Semi-Supervised Machine Learning Approaches for Thyroid Disease Prediction and its Integration With the Internet of Everything

Melih Agraz

Division of Applied Math, Brown University (USA)

Department of Statistics, Giresun University (Turkey)

* Corresponding author: melih_agraz@brown.edu; melih.agraz@giresun.edu.tr

Received 29 June 2023 | Accepted 7 June 2024 | Early Access 26 July 2024

Abstract

Thyroid disorders are critical conditions that considerably affect a person's general health, and may lead to additional health complications. Notably, these conditions often remain undetected in individuals who show "normal" results on traditional thyroid function tests. To enhance the diagnostic accuracy for thyroid disorders, such as hypothyroidism and hyperthyroidism, this study leveraged digital health records and explored semi-supervised learning methods. We intentionally removed the labels from subjects initially categorized as "normal," incorporating them into our dataset as unlabeled data. The goal was to overcome the limitations of conventional diagnostic techniques, which may fail to detect subtle imbalances in thyroid hormones. In pursuit of this objective, we employed a combination of semi-supervised learning methods, namely FixMatch, Co-training, and self-training, in conjunction with supervised learning algorithms, specifically Naive Bayes and logistic regression. Our findings indicate that the FixMatch algorithm surpassed traditional supervised learning methods in various metrics, including accuracy (0.9054), sensitivity (0.9494), negative predictive value (0.9365), and F1 score (0.9146). Additionally, we propose a framework for integrating these diagnostic tools into the Internet of Everything (IoE) to promote early detection and facilitate improved healthcare outcomes. This research highlights the potential of semi-supervised learning techniques in the diagnosis of thyroid disorders and offers a roadmap for harnessing the IoE in healthcare advancement.

Keywords

FixMatch, IoE Medical Systems, Machine Learning, Thyroid Disease.

DOI: 10.9781/ijimai.2024.07.006

I. Introduction

A. Research Background and Motivations

The Internet of Everything (IoE) is increasingly recognized as the future of the internet. It enables intelligent interconnections among diverse elements, including things (devices), processes, people, and data, as illustrated in Fig. 1. This concept harnesses advanced technologies like 5G and artificial intelligence (AI) to enhance internet connectivity, making it faster, smarter, and more tailored to individual needs [1], [2]. Furthermore, the advancement in smart devices has given rise to a new dimension of the IoE, emphasizing the connection of all devices to the internet [3].

The Internet of Everything (IoE) holds the potential to revolutionize numerous industries and various aspects of daily life. For instance, it can be utilized to enhance transportation efficiency by integrating vehicles, traffic signals, and road sensors. Similarly, IoE can aid in improving energy efficiency through the interconnectedness of buildings, appliances, and power grids. Moreover, it has the capability to transform healthcare by establishing connections among patients, doctors, and medical devices. This study discusses the development of machine learning methods, specifically semi-supervised learning (SSL) and supervised learning (SL) algorithms, for monitoring and detecting thyroid diseases. Additionally, it delves into the integration of these methods with the IoE, highlighting their potential synergistic benefits.

B. Research Objectives

In this study, we aimed to investigate several critical questions: Can we predict subtypes of thyoid diseases? Can individuals with normal thyroid levels still be suffering from thyroid disorders? Is it possible to use pseudo-labels to develop robust machine learning models for early detection? Beyond addressing these medical queries, we are also proposing IoE models for researchers in the field of thyroid diseases. Specifically, we explore how pseudo-labels can enhance

Please cite this article as:
M. Agraz. Semi-Supervised Machine Learning Approaches for Thyroid Disease Prediction and its Integration With the Internet of Everything, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 7, pp. 38-47, 2024, http://dx.doi.org/10.9781/ijimai.2024.07.006

the effectiveness of machine learning models. As a result, this study focuses on predicting hypothyroid and hyperthyroid diseases in thyroid patients. We utilized the UC Irvine (UCI) thyroid dataset [4], comprising three classes (hypo/hyper/normal), for this prediction. To consider the possibility of individuals with normal thyroid function actually being euthyroid, we treated them as unlabeled data in our analysis i.e., we removed 'normal' labels, we introduced pseudo-labels during the semi-supervised learning process. This approach allowed us to validate the model's performance by comparing predicted labels with the true class distributions identified through subsequent validation steps. Our comprehensive study employs both SL techniques, including Naive Bayes and Logistic Regression, and SSL approaches, such as Fixmatch, Co-training, and self-training. We conducted experiments using both labeled (hypo/hyper) and unlabeled data to evaluate the effectiveness of our proposed SSL methods in comparison with traditional SL methods. Additionally, we encourage researchers to utilize our findings in their IoE research studies, as the insights gained could be pivotal in developing early detection tools for thyroid disorders.

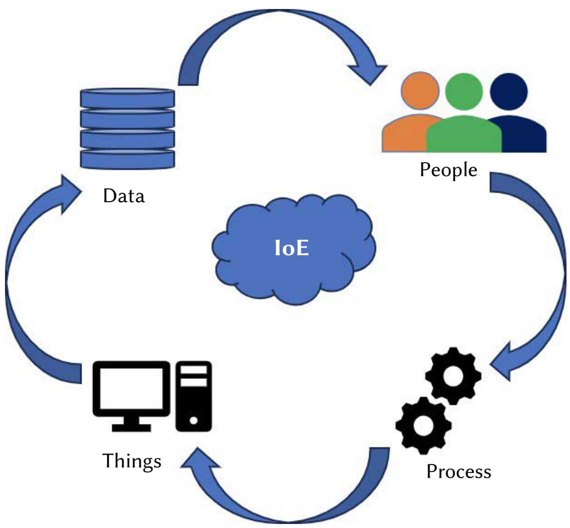

Fig. 1. Integral Elements of the Internet of Everything (IoE). This illustration depicts the interconnected nature of IoE, highlighting the continuous interaction among four critical components: Data, People, Processes, and Things. It emphasizes the dynamic flow of information and the synergy that fuels the functionality of IoE ecosystems.

The structure of the study is as follows: Section II presents a literature review. Section III explains the workflow of thyroid prediction and the proposed IoE application for thyroid disorders. Section IV details the data, features, machine learning methods used in the study, and performance measures, including a subsection for Discussion where the results are analyzed. Finally, Section V discusses the conclusions and future directions of this research.

