JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 6, No. 4. September 2024 DOI: https://doi.org/10.34288/jri.v6i4.348
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

IMPROVING IMAGE CLASSIFICATION ACCURACY WITH OVERSAMPLING AND DATA AUGMENTATION USING DEEP LEARNING: A CASE STUDY ON THE SIMPSONS CHARACTERS DATASET

Ilham Maulana[-1],Siti Ernawati[-2], Muhammad Indra[-3]

[1,3]Ilmu Komputer / Fakultas Teknologi Informasi
Universitas Nusa Mandiri
[1]k4ilham@gmail.com, [3]indra.orion@gmail.com

[2]Sistem Informasi / Fakultas Teknologi Informasi
Universitas Nusa Mandiri
[2]siti.ste@nusamandiri.ac.id

Abstract

The issue of data imbalance in image classification often hinders deep learning models from making accurate predictions, especially for minority classes. This study introduces AugOS-CNN (Augmentation and Over Sampling with CNN), a novel approach that combines oversampling and data augmentation techniques to address data imbalance. The *The Simpsons Characters* dataset is used in this study, featuring five main character classes: Bart, Homer, Agnes, Carl, and Apu. The number of samples in each class is balanced to 2,067 using an augmentation method based on *Augmentor*. The proposed model integrates oversampling and augmentation steps with a Convolutional Neural Network (CNN) architecture to improve classification accuracy. Evaluation results show that the AugOS-CNN model achieves the highest accuracy of 96%, outperforming the baseline CNN approach without data balancing techniques, which only reaches 91%. These findings demonstrate that the AugOS-CNN model effectively enhances image classification performance on datasets with imbalanced class distributions, contributing to the development of more robust deep learning methods for addressing data imbalance issues.

Keywords: Imbalanced Data; Oversampling; Data Augmentation; Convolutional Neural Network (CNN)

Abstrak

Masalah ketidakseimbangan data dalam klasifikasi gambar sering kali menghambat model deep learning dalam membuat prediksi yang akurat, terutama untuk kelas minoritas. Penelitian ini memperkenalkan AugOS-CNN (Augmentation and Over Sampling with CNN), sebuah pendekatan baru yang menggabungkan teknik oversampling dan augmentasi data untuk mengatasi ketidakseimbangan data. Dataset The Simpsons Characters *digunakan dalam penelitian ini, dengan lima kelas karakter utama, yaitu Bart, Homer, Agnes, Carl, dan Apu. Jumlah data dalam setiap kelas disamakan menjadi 2.067 sampel menggunakan metode augmentasi berbasis* Augmentor*. Model yang diusulkan mengintegrasikan langkah-langkah oversampling dan augmentasi dengan arsitektur Convolutional Neural Network (CNN) untuk meningkatkan akurasi klasifikasi. Hasil evaluasi menunjukkan bahwa model AugOS-CNN mencapai akurasi tertinggi sebesar 96%, mengungguli pendekatan CNN dasar tanpa teknik penyeimbangan data yang hanya mencapai 91%. Temuan ini membuktikan bahwa model AugOS-CNN secara efektif meningkatkan performa klasifikasi gambar pada dataset dengan distribusi kelas yang tidak seimbang, sehingga berkontribusi pada pengembangan metode deep learning yang lebih robust dalam menghadapi permasalahan ketidakseimbangan data.*

Kata Kunci: Imbalanced Data; Oversampling; Data Augmentation; Convolutional Neural Network (CNN)

INTRODUCTION

Image classification plays a crucial role in various applications, ranging from agriculture to healthcare and information technology. In agriculture, image classification techniques enable efficient monitoring and classification of agricultural lands using time-series satellite imagery (Meng et al., 2023). This improves accuracy in crop management and yield prediction. Additionally, in the healthcare sector, medical image classification using Convolutional Neural

Networks (CNN) has proven effective in disease diagnosis, such as in automated diagnostic systems for diabetic retinopathy (Iskandar & Salam, 2024).

In the field of information technology, image classification is also highly relevant, particularly in object recognition and image processing. CNN can be used for dog breed classification, demonstrating the network's ability to learn features from data during training (Harsa Pratama et al., n.d.). Additionally, machine learning techniques provide models that can enhance the efficiency of robot sensor information processing, including image recognition (Sichevskyi, 2022). Thus, image classification not only improves operational efficiency but also opens up new opportunities for innovation across various fields.

