Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Image Augmentation for BreaKHis Medical Data using Convolutional Neural Networks

Maie Istighosah[1)*], **Andi Sunyoto**[2)], **Tonny Hidayat**[3)]
[1,2,3)]University Amikom Yogyakarta, Indonesia
[1)]maie.is@students.amikom.ac.id, [2)]andi@amikom.ac.id, [3)]tonny@amikom.ac.id

Submitted : Aug 14, 2023 | **Accepted** : Aug 31, 2023 | **Published** : Oct 1, 2023

Abstract: In applying Convolutional Neural Network (CNN) to computer vision tasks in the medical domain, it is necessary to have sufficient datasets to train models with high accuracy and good general ability in identifying important patterns in medical data. This overfitting is exacerbated by data imbalances, where some classes may have a smaller sample size than others, leading to biased predictive results. The purpose of this augmentation is to create variation in the training data, which in turn can help reduce overfitting and increase the ability of the model to generalize. Therefore, comparing augmentation techniques becomes essential to assess and understand the relative effectiveness of each method in addressing the challenges of overfitting and data imbalance in the medical domain. In the context of the research described, namely a comparative analysis of augmentation performance on CNN models using the ResNet101 architecture, a comparison of augmentation techniques such as Image Generator, SMOTE, and ADASYN provides insight into which technique is most suitable for improving model performance on limited medical data. By comparing these techniques' accuracy, recall, and overall performance results, research can identify the most effective and relevant techniques in addressing the challenges of complex medical datasets. This provides a valuable guide for developing better CNN models in the future and may encourage further research in developing more innovative augmentation methods suitable for the medical domain.

Keywords: Augmentation; Convolutional Neural Network; Image Generator; SMOTE; ADASYN

INTRODUCTION

The importance of data augmentation in computer vision, especially in the medical field, is shown by applying this technique to increase the amount and variety of training data (Saber et al., 2021). In data enhancement, new variations are created by modifying or transforming the original data, such as moving, rotating, cropping, scaling, mirroring, and turning colors. This set of convolutional transformations creates a much lower-dimensional and more useful image than can be created manually (Shorten & Khoshgoftaar, 2019). This advantage of refinement is particularly relevant in the context of training deep learning models, which often require large amounts of data. This technique is attractive due to its simplicity and low cost, which can practically increase the model's performance. In addition to improving accuracy, enhancement also plays an important role in reducing bias in deep learning-based classification models (Tiwari, Dixit, & Gupta, 2021).. Especially with the prevalence of complex neural networks where training requires a lot of data, image enhancement is becoming increasingly important. During the development of predictive models for biomedical imaging, where labeled data is challenging to obtain, image enhancement is of great value.

In practice, however, medical imaging is often limited in the amount and variety of data (Saber et al., 2021)(Shorten & Khoshgoftaar, 2019)(Zhao, Balakrishnan, Durand, Guttag, & Dalca, 2019). Asymmetric data distribution and class imbalance can cause problems in the model's ability to generalize effectively (Chowdhury et al., 2022)(Bloice, Roth, & Holzinger, 2019). This is often caused

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

by the learning process, which is more focused on the majority class and is under-represented by minority groups, resulting in unequal results and less-than-optimal performance. Therefore, overcoming the problem of data imbalance and limitations is an essential goal in developing a reliable model.

In the context of this research, where refinements will be applied to the complex and unbalanced BreaKHis data, there are real problems regarding data availability and diversity (Yang et al., 2022). Data enhancement techniques such as Image Generator, SMOTE, and ADASYN are used to overcome this challenge. However, the relative effectiveness of each method is unclear, which is why comparisons are necessary. By combining these techniques with the ResNet101 architecture and equalization techniques, this study aims to determine which enhancement techniques are most effective in overcoming the problem of data limitations and imbalances in the Complex BreakKHis dataset. The results of this study will help determine the most appropriate technique to improve model accuracy and performance on complex medical data sets and ensure good generalizability.

