International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Thyroid Cancer Diagnosis with Machine Learning: A Multimodal Ensemble Approach for Clinical Decision Support

Sujithra Sankar*[1], S. Sathyalakshmi[2]

Submitted: 16/09/2023 **Revised**: 17/11/2023 **Accepted**: 29/11/2023

Abstract: Despite ongoing research, the current diagnostic methods of thyroid cancer may still have limitations, leading to potential errors in determining the malignancy of thyroid nodules. To address these challenges, this research introduces a cutting-edge multimodal thyroid cancer diagnosis framework that integrates data from multiple sources, including both ultrasound images and its clinical data. To carry out the experiments, the researchers utilised the Thyroid Nodule Ultrasound Images Dataset (TDID), an open-access public dataset. The research is divided into two phases. In phase I, a wide range of machine learning and deep learning models are employed to create thyroid cancer prediction models using thyroid US image and the corresponding clinical data. The models that consistently demonstrated high prediction accuracy are selected for further consideration. Phase II focuses on creating ensemble combinations of models to perform multimodal prediction of thyroid nodules. The results of the experiments are highly promising, with certain ensemble combinations achieving impressive accuracy. A web app interface, ThyroPredict, has been developed, to generate a final prediction from the uploaded ultrasound image and initial radiologist's input. ThyroPredict app is powered by multimodal ensemble framework.

Keywords: *Deep Learning, Machine Learning, Ensemble Methods, Thyroid ultrasound, Thyroid cancer diagnosis*

1. Introduction

With the fast-approaching end of the first quarter of the twenty-first century, the healthcare industry is undergoing transformative changes fueled by the application of Artificial Intelligence (AI) technology. This technological advancement has ushered in a new era of disease diagnosis, where the analysis of big data using AI algorithms has shown great promise in improving the accuracy and efficiency of diagnostic processes. Within this context, recent research conducted on the female population in India has shed light on a high prevalence of thyroid disorders among women of reproductive age. These studies have indicated a notable increase in the incidence of thyroid diseases with age in India.

The focus of this paper is on thyroid cancer, a form of cancer that affects the thyroid gland, a small, butterfly-shaped organ responsible for producing hormones that regulate various body functions. The disruption in the balance of these hormones can lead to thyroid diseases, making accurate diagnosis crucial for effective treatment and management. In the United States, the diagnosis of thyroid cancer has been on the rise, largely attributed to the widespread adoption of thyroid ultrasound, enabling the detection of small thyroid nodules that might have previously gone unnoticed. Moreover, women are found to be four times more likely to be affected by thyroid cancer com-pared to men, as suggested by recent studies. Thyroid cancer is a significant global health concern, affecting individuals of all ages and genders. Achieving accurate diagnosis in a timely manner is of utmost importance to provide appropriate treatment and management. However, diagnosing thyroid cancer can be challenging due to the complexity of the disease, the various imaging modalities used, and the heterogeneity of the patient population. To address these challenges, this paper proposes a novel approach for multimodal thyroid cancer diagnosis, leveraging deep learning and machine learning ensembles.

To detect malignant nodules accurately, the Thyroid Imaging Reporting and Data System (TIRADS) score is considered as a benchmark in this study. Physicians rely on ultrasound imaging to assess thyroid nodules, evaluating visual characteristics observed in the scans. The TIRADS approach facilitates risk stratification and differentiation between benign and malignant nodules, and various versions of TIRADS exist with slight variations in their scoring criteria. The primary research question addressed in this study is whether multimodal ensemble prediction models can effectively and efficiently predict the malignancy of thyroid nodules compared to single modality prediction methods. By combining data from multiple sources, including ultrasound images and clinical data, the proposed approach aims to achieve higher diagnostic accuracy and reliability.

In the subsequent section of the paper, a comprehensive review of the literature on thyroid cancer diagnosis and the utilisation of deep learning and machine learning tools in

[1, 2] *Department of CSE, Hindustan Institute of Technology and Science, Chennai, Tamil Nadu, India*
ORCID ID: 0009-0005-3596-0305
** Corresponding Author Email: sujithra.sankar@gmail.com*

this field is presented. Following this, the dataset and methodology employed in the study, including the architecture of the models and the evaluation metrics, are described in detail. The research delves into experiments with various ensembles of models using both ultrasound images and clinical datasets, highlighting further improvements in performance. The ensemble classifier achieves a remarkable accuracy of 94.78%, underscoring the potential of this approach for precise thyroid cancer diagnosis. The ThyroPredict app powered by the ensemble prediction model, as an aid to radiologists is discussed further in the results.

This research contributes to the healthcare industry by exploring innovative techniques for multimodal thyroid cancer diagnosis, demonstrating the pivotal role of AI technology in the advancement of disease diagnosis. The paper concludes with a discussion of the findings and outlines the scope for future research and improvements on the ensemble framework and ThyroPredict app.

2. Related Work

Thyroid cancer diagnosis has been a topic of intensive research in recent years. The use of deep learning and machine learning algorithms are popular models in the pipeline of clinical prediction of thyroid diseases. In particular, the use of multimodal imaging and the integration of clinical data have shown promising results in improving the accuracy of thyroid cancer diagnosis.

Several studies have investigated the use of CNN for thyroid cancer ultrasound image classification. CNN is a deep learning technique that has shown success in image analysis tasks. In a study by Wenfen Song, Shuai Li et al. in 2019 [7], a CNN-based model was developed to classify thyroid nodules as malignant or benign using ultrasound images. The model achieved a high accuracy of 98.2% in diagnosing malignant nodules. Another study by Yi-Cheng Zhu et al. which came out in 2021 proposed a generic DCNN architecture with transfer learning for thyroid cancer diagnosis using ultrasonographic images [8]. They achieved an average accuracy of 86.5% in classification.

RNN is another deep learning technique that has shown potential in time-series analysis. In a study [9] by Sobhanan Warrier, Gayathry et al. in 2022, an RNN-based model was proposed for the diagnosis of cancer. The objective of their hybrid optimisation approach is to utilise multispectral photoacoustic imaging and transfer learning techniques for the purpose of cancer detection and classification using ultrasound images. The model displayed excellent accuracy above 90%. In a research paper [10] by Elmer et.al in 2020, each geometric and morphological feature was labelled as benign and malignant and the performance of the selected features was evaluated using machine learning. In a three-stage expert system called FS-PSO-SVM, with feature selection methods, optimisation and finally machine learning with SVM in a study [11], researchers obtained an optimal SVM model with the most discriminative feature subset and the optimal parameters.

In another study with a new dataset [12], researchers Khalid Salman et.al. were able to achieve the classification of thyroid diseases as hypothyroidism, hyperthyroidism and normal, with high accuracy with the ensemble method of random forest algorithm. Koundal et al. [13] introduced a comprehensive approach for detecting thyroid nodules based on image analysis. Their method achieved a high accuracy of up to 93.88% in identifying thyroid nodules.

