Gazi AKADEMİK YAYINCILIK

AIS ARTIFICIAL INTELLIGENCE STUDIES

ARTIFICIAL INTELLIGENCE STUDIES

Use of Machine Learning and Deep Learning Algorithms in the Diagnosis of Thyroid Nodules

Süleyman Menderes[*a,b], **İsmail Şahin**[a,c]

ABSTRACT

This research investigates the effectiveness of machine learning and deep learning algorithms in evaluating thyroid nodules, focusing on their role in supporting diagnostic decisions. Experimental studies assess the classification success of various models, analyzing Positive Predictive Value (PPV) and False Discovery Rate (FDR) across benign, malignant, and normal classes. Model-1 performed well in malignant and normal classes but showed scope for improvement in the benign class. Model-2 enhanced performance in benign and normal classes but still required refinement for the benign class. Notably, Model-3 and Model-4 demonstrated high classification accuracy, achieving superior PPV and low FDR values across all classes, particularly excelling in the benign category. These findings emphasize the potential of deep learning algorithms in enhancing thyroid nodule assessment and improving diagnostic accuracy. However, there remains a need for further advancements, particularly in the benign class. Future research should explore larger datasets and refined methodologies to bolster model performance, ensuring greater reliability in clinical applications. This study underscores the promise of deep learning in transforming thyroid diagnostics while highlighting the need for continued innovation to address existing challenges.

[a,*] Gazi University,
Informatics Institute,
Department of Information Systems,
Tunus Street No: 35 Çankaya/Ankara

[b] Turkcell Ankara Plaza,
Eskisehir Road 9th Km.
No:264 06510 Çankaya / Ankara

[c] Gazi University,
Faculty of Technology,
Department of Industrial Design Engineering,
06560 – Yenimahalle/Ankara/Turkey
ORCID: 0000-0001-8566-3433

[*] Corresponding author.
e-mail: suleyman.menderes@gmail.com

Keywords: Machine Learning, Deep Learning, Ultrasound Image Analysis, Thyroid Nodules, Data Analysis

Submitted: 04.09.2024
Revised: 19.10.2024
Accepted: 08.12.2024

doi: 10.30855/AIS.2024.07.02.03

1. Introduction

This type of cancer ranks seventh in prevalence among malignant endocrine tumors worldwide, while it maintains the same rank among women and ranks fifteenth among men[8]. These data, although they differ by gender, reveal that the disease is an important health problem in the general population[9].

Thyroid cancer diagnosis is often a meticulous and complex process. Most cases occur as thyroid nodules, which are incidentally detected during neck imaging studies[10]. High-resolution ultrasound stands out as one of the most effective methods for the detection and evaluation of thyroid nodules and plays a critical role in the diagnosis of thyroid diseases. Ultrasound images allow radiologists to examine important parameters such as the structure of nodules, echogenicity, presence of calcification, border characteristics and size. In light of this information, radiologists can evaluate the risk of nodules becoming malignant using standard scoring systems. This evaluation process is of great importance for the early diagnosis of thyroid cancer and the planning of correct treatment strategies[11].

Research in this field is carried out with the aim of developing computer-aided diagnostic systems, using neural networks and other machine learning techniques, and improving clinical decisions. In particular, the use of machine learning and deep learning algorithms has the potential to support clinical decisions and increase the effectiveness of the Feinnadel Aspiration (FNA) test[1].

Computer-aided diagnostic (CAD) systems developed for automatic thyroid nodule detection offer revolutionary innovations in medical imaging. Among these methods, support vector machine (SVM) based systems based on extracting image orientation patterns and powerful classifier algorithms such as random forests and support vector machines that evaluate various features such as histogram parameters and fractal dimension stand out[17]. In the field of machine learning, techniques such as Dynamic Mutation-based Glowworm Swarm Optimization (DMGSO) algorithm, Long Short-Term Memory (LSTM) model, and logistic regression have shown remarkable success in the selection of the best features and nodule classification. These methods have significantly improved the efficiency of the processes by increasing the correct identification and classification rates[20]. These advanced techniques play a critical role not only for nodule detection and classification, but also for obtaining more sensitive and reliable results in clinical decision support systems.

Deep learning is a learning method used in the field of artificial intelligence. This technology uses neural networks that mimic the way the human brain works and can recognize complex patterns on large amounts of data. Deep learning involves a series of mathematical operations used to improve and recurse the representatives learned by a system. Deep learning models, which are usually trained on large data sets, can achieve high success in tasks such as image recognition, natural language processing, speech recognition, and similar tasks[15]. Deep learning generally includes special neural network architectures such as convolutional neural networks (CNN) and recurrent neural networks (RNN)[15]. The method we use in this study, Residual Network(ResNet), is a type of CNN. ResNet has an architecture developed specifically to solve the problems encountered in the training of deep neural networks[16].

In recent years, significant successes have been achieved in nodule classification and malignancy prediction with deep learning model training on large data sets in research in this field[12-14]. Deep learning algorithms have strong capabilities in image processing and feature extraction and can effectively analyze data obtained from imaging techniques such as ultrasonography or computed tomography. These models can help doctors determine the benign or malignant probability of thyroid nodules and help determine treatment plans by providing objective and reliable support[1].

However, the integration of these technological developments into clinical practice also brings with it important issues such as ethics, safety and standardization. Machine learning techniques used in the evaluation of thyroid nodules need to be carefully examined and correctly applied in order to improve clinical practice and optimize patient health outcomes. In this context, adopting a balanced approach on how to use technological advances in the field of health is of critical importance in terms of achieving the most effective results for both healthcare professionals and patients.

The aim of this study is to understand the potential of deep learning algorithms in the evaluation of thyroid nodules and to evaluate the effectiveness of these algorithms in clinical applications. The

findings obtained are based on the technology used in medical diagnosis processes.