II. Literature Review

Numerous studies have combined the Internet of Everything (IoE) with medicine [5]–[10]. For instance, Roy and Singh [9] explored the use of the Internet of Medical Things (IoMT) for remote patient monitoring during the COVID-19 pandemic. Their machine learning model achieved a remarkable F1-score of 0.997 in predicting infections. Ahamed et al. [10] developed a Cardiovascular Disease Prediction System using IoT and machine learning for early diagnosis of heart diseases, achieving notable effectiveness with a hyperparameter-tuned Random Forest model. Sivaparthipan et al. [8] proposed a robotic system employing machine learning for treating Parkinson's disease, focusing on processing large datasets to predict patient mobility patterns. In their study on predictive analytics, Rghioui et al. [5] introduced a unique system for monitoring diabetes patients. They evaluated four different machine learning algorithms (Naive Bayes, Random Forest, OneR, and J48), ultimately selecting Random Forest as the most effective prediction algorithm. Maghawry et al. [6]'s study allowed for the development of advisory systems that combine biosensor data with historical medical and social network data to provide accurate alerts and recommendations for various diseases. Additionally, Raja and Chakraborty [7] enhanced medical access in remote areas by using wearable sensors to collect health metrics, stored in cloud storage, and monitored by doctors for effective treatment.

Thyroid gland disorders are significant and frequently encountered conditions that are often overlooked in clinical diagnoses. In areas where goiters are common, about 15 to 30% of adults suffer from this issue [11], [12]. Based on hormonal levels, individuals with thyroid dysfunction are categorized into hypothyroidism, euthyroidism, and hyperthyroidism [13]. These disorders are typically associated with either excessive (hyper) or insufficient (hypo) secretion of thyroid hormones. Factors such as previous thyroid surgery, ionizing radiation exposure, chronic thyroid inflammation, iodine deficiency, enzyme deficiencies, and certain medications can lead to hypothyroidism [14]. Graves' disease is a common cause of hypothyroidism, characterized by the body's production of proteins that stimulate excessive thyroid hormone production [14], [15]. Euthyroidism refers to a state of normal thyroid hormone production and serum levels [16], [17]. In this context, understanding thyroid gland functional data is crucial for accurate diagnosis. Ozyılmaz and Yıldırım [14] highlighted the importance of interpreting this data in diagnosing thyroid diseases, demonstrating the effectiveness of feedforward neural network structures. There have been numerous studies in machine learning to explore both SL [10], [14], [18]–[21] and SSL [22], [22]–[26] approaches. Keramidas et al. [18] proposed a k-nearest neighborhood (k-NN) algorithm to detect thyroid nodules in ultrasound images. Razia [19] introduced a model using both unsupervised and SL methods for thyroid disease diagnosis. Zhang et al. [24] developed a semi-supervised graph convolutional deep learning model, Semi-GCNs-DA, for cross-device adaptation in identifying thyroid nodules. Turk et al. [23] suggested semi-supervised methods for detecting thyroid nodules in ultrasound data, proposing an encoder-based neural network model with high recall and sensitivity. Yang et al. [25] developed a dual-path semi-supervised conditional generative adversarial network (DScGAN) model for thyroid nodule detection, demonstrating its effectiveness in SSL with limited medical datasets and insufficient labels. Requena et al. [26] introduced an innovative SSL approach using Encoder-Decoder Convolutional Neural Networks for Human Activity Recognition (HAR) in healthcare. This method combines public labeled and extensive private unlabeled raw sensor data, enhancing the model's generalization to real-world scenarios. They demonstrated its effectiveness in a case study involving overweight patients, accurately classifying movement patterns from large volumes of accelerometer data. Martin et al. [22] addressed the challenge of selecting an appropriate distance metric for clustering algorithms like k-means. They introduced a semi-supervised clustering algorithm that learns a linear combination of multiple dissimilarities, utilizing incomplete knowledge through pairwise constraints. Enhanced with a regularization term to prevent overfitting, this method showed superior performance in identifying tumor samples using gene expression profiles, outperforming standard semi-supervised techniques and those relying on a single dissimilarity, particularly in noisy environments.

In this study, we achieved the highest accuracy of 90.54% using the FixMatch method, which incorporates several improvements from previous studies. For instance, Razia et al. [27]achieved an accuracy of 73.29% using a combination of SVM, Multiple Linear Regression, Naïve Bayes, and Decision Trees with 22 attributes. Chaubey et al. [28] reached 88.54% using Decision Trees, KNN, and Logistic Regression with a reduced attribute set of 5. Most recently, Peya et al. [29] reported an impressive accuracy of 95.85% using KNN, Naïve Bayes, and Decision Trees with 22 attributes. Our use of the semi-supervised learning approach, particularly FixMatch, allows for significant improvements in scenarios where labeled data is limited, thereby providing a robust solution in resource-constrained settings.

Despite significant advancements in combining IoE with medical applications, current studies predominantly rely on fully supervised learning methods, which require extensive labeled data. This limitation restricts their applicability in scenarios with limited labeled data. Moreover, few studies have explored SSL techniques, which can leverage unlabeled data to improve diagnostic accuracy. Additionally, the integration of SSL methods with IoE frameworks remains underexplored, particularly in the context of real-time monitoring and early detection of thyroid diseases. These gaps highlight the need for research that combines SSL methods with IoE to enhance diagnostic accuracy and enable continuous health monitoring.

III. Research Methodology

This study established a comprehensive research model aimed at addressing the challenges associated with thyroid disease prediction. Our methodology encompasses experimental design, performance evaluation, and the implementation of various machine learning techniques.

Workflow of the Thyroid Prediction Study. The study's workflow involved a comparative analysis of SL and SSL models, as depicted in Fig. 2. In this process, we discarded the normal labels from the thyroid dataset, which comprises three labels (hypo/hyper/normal), resulting in the creation of unlabeled data. Consequently, this left us with labeled data (hypo/hyper) and unlabeled data. The SL approach was exclusively applied to the labeled data, while the SSL approach was utilized for both labeled and unlabeled data. We compared the performance of these two methods to assess their effectiveness. To address the issue of significant imbalance within the model, random undersampling was conducted during the training phase to balance the training data. The machine learning methods employed in the SL approach included Logistic Regression and Naive Bayes. In contrast, the SSL approach incorporated FixMatch, Co-training, and self-training techniques.

Workflow of the Proposed Internet of Everything Thyroid Study. While our current machine learning model operates with laboratory data (electronic health records in the Fig. 3), we propose a future-oriented model that integrates data from wearable devices, as depicted in Fig. 3.

