Improving the Diagnosis of Thyroid Cancer using Machine Learning

A Project Report Submitted in
Partial Fulfilment of the Requirements for the
7^{th} Semester B.Tech. Project

by

Aditya Kumar 2012021
Sanjeeb Kumar Rai 2012052
Rahul Ravel 2012063

Under the Supervision of

Dr. Saroj Kumar Biswas

Computer Science & Engineering Department
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
December, 2023

COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

Declaration

Thesis Title: **Improving the diagnosis of Thyroid Cancer using Machine Learning.**
Degree for which the Thesis is submitted: **Bachelor of Technology**

We declare that the presented thesis represents largely my own ideas and work in my own words. Where others ideas or words have been included, We have adequately cited and listed in the reference materials. The thesis has been prepared without resorting to plagiarism. We have adhered to all principles of academic honesty and integrity. No falsified or fabricated data have been presented in the thesis. We understand that any violation of the above will cause for disciplinary action by the Institute, including revoking the conferred degree, if conferred, and can also evoke penal action from the sources which have not been properly cited or from whom proper permission has not been taken.

Signed: _____

Date: _____

YOUR DEPARTMENT NAME
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

It is certified that the work contained in this thesis entitled **"Improving the diagonosis of Thyroid Cancer using Machine Learning "** submitted by , Registration no (Registration No.) for the B.Tech. End Semester Project Examination December, 2023 is absolutely based on his own work carried out under my supervision.

Place:
Date:

Dr. Saroj Kumar Biswas
Computer Science & Engineering
National Institute of Technology Silchar

"You have to dream before your dreams can come true."

A. P. J. Abdul Kalam

Abstract

This project focuses on enhancing the accuracy and efficiency of thyroid cancer diagnosis through the application of machine learning techniques. Leveraging a dataset of thyroid nodule clinical dataset, we employ advanced machine learning algorithms for feature extraction and classification. Our methodology involves preprocessing the data, selecting relevant features, and training a predictive model. Through rigorous evaluation and validation, our results demonstrate improved diagnostic precision compared to traditional methods. This project contributes to the advancement of thyroid cancer diagnosis, showcasing the potential of machine learning in enhancing medical decision-making processes.

Acknowledgements

We take this opportunity to express our sincere gratitude and heartily thanks to our supervisor Dr Saroj Kumar Biswas, CSE Department, National Institute of Technology Silchar for his continuous inspiration and valuable guidance at every stage of my research work. We would like to thank our Doctoral Committee Chairman and other members for their continuous evaluation and valuable constructive suggestions during this work. We would like to also thank all the faculty members of the CSE Department of National Institute of Technology Silchar, for their administrative support during various phases of this work.

Aditya Kumar, Sanjeeb Kumar Rai, Rahul Ravel

Contents

List of Figures

List of Tables

List of Algorithms

CHAPTER 1

Introduction

Thyroid cancer, a prevalent form of endocrine carcinoma impacting the thyroid gland, presents a formidable diagnostic challenge due to the limitations of existing detection methods. The current approaches often struggle to deliver precise and reliable results, underscoring the need for innovative solutions to enhance diagnostic accuracy. In response to this critical issue, our research endeavors to revolutionize the diagnosis of thyroid cancer by harnessing the power of machine learning in conjunction with valuable clinical data.

Recognizing the shortcomings of conventional diagnostic methods, our study seeks to introduce a novel approach that leverages the capabilities of machine learning algorithms. By integrating advanced computational techniques with comprehensive clinical information, we aim to overcome the limitations that have hindered accurate thyroid cancer diagnosis. The primary objective of our project is not only to improve diagnostic precision but also to establish a foundation for more effective and informed medical decision-making.

This innovative initiative represents a significant step towards addressing the challenges associated with thyroid cancer diagnosis. By embracing the potential of machine learning, we aspire to contribute to the advancement of medical practices, ultimately leading to better patient outcomes and more personalized treatment strategies. Through this collaborative effort between technology and clinical expertise, our research aims to pave the way for a transformative era in the field of thyroid cancer diagnosis.

1.0.1 Motivation

Our motivation for undertaking this pioneering project lies in our unwavering commitment to revolutionize healthcare, with a specific focus on optimizing the treatment planning for thyroid cancer. Recognizing the diagnostic challenges posed by this frequently encountered endocrine carcinoma, we are driven by a profound desire to improve patient outcomes through the integration of advanced technologies and data-driven approaches.

