JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Lung Cancer Detection using ProCAN with Enhanced Data Balancing Techniques

T.KALA
Department of Computer Science and Engineering, University College of Engineering, Kancheepuram, Kanchipuram, India

Abstract: Lung cancer is a major health concern, and its early detection is crucial for improving patient survival rates. However, identifying malignant nodules in computed tomography (CT) scans remains challenging due to the complexity of tumor patterns and class imbalance in medical datasets. This study proposes an advanced deep learning approach using the Progressive Growing Channel Attentive Non-Local (ProCAN) network, combined with the Synthetic Minority Over-sampling Technique (SMOTE) to enhance detection accuracy. ProCAN leverages progressive feature learning, channel attention mechanisms, and non-local dependencies, allowing the model to focus on critical tumor regions while capturing both local and global patterns. SMOTE addresses the imbalance issue by generating synthetic samples for underrepresented classes, ensuring a more robust training process. Experimental results demonstrate improved classification accuracy, enhanced sensitivity to malignant cases, and a reduced false-negative rate. This research highlights the effectiveness of ProCAN with SMOTE in lung cancer detection, offering potential benefits for early diagnosis and treatment planning.

Keywords: Lung cancer detection, Deep learning, Progressive Growing Channel Attentive Non-Local network (ProCAN), Synthetic Minority Over-sampling Technique (SMOTE), Lung Segmentation, Class imbalance, Computed Tomography scan analysis, Channel attention mechanism.

1. Introduction

Lung cancer is one of the most prevalent and life-threatening diseases, with early detection playing a crucial role in improving survival rates. Computed tomography (CT) scans are widely used for lung cancer diagnosis, offering detailed imaging of pulmonary structures. However, the manual interpretation of these images is highly complex and requires extensive expertise. Subtle or small malignant nodules can be overlooked, leading to misdiagnosis or delayed treatment. To address these challenges, deep learning has emerged as a powerful tool, offering automated and efficient analysis of medical images to assist in cancer detection [1].

Despite the advancements in deep learning models, a significant challenge in lung cancer detection is the imbalance in medical datasets. Malignant cases are significantly fewer than benign ones, causing models to favour the majority class while failing to accurately detect cancerous nodules.

This research introduces a Progressive Growing Channel Attentive Non-Local (ProCAN) network integrated with the Synthetic Minority Over-sampling Technique (SMOTE) to enhance lung nodule detection. ProCAN progressively refines feature learning, applies channel attention mechanisms to prioritize critical regions, and captures both local and global patterns using non-local operations. Meanwhile, SMOTE generates synthetic instances of the underrepresented malignant class, balancing the dataset and improving model learning. By combining these techniques, this study aims to enhance detection accuracy, reduce false negatives, and improve early diagnosis. The proposed approach can support radiologists in clinical decision-making, ultimately contributing to improved lung cancer outcomes.

2. Background and Related Work

Osadebey et al. [2] proposed a three-stage deep learning-based lung segmentation method to improve accuracy in CT scans. Their framework consists of preprocessing, segmentation, and postprocessing. A CNN classifier in preprocessing removes non-lung slices, reducing false positives. The U-Net model in segmentation extracts lung regions, while postprocessing refines contours using another U-Net and CNN. Their method achieved Dice scores between 0.76 and 0.95, demonstrating robustness across datasets.

Our approach builds on this by integrating Hounsfield Unit (HU)-based thresholding and morphological operations to enhance segmentation accuracy. These refinements improve lung boundary delineation and reduce segmentation errors, ensuring better performance in lung cancer detection.

Harale et al. [3] proposed an improved YOLOv5 (iYOLOv5) model for lung nodule detection and a CNN-SVM-based classification framework to differentiate nodule types. Their approach focused on enhancing detection accuracy using object detection techniques like YOLOv5 and Faster R-CNN. The proposed iYOLOv5 model outperformed its baseline version, achieving a mean Average Precision (mAP) of 0.864 and demonstrating improved convergence during training. For classification, the CNN-SVM model categorized nodules into well-circumscribed, juxta-vascular, and juxta-pleural types, achieving 98.17% accuracy on the LIDC-IDRI dataset.