2. Literature Review

The evaluation of thyroid nodules is of great importance in both medical diagnosis and treatment processes. In recent years, research has revealed the potential of neural networks and other machine learning techniques in developing computer-aided diagnosis systems. This literature review discusses the success and difficulties encountered in the evaluation of thyroid nodules by various neural network models and classifier algorithms.

Many CAD methods have been developed for the automatic detection of thyroid nodules. In 2007 Savelonas and his team developed a CAD system based on radon transformation and SVM, achieving a high classification accuracy of 89% based on data from 66 patients. [17]. In 2011, Ding and his team used SVM to classify thyroid nodules and achieved successful results on a dataset of 125 patients[18]. In 2019, Prochazka and his team achieved a breakthrough in differentiating between malignant and benign nodules by utilizing random forests and support vector machines (SVM). They focused on key image analysis features like histogram parameters, fractal dimension, and mean brightness value, which were instrumental in their successful classification[19].

The year 2020 saw the development of various innovative approaches for the classification of thyroid nodules. Sathyapriya and Anitha used the DMGSO algorithm to select optimal features and employed the LSTM model to classify nodules[20]. In the same year, Ma and his team evaluated five different machine learning methods and tested the performance of these models with four different distance measures[21]. Miao and his team analyzed the variables affecting malignant nodules using logistic regression and worked on ultrasound imaging reporting and data system classification results[22].

In recent years, the potential of deep learning-based CAD systems for thyroid nodule detection has increased. In 2017, Chi and his team extracted features from thyroid ultrasound images using deep convolutional neural networks and transferred these features to a cost-sensitive random forest classifier, achieving high accuracy[24]. In 2019, Ouyang and his team compared non-linear and linear machine learning algorithms and showed that non-linear algorithms are also effective[25]. In the same year, Zhang and his team developed a diagnostic model based on conventional ultrasound and real-time elastography, demonstrating that this model outperformed the diagnostic accuracy of radiologists[26].

Wang and his team, in 2020, showed that deep learning-based methods outperformed radiomics when they compared the performance of the two approaches[27]. In the same year, Song and his team developed a hybrid multi-branch convolutional neural network based on feature cropping[28]. Xie and Yuan combined deep neural networks with traditional features and achieved successful results in classifying nodules[13, 14].

In 2021 and 2022, notable advancements were made in the classification of thyroid nodules. Liu et al. introduced an innovative information fusion technique using a dual-branch convolutional neural network, while Vadhiraj et al. evaluated and compared the performance of support vector machines and artificial neural networks, focusing on metrics such as accuracy, sensitivity, and specificity[29, 30]. Employing various machine learning techniques, Luong et al. in 2022 reached an accuracy of 79.10% in forecasting the malignancy of thyroid nodules with uncertain diagnoses[32].

Xu et al. (2022) proposed C-LSTM, an ultrasound image diagnosis model based on long and short-term memory neural network (LSTM). They obtained an AUC value of 0.86[38].

Jiang et al. in 2022 used the "Attention U-Net" architecture, a deep learning-based approach to classify thyroid nodules. With this method, they achieved a higher accuracy rate than the traditional U-Net model, and this success demonstrated the effectiveness of attention-based networks, especially in small nodule segmentation. In their research, they showed that attention mechanisms have an important contribution in the process of selecting image features[39].

In 2023, Tianlei et al. DSRU-Net achieved an average Intersection coefficient of 85.8%, average dice coefficient of 92.5% and nodule dice coefficient of 94.1% over Union, which increased by 1.8%, 1.3%

and 1.9% compared to U-Net, respectively[40].

In this literature review, we have reviewed many studies that have effectively applied CAD methods, especially neural networks and other machine learning algorithms, to the evaluation of thyroid nodules. Research from 2007 to the present has evaluated the performance metrics of various algorithms and techniques, including accuracy, sensitivity, and specificity, presenting the strengths and weaknesses of each method. While CAD systems appear to have significant potential in the classification of thyroid nodules as malignant or benign, the performance differences between different approaches indicate that further research and improvements are still needed in this area. Deep learning-based methods are expected to outperform traditional methods, and they may become a standard tool in the diagnosis of thyroid nodules in the future. Furthermore, the integration of these systems into clinical decision support systems can contribute to more accurate and consistent diagnosis processes that are free from subjectivity. However, the performance of deep learning-based systems is often dependent on large amounts of medical image data, and thyroid nodule image data is often limited and expensive.

Therefore, in this study, a new method is developed based on image enhancement technology to fully utilize the limited image data information. Four metrics that are important for clinical usability such as Accuracy, PPV and NPV are used as the prominent performance metrics in this study and it is observed that the proposed method outperforms the existing thyroid nodule diagnosis systems in these metrics.

3. Methodology

This study aims to develop an image analysis model for the evaluation of thyroid nodules using machine and deep learning algorithms. First, we focus on the FNA procedure, which is widely used in the evaluation of thyroid nodules. FNA is a preferred method, especially for nodules over 2 cm, even if the suspicion of malignancy is low[1]. However, there are risks and costs associated with this procedure. Therefore, this study aims to develop a model for evaluating FNA results and predicting the benign or malignant status of nodules using deep learning algorithms[2].

The use of machine and deep learning algorithms has led to significant advances in the field of medical image analysis, especially in recent years. These algorithms are known for their ability to recognize and learn complex patterns when trained on large data sets. The process of analyzing and classifying thyroid nodules using deep learning algorithms is illustrated in Figure 1. This approach leverages advanced computational methods to enhance accuracy and efficiency in medical diagnostics.