One of our primary objectives is to reduce surgical complications associated with thyroid cancer interventions. By harnessing the power of machine learning and comprehensive clinical data, we aim to provide clinicians with a more precise understanding of the disease, enabling them to tailor treatment plans with greater accuracy. This, in turn, holds the potential to enhance the overall quality of medical interventions, contributing to improved patient recovery and well-being.

In addition to our commitment to precision in medical interventions, we are cognizant of the economic considerations in healthcare. Our focus on cost-effective solutions aligns with the goal of making high-quality medical care accessible and sustainable. By incorporating machine learning into the diagnostic process, we not only aim to improve accuracy but also to streamline healthcare practices, reducing unnecessary costs associated with misdiagnosis or ineffective treatments.

1.0.2 Objective

give objective in bulletin.

i To improve the accuracy of thyroid cancer diagnosis.

ii To reduce the time and effort required for diagnosis.

iii To provide more informed treatment decisions and better patient outcomes.

iv To make the diagnosis of thyroid cancer more accessible to patients in all areas.

CHAPTER 2

Literature survey

In [1], there is a study about a deep learning model to assist thyroid nodule diagnosis and management. It discusses the development and testing of a deep learning model called ThyNet. ThyNet was trained on a large dataset of images of thyroid nodules and was able to accurately differentiate between malignant and benign nodules. In a clinical trial, ThyNet was able to improve the diagnostic performance of radiologists. The authors conclude that ThyNet is a promising tool for improving the diagnosis and management of thyroid nodules. You can find more details in the article

In [2], the author mentions about the use of deep convolutional neural networks (DCNNs) to diagnose thyroid cancer. It discusses the rising incidence of thyroid cancer and the need for more accurate diagnostic methods. The authors developed a DCNN model and trained it on a large dataset of ultrasound images. They then tested the model on three validation sets and found that it was able to accurately diagnose thyroid cancer with high sensitivity and specificity. The authors conclude that DCNNs are a promising tool for the diagnosis of thyroid cancer.

In [3], the investigation was conducted and the classification of thyroid disorders was approached through various models utilizing parameters such as TSH, T4U, and the presence of goiter. The study employed diverse grouping techniques, including K-nearest neighbor, to analyze and categorize thyroid disorders. Notably, Naive Bayes and support vector machines algorithms were applied in the study, conducted using the RapidMiner tool.

The study highlighted that the proposed KNN technique contributed to an enhancement in classification accuracy, thereby leading to improved overall results. It was emphasized that Naive Bayes, due to its nature, can only establish linear, elliptic, or parabolic decision boundaries, while the K-nearest neighbor's decision boundary consistency was identified as a significant advantage. The superior performance of KNN was attributed to the interdependence of factors, surpassing most other methods in the process.

AKGÜL, Göksu, et al [4], aimed to enhance the precision of hypothyroidism diagnosis by proposing a data mining-based method that integrates patient-generated questions with test results in the diagnostic process. Another objective was to mitigate risks associated with dialysis interventional trials. The study reached a logical conclusion by determining the hypothyroid status of new samples, utilizing data from the UCI machine learning database, comprising 3163 samples, of which 151 were hypothyroid and the remainder were not.

To address the issue of unbalanced distribution in the dataset, various sampling techniques were employed during data collection. Models were then developed to diagnose hypothyroidism, utilizing classifiers such as Logistic Regression, K Nearest Neighbor, and Support Vector Machine. The thesis underscored the significance of sampling techniques in influencing the accuracy of hypothyroidism diagnosis in this context. Through this approach, the researchers aimed to contribute to the refinement of diagnostic processes and the reduction of risks associated with dialysis interventional trials.

This article discusses treatment prediction for thyroid illness. It talks about how machine learning can be used to forecast when levothyroxine dosage adjustments are necessary. Levothyroxine is a medicine used to treat hypothyroidism. A dataset of patient data from a hospital in Naples, Italy was created by the authors. After evaluating ten distinct machine learning algorithms for treatment trend prediction, they discovered that the Extra-Tree Classifier, was the most precise.