Although their method showed improved detection and classification, it primarily relied on local feature extraction, which may not effectively capture complex tumor patterns. Additionally, object detection models like YOLOv5 perform well for nodule localization, but their reliance on predefined anchor boxes limits their ability to model variations in lung nodule morphology. In contrast, the ProCAN (Progressive Growing Channel Attentive Non-Local) network addresses these challenges by incorporating progressive feature learning, channel attention mechanisms, and non-local dependencies. This allows the model to focus on critical tumor regions while capturing both local and global patterns, improving the classification of malignant nodules. Unlike CNN-SVM, ProCAN dynamically refines its feature representation, leading to better discrimination between benign and malignant nodules without requiring manual feature engineering.

Tang et al. [4] proposed a method to address class imbalance in lung nodule classification by integrating SMOTE into their Multiresolution 3D DPSECN framework. Their approach involved a semi-supervised clustering algorithm to assign pseudo-labels to unlabelled samples before applying SMOTE to generate synthetic malignant nodules. This technique aimed to expand the minority class distribution, ensuring a more balanced dataset for training. The method was tested on the LIDC-IDRI dataset, achieving 94.4% accuracy and an AUC of 0.931, demonstrating its effectiveness in reducing model bias toward the majority class.

In their implementation, SMOTE was applied after pseudo-labelling, which introduced potential errors if misclassified samples were oversampled. This could distort the spatial, textural, and intensity-based features of malignant nodules, affecting classifier performance. Additionally, their k-nearest neighbour (KNN) selection lacked precise distance constraints, which could result in synthetic samples that deviated from real malignant characteristics, impacting model generalization. Furthermore, their approach adjusted the classifier's decision boundary without explicitly controlling distribution overlap between synthetic and real malignant nodules, increasing the risk of misclassification.

Our proposed method improves upon this by applying SMOTE in a fully supervised setting, ensuring that only correctly labelled malignant samples are used for synthetic data generation. We optimize nearest neighbour selection using refined Euclidean distance constraints, preserving the natural feature distribution of malignant nodules. Additionally, This approach controls the decision boundary shift, preventing overgeneralization and improving class separability. These refinements enhance the model's ability to distinguish between benign and malignant cases, reducing false negatives in lung cancer detection.

3. Proposed Methodology

3.1 Dataset and Preprocessing

The dataset used in this study consists of computed tomography (CT) scan images of lung nodules, obtained from publicly available sources such as LIDC-IDRI or clinical datasets [5]. These datasets provide a diverse range of annotated lung nodule samples, making them suitable for training deep-learning models. However, raw medical images often contain noise, variations in intensity, and irrelevant anatomical structures that can affect model performance. To ensure accurate lung cancer detection, data preprocessing is applied to refine input images before they are processed by the deep learning pipeline.

The first step in preprocessing involves normalizing CT scans using trilinear interpolation to achieve isotropic resolution across the x, y, and z dimensions. This ensures that the spatial scale remains consistent across different scans, allowing the model to learn relevant patterns effectively. After normalization, a $32\times32\times32$ mm^3 volume is cropped around the detected nodule, as this is sufficient to accommodate the largest nodule found in the LIDC-IDRI dataset (which has a maximum diameter of 30 mm).

3.2 Lung Segmentation

Lung segmentation is a crucial preprocessing step in lung nodule detection, as it isolates the lung region from surrounding anatomical structures such as the ribcage, spine, and air-filled spaces. Precise segmentation ensures that the deep learning model focuses on relevant pulmonary regions, reducing computational complexity and enhancing classification accuracy.

Each patient sample contains approximately 40-160 CT scan slices in 512×512 resolution, forming a 3D composite of the lung. The pixel values in these images represent radio density in Hounsfield Units (HU), ranging from -1000 HU for air to 700-3000 HU for bones. Most lung tissue falls within the range of -500 to 0 HU, while nodules can

range from -500 HU to over 200 HU. These HU values provide essential diagnostic information, as the composition of nodules plays a critical role in distinguishing between benign and malignant cases [6].

To achieve effective segmentation, the HU thresholding method is applied to differentiate lung tissue from non-lung regions. The HU values are clamped between -1000 HU and 400 HU, where:

- HU values below **-1000** correspond to air, which is excluded.
- HU values above **400** represent bones and other dense structures, which are also removed.