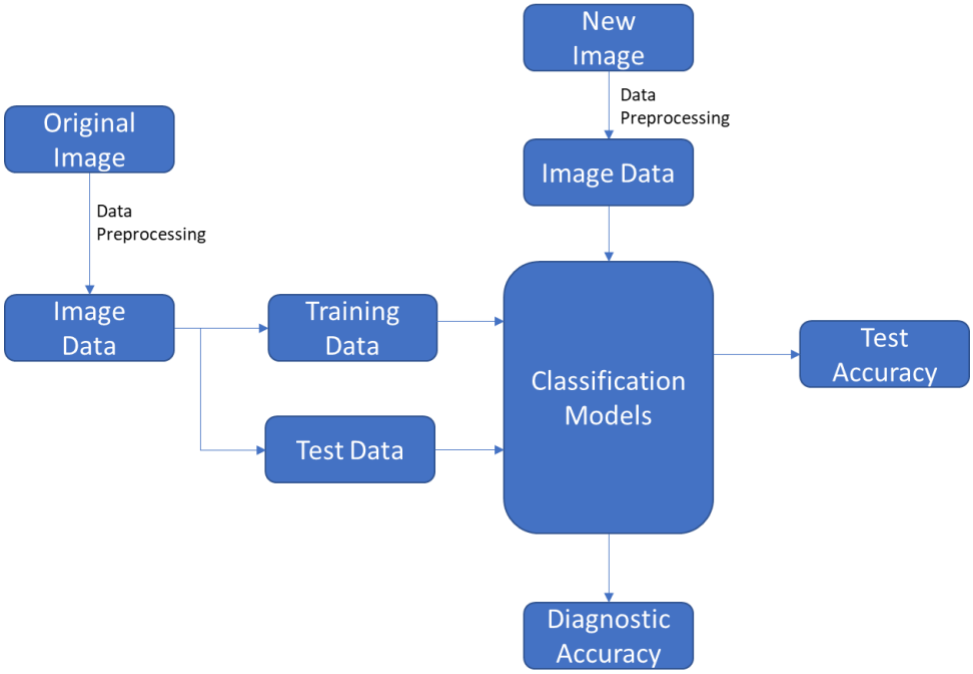

Figure 1. Thyroid nodule processing classification process in deep learning algorithms

In the classification process shown in Figure 1, the image data obtained from the original image is used for training and testing the classification models. In this process, some of the image data is utilized as training data, while the remaining image data is employed as test data. The training data is used for learning the classification models, while the test data is used to evaluate the performance of the models. After the training process, the classification models are run on the test data and the diagnostic accuracy of these models is measured. When a new image is obtained, the data extracted from this image is also input into the classification models and the diagnostic accuracy of the models on this new data is evaluated. This process focuses on training and testing the classification models and improving their ability to make accurate diagnoses on new images.

3.1. Dataset

In this study, an open access dataset called DDTI (Digital Database of Thyroid Ultrasound Images) was used in the examination of thyroid nodules. This valuable dataset was created in collaboration with Universidad Nacional de Colombia, CIM@LAB and IDIME (Instituto de Diagnostico Medico) and provides an important resource for scientific research. The main purpose of the DDTI dataset is to create algorithms for the development of CAD systems in the evaluation of thyroid nodules. In addition, the dataset is designed as an effective tool for the education and training of new radiologists[37].

The DDTI dataset contains 375 cases and 440 images, and each case is presented with an XML file containing expert markings and patient information. The 375 cases examined in this study were divided into three groups according to the TIRADS classification. This dataset is expected to help us better understand thyroid nodules and improve diagnostic processes[7].

Benign: Cases in this group include conditions that indicate that the thyroid nodules are benign. Benign nodules are usually harmless and have a low likelihood of cancer. Analysis of cases in this category is important for understanding positive thyroid health outcomes.

Malignant (Suspicious Feature): Cases in this group include conditions that indicate that thyroid nodules have suspicious features. Nodules with suspicious features indicate a higher likelihood of cancer and may require further evaluation or follow-up. Examination of these cases is critical for early detection and treatment.

Normal: This category includes cases where thyroid nodules do not show any suspicious or malignant features. That is, thyroid nodules are classified as normal and healthy.

Table 1. TIRADS Classification

TIRADS	Interpretation
1	Normal thyroid gland
2	Benign lesion
3	Probably benign lesion
4 a,b,c	Suspicious of malignancy
5	Probably malignant(>80% risk)
6	Biopsy proven malignancy

TIRADS classification is a very important and widespread system used in the evaluation of thyroid nodules. This classification allows nodules to be categorized according to their different characteristics, helping doctors make more accurate diagnoses and patients avoid unnecessary procedures. In particular, TIRADS, which plays a critical role in determining the malignancy risk of nodules, provides valuable guidance to clinical practice. In this way, the most appropriate treatment plan for patients can be determined while avoiding unnecessary biopsies and surgical interventions. (Table 1). This classification is an important tool used in clinical applications and in the development of CAD systems. The distribution of the images in the dataset into these groups is given in Table 2 below.

Table 2. Distribution of images according to classification in the dataset

Classification	Number of Images in the Data Set
Benign	59
Malignant	261
Normal	100

3.2. Data Preprocessing

CNN based models usually expect a certain input size and frame size. In accordance with the Inception-ResNet-v2 model that we will use in this study, we resized the images in our dataset as (224x224) frames[16, 33].

3.3. Data Augmentation

In order to effectively compare the results of our developed learning model, two distinct methods were implemented: one with data augmentation techniques and one without. The techniques applied included modifications in brightness, contrast, saturation, and hue. These data augmentation techniques are particularly valuable for enhancing model performance by effectively increasing the variety and size of training datasets. Since deep learning models benefit significantly from large datasets, data augmentation provides an alternative to data collection by transforming a limited dataset into a more extensive, varied one, improving the model's ability to generalize.

In our study, given the constraints of a limited dataset, data augmentation played a crucial role. By creating new data through deformation operations, these techniques enabled a more comprehensive training process, allowing our model to learn from a more diverse dataset. The augmented data proved instrumental in enhancing model performance by reducing overfitting and improving accuracy, particularly by simulating various real-world conditions that the model might encounter. This aligns with the principle that expanded datasets provide richer information for deep learning architectures, ultimately leading to better outcomes with minimal data collection efforts[34].

