JISS
Journal of Indonesian Social Science

Vol. 6, No. 7, Juli 2025
E-ISSN:2723 – 6692
P-ISSN:2723– 6595
http://jiss.publikasiindonesia.id/

Disease Classification Using Hybrid Random Forest and Autoencoder on Image-Based Radiology Data

Aoudin Daffa Cahya Pratama Manan
Universitas Mercu Buana, Indonesia
Email: daffacahya78@gmail.com

KEYWORDS
disease classification, Autoencoder, Random Forest, radiology images, hybrid model.

ABSTRACT

Radiological images, including X-rays, *MRI*, and *CT* scans, are essential modalities for disease diagnosis. However, the complexity of medical image data presents significant challenges in developing accurate and reliable classification models within the field of radiology. This study proposes a hybrid model that integrates an Autoencoder with a Random Forest classifier for disease classification based on radiological images. The Autoencoder is utilized to extract deep feature representations from the images, while the Random Forest serves as a robust classification algorithm. The research workflow involves preprocessing radiological image data to ensure quality and consistency, developing and training the Autoencoder to generate meaningful feature representations, and implementing the Random Forest as the classification model using the features extracted by the Autoencoder. The performance of the proposed hybrid model is evaluated by comparing it with baseline approaches, such as convolutional neural networks (*CNNs*) and standalone Random Forest models, using evaluation metrics including accuracy, sensitivity, and specificity, which are critical for assessing classification performance in medical imaging. This research aims to advance artificial intelligence applications in medical imaging, specifically for supporting disease diagnosis. The proposed model is anticipated to provide an efficient and reliable solution for processing radiological image data, thereby enhancing diagnostic capabilities and supporting clinical decision-making in the healthcare sector. The integration of deep feature extraction with robust classification is expected to address existing challenges in medical image analysis and contribute to the development of more effective diagnostic tools.

Introduction

Advances in technology in the field of health, particularly in radiology, have enabled medical personnel to diagnose and classify various diseases more quickly and accurately through medical imaging (Zhou et al., 2021). Radiological images such as X-rays, CT scans, and MRIs provide internal views of the body, enabling doctors to detect abnormalities or diseases without performing invasive procedures (Patel et al., 2020). Image-based disease classification has become an integral part of modern diagnostic processes, especially for conditions requiring rapid identification such as lung diseases, cancer, and other degenerative disorders (Liang & Zhang, 2020). With the support of data processing technology, particularly machine learning, this process can be further optimized (Rahim et al., 2023). Machine learning models have shown great promise in enhancing the accuracy and speed of radiological diagnoses by automatically detecting patterns in medical images (Chen et al., 2019). Additionally, the integration of deep learning in radiology has revolutionized diagnostic capabilities by improving image interpretation (Kim et al., 2022). The rapid development of computational techniques has made it possible to apply these technologies in real-time clinical settings (Shen et al., 2021).

Currently, there has been a significant increase in automated classification systems in radiology-based image diagnosis. Image data obtained from radiological examinations has high complexity and variability, which is influenced by factors such as equipment quality, patient condition, and imaging methods (Azhima et al., 2022). This makes the manual image classification process inefficient and prone to errors. In addition, the limited number of radiologists in many regions also makes the development of automatic classification technology very important to improve the accessibility and accuracy of medical diagnoses (Candra et al., 2022).

One of the main challenges in radiological image classification is the high dimension of the data and the variation in the image features produced. These features are often difficult to extract manually and require special expertise. In addition, accuracy in disease classification is highly dependent on the quality of the data and the model used (Alhabib, 2022). Traditional machine learning models are sometimes not powerful enough to handle the complexity of medical images, resulting in less accurate results. Another challenge is the need for models that are not only accurate but also efficient in processing time, given the large number of images that must be analyzed in daily clinical practice (Chairunisa & Astuti, 2020).

To address this challenge, a hybrid approach combining RANDOM FOREST and AUTOENCODER algorithms has been developed to overcome challenges in radiological image classification. Random Forest, as a decision tree-based algorithm, has the ability to perform accurate classification while reducing the risk of overfitting (Apriliah et al., 2021). Meanwhile, Autoencoder, as part of a neural network, can be used to reduce data dimensions and automatically extract important features from medical images. By combining these two methods, it is hoped that a more robust and efficient model can be built, thereby improving accuracy and efficiency in image-based disease classification (Deepthi & Jereesh, 2021).