Machine learning models linear model Ridge, LR and XG Boost models are used in developing a free online tool for clinicians to predict patients with high thyroid cancer risk by Jianhua Gu et.al. in a study conducted in 2022 [14]. In their research on the detection of cancerous thyroid nodules, Nanda S, and M Sukumar utilised local binary pattern variants (LBPV) for feature extraction [15], employing which, they were able to accurately classify between benign and malignant thyroid nodules with impressive accuracy of 94.5%.

In research conducted by Xi, N.M. [16], Random Forest, and GBM exhibit better overall diagnostic accuracy in the range of 78% to 85% and the capacity to identify malignant nodules. In a study conducted in 2021 by Wenjun Li et.al, [17] a low-level and high-level feature fusion classification network CNN-F is proposed to classify the benign and malignant nodules. In diagnosing thyroid diseases their classification accuracy reaches 85.92%. These methods can potentially improve the accuracy of thyroid cancer diagnosis, which is crucial for effective management and treatment of the disease. However, further research is needed to validate and optimise the performance of these models in real-world clinical settings.

3. Novelty

This research represents a significant advancement in the field of thyroid cancer diagnosis, where accurate assessment is crucial before undertaking the primary treatment of thyroidectomy.

This showcases the potential of the proposed multimodal approach for accurate thyroid cancer diagnosis. By leveraging the power of deep learning and machine learning, integrating multiple data sources, and utilising a standardised scoring system, the research demonstrates the importance of AI technology in advancing disease diagnosis in the healthcare industry. The novelty of this research lies in its comprehensive approach, which goes beyond traditional single-modality diagnosis methods. By integrating ultrasound images and clinical data and employing an ensemble of models, this study provides a more holistic and accurate assessment of thyroid cancer.

The potential impact of this research on healthcare is significant, as it offers a valuable tool for physicians to make more informed decisions, leading to improved patient outcomes and better prognosis of thyroid cancer cases. As the healthcare industry continues to embrace AI technology, this research sets a promising precedent for the future of disease diagnosis and treatment.

4. Methodology

This section is divided into five major subsections. The first two subsections show the datasets, feature engineering and preprocessing on the clinical and US image datasets. The subsequent subsections introduce the architecture of deep learning and machine learning models. Last subsection introduces the proposed ensemble prediction framework which powers the ThyroPredict app.

4.1 Dataset Overview

The dataset used in our study is TDID [18] which is open access and provided ultrasound (US) images of jpeg format, and the corresponding ground truth by expert radiologists in xml format. The TDID US image dataset, which is publicly available and is created by the Universidad Nacional de Colombia, includes a total of 380 thyroid cancer cases. Each case has single or multiple image files. Thus, the thyroid US images for 380 cases has a total of 455 image files. All the images are in RGB format with the size of 560X360 pixels. The xml file contains information about each patient, and provide details about nodule characteristics, location coordinates (ground truth), and TIRADS scores. The TIRADS score [19] ranges from 2 to 5, with scores of 2 and 3 indicating benign nodules and scores of 4a, 4b, 4c, and 5 indicating malignant nodules. The dataset includes 72 benign cases and rest malignant cases. If the 4a,4b, and 4c are considered as intermediate label, then there are a total of 325 intermediate cases. The xml file provides the clinical data which includes case number, gender, age, composition, echogenicity, margins, calcifications, TIRADS score, image number and the x and y coordinates to mark the Region of Interest (RoI) or ground truth in the ultrasound image.

These attributes from the xml files are first captured to a csv file to create a clinical dataset corresponding to the thyroid US image dataset. Apart from this primary data, x and y coordinates in the xml file is used to compute thyroid nodule’s width, height, L/S ratio and the taller than wide attribute. These are additional columns added to clinical dataset. Thus, the dataset has a total of 10 attributes and 455 tuples.

4.2 Pre-processing and Feature Engineering

Data Preprocessing involves tasks such as handling missing values, removing outliers, normalising, or scaling features, and encoding categorical variables. Feature Engineering step involves tasks such as dimensionality reduction, feature scaling, or creating new features based on domain knowledge. These steps enhance data quality, and removes noise and inconsistencies, helps to standardise the format, and extracts relevant features.

4.2.1 Thyroid Clinical Dataset

Data Cleaning: Pre-processing techniques such as handling missing data, outlier detection, and data imputation are performed to ensure data integrity and reduce bias. To handle missing values or NaNs (Not a Number) in the data set, the missing rows are dropped or filled them with mean or most frequent domain values (imputation by mode).

Feature Scaling and Normalization: Pre-processing by scaling and normalising variables, are carried out to avoid dominance of certain features. Continuous attributes such as age, width, height, and L/S ratio (Long to Short ratio) values are transformed into discrete values.

Encoding Categorical Variables: All categorical variables are encoded into numerical representations (one-hot encoding). For example, the simple integer attribute of age is first converted into categorical data with age ranges of 20. This categorical data is encoded to numerical representations by encoder.

Label Encoder is used to encode the categorical variables of the feature matrix. After encoding the categorical variable, the features are scaled using standard scaler.

The TIRADS score is the class label in our experiments. For binary classification, labels are benign, and malignant. For multiclass classification, labels are benign, intermediate, and malignant.

Handling Imbalanced Data: The distribution of target classes is imbalanced, since it has fewer benign cases compared to malignant cases. oversampling techniques like SMOTE are applied to balance the dataset.

4.2.2 Thyroid US Image Dataset

The thyroid US images are preprocessed using the annotation information or the ground truth provided in xml files of the TDID dataset. The image files have their corresponding xml file which provides x, y coordinates of the ground truth. The US images are annotated, and cropped around the RoI. This reduces unnecessary background noise. Image augmentation methods are applied to increase the dataset size. Each image is flipped, rotated, translated, zoomed, and shifted to create a new image file. Thus, the number of images is increased many folds.

The image size is reduced before they are trained by the model. These image pre-processing steps are crucial to the efficient performance of the deep learning and machine learning models. The original ultrasound thyroid nodule image, and its preprocessed images are shown in Figures 1a, 1b, and 1c.

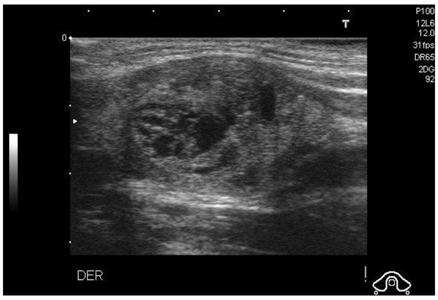


Fig 1a: Thyroid US original image

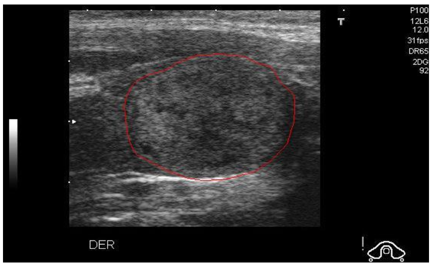


Fig 1b: Thyroid US image with RoI outlined

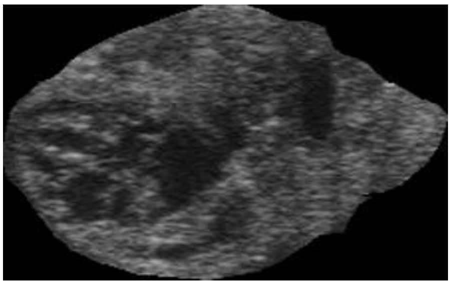


Fig 1c: Thyroid US image- annotated around RoI

TIRADS score is extracted from xml file as class label of classification task. Both binary and multiclass classification are tested on the models.