After HU-based filtering, morphological operations such as erosion and dilation are used to refine the segmented lung masks, removing small artifacts and improving boundary accuracy [7]. Once the lung regions are successfully segmented, the extracted lung mask is applied to the original CT scan, preserving only the relevant pulmonary structures for further processing. This segmentation step ensures that the subsequent feature extraction and classification tasks are performed on a clean, focused lung region, minimizing interference from non-essential anatomical components.

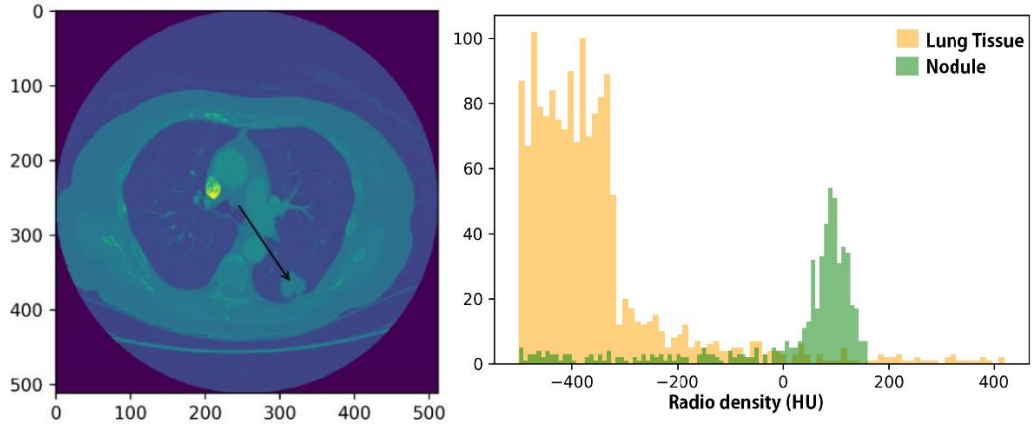

Fig.1. Hounsfield Unit (HU) Distribution of Lung Tissue and Nodule

3.3 Resolving Class Imbalance with SMOTE

In lung nodule detection, deep learning models often face challenges due to class imbalance, where malignant nodules are significantly underrepresented compared to benign cases. This imbalance causes the model to be biased towards the majority class, increasing the likelihood of false negatives, a critical issue in early cancer detection [8]. To address this, the Synthetic Minority Oversampling Technique (SMOTE) is applied to generate additional synthetic malignant nodule samples, ensuring a more balanced dataset for training.

Instead of simply duplicating existing malignant samples, SMOTE generates synthetic samples by interpolating between real malignant cases. This approach preserves the feature distribution of malignant nodules and allows the model to learn a more generalized representation, ultimately improving detection accuracy.

