat Jenderal Aplikasi Informatika, *Satu dekade pembangunan digital Indonesia 2014–2024*. Jakarta, 2024. Accessed: Sep. 09, 2025. [Online]. Available: https://aptika.kominfo.go.id

[2] A. Fachmi, "Peningkatan Kompetensi Pengelola Arsip Dengan Keahlian Dokumen Forensik," *Info Bibliotheca: Jurnal Perpustakaan dan Ilmu Informasi*, vol. 5, no. 1, pp. 72–88, Dec. 2023, doi: 10.24036/ib.v5i1.447.

[3] W. Anggara, P. Hafidati, and M. Kamil, "Pertanggungjawaban Tindak Pidana Pemalsuan Surat Ktp Yang Dapat Mengakibatkan Kerugian Pada Orang Lain," Jurnal Pemandhu, vol. 5, no. 1, pp. 229–250, 2024.

[4] K. E. Purnama, C. Rozikin, and A. A. Ridha, "Analisis Forensic Citra Digital Menggunakan Teknik Error Level Analysis Dan Metadata Berdasarkan Metode NIST," 2023.

[5] R. Abrori, M. Fitria, and H. Bullah, "Mendeteksi Orisinalitas Citra Menggunakan Teknik Error Level Analysis dan Metadata," Indo-Fintech Intellectuals: Journal of Economics and Business, vol. 4, no. 2, pp. 170–178, Jun. 2024, doi: 10.54373/ifijeb.v4i2.1169.

[6] W. D. Prastowo, D. Danianti, and A. Pramuntadi, "Analisis Risiko Pada Pengembangan Perangkat Lunak Menggunakan Metode Agile Dan RAD (Rapid Application Development)," Citizen : Jurnal Ilmiah Multidisiplin Indonesia, vol. 3, no. 3, pp. 169–174, Aug. 2023, doi: 10.53866/jimi.v3i3.388.

[7] M. F. Rahman, R. Rachman, J. Putra, and Y. Wihardi, "Analisis Statistik dan Implementasi Image Masking Berdasarkan Hasil Error Level Analysis Pada Gambar Digital Statistical Analysis and Image Masking Implementation Based on the Results of Error Level Analysis on Digital Images," 2020. [Online]. Available: https://ejournal.upi.edu/index.php/JATIKOM

[8] F. Harahap, "Deteksi Foto Manipulasi Dengan Tools Forensicallybeta dan Imageforensic.org Dengan Metode Error Level Analysis (ELA)," 2021.

[9] M. Muchtar and R. A. Muchtar, "Perbandingan Metode Knn Dan Svm Dalam Klasifikasi Kematangan Buah Mangga Berdasarkan Citra Hsv Dan Fitur Statistik," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4010.

[10] D. Nasution, "Prosiding SNASTIKOM: Seminar Nasional Teknologi Informasi & Komunikasi Paper Klasifikasi Objek Menggunakan Metode Convolutional Neural Network (CNN)," 2022.

[11] S. Yuliany and A. Nur Rachman, "Implementasi Deep Learning pada Sistem Klasifikasi Hama Tanaman Padi Menggunakan Metode Convolutional Neural Network (CNN)," 2022.

[12] E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," JOMTA Journal of Mathematics: Theory and Applications, vol. 4, no. 1, 2022.