4.3 Deep Learning Architecture

We experimented on the architecture, hyperparameters, and optimisation methods used in deep learning models CNN, Inception-v3 model, ANN and RNN for classification task on thyroid US images. RNN and ANN are used in the classification using both thyroid clinical dataset as well as US images. An overview of deep learning model configurations used in these experiments are given in the subsections.

4.3.1 CNNs and Inception-v3

Convolutional Neural Networks are widely used for image analysis tasks, including medical image segmentation and diagnosis as can be seen from these studies [7], [8], [20], [21], and [22]. In a typical CNN architecture, the input image is first passed through a series of convolutional layers, each of which applies a set of learnable filters to the input image to extract relevant features.

The image pre-processing steps mentioned in previous section 4.2.2 are crucial to the efficient performance of the CNN model to extract relevant features from the image in the early layers. The architecture of CNN model used in this study is as follows:

1. Input layer: It accepts the grayscale ultrasound image as an input. The US images that are procured from the TDID dataset are of the size 560 x 360 RGB images. The image is resized to fit the model.
2. The first layer is a Conv2D layer with 128 filters, each having a size of 3x3. The activation function is Rectified Linear Unit (ReLU). After each convolutional layer, an activation layer is applied to introduce non-linearity into the model. This layer takes input images with a shape of (224, 224, 3) where 3 represents the RGB color channels.
3. A MaxPooling2D layer with a pool size of (2, 2). Max pooling reduces the spatial dimensions of the input by taking the maximum value in each 2x2 region.
4. Conv2D layer with 256 filters of size 3x3 and ReLU activation.
5. A MaxPooling2D layer with a pool size of (2, 2).
6. A Flatten layer, which flattens the multi-dimensional output of the previous layer into a 1D vector, preparing it for the fully connected layers.
7. A fully connected layer with 512 units and ReLU activation.
8. Final fully connected layer with either 2 or 3 units, representing the output classes of the model. The activation function used is softmax.

The CNN model architecture is given in Figure 2.

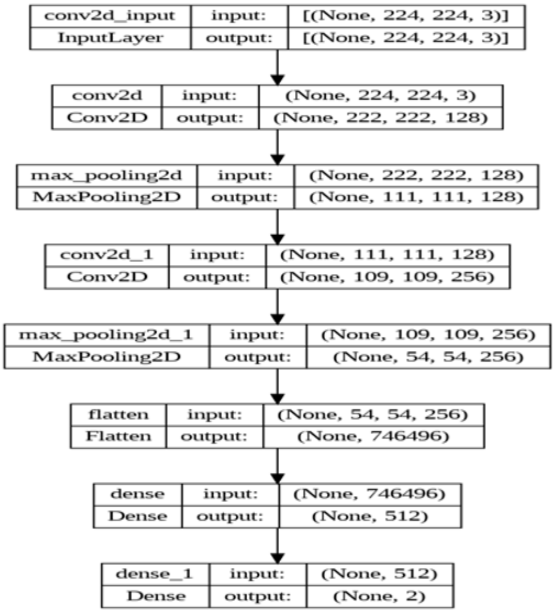

conv2d_input | input: | [(None, 224, 224, 3)]
InputLayer | output: | [(None, 224, 224, 3)]
conv2d | input: | (None, 224, 224, 3)
Conv2D | output: | (None, 222, 222, 128)
max_pooling2d | input: | (None, 222, 222, 128)
MaxPooling2D | output: | (None, 111, 111, 128)
conv2d_1 | input: | (None, 111, 111, 128)
Conv2D | output: | (None, 109, 109, 256)
max_pooling2d_1 | input: | (None, 109, 109, 256)
MaxPooling2D | output: | (None, 54, 54, 256)
flatten | input: | (None, 54, 54, 256)
Flatten | output: | (None, 746496)
dense | input: | (None, 746496)
Dense | output: | (None, 512)
dense_1 | input: | (None, 512)
Dense | output: | (None, 2)

Fig 2: CNN model visualisation: Conv2D layer 128 filters size 3x3 →Maxpool layer (2,2)→Conv2D layer 256 filters of size 3x3→ Maxpooling (2,2)→Flatten layer→Fully Connected layer of 512 units → Dense fully connected layer of 2 Units→Output binary Class

The InceptionV3 is a pre-trained model of convolutional neural network (CNN) architecture designed for image classification tasks. InceptionV3 [23] has been trained on the ImageNet dataset, which consists of millions of labeled images. A brief description of the architecture used in this study:

1. First layer: The InceptionV3 base model is the first layer in the sequential model. This allows the model to leverage the pre-trained features extracted by the InceptionV3 model.
2. Flattening Layer: The output of the base model is flattened, converting it into a 1-dimensional array.
3. Fully connected layers (Dense layers): Dense layer with 512 neurons and ReLU activation function. This layer helps in learning complex patterns in the data.
4. The final Dense layer: Output layer of either 2 or 3 neurons depending on the binary or multiclass scenario. It uses the sigmoid/softmax activation function. Sigmoid is suitable for binary, while softmax is suitable for multiclass classifications.

Five instances of this model are appended to form an ensemble of Inception-v3 model. The architecture is as shown in Figure 3.

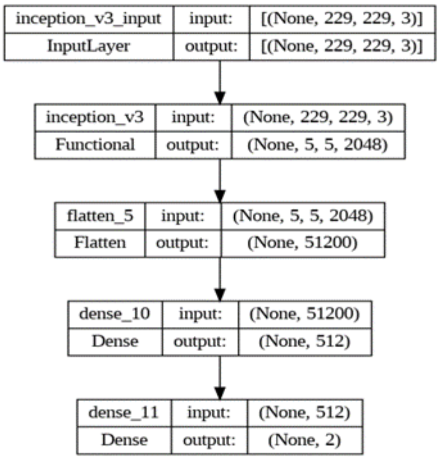

inception_v3_input | input: | [(None, 229, 229, 3)]
InputLayer | output: | [(None, 229, 229, 3)]
inception_v3 | input: | (None, 229, 229, 3)
Functional | output: | (None, 5, 5, 2048)
flatten_5 | input: | (None, 5, 5, 2048)
Flatten | output: | (None, 51200)
dense_10 | input: | (None, 51200)
Dense | output: | (None, 512)
dense_11 | input: | (None, 512)
Dense | output: | (None, 2)

Fig 3: Inception-v3 single model architecture: Base model→Flatten layer→Fully Connected layer of 512 units → Dense fully connected layer of 2 Units→Output binary Class

4.3.2 Recurrent Neural Networks (RNNs) and ANN

Recurrent Neural Networks are another type of neural network commonly used for time-series data analysis, including medical image analysis tasks such as tumor growth prediction.