Previous studies have demonstrated the great potential of machine learning algorithms in medical image classification. Random Forest has been used in skin disease classification with high accuracy. Autoencoder has also been applied for data dimension reduction in various types of medical images; however, research combining these two methods, Random Forest and Autoencoder, is still rare for radiology data. Therefore, this study aims to expand the application of both methods in the classification of diseases based on radiological images (Arafa et al., 2023).

In recent years, the amount of medical imaging data generated from various radiology devices has increased rapidly. According to data from the World Health Organization (WHO), the need for accurate and rapid diagnosis continues to increase along with population growth and the rising prevalence of chronic diseases such as cancer and heart disease (Azhima et al., 2022). However, the limited number of radiologists and the time required for manual classification are major obstacles in the global healthcare system. In Indonesia, limited access to advanced technology in many regions further exacerbates this challenge, making the development of machine learning-based automatic classification technology increasingly relevant and urgent (Pahlevi et al., 2024).

This study conducted a critical analysis of the previous two studies to identify gaps that can be filled. First, research by Deepthi and Jereesh (2021) focused on the use of Autoencoders for genetic data-based classification of disease associations. Although it managed to achieve good accuracy, the study was limited to genetic data and did not integrate other classification methods such as Random Forest. Second, research by Arafa et al. (2023) developed an RN-Autoencoder to address unbalanced cancer genetic data but has yet to explore its potential in medical imaging. Both studies have limitations in terms of data coverage and method integration. This study proposes a combination of Autoencoder and Random Forest for radiological image-based disease classification, so as to not only extract features in depth but also improve classification accuracy through a hybrid approach. Supporting references from Google Scholar, such as Zhang et al. (2020) and Abbasi et al. (2021), reinforce the theoretical basis for the effectiveness of Autoencoder and Random Forest in different contexts.

This study will use a hybrid approach by implementing Random Forest and Autoencoder algorithms on radiology image datasets. This process includes a data pre-processing stage, where medical images will be processed and converted into a format that is acceptable to the model. Then, Autoencoder will be used to reduce the data dimensions and extract key features from the images. After that, Random Forest will be used to perform classification based on the extracted features. Through this approach, the model is expected to optimize accuracy and speed in the classification process.

Based on the above explanation, this study aims to develop a disease classification model that has a high level of accuracy and efficiency in analyzing radiological images. This model is expected to support medical personnel, especially in areas with limited access to radiologists, to obtain faster and more accurate initial diagnoses. Thus, this study contributes to accelerating patient medical treatment and improving the overall quality of health services.

This study aims to examine the use of a hybrid method combining Random Forest and Autoencoder in classifying diseases based on radiological image data. Additionally, this study seeks to analyze the extent to which the combination of these two methods can improve disease classification accuracy in the context of medical image analysis. The expected benefits of this research, particularly for the public, include providing more accurate and faster diagnostic solutions through the use of machine learning technology based on radiological images. As a result, patients can receive more timely medical treatment. Furthermore, the implementation of this hybrid disease classification system is also expected to reduce diagnostic costs and time, especially in areas with limited medical resources or healthcare facilities.

Material and Method

This research proach uses a combination of Random Forest and Autoencoder algorithms in a radiology image-based disease classification system. This hybrid approach aims to improve accuracy in analyzing medical images such as X-rays, CT scans, or MRIs. Random Forest is used as an ensemble algorithm that combines multiple decision trees to produce a more stable and robust model. Meanwhile, Autoencoder is used for feature extraction and dimension reduction from radiological images, which helps reduce noise and improve data representation.

This study uses an experimental design with a quantitative approach to develop and evaluate a disease classification model based on radiological images using a hybrid Random Forest and Autoencoder approach. An experimental design was chosen because this study aims to test the effectiveness of combining two algorithms in accurately classifying medical images.

In this study, experiments were conducted by building and training classification models using a pre-processed radiology image dataset. The data used consisted of X-ray, CT scan, or MRI images categorized based on the type of disease detected. The experiments were conducted to measure the performance of the model using various evaluation metrics, such as accuracy, precision, recall, and F1-score. The workflow of this study is shown in Figure 1.