4. Experimental Setup

Neural network training was performed to classify benign and malignant nodules using preprocessed US images. In this context, the most popular image recognition model Inception-ResNet-v2 was used by applying a transfer learning method.
Inception-ResNet-v2 is a deep learning model that combines the Inception and Residual Network (ResNet) architectures. The Inception architecture has parallel convolution layers that allow processing image features with filters of different sizes, thus enabling the model to have a wider information coverage. Inception-ResNet-v2, which also incorporates the “skip connection” structure, a core component of ResNet, prevents the loss of information from the deep layers, enabling faster and more effective learning.
With the combination of these two powerful architectures, Inception-ResNet-v2 can learn richer features from data and at the same time work with a deeper network structure. The model has been tested on large datasets such as ImageNet and outperforms many other deep learning models in terms of accuracy. In particular, it stands out for its lower computational costs despite having more complex architectures. This makes the model effectively usable even on devices with limited resources, making it a preferred choice for a variety of applications.
ResNet, introduced by Kaiming He and colleagues in 2015, represents a significant advancement in CNNs. Despite being eight times deeper than VGGNet[35], ResNet manages to maintain a lower computational cost[16]. This innovative architecture led to a remarkable achievement in the 2015 ImageNet competition, where it achieved an impressive error rate of just 3.57%[36].
ResNet features an architecture specifically engineered to enhance both the training process and overall performance of very deep neural networks[16]. In traditional neural networks, performance issues can arise as they get deeper because training deeper models can become difficult. ResNet solves this problem by adding layers that cross connections. This is done by adding a connection to the output of the previous layer, instead of the network forwarding the output of one layer directly. These “skip connections” allow the gradient to be transmitted more efficiently during backward traversal and make deep networks easier to learn. In this way, ResNet makes it possible to train much deeper and more efficient neural networks.
Two different models, ResNet101v2 and ResNet151v2, were implemented with different hyperparameters. In the studies, the data set was randomly divided into 80% training data and 20% test data.

4.1. Experimental Setup

In machine learning, evaluating the success of a model is done through performance measurement. This step is critical to understanding how effective the model is because only a well-performing model can produce successful results in the real world. Performance measurement provides the feedback needed to improve the model and minimize errors. Therefore, evaluating a model is an essential part of the process. As these performance metrics, 4 metrics are considered: Accuracy, PPV, NPV and FDR. Setting PPV, FDR and NPV as the main criteria instead of accuracy rate to evaluate the performance of the models evaluated on health data reflects the aim of measuring accurate diagnosis and prediction capabilities in the health field. In this context, in order to understand and evaluate the results obtained, the PPV value obtained from the Confusion Matrix indicates the ratio of positive predictions to true positives, while the FDR value indicates the ratio of false positive predictions to all positive predictions. NPV, on the other hand, stands out as valuable metrics that measure the success of negative predictions in accurately predicting true negative situations. These metrics were employed to further assess the model's performance on health data and to gauge the clinical significance of the results.

$$Accuracy_i = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \quad (1)$$

$$PPV_i = \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$FDR_i = \frac{FP_i}{FP_i + TP_i} \text{ or } FDR_i = 1 - PPV_i \quad (3)$$

$$NPV_i = \frac{TN_i}{TN_i + FN_i} \quad (4)$$

Accuracy(eq. 1) is a measure of how accurately a model classifies thyroid nodules. In other words, accuracy is the proportion of cases that the model correctly predicts out of all cases. However, accuracy alone may not always be a sufficient measure of performance. For example, if there are many cases without nodules, the model may achieve a high accuracy rate by correctly predicting these cases. In this case, evaluation with other performance metrics is necessary to better understand the true performance of the model. PPV (eq. 2) and NPV (eq. 3) are the ratio of true positive and true negative results, positive and negative results respectively.

The results obtained with different deep learning algorithms used in this study are given in Table 3. These results were achieved with Intel Core i7, 2.0 GHz CPU, 16 GB RAM and NVDIA GeForce GTX 1050M 4GB DDR5 GPU, Windows® 10 operating system. Uses the Python (3.11) programming language. Keras(3.0) and Tensorflow(2.15) libraries were used for the creation, training and evaluation of deep learning models (Resnet).

Table 3. Results obtained with different deep learning algorithms

Name	Model	Learning Rate	Optimization Algorithm	Data Augmentation	Accuracy
Model-1	ResNet151v2	0.0006	Adam	No	90%
Model-2	ResNet101v2	0.0001	Adam	No	85%
Model-3	ResNet151v2	0.0006	Adam	Yes	91%
Model-4	ResNet101v2	0.0001	Adam	Yes	92%

The Learning Rate information shared in the chart is a hyperparameter used to determine the updated parameter values of a machine learning model. This hyperparameter controls how many “steps to take” at each update step during the training of a model. That is, the learning rate determines the size of the update steps to minimize the loss function of the model.

The ResNet deep learning architecture generally uses Adam (Adaptive Moment Estimation) as an optimization algorithm for training deep learning models. Adam tries to combine the advantages of other popular optimization algorithms and combines momentum-based methods with adaptive learning speed.

Considering the results in Table 3, Adam was used as the optimization algorithm in all studies. The experimental results in the table can be expressed as follows. When the data augmentation techniques mentioned in section “4.3. Data Augmentation” are applied to the same models with the same hyperparameters, the effect on the results is measured. As can be seen in the table, while data augmentation techniques were not used in experimental studies 1 and 2, they were used in experimental studies 3 and 4. It was observed that accuracy values increased positively in both models.

The tables below present the accuracy performance of the four models in the benign, malignant and normal classes and the associated PPV, FDR and NPV values. Comments on the performance of each model are given immediately after the results.