CHAPTER 3

Proposed System

Algorithm 1 Model Performance Prediction

Input: Dataset X
Output: Model Performance Prediction
procedure MAIN PROCEDURE
    Step 1: Data Cleaning
    Step 2: Data Preprocessing
        Feature Selection
        Balancing
    Step 3: Splitting the Dataset
    Perform k-fold cross Validation
        Randomly Split X into k partitions:
        do: for k times
            $data.test = X'$
            $data.train = X - X'$
            Train a machine learning model m on $data.train$
            Predict the nodule malignancy on $data.test$ using m
            Save the current prediction result $Y', Y'', ...$
        end :
        Predict the nodule malignancy on $data.test$ using m
        Calculate the Prediction result Y
    Step 4: Measure the model prediction by comparing Y with the true nodule malignancy
end procedure

3.1 Data Collection and Preprocessing

3.1.1 Data Collection: Thyroid Clean Dataset

The initial step involves acquiring a comprehensive dataset for thyroid cancer prediction. In this study, the dataset used is named "Thyroid Clean," obtained from a reliable reference source (to be appropriately cited). The dataset is meticulously curated and preprocessed to ensure its cleanliness and reliability in capturing relevant features for thyroid cancer analysis.

3.1.2 The Proposed Model Feature Selection and Balancing

Following data collection, the dataset undergoes a rigorous preprocessing phase, consisting of feature selection and balancing procedures.

3.1.2.1 Feature Selection

Feature selection is a critical aspect of building an effective predictive model. It involves choosing the most relevant variables or features that significantly contribute to the prediction task. In this context, the features relevant to predicting thyroid malignancy are carefully identified and selected. The selection process is driven by the need to focus on the most informative aspects of the dataset, optimizing the model's ability to discern patterns associated with cancer presence.

3.1.2.2 Balancing

Imbalance in the dataset, where one class (e.g., malignant or benign thyroid nodules) is significantly underrepresented compared to the other, can pose challenges for machine learning models. To address this, balancing techniques are applied. These techniques ensure that the model is exposed to an equitable representation of both classes, preventing bias and enhancing its ability to generalize across different scenarios.

3.1.3 Dataset Splitting and Model Training

The preprocessed dataset is split into training and testing subsets. To ensure a comprehensive evaluation, a variant of k-fold cross-validation is employed. The dataset is randomly partitioned into k subsets, and the algorithm iterates k times, each time designating one of the subsets as the testing set (denoted as X') and the remaining data (X - X') as the training set. This approach allows for an unbiased assessment of the model's performance on unseen data.

3.1.4 Model Prediction and Evaluation

A machine learning model, denoted as m, is trained on the training dataset and subsequently used to predict nodule malignancy on the reserved test dataset (X').Multiple result are combined to obtained the model predictions (Y). The model's predictions (Y) are then compared with the true nodule malignancy labels to evaluate the accuracy, precision, recall, and other relevant metrics indicative of its predictive performance.

Through this systematic approach of data collection, preprocessing, feature selection, balancing, and model evaluation, the study aims to develop a robust machine learning model for thyroid cancer prediction, contributing to advancements in accurate diagnosis and treatment planning.

CHAPTER 4

Experimental Results and Discussions

To ascertain the accuracy or efficacy of a machine learning model, it is imperative to conduct a comprehensive assessment of its performance when applied to generate predictions on real-world data.

The metrics used for model validation include accuracy, precision, recall, F1 score, and support. In detail, accuracy indicates the model's correctness, representing the fraction of the test dataset for which the model provides correct predictions. The formula for accuracy is defined as follows:

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

Precision and recall quantify the rates of True Positive (TP) and True Negative (TN) respectively. Precision assesses the classifier's ability to avoid labeling a genuinely negative instance as positive, given by:

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

Recall measures the model's sensitivity, representing the ratio of correct predictions for a class to the total occurrences of that class:

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

The F1 score combines precision and recall into a single metric and is defined as:

$$F1 \text{ score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Additionally, the support metric indicates the number of occurrences of each class in the true responses, providing context for the other metrics.

	Accuracy	Precision	Recall	F1-score
LR	0.7470	0.79	0.8383	0.8124
SVM	0.7329	0.8113	0.7818	0.7962
KNN	0.6882	0.7500	0.8	0.7741
Random Forest	0.7620	0.80	0.85	0.82
AdaBoost	0.7652	0.78	0.92	0.84
GBM	0.7532	0.7651	0.8363	0.8263

FIGURE 4.1: Performance.

