Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Ensemble of Random Forest and Adaboost Algorithms for Human Skin Disease Classification

Deni Serviana[1]*, Octara Pribadi[2], Robet[3]

[1,2,3]STMIK TIME Medan, Indonesia
deniserviana02@gmail.com[1*], octarapribadi@gmail.com[2], robertdetime@gmail.com[3]

Abstract

Skin diseases are a common health issue that is often overlooked, despite the fact that they can develop into serious conditions if not addressed early. To assist in early detection, this study developed a skin disease classification system based on digital images using an ensemble learning approach with Random Forest and AdaBoost algorithms. The dataset used was sourced from Kaggle and includes 9 types of skin diseases: Actinic Keratosis, Atopic Dermatitis, Benign Keratosis, Dermatofibroma, Melanocytic Nevus, Melanoma, Squamous Cell Carcinoma, Tinea Ringworm Candidiasis, and Vascular Lesion. The system was developed as a website that allows users to upload images for automatic classification. Test results show that the Random Forest algorithm achieved an accuracy of 65.66%, precision of 67.93%, recall of 65.66%, and a misclassification error of 34.34%. Meanwhile, the combination of Random Forest and AdaBoost improved the accuracy to 71.69%, precision to 75.23%, recall to 75.23%, and reduced the misclassification error to 28.31%. The 6.03% increase in accuracy indicates that the ensemble approach provides better and more stable classification compared to single algorithms. This research is expected to support the development of early detection systems for skin diseases based on artificial intelligence.

Keywords: *Skin Disease, Image Classification, Machine Learning, Random Forest, AdaBoost, Ensemble Learning, Early Detection*

1. Introduction

Health is one of the most important and valuable aspects of human life, as anyone can experience health problems [1]. One particularly troublesome health issue is skin disease. Skin disease refers to disorders affecting the outer part of the body, characterized by symptoms such as itching, pain, numbness, and redness. Causes may include sun exposure, viruses, a weakened immune system, microorganisms, microbes, fungi, and personal hygiene factors [2]. Indonesia, as a country with a tropical climate, is vulnerable to bacterial, viral, and fungal infections that can affect human skin. However, many people tend to underestimate skin diseases, considering them harmless and non-lethal. This view is misleading, as untreated skin diseases can spread and become more difficult to cure [3].In reality, many people are reluctant to consult a doctor and often choose to ignore their skin problems. Several barriers contribute to this, including the limited number of specialists, long patient queues, unavailability of doctors in certain areas, and the distance required to access healthcare. This situation is exacerbated by the uneven distribution of doctors. Moreover, the treatment of skin diseases highly depends on the type and severity of the condition. These factors indicate that skin disease remains a significant public health issue, primarily due to low awareness about skin health and environmental hygiene [4].

In today's digital era, everything has become increasingly automated and interconnected, with computers designed to mimic human abilities. Modern computers can now read and recognize objects, offering many applications—especially in the healthcare sector. As this technology evolves, two primary learning methods have emerged: supervised learning and unsupervised learning. This study applies supervised learning through image classification to distinguish different types of skin diseases, enabling early detection and timely treatment before the conditions worsen [5]. Previous studies have implemented supervised learning methods for skin disease image classification as an early detection tool. For example, a study conducted in 2023 used the Random Forest algorithm for skin cancer classification. The results showed that the testing process using segmented data successfully classified the images with an accuracy of 81%, which falls into the category of good classification. A more recent study in 2024 applied the Convolutional Neural Network (CNN) algorithm to classify skin cancer. This study reported that at the 50th epoch, the model achieved an accuracy of 99.01% [6].

Based on the previous research, this study applies ensemble learning by combining multiple algorithm models to improve classification accuracy. The algorithms used in this study are Random Forest and AdaBoost. By integrating Random Forest with AdaBoost, it is expected that the classification accuracy will improve, as previous research has shown that incorporating AdaBoost can increase classification accuracy by 1–2%. Therefore, based on the aforementioned explanation, a comparative study will be conducted under the title: "Ensemble of Random Forest and AdaBoost Algorithms for Human Skin Disease Classification." [7]

2. Literature Review

2.1. Digital Image

A digital image is an image formed from elements called pixels. Each pixel has a specific numeric value that represents intensity, which is displayed in grayscale or color [8].