Table 4. Performance Results for Model-1, Model-2, Model-3 and Model-4

Model	Confusion Matrix	PPV(%)	FDR(%)	NPV(%)
Model-1	Benign: (3, 6, 0)	Benign: 33	Benign: 67	Benign: 75
	Malignant: (1, 41, 0)	Malignant: 97.6	Malignant: 2.4	Malignant: 100
	Normal: (0, 0, 19)	Normal: 100	Normal: 0	Normal: 100
Model-2	Benign: (4, 5, 0)	Benign: 44.4	Benign: 65.6	Benign: 80
	Malignant: (1, 40, 1)	Malignant: 97.6	Malignant: 2.4	Malignant: 95.2
	Normal: (1, 2, 16)	Normal: 88.9	Normal: 11.1	Normal: 94.1
Model-3	Benign: (48, 10, 6)	Benign: 82.7	Benign: 17.3	Benign: 100
	Malignant: (0, 269, 8)	Malignant: 97.1	Malignant: 2.9	Malignant: 91.5
	Normal: (0, 25, 98)	Normal: 79.7	Normal: 20.3	Normal: 94.2
Model-4	Benign: (54, 8, 2)	Benign: 87.1	Benign: 12.9	Benign: 85.7
	Malignant: (9, 266, 2)	Malignant: 96.7	Malignant: 3.3	Malignant: 99.25
	Normal: (3, 19, 101)	Normal: 98.3	Normal: 1.7	Normal: 84.1

Model-1 shows a low PPV in benign classes, but has a very good accuracy in malignant and normal classes.
While Model-2 achieves higher PPV values in the benign and normal classes, its accuracy in the malignant class is similar to Model-1.
Model-3 shows high accuracy in benign and malignant classes, while it achieves lower accuracy in the normal class.
Model-4 stands out as the model with the best overall performance, offering high accuracy rates in benign and normal classes.

Figure 2 and Figure 3 show the images obtained from the images separated as test data of Model-4 given in Table 3 and correctly classified as benign. Figure 4 and Figure 5 show the images obtained from the images separated as test data of Model-4 and correctly classified as malignant. The images in all these figures were randomly selected from among 54 images from the benign class and 266 images from the malignant class that Model-4 successfully predicted. While the PPV value for the benign class was 87,6, the model proved its success with a high prediction ability of 96,7 for the malignant class.

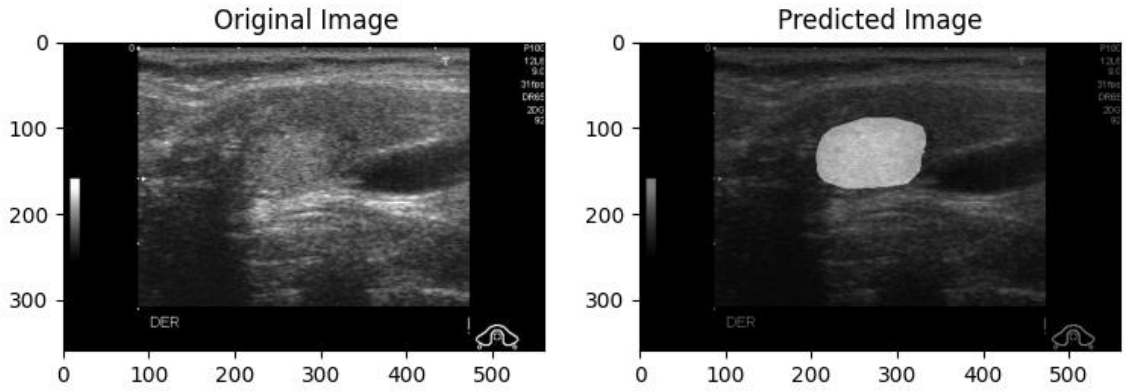

Figure 2. According to Model-4 results, the correctly predicted example from the benign class is Image-1

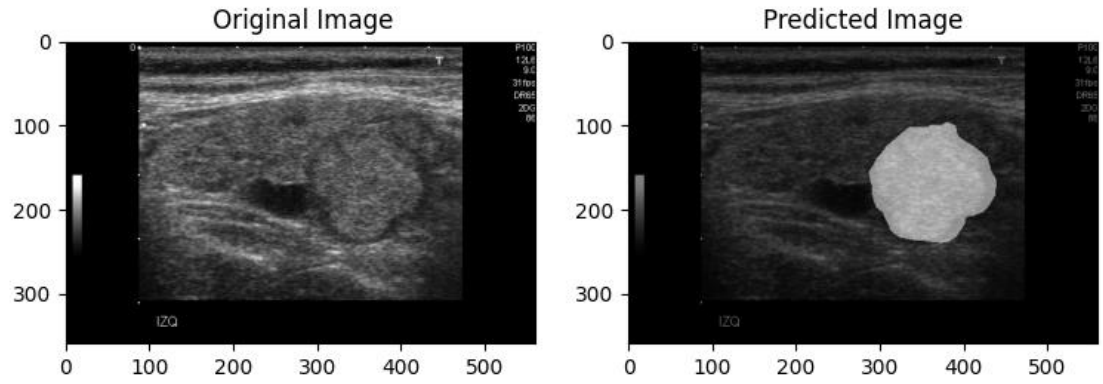

Figure 3. According to Model-4 results, the correctly predicted example from the benign class is Image-2

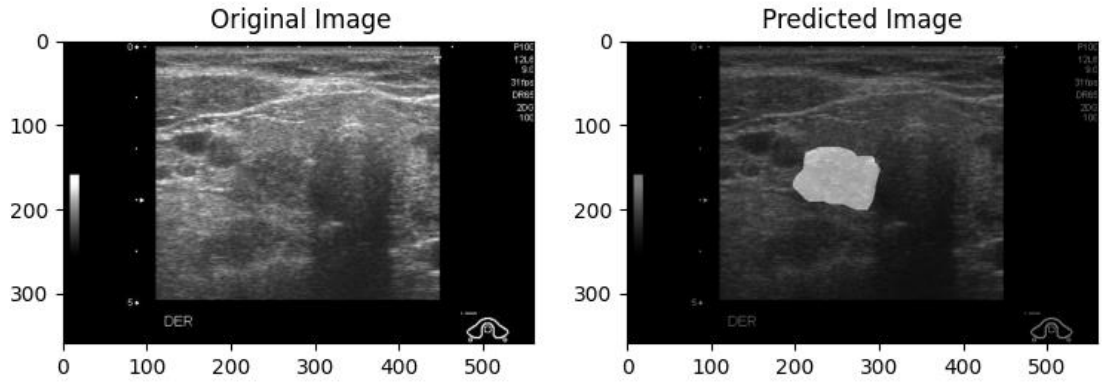

Figure 4. According to Model-4 results, the correctly predicted example from the malignant class is Image-1