As we can see, the AdaBoost model has the highest accuracy, precision, recall, and F1-score of all the models. This means that it is the best model for predicting the correct outcome.

Here are some of the reasons why AdaBoost may be performing better than the other models:

AdaBoost is an ensemble method, which means that it combines the predictions of multiple weak learners to create a single strong learner. This can help to reduce overfitting and improve the overall performance of the model. AdaBoost is also able to learn from its mistakes. When a weak learner makes an incorrect prediction, AdaBoost increases the weight of that learner so that it has more influence on the final prediction. This

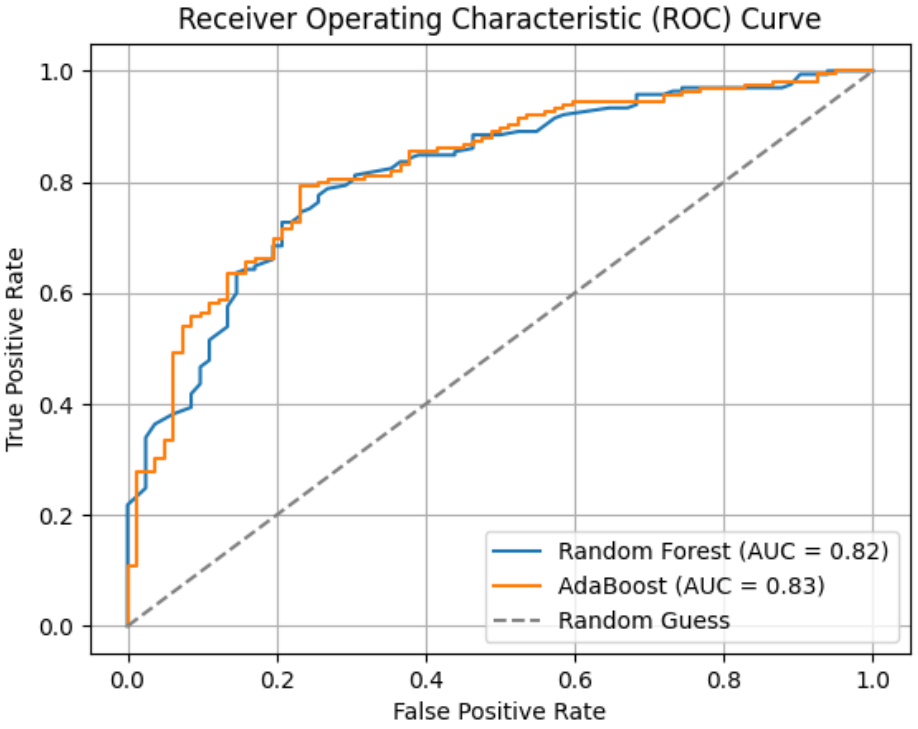

FIGURE 4.2: random forest and adaboost roc.

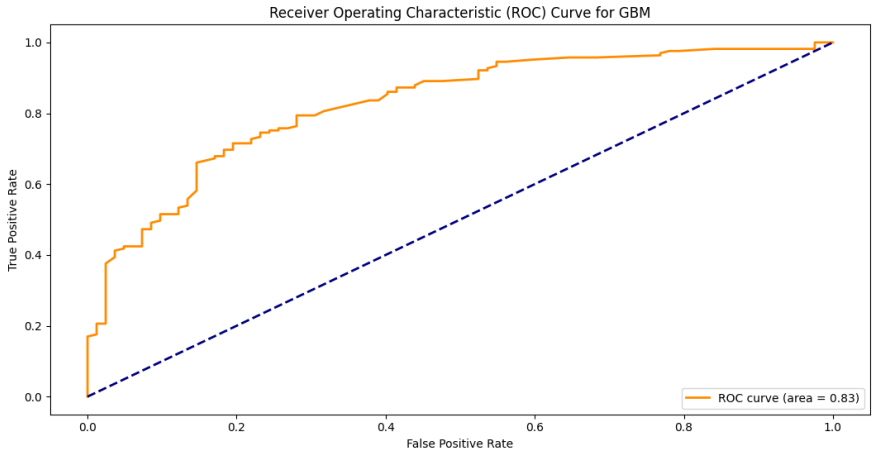

FIGURE 4.3: gbm roc.

