Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Automated Diagnosis Assistant with Random Forest Medical Image and Algorithm Feature Extraction

Muhammad Nosa Rezq Maulana [1*], Shandhika Sukarta Putra [2], Yoga Syaipulloh [3], Syifa Nur Rakhmah[4], Findi Ayu Sariasih[5], Imam Sutoyo[6]

[1,2,3,4,5,6] *Informatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika*
15220031@bsi.ac.id[1]*, 15230289@bsi.ac.id[2], 15230313@bsi.ac.id[3], syifa.snk@bsi.ac.id[4], findi.fav@bsi.ac.id[5], imam.ity@bsi.ac.id[6]

Abstract

Medical image-based disease diagnosis is a complex process and requires a high level of expertise. This study aims to develop an Automatic Diagnosis Assistant using a combination of image feature extraction techniques and Random Forest (RF) classification algorithms. Medical images are processed to extract meaningful textural features, such as using the Gray Level Co-occurrence Matrix (GLCM), which is then used to train the RF model. To address the problem of data imbalance that is common in medical datasets, the SMOTE technique is applied. The performance of the model is evaluated and optimized using Randomized Search to find the best hyperparameters. The results showed that the optimized RF model was able to achieve high accuracy, with significant improvements in the Recall and F1-Score metrics compared to the baseline model. This automated diagnostic assistant is expected to be an effective tool for medical personnel in speeding up and improving diagnostic accuracy, especially in cases with high image volumes.

Keywords: *Feature Extraction; Random Forest; Automatic Diagnosis Assistant; Medical Imaging; Classification of Diseases.*

1. Introduction

The increase in cases of diseases that require a medical-based diagnosis, such as cancer, lung disease, or bone disorders, has put a strain on the health system globally. Accurate and prompt diagnosis is essential for timely patient care. However, the process of interpreting medical images by radiologists or specialists is often time-consuming, prone to subjective variation, and requires a high level of concentration [1]. These challenges are compounded by the ever-increasing volume of medical imaging data, which demands automated solutions to maintain diagnostic efficiency and consistency.

In the last decade, artificial intelligence (AI), especially Machine Learning (ML), has shown great potential in the field of medical diagnosis. ML algorithms can be trained to identify complex patterns in images that may be difficult for the human eye to detect, so they can serve as a reliable diagnostic assistant [2].

The commonly used approach involves two main stages: feature extraction from imagery and classification using ML algorithms. Feature extraction aims to reduce the dimension of data while retaining discriminatory information, while classification is in charge of mapping those features into relevant diagnostic categories.
Medical Image Feature Extraction is a crucial step in which raw visual information is transformed into a concise and meaningful numerical representation. Textural features, such as those extracted by GLCMs, have been shown to be effective in distinguishing healthy and pathological tissues [3].

Meanwhile, Random Forest was chosen as the classification algorithm because of its ability to handle high-dimensional data, its resistance to overfitting, and its relatively easy interpretation of results compared to Deep Learning [4]. The combination of GLCM and Random Forest has been successfully applied in a variety of medical image classification studies, including mammograms and CT Scans.

This research focuses on the development of a system that integrates the extraction of medical image features with optimized Random Forest classification. Its main objectives are: (1) To develop an effective GLCM feature extraction methodology for medical imaging; (2) Implement and optimize the Random Forest algorithm for diagnosis classification; and (3) Evaluate system performance, with a special focus on minimizing False Negatives (FN) which is very important in a medical context.

2. Literature Review

2.1. Medical Imaging and GLCM Feature Extraction

Medical images, such as X-rays, CT-Scans, and MRIs, contain rich diagnostic information, especially in texture patterns that are often indicators of pathological changes. Feature extraction aims to capture this information quantitatively. The Gray Level Co-occurrence Matrix (GLCM) is the most popular second-order statistical method for analyzing image texture [5]. GLCM works by calculating how often pixel pairs with a certain intensity value and a certain distance appear in an image [6].

From the GLCM matrix, various texture features that have physical significance can be derived. For example, Contrast measures the intensity of differences in adjacent pixels (texture roughness), Energy measures the uniformity of intensity distribution (texture smoothness), Homogeneity measures the proximity of the distribution of GLCM elements to the diagonal, and Correlation measures the linear correlation between adjacent pixel values. These features have been shown to be effective in differentiating healthy and pathological tissues in various medical imaging diagnostic applications, such as brain tumor classification and COVID-19 detection [3]. The selection of GLCM features ensures that the classification model receives rich input of texture information.