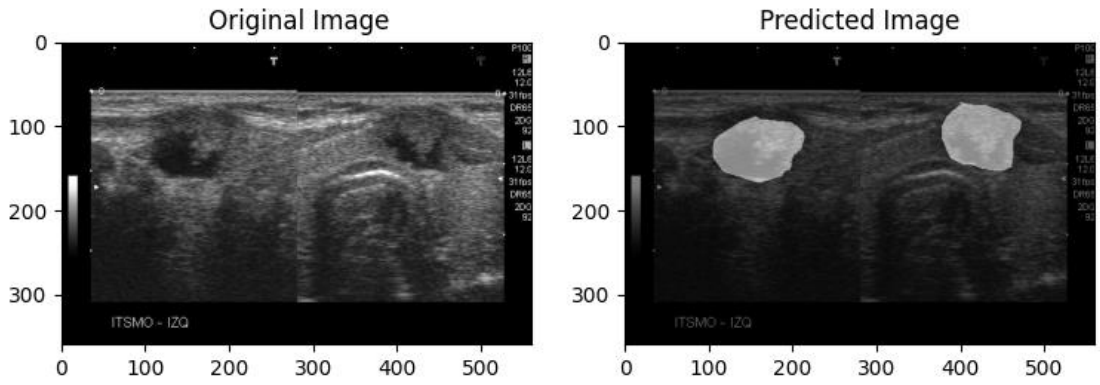

Figure 5. According to Model-4 results, the correctly predicted example from the malignant class is Image-2

The images in all these figures are presented by combining the original image and the mask image created by the prediction ability of the model. This situation tries to increase the success of the model we obtained and at the same time, it helps clinicians who perform the examination by marking the relevant regions in diagnosis and allows them to save more time and optimize the treatment processes.

4.2. Discusssion of Experimental Setup

In this study, the performance of four different deep learning models evaluated on thyroid nodules data was comprehensively investigated. Among the criteria used to evaluate model performance, PPV, FDR and NPV stand out. These metrics reflect the aim of measuring accurate diagnosis and prediction

capabilities in the health field and are supported by evaluations obtained through the Confusion Matrix.

When the results obtained for the first model (Model-1) are examined, the high PPV value (97,6%) and low FDR value (2,4%) in the Malignant class are remarkable. This shows that the model correctly classifies malignant lesions and false positive predictions are at a minimum level. In the Benign class, the PPV value (33%) is low, while the FDR value (67%) is high. This shows that benign lesions are predicted with lower accuracy by the model.

For the second model (Model-2), high PPV (97,6%) and low FDR (2,4%) values were observed in the Malignant class, while PPV (44,4%) and FDR (65,6%) values in the Benign class improved compared to the previous model. This shows that the second model classifies benign lesions more accurately.

For the third model (Model-3), high PPV (97,1%) and low FDR (2,9%) values were maintained in the Malignant class. In the Benign class, PPV (82,7%) and FDR (17,3%) values show that the model classifies benign lesions more effectively.

For the last model (Model-4), high PPV (96,7%) and low FDR (3,3%) values in the Malignant class continued in a similar manner. In the Benign class, PPV (87,1%) and FDR (12,9%) values show that the model classifies benign lesions with higher accuracy. When NPV values are examined, it is seen that the models generally achieve high accuracy in negative predictions, highlighting the ability of the models to correctly identify healthy situations.

When the obtained results are evaluated, it is seen that deep learning algorithms, especially CNN, provide high accuracy rates in the diagnosis of thyroid nodules. However, low accuracy rates in the benign class reveal that more data and more sophisticated algorithms are needed to correctly classify these classes. In addition, the limited data sets used limit the generalization ability of the models, and this situation emphasizes the need for larger and more diverse data sets. At this point, it is seen that data augmentation techniques contribute to performance improvements, but may not be sufficient to obtain more comprehensive results in larger-scale studies. Therefore, expanding the data sets and training the models on these larger data sets in future studies may provide more reliable and generalizable results.
As a result, when the performance of four different models is examined, it is seen that each of them achieves high accuracy in certain classes but has potential for improvement in some classes. These findings provide an important basis for a detailed evaluation of model performance on health data and for understanding clinical significance.

5. Conclusions

This study examines the performance of machine and deep learning algorithms used in the evaluation of thyroid nodules, particularly evaluating their potential to support diagnostic decisions. The experimental studies conducted evaluate the success of various models in classifying thyroid nodules.

The obtained results include a detailed analysis based on PPV and FDR values performed on benign, malignant and normal classes. Although Model 1 has potential for improvement in the benign class, it exhibited high performance in the malignant and normal classes. Model 2 increased the success in the benign and normal classes, but there are still areas for improvement in the benign class. In particular, Model 3 and Model 4 achieved high PPV values in all three classes, demonstrating successful classification performance. These models are particularly notable for their high PPV and low FDR values in the benign class.

The findings reveal that deep learning algorithms hold considerable promise for assessing thyroid nodules. However, the need for improvement observed in the benign class highlights the need for further work and development in this area. Future research should focus on the use of larger datasets and advanced methodologies to improve model performance and obtain more reliable results in clinical applications. In addition, the use of diverse and correctly labeled datasets is recommended to strengthen the generalization ability of existing models. Such studies may contribute to the development of a potential tool that can be used for rapid and accurate evaluation of thyroid nodules, thereby reducing the risks and costs resulting from unnecessary FNA procedures.