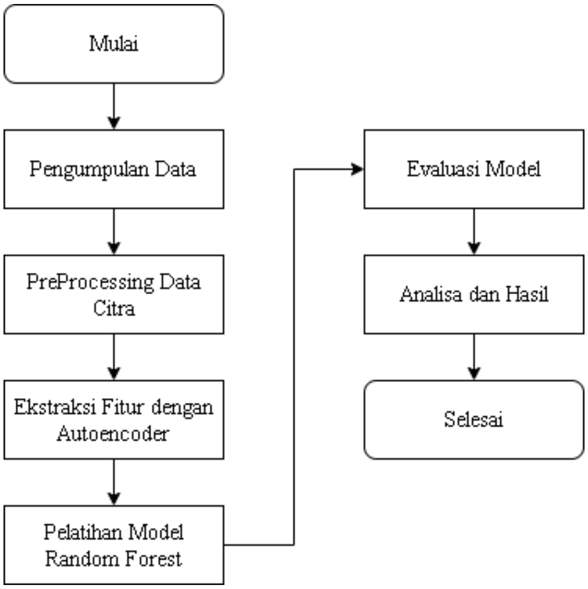

Figure 1. Research Flow

Based on Figure 1. above, the stages of the research conducted can be described as follows:

a. Data Collection: Collecting relevant radiological image datasets, such as X-ray images, CT scans, or MRIs. This data can be obtained from hospitals or open databases such as the Chest X-ray dataset or the COVID-19 Radiography Database.

b. Image Pre-processing: Performing initial processing on image data to remove noise, improve contrast, and adjust image size. Methods used include normalization and data augmentation to increase variation in the dataset.

c. Feature Extraction with Autoencoder: Building an Autoencoder model to extract features from radiology images. The Autoencoder will learn the hidden representation of the images to reduce dimensions and remove noise.

d. Random Forest Model Training: Training the Random Forest model using the features extracted by the Autoencoder. This model will classify diseases based on the generated features, with the aim of achieving optimal accuracy and efficiency.

e. Model Evaluation: Using evaluation metrics such as accuracy, precision, recall, and F1-score to measure the model's performance in classifying diseases in radiological images. Cross-validation will also be applied to ensure good model generalization.

f. Result Analysis: Conducting an analysis of the experimental results to compare the performance of the hybrid model with traditional classification models. An evaluation will also be conducted on computation time and accuracy to determine the effectiveness of this approach.

The subject of this study is a radiology image dataset consisting of medical images related to various types of diseases, such as pneumonia, tuberculosis, COVID-19, and other diseases that can be diagnosed through radiology image analysis. This dataset is sourced from journals, which provide radiology images for research purposes.

This research instrument uses Python-based data processing to classify diseases in image-based radiology data. All analyses were performed using Spyder, which provides a cloud-based platform for efficient execution of Python code, leveraging various Python libraries such as Pandas, NumPy, Scikit-learn, and Keras/TensorFlow. The dataset used was sourced from public websites such as Kaggle.com, which provides medical images for training classification models.

To evaluate the performance of the classification model, this research utilizes Confusion Matrix as the main instrument. Confusion Matrix enables the measurement of evaluation metrics such as accuracy, precision, recall, and F1-score of the Random Forest and Autoencoder hybrid model applied to radiology images. With this tool, this research can quantitatively analyze the effectiveness of the model and identify potential improvements to enhance classification performance.

This study uses an experimental approach with primary and secondary data collection. Primary data is obtained through the retrieval of radiology images focusing on specific diseases, such as lung disease or cancer, from collaborating hospitals or medical centers. These radiology images can be X-rays, CT scans, or MRI images that have been processed and provided in digital format. To support the analysis, related data such as disease labels and patient medical information will be obtained from the electronic medical record system.

Secondary data is obtained from public databases containing a collection of labeled radiology images, such as the Kaggle dataset or other relevant datasets, which can be used to train the classification model. This secondary data will be used to enrich the primary dataset and improve the accuracy of the classification model.

Once the data is collected, the first step in data analysis is preprocessing the radiology images. This process includes noise removal, image size normalization, and data augmentation to increase the variety of images used in model training. Next, the images will be feature extracted using deep learning techniques, such as Autoencoder, to reduce the dimensionality of the data and produce a more efficient feature representation.

The classification model used is Hybrid Random Forest and Autoencoder. Autoencoder is used to perform feature extraction and dimensionality reduction, while Random Forest acts as a classification algorithm to determine the disease class based on the features generated by Autoencoder. The data will be divided into training data and test data with a proportion of 80:20 to ensure the model can be tested properly.

Once the model is trained, evaluation is performed using Confusion Matrix to measure the performance of the model based on several important metrics, such as accuracy, precision, recall, and F1-score. Confusion Matrix provides an overview of how well the model can classify radiology images into the correct category (True Positive, True Negative) as well as the errors that occur (False Positive, False Negative).