2.2. Digital Image Processing

Image processing opens the process of processing pixels in digital images for specific purposes. At the beginning of processing, this is done to improve image quality and extract information using convolution kernels applied through matrix operations based on the processing system. The process produces numerical values that computers understand and continue with computational processes to retrieve useful information from the image. The image processing process is in line with the process of information extraction and image quality enhancement [8].

2.3. Computer Vision

Image processing opens the process of processing pixels in digital images for specific purposes. At the beginning of processing, this is done to improve image quality and extract information using convolution kernels applied through matrix operations based on the processing system. The process produces numerical values that computers understand and continue with computational processes to retrieve useful information from the image. The image processing process is in line with the process of information extraction and image quality enhancement [8].

2.4. Machine Learning

Machine Learning (ML) is one branch of computer science based on pattern recognition and learning theory in artificial intelligence. ML is a study that provides computers with the ability to learn without being explicitly programmed. In ML, there are various algorithms that are developed according to the data provided, which will then be used in systems that do not use static programs [9].

2.5. Knowledge Discovery

Knowledge discovery, or knowledge discovery in databases (KDD), is one method for analyzing useful patterns in large, raw, and noisy data. KDD consists of several stages and is often confused with data mining, although data mining is only one part of the KDD process. Below in Figure 2.7 is a description of the KDD process [10].

2.6. Data Mining

Data mining is the process of discovering information or patterns from large datasets, where the data is very raw, redundant, and not yet ready to be used. Data mining has an important role in machine learning and artificial intelligence, as it is able to find hidden knowledge with very high accuracy. Data mining is part of knowledge discovery in databases (KDD) [10].

2.7. Random Forest

Random Forest is an ensemble method consisting of a collection of decision trees used to classify data into certain classes. The first step in building a decision tree is by determining the entropy and information gain values. There are two main formulas used by the Random Forest algorithm to construct decision trees, including the following [11]:

1. Entropy

Entropy is used as an approach to determine attribute impurity. The calculation of entropy values can be seen in the following formula:

$$Entropy(S) = \sum_{i=1}^{n} - p_i log_2 \, p_i \quad (1)$$

Explanation:

S: Dataset
n: Number of classes
p_i: Proportion of class i in dataset S

2. Information Gain

Information gain is the value of information obtained when selecting a node. The gain calculation is very influential on the root node and the splitting node. If the result is still in the form of zero or negative, the process will stop because the output value is considered insignificant.

$$Information\ Gain\ (A) = Entropy(S) = \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2)$$

Explanation:

A: Attribute
S: Dataset
S_i: Number of samples for value i
n: Number of class values

2.8. Adaptive Boosting

The Adaptive Boosting algorithm is an algorithm designed to improve decision-making accuracy by combining several weak learners into a strong learner using the boosting platform. This algorithm is intended for supervised learning that takes attributes and labels from the dataset and splits them into subsets that are then evaluated and corrected based on performance. The step-by-step process of the AdaBoost algorithm is as follows [12]:

1. Initialize the weights of the dataset with equal values:

$$W_i^1 = \frac{1}{n} \tag{3}$$

2. For each m = 1, 2, 3, …., M, do:

3. Build a classification model $G_m(X_i)$ using the weights $W_i^{(m)}$.

4. Calculate the classification error using the formula:

$$\varepsilon_m = \frac{\sum_{i=1}^{n} w_i^{(m)}, G_m(X_i) \neq Y_i}{\sum_{i=1}^{n} w_i^{(m)}} \tag{4}$$

5. Compute the value of αm using the formula:

$$\alpha_m = log\frac{1 - \varepsilon_m}{\varepsilon_m} \tag{5}$$

6. Update the misclassification weight coefficient using the formula:

$$\begin{aligned} w_i^{m+1} &= w_i^m \exp(-\alpha_m), G_m(X_i) = Y_i \\ w_i^{m+1} &= w_i^m \exp(\alpha_m), G_m(X_i) \neq Y_i \end{aligned} \tag{6}$$

7. Final classification decision is made using the formula:

$$Y = Sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right) \tag{7}$$

2.9. Confusion Matrix

The Confusion Matrix is one of the common ways used in the evaluation process of models in data mining classification to determine decision accuracy. The Confusion Matrix is presented in a table that shows the number of correctly and incorrectly classified data and the total data used for evaluation [13].