In the future, it is anticipated that the use of machine learning and deep learning technologies in the diagnosis and evaluation of thyroid nodules will continue to increase. In particular, it is thought that by training deep learning algorithms on larger and more diverse datasets, diagnostic accuracy and overall performance can be further improved. In addition, the integration of artificial intelligence and machine learning models into clinical applications will allow physicians to use decision support systems more effectively.

In future studies, the performance of the model can be improved, and its generalization ability can be strengthened by using a larger variety of correctly labelled data sets. The managerial implications of these studies and future research can contribute to the development of a potential tool that can be used for rapid and accurate assessment of thyroid nodules. This can provide faster diagnosis and treatment planning, reducing the risks and costs resulting from unnecessary FNA procedures. Finally, the regulation and standardization processes of AI-based diagnostic systems will need to be improved. The establishment of legal and ethical frameworks for the safe and effective use of these technologies in clinical applications will be one of the most important trends in the future. In this direction, the importance of multidisciplinary studies and collaborations will increase.

References

[1] J. Song, Y. J. Chai, H. Masuoka, S.-W. Park, S. Kim, J. Y. Choi, H.-J. Kong, K. E. Lee, J. Lee, N. Kwak, K. H. Yi, and A. Miyauchi, "Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules," *Medicine (Baltimore)*, vol. 98, no. 15, Apr. 2019. doi:10.1097/MD.0000000000015133.

[2] C. Durante, G. Grani, L. Lamartina, S. Filetti, S. J. Mandel, and D. S. Cooper, "The diagnosis and management of thyroid cancer," *JAMA*, vol. 319, no. 9, pp. 914–924, Mar. 2018. doi:10.1001/jama.2018.0898.

[3] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," *Expert Systems with Applications*, vol. 36, no. 1, pp. 944–949, Jan. 2009. doi:10.1016/j.eswa.2007.10.010.

[4] K. Salman and E. Sonuç, "Thyroid disease classification using machine learning algorithms," in *Proc. of the 2021 Int. Conf. on Physics and its Applications (ICPA)*, Jan. 2021, Istanbul, Turkey [Online]. Available: IOPScience, doi:10.1088/1742-6596/1963/1/012140. [Accessed: Nov. 27, 2024].

[5] Q. T. Nguyen, E. J. Lee, M. G. Huang, Y. I. Park, A. Khullar, and R. A. Plodkowski, "Diagnosis and treatment of patients with thyroid cancer," *American Health and Drug Benefits*, vol. 8, no. 1, pp. 30–40, 2015. Available: PubMed.

[6] H. A. Ur Rehman, C.-Y. Lin, Z. Mushtaq, and S.-F. Su, "Performance analysis of machine learning algorithms for thyroid disease," *Arabian Journal for Science and Engineering*, vol. 46, pp. 5337–5349, Aug. 2021. doi:10.1007/s13369-020-05206-x.

[7] "TIRADS for Thyroid," *ACE Imaging*. [Online]. Available: https://www.aceimaging.in/tirads-for-thyroid/.

[8] B. R. Haugen, E. K. Alexander, K. C. Bible, et al., "American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer," *Thyroid*, vol. 26, no. 1, pp. 1–133, Jan. 2016. doi:10.1089/thy.2015.0020.

[9] S. Vaccarella, S. Franceschi, F. Bray, C. P. Wild, M. Plummer, and L. Dal Maso, "Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis," *New England Journal of Medicine*, vol. 375, no. 7, pp. 614–617, Aug. 2016. doi:10.1056/nejmp1604412.

[10] Y. R. Hong, C. X. Yan, G. Q. Mo, et al., "Conventional US, elastography, and contrast-enhanced US features of papillary thyroid microcarcinoma predict central compartment lymph node metastases," *Scientific Reports*, vol. 5, art. 7748, Jan. 2015. doi:10.1038/srep07748.

[11] E. J. Ha and J. H. Baek, "Applications of machine learning and deep learning to thyroid imaging," *Ultrasonography*, vol. 39, no. 4, pp. 357–369, Oct. 2020. doi:10.14366/usg.20068.

[12] H. Zhang, C. Zhao, L. Guo, et al., "Diagnosis of thyroid nodules in ultrasound images using two combined classification modules," in *Proc. of the 2019 Int. Conf. on Image and Signal Processing (CISP-BMEI)*, Suzhou, China, Oct. 2019, pp. 1234–1239. doi:10.1109/CISP-BMEI48845.2019.8965903.

[13] J. Xie, L. Guo, C. Zhao, X. Li, Y. Luo, and J. Liu, "A hybrid deep learning and handcrafted features-based approach for thyroid nodule classification in ultrasound images," *Journal of Physics: Conference Series*, vol. 1693, no. 1, p. 012160, Dec. 2020. doi:10.1088/1742-6596/1693/1/012160.

[14] H. Yuan, "Thyroid nodule classification in ultrasound images by fusion of conventional features and res-GAN deep features," *International Journal of Biomedical Imaging*, vol. 2021, art. 9917538, 2021. doi:10.1155/2021/9917538.

[15] "Deep Learning Foundations," *Dive into Deep Learning*. [Online]. Available: https://d2l.ai/chapter_introduction/index.html.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, Jun. 2016, pp. 770–778. doi:10.48550/arXiv.1512.03385.

[17] M. A. Savelonas, D. K. Iakovidis, N. Dimitropoulos, and D. Maroulis, "Computational characterization of thyroid tissue in the radon domain," in *Proc. of the 2007 IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*, Maribor, Slovenia, Jun. 2007, pp. 179–184. doi:10.1109/cbms.2007.33.