In the context of image classification, data imbalance is a significant challenge that can affect the performance of machine learning models, especially in applications involving pattern recognition and image classification. The proposed novel approach, AugOS-CNN, combines oversampling and data augmentation techniques to address this issue. Previous research has shown that oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling) can significantly improve classification accuracy for minority classes in imbalanced datasets (Gao et al., 2024)(Mahmudah et al., 2021)(Mirs, 2010). The combination of CNN models with ensemble techniques and oversampling methods, including SMOTE, can enhance overall classification performance, especially in situations with high class imbalance (Gao et al., 2024). This suggests that converting binary features to numeric values through feature extraction methods allows oversampling techniques to fully realize their potential in improving classification performance (Mahmudah et al., 2021). Other studies also support the use of SMOTE in the context of skin disease recognition, where this technique successfully generated a more balanced dataset and improved evaluation results (Mirs, 2010) (Iskandar & Salam, 2024).

Oversampling is a technique used to address data imbalance in machine learning, where the number of samples in the minority class is much smaller compared to the majority class. This imbalance often leads to the model prioritizing the majority class, resulting in poor classification performance for the minority class. Proper application of oversampling is key to improving model performance on imbalanced datasets, supporting better decision-making across various fields (Eom & Byeon, 2023). Data augmentation also plays a crucial role in enhancing model performance. Applying image augmentation techniques, alongside oversampling, can significantly improve accuracy, recall, and F1-score of the model (Iskandar & Salam, 2024). Data augmentation not only helps in expanding the dataset but also in increasing data diversity, which is important for training more robust models. It is also essential to address data shortages during experimental processes (Abayomi-Alli et al., 2021).

Overall, AugOS-CNN integrates Augmentor-based data augmentation with oversampling techniques to address the data imbalance issue in the Simpsons Characters Dataset. This method demonstrates great potential in handling data imbalance challenges in image classification. By leveraging a combination of augmentation and oversampling techniques, the model can be trained on a more balanced and diverse dataset, ultimately improving the accuracy and reliability of classification in various real-world applications.

RESEARCH METHODS

This research aims to develop the AugOS-CNN approach, which combines oversampling and data augmentation techniques with a Convolutional Neural Network (CNN) model to address the data imbalance issue in image classification. Figure 1 shows the flow of the AugOS-CNN model research. The study begins with the data preprocessing stage, aimed at balancing the dataset distribution and ensuring the data is ready for model training. After the data is balanced, the CNN model is designed and trained using the preprocessed data. The final stage is the model performance evaluation using a separate test dataset, to obtain a comprehensive assessment of the model's reliability.

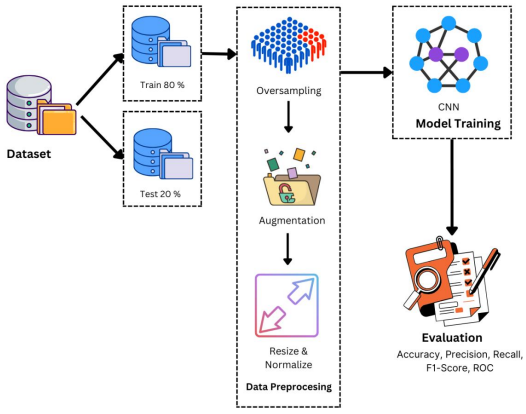

Figure 1. Research Stages

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v6i4.348
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

3.1. Dataset

The dataset used is the Simpsons 50 Characters Images - Unbalanced, sourced from Kaggle. This dataset contains images of 50 characters from the TV show The Simpsons. The research focuses on the classification of 5 main characters, namely Bart, Homer, Agnes, Carl, and Apu (as shown in Figure 2).

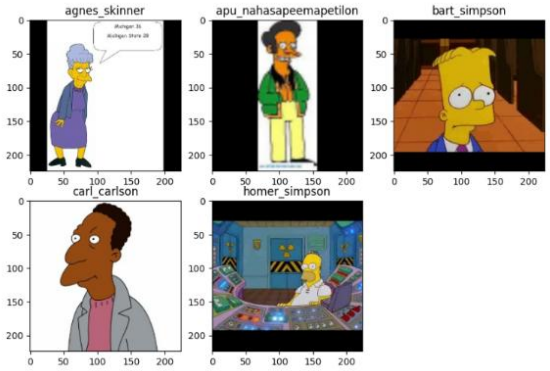

Figure 2. Dataset of Main Character

The main challenge of the dataset used is the imbalanced distribution among classes, with a significant variation in the number of images for each character. Figure 3 shows the dataset quantity per class: Bart with 1,257 samples, Homer with 2,067 samples, Agnes with 99 samples, Carl with 116 samples, and Apu with 593 samples.

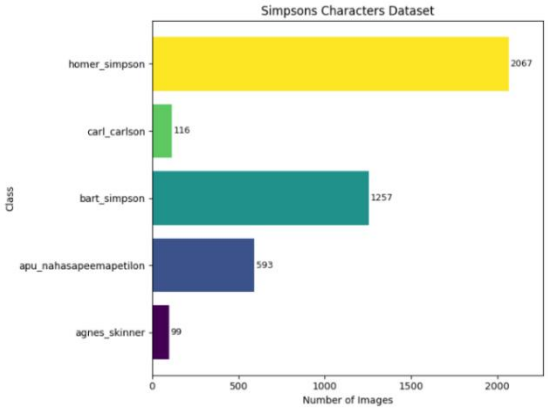

Figure 3. Dataset Quantity Per ClassAuthor Format

3.2. Data Preprocessing

One important step in preprocessing is resizing the images. In this study, all images were resized to 180x180 pixels to ensure consistent input for the CNN. This resizing is crucial because CNN require uniform input dimensions to perform training and inference effectively.