LITERATURE REVIEW

In computer vision, several deep learning-based data augmentation techniques have been explored to increase the quantity of data. In particular, (Saber et al., 2021)(Tripathi, Khatri, & Greunen, 2022) applied data augmentation to add variety to MIAS data. Data augmentation was proven to increase data variation and reduce overfitting and accuracy. Furthermore, (Chen, Chang, & Guo, 2021) applied SMOTE to the DEAP dataset. SMOTE improved the model performance by 50%(Kummer, Ruppert, Medvegy, & Abonyi, 2022) and performed well (Apostolopoulos, 2020) in some instances, such as Cardiovascular Disease classification. SMOTE is not only easy to use but also effective (Dablain, Krawczyk, & Chawla, 2022). The SMOTE algorithm uses the closest sign of the data samples in the minority class to construct synthetic data from the minority class. We can see that this algorithm excels at detecting emotions. To solve the Alzheimer's disease data problem, the study applied the ADASYN technique to solve the data imbalance problem. ADASYN's performance uses weighted distributions for minor class examples with varying learning difficulties. As a result, the ADASYN technique reduces the bias caused by class imbalance and improves distribution learning by adaptively changing the classification decision limits for complex samples. Although the application of SMOTE and ADASYN oversampling can reduce the bias towards the majority class (Dablain et al., 2022)(Chan, Kelly, & Schnabel, 2021), it is possible that the images will add noise to the minority class and overfitting if the number of synthetic samples generated is too large. In synthetic data, the resulting image is not the same as the original image and can be seen as dots around the distribution (M. Xu, Yoon, Fuentes, & Park, 2023)(Y. Xu & Wali, 2023). This may affect the quality of the synthetic data generated and thus affect the generalization ability of the model.

In this research, ResNet101 will be used. ResNet101's strengths lie in its ability to solve deep training problems (Wu, Xin, Fang, Hu, & Hu, 2019), overcome missing gradient problems (Montaha et al., 2021), generalize complex data sets (Fazari et al., 2021) (Montaha et al., 2021) (Mahmud & Abdelgawad, 2023), and have substantial transfer learning potential (Uzen, Turkoglu, & Hanbay, 2021). This combination of features makes ResNet101 an ideal choice for various research and computer vision tasks in many fields, including the medical field.

Our research will build upon these insights to address the challenges posed by the limited and imbalanced medical image dataset, specifically the BreaKHis dataset. The ResNet101 architecture will serve as our base model. To tackle these challenges, we will combine augmentative techniques, namely the Image Generator, SMOTE, and ADASYN. These techniques have shown promise in enhancing model performance and mitigating the effects of class imbalance. We will evaluate and compare the effectiveness of each method in improving model accuracy and its ability to generalize on the complex and imbalanced BreaKHis dataset. By incorporating these techniques, our study aims to provide a comprehensive analysis of how different augmentation strategies influence model performance, offering valuable insights into tackling the challenges of limited and imbalanced medical image data in the context of computer-aided diagnosis.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12878 p-ISSN : 2541-044X

METHOD

The framework of this study is divided into four steps: Identify the breast cancer histopathology dataset, preprocess the data, train and validate the ResNet101 algorithm, then test and evaluate the model as seen Figure 1. This study separated the dataset into three parts, namely training data, validation data, and testing data, and was carried out proportionally to ensure accurate and consistent model evaluation. This grouping is very important in objectively trying the performance of the model. The BreaKHis dataset has been divided according to the ratio of 80:10:10, where 80% of the data is used as training data, 10% as validation data, and 10% as testing data. Using training data helps the model learn and adjust its internal parameters. Validation data is used to optimize hyperparameters and prevent overfitting during training. Meanwhile, the test data measures the final model's performance that has been trained on data that has never been seen before, thus providing a more accurate picture of the model's ability to generalize to new data. By separating the dataset according to this ratio, this research can provide more reliable and meaningful results regarding the performance and effectiveness of the augmentation techniques applied to the CNN model.

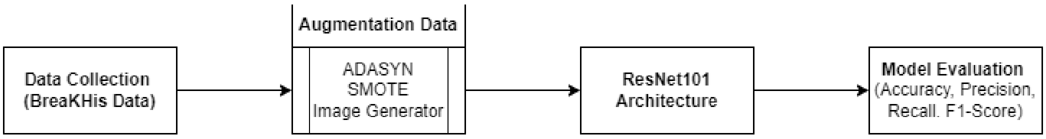

Fig. 1. Research Flow

Data Collection

Data collection is carried out to obtain as much information as is necessary for the research activity. The dataset for this study included medical images of breast cancer histopathology from previous research (Spanhol, Oliveira, Petitjean, & Heutte, 2016). Histopathological medical images are generated from the cytology of the biopsy results, which are then stained with hematoxylin and eosin (H&E). Hematoxylin and eosin staining is commonly used to study cancer, with tissues stained pink and cell nuclei blue.Until now, histopathological procedures still required the pathologist to examine with a microscopy by hand. BreakKHis medical imaging has limited data. Data set limitations may limit the generalizability of existing models (Chowdhury et al., 2022). The BreaKHis data themselves show differences in contrast and color (Elmannai, Hamdi, & Algarni, 2021). D Due to different histopathological data collection procedures, this causes noise and blur in medical images (Rashmi, Prasad, & Udupa, 2022). Figure 2 shows an example of the BreaKHis dataset.