Optimisation methods such as back propagation through time (BPTT), which is a variant of gradient descent, is used to train the model. To preprocess the thyroid US images, all the preprocessing steps in 4.2.2 are carried out. In this study, recurrent neural network (RNN) model architecture using the Keras framework is employed. The model is created using the Sequential function, which allows stacking layers on top of each other in a sequential manner.

RNN model architecture used in the study is as follows:

1. The first layer is a Bidirectional layer with an LSTM (Long Short-Term Memory) layer inside it. The LSTM layer has 64 units, which control the memory and processing of sequential data. We employed 3 such Bidirectional layers with LSTM layer with 64 units.

2. The final layer is a fully connected output layer. The activation function used is softmax.

This model consists of multiple bidirectional LSTM layers, which capture information from both past and future timesteps in the input sequences. The Dense layer at the end maps the LSTM outputs to the binary or multiclass output.

The RNN architecture used in the research is shown in figure 4 as follows:

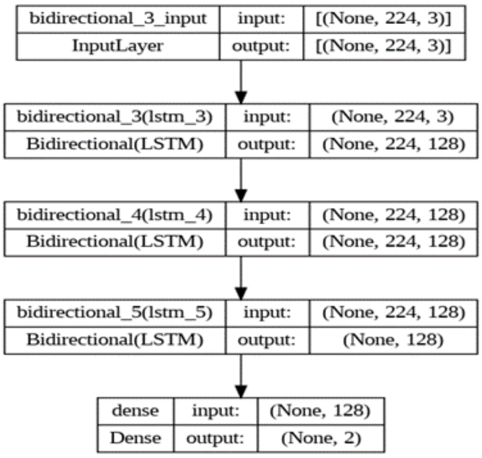

bidirectional_3_input | input: | [(None, 224, 3)]
InputLayer | output: | [(None, 224, 3)]
bidirectional_3(lstm_3) | input: | (None, 224, 3)
Bidirectional(LSTM) | output: | (None, 224, 128)
bidirectional_4(lstm_4) | input: | (None, 224, 128)
Bidirectional(LSTM) | output: | (None, 224, 128)
bidirectional_5(lstm_5) | input: | (None, 224, 128)
Bidirectional(LSTM) | output: | (None, 128)
dense | input: | (None, 128)
Dense | output: | (None, 2)

Fig 4: RNN model: Input Layer → Bidirectional layer with an LSTM with 64 units→ Bidirectional layer with an LSTM with 64 units → Bidirectional layer with an LSTM with 64 units → Dense fully connected layer of 2 Units→Output Class (binary class output scenario)

The clinical dataset extracted from the xml files, is used for creating a thyroid cancer prediction model using RNN, and ANN deep learning classifiers. CNNs are more commonly associated with image processing tasks due to their ability to effectively capture local patterns and spatial relationships in two-dimensional data. So, in our experiment, CNN, and Inception-v3 models are not used for classification using clinical dataset. Dataset pre-processing is an important prerequisite for deep learning algorithms to work effectively. RNN and ANN models are applied on clinical data as well as US image data and yielded good accuracy.

4.4 Machine Learning Architecture

This study proposes a multimodal thyroid cancer diagnosis approach that combines 2 datasets; ultrasound image dataset, and its corresponding clinical dataset. Support Vector Machine (SVM), eXtreme Gradient Boosting (XGB), KNN, Decision trees (DT), Naïve Bayes' (NB), and Random Forest (RF) models are used to train the datasets.

4.4.1 Machine Learning Model with Clinical Data

The clinical dataset of 10 attributes and 455 tuples, is prepared for machine learning. This dataset is preprocessed as mentioned in the clinical dataset pre-processing section 4.2.1 above.

A 10-fold validation is performed to improve the efficiency before it is tested on test dataset. The prepared dataset is divided into train and test subsets in 80:20 ratio.

4.4.2 Machine Learning Model with Image Data

Preprocessing step is a crucial one and the accuracy of the classification depends on effective preprocessing of image. For detailed preprocessing of image data refer to section 4.2.2.

Feature Extraction is achieved by Discrete Fourier Transform (DFT) that converts a signal from the spatial domain to the frequency domain. The frequency spectrum is used to extract features that are relevant for diagnosing thyroid nodules. Frequency components in the spectrum correspond to the various features of the image such as edges or texture of the nodule, overall shape, and size of the nodule. We extracted features from the frequency spectrum, by using statistical measures like mean, variance, skewness, and kurtosis. Using DFT for feature extraction improved the accuracy of the model in diagnosing malignant thyroid nodules in ultrasound images.

DFT can be represented as follows:

Let x(t) be input signal representing the US image of a thyroid nodule in the spatial domain, where t is the spatial domain variable.

Let X(f) be discrete Fourier Transform (DFT) of x(t) representing the frequency spectrum, where f is the frequency domain variable.

The DFT of the signal x(t) is,

$$X(f)= \sum^{N-1} x(t)\cdot \underset{t=0}{e}^{-j2\pi ft/N} \qquad (1)$$

Where N is the number of samples in the input signal, and

j is the imaginary unit.

The frequency spectrum X(f) captures the frequency components of the image, which correspond to various features such as edges, texture, shape, and size of the thyroid nodule.

To extract features from the frequency spectrum X(f), statistical measures are applied.

$$\text{Mean, } \mu = N1 \sum_{f=0} \underset{N-1}{X(f)} \qquad (2)$$

$$\text{Variance, } \sigma^2 = 1/N \sum_{f=0} |\underset{N-1}{X(f)}-\mu|^2 \qquad (3)$$

$$\text{Skewness, } \gamma = (1/N \sum_{f=0} |\underset{N-1}{X(f)}-\mu|^3) / \sigma^3 \qquad (4)$$

$$\text{Kurtosis, } \kappa = (1/N \sum_{f=0} |\underset{N-1}{X(f)}-\mu|^4) / \sigma^4 \qquad (5)$$

These statistical measures provide quantitative information about the distribution of frequency components in the spectrum. The extracted statistical features mean, variance, skewness and kurtosis from the frequency spectrum are then used as inputs for a diagnostic model to distinguish between malignant and non-malignant thyroid nodules. This DFT-based feature extraction is found to increase the accuracy of thre model in thyroid nodule.

For training the SVM model, a radial basis function (RBF) kernel was used. For gradient boost method, we used the XG Boost library. For Random Forest algorithm scikit-learn library is used.

Once the models are trained, their performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. A 10-fold cross-validation is used to prevent overfitting on training data. The trained models are tested using new US images of thyroid nodules which are not used during training. The models are trained for both binary or multiclass classifications scenarios.

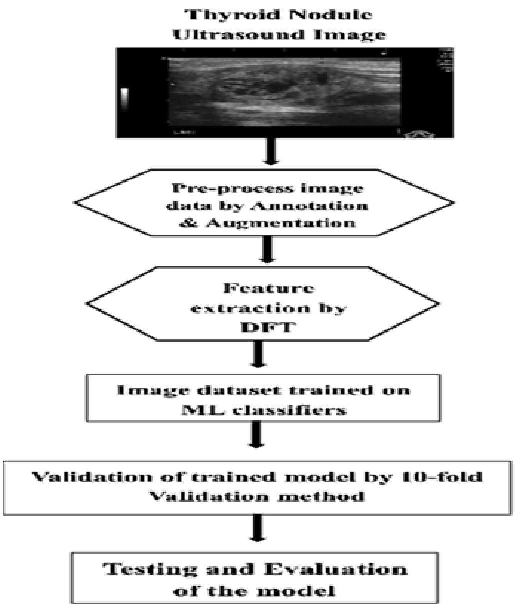

Thyroid Nodule Ultrasound Image
Pre-process image data by Annotation & Augmentation
Feature extraction by DFT
Image dataset trained on ML classifiers
Validation of trained model by 10-fold Validation method
Testing and Evaluation of the model

Fig 5: Machine learning model architecture with thyroid US.