Resizing images can affect the performance of the CNN model, so it is important to apply the appropriate technique for resizing to maintain the quality of visual information (Kato et al., 2020). A good representation of image data is crucial for success in classification, and consistent resizing helps achieve the goals of the research (Liu et al., 2019).

The normalization stage aims to scale the pixel values of the images to a range of 0 to 1, which is done using the built-in Rescaling layer from TensorFlow. This process is crucial because it helps speed up convergence during model training. Normalization allows the model to learn more efficiently by reducing unnecessary variability in the data, thereby accelerating the training process and improving accuracy.

Data normalization is a key step in improving the accuracy and efficiency of models in image segmentation tasks (Li et al., 2022). This is particularly important in the context of deep learning, where scale differences between features can cause the model to struggle during learning. By applying normalization, the model can focus more on relevant patterns and features in the data, which in turn enhances the model's ability to classify characters in the dataset. Therefore, normalization is an essential step in data preprocessing that supports optimal model training.

The dataset was balanced using the oversampling technique with the help of the Augmentor tool, which generates additional variations of the minority class images. The techniques applied in this process include rotation, cropping, flipping, and other geometric transformations. By using this technique, the data for each class was balanced to 2,067 images per class, thereby improving the representation of the minority class in the dataset.

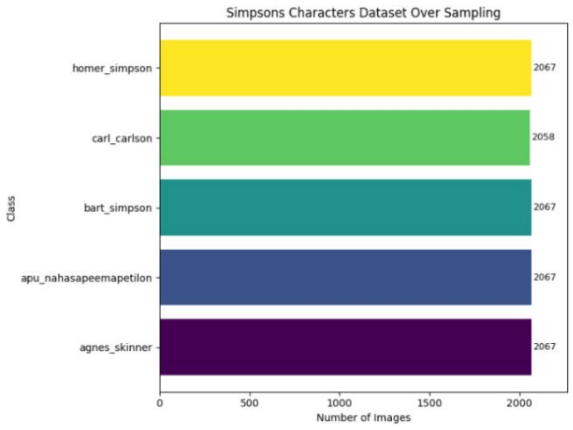

Figure 4. Dataset Quantity After Oversampling with Augmentor

Class balancing is crucial in the context of machine learning, especially when dealing with imbalanced datasets. Oversampling can help address data imbalance by increasing the number of samples from the minority class, allowing the model to learn better and reducing the risk of bias toward the majority class (Hayaty et al., 2021). Additionally, oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) can be used to generate synthetic samples that help create balance between the classes (Rustam et al., 2019). To tackle the dataset imbalance, a combination of oversampling and data augmentation was applied:

1) Oversampling

Oversampling is performed by adding data to the minority class to equalize the number of images across all classes. This technique is important to ensure that the model is not biased toward the majority class, which could result in poor performance on the minority class. Oversampling can improve the model's accuracy by providing more examples from the underrepresented class, helping the model learn relevant patterns from the data (Díez López et al., 2022). Oversampling is particularly effective in the context of imbalanced datasets, such as in rice pest classification, where collecting images for the minority class is very challenging (Li et al., 2022). By using this technique, the number of images for each class can be balanced, which in turn enhances the model's ability to classify characters in the dataset.

2) Augmentation With Augmentor

In this research, data augmentation techniques were applied using the Augmentor tool to increase the diversity of images and enrich the visual information in the dataset. The techniques used include rotation, cropping, flipping, and small-scale distortions.

Table 1. Augmentation Results with Augmentor

#	Bart	Homer	Agnes	Carl	Apu
Original					
Aug 1					
Aug 2					

Data augmentation is crucial for increasing the amount and variety of available data, allowing the model to learn from different perspectives and conditions. Data augmentation is a necessary step in deep learning model training, as it helps reduce overfitting and improves the model's generalization on unseen data (Fuadi et al., 2024). Additionally, transformations such as rotation and flipping are effective in enhancing image classification performance, as they help the model learn invariance to changes in position and orientation (Fuadi et al., 2024).

3.3. Clasification Model

The classification model used in this research is the Convolutional Neural Network (CNN), which is designed to recognize visual patterns in images and learn the features that distinguish between classes. CNNs have proven to be effective in various image recognition applications, including character classification in the Simpsons dataset. The structure of this model includes several convolutional layers followed by pooling layers, which aim to extract features from the input images.