1. **Continuous Monitoring of Thyroid-Related Physiological Parameters**: Our envisioned device is designed for the ongoing monitoring of various physiological indicators critical to thyroid function. These include heart rate, body temperature, and physical activity levels. This continuous data stream provides a solid foundation for in-depth analysis, making it an effective tool for real-time health tracking.
2. **Data-Driven Analysis through Machine Learning Algorithms**: The device goes beyond simple data gathering. It incorporates sophisticated machine learning algorithms to analyze and interpret the collected data. These algorithms are specifically tailored to evaluate the thyroid's functional status, offering an automated yet highly accurate analysis.
3. **Immediate Alerting Mechanisms for Anomalies**: In case of detection of abnormalities indicative of thyroid dysfunction, the device promptly issues alerts to the user. This rapid response facilitates early lifestyle adjustments or urgent medical consultations, helping to prevent potential health complications.
4. **Secure Data Management via Cloud Infrastructure**: Recognizing the paramount importance of data security in healthcare, our system employs robust cloud storage for data management. This approach guarantees the safety and privacy of user data and supports long-term health monitoring and analysis.
5. **User Interface Designed for Accessibility and Customization**: To enhance user interaction, the system features a user-friendly mobile app and web dashboard. These interfaces not only provide access to real-time and historical data but also enable users to add additional information, like medication and symptoms. This extra layer of data enriches the precision and dependability of the predictive models used by the device.

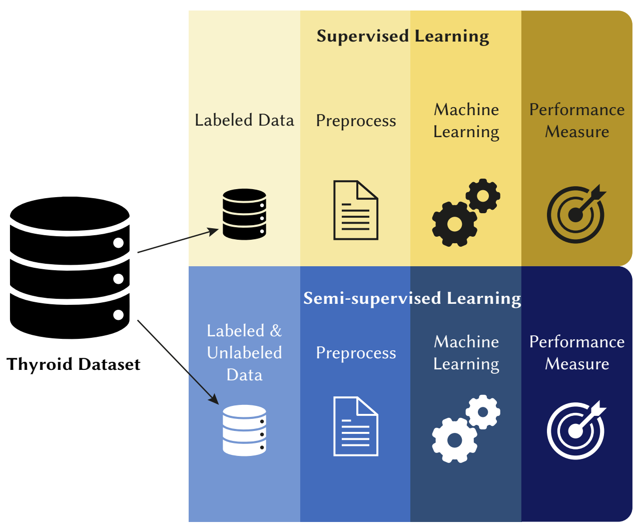

Fig. 2. Comparative Main Workflow of Supervised vs. Semi-Supervised Learning Using the Thyroid Dataset. This flowchart illustrates the study's methodology, highlighting the differences in data preprocessing, machine learning application, and performance measurement stages between the supervised and semi-supervised learning paradigms when applied to thyroid health data.

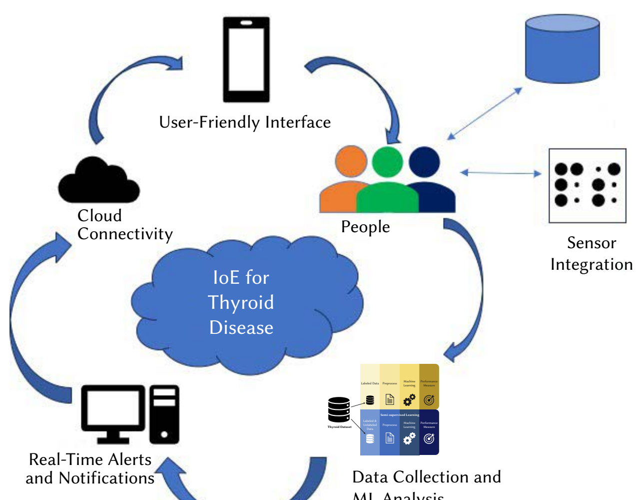

Fig. 3. Workflow of the IoE Framework for Thyroid Disease Management. This diagram presents a comprehensive IoE strategy, highlighting the essential steps from sensor- based data acquisition, through machine learning-powered analysis, to timely alerts and notifications, ultimately leading to a user-centric interface enhanced by cloud technology.

IV. Experimental Design and Performance Evaluation

A. Dataset Collection

In this study, we leveraged the thyroid disease dataset, freely accessible to the public, from the UCI repository [4]. This dataset comprises 10 distinct databases, all sourced from the Garvan Institute in Sydney, Australia. Our specific focus was on the thyroid-related data within this collection of datasets available on the UCI repository [4]. Additionally, users can access the data at Github repository at https://github.com/melihagraz/SSL_AND_SL/tree/main/data.

The dataset offers comprehensive information on 7,200 patients and encompasses 21 features, as outlined in Table I. Patients are categorized into three groups: normal (1), hyperthyroidism (2), and hypothyroidism (3). The distribution of these categories is as follows: 166 patients are classified as normal, 368 as having hyperthyroidism, and 6666 as having hypothyroidism, as illustrated in Fig. 4. Regarding data preprocessing, the dataset was found to be complete, requiring no imputation for missing values. Additionally, we retained all observations, and we worked with the normalized data.

1. Data Security and Ethics

While the IoE offers unprecedented opportunities for real-time data collection and analysis in thyroid disease research, it also poses challenges in terms of data compatibility and scalability. Existing medical databases may require substantial modifications to integrate with IoE devices. Additionally, the sheer volume of data generated could potentially overwhelm current data storage solutions.

The growing digitalization of healthcare raises a variety of ethical opportunities and challenges [30]. As noted by Jacquemard et al. [31], ethical values should inform all stages of the electronic health records lifecycle, from design and development to implementation and practical application. These ethical considerations become particularly important when discussing the integration of the IoE with medical data. Obtaining informed consent for data collection becomes increasingly complex, especially when dealing with sensitive health information. Moreover, there is a risk that the benefits of this advanced technology may not be equally accessible across all populations, potentially exacerbating existing healthcare inequalities.

Data security is a critical concern when integrating IoE technologies into medical research. This integration involves the transmission and storage of sensitive personal health information, which may be vulnerable to unauthorized access or misuse [32]. The expanded network of interconnected devices amplifies the potential points of failure or unauthorized access, making the protection of sensitive health data increasingly challenging. Therefore, rigorous encryption methods and robust cybersecurity protocols are essential to mitigate these risks.

2. Potential Challenges

There could be potential challenges of IoE with thyroid disease research.

1. Technical Challenges: Integrating IoE in thyroid disease research involves handling various types of devices and data sources. Ensuring compatibility, interoperability, and continuous data flow between these devices and systems can be complex. There may also be issues related to data accuracy, reliability, and real-time processing.
2. Data Management and Analysis: The sheer volume of data generated by IoE devices poses a significant challenge. Efficiently storing, managing, and analyzing this data to extract meaningful insights for thyroid disease research requires advanced data analytics tools and expertise in big data management.