The machine learning architecture of thyroid image data is shown in Figure 5.

4.5 Proposed Ensemble Prediction framework

ML and DL models are trained individually on image and clinical datasets and their performance is evaluated in phase I. The models which demonstrated high accuracy with good performance consistency are selected to the next phase. In phase II, multiple ensemble prediction models are trained on thyroid US image data and clinical data. The ensemble which showed highest prediction accuracy is used in ThyroPredict web app framework.

Our proposed ThyroPredict web app framework is as follows:

1. Clinical data extraction: In the first step, radiologist can upload the US image in the interface provided in the app, and input their first assessment of the US image. The RoI can be marked on the image. The clinical data input from the radiologist, as well as the RoI will be converted into an xml file and stored for future reference. Clinical data is extracted from this xml file. Additional attributes such as nodule width, height, L/S ratio, and ttw are also calculated from RoI and added to clinical data.

2. Preprocess data: The image and clinical data are preprocessed and prepared for prediction on the ensemble model. The preprocessing steps mentioned in section 4.2 are applied on both image and clinical data. RoI extracted from the xml file is used in the annotation of image file in this step.

3. Classification: ThyroPredict is an ensemble of two types of classifiers. Image data is loaded into image classifier ensemble. Clinical data is loaded into clinical data classifier ensemble. These ensembles are trained on thyroid nodules image and clinical datasets respectively.

4. Prediction Vector: When a new thyroid US image data with its RoI is given to this ensemble, individual predictions are generated by models and are added to a prediction vector.

5. Majority voting: A majority vote on prediction vector is performed. The mode (most frequent value) across the predictions is the final prediction. For our proposed model, an odd number of models is combined to avoid tie breaker scenario.

6. Prediction: Final prediction is, either a benign or a malignant class.

The mathematical model of the majority voting ensemble

can be expressed as follows: We have M individual models, each producing a binary prediction (1 or 0) for a given instance.

Each instance is represented by an index j, where 1≤ j ≤N,

and N is the total number of instances.

For each instance j, let $\mathbf{p_{i,j}}$ represent the prediction of the ith model, where 1≤ i ≤M.

$\mathbf{p_{i,j}}$ can take values of 1 (positive prediction) or 0 (negative prediction).

The majority voting ensemble prediction p_j for the jth instance can be calculated using the Heaviside step function (H) as shown in eqation no. 6.

$$p_j=H\left(\sum_{i=1}^{M} p_{i,\,j} - M/2\right) \tag{6}$$

Here, H(x) is the Heaviside step function, defined as:

1, if x >=0

0, if x < 0

In this representation, the sum of positive predictions from the individual models sum p_{ij} is calculated, and then subtract M/2 from it.

The Heaviside step function is then applied to determine whether the result is positive or negative, leading to the final ensemble prediction 1 or 0.

The majority voting ensemble predicts the positive class if the sum of positive predictions from individual models is at least half of the total number of models. Otherwise, it predicts the negative class. Figure 6 shows the proposed ThyroPredict framework.

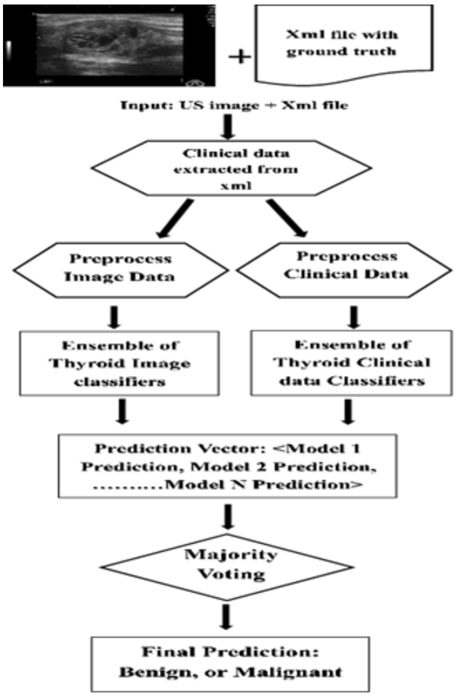

Xml file with ground truth
+
Input: US image + Xml file
Clinical data extracted from xml
Preprocess Image Data
Preprocess Clinical Data
Ensemble of Thyroid Image classifiers
Ensemble of Thyroid Clinical data Classifiers
Prediction Vector: <Model 1 Prediction, Model 2 Prediction, ...Model N Prediction>
Majority Voting
Final Prediction: Benign, or Malignant

Fig 6: Proposed ThyroPredict web app framework.

5. Results and Discussion

5.1 Experimental Setup

The experiments on clinical dataset with ML models are conducted on a personal computer with an Intel(R) Core (TM) i5-1035G1 CPU, 1.19 GHz Processor and 8GB RAM. The operating system is 64-bit Windows 11. The environment for deep learning is Python 3.8.5 and Keras 2.4.3 with TensorFlow 2.4.0. The experiments with ML, DL and Inception-v3 models on image datasets are conducted on Google Colab Pro with V100 / A100 Nvidia GPU, Python 3.10.11 and Keras 2.12.0 with TensorFlow 2.12.0.

In the DL model training, the loss function used are Binary and Categorical Cross Entropy, the optimizer is adam, and the performance evaluation indices are accuracy and loss. The learning rate of 0.001 is used in DL models. Some more specifications of the experiment are: dropout rate is 0.3; activation functions are ReLU, Sigmoid, and softmax; batch size is 32, and epochs for all deep learning models is fixed at 200.

In phase I of this experiment, following scenarios are tested:

Datasets: Clinical dataset, and Image dataset

Model Architectures: ML models, and DL models.

Class Labels: Binary and Multiclass

Thus, there is a total of 8 scenarios possible with these combinations.

In phase II of our experiment, model ensembles are trained on both image and clinical datasets to give a multimodal prediction output. Majority voting is employed for final output of the ensemble of model combinations.

To develop the ThyroPredict web app, Python Flask, javscript and html are used. Flask is a python web framework. The app was hosted locally in the Flask web server.

5.1.1. Phase I: Performance Evaluation of Models

In phase I of the experiment, all ML and DL models are evaluated on multiple scenario combinations of datasets, and class labels. The performance of each model is evaluated in terms of accuracy metric. The best models are selected for the next phase. The comparison of model performances in accuracy metric is shown in the corresponding charts in figures 7 to 9. All ML models showed a jump in accuracy when scenario changed from multiclass to binary classification.