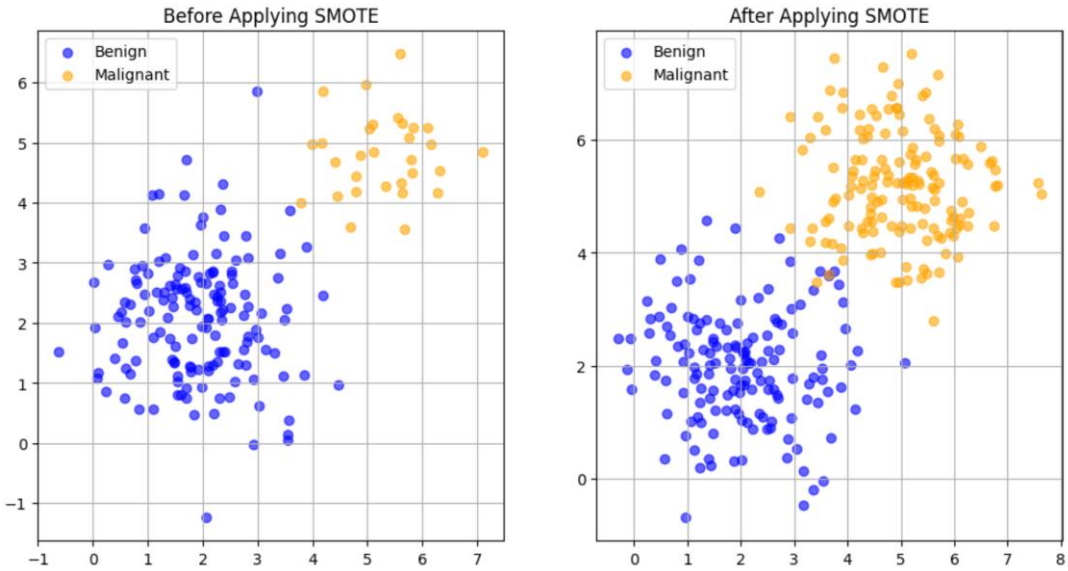

Fig.2. Class Distribution Before and After Applying SMOTE

3.3.1 Feature Space Representation and Minority Class Identification

Lung nodules are characterized by spatial, textural, and intensity-based features, extracted from medical imaging data. These features form a high-dimensional space in which classification is performed. Due to the inherent imbalance in medical datasets, malignant nodules are identified as the minority class, requiring oversampling to prevent biased learning [8]. Since an imbalanced dataset shifts the classifier's decision boundary toward the majority class, SMOTE is introduced to expand the minority class distribution and enhance the model's ability to distinguish malignant nodules.

3.3.2 Nearest Neighbour Selection and Synthetic Sample Generation

SMOTE generates synthetic samples by analysing the k-nearest neighbours of a given malignant nodule and creating new samples through interpolation. The selection of k is a crucial factor that influences the variability and realism of the synthetic samples. Euclidean distance is used to determine the similarity between malignant samples, ensuring that the newly generated instances align with the natural feature distribution of real malignant nodules.

For each minority class instance X_i , one of its k-nearest neighbours X_k is randomly selected, and a new synthetic sample is generated as follows:

$$X_{new} = X_i + \lambda \cdot (X_k - X_i) \quad (1)$$

Where X_i is a minority class instance, X_k is a randomly selected neighbour, and λ is a random value between 0 and 1 [8]. This ensures that the synthetic samples remain within the natural feature space of the minority class

3.3.3 Impact on Classifier Decision Boundary

In an imbalanced dataset, the classifier's decision boundary is skewed toward the majority class, making it difficult to accurately classify malignant cases. By introducing synthetic malignant samples, SMOTE adjusts the decision boundary so that it better separates benign and malignant nodules. This shift can be represented mathematically as:

$$D_{SMOTE} = D_{original} + \Delta D \quad (2)$$

Where $D_{original}$ represents the decision boundary of the original imbalanced dataset, ΔD is the adjustment due to the synthetic samples, D_{SMOTE} represents the final decision boundary after applying SMOTE [9].

3.4 Progressive Growing Channel Attentive Non-Local (ProCAN) Network for Lung Nodule Classification

The Progressive Growing Channel Attentive Non-Local (ProCAN) network is designed to improve lung nodule classification by progressively refining feature learning while incorporating channel attention and non-local dependencies. Unlike conventional deep learning models that learn all features simultaneously, ProCAN follows a progressive growing strategy where the network structure is expanded dynamically during training, enabling a more stable optimization process and better feature representation for complex nodules. The channel attention mechanism enhances the network's ability to focus on relevant lung regions, while non-local operations capture long-range dependencies that are crucial for detecting subtle malignant patterns in CT scans.

The ProCAN architecture consists of multiple Channel Attentive Non-Local (CAN) blocks, where each block integrates spatial and channel-wise attention mechanisms. A given input feature map $\boldsymbol{X}$ of shape $\boldsymbol{C_{in}} \times \boldsymbol{H} \times \boldsymbol{W}$ (where $\boldsymbol{C_{in}}$ is the number of channels, and $\boldsymbol{H}$, $\boldsymbol{W}$ represent height and width) is reshaped into a matrix form where each pixel is treated as a separate entity. This transformation helps in computing spatial attention across the entire image

$$X \in R^{C_{in} \times H \times W} \rightarrow X \in R^{C_{in} \times N} \quad (3)$$