CNN architectures like AlexNet have revolutionized image classification by applying convolutional networks to large datasets such as ImageNet, with data augmentation techniques used to significantly increase the size of the dataset (Shorten & Khoshgoftaar, 2019). The data augmentation performed earlier helps the model learn from image variations, thereby improving its generalization capability. The importance of the encoder-decoder architecture in the development of more complex models, which can enhance accuracy in pattern recognition, is also emphasized (ARPACI & VARLI, 2021).

In CNN architecture, convolutional layers play a crucial role in extracting features from the input images. The use of 3x3 kernels is a common

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v6i4.348
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

practice due to their ability to capture local details efficiently, contributing to more complex feature representations (Fernández et al., 2018). Additionally, the Rectified Linear Unit (ReLU) activation function is applied to accelerate convergence and reduce the vanishing gradient problem, which is often encountered during the training of deep neural networks. The combination of multiple convolutional layers with small kernels and ReLU activation has proven to be effective in enhancing the performance of CNN models in various applications, including image classification and object detection. Research shows that architectures integrating several convolutional layers can produce deeper feature representations, contributing to higher model accuracy (Shafique & Tehsin, 2018).

The dropout technique is also applied in the CNN architecture to prevent overfitting, particularly in deeper layers. Dropout works by randomly removing a number of neurons during the training process, which helps reduce dependency between neurons and enhances the model's generalization ability. Research indicates that setting the dropout rate between 0.2 and 0.5 can yield optimal results, depending on the complexity and size of the dataset used. The use of dropout has been shown to be effective in improving the performance of CNN models in various applications, including object recognition and image classification. By applying dropout progressively, the model can learn to rely on more robust features and avoid focusing on a specific subset of neurons, which often leads to overfitting.

Therefore, dropout becomes one of the important regularization methods in the training of neural networks in the context of deep learning. A fully connected layer with 1,024 neurons is activated using ReLU, followed by an output layer with a softmax activation function for multi-class classification. The model is compiled with Sparse Categorical Crossentropy as the loss function, the Adam optimizer (with a learning rate of 0.0001), and accuracy as the evaluation metric.

Table 2. AugOS-CNN Architecture

Layer	Kernel Size	Stride	Padding	Channel	Activation
Input				3	
Rescalling					
Convolution 2D	3x3	1	SAME	32	ReLU
MaxPooling 2D	2x2	2	VALID	32	
Dropout					
Convolution 2D	3x3	1	SAME	64	ReLU
MaxPooling 2D	2x2	2	VALID	64	
Dropout					
Convolution 2D	3x3	1	SAME	128	ReLU
MaxPooling 2D	2x2	2	VALID	128	
Dropout					
Convolution 2D	3x3	1	SAME	256	ReLU
MaxPooling 2D	2x2	2	VALID	256	
Dropout					
Flatten					
Dense	1x1	1		1024	ReLU
Dropout					
Dense	1x1	1		5	SoftMax

The CNN model shown in Table 2 represents the architecture used in this research, consisting of several convolutional and pooling layers. Starting with an input layer that receives color images (3 channels), the model performs rescaling for data normalization. Next, convolutional layers with 3x3 kernels and ReLU activation functions are used to extract features from the images, followed by MaxPooling layers to reduce the dimensionality and complexity of the data. Dropout is applied after several layers to prevent overfitting. This process is repeated with an increasing number of channels (32, 64, 128, and 256) at each convolutional layer. After several iterations of convolution and pooling, the data is flattened before passing through a dense layer with 1,024 neurons, culminating in the output layer, which uses the Softmax activation function for multi-class classification. This architecture demonstrates an effective approach for recognizing visual patterns and improving model accuracy in image recognition.

3.4. Evaluation

Model performance evaluation is conducted using a confusion matrix to provide an overview of the distribution of correct and incorrect predictions for each class, ensuring the effectiveness of AugOS-CNN in handling data imbalance. Research indicates that combining various evaluation metrics can provide better insights into the model's effectiveness in real-world application contexts. Thus, while accuracy is

a useful metric, it is important to complement it with deeper analysis to ensure that the model not only performs well statistically but is also relevant in the context of the desired application (NAHZAT & YAĞANOĞLU, 2021).

Accuracy is an important metric in model evaluation, which measures the percentage of correct predictions out of the total dataset. In the context of machine learning models, accuracy is often used as an initial indicator to assess model performance (NAHZAT & YAĞANOĞLU, 2021). However, it is important to note that accuracy can be misleading, especially in the case of imbalanced datasets, where the model might show high accuracy by ignoring the minority class (NAHZAT & YAĞANOĞLU, 2021). Therefore, it is recommended to use additional metrics such as precision, recall, and F1-score to gain a more comprehensive understanding of the model's performance (Díez López et al., 2022).