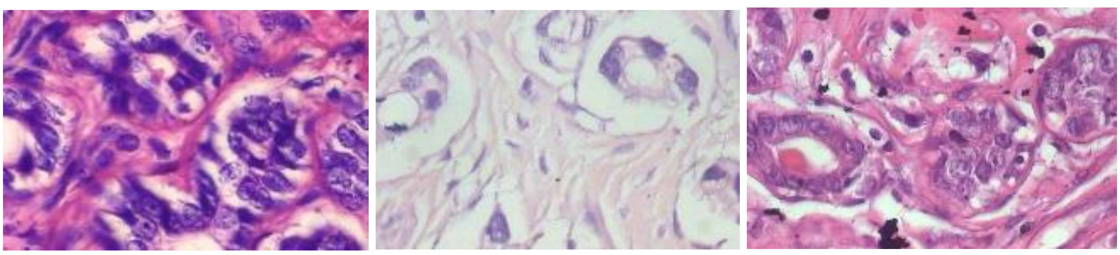

Fig. 2. Sample Dataset BreaKHis

The BreakKHis dataset is divided into two classes: benign and malignant. Benign tumors are lesions that do not meet the histological criteria for malignancy. In general, benign tumors are largely harmless, slow-growing, and limited. Cancer is synonymous with malignancy. Cancerous lesions have the ability to invade and destroy adjacent structures (locally invasive), spread to distant regions (metastasize), and eventually destroy. Breast Cancer Histopathology Imaging (BreaKHis) is a collection of 9,109 microscopic images of breast cancer tissue with different discriminations (40x, 100x, 200x, and 400x) collected from 82 patients. To date, the database contains 2480 benign samples and 5429 malignant samples (700 x 460 pixels, 3 RGB channels, 8-bit depth per channel, PNG format). However, in this study, only 400x magnification images were used, 1146 malignant data, and 547 benign data were used.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data Augmentation

This study's main objective is to analyze and compare the effectiveness of various data enhancement techniques, namely SMOTE, ADASYN, and Image Generator, to improve the performance of the CNN (Neural Network) model when using the ResNet101 architecture. This research focuses on applying data enhancement in the context of complex and unbalanced medical datasets, explicitly using the BreakKHis dataset. The first step of this research was to collect and prepare the BreaKHis dataset as a basis for the experiment. This data set will then be divided into training and validation subsets. As a medical data set, BreaKHis may suffer from class imbalance, where others may overrepresent certain types of pathology. Next, three different data augmentation techniques will be applied to the training data set:

i. SMOTE (Synthetic Minority Sampling Technique):
   This technique will produce a composite sample from the minority class based on nearest neighbors. The goal is to strengthen minority groups and reduce inequality.
ii. ADASYN (adaptive synthetic sampling):
   This technique also addresses class imbalances but with a more adaptive approach. ADASYN emphasizes minority samples, which are more difficult to classify.
iii. Image Generator
   This technique creates new image variations from the original data by applying transformations such as rotate, move, pan, etc. This is intended to increase the variety and variety of training data. The transformations carried out in image generator can be seen in Table I.

TABLE I. TRANSFORMATION IN THE IMAGE GENERATOR

Transformation	Specifications
Rescale	1.0 / 255
Channel Shift Range	20
Rotation Range	15
Brightness Range	0.1, 1.5
Horizontal Flip	True
Vertical Flip	True

Other oversampling techniques used to address class imbalance are SMOTE and ADASYN. Both techniques focus on increasing the number of data samples by creating new synthetic data. SMOTE creates synthetic samples of the minority class from the nearest neighbors. By doing so, SMOTE can generate synthetic data for the minority data that is close to equal to the majority class. Creating "synthetic" samples allows minority groups to perform over-sampling, instead of using duplicates of real data entries (Torres, Oliveira, & Gomes, 2022)(Babu & Rao, 2023). The k-nearest neighbors from the list are randomly selected depending on the number of over-samples required (Chlap et al., 2021). Our five nearest neighbors are currently used in our implementation. SMOTE's main process is to find the k-element nearest neighbors, which belong to each minority class sample. Then, it randomly selects one of these K-elements of neighbors. To avoid duplication of random samples, new samples will be generated using interpolation theory.

ADASYN is a development of SMOTE that considers the value distribution of data. By generating more examples in the lower-density specialized space and fewer examples in the feature space, this algorithm differs from other data augmentation algorithms. Adaptively shifting the decision boundaries to hard-to-learn samples is an advantage of this feature, which makes ADASYN more suitable than other data augmentation algorithms for handling network traffic with significant data imbalance. Therefore, before entering the model training, a resize process is performed to equalize the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

overall size of the dataset. The data size, which was initially 700x460 pixels, was changed to 224x224 pixels according to the input size of ResNet101. With these preprocessing steps, histopathology data can be used to train and test the ResNet101 model in breast cancer detection and diagnosis. After that, the results are used to determine the model's performance. The stages in data preprocessing are performed as critical steps to ensure the model can achieve optimal performance and provide accurate and reliable results.