TABLE I. Dataset Variables Overview. This Table Lists the Features of the Thyroid UCI Dataset, Including Variable Names, Their Descriptions, and Classification as Continuous or Discrete Data Types. The Dataset Encompasses Demographic Details, Treatment History, and Clinical Test Results, Which Are Pivotal for Thyroid Disease Analysis [4]

Variable	Description	Data types
Age	Age of the patient	Continuous
Sex	Sex of the patient	Discrete
On_thyroxine	On thyroxine	Discrete
Query_on_thyroxine	Query on thyroxine	Discrete
On_antithyroid _medication	On antithyroid medication	Discrete
Sick	Sick	Discrete
Pregnant	Pregnant	Discrete
Thyroid_surgery	Thyroid surgery	Discrete
I131_treatment	I131 treatment	Discrete
Query_hypothyroid	Query hypothyroid	Discrete
Query_hyperthyroid	Query hyperthyroid	Discrete
Lithium	Lithium	Discrete
Goitre	Goitre	Discrete
Tumor	Tumor	Discrete
Hypopituitary	Hypopituitary	Discrete
Psych	Psych	Discrete
TSH	Amount of TSH	Continuous
T3	Amount of T3	Continuous
TT4	Amount of TT4	Continuous
T4U	Amount of T4U	Continuous
FTI	Amount of FTI	Continuous

3. Infrastructure and Cost: Setting up the infrastructure for IoE integration in thyroid disease research can be costly. This includes the cost of devices, data storage solutions, analytics software, and securing the network. Additionally, there is a need for continuous maintenance and upgrades.
4. User Acceptance and Training: For IoE to be effective in research, it's essential that all collaborators (researchers, clinicians, patients) understand and accept the technology. This might require extensive training and education to ensure proper usage and interpretation of IoE-generated data.

B. Outcome and Features

1. Outcome

In this study, the output of the raw data has three labels, due to the possibility of hypo/hyperthyroidism in normal individuals, the labels of individuals with a normal condition were removed. Therefore, only labeled (hypo/hyper) for SL analysis and labeled and unlabeled data for SSL analysis were used.

2. Features

The analyzes performed in this study took advantage of the comprehensive set of features listed in Table I, allowing for a more comprehensive and in-depth exploration of the research objectives. In addition, considering that the dataset is sufficient, feature selection or feature reduction methods were not applied.

We can explain the variables and classify them as the following format.

- **Patient Details**
  - **Age:** Age of the patient. This is a continuous variable, meaning it can take any numerical value within a given range.
  - **Sex:** Sex of the patient (usually Male or Female). This is a discrete variable, meaning it can only take certain predefined values.
- **Treatment and Medication**
  - **On_thyroxine:** Indicates whether the patient is on thyroxine medication. Discrete variable (usually Yes or No).
  - **On_antithyroid_medication:** Specifies if the patient is taking medication for thyroid issues. Discrete variable (usually Yes or No).
- **Health Conditions**
  - **Sick:** Indicates whether the patient is currently sick. Discrete variable (usually Yes or No).
  - **Pregnant:** Indicates whether the patient is pregnant. Discrete variable (usually Yes or No).
  - **Thyroid_surgery:** Indicates whether the patient has had thyroid surgery. Discrete variable (usually Yes or No).
  - **I131_treatment:** Indicates whether the patient has undergone Iodine-131 treatment, usually for thyroid cancer or hyperthyroidism. Discrete variable (usually Yes or No).
- **Queries and Concerns**
  - **Query_hypothyroid:** Indicates whether there is a question or query about the patient having hypothyroidism. Discrete variable (usually Yes or No).
  - **Query_hyperthyroid:** Indicates whether there is a question or query about the patient having hyperthyroidism. Discrete variable (usually Yes or No).
  - **Query_on_thyroxine:** Indicates whether there is a question or query about the patient being on thyroxine. Discrete variable (usually Yes or No).
- **Other Medications and Conditions**
  - **Lithium:** Indicates whether the patient is on lithium, which can affect the thyroid. Discrete variable (usually Yes or No).
  - **Goitre:** Indicates whether the patient has a goitre (enlarged thyroid). Discrete variable (usually Yes or No).
  - **Tumor:** Indicates whether the patient has a tumor. Discrete variable (usually Yes or No).
  - **Hypopituitary:** Indicates whether the patient has hypopituitarism (underactive pituitary gland). Discrete variable (usually Yes or No).
  - **Psych:** Likely indicates whether the patient has a psychiatric condition. Discrete variable (usually Yes or No).
- **Hormone Levels**
  - **TSH (Thyroid-Stimulating Hormone):** Amount of TSH in the blood. Continuous variable.
  - **T3 (Triiodothyronine):** Amount of T3 in the blood. Continuous variable.
  - **TT4 (Total Thyroxine):** Amount of TT4 in the blood. Continuous variable.
  - **T4U (Thyroxine Uptake):** Amount of T4U in the blood. Continuous variable.
  - **FTI (Free Thyroxine Index):** Amount of FTI in the blood. Continuous variable.

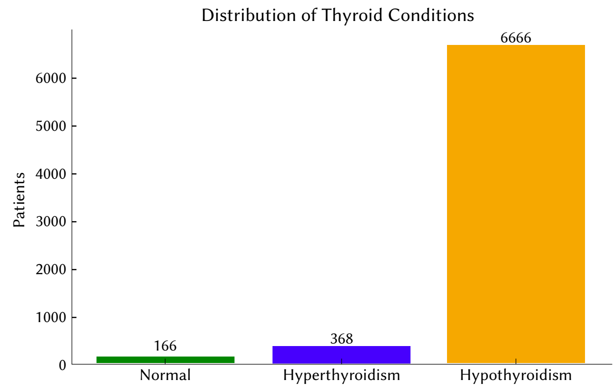

Fig. 4. Distribution of Patient Diagnoses in the Thyroid Dataset. The bar chart presents the number of patients categorized into three groups: those diagnosed with hyperthyroidism, hypothyroidism, and normal (unlabeled) labels group.

C. Data Imbalance

We unlabeled normal patients from the thyroid data, and we observed an imbalance in the ratio of Hypo/Hyperthyroid patients, with a ratio of 1:18. To address this imbalance, random undersampling was performed during the training phase by randomly selecting samples from the majority class, resulting in a balanced ratio of 1:1 in the training of the machine learning algorithm.