2.10. Unified Modeling Language

Unified Modeling Language (UML) is a standard modeling language designed to visualize, specify, construct, and document software system artifacts. UML is a development of three methods created by Grady Booch, Ivar Jacobson, and James Rumbaugh under Rational Software in the 1990s. UML has become the industry standard maintained by the Object Management Group (OMG) [14].

2.11. Use Case Diagram

A Use Case Diagram is one of the modeling tools in systems, part of the Unified Modeling Language (UML), used to describe interactions between users (actors) and systems. The creation of Use Case Diagrams aims to define the behavior of a system in a specific context, which makes the system easier to understand by clearly defining the language used [15].

2.12. Entity Relationship Diagram

An Entity Relationship Diagram (ERD) is a top-down technique in database design, typically used to identify entities and the relationships between those entities in a system. The Entity Relationship Diagram is used to visualize a conceptual data model, where each entity represents a table in the system and each relationship represents the relationship between entities [15].

3. Research Methodology

3.1. Data Collection Method

The data collection methods used in this study include the following:

1. Experimental Method
   Conducting experiments by searching for research datasets through the Kaggle website.

2. Literature Study Method
   Collecting theoretical data through journals, print media, or reference sources from the internet.

3.2. System Analysis

System analysis in this study is divided into three process stages, namely:

1. Process analysis based on field conditions, which includes analyzing the existing system used in the field, specifically the system used for early detection of human skin diseases and its associated issues.
2. Proposed method analysis, namely the Random Forest and Adaptive Boosting algorithms, by presenting simple case examples and demonstrating how these algorithms perform skin disease classification in humans.
3. Proposed system analysis, which includes system visualization or design modeling, including the features to be provided. This system is modeled using modeling tools such as a Use Case Diagram.

3.3. System Design

System design in this study is divided into two stages, namely:
1. Designing a user interface prototype using the Balsamiq Mockup 3 software.
2. Designing data storage, showing relationships between data to be modeled using the Entity Relationship Diagram (ERD).

3.4. Research Flowchart

The following flowchart illustrates the stages of the research process, from data collection to evaluation:

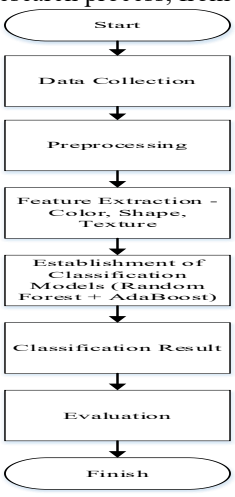

Fig. 1: Flowchart Illustrates

4. Research Results

The research results consist of the development of a website for ensemble-based analysis using the Random Forest and AdaBoost algorithms to classify human skin diseases. Below is a summary of the design and appearance of the skin disease classification system using the ensemble of Random Forest and AdaBoost algorithms:

4.1. Login Admin Page

The admin login page provides an authentication form with fields for email and password that ensure the user is a registered user who has the authority to access the system.