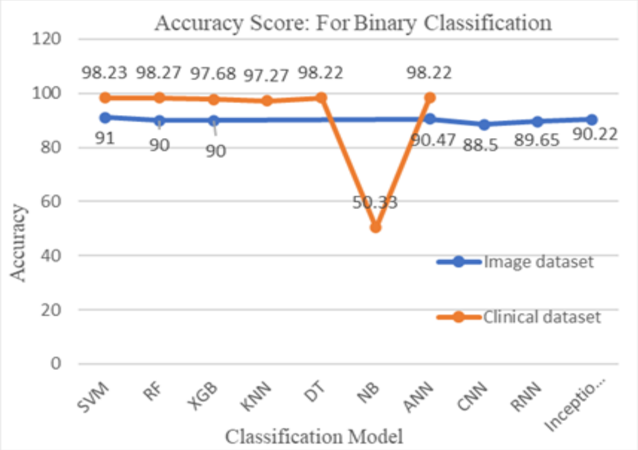


Fig 7: Prediction accuracy scores of Binary classifications

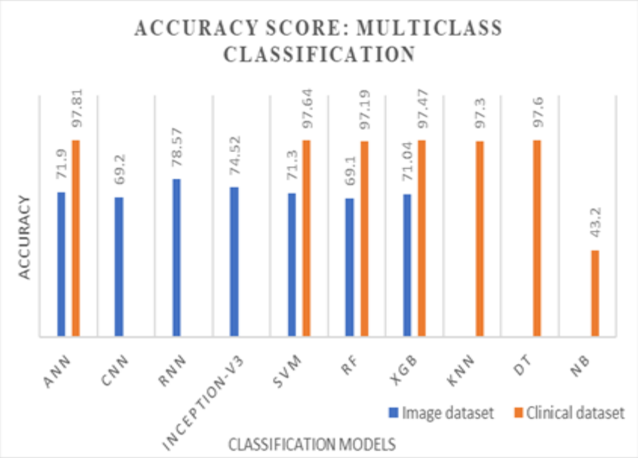


Fig 8: Prediction accuracy scores Multiclass Classifications**.**

In phase I, DL models CNN and Inception-v3 are not trained and evaluated on thyroid clinical dataset, since these models are primarily designed for image data analysis, and they are not the most common or effective choice for this task.

5.1.2. Phase II: Our Proposed Ensemble Prediction Model

The models with highest accuracies are selected to phase II for creating ensembles to experiment on the combination of image and clinical data in only binary classification scenario. Ensemble of models are created and first evaluated in terms of accuracy. To select the best performing model, precision, recall, and F1 score metrics are used to further evaluate the performance. While accuracy provides an overall measure of the model's correctness, precision, recall, and F1 score helps to assess a model's performance [24], particularly in imbalanced or skewed datasets. These parameters are calculated as follows:

Accuracy = TP+TN / TP+TN+FP+FN (7)

Precision = TP / TP+FP (8)

Recall = TP / TP+FN (9)

Where, TP, True Positive: number of malignant cases correctly classified as malignant.

FP, False Positive: number of benign cases misclassified as malignant.

TN, True Negative: number of benign cases correctly classified as benign.

FN, False Negative: number of malignant cases misclassified as benign.

Precision, also known as Positive Predictive Value or PPV, is calculated as the ratio of true positives to the sum of true positives and false positives. Precision provides insight into the model's accuracy when it predicts a positive class, as in equation no. 8.

Recall, also known as sensitivity or true positive rate, measures the model's ability to correctly identify positive instances out of all the actual positive instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives, as in equation no. 9.

The F1 score calculation is based on precision and recall, which is shown in equation no. 10. F1 score is useful when there is an uneven distribution of classes in the dataset.

F1 Score = 2 * ((precision * recall) / (precision + recall)) (10)

The ensemble prediction model with ANN and SVM on image dataset in combination with SVM, DT and RF on clinical dataset produced a prediction accuracy of 94.78%, and precision of 95.5% and recall of 71.4%. This ensemble showed the best performance in our experiments.

All the tested and passed ensemble prediction model combinations and their performance metrics are shown graphically in Figure 9.

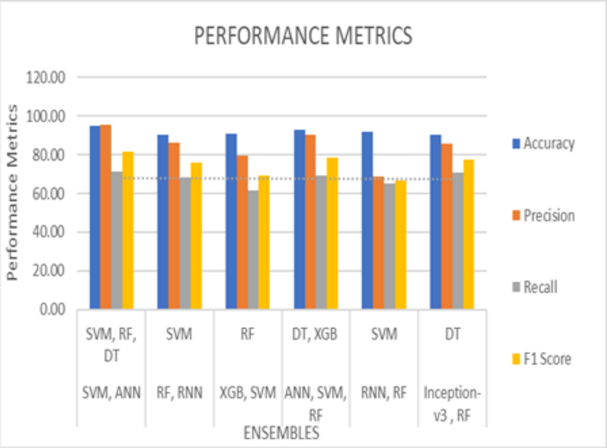
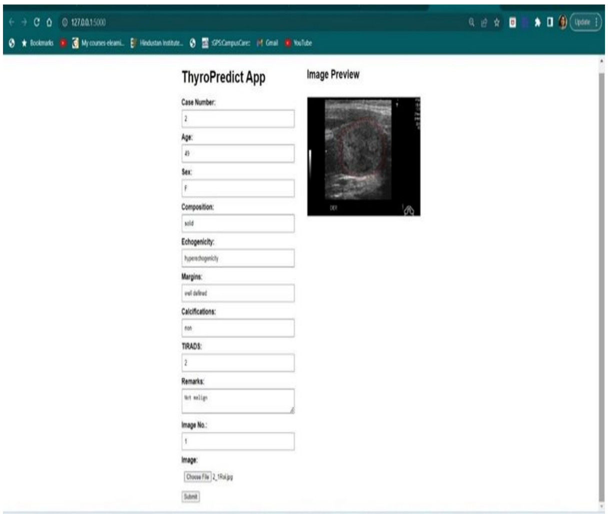


Fig 9: Performance metrics of ensemble prediction models

5.2. ThyroPredict App

The proposed ensemble prediction model is used to power the ThyroPredict app, which uses Flask as web server in the local machine, and is developed using Python, HTML, CSS and JavaScript. The user interface of ThyroPredict is easy to navigate. The radiologist can upload the ultrasound image and input the initial details and assessments. On submission of the data in the home page, image and corresponding clinical data of the patient is processed by the ensemble prediction model. The RoI marked on the image will be used in annotating the image in pre-processing. The x-y coordinates of RoI will be used to compute some additional clinical data such as Long to Short ratio, taller than wide etc. The home page and the final prediction of the app are shown in Figures 10, and 11.