Precision and recall are two key metrics used to evaluate the performance of classification models, particularly in the context of imbalanced data. Precision measures the accuracy of the model's predictions for a specific class by calculating the proportion of correct predictions out of all the predictions made for that class. This is important to ensure that the model not only generates many positive predictions but also that those predictions are correct (Tan et al., 2021). Recall, on the other hand, measures how well the model is able to identify all the true instances of that class, providing an insight into the model's sensitivity to the minority class (Okawa et al., 2022). The combination of precision and recall provides a more comprehensive view of the model's performance in classification tasks. These metrics are also used to evaluate anomaly detection models, indicating that focusing on both metrics is crucial in the context of imbalanced datasets (Huang et al., 2023).

Additionally, the F1-score, which is the harmonic mean of precision and recall, is often used to provide a more balanced view of the model's performance (Xie et al., 2021). The F1-score helps assess the trade-off between precision and recall. Furthermore, performance is also evaluated using the ROC curve, which plots the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) to assess classification performance for each class.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\text{-}Score = \frac{2TP}{2TP + FP + FN} \quad (4)$$

Where the terms TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively [4].

RESULTS AND DISCUSSION

The experiment was conducted with training for 25 epochs using a batch size of 32. The dataset was split into 80% for training and 20% for testing, with 20% of the training data used for validation. Callbacks, such as Early Stopping, stopped the training if the validation loss did not improve for 10 consecutive epochs, and Model Checkpoint saved the best model based on validation loss. The entire training process was carried out on a GPU to accelerate computation.

Character classification in The Simpsons dataset shows that the proposed model, AugOS-CNN, successfully addressed the data imbalance issue and performed better compared to the baseline model. By utilizing data augmentation and oversampling techniques, the AugOS-CNN model significantly improved the image classification accuracy of character images. The character classification in the dataset.

The performance comparison between the AugOS-CNN model and other approaches, such as CNN Imbalance and CNN with oversampling techniques, is shown in Figure 5. The results indicate that AugOS-CNN provides more optimal outcomes. The AugOS-CNN model achieved the highest accuracy of 96%, precision of 96%, recall of 96%, and F1-Score of 96%, outperforming the CNN Imbalance model with accuracy of 91%, precision of 88%, recall of 80%, and F1-Score of 83%. These results demonstrate the effectiveness of AugOS-CNN in handling imbalanced class distributions and improving classification quality.

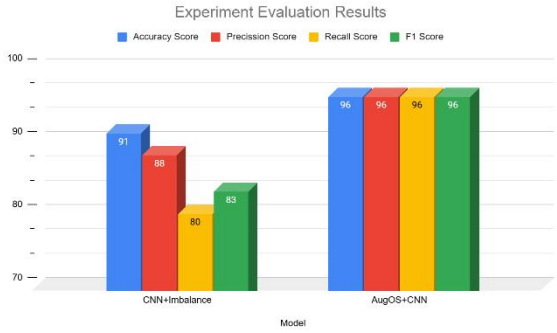

Figure 5. Experiment Evaluation Results

Figure 6 shows the evaluation results of the AugOS-CNN model compared to the standard CNN model, highlighting a significant improvement in precision, recall, and F1-score for the minority class. For example, for the "Agnes" class, the AugOS-CNN model achieved 96% precision and 98% recall, while the CNN model without imbalance handling only obtained 86% precision and 55% recall. Overall, AugOS-CNN outperformed in all classes, with the highest F1-score of 97% for the "Agnes" class, indicating that this approach effectively addresses the class imbalance issue and enhances classification accuracy.

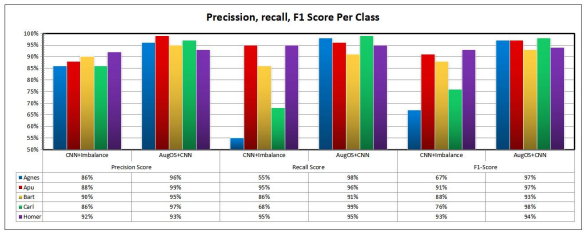

Figure 6. Precision, Recall, F1 Score per Class

The graph in Figure 7 compares the CNN+Imbalance model with the AugOS-CNN model that applies oversampling and data augmentation. At the top, the accuracy graph shows that the CNN+Imbalance model achieves lower accuracy and tends to stagnate, while the loss graph indicates that validation loss remains high, signaling overfitting. In contrast, at the bottom, the AugOS-CNN model shows a more consistent increase in accuracy, approaching 100%, with a significant decrease in loss, indicating that this approach is more effective in addressing data imbalance and improving overall model performance.