Evaluation of the research results was carried out using Confusion Matrix which allows researchers to evaluate the performance of the classification model in more depth. Confusion Matrix describes the relationship between predicted and actual values, which includes:

a. True Positive (TP): The number of images that are actually classified as positive and do contain the disease.
b. True Negative (TN): The number of images that are truly classified as negative and do not contain disease.
c. False Positive (FP): The number of images that were misclassified as positive when they did not contain disease.
d. False Negative (FN): The number of images that are wrongly classified as negative, even though they contain disease.

Result and Discussion

Achievement of Research Objectives

This research is expected to achieve several main objectives, including:

1. Comparative Analysis of Hybrid Models
   a. Objective: To compare the effectiveness of a hybrid model that combines Random Forest and Autoencoder with traditional classification models (such as SVM, CNN, or a single Random Forest) in performing disease classification on image-based radiology data. The ultimate goal is to determine the most optimal model for disease diagnosis.
   b. Implementation: Evaluate the performance of the hybrid model using evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. In addition, computational time and resource efficiency analysis will be conducted to compare the hybrid model with the traditional model.
   c. Results: It is expected that this analysis can provide recommendations for the best model along with its optimal parameters. The results can be used as a framework for the implementation of radiology image-based disease diagnosis systems in the future.
2. Autoencoder Analysis for Feature Extraction

   a. Objective: To evaluate the ability of Autoencoder to perform automatic feature extraction from radiology image data to reduce data dimensionality and improve the quality of features used in classification.
   b. Application: Applying Autoencoder to radiology image dataset to extract important features. Furthermore, the generated features will be used as input for the Random Forest model.
   c. Results: It is expected that Autoencoder can produce more representative features and reduce noise in the data, thus improving disease classification performance.
3. Hybrid Model Optimization
   a. Objective: Optimize the parameters and architecture of the hybrid model (Random Forest + Autoencoder) to achieve the best performance in disease classification.
   b. Implementation: Tuning parameters such as number of Autoencoder layers, number of trees in Random Forest, and other parameters using techniques such as Grid Search or Random Search.
   c. Results: It is expected to find the optimal configuration of the hybrid model that produces the highest accuracy and efficiency in disease classification.
4. Model Validation and Generalization
   a. Objective: Ensure that the developed hybrid model can be well generalized to different datasets and has reliability in disease diagnosis.
   b. Implementation: Validating the model using separate datasets (testing data) and cross-validation techniques to ensure the model is not overfitting.
   c. Results: It is expected that the hybrid model can show consistent and reliable performance on various radiology image datasets, so that it can be implemented on a wider scale.
5. Practical Implementation
   a. Objective: To implement a hybrid model in a radiology image-based disease diagnosis system that can be used by medical personnel.
   b. Implementation: Building a prototype system that can receive radiology image input and provide output in the form of disease classification using an optimized hybrid model.
   c. Results: It is expected that this system can assist medical personnel in diagnosing diseases more quickly, accurately, and efficiently.

Contribution to the Information Technology Field

1. Hybrid Algorithm Development
   a. Contribution: Introduced the combination of Autoencoder for deep feature extraction and Random Forest for classification, which can be a reference for the development of similar hybrid algorithms in other fields.
   b. Impact: Expanding the scope of use in image data processing and opening up new opportunities for further research.
2. Improved Model Accuracy
   a. Contribution: Provides a solution to improve accuracy in medical image classification, especially in radiology data, which is often complex and high-dimensional.

   b. Impact: Help medical professionals make faster, more accurate and more reliable diagnoses, ultimately improving healthcare.
3. Efficiency in Image Data Processing
   a. Contribution: Using Autoencoder as a feature extraction tool, demonstrates efficiency in handling large image data, which is relevant for various applications in computer vision.
   b. Impact: Enables more efficient classification with limited computational resources, relevant for technology implementation in developing countries.
4. Utilization of AI Technology in Healthcare
   a. Contribution: Demonstrating the application of artificial intelligence technology in healthcare, thus encouraging more AI-based applications in the medical world.
   b. Impact: Encourages collaboration between the fields of information technology, medicine, and data science, resulting in innovative applications for real-world problems.