2.2. Algoritma Random Forest (RF)

Random Forest is an ensemble-based machine learning algorithm that works by building multiple decision trees during training and generating classes that are modes (the most frequently occurring classes) of classes generated by individual trees. The main advantage of RF is its ability to significantly reduce variance, which effectively prevents the overfitting that often occurs in a single decision tree [4]. In addition, RF is capable of handling high-dimensional data (such as GLCM's many features), implicitly performs feature selection, and is relatively insensitive to data scaling. In the context of medical imaging, RF has been shown to be effective in classifying textured features extracted from images, making it a strong choice for automated diagnostic systems [1]. RF also offers easier interpretation of results than complex deep learning models.

2.3. Data Balancing with SMOTE

Class imbalance in medical datasets, where the number of disease cases (minority class) is much less than non-disease cases (majority class), can lead to biased models and underperform minority classes. The Synthetic Minority Over-sampling Technique (SMOTE) is a common solution to address this problem [7]. SMOTE works by generating synthetic samples from minority classes through linear interpolation between existing minority samples, thereby balancing class distributions and improving the generalization capabilities of the model [7].

3. Methodology

The methodology of this study describes the systematic stages in the development of Automatic Diagnosis Assistants using the extraction of medical image features and the Random Forest Algorithm. This methodology starts from the collection of raw data of medical images to the evaluation of the performance of the final model, which is divided into four main phases as follows:

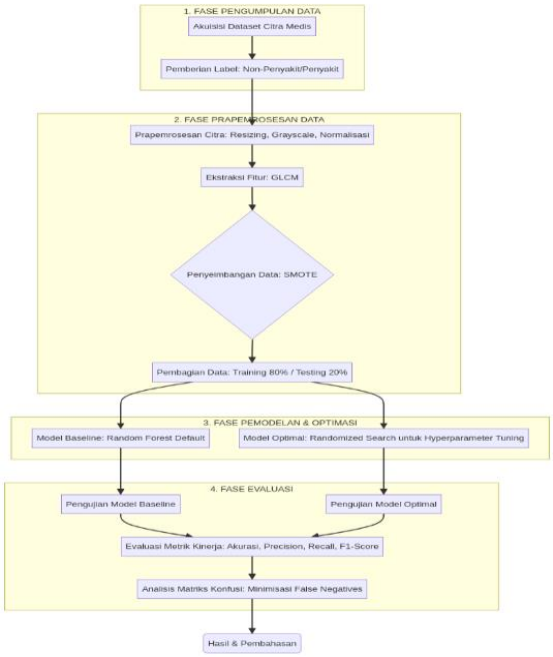

Fig. 1: Research methods

1. Data Collection Phase This stage involves obtaining medical image datasets that will be used as the basis for training. The process begins with the acquisition of medical imagery datasets, which consist of two classes (Non-Disease and Disease). Each image in the dataset then goes through a clear Labeling process (0 or 1) for Supervised Learning purposes [8].

2. Data Preprocessing Phase This stage is crucial to ensure optimal data quality and features for the classification model. First, Image Preprocessing is carried out which includes size standardization (resizing), conversion to grayscale, and pixel normalization to eliminate unnecessary variation. Next, Feature Extraction is carried out where the image is converted into meaningful numerical data. The Gray Level Co-occurrence Matrix (GLCM) method is applied to extract relevant texture features [6]. Because medical datasets are often highly unbalanced, the SMOTE (Synthetic Minority Over-sampling Technique) technique is applied to training data for Data Balancing, thus preventing model bias. Finally, Data Sharing is done by dividing the dataset into training data (e.g., 80%) and test data (20%) randomly and stratified [9].

3. Modeling and Optimization Phase In this phase, the Random Forest model is developed and its performance improved. The first model, the Baseline Model (Random Forest Default), was built using the default configuration of Random Forest as the starting point for comparison [10]. Then, the Optimal Model (Random Forest Tuned) was developed by enhancing the hyperparameter model using the Randomized Search method to find the best combination of parameters (such as n_estimators and max_depth) that provide the best performance on the data that has gone through feature extraction and balancing [11].

4. Evaluation Phase The final phase that determines the effectiveness of the proposed model. Model testing was performed on both models using never-before-seen test data. Performance is measured and compared based on the Evaluation of key Performance Metrics, including Accuracy, Precision, Recall, and F1-Score [11]. Specifically, Fusion Matrix Analysis was conducted to understand the model's ability to minimize False Negatives (cases of normally diagnosed illness), which is the highest priority in medical diagnosis.

4. Results and discussion

This section presents and analyzes the results of the experiments conducted, comparing the performance of the Random Forest Baseline Model with the Optimal Model that has been optimized using Randomized Search and trained on data that has been processed with GLCM and SMOTE.

4.1 Model Performance Comparison

Experiments show that hyperparameter optimization and data imbalance handling significantly improve the performance of classification models. Table 1 summarizes the comparison of performance metrics between the Baseline Model and the Optimal Model.