[18] J. Ding, H. Cheng, C. Ning, J. Huang, and Y. Zhang, "Quantitative measurement for thyroid cancer characterization based on elastography," *Journal of Ultrasound in Medicine*, vol. 30, no. 9, pp. 1259–1265, Sep. 2011. doi:10.7863/jum.2011.30.9.1259.

[19] A. Prochazka, S. Gulati, S. Holinka, and D. Smutek, "Classification of thyroid nodules in ultrasound images using direction-independent features extracted by two-threshold binary decomposition," *Technology in Cancer Research & Treatment*, vol. 18, pp. 1–12, Mar. 2019. doi:10.1177/1533033819830748.

[20] S. S. Anitha and D. Dynamic, "Mutation-based glowworm swarm optimization with long short-term memory approaches for thyroid nodule classification," *Indian Journal of Science and Technology*, vol. 13, no. 14, pp. 1805–1812, Apr. 2020. doi:10.17485/ijst/v13i14.38.

[21] X. Ma, B. Xi, Y. Zhang, et al., "A machine learning-based diagnosis of thyroid cancer using thyroid nodules ultrasound images," *Current Cancer Drug Targets*, vol. 20, no. 10, pp. 872–882, Oct. 2020. doi:10.2174/1574893614666191017091959.

[22] S. Miao, M. Jing, R. Sheng, et al., "The analysis of differential diagnosis of benign and malignant thyroid nodules based on ultrasound reports," *Gland Surgery*, vol. 9, no. 3, pp. 314–321, Jun. 2020. doi:10.21037/gs.2020.04.03.

[23] G. Ataide, N. Ponugoti, A. Illanes, A. Schenke, M. Kreissl, and M. Friebe, "Thyroid nodule classification for physician decision support using machine learning-evaluated geometric and morphological features," *Sensors*, vol. 20, no. 21, p. 6110, Oct. 2020. doi:10.3390/s20216110.

[24] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural networks," *Journal of Digital Imaging*, vol. 30, no. 4, pp. 477–486, Aug. 2017. doi:10.1007/s10278-017-9997-y.

[25] F. S. Ouyang, B. L. Guo, L. Z. Ouyang, et al., "Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules," *European Journal of Radiology*, vol. 110, pp. 81–86, Feb. 2019. doi:10.1016/j.ejrad.2019.02.029.

[26] B. Zhang, J. Tian, S. Pei, et al., "Machine learning-assisted system for thyroid nodule diagnosis," *Thyroid*, vol. 28, no. 8, pp. 1035–1042, Aug. 2018. doi:10.1089/thy.2018.0380.

[27] Y. Wang, W. Yue, X. Li, et al., "Comparison study of radiomics and deep learning-based methods for thyroid nodules classification using ultrasound images," *IEEE Access*, vol. 8, pp. 126896–126906, Apr. 2020. doi:10.1109/access.2020.2980290.

[28] R. Song, L. Zhang, C. Zhu, J. Liu, J. Yang, and T. Zhang, "Thyroid nodule ultrasound image classification through hybrid feature cropping network," *IEEE Access*, vol. 8, pp. 212407–212418, Apr. 2020. doi:10.1109/access.2020.2982767.

[29] Z. Liu, S. Zhong, Q. Liu, et al., "Thyroid nodule recognition using a joint convolutional neural network with information fusion of ultrasound images and radiofrequency data," *European Radiology*, vol. 31, pp. 1426–1436, Jan. 2021. doi:10.1007/s00330-020-07585-z.

[30] V. V. Vadhiraj, A. Simpkin, J. O'Connell, N. Singh Ospina, S. Maraka, and D. T. O'Keeffe, "Ultrasound image classification of thyroid nodules using machine learning techniques," *Medicina*, vol. 57, no. 6, art. 527, Jun. 2021. doi:10.3390/medicina57060527.

[31] W. Li, S. Cheng, K. Qian, K. Yue, and H. Liu, "Automatic recognition and classification system of thyroid nodules in CT images based on CNN," *International Journal of Biomedical Imaging*, vol. 2021, art. 5540186, 2021. doi:10.1155/2021/5540186.

[32] G. Luong, A. J. Idarraga, V. Hsiao, and D. F. Schneider, "Risk stratifying indeterminate thyroid nodules with machine learning," *Journal of Surgical Research*, vol. 267, pp. 23–30, Dec. 2021. doi:10.1016/j.jss.2021.09.015.

[33] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017. doi:10.1145/3065386.

[34] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, Mar. 2017. doi:10.1109/LSP.2017.2657381.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, Sep. 2014. doi:10.48550/arXiv.1409.1556.

[36] "ImageNet Large Scale Visual Recognition Challenge Results," *ImageNet*. [Online]. Available: https://image-net.org/challenges/LSVRC/2015/results.

[37] "DDTI Thyroid Ultrasound Images Dataset," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/dasmehdixtr/ddti-thyroid-ultrasound-images/data.

[38] P. Xu, Z. Du, L. Sun, Y. Zhang, J. Zhang, and Q. Qiu, "Diagnostic value of contrast-enhanced ultrasound image features under deep learning in benign and malignant thyroid lesions," *Journal of Healthcare Engineering*, vol. 2022, art. 6786966, 2022. doi:10.1155/2022/6786966.

[39] Z. Hong, H. Xi, W. Hu, Q. Wang, J. Wang, and L. Luo, "Multi-attentional U-Net for medical image segmentation," in *Proc. of the 2022 IEEE International Symposium on Artificial Intelligence and Applications (ISAIAM)*, Chongqing, China, Jan. 2022, pp. 33–40. doi:10.1109/ISAIAM55748.2022.00033.

[40] T. Zheng, H. Qin, Y. Cui, R. Wang, W. Zhao, S. Zhang, S. Geng, and L. Zhao, "Segmentation of thyroid glands and nodules in ultrasound images using the improved U-Net architecture," *BMC Medical Imaging*, vol. 23, art. 11, Jan. 2023. doi:10.1186/s12880-023-01011-8.