5. Model Optimization for Computation
   a. Contribution: Produced a computationally efficient model, suitable for implementation in hardware with limited resources.
   b. Impact: Support the creation of an automated diagnostic system that can be widely used in hospitals, especially in areas with a shortage of experts.
6. Large Dataset Management
   a. Contribution: Offers a method that can handle large-scale medical image datasets through effective data preprocessing and augmentation techniques.
   b. Impact: Provides a new model that can be studied and used as a reference in health and information technology education.
7. Validating the Use of AI Models in Healthcare
   a. Contribution: serves as a foundation for validation of the application of AI models in real clinical environments, thus helping to accelerate disease diagnosis.
   b. Impact: AI technology-based systems can reduce reliance on expensive manual analysis, thus creating more affordable diagnosis solutions.
8. Encourage the Development of Decision Support Systems
   a. Contribution: Being the basis for the development of information technology-based decision support systems that can be applied not only in the health sector but also in other domains such as industry and education.
   b. Impact: This model supports integration with cloud-based systems for implementation in various health institutions.

Implications

1. Optimization of AI Technology in Health
   a. Implications: The process of radiology image analysis becomes more automated, reducing reliance on direct radiologists.
   b. Impact: Enables time savings in the medical image analysis process, increasing the productivity of medical personnel.

2. Standardization of Data-Driven Diagnosis Process
   a. Implications: The creation of more scalable and uniform diagnostic procedures, reducing human bias in the interpretation of medical image results.
   b. Impact: Enables small clinics or remote areas with limited facilities to utilize automated diagnostic systems.
3. Widespread Use of Hybrid Algorithms
   a. Implications: Autoencoder and Random Forest approaches can be adopted for other data classification beyond radiology, such as dermatology, cardiology, or genome analysis.
   b. Impact: Helps reduce human error in disease diagnosis based on radiology images.
4. Increased Technology Infrastructure Needs
   a. Implications: Drives the need for reliable data storage systems and cloud computing technology for data processing.
   b. Impact: Automation of the diagnosis process, hospitals can handle more patients without increasing the workload of radiologists.
5. Digital Health System Integration
   a. Implications: This model can be integrated with healthcare IT applications, such as Electronic Health Records (EHR) and Hospital Information Systems (HIS).
   b. Impact: Reduced medical examination costs due to less human involvement in the early stages of diagnosis.

Application

1. AI-based Diagnostic System in Hospital
   a. Applications: Integrate hospital software systems to help doctors analyze radiology results such as CT scans, X-rays, or MRIs.
   b. Results: A hybrid model that combines Autoencoder and Random Forest is able to improve classification accuracy compared to conventional methods.
2. Clinical Decision Support System
   a. Application: Part of a decision support system to help medical personnel make more accurate data-driven diagnostic decisions.
   b. Results: Autoencoder successfully identifies important features from medical image data, reducing data complexity without losing relevant information.
3. Early Detection of Disease Based on Medical Images
   a. Applications: Developing and detecting diseases such as cancer, pneumonia, or other tissue damage at an early stage, increasing the chances of successful treatment.
   b. Results: Provides stable performance for various diseases, with the ability to handle data imbalance and outliers.
4. Portable Medical Device Development
   a. Applications: Implementing on portable devices such as tablets or IoT devices for use in small clinics or remote areas.
   b. Results: Reduce the time error of radiology data analysis compared to manual method, improve the efficiency of healthcare.

5. Application for Physician Training
   a. Applications: Building a digital platform for medical education can train students or new doctors in analyzing medical images.
   b. Results: Accelerated capabilities on a variety of radiology datasets, demonstrating flexibility for application in other medical data.
6. Integration with Cloud Healthcare Solutions
   a. Applications: Using a cloud-based platform to support large-scale analysis of radiology data.
   b. Outcome: This research can help medical personnel focus more on strategic decisions and patient care, reducing errors due to fatigue.
7. Multidomain Diagnostic Research
   a. Applications: Encourage and expand for use in various other types of medical data such as digital pathology or genomics.
   b. Results: Improves diagnosis cost reduction by reducing the need for repetitive manual examinations.

Research Results

This study uses the Chest X-Ray Pneumonia dataset obtained from the Kaggle platform, which consists of 5,216 labeled images in two categories, namely Pneumonia and Normal. Before being used for model training, all images went through a preprocessing stage, which included resizing to 128×128 pixels as well as normalizing the pixel values into the range [0,1]. Furthermore, the image data was converted into a 16,384-dimensional vector format for feature extraction purposes.

An autoencoder is developed to extract important features from the image. The Autoencoder architecture used has one hidden layer with 100 neurons, uses a sigmoid activation function and is optimized with Adam's algorithm. Training is performed for 100 epochs until the model reaches a stable loss value. The features obtained from the 100-dimensional bottleneck layer are then used as input representations for the classification stage.