Fig. 1: Login Admin View

4.2. Skin Disease Classification Form Page

The skin disease classification form contains a form that allows users to upload a skin image to be classified. After the image is uploaded, the user can choose the classification option using one of the available ensemble algorithms, namely Random Forest or AdaBoost. This form also includes a button to run the classification process and display the predicted skin disease based on the uploaded image. The design of this page is intended to provide a user-friendly experience with an intuitive layout and focus on its primary function, which is classification. At the bottom, the page displays the classification result summary based on the input image

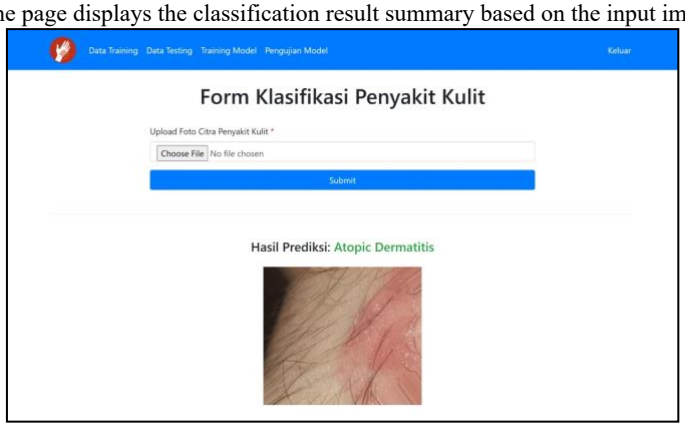

Fig. 3: Skin Disease Classification Form Display

4.3. Training Data Management View

The training data management view contains a list of available training data information. Additionally, the administrator can add, edit, and delete training data using the available buttons.

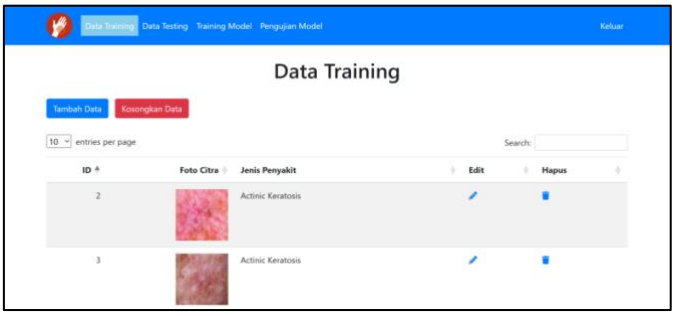

Fig. 4: Training Data Management View

4.4. Add Training Data Form View

The add training data form view contains input fields in a form to fill in training data information. The administrator simply needs to fill in the form fields and then click the submit button to add the training data.

Fig. 5: Add Training Data Form View

4.5. Edit Training Data Form View

The edit training data form view contains input fields in a form to edit training data information. The administrator can edit the form fields that are already filled in and then click the submit button to update the training data.

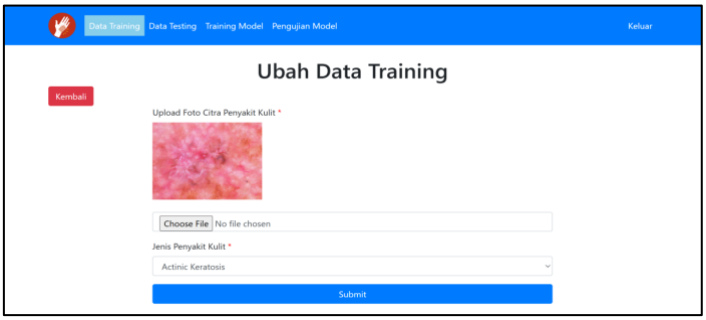

Fig. 6: Edit Training Data Form View

4.6. Testing Data Management View

The testing data management view displays a list of available testing data information. Additionally, the administrator can add, edit, and delete testing data using the available buttons.

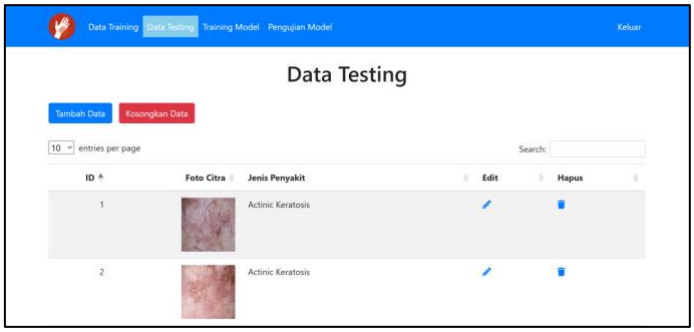

Fig. 7: Testing Data Management View

4.7. Add Testing Data Form View

The add testing data form view contains form input fields to fill in testing data information. The administrator simply needs to complete the form and then click the submit button to add the testing data.