After data enhancement, the CNN model with ResNet101 architecture will be drilled using the enriched data set. The model will be tested on a validation subset to measure its accuracy and other evaluation measures, such as recoverability. These experiments will provide insight into how eac**h enhancement technique improves model performance in the face of data limitations and constraints.**

Convolutional Neural Networks

Convolutional Neural Network (CNN) is a deep learning algorithm specifically designed to analyze and process images. CNN differs from other deep learning algorithms because it focuses on processing and extracting image information (Kumar, Gandhi, Zhou, Kumar, & Xiang, 2020). CNNs receive an image as input and determine weights and offsets for the various objects in the picture. This allows CNNs to recognize and distinguish images from others based on features learned during model training (El-Amir & Hamdy, 2020). In short, CNNs combine multiple processing layers with parallel elements to achieve better visualization and abstraction (Arrofiqoh & Harintaka, 2018). By combining these layers, convolutional neural networks can process and understand image data efficiently, making them a potent tool for object detection, classification, segmentation, and many other applications related to image analysis and image processing.

Transfer learning is a prominent computer vision technique that applies information gained from dealing with various network difficulties to new response settings. Research [18] used deep learning as a feature extractor along with Convolutional Neural Networks to reduce the complexity in complex images such as histopathology images. Transfer learning models are essential in improving diagnosis performance in the medical field, especially on complex images (Montaha et al., 2021).

The majority of the architectural models, including VGG16, InceptionNet, and ResNet, were trained on the same Imagenet large data set. During training, better models are developed. Transfer-learning from CNN models that have been previously trained using natural datasets such as ImageNet to medical images can provide significant improvement in model performance. When the training data size is small enough, the performance of a well-tuned network will not decrease much (Shallu & Mehra, 2018). So, it is necessary to pay attention to choosing the right parameters for each model to be able to create a model that generalizes well.

ResNet101

This study uses Resnet101 which consists of 101 convolutional layers organized into 33 blocks, where each block has a residual connection connecting the initial block to the final block. This innovative approach optimizes the residuals between the desired convolution and the input function. Remaining or omitted connections allow the model to recycle the last layer, thereby reducing training time and improving model performance (Eldin, Hamdy, & Adnan, 2021). ResNet101 is designed to use concatenation hops, allowing data to jump blocks of convolutional layers (Mahmud & Abdelgawad, 2023). This approach makes it easier and more efficient to achieve the desired functions, and also speeds up the neural network training process, thereby improving the accuracy and speed of the model (Aksoy & Salman, 2022). Therefore, ResNet101 is a reliable and efficient transfer learning model for medical image analysis and other applications. The present study examines the association between ResNet101 knowledge and the BreaKHis dataset and evaluates this expertise. In this case, the ResNet101 model is first trained using the ImageNet dataset. With the Transfer Learning method, the model will save time and computational resources because there is no need to train the model from scratch. The ability to use ResNet101 with transfer learning will allow the extraction of high-level features from histopathological images. During this process, flattening, thickening, and suppression

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

layers are applied to improve model performance. The inclusion of classes allows the trained model to do the job optimally. The parameters are changed during the training phase to achieve optimal model performance. This setting is made so that the model can better generalize the BreaKHis data.

Model Evaluation

In this study, we performed tests comparing the performance results of the model before and after applying data enhancement. Model evaluation is performed using a set of tests generated by splitting the dataset before training the model. The use of several matrices ensures the robustness of the ResNet101 model. We used different metrics to compare the model's performance. The following metrics are used to evaluate the ResNet101 model.

Accuracy is used as the evaluation model. Accuracy is a measure of the number of correct predictions from the trained model. The exact formula can be seen in the equation below:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

Precision is expressed as a percentage of total cases predicted. Precision indicates the percentage of correct model predictions if said to be correct. The Precision formula can be seen in equation below:

$$Precision = \frac{TP}{TP+FP}$$

The recall indicates how often the model is correctly predicted. The callback displays the percentage of positive cases out of all actual cases in the dataset. The callback formula can be seen in the equation below:

$$Recall = \frac{TP}{TP+FN}$$

The F score is used to score the model. The higher the F score, the better the model. The Formula F1 Score can be seen in the equation below:

$$F1 - Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

The confusion matrix is used as an indication for the attribute of the taxonomy (discrimination) rule containing the number of correctly or incorrectly classified items for each class.