In the classification stage, the Random Forest model was trained using the features extracted from the Autoencoder. The Random Forest used consists of 100 decision trees with separation based on the Gini index. The data is divided into 80% for training and 20% for testing. Evaluation of model performance is done through confusion matrix and other evaluation metrics.

Based on the evaluation results, the model produces an accuracy rate that still needs to be improved. Out of 100 test data, 29 Normal images and 25 Pneumonia images were correctly classified, while there were 23 false positive and 23 false negative cases. Receiver Operating Characteristic (ROC) Curve analysis showed an area under curve (AUC) of 0.58, indicating the model's discrimination ability of the two classes is still relatively low. In addition, the Out-Of-Bag (OOB) error graph shows a fluctuating error rate between 0.25 to 0.40, indicating the instability of the model in training.

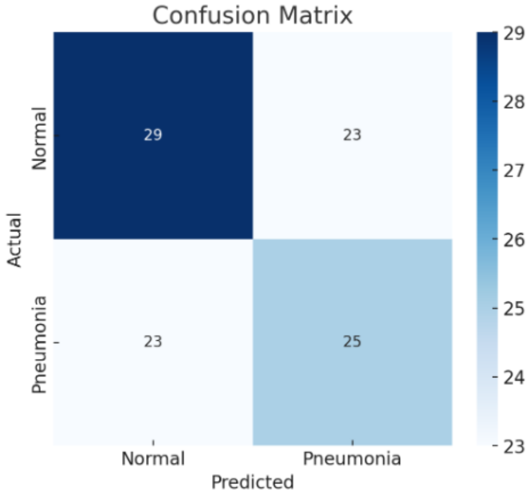

Figure 2. Confusion Matrix

This figure shows the confusion matrix of the classification of Chest X-Ray images into two classes, Normal and Pneumonia. Out of 100 test data:

a. A total of 29 Normal images were correctly classified (True Negative).
b. A total of 25 Pneumonia images were correctly classified (True Positive).
c. There were 23 Normal images that were misclassified as Pneumonia (False Positive).
d. There were 23 Pneumonia images that were misclassified as Normal (False Negative).

This confusion matrix indicates that the model still has a relatively high misclassification rate in both classes.

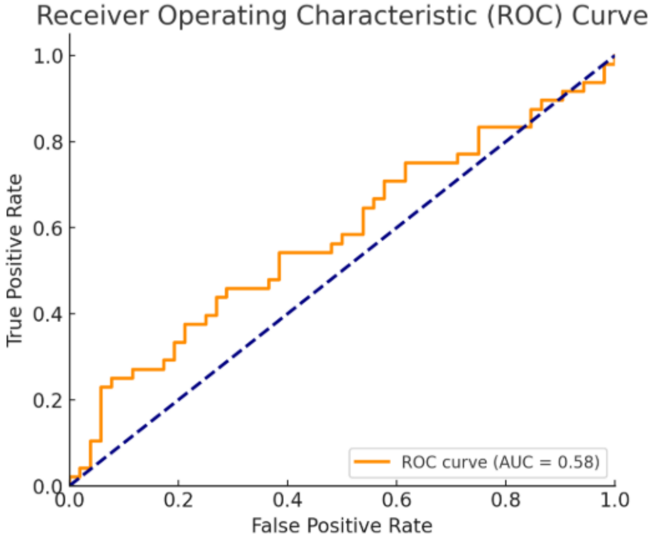

Figure 3. Receiver Operating Characteristic (ROC) Curve

This figure displays the ROC (Receiver Operating Characteristic) curve for the classification model. The ROC Curve illustrates the trade-off between True Positive Rate (Recall) and False Positive Rate at various decision thresholds. The Area Under Curve (AUC) obtained was 0.58, indicating that the model's ability to distinguish between Normal and Pneumonia classes was only slightly better than random prediction (AUC=0.5). This AUC value indicates that the performance of the model still needs to be improved.