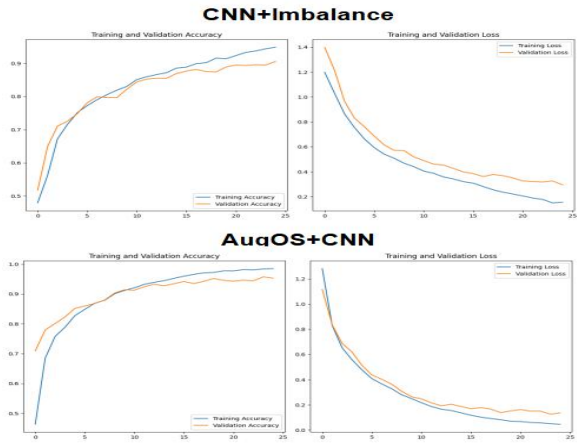

Figure 7. Training and Validation Accuracy

The comparison of the confusion matrix between the CNN+Imbalance model (on the left) and the AugOS-CNN model (on the right) is shown in Figure 8. In the CNN+Imbalance model, it can be observed that the model struggles to recognize minority classes such as Agnes Skinner and Carl Carlson, with a much lower number of correct predictions compared to the majority class, Homer Simpson. In contrast, the AugOS-CNN model shows more balanced predictions across all classes, with a significantly higher number of correct predictions, including for the minority classes. This demonstrates the effectiveness of the data augmentation and oversampling techniques used by AugOS-CNN in improving the model's accuracy and sensitivity towards classes with fewer data points. The AugOS-CNN model is able to minimize prediction errors between classes, resulting in much more optimal performance.

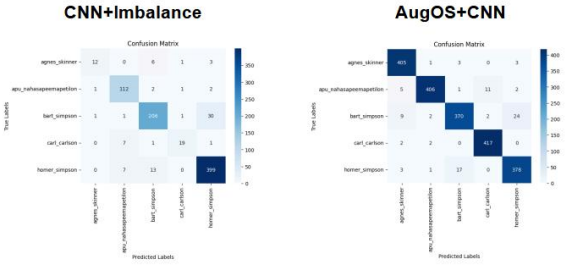

Figure 8. Confusion Matrix

The ROC curve in Figure 9 compares the performance of the CNN model trained with imbalanced data (CNN+Imbalance) and the AugOS-CNN model, which uses oversampling and data augmentation. The CNN+Imbalance model shows varying AUC values, with the class "apu_nahasapeemapetilon" achieving an AUC of 1.00, but other classes like "agnes_skinner" only

reach 0.97, indicating instability in differentiating classes. In contrast, the AugOS-CNN model demonstrates higher and more consistent AUC across all classes, with some classes, such as "carl_carlson," achieving AUC of 1.00. This indicates that AugOS-CNN is more effective in addressing data imbalance and improving classification accuracy.

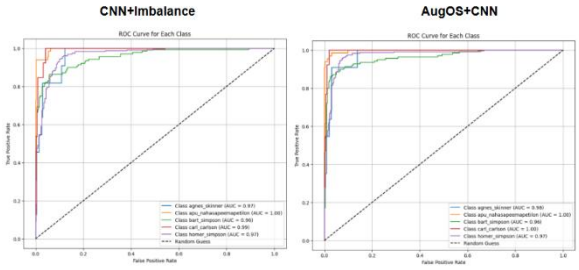

Figure 9. ROC Curve

Figure 10, on the left, shows per-class accuracy for the CNN model trained on imbalanced data (CNN+Imbalance), where the class "agnes_skinner" has very low accuracy at only 0.36, while the class "apu_nahasapeemapetilon" achieves high accuracy at 0.94. Other classes, such as "bart_simpson" and "homer_simpson," also show good accuracy, but the class "carl_carlson" only reaches 0.54, indicating uneven performance across classes.

In contrast, the second image represents the AugOS-CNN model, which shows improved accuracy across all classes. The accuracy for "agnes_skinner" rises to 0.64, and "carl_carlson" improves to 0.77. The classes "homer_simpson" and "bart_simpson" also show better accuracy, with scores of 0.89 each. This indicates that the AugOS-CNN approach is more effective in handling data imbalance and improving overall classification performance.

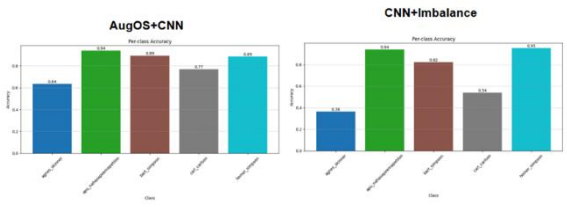

Figure 10. Per Class Accuracy

The experimental results show that AugOS-CNN significantly addresses the issue of data imbalance, which typically reduces model accuracy, especially for minority classes. Data imbalance can cause the model to focus more on the majority class, reducing its ability to recognize the minority class. By combining data augmentation techniques using Augmentor and oversampling, AugOS-CNN successfully produces a more balanced dataset and improves overall model performance. The data augmentation technique adds diversity to the dataset, while oversampling helps ensure that the minority classes are adequately represented during training.