Where $\mathbf{N} = \mathbf{H} \times \mathbf{W}$ represents the total number of spatial locations. The network then transforms the feature space into three key representations:

$$Q_{j,i} = X_{c,j} M^{q}_{i,c}, \quad K_{i,j} = M^{k}_{i,c} X_{c,j}, \quad V_{i,j} = M^{v}_{i,c} X_{c,j} \quad (4)$$

where **M** matrices are learnable parameters, and Q, K, and V represent query, key, and value embeddings. The spatial attention matrix **B** is computed as:

$$S_{i,j} = Q_{i,z} K_{z,j}, \quad B_{i,j} = \text{Softmax}(S_{i,j}) \quad (5)$$

The attended feature representation is then obtained by multiplying **B** with **V**:

$$A_{c,j} = V_{c,i} B_{i,j} \quad (6)$$

To enhance feature discrimination across channels, ProCAN introduces a channel attention mechanism, computed as:

$$g_c = \text{Sigmoid}\left(X_{c,j} e_j\right), \quad \Phi_{c,n} = A_{c,n} g_c + X_{c,n} \qquad (7)$$

where g_c represents channel-wise attention values that adjust feature importance dynamically. Finally, a **3×3 convolution** with *ReLU* activation is applied:

$$O = \text{ReLU}(\Phi * M_o) \qquad (8)$$

where $\mathbf{M_o}$ represents the convolution filter, refining the extracted features before classification.

ProCAN employs a progressive growing strategy to dynamically increase the number of CAN blocks during training. Instead of introducing new layers abruptly, which can disrupt convergence, a **Bernoulli gating mechanism** is used to gradually integrate additional CAN blocks:

$$f_i(x) = \Omega_i \theta_i\left(f_{i-1}(x)\right) + (1 - \Omega_i) f_{i-1}(x) \qquad (9)$$

where $\boldsymbol{\theta_i}$ is a newly introduced CAN block, $\boldsymbol{\Omega_i}$ is a **2D Bernoulli matrix** controlling feature propagation, and $\mathbf{f_i(x)}$ represents the progressively updated feature extractor. Initially, $\boldsymbol{\Omega_i}$ is set to $\mathbf{0}$, preventing the new block from affecting training. As training progresses, $\boldsymbol{\Omega_i}$ is gradually increased, allowing the network to integrate the new feature extractor without disrupting previously learned representations [10].

For lung nodule classification, ProCAN extracts both **local and global features**, where local features capture fine-grained textures and intensity variations, while global features provide a broader understanding of nodule structure and spatial relationships. By leveraging progressive learning and attention mechanisms, ProCAN improves classification accuracy, particularly for hard-to-distinguish malignant nodules.