ThyroPredict App

Case Number: 2

Age: 49

Sex: F

Composition: solid

Echogenicity: hyperechogenicity

Margins: well defined

Calcifications: non

TIRADS: 2

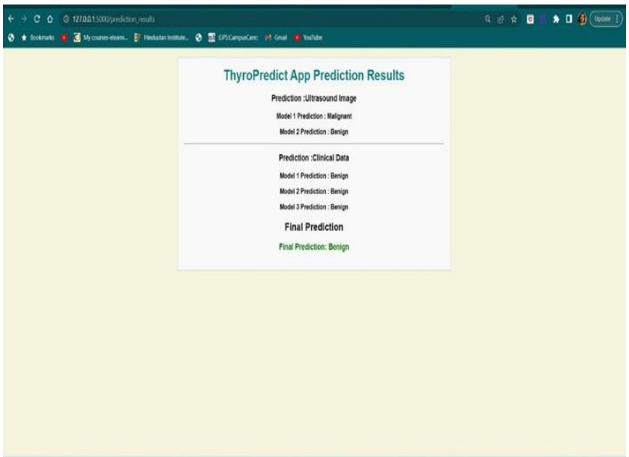
Remarks: Not malign

Image No.: 1

Image: 2_1RoI.jpg

Image Preview

Fig 10: ThyroPredict app homepage after data entry and image upload



ThyroPredict App Prediction Results

Prediction :Ultrasound Image

Model 1 Prediction : Malignant

Model 2 Prediction : Benign

Prediction :Clinical Data

Model 1 Prediction : Benign

Model 2 Prediction : Benign

Model 3 Prediction : Benign

Final Prediction

Final Prediction: Benign

Fig 11: ThyroPredict app final prediction page

5.3. Performance Comparisons of Proposed Ensemble Prediction Model with other Similar Models

To evaluate and compare our proposed model with other similar works, we have selected studies which worked on TDID dataset. In a study using deep learning and machine learning methods [20], a prominent work by Dat Tien Nguyen in 2019 with CNN-based network for classifying ultrasound thyroid images yielded an overall accuracy of close to 87% using the TDID dataset.

The study by Nanda and Sukumar [15] mentioned in the literature, used deep learning-based and LBPV-based methods using feature extraction and classification techniques that are similar to our proposed approach.

The overall accuracy reported by Zhu et.al., [21] using the original TDID dataset is 93.75%, which is obtained by performing the data augmentation on both training and testing datasets, which is like our experiments. They obtained a classification accuracy of about 84.0% using Resnet18-based network with the TDID dataset. In our experiments, the ensemble of five Inception-v3 models achieved an accuracy of 90.22% on TDID thyroid US image dataset, after data augmentation on training and test data subsets.

Additionally, it should be noted that the studies conducted by Chi et al. [22] and Sundar et al. [25] utilised multiple datasets other than TDID dataset for their researches. Their accuracy scores are 79.36 and 77.57% respectively. A study by Guan Q. et.al. [26] exploited a VGG-16 deep convolutional neural network (DCNN) model to differentiate papillary thyroid carcinoma (PTC) from cytological images. They tested the dataset on VGG-16 network and Inception-V3. The accuracy comparisons are given below in the table.

In a recent study [27] by Luoyan Wang, et.al, in 2022 proposed a CNN model, called n-ClsNet, for thyroid nodule classification. Their model n-ClsNet achieved an average accuracy score of 93.8% in the thyroid nodule classification task. In their study in 2020, Dat Tien Nguyen et.al. [28] have achieved an accuracy score of 92.05%, with their model using Min, Max and Sum methods for thyroid nodule classification.

This is presented in Tables 1 and 2 below. There are 2 types of comparisons. Table 1 shows the comparison between our proposed method and the other state-of-the-art methods which used the same dataset TDID. Table 2 compares between the proposed method and other thyroid nodule ensemble classification approaches in other related studies.

Table 1: Accuracy Comparison table of models which used other data sources, and our proposed model

Reference	*Accuracy*	*Network/ Model*
[13]	93.88	
[15]	94.5	LBVP
[20]	87	
[28]	92.05	Min, Max, Sum Method
[21]	84	Inception Model
[21]	93.75	Overall Accuracy
[22]	79.36	FDCNN
[23]	77.57	
Our Proposed Model	94.8	Ensemble of ANN, SVM, DT and RF

Table 2: Accuracy Comparison of other models which used TDID dataset, and our proposed model.

Reference	*Accuracy*	*Network/ Model*
[12]	90.91	Average Accuracy of ML Models
[16]	85	RF/ XGB
[17]	85.92	CNN - F
[24]	87.5	Inception V3
[26]	87.5	Inception, VGGNet-16

[27]	93.8	TNUI -21 dataset, nCls-Net
Our Proposed Model	94.8	Ensemble of ANN, SVM, DT and RF

6. Conclusions and Future Scope of the Study

The main goal of this study is to develop a state-of-the-art ensemble model which will classify and predict thyroid nodules based on multiple modalities, clinical and US image data.

From the experiments, an ideal model for thyroid nodule classification has emerged: a majority voting ensemble of ANN, SVM model trained on image dataset in combination with SVM, RF, and DT models trained on clinical dataset.

An app developed based on this ensemble prediction model have demonstrated that our proposed ensemble method surpasses the classification accuracy achieved by many state-of-the-art models. The ThyroPredict app is specifically designed to offer a second opinion to doctors, particularly radiologists, in the diagnosis of thyroid nodules.

The Recall (sensitivity) metric of our ensemble is found to be a bit weak (71.4%). Recall, also known as True Positive Rate (TPR) gives us the correctly classified positive instances out of all positive instances. It implies that some of the benign cases are also misclassified as malignant. A major future scope of this research is to reduce the false negatives and thus increase the sensitivity of the prediction app.

The accuracy result analysis of models shows that the prediction accuracies of DL and ML models trained on image dataset are relatively lower than the same models trained on their corresponding clinical dataset. Thus, the future scope of this study is to improve the image dataset prediction accuracy.

Another observation in the study is that the predictions with multiple class (viz. benign, intermediate, and malign) are less accurate compared to binary class (viz. benign, and malign) predictions. Thus, another key area of improvement is multiclass classification accuracy of thyroid nodule diagnosis.

Key findings of our study of multimodal ensemble prediction can be used to identify areas for future research and development, and recommendations for clinicians and researchers. In conclusion, our ensemble model powered ThyroPredict app can be effectively used in thyroid cancer diagnosis and has shown promising results.

Data Availability Statement

Publicly available dataset is analysed in this study. This data can be found here: www.cimalab.unal.edu.co.

Acknowledgements

We thank our colleagues from Hindustan Institute of Technology and Science who provided insight and expertise that greatly assisted the research.

Author contributions

Sujithra Sankar: Conceptualization, Methodology, Software, Field study, Data curation, Writing-Original draft preparation, Software, Validation., Field study, and Visualization **Sathyalakshmi S:** Investigation, Reviewing and Editing.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Parul Puri et al., "Burden and determinants of multimorbidity among women in reproductive age group: a cross-sectional study based in India.," 2021 Feb 18. doi: 10.12688/wellcomeopenres.16398.2

[2] American Cancer Society. Cancer Statistics Center. https://cancerstatisticscenter.cancer.org/ (2023).

[3] Kitahara CM, Sosa JA. "The changing incidence of thyroid cancer", Nat Rev Endocrinol 2016 Nov;12(11):646-653.

[4] Davies L, Welch HG. "Current thyroid cancer trends in the United States", JAMA Otolaryngol Head Neck Surg. 2014 Apr;140(4):317-22.