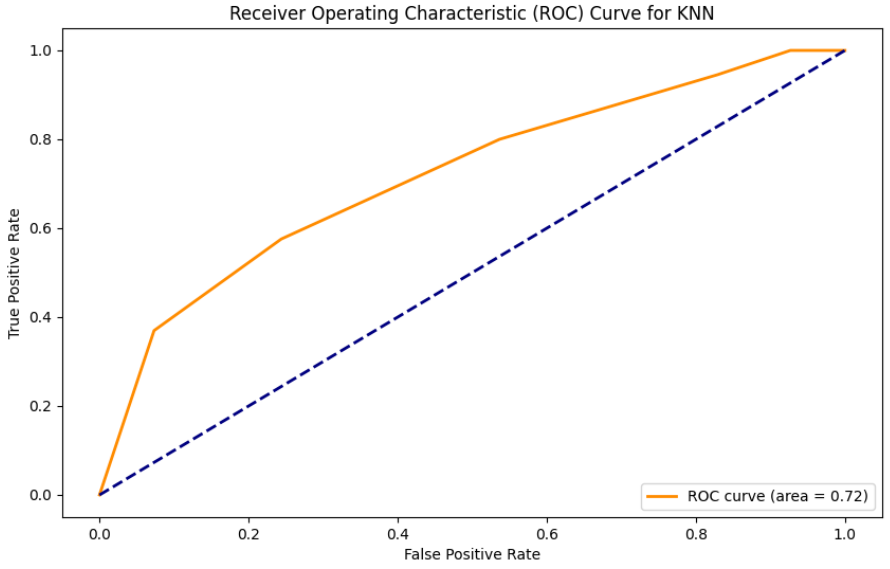

FIGURE 4.4: knn roc curve.

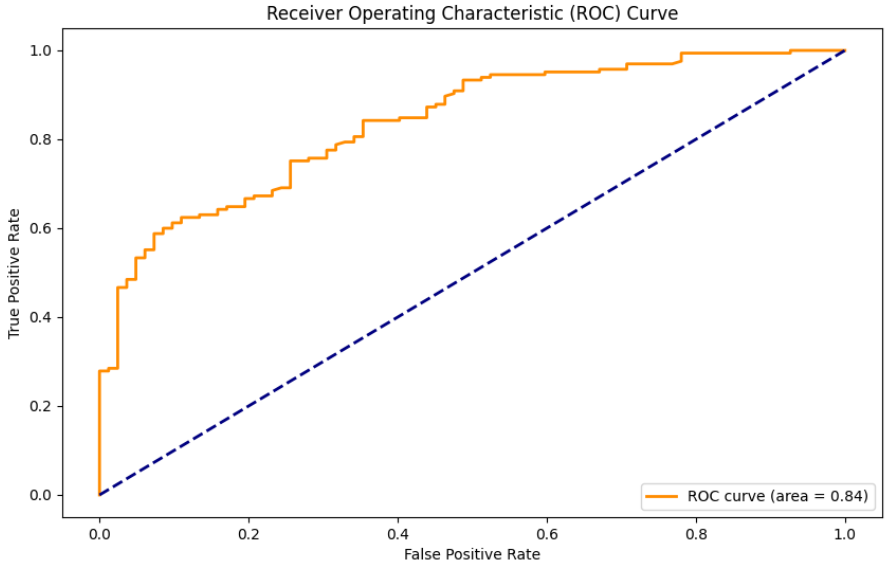

FIGURE 4.5: LR Roc Curve.