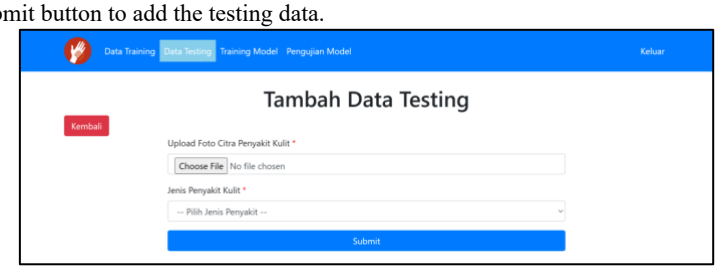

Fig. 8: Add Testing Data Form View

4.8. Edit Testing Data Form View

The edit testing data form view contains form input fields to modify the testing data information. The administrator can make changes to the pre-filled form and then click the submit button to update the testing data.

Fig. 9: Edit Testing Data Form View

4.9. Training Model Display

The training model view displays a pop-up form that shows the system loading process during model training. When the training process is completed, the model will be saved in several files in .joblib format.

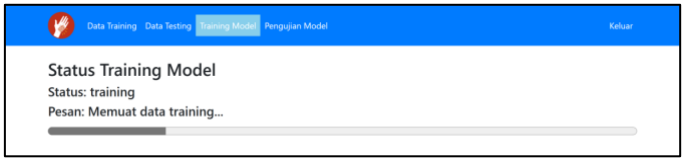

Fig. 10: Training Model Display in Progress

When the training process is completed, a display like the following image will appear:

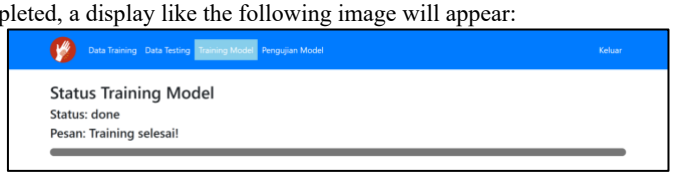

Fig. 11: Completed Training Model Display

4.10. Model Testing Display

The model testing view contains an image of the Confusion Matrix Plot results from testing the ensemble algorithm of Random Forest and AdaBoost, along with accuracy, precision, sensitivity, and misclassification error information from the model used in this research.

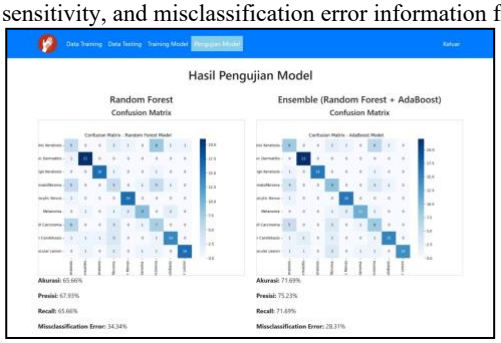

Fig. 12: Model Testing Display

5. Conclusions

The final stage outlines the conclusions of the research that has been conducted:

1. The website developed successfully implemented a human skin disease classification system using machine learning algorithms. This system is capable of early detection of 9 types of skin diseases based on images uploaded by users.
2. Based on the test results, the Random Forest algorithm achieved an accuracy of 65.66%, precision of 67.93%, recall of 65.66%, and a misclassification error of 34.34%. Meanwhile, the ensemble combination of Random Forest and AdaBoost showed a slight improvement in performance with an accuracy of 71.69%, precision of 75.23%, recall of 75.23%, and misclassification error of 28.31%. Although the increase in accuracy was only 6.03%, this ensemble approach proved to offer better stability and slightly improved performance in skin disease classification. These results address the second research question, which concerns the performance analysis of the two algorithms in classifying human skin diseases.