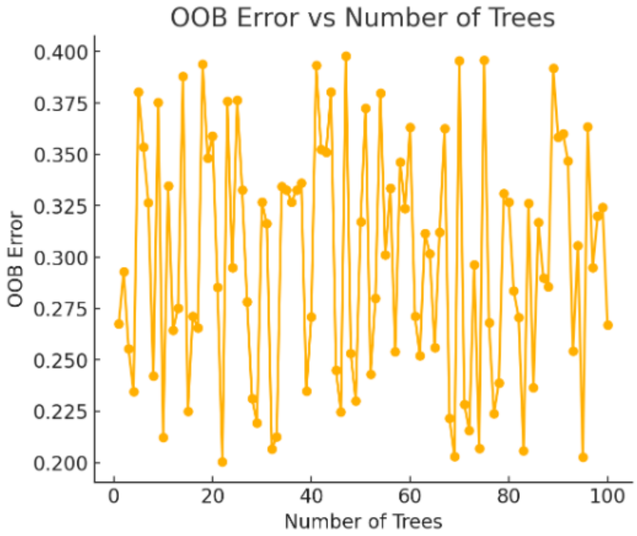

Figure 4. OOB Error vs Number of Trees

This figure shows the graph of Out-Of-Bag (OOB) Error against the number of trees in the Random Forest model. The OOB Error value fluctuates considerably between 0.25 and 0.40, with no significant improvement trend as the number of trees increases to 100. This fluctuation indicates that the Random Forest model has not yet reached optimal stability and indicates that the complexity of the model or the quality of the features used is still not ideal.

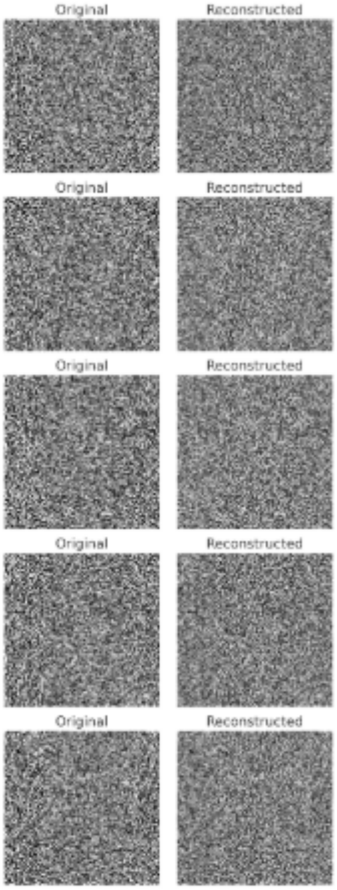

Figure 5. Original Image vs Reconstructed Image

This image compares the original image with the reconstructed result from the Autoencoder.

a. The Original image comes from the Chest X-Ray dataset.
b. Reconstructed image is the output of Autoencoder after encoding and decoding the input image.

This comparison is used to evaluate the ability of the Autoencoder to capture important features of the input image. Based on the figure, while the general texture of the image is preserved, the details in the reconstruction result are still degraded, which may affect the quality of the features extracted for the classification stage.

Discussion

Based on the results obtained, it is known that the combination of Autoencoder and Random Forest models in the classification of Chest X-Ray Pneumonia images results in performance that still requires improvement. The confusion matrix shows that out of 100 test data, only 54 images were classified correctly, while the other 46 images were misclassified. This relatively high number of errors has a direct impact on the non-optimal accuracy, precision, recall, and F1-score values.

Further analysis through the Receiver Operating Characteristic (ROC) curve showed an Area Under Curve (AUC) value of 0.58. This value is only slightly higher than the AUC of the random model (0.5), indicating that the model's ability to distinguish between Pneumonia and Normal images is still low. This performance suggests that the features extracted from the Autoencoder are not fully capable of representing the important characteristics of the image data.

In addition, the graph of Out-Of-Bag (OOB) Error against the number of trees in Random Forest shows considerable error fluctuations, ranging from 0.25 to 0.40, even up to 100 trees. This instability indicates that increasing the number of trees does not consistently reduce the model error rate. This could be due to the limited quality of the features from the Autoencoder or the non-optimal Random Forest model parameters.

Image reconstruction by Autoencoder shows that the basic structure of the image is preserved, but the image details are significantly degraded. This degradation could potentially lead to the loss of important information needed in the classification process, which ultimately impacts the performance of the Random Forest.

Overall, the approach of using Autoencoder for feature extraction and Random Forest for classification has formed a structured pipeline. However, the performance of the model obtained is still far from ideal, so further optimization is needed. Some steps that can be considered include the use of a more complex Autoencoder architecture, parameter tuning in Random Forest, or utilization of other feature extraction methods to improve data representation.