[5] Tessler, F.N.; Middleton, W.D.; Grant, E.G.; Hoang, J.K.; Berland, L.L.; Teefey, S.A.; Cronan, J.J.; Beland, M.D.; Desser, T.S.; Frates, M.C.; et al. "ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee." J. Am. Coll. Radiol. 2017, 14, 587–595. [CrossRef] [PubMed]

[6] Kwak, J.Y.; Han, K.H.; Yoon, J.H.; Moon, H.J.; Son, E.J.; Park, S.H.; Jung, H.K.; Choi, J.S.; Kim, B.M.; Kim, E.-K. "Thyroid Imaging Reporting and Data System for US Features of Nodules: A Step in Establishing Better Stratification of Cancer Risk", Radiology 2011, 260, 892–899. [CrossRef]

[7] W. Song et al., "Multitask Cascade Convolution Neural Networks for Automatic Thyroid Nodule

Detection and Recognition," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 3, pp. 1215-1224, May 2019, doi: 10.1109/JBHI.2018.2852718.

[8] Yi-Cheng Zhu, Alaa AlZoubi, Sabah Jassim, Quan Jiang, Yuan Zhang, Yong-Bing Wang, Xian-De Ye, Hongbo DU, “A generic deep learning framework to classify thyroid and breast lesions in ultrasound images”, Ultrasonics, Volume 110, 2021, 106300, ISSN 0041-624X, https://doi.org/10.1016/j.ultras.2020.106300.

[9] Gayathry Sobhanan Warrier ,1T. M. Amirthalakshmi, K. Nimala, T. Thaj Mary Delsy, P. Stella Rose Malar, G. Ramkumar, and Raja Raju, Yuvaraja Teekaraman, “Automated Recognition of Cancer Tissues through Deep Learning Framework from the Photoacoustic Specimen”, Hindawi, Volume 2022 | Article ID 4356744

[10] Elmer Jeto Gomes Ataide, Nikhila Ponugoti, Alfredo Illanes, Simone Schenke, Michael Kreissl, and Michael Friebe, “Thyroid Nodule Classification for Physician Decision Support Using Machine Learning-Evaluated Geometric and Morphological Features”, Sensors 2020, 20, 6110; doi:10.3390/s20216110

[11] Chen HL, Yang B, Wang G, Liu J, Chen YD, Liu DY. “A three-stage expert system based on support vector machines for thyroid disease diagnosis.”, J Med Syst. 2012 Jun;36(3):1953-63. doi: 10.1007/s10916-011-9655-8. Epub 2011 Feb 1. PMID: 21286792.

[12] Khalid Salman and Emrullah Sonuç, “Thyroid Disease Classification Using Machine Learning Algorithms”, 2021 J. Phys.: Conf. Ser. 1963 012140 doi:10.1088/1742-6596/1963/1/012140

[13] Koundal, D.; Gupta, S.; Singh, S., “Computer aided thyroid nodule detection system using medical ultrasound images”; Biomed. Signal Process. Control 2018, 40, 117–130. [CrossRef]

[14] Gu Jianhua, Xie Rongli, Zhao Yanna, Zhao Zhifeng, Xu Dan, Ding Min, Lin Tingyu, Xu Wenjuan, Nie Zihuai, Miao Enjun, Tan Dan, Zhu Sibo, Shen Dongjie, Fei Jian, “A machine learning-based approach to predicting the malignant and metastasis of thyroid cancer “, Frontiers in Oncology 12 2022, https://www.frontiersin.org/articles/10.3389/fonc.2022.938292, doi=10.3389/fonc.2022.938292

[15] Nanda S, M Sukumar "Identification of Thyroid Cancerous Nodule using Local Binary Pattern Variants in Ultrasound Images", International Journal of Engineering Trends and Technology (IJETT), V49(6),369-374 July 2017. ISSN:2231-5381. www.ijettjournal.org. published by seventh sense research group. [CrossRef]

[16] Xi, N.M., Wang, L. & Yang, C. Improving the diagnosis of thyroid cancer by machine learning and clinical data. Sci Rep 12, 11143 (2022). https://doi.org/10.1038/s41598-022-15342-z

[17] Li W, Cheng S, Qian K, Yue K, Liu H. Automatic Recognition and Classification System of Thyroid Nodules in CT Images Based on CNN. Comput Intell Neurosci. 2021 May 27; 2021:5540186. doi: 10.1155/2021/5540186. PMID: 34135949; PMCID: PMC8175135.

[18] Pedraza L., Vargas C., Narvaez F., Duran O., Munoz E., Romero E. An open access thyroid ultrasound-image database; Proceedings of the 10th International Symposium on Medical Information Processing and Analysis; Cartagena de Indias, Colombia. 28 January 2015; pp. 1–6. [Google Scholar]

[19] Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A, et al. An ultrasonogram reporting system for thyroid nodulesstratifying cancer risk for clinical management. J Clin EndocrinolMetab 2009; 94(5):1748-51

[20] Nguyen DT, Pham TD, Batchuluun G, Yoon HS, Park KR. Artificial Intelligence-Based Thyroid Nodule Classification Using Information from Spatial and Frequency Domains. J Clin Med. 2019 Nov 14;8(11):1976. doi: 10.3390/jcm8111976. PMID: 31739517; PMCID: PMC6912332.

[21] Zhu Y., Fu Z., Fei J. An image augmentation method using convolutional network for thyroid nodule classification by transfer learning; Proceedings of the 3rd IEEE International Conference on Computer and Communication; Chengdu, China. 13–16 December 2017; pp. 1819–1823. [Google Scholar]

[22] Chi J., Walia E., Babyn P., Wang J., Groot G., Eramian M. “Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network.” J. Digit. Imaging. 2017;30:477–486. doi: 10.1007/s10278-017-9997-y. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[23] Szegedy C, Vanhoucke V, Ioffe S, “Rethinking the Inception architecture for computer vision”, [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; 2818-26. [Ref list]

[24] Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies.

[25] Sundar K.V.S., Rajamani K.T., Sai S.-S.S. Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering. Springer; Cham, Germany: 2018. “Exploring image classification of thyroid ultrasound images using deep learning”, pp. 1635–1641. [Google Scholar]

[26] Guan Q, Wang Y, Ping B, Li D, Du J, Qin Y, Lu H, Wan X, Xiang J.;” Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study.”; J Cancer. 2019 Aug 27; 10(20):4876-4882. doi: 10.7150/jca.28769. PMID: 31598159; PMCID: PMC6775529.

[27] Luoyan Wang, et.al, “A Multi-Scale Densely Connected Convolutional Neural Network for Automated Thyroid Nodule Classification,” Front Neurosci. 2022; 16: 878718., Published online 2022 May 19. doi:10.3389/fnins.2022. 878718.

[28] Nguyen DT, Kang JK, Pham TD, Batchuluun G, Park KR. “Ultrasound Image-Based Diagnosis of Malignant Thyroid Nodule Using Artificial Intelligence.,” Sensors (Basel). 2020 Mar 25;20(7):1822. doi: 10.3390/s20071822. PMID: 32218230; PMCID: PMC7180806.