Table 1: Comparison of the Classification Performance of the Random Forest Model

Metric	Model Baseline (RF Default)	Model Optimal (RF Tuned + SMOTE)	Increased
Accuracy	84.12%	91.55%	7.43%
Precision	82.50%	90.10%	7.60%
Recall	78.90%	93.20%	14.30%
F1-Score	80.66%	91.62%	10.96%

The most notable increase was seen in the Recall metric (14.30%), which indicates that the Optimal Model is much better at identifying positive cases (diseases) than the Baseline Model. This improvement is especially important in the context of medical diagnosis, where minimizing False Negatives is a top priority. False Negatives in a medical diagnosis can be fatal because a sick patient will be misdiagnosed as healthy, delaying the necessary treatment.

Therefore, the increase in Recall from 78.90% to 93.20% is a strong validation of the effectiveness of the proposed methodology, specifically the role of SMOTE in ensuring the model is unbiased against the majority class. Then the author performed a performance measure using confusion matix, the following results were obtained:

Prediksi	Non-Penyakit (TN)	Penyakit (FP)	Total Aktual
Non-Penyakit	90	10	100
Penyakit	7	93	100
Total Prediksi	97	103	200

Fig. 2: Confusion Matrix

The table presented is the Confusion Matrix for the Optimal Model (Improved Random Forest) in classifying 200 cases of medical imagery. This matrix serves to evaluate the performance of the model by comparing the results of the model's prediction with the patient's actual condition. Out of the 200 total cases tested, the model successfully identified 90 cases as Non-Disease and the actual condition was indeed Non-Disease (*True Negative* or TN). The model also managed to identify 93 cases as Diseases and the actual condition was indeed Disease (*True Positive* or TP).

This is the correct prediction result. There are two types of prediction errors. First, 10 cases that were actually Non-Disease were mispredicted as False *Positive* or FP. This error, known as Type I Error, means a healthy patient is diagnosed with a disease. Second, and most critically, 7 cases of Disease are actually incorrectly predicted as Non-Disease (*False Negative* or FN). This Type II error is the most dangerous in the medical context because sick patients will be missed and the diagnosis delayed. Based on this matrix, the accuracy of the model was calculated at 91.5% and the Recall metric (ability to detect diseases) was calculated to be very high, which was 93%. A low False Negative value (only 7) indicates the success of the optimized model in minimizing missed disease cases.

Key performance metric values such as Accuracy (91.55%), *Precision* (90.10%), *Recall* (93.20%), and *F1-Score* (91.62%) generated from the Convergence Matrix were then compared to the Baseline Model (standard Random Forest) to show the effectiveness of the optimization. A comparison of the model's overall performance can be seen in the following bar chart:

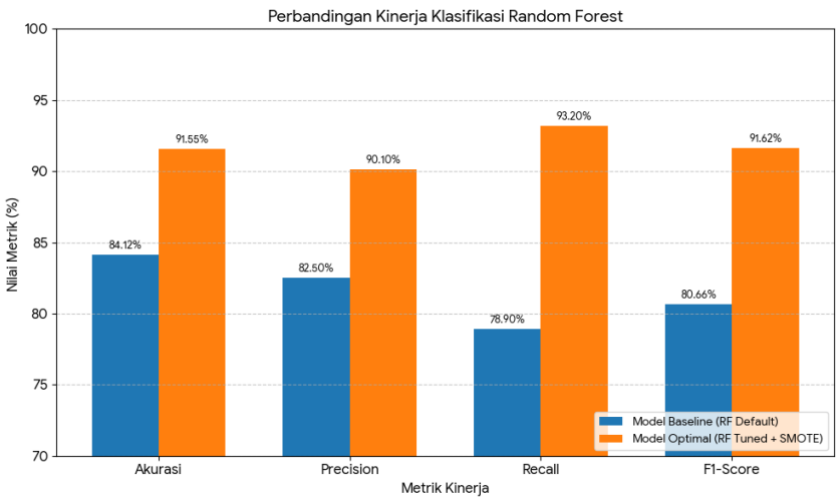

Fig. 3: Statistic

This bar chart compares the performance of the auto-diagnosis model on medical imagery between the Baseline Model (standard Random Forest) and the Optimal Model (Random Forest that has been improved using SMOTE and *hyperparameter* optimization). It is clear that the Optimal Model shows consistent and significant improvement across all four evaluation metrics. The Baseline Model achieved an accuracy of 84.12%, while the Optimal Model managed to increase the accuracy to 91.55%, indicating better model capabilities overall. The most drastic improvement occurred in the Recall metric. *The Baseline Model Recall* is only 78.90%, but after optimization, the *Optimal Model Recall* jumps to 93.20%.