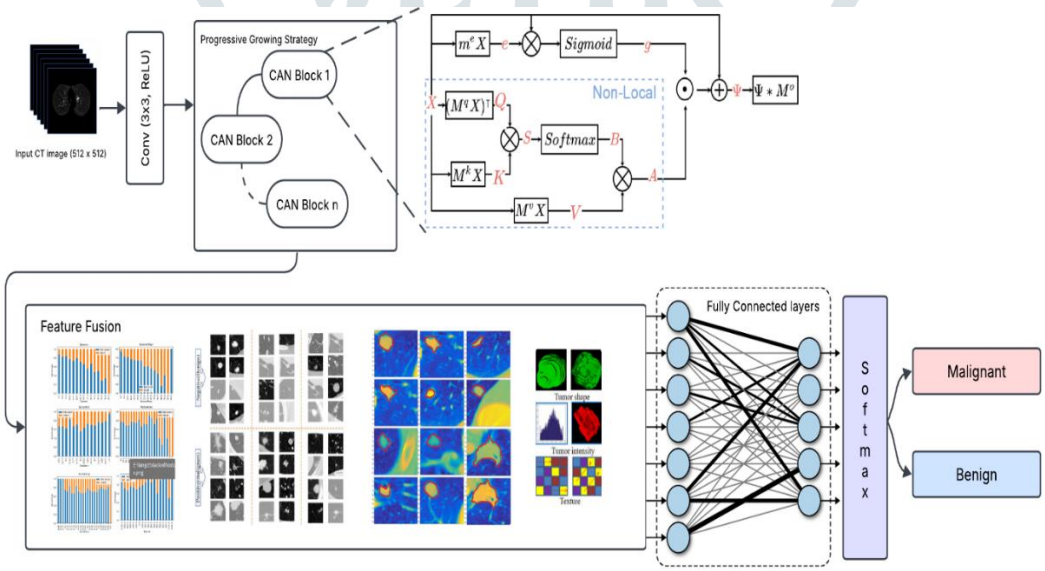

Fig.3. ProCAN Network Architecture for Lung Nodule Classification

4. Experiment Results

4.1 Confusion Matrix

The confusion matrix for the ProCAN model alone indicates an overall classification accuracy of 94.86%. The model correctly identifies 339 benign cases, while 21 benign nodules are misclassified as malignant. Similarly, it accurately classifies 344 malignant samples, but 16 malignant cases are incorrectly predicted as benign. While the model performs effectively in distinguishing between benign and malignant nodules, the relatively higher number of benign misclassifications suggests the presence of a slight class imbalance or borderline cases that make decision boundaries less precise.

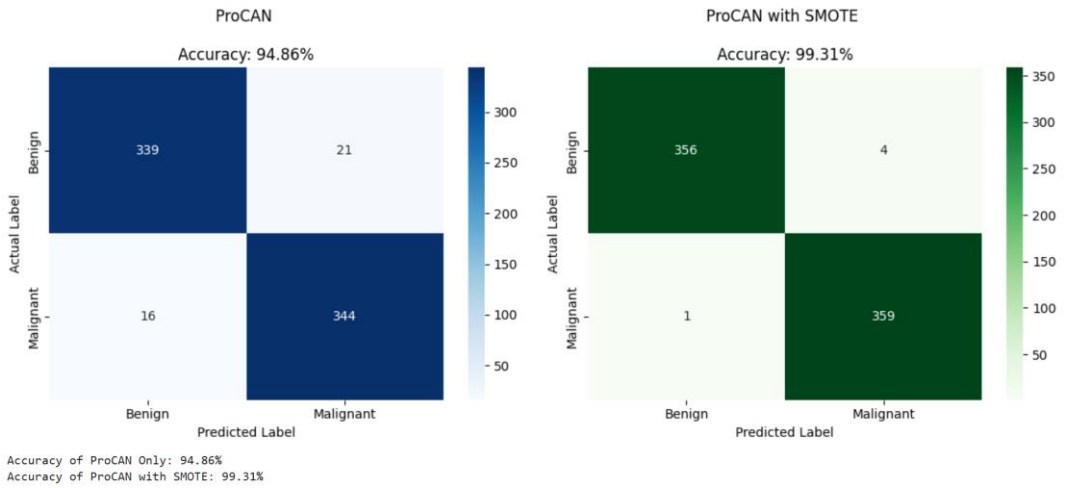

Fig. 4. Confusion Matrices of ProCAN with and without SMOTE

After applying SMOTE to balance the dataset, the model exhibits a significant improvement, achieving an impressive accuracy of 99.31%. The confusion matrix reveals a drastic reduction in misclassification errors, with only 4 benign samples incorrectly classified and just 1 malignant sample misclassified. This improvement highlights the effectiveness of SMOTE in creating a more balanced dataset, leading to a better-defined decision boundary and enhancing the model's ability to generalize across both classes.

4.2 Performance Evaluation

To assess the effectiveness of the Progressive Growing Channel Attentive NonLocal Network (ProCAN) for lung nodule classification, key performance metrics are analysed, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics offer a comprehensive evaluation of the model's capability to differentiate between benign and malignant cases.

Table 4.1 presents the comparative results for the ProCAN model alone and ProCAN with SMOTE. The baseline ProCAN model achieves a high classification accuracy of 94.86%, with an F1-score of 94.84% and an AUC of 0.95, demonstrating strong discriminatory power. By applying SMOTE, the model achieves 99.31% accuracy. The AUC also increases to 0.99, indicating near-perfect separability between benign and malignant cases. These enhancements validate the effectiveness of data balancing techniques in strengthening the model's Performance.