In Table 3, Xiaoling Gao used the CIFAR-10 dataset with the CNN-AdaBoost algorithm, achieving an accuracy of 86.36%. E. Mirs applied CNN-SMOTE on the Plant Pathology 2020 - FGVC7 dataset, with an accuracy of 92%. N. H. Pratama used ResNet 101 on the Tsinghua Dogs dataset, obtaining an accuracy of 93%. M. Hayaty, working with images of medicinal plant leaves using MobileNetV2, achieved an impressive accuracy of 97.74%. L. Riyadi used CNN for X-ray images, reaching an accuracy of 95.67%. The proposed model, which focuses on the Simpsons 50 Characters Images dataset, achieved a good accuracy of 96% using the AugOS-CNN algorithm.

Table 3. Comparison of Deep Learning Models

Authors	Dataset	Best Algoritm	Accuracy(%)
Xiaoling Gao (Gao et al., 2024)	CIFAR-10	CNN-AdaBoost	86.36%
E. Mirs (Mirs, 2010)	Plant Pathology 2020 - FGVC7	CNN-SMOTE	92%
N. H. Pratama (Harsa Pratama et al., n.d.)	Tsinghua Dogs dataset	ResNet 101	93%
M. Hayaty	Medicinal Plant Leaf Images	MobileNetV2	97.74%
L. Riyadi	X-ray images	CNN	95.67%
Proposed Model	**Simpsons 50 Characters Images**	**AugOS-CNN**	**96%**

The results obtained from the comparison between AugOS-CNN and other approaches, such as the basic CNN, show that AugOS-CNN achieves significantly higher accuracy, with 96% compared to 91% in the basic CNN. The data augmentation and oversampling applied in AugOS-CNN not only preserve data diversity but also enhance the model's ability to better recognize all classes.

Thus, AugOS-CNN demonstrates excellent potential in improving image classification performance, especially in datasets with imbalanced class distributions. This technique can be applied to various fields, including face

recognition, medical image processing, and object classification tasks facing similar challenges.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This research demonstrates that AugOS-CNN, which combines data augmentation and oversampling, successfully addresses the issue of data imbalance in image classification. Using the "The Simpsons" dataset, the AugOS-CNN model achieved a top accuracy of 96%, surpassing the baseline CNN model (91%). These results confirm that the AugOS-CNN method is effective in improving model performance on imbalanced datasets, enhancing prediction accuracy, and increasing the reliability of the model in image classification. This approach holds great potential for applications in real-world image classification, where data imbalance is often a major challenge.

Suggestion

Future research is expected to explore the use of other, more varied augmentation and oversampling techniques to further improve model performance on more complex datasets. Additionally, testing the AugOS-CNN model on various datasets with higher or lower imbalance levels could provide further insights into the effectiveness of this method in broader contexts. Future studies could also optimize this model by applying regularization techniques or more advanced network architectures to reduce potential overfitting and improve performance on unseen data.

REFERENCES

Abayomi-Alli, O. O., Damaševičius, R., Misra, S., Maskeliūnas, R., & Abayomi-Alli, A. (2021). Malignant skin melanoma detection using image augmentation by oversampling in nonlinear lower-dimensional embedding manifold. *Turkish Journal of Electrical Engineering and Computer Sciences*, *29*(8), 2600–2614. https://doi.org/10.3906/elk-2101-133

ARPACI, S. A., & VARLI, S. (2021). LUPU-Net: a new improvement proposal for encoder-decoder architecture. *International Advanced Researches and Engineering Journal*, *5*(3), 352–361. https://doi.org/10.35860/iarej.939243

Díez López, C., Montiel González, D., Vidaki, A., & Kayser, M. (2022). Prediction of Smoking Habits From Class-Imbalanced Saliva Microbiome Data Using Data Augmentation and Machine Learning. *Frontiers in Microbiology*, *13*(July), 1–12. https://doi.org/10.3389/fmicb.2022.886201

Eom, G., & Byeon, H. (2023). Searching for Optimal Oversampling to Process Imbalanced Data: Generative Adversarial Networks and Synthetic Minority Over-Sampling Technique. *Mathematics-MDPI*, *11*(16), 3605. https://doi.org/10.3390/math11163605

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Fuadi, E. H., Ruslim, A. R., Wardhana, P. W. K., & Yudistira, N. (2024). Gated Self-supervised Learning for Improving Supervised Learning. *Proceedings - 2024 IEEE Conference on Artificial Intelligence, CAI 2024*, 611–615. https://doi.org/10.1109/CAI59869.2024.00120

Gao, X., Jamil, N., Ramli, M. I., & Ariffin, S. M. Z. S. Z. (2024). A Comparative Analysis of Combination of CNN-Based Models with Ensemble Learning on Imbalanced Data. *International Journal on Informatics Visualization*, *8*(1), 456–464. https://doi.org/10.62527/joiv.8.1.2194

Harsa Pratama, N., Rachmawati, E., & Kosala, G. (n.d.). *CLASSIFICATION OF DOG BREEDS FROM SPORTING GROUPS USING CONVOLUTIONAL NEURAL NETWORK*.