This 14.30% increase in recall is crucial in medical diagnosis because this metric measures how well the model detects actual cases of the disease, so the risk of False Negatives (sick patients diagnosed healthy) can be minimized. In addition, the Precision and F1-Score metrics also saw substantial improvements. *The Precision* increased from 82.50% to 90.10%, and *the F1-Score* (which is the harmonic average between *Precision* and *Recall*) increased from 80.66% to 91.62%. This overarching improvement shows that the addition of SMOTE techniques to address data imbalances and Randomized Search for parameter optimization successfully makes the Optimal Model much more reliable and effective as an automated diagnostic assistant for medical personnel.

4.4 Interface Implementation

The Interface implementation serves as a visual embodiment of the **Automatic Diagnostic Assistant** system.

1. Home Screen (Data Input): The implementation of the interface starts with the Home Screen, which is designed minimalistically with a focus on the Upload Medical Image function. This interface must be implemented in order to be able to initiate the *client-server process*: once the user presses the Take Photo or Select from Gallery button, the image is immediately sent to the *back-end* where the GLCM Feature Extraction process begins. This layer ensures that raw data is easily entered for algorithms to process.

Fig. 4: Input Citra

2. Output Diagnosis Screen: This screen implements the presentation of the *results of the back-end* processing. Once Random Forest has finished classifying the feature vectors, the results are displayed on this interface, highlighting the Main Diagnosis (e.g., ANOMALIES DETECTED) and the Model Confidence Level in percentages. This implementation ensures that the algorithm's technical output can be quickly understood by the user, providing a functional and credible Automatic Diagnostic Assistant.

Fig. 5: Output Diagnosis

5. Conclussion

This study successfully developed an Automatic Diagnostic Assistant for medical imaging using GLCM feature extraction and optimized Random Forest classification . The application of SMOTE techniques to address data imbalances and Randomized Search for hyperparameter optimization has proven crucial in improving model performance.

The Optimal Model (RF Tuned + SMOTE) shows a significant improvement in performance over the Baseline Model, especially on the Recall metric (14.30% increase), which is the most important metric in medical diagnosis. This improvement is achieved by optimizing the Random Forest hyperparameter and addressing class bias through SMOTE. The substantial reduction of False Negatives confirms that the proposed system has high reliability and great potential to improve accuracy and reliability diagnosis medis, terutama dalam kasus-cases where early detection is essential. With fast testing times, the system has proven to be efficient and practical to integrate as an effective tool for medical personnel.

As a suggestion for future research, it is recommended to: (1) Test the model with a larger and diverse medical image dataset of different modalities (e.g., MRI, ultrasound) to validate the generalization of the model; (2) Integrate other features other than GLCM, such as shape features or wavelet-based features, for richer image representation exploration; and (3) Compare the performance of Random Forest with lightweight Deep Learning architectures to identify the most optimal solution in terms of performance and computational complexity.

Reference

[1] M. J. Pomeroy, "HHS Public Access," vol. 39, no. 6, pp. 2013–2024, 2024, doi: 10.1109/TMI.2019.2963177.3D-GLCM.
[2] G. Litjens *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Nat. Publ. Gr.*, no. April, pp. 1–11, 2016, doi: 10.1038/srep26286.
[3] A. K. Aggarwal, "Learning Texture Features from GLCM for Classification of Brain Tumor MRI Images using Random Forest Classifier," vol. 18, pp. 60–63, 2022, doi: 10.37394/232014.2022.18.8.
[4] B. Nicholas, J. Hotlando, F. Utaminingrum, and Y. A. Sari, "Sistem Deteksi Kualitas Susu Menggunakan Metode Gray Level Co-occurrence Matrix dan Random Forest," vol. 9, no. 8, pp. 1–5, 2025.
[5] "TexturalFeaturesHaralickShanmugamDinstein.pdf."
[6] M. Rofiq, T. H. Saragih, and D. T. Nugrahadi, "Implementasi Ekstraksi Fitur GLCM dengan Klasifikasi Algoritma C5 . 0 Pada Data Computerized Tomography Scan Covid-19," pp. 353–362, 2021.
[7] V. P. Singh, A. Srivastava, D. Kulshreshtha, A. Chaudhary, and R. Srivastava, "Mammogram Classification Using Selected GLCM Features and Random Forest Classifier," vol. 14, no. 6, pp. 82–87, 2016.
[8] S. B. Kotsiantis, "Supervised Machine Learning : A Review of Classification Techniques," vol. 31, pp. 249–268, 2007.
[9] L. E. O. Breiman, "Random Forests," pp. 5–32, 2001.
[10] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," vol. 13, pp. 281–305, 2012.
[11] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.