D. Machine Learning Methods

1. Supervised Learning

Supervised learning (SL) is a widely used learning method in the field of data science. In this method, a machine learning model learns to predict the associated output value using a given input data. In this study, the following two SL methods were preferred.

1. Logistic regression.
2. Naive Bayes.

Consensus methods for choosing the type of statistical model have not been established [33]. As a result, we selected logistic regression and Naive Bayes for their respective advantages. Logistic regression is highly valued in medical research for its interpretability, suitability for binary outcomes, ability to adjust for confounders, flexibility, and the appropriateness of predicted probabilities. It is also recognized for its robustness in assessing model fit and accuracy. These attributes make logistic regression an invaluable tool for handling the complexities of medical data effectively [34]. On the other hand, the Naive Bayes algorithm is acclaimed for requiring minimal training data, straightforward computation, ease of implementation, time efficiency, and its ability to handle large datasets and incomplete data. It is also resilient to irrelevant features and data noise [35]–[38]. For these reasons, we chose logistic regression and Naive Bayes. Given the limited variety of models in semi-supervised methods compared to supervised methods, we tried to select effective methods such as FixMatch and self-training among the limited options available.

2. Semi-Supervised Learning

Semi-supervised learning (SSL) is a learning method used in the field of machine learning. This method works on a dataset that includes both labeled (with correct output values) and unlabeled (without correct output values) data samples. In this system, unlabeled data is used either to improve the model's performance or to augment labeled data.

Self-training method. This method serves as the foundational approach for SSL. As illustrated in Fig. 5, the model is initially trained using labeled data, same to the procedure in SL. Subsequently, this trained model is utilized to predict the labels for the unlabeled data.

Observations with a prediction confidence exceeding a predetermined threshold are identified and assigned pseudo-labels. These pseudo-labels are incorporated into the existing pool of labeled data, and the model undergoes retraining. The iterative process persists until a stopping criterion is met, when no more pseudo-labels can be predicted in this study. The final trained model is then used to assess its performance on test data.

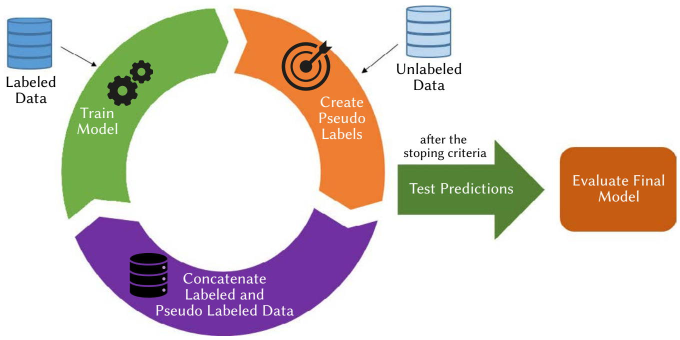

Fig. 5. Overview of the Self-training method workflow. This diagram illustrates the cyclical process of training a model with labeled data, generating pseudo-labels for unlabeled data, testing predictions after meeting the stopping criteria, and evaluating the final model performance, followed by concatenation of labeled and pseudo-labeled data for further iterations.

FixMatch method. FixMatch is one of the SSL methods that focuses on effectively utilizing a small amount of labeled data along with a larger amount of unlabeled data. It trains the model using both weak and strong augmentations, enabling the model to make more general and robust predictions. In FixMatch, "augmentation" refers to data augmentation, which is typically used to increase data diversity. The concept of "weak" and "strong" augmentations is employed in FixMatch. Weak augmentation makes slight modifications to the data, such as rotating or slightly zooming an image, while preserving its basic characteristics. Strong augmentation, on the other hand, introduces more radical changes like cropping, flipping, or significant color adjustments, exposing the model to a greater variety of data. In our approach, we apply weak augmentation to add a low level of Gaussian noise to the data, and strong augmentation to introduce a higher level of Gaussian noise. Conventional SSL methods use both labeled and unlabeled data, while this method encourages consistency between predictions obtained by applying different data augmentations on the same data points. This method, developed by Sohn et al. [39], is a highly effective SSL method that enables the learning of the system even with a single labeled data point. It has been proven to be more effective than other SSL methods, and it is used to build high-accuracy models with a limited number of labeled data.

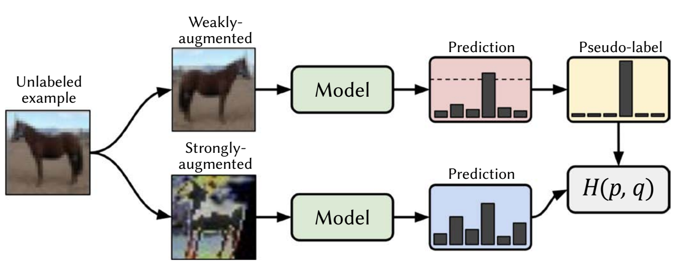

Fig. 6. Schematic representation of the FixMatch algorithm applied in semi-supervised learning [39].

The FixMatch method, as applied in our study of thyroid tabulated data [39], consists of the following steps, as seen in the general form of FixMatch in Fig. 6 and the adaptation of the FixMatch method for thyroid disease prediction in Fig. 7:

1. **Data Augmentation Function:** The most critical aspect of FixMatch is the data augmentation phase. Unlabeled data are augmented using two distinct strategies: strong and weak augmentation. Strong augmentation aims to make the data more challenging to learn, while weak augmentation aims to simplify the learning process.
2. **Training on Labeled Data:** The model is initially trained on labeled data, and the loss, denoted as L_s, is calculated.
3. **Pseudo-labeling:** The $H(p, q)$ (cross entropy loss) component shown in Fig. 6 is generated during this step and the subsequent one. The classification model processes the weakly augmented data, and values exceeding a predetermined threshold are assigned pseudo-labels.
4. **Loss Calculation:** Data that were strongly augmented are processed through the classification model, and the loss, L_{us}, is calculated. This loss is computed using data identified in the previous step as pseudo-labels. The total loss, L, is computed as the sum of L_s and L_{us} as seen in Fig. 7.
5. **Iteration:** Steps 2-4 are iteratively repeated until the model converges.

Additionally, we updated the working schema as represented in Fig. 7. Originally, the FixMatch method was developed for image data. However, in our study, we adapted it for tabulated data by employing data augmentation techniques based on a Gaussian distribution, which could be another novelty of the study.