Table 1. Evaluation Metrics for ProCAN and ProCAN with SMOTE

Configuration	Accuracy	Precision	Recall	F1-score
ProCAN	94.86%	94.92%	94.76%	94.84%
ProCAN With SMOTE	99.31%	98.89%	99.72%	99.30%

4.3 Sensitivity and Specificity

Sensitivity and specificity are performance metrics in lung cancer detection, as it measure the model's ability to correctly identify malignant nodules while minimizing false-positive diagnoses. Sensitivity, also referred to as recall, represents the proportion of actual malignant cases that the model correctly identifies. Conversely, specificity measures the proportion of benign cases that are correctly classified, reducing unnecessary interventions and misdiagnoses.

$$Sensitivity = \frac{(TP)}{(TP + FN)} * 100 \quad (10)$$

$$Specificity = \frac{(TN)}{(TN + FP)} * 100 \quad (11)$$

For the baseline ProCAN model, the sensitivity (recall) is 94.76%, indicating that the model effectively detects a majority of malignant nodules. However, the specificity is slightly affected due to class imbalance, leading to 21 benign cases being misclassified as malignant.

After applying SMOTE to balance the dataset, the model achieves a sensitivity of 99.72%, indicating a high ability to detect malignant cases. Specificity also improves, reducing false positives and enhancing classification accuracy. These

results indicate that SMOTE effectively addresses dataset imbalance, improving the model's overall performance in lung nodule classification.

5. Conclusion

This study introduced the Progressive Growing Channel Attentive Non-Local (ProCAN) Network, integrated with Synthetic Minority Over-sampling Technique (SMOTE), for improved lung nodule classification. The results demonstrate that ProCAN effectively captures local and global features using progressive feature learning, channel attention, and non-local operations, contributing to high classification accuracy. However, initial model performance was affected by class imbalance, leading to misclassification of some benign nodules.

The application of SMOTE successfully addressed this limitation by generating synthetic malignant samples, significantly improving classification performance. With SMOTE, the model achieved an accuracy of 99.31%, a sensitivity of 99.72%, and a specificity improvement, minimizing false positives and false negatives. These results highlight the impact of data balancing techniques in enhancing deep learning models for medical image analysis. The findings suggest that the ProCAN model, combined with SMOTE, can serve as a good tool for lung nodule classification, supporting early lung cancer detection.

References

[1] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, R. H. Mak, R. M. Tamimi, C. M. Tempany, C. Swanton, U. Hoffmann, L. H. Schwartz, R. J. Gillies, R. Y. Huang, and H. J. W. L. Aerts, "Artificial intelligence in cancer imaging: Clinical challenges and applications," CA Cancer J. Clin., vol. 69, no. 2, pp. 127–157, Mar. 2019, doi: 10.3322/caac.21552.

[2] Armato III, Samuel G. et al., "Data From LIDC-IDRI." The Cancer Imaging Archive, 2015, doi: 10.7937/K9/TCIA.2015.LO9QL9SX.

[3] Hu, Shiying & Hoffman, Eric & Reinhardt, Joseph. (2001). Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. IEEE transactions on medical imaging. 20. 490-8. 10.1109/42.929615.

[4] Sui, Yuan, Wei, Ying, Zhao, Dazhe, Computer-Aided Lung Nodule Recognition by SVM Classifier Based on Combination of Random Undersampling and SMOTE, Computational and Mathematical Methods in Medicine, 2015, 368674, 13 pages, 2015. https://doi.org/10.1155/2015/368674.

[5] A. Fauzi, R. Supriyadi, and N. Maulidah, "Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest," Jurnal Infortech, vol. 2, no. 1, pp. 96–101, 2020.

[6] Hairani, M. N. Abdillah, and M. Innuddin, "An Expert System for Diagnosis of Rheumatic Disease Types Using Forward Chaining Inference and Certainty Factor Method," in 2019 International Conference on Sustainable Information Engineering and Technology (SIET), 2019, pp. 104–109.

[7] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes," Jurnal Teknologi dan Sistem Komputer, vol. 8, no. 2, pp. 89–93, apr 2020.

[8] A. J. Mohammed, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 3, pp. 3161–3172, jun 2020.

[9] Al-Shabi, Mundher & Shak, Kelvin & Tan, Maxine. (2021). ProCAN: Progressive Growing Channel Attentive Non-Local Network for Lung Nodule Classification. Pattern Recognition. 122. 108309. 10.1016/j.patcog.2021.108309.