References

[1] Dona, H. Maradona, and Masdewi, "Sistem Pakar Diagnosa Penyakit Jantung Dengan Metode Case Based Reasoning (Cbr)," *Zo. J. Sist. Inf.*, vol. 3, no. 1, pp. 1–12, 2021, doi: 10.31849/zn.v3i1.6442.

[2] T. Srisantyorini and N. F. Cahyaningsih, "Analisis Kejadian Penyakit Kulit pada Pemulung di Tempat Pengolahan Sampah Terpadu (TPST) Kelurahan Sumur Batu Kecamatan Bantar Gebang Kota Bekasi," *J. Kedokt. dan Kesehat.*, vol. 15, no. 2, p. 135, 2019, doi: 10.24853/jkk.15.2.135-147.

[3] R. Rismanto, Y. Yunhasnawa, and M. Mauliwidya, "Pengembangan Sistem Pakar Untuk Diagnosa Penyakit Kulit Pada Manusia Menggunakan Metode Naive Bayes," *J. Ilm. Teknol. Inf. dan Robot.*, vol. 1, no. 1, pp. 18–24, 2019, doi: 10.33005/jifti.v1i1.8.

[4] K. P. Pohan and C. Chairunisah, "Sistem Pakar Mendiagnosa Penyakit Kulit Pada Manusia Menggunakan Metode Naive Bayes Classifier Berbasis Web," *J. SAINTIKOM (Jurnal Sains Manaj. Inform. dan Komputer)*, vol. 23, no. 1, p. 204, 2024, doi: 10.53513/jis.v23i1.9521.

[5] N. A. Winanti, D. P. Martiyaningsih, C. A. A. Soemedhy, and U. Athiyah, "Analisis Klasifikasi Citra Kanker Kulit dengan Random Forest," *Remik*, vol. 7, no. 1, pp. 506–515, 2023, doi: 10.33395/remik.v7i1.12102.

[6] F. Findry and Rizal Adi Saputra, "Klasifikasi Kanker Kulit Berdasarkan Data Citra Benign Dan Malignant Menggunakan Convolutional Neural Network," *J. Mhs. Tek. Inform.*, vol. 3, no. 1, pp. 1–9, 2024, doi: 10.35473/jamastika.v3i1.2417.

[7] N. Novianti, M. Zarlis, and P. Sihombing, "Penerapan Algoritma Adaboost Untuk Peningkatan Kinerja Klasifikasi Data Mining Pada Imbalance Dataset Diabetes," *J. Media Inform. Budidarma*, vol. 6, no. 2, p. 1200, 2022, doi: 10.30865/mib.v6i2.4017.

[8] F. D. Marleny, *Pengolahan Citra Digital Menggunakan Python*. Jawa Tengah: CV. Pena Persada, 2021.

[9] A. A. Permana *et al.*, *Machine Learning*. Padang: PT Global Eksekutif Teknologi, 2022.

[10] Amna *et al.*, *Data Mining*. Padang: PT Global Eksekutif Teknologi, 2023.

[11] F. Diba, M. S. Lydia, and P. Sihombing, "Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi," *Indones. J. Comput. Sci.*, vol. 12, no. 4, pp. 2152–2160, 2023, doi: 10.33022/ijcs.v12i4.3329.

[12] M. R. Raihan, Y. H. Chrisnanto, and A. K. Ningsih, "KLASIFIKASI PENENTUAN KELAYAKAN PINJAMAN KOPERASI DENGAN ALGORITMA CART MENGGUNAKAN ALGORITMA ADABOOST," *INFOTECH J.*, vol. 8, no. 2, pp. 74–83, 2022.

[13] M. A. Muslim *et al.*, *Data Mining Algoritma C4.5*. 2019.

[14] E. Widarti *et al.*, *Buku Ajar Pengantar Sistem Informasi*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.

[15] R. A. Sukamto and M. Shalahuddin, "Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek." Informatika, Bandung, 2018.