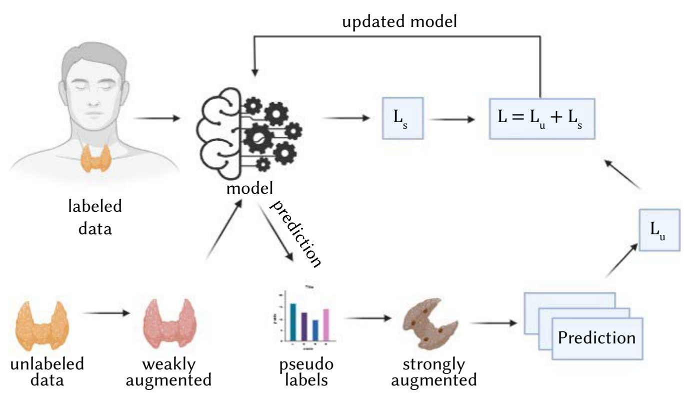

Fig. 7. Adaptation of the FixMatch method for thyroid disease prediction, depicting the process from labeled data through model refinement to loss calculation, highlighting the semi-supervised learning cycle.

Co-Traininig method. The Co-Training method, first proposed by Blum [40], is a semi-supervised machine learning technique that involves splitting labeled data and generating the most effective pseudo-labels from unlabeled data through two different models as represented in Fig. 8. According to the Fig. 8, the data is first split into two distinct views, X1 and X2, each having the same outputs. Each set of view is trained using a separate classifier (Classifier 1 (M1) and Classifier 2 (M2)). After the training, these classifiers are applied to unlabeled data and pseudo-labels are predicted. The common pseudo-labels selected from M1 and M2 are identified and added to the labeled views X1 and X2. The model repeats this process either until a predetermined number of iterations is reached or until there is no more unlabeled data. In this study, we have set the number of iterations to 30, following Blum's approach [40]. In some cases, SSL is used to augment unlabeled data, while in others, it enhances the performance of labeled data. In this study, it is preferred to increase the performance of the algorithms. Performance measures of the final iterative results are listed. This study explores the following structure.

TABLE II. Performance Metrics of Machine Learning Algorithms. Co-Training M1: Co-Training Classifier1, Co-Training M2: Co-Training Classifier2, NB: Naive Bayes, LR: Logistic Regression; Sens: Sensitivity; Spec: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; F1: F1-measure; Acc: Accuracy. Best Performance Measures Are Showed in Bold

Metric	FixMatch	Self-training	NB	LR	Co-Training M1	Co-Training M2
Sens	**0.9494**	0.2024	0.0642	0.6794	0.1961	0.2487
Spec	0.8551	0.9543	**0.9865**	0.7978	0.9605	0.9605
PPV	0.8824	0.9883	0.9891	0.7706	0.9886	**0.9910**
NPV	**0.9365**	0.0619	0.0549	0.7133	0.0639	0.0680
F1	**0.9146**	0.2420	0.1201	0.7222	0.3273	0.3976
Acc	**0.9054**	0.3350	0.1126	0.7386	0.2374	0.2871

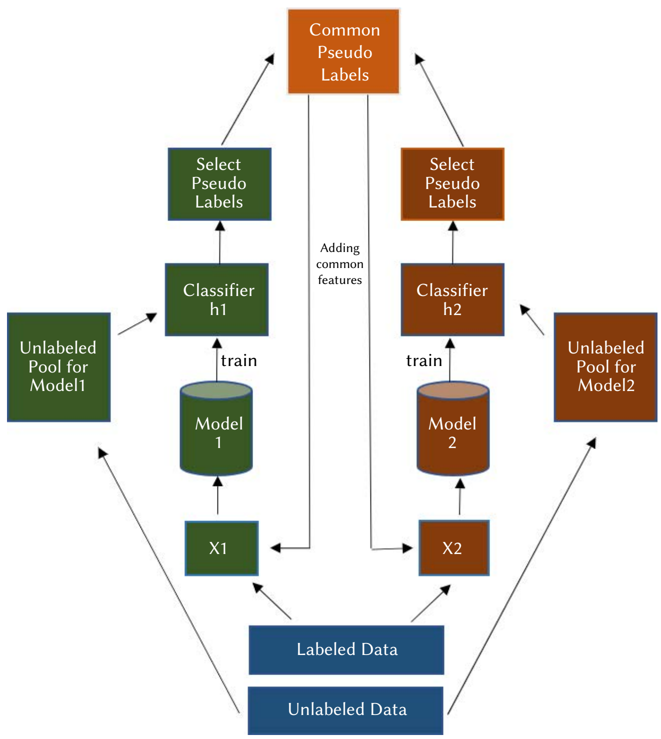

Fig. 8. Workflow of the Co-training method, illustrating the synergistic training of two models on labeled and unlabeled data with a focus on the iterative selection and refinement of pseudo-labels to improve classification performance.

1. **Split Labeled Data**
- Split the labeled data into two parts: a training set and a test set. Use the training set for training the model, while the test set is utilized for assessing its performance.
2. **Prepare Data for Co-training**
- Divide the features of the training set into two subsets views (X1 and X2). Each subset will be used to train a separate model.
3. **Co-training Process**
- Initial Training: Train two separate models, each on a different view of features from the labeled training data.
- Iteration Loop: Repeat the following steps for a set number of iterations (e.g., 30 times):
  - Predict on Unlabeled Data: Use both models to make predictions on the unlabeled data, utilizing their respective feature subsets.
  - Identify Agreements: Find instances where both models agree on the prediction. These are treated as pseudo-labeled data.
  - Update Training Sets: Add the pseudo-labeled instances to the respective training sets of each model, effectively increasing the labeled dataset.
  - Retrain Models: Re-train both models on their updated training sets, which now include the newly pseudo-labeled data.
  - Evaluate Models: At each iteration, evaluate the performance of both models on the separate test set and track the accuracy.
4. **Model Evaluation**
- After the final iteration, assess the performance of the models using the test set. Calculate sensitivity, specificity, positive predictive value, negative predictive value, F1 and accuracy.

E. Performance Evaluation

The core of our study involved comparing the performance of the SL and SSL models, as depicted in Fig. 2. The SL method was applied to labeled data, while the SSL method incorporated both labeled and unlabeled data. Performance metrics were meticulously computed and are outlined in Table II, these include Sens: Sensitivity; Spec: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; F1: F1-measure; Acc: Accuracy.