RESULT

The confusion matrix values in Figures 2,3,4,5 is used to establish the value of accuracy, precision, recall, and F1-Score. Table II contains the results of the confusion matrix. Table II shows the performance metrics on the data before and after applying augmentation. In this section, we will discuss more explicit theories and calculations for each model performance evaluation method, namely accuracy, precision, recall, and F1-Score, based on the confusion matrix results listed in Table II. Accuracy is a metric that measures how number of correct predictions compared to the total amount of data. Precision measures how many optimistic predictions are accurate compared to the number of positive predictions. Precision shows how much the model can avoid showing false positive results. Recall (Sensitivity or True Positive Rate) measures how many optimistic predictions are accurate compared to the total number of positive data. Remember to show the ability model to find all positive cases. F1-Score is the harmonic mean between precision and recall, which provides information about the balance between the two. The F1-Score is useful when there is continuity between the positive and negative classes.

By using ADASYN as the data augmentation technique, the model experienced significant improvements in all evaluation metrics. The accuracy increased to 85.22%, and the F1-score reaching 85% indicates that the model has a good balance between precision and recall. SMOTE technique as

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

data augmentation, the model experienced improved performance compared to before. Although the accuracy (80.43%) increased, the F1-score (79.79%) showed that there was still a slight imbalance between precision and recall. Image Generator resulted in the most significant improvement in model performance. With an accuracy of 96.09% and an F1-score of 96.08%, the model has achieved excellent performance with an optimal balance between precision and recall. Th use of Image Generator as a data augmentation technique successfully improved the model performance with a good balance between precision and recall. In this case, Image Generator is the most effective technique in improving the model's overall performance.

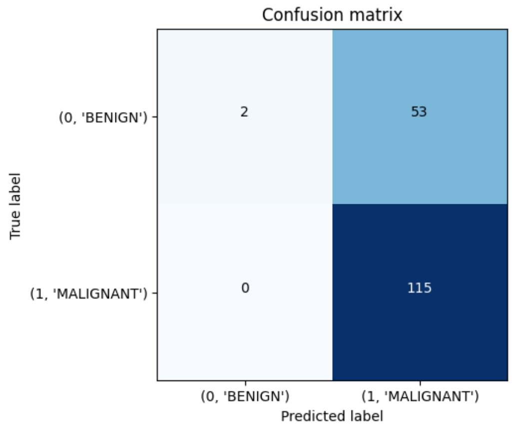

Fig. 3. Confusion Matrix Illustrating the Robustness of the ResNet101 in Dataset before Augmentation

Figure 3 presents the Confusion Matrix that illustrates the robustness of the ResNet101 model on the dataset before the application of augmentation techniques. Before the augmentation was applied, the model failed to predict the benign class and produced many false negatives.

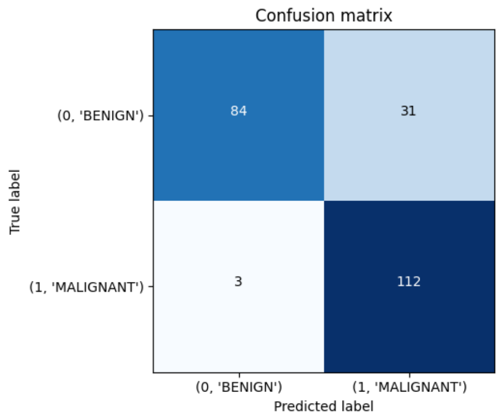

Fig. 4. Confusion Matrix Illustrating the Robustness of the ResNet101 in Dataset After Augmentation use ADASYN.

The confusion matrix is presented in illustration 4 the robustness of the ResNet101 model on the dataset after applying the ADASYN enhancement technique. This matrix provides a visual representation of the model's performance in classifying the different categories after the upgrade process. This matrix presents the distribution of predicted class labels relative to the actual truth labels in the data set, reflecting the impact of increasing ADASYN on the classifierability of the model.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

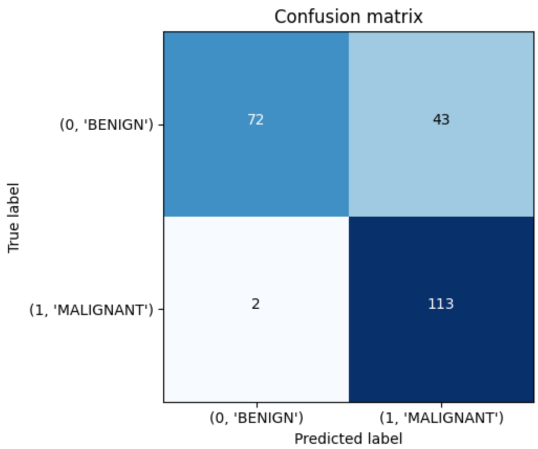

Fig. 5. Confusion Matrix Illustrating the Robustness of the ResNet101 in Dataset After Augmentation use SMOTE.