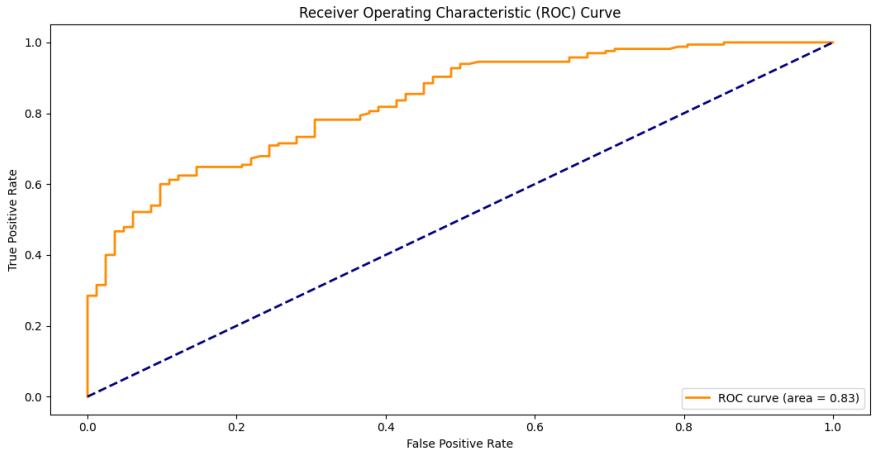

FIGURE 4.6: svm roc curve.

helps to improve the accuracy of the model over time. Overall, the AdaBoost model is a powerful and versatile machine learning algorithm that can be used to solve a wide variety of problems. If you are looking for a model that can provide high accuracy and performance, AdaBoost is a great option to consider.

In addition to the AdaBoost model, the Random Forest and GBM models also performed well. These models are both ensemble methods that are able to learn from their mistakes. The Random Forest model is particularly well-suited for handling high-dimensional data, while the GBM model is known for its ability to handle complex nonlinear relationships.

The Logistic Regression and SVM models performed well in terms of accuracy, but they were not as good as the AdaBoost, Random Forest, or GBM models in terms of precision, recall, and F1-score. This suggests that these models may be more prone to overfitting or underfitting.

The KNN model performed the worst of all the models. This is likely because it is a non-parametric model that does not make any assumptions about the underlying data distribution. This can make it difficult to learn complex relationships between the input features and the target variable.

Overall, the AdaBoost model is the best model for predicting the correct outcome. However, the Random Forest and GBM models are also good options, and they may

be a better choice for some tasks. The Logistic Regression and SVM models are also good choices, but they may be more prone to overfitting or underfitting. The KNN model is not a good choice for this task.

CHAPTER 5

Conclusion and Future Work

The findings of this study hold significant promise for improving the early diagnosis and treatment of thyroid cancer. Accurate prediction of thyroid cancer can aid in timely intervention, personalized treatment plans, and improved patient outcomes. The AdaBoost model, in particular, could be implemented as a decision-support tool for clinicians, assisting in the evaluation of thyroid nodules and guiding appropriate diagnostic procedures.

Future Directions

Further research could focus on:

Expanding the dataset: Incorporating a larger and more diverse dataset could enhance the generalizability and robustness of the machine learning models.

Refining feature selection: Identifying the most relevant clinical features could further improve the predictive accuracy of the models.

Exploring deep learning approaches: Investigating the application of deep learning techniques, particularly convolutional neural networks (CNNs), could enhance the models' ability to extract complex patterns from medical images.

Developing a clinical decision-support system: Integrating the AdaBoost model into a clinical decision-support system could facilitate its seamless implementation in medical practice.

Evaluating the impact on patient outcomes: Conducting clinical studies to assess the impact of the machine learning models on patient outcomes could provide valuable insights into their real-world effectiveness

Overall, this study represents a significant step forward in the application of machine learning for thyroid cancer prediction. The potential benefits of these models for patient care and healthcare decision-making are substantial, and further research is warranted to fully realize their impact.

References

[1] S. Peng, Y. Liu, W. Lv, L. Liu, Q. Zhou, H. Yang, J. Ren, G. Liu, X. Wang, X. Zhang, *et al.*, "Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study," *The Lancet Digital Health*, vol. 3, no. 4, pp. e250–e259, 2021.

[2] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, X. Xin, C. Qin, X. Wang, J. Li, *et al.*, "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study," *The Lancet Oncology*, vol. 20, no. 2, pp. 193–201, 2019.

[3] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, and S. Mukherjee, "A comparative study on thyroid disease detection using k-nearest neighbor and naive bayes classification techniques," *CSI transactions on ICT*, vol. 4, pp. 313–319, 2016.

[4] G. Akgül, A. A. Çelik, Z. E. AYDIN, and Z. K. ÖZTÜRK, "Hipotiroidi hastalığı teşhisinde sınıflandırma algoritmalarının kullanımı," *Bilişim Teknolojileri Dergisi*, vol. 13, no. 3, pp. 255–268, 2020.