Hayaty, M., Muthmainah, S., & Ghufran, S. M. (2021). Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification. *International Journal of Artificial Intelligence Research*, *4*(2), 86. https://doi.org/10.29099/ijair.v4i2.152

Huang, P., Shang, J., Xu, Y., Hu, Z., Zhang, K., Dai, J., & Yan, H. (2023). Anomaly detection in radiotherapy plans using deep autoencoder networks. *Frontiers in Oncology*, *13*(March), 1–8. https://doi.org/10.3389/fonc.2023.1142947

Iskandar, D. A., & Salam, A. (2024). Evaluasi Performa Oversampling dan Augmentasi pada Klasifikasi Penyakit Kulit Menerapkan Convolutional Neural Network. *Jurnal Media Informatika Budidarma*, *8*(1), 240. https://doi.org/10.30865/mib.v8i1.7119

Kato, H., Osuge, K., Haruta, S., & Sasase, I. (2020). A preprocessing by using multiple steganography for intentional image downsampling on CNN-based steganalysis.

IEEE Access, *8*, 195578–195593. https://doi.org/10.1109/ACCESS.2020.3033814

Li, Z., Jiang, X., Jia, X., Duan, X., Wang, Y., & Mu, J. (2022). Classification Method of Significant Rice Pests Based on Deep Learning. *Agronomy*, *12*(9). https://doi.org/10.3390/agronomy12092096

Liu, L., Chen, J., Fieguth, P., Zhao, G., Chellappa, R., & Pietikäinen, M. (2019). From BoW to CNN: Two Decades of Texture Representation for Texture Classification. *International Journal of Computer Vision*, *127*(1), 74–109. https://doi.org/10.1007/s11263-018-1125-z

Mahmudah, K. R., Indriani, F., Takemori-sakai, Y., Iwata, Y., Wada, T., & Satou, K. (2021). Classification of imbalanced data represented as binary features. *Applied Sciences (Switzerland)*, *11*(17). https://doi.org/10.3390/app11177825

Meng, H., Li, C., Liu, Y., Gong, Y., He, W., & Zou, M. (2023). Corn Land Extraction Based on Integrating Optical and SAR Remote Sensing Images. *Land*, *12*(2). https://doi.org/10.3390/land12020398

Mirs, E. (2010). Oversampled-Based Approach to Overcome Imbalance Data in the Classification of Apple Leaf Disease with SMOTE. *Romanian Journal Ofapplied Science and Technology*, *XIII*(3), 254–260.

NAHZAT, S., & YAĞANOĞLU, M. (2021). Makine Öğrenimi Sınıflandırma Algoritmalarını Kullanarak Diyabet Tahmini. *European Journal of Science and Technology*, *24*, 53–59. https://doi.org/10.31590/ejosat.899716

Okawa, T., Mizuno, T., Hanabusa, S., Ikeda, T., Mizokami, F., Koseki, T., Takahashi, K., Yuzawa, Y., Tsuboi, N., Yamada, S., & Kameya, Y. (2022). Prediction model of acute kidney injury induced by cisplatin in older adults using a machine learning algorithm. *PLoS ONE*, *17*(1 January), 1–10. https://doi.org/10.1371/journal.pone.0262021

Rustam, Z., Utami, D. A., Hidayat, R., Pandelaki, J., & Nugroho, W. A. (2019). Hybrid preprocessing method for support vector machine for classification of imbalanced cerebral infarction datasets. *International Journal on Advanced Science, Engineering and Information Technology*, *9*(2), 685–691. https://doi.org/10.18517/ijaseit.9.2.8615

Shafique, S., & Tehsin, S. (2018). Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. *Technology in Cancer Research and Treatment*, *17*, 1–7. https://doi.org/10.1177/1533033818802789

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

Sichevskyi, S. (2022). Machine Learning Techniques for Increasing Efficiency of the Robot's Sensor and Control Information Processing †. *Sensors - MDPI*, *22*, 2–31. https://doi.org/https://doi.org/10.3390/s22031062

Tan, L., Lu, J., & Jiang, H. (2021). Tomato Leaf Diseases Classification Based on Leaf Images: A Comparison between Classical Machine Learning and Deep Learning Methods. *AgriEngineering*, *3*(3), 542–558. https://doi.org/10.3390/agriengineering3030035

Xie, J., Wang, Z., Yu, Z., Guo, B., & Zhou, X. (2021). Ischemic stroke prediction by exploring sleep related features. *Applied Sciences (Switzerland)*, *11*(5), 1–25. https://doi.org/10.3390/app11052083