F. Discussion

Thyroid disease is a significant condition that can lead to important consequences if left untreated. The thyroid gland, an essential part of the body's endocrine system, is in charge of producing the hormones that regulate metabolism. When thyroid disorders occur, they disrupt the normal functioning of the thyroid gland, resulting in imbalances in hormone production. These imbalances can have a profound impact on individuals' overall health and quality of life. Common symptoms include fatigue, weight changes, mood swings, and cognitive issues [41]. It is crucial to promptly address thyroid disorders as untreated cases can lead to serious complications. Cardiovascular problems, fertility issues, and mental health concerns are among the potential outcomes of untreated or poorly managed thyroid disease. Thyroid disorders can be diagnosed and understood through a combination of laboratory results and imaging techniques. The term "euthyroid" refers to the enlargement of the thyroid gland despite the values obtained from thyroid function tests being within normal limits. It highlights the importance of considering additional factors beyond laboratory results when evaluating thyroid health. Thyroid disorders can also manifest in individuals who have normal results from laboratory tests, emphasizing the need to take into account factors beyond laboratory results when assessing thyroid health.

In this study, we propose the use of SSL methods in conjunction with IoE for the diagnosis of thyroid diseases. Initially, labels of individuals classified as normal were discarded. These individuals were then included in the analysis as unlabeled data in the SSL methods. In the self-training process, pseudo-labels were assigned to this unlabeled data, enabling the identification of individuals with hypo- or hyperthyroidism within the group initially labeled as normal. In the Co-training method, the dataset was split into two different views, and each dataset was trained separately. Common pseudo-labels were created, obtained from each view, and the model's performance was increased. In the FixMatch method, unlabeled data was utilized to enhance the model's performance through data augmentation, specifically by applying Gaussian noise in this study. This approach was aimed at improving the prediction accuracy for hypo/hyperthyroidism using the SSL method.

As shown in Table II and Fig. 9, the SSL method, particularly the FixMatch approach, yielded the highest values for Accuracy, Sensitivity, NPV, and F1 score. Additionally, Naive Bayes provided the highest values for Specificity, and the highest PPV result was obtained from the Co-training method. Accordingly, the best overall results were achieved using the SSL method, especially the FixMatch method.

V. Conclusion

A. Research Contribution

In this study, thyroid disease prediction was performed using SL and SSL methods based on data generated from individuals who do not have thyroid disease. The model performances were compared with the outputs of the SL method. According to the results obtained from SSL, especially in the FixMatch method, high values were achieved for accuracy, sensitivity, NPV and F1 score. However, in the SL methods, Naive Bayes yielded high values for Specificity and the highest PPV was ontained from the Co-training algorithms. In summary, it can be said that the FixMatch method demonstrated better performance in hypo/hyperthyroid prediction by utilizing unlabeled data from normal individuals. For this reason, based on the machine learning analysis the device could generate real-time alerts and notifications to the user, indicating any abnormalities or potential thyroid-related problems and those with euthyroid can be detected, so that the patients labeled as normal can be double checked to see if they have thyroid. In addition, hypo/hyper thyroid diseases of the people can be predicted in the same way, and people could be warned and take precautions beforehand.

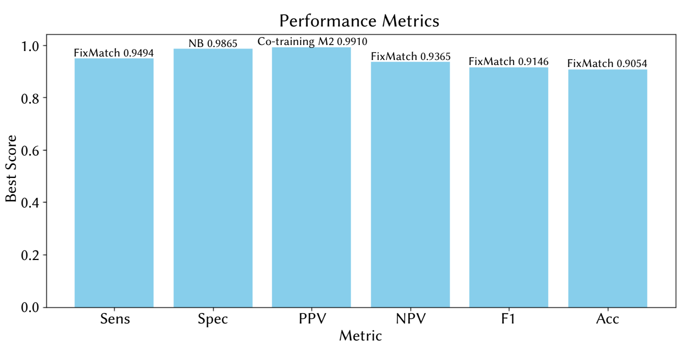

Fig. 9. Best performance measures obtained from the thyroid dataset, displaying the highest scores for each metric: Sensitivity (Sens), Specificity (Spec), Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1 Score (F1), and Accuracy (Acc). The Naive Bayes (NB) algorithm achieved a Spec score of 0.9865, while the Co-Training M2 algorithm showed superior performance in PPV with a score of 0.9910. Fix Match scores were notable for F1 (0.9146), and NPV (0.9365), with an overall Accuracy (Acc) of 0.9054.

B. Future Works and Research Limitations

Although numerous research efforts have investigated the use of ultrasound imaging to detect thyroid nodules, the datasets utilized in these studies are predominantly inaccessible to the public. Accumulating a substantial dataset is hindered by several factors, including time constraints, the specialized nature of medical practices, patient involvement, and the costs associated with ultrasound equipment acquisition [42]. This study aims not only to develop a model that can differentiate between thyroid subtypes but also to leverage labels of cases that are marked as normal yet may have a high likelihood of thyroid disease to enhance model performance. Additionally, we considered integrating this model with an IoT system, initially using publicly available datasets from the UCI repository. Our future goal is to refine our model using sensor data suitable for IoT applications.

Additionally, in our forthcoming research endeavors, our primary emphasis will be on the development of the proposed IoE methodology for thyroid monitoring using wearable devices. Our aim here will be to monitor and detect various thyroid-related diseases, including thyroid cancer. As an essential initial step, we will collect extensive thyroid-related data through these wearable devices, including image datasets. Subsequently, our objective is to delve into the realm of advanced deep learning algorithms.

The choice to explore deep learning algorithms is driven by the complex and multifaceted nature of thyroid data. Deep learning is highly effective at identifying complex relationships and drawing meaningful conclusions from extensive and complicated data sets. With thyroid health being influenced by numerous factors and characterized by subtle variations, deep learning techniques, such as convolutional neural networks (CNNs) [43], [44], long-short term memory (LSTM) [45] and recurrent neural networks (RNNs) [46], hold the potential to uncover hidden relationships within the data.

Moreover, we aim to undertake a comparative analysis between SSL methods, which we have already employed, and deep learning algorithms. This comparative study will allow us to evaluate the efficacy of deep learning in extracting valuable information from our thyroid dataset. By doing so, we hope to gain a comprehensive understanding of which approaches are most suitable for thyroid monitoring, ultimately advancing the field of thyroid healthcare.

Appendix

The SSL and SL code used in this study are available at https://github.com/melihagraz/SSL_AND_SL. Additionally, dataset is reachable at https://github.com/melihagraz/SSL_AND_SL/tree/2af0fecb9adcdd53000a3fac8ef97cc59471b242/data.

References

[1] Q. Yu, J. Ren, Y. Fu, Y. Li, W. Zhang, "Cybertwin: An origin of next generation network architecture," *IEEE Wireless Communications*, vol. 26, no. 6, pp. 111–117, 2019.

[2] S. Abdelwahab, B. Hamdaoui, M. Guizani, A. Rayes, "Enabling smart cloud services through remote sensing: An internet of everything enabler," *IEEE Internet of Things Journal*, vol. 1, no. 3, pp. 276–288, 2014.