The provided confusion matrix provides a visualization of the robustness of the ResNet101 model on the dataset after applying the SMOTE enhancement technique in Figure 4. This matrix contains four key values. True positive (TP), true negative (TN), false negative (FN), and false positive (FP). True positives indicate the cases where the model correctly classified a positive case, and in this case, it was scored as 113. True negative indicates the correct classification of negative instances, with a count is 72. False negatives mean the cases in which the model incorrectly predicted a case negative. Class when it should have been positive, for 43 cases. Finally, a False Positive indicates an incorrect prediction of the positive class when it should have been negative by a count of 2.

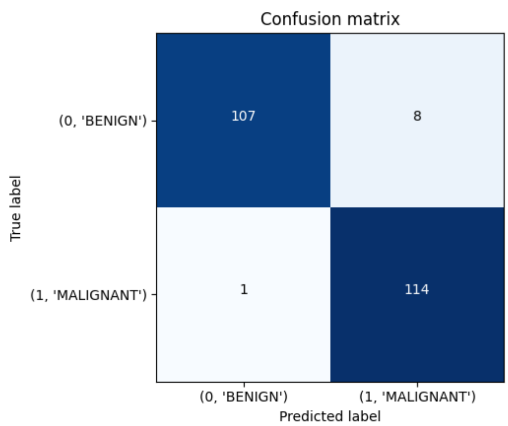

Fig. 6. Confusion Matrix Illustrating the Robustness of the ResNet101 in Dataset After Augmentation use Image Generator

The presented confusion matrix visualizes the resilience of the ResNet101 model on the dataset after incorporating the image generator enhancement technique. In this matrix, two important values, True Negative (TN) and True Negative (TP), provide insight into the classification performance of the model. True negatives, denoted by 107, mean cases where the model correctly identifies negative cases. On the other hand, True Positive, with a value of 114, indicates that the model correctly identified positive cases. These two values highlight the model's ability to correctly classify positive and negative cases after increasing the image generator. The confusion matrix helps to understand the model's ability to perform accurate classification and shows how increasing the image generator has improved the robustness and classification accuracy of the model.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

TABLE II. PERFORMANCE METRICS COMPARISON FOR ALL MODELS

Augmentation	***Accuracy***	***Precision***	***Recall***	***F1_score***
Before Augmentation	*0.6882*	*0.7866*	*0.6882*	*0.5725*
Image Generator	***0.9609***	***0.9626***	***0.9609***	***0.9608***
SMOTE	***0.8043***	***0.8487***	***0.8043***	***0.7979***
ADASYN	***0.8522***	***0.8744***	***0.8522***	***0.8500***

Based on the calculation results of the confusion matrix in Table II, it can be seen how each data augmentation method (ADASYN, SMOTE, and Image Generator) affects model performance in terms of accuracy, precision, accuracy, recall, and F1 score. We can analyze the improvements or changes to these metrics after implementing advanced techniques. This research shows that the Image Generator provides the most significant performance improvement with a good balance between accuracy and gain, resulting in the highest precision and F1 score. This shows that Image Generator, as a data enhancement technique, is effective in fixing problems to synchronization and improving overall model performance. At the same time, ADASYN and SMOTE offer increased performance but have a different ratio between accuracy and gain.

DISCUSSIONS

The following are various studies conducted on breast histopathology classification using data augmentation techniques. Table III lists the architecture along with the evaluation results achieved. We may not be able to make a fair comparison due to varying training and testing data. Therefore, we selected several CNN models that reported evaluation results on the BreaKHis dataset using data balancing and augmentation and compared them with our approach. n this method, our results are compared to the same evaluation to test the models' performance. With an accuracy rating of 96.09%, our suggested model outperforms all existing studies. However, the recall value of ViT is still better than our research. In the end, image generators are effective methods for boosting data quality and variety, which can enhance ResNet101 models' performance on BreaKHis data. Furthermore, selecting a complicated model is appropriate for complex images. Therefore, it's important to pay attention to choosing the correct CNN model for the task at hand.