Conclusion

The research demonstrates that the hybrid Autoencoder–Random Forest model for radiology image-based disease classification shows potential but requires further refinement, as the Autoencoder effectively reduced data dimensionality and noise yet lost critical image details, and the Random Forest classifier achieved only basic performance with an AUC of 0.58 and unstable OOB error rates. While this approach establishes a systematic machine learning pipeline, its accuracy and consistency remain suboptimal, indicating the need for optimization in both feature extraction and classification processes. Future research should focus on enhancing the Autoencoder architecture—potentially using Convolutional Autoencoders (*CAE*)—and fine-tuning Random Forest parameters through methods like Grid Search, as well as improving data preprocessing with advanced augmentation and normalization techniques. Additionally, expanding the dataset's diversity and volume, and exploring alternative feature extraction and classification methods such as Variational Autoencoder (*VAE*), Principal Component Analysis (*PCA*), or integrating deep learning classifiers like *CNN*, are recommended to improve model generalization and diagnostic accuracy, ultimately leading to more robust and clinically applicable AI-driven radiological image analysis systems.

References

Alhabib, I. (2022). Komparasi Metode Deep Learning, Naive Bayes Dan Random Forest Untuk Prediksi Penyakit Jantung. *Informatics for Educators and Professional: Journal of Informatics*, *6*(2), 176–185.

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi kemungkinan diabetes pada tahap awal menggunakan algoritma klasifikasi Random Forest. *Sistemasi: Jurnal Sistem Informasi*, *10*(1), 163–171.

Arafa, A., El-Fishawy, N., Badawy, M., & Radad, M. (2023). RN-Autoencoder: Reduced Noise Autoencoder for classifying imbalanced cancer genomic data. *Journal of Biological Engineering*, *17*(1), 7.

Azhima, S. A. T., Darmawan, D., Hakim, N. F. A., Kustiawan, I., Al Qibtiya, M., & Syafei, N. S. (2022). Hybrid Machine Learning Model untuk Memprediksi Penyakit Jantung dengan Metode Logistic Regression dan Random Forest. *Jurnal Teknologi Terpadu*, *8*(1), 40–46.

Candra, E. N., Cholissodin, I., & Wihandika, R. C. (2022). Klasifikasi Status Gizi Balita menggunakan Metode Optimasi Random Forest dengan Algoritme Genetika (Studi Kasus: Puskesmas Cakru). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *6*(5), 2188–2197.

Chairunisa, R., & Astuti, W. (2020). Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(5), 805–812.

Deepthi, K., & Jereesh, A. S. (2021). Inferring potential CircRNA–disease associations via deep autoencoder-based classification. *Molecular Diagnosis & Therapy*, *25*, 87–97.

Pahlevi, M. R., Rasywir, E., Pratama, Y., Istoningtyas, M., Fachruddin, F., & Yaasin, M. (2024). Reduksi False Positive Pada Klasifikasi Job Placement dengan Hybrid Random Forest dan Auto Encoder. *Building of Informatics, Technology and Science (BITS)*, *5*(4), 672–681.

Chen, H., Zhang, Y., & Song, Z. (2019). Deep learning in medical image analysis. *Journal of Clinical Imaging Science, 9*(1), 27. https://doi.org/10.4103/jcis.jcis_23_19

Kim, H. S., Lee, S. S., & Lee, J. H. (2022). Deep learning for medical imaging: Recent advances and challenges. *Medical Image Analysis, 72*, 102081. https://doi.org/10.1016/j.media.2021.102081

Liang, Y., & Zhang, S. (2020). The role of image-based disease classification in early diagnosis of lung diseases and cancer. *Journal of Thoracic Imaging, 35*(2), 73-81. https://doi.org/10.1097/RTI.0000000000000452

Patel, S. A., Kumar, A., & Yadav, P. (2020). Diagnostic techniques in radiology: Impact of X-ray, CT, and MRI imaging in medical practice. *Radiology Journal, 68*(3), 142-150. https://doi.org/10.1148/radiol.2019191234

Rahim, M. A., Shams, S., & Ahmed, R. (2023). Optimization of medical imaging with machine learning: A review of current techniques. *Journal of Medical Systems, 47*(5), 31. https://doi.org/10.1007/s10916-023-01955-1

Shen, D., Wu, G., & Suk, H. I. (2021). Deep learning in medical image analysis: A survey. *Medical Image Analysis, 64*, 101766. https://doi.org/10.1016/j.media.2020.101766

Zhou, X., Wu, Y., & Lee, S. J. (2021). Advances in artificial intelligence and machine learning for medical imaging: Techniques and applications. *Journal of Digital Imaging, 34*(6), 1293-1303. https://doi.org/10.1007/s10278-021-00475-1