TABLE III. COMPARISON OF THE PROPOSED MODEL WITH STATE-OF-THE-ART

Paper	***Method***	***Accuracy***	***Precision***	***Recall***	***F1_score***
Maistry et al (Maistry & Ezugwu, 2021)	***ViT + Augmentation***	***0.8896***	***0.8772***	***0.9658***	***0.9174***
S. Sharma et al(Sharma & Kumar, 2022)	***Xception + SVM + Augmentation***	***0.9411***	***0.9500***	***0.9300***	***0.9300***
S. Khrisna (Krishna, S, Krishnamoorthy, & Bhavsar, 2022)	***EfficicentNetB0+ Augmentation***	***0.9457***	***0.9434***	***0.8929***	***0.9636***
Proposed Model	***ResNet101+ Augmentation***	***0.9609***	***0.9626***	***0.9609***	***0.9608***

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12878
e-ISSN : 2541-2019
p-ISSN : 2541-044X

CONCLUSION

In classification tasks, evaluation metrics such as recall are very important to determine the model's overall performance, especially in medical data. Before implementing data augmentation, the recall was only 68.83%. This was influenced by the unbalanced dataset. After the augmentation technique was applied, the performance of the model increased. The combination of ResNet101 model with Image Generator gives the best model performance than other augmentation techniques. Accuracy, recall 96.09%, F1 Score 96.08%, and Precision 96.26%.

The analysis results provide valuable insights into the application of augmentation. Augmentation applied to BreaKHis data can help model performance, especially in the case of imbalanced data. This is because the model is trained with more additional data and variations without collecting time-consuming and expensive physical data. In addition, increasing variation in minority data can help the model learn valuable features in the dataset. The selection of data augmentation should also be tailored to the problem faced by the dataset and the goal of the developed model. This is because data augmentation plays an important role in the dataset used and will affect the future performance of the model. In model selection, it is necessary to consider the goal of the classification task and the trade-off between precision and recall. In some cases, such as in medical images, focusing on a high recall value is better to ensure that the model can detect as many positive samples as possible. It is important to analyze the evaluation metrics as a whole and specifically on relevant cases in order to get a clear picture of the model's performance and improve the model's capabilities based on the given task.

Although oversampling using ADASYN and SMOTE techniques is a useful approach to balance unbalanced data, the noise generated from synthetic data causes the model to be less stable and disrupts the learning process. This is because the synthetic data produced does not match or is less realistic than the characteristics of the original data. Thus, the data patterns are blurred, making it difficult for the model to understand the dataset's characteristics. Future research needs to focus on the quality of the dataset used to improve the evaluation's accuracy. Future research can focus on developing more sophisticated and adaptive data augmentation techniques such as GANs or VAEs to produce more diverse and realistic synthetic data. Future research should focus on developing more appropriate, efficient, and effective techniques for different tasks and domains. It is also necessary to consider the impact on model performance in real applications.

REFERENCES

Aksoy, B., & Salman, O. K. M. (2022). Prediction of Covid-19 disease with Resnet-101 deep learning architecture using Computerized Tomography images. *Türk Doğa ve Fen Dergisi*, *11*(2), 36–42. https://doi.org/10.46810/tdfd.1095624

Apostolopoulos, I. D. (2020). Investigating the Synthetic Minority Class Oversampling Technique (Smote) on an Imbalanced Cardiovascular Disease (Cvd) Dataset. *International Journal of Engineering Applied Sciences and Technology*, *04*(09), 431–434. https://doi.org/10.33564/ijeast.2020.v04i09.058

Arrofiqoh, E. N., & Harintaka, H. (2018). Implementasi Metode Convolutional Neural Network Untuk Klasifikasi Tanaman Pada Citra Resolusi Tinggi. *Geomatika*, *24*(2), 61. https://doi.org/10.24895/jig.2018.24-2.810

Babu, K. S., & Rao, Y. N. (2023). A Study on Imbalanced Data Classification for Various Applications. *Revue d'Intelligence Artificielle*, *37*(2), 517–524. https://doi.org/10.18280/ria.370229

Bloice, M. D., Roth, P. M., & Holzinger, A. (2019). Biomedical image augmentation using Augmentor. *Bioinformatics*, *35*(21), 4522–4524. https://doi.org/10.1093/bioinformatics/btz259

Chan, E., Kelly, M., & Schnabel, J. A. (2021). Comparison of classical machine learning deep learning to characterise fibrosis inflammation using quantitative MRI. *Proceedings - International Symposium on Biomedical Imaging*, *2021-April*, 729–732. https://doi.org/10.1109/ISBI48211.2021.9433962

Chen, Y., Chang, R., & Guo, J. (2021). Effects of Data Augmentation Method Borderline-SMOTE on Emotion Recognition of EEG Signals Based on Convolutional Neural Network. *IEEE Access*, *9*, 47491–47502. https://doi.org/10.1109/ACCESS.2021.3068316

*name of corresponding author

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 545–563. https://doi.org/10.1111/1754-9485.13261

Chowdhury, D., Das, A., Dey, A., Shreya, S., Dwivedi, A. D., Mukkamala, R. R., & Murmu, L. (2022). *ABCanDroid : A Cloud Integrated Android App for*.

Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, *PP*, 1–15. https://doi.org/10.1109/TNNLS.2021.3136503

El-Amir, H., & Hamdy, M. (2020). Deep Learning Pipeline. In *Deep Learning Pipeline*. https://doi.org/10.1007/978-1-4842-5349-6

Eldin, S. N., Hamdy, J. K., & Adnan, G. T. (2021). *Deep Learning Approach for Breast Cancer Diagnosis from Microscopy Biopsy Images*. 216–222. https://doi.org/10.1109/MIUCC52538.2021.9447653

Elmannai, H., Hamdi, M., & Algarni, A. (2021). Deep learning models combining for breast cancer histopathology image classification. *International Journal of Computational Intelligence Systems*, *14*(1), 1003–1013. https://doi.org/10.2991/ijcis.d.210301.002

Krishna, S., S, S. S., Krishnamoorthy, S., & Bhavsar, A. (2022). *Stain Normalized Breast Histopathology Image Recognition using Convolutional Neural Networks for Cancer Detection*. (1).

Kumar, A., Gandhi, C. P., Zhou, Y., Kumar, R., & Xiang, J. (2020). Improved deep convolution neural network (CNN) for the identification of defects in the centrifugal pump using acoustic images. *Applied Acoustics*, *167*, 107399. https://doi.org/10.1016/j.apacoust.2020.107399

Kummer, A., Ruppert, T., Medvegy, T., & Abonyi, J. (2022). Machine learning-based software sensors for machine state monitoring - The role of SMOTE-based data augmentation. *Results in Engineering*, *16*(November). https://doi.org/10.1016/j.rineng.2022.100778

Mahmud, I., & Abdelgawad, A. (2023). *A Deep Analysis of Transfer Learning Based Breast Cancer Detection Using Histopathology Images*.

Maistry, B., & Ezugwu, A. E. (2021). *Breast Cancer Detection and Diagnosis: A comparative study of state-of-the-arts deep learning architectures*. 1–20.

Montaha, S., Azam, S., Kalam, A., Rakibul, M., Rafid, H., Ghosh, P., … Boer, F. De. (2021). *BreastNet18 : A High Accuracy Fine-Tuned VGG16 Model Evaluated Using Ablation Study for Diagnosing Breast Cancer from Enhanced Mammography Images*.

Rashmi, R., Prasad, K., & Udupa, C. B. K. (2022). Breast histopathological image analysis using image processing techniques for diagnostic puposes: A methodological review. *Journal of Medical Systems*, *46*(1). https://doi.org/10.1007/s10916-021-01786-9

Saber, A., Sakr, M., Abo-seida, O. M., Keshk, A., Chen, H., & Member, A. (2021). *A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique*. *9*. https://doi.org/10.1109/ACCESS.2021.3079204

Shallu, & Mehra, R. (2018). Breast cancer histology images classification: Training from scratch or transfer learning? *ICT Express*, *4*(4), 247–254. https://doi.org/10.1016/j.icte.2018.10.007

Sharma, S., & Kumar, S. (2022). The Xception model: A potential feature extractor in breast cancer histology images classification. *ICT Express*, *8*(1), 101–108. https://doi.org/10.1016/j.icte.2021.11.010

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0197-0

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, *63*(7), 1455–1462. https://doi.org/10.1109/TBME.2015.2496264

Tiwari, D., Dixit, M., & Gupta, K. (2021). *Traitement du Signal Deep Multi-View Breast Cancer Detection : A Multi-View Concatenated Infrared Thermal Images Based Breast Cancer Detection System Using Deep Transfer Learning*. *38*(6), 1699–1711.

Torres, J., Oliveira, J., & Gomes, E. (2022). *The Usage of Data Augmentation Strategies on the Detection of Murmur Waves in a PCG Signal*. *4*(Biostec), 128–132.

*name of corresponding author

https://doi.org/10.5220/0010784500003123

Tripathi, P., Khatri, S. K., & Greunen, D. Van. (2022). *A transfer learning approach to implementation of pretrained CNN models for Breast cancer diagnosis*. *6*(4), 5816–5830.

Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2023). A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognition*, *137*, 109347. https://doi.org/10.1016/j.patcog.2023.109347

Xu, Y., & Wali, A. (2023). Handwritten Pattern Recognition using Birds-Flocking Inspired Data Augmentation Technique. *IEEE Access*, *11*(May), 71426–71434. https://doi.org/10.1109/ACCESS.2023.3294566

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. (2022). *Image Data Augmentation for Deep Learning: A Survey*. Retrieved from http://arxiv.org/abs/2204.08610

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., & Dalca, A. V. (2019). Data augmentation using learned transformations for one-shot medical image segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 8535–8545. https://doi.org/10.1109/CVPR.2019.00874