[3] M. Deore, U. Kulkarni, "MDFRCNN: Malware detection using faster region proposals convolution neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 146–162, 2022.

[4] R. Quinlan, "Thyroid Disease." UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5D010.

[5] A. Rghioui, A. Naja, J. L. Mauri, A. Oumnad, "An iot based diabetic patient monitoring system using machine learning and node mcu," in *Journal of Physics: Conference Series*, vol. 1743, 2021, p. 012035, IOP Publishing.

[6] N. E. Maghawry, S. Ghoniemy, "A proposed internet of everything framework for disease prediction," *International Journal of Online & Biomedical Engineering*, vol. 15, no. 4, 2019.

[7] G. B. Raja, C. Chakraborty, "Internet of things based effective wearable healthcare monitoring system for remote areas," in *Implementation of Smart Healthcare Systems using AI, IoT, and Blockchain*, Elsevier, 2023, pp. 193–218.

[8] C. Sivaparthipan, B. A. Muthu, G. Manogaran, B. Maram, R. Sundarasekar, S. Krishnamoorthy, C.-H. Hsu, K. Chandran, "Innovative and efficient method of robotics for helping the parkinson's disease patient using iot in big data analytics," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 12, p. e3838, 2020.

[9] A. Singh, A. Kumar, S. Namasudra, "Dnacds: Cloud ioe big data security and accessing scheme based on dna cryptography," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 18, no. 1, pp. 181–801, 2023.

[10] J. Ahamed, A. Manan Koli, K. Ahmad, A. Jamal, B. Gupta, *et al.*, "CDPS-IoT: cardiovascular disease prediction system based on iot using machine learning," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 78–86, 2022.

[11] A. Mostbeck, G. Galvan, P. Bauer, O. Eber, K. Atefie, K. Dam, H. Feichtinger, H. Fritzsche, H. Haydl, H. Köhn, *et al.*, "The incidence of hyperthyroidism in austria from 1987 to 1995 before and after an increase in salt iodization in 1990," *European Journal of Nuclear Medicine*, vol. 25, pp. 367–374, 1998.

[12] P. A. Farling, "Thyroid disease," *British Journal of Anaesthesia*, vol. 85, no. 1, pp. 15–28, 2000.

[13] J. Muñoz-Ortiz, M. C. Sierra-Cote, E. Zapata-Bravo, L. Valenzuela-Vallejo, M. A. Marin-Noriega, P. Uribe- Reina, J. P. Terreros-Dorado, M. Gómez-Suarez, K. Arteaga-Rivera, A. De-La-Torre, "Prevalence of hyperthyroidism, hypothyroidism, and euthyroidism in thyroid eye disease: a systematic review of the literature," *Systematic Reviews*, vol. 9, no. 1, pp. 1–12, 2020.

[14] L. Ozyilmaz, T. Yildirim, "Diagnosis of thyroid disease using artificial neural network methods," in *Proceedings of the 9th International Conference on Neural Information Processing, 2002, ICONIP'02*, vol. 4, 2002, pp. 2033–2036, IEEE.

[15] I. Kravets, "Hyperthyroidism: diagnosis and treatment," *American family physician*, vol. 93, no. 5, pp. 363–370, 2016.

[16] J. Kvetny, "The significance of clinical euthyroidism on reference range for thyroid hormones," *European Journal of Internal Medicine*, vol. 14, no. 5, pp. 315–320, 2003.

[17] K. Hedayat, J. Lapraz, "A new approach to biological modeling: Introduction to the biology of functions," *Theory Endobiogeny*, vol. 15, pp. 215–254, 2019.

[18] E. G. Keramidas, D. K. Iakovidis, D. Maroulis, S. Karkanis, "Efficient and effective ultrasound image analysis scheme for thyroid nodule detection," in *Image Analysis and Recognition: 4th International Conference, ICIAR 2007, Montreal, Canada, August 22-24, 2007. Proceedings 4*, 2007, pp. 1052–1060, Springer.

[19] S. Razia, M. Rama Narasingarao, "A neuro computing frame work for thyroid disease diagnosis using machine learning techniques," *Journal of Theoretical and Applied Information Technology*, vol. 95, pp. 1996–2005, 05 2017.

[20] A. Andueza, M. Arco-Osuna, B. Fornés, R. Gonzalez Crespo, J. Martín Álvarez, "Using the statistical machine learning models arima and sarima to measure the impact of covid-19 on official provincial sales of cigarettes in spain," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 73–87, 2023.

[21] M. Qasim Gandapur, E. Verdú, "ConvGRU-CNN: Spatiotemporal deep learning for real-world anomaly detection in video surveillance system," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 88–95, 2023.

[22] M. Martín Merino, A. J. López Rivero, V. Alonso, M. Vallejo, A. Ferreras, "A clustering algorithm based on an ensemble of dissimilarities: An application in the bioinformatics domain," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 6, pp. 6–13, 2022.

[23] G. Turk, M. Ozdemir, R. Zeydan, Y. Turk, Z. Bilgin, E. Zeydan, "On the identification of thyroid nodules using semi-supervised deep learning," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 37, no. 3, p. e3433, 2021.

[24] K. Zhang, Z. Li, C. Cai, J. Liu, D. Xu, C. Fang, P. Huang, Y. Wang, M. Yang, S. Chang, "Semi- supervised graph convolutional networks for the domain adaptive recognition of thyroid nodules in cross-device ultrasound images," *Medical Physics*, vol. 50, pp. 7806–7821, 2023.

[25] W. Yang, J. Zhao, Y. Qiang, X. Yang, Y. Dong, Q. Du, G. Shi, M. B. Zia, "Dscgans: Integrate domain knowledge in training dual-path semi-supervised conditional generative adversarial networks and s3vm for ultrasonography thyroid nodules classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, 2019, pp. 558–566, Springer.

[26] S. H. Requena, J. M. G. Nieto, A. Popov, I. N. Delgado, "Human activity recognition from sensorised patient's data in healthcare: A streaming deep learning-based approach," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 23–37, 2023.

[27] S. Razia, M. N. Rao, "Machine learning techniques for thyroid disease diagnosis-a review," *Indian Journal of Science and Technology*, vol. 9, no. 28, pp. 1–9, 2016.

[28] G. Chaubey, D. Bisen, S. Arjaria, V. Yadav, "Thyroid disease prediction using machine learning approaches," *National Academy Science Letters*, vol. 44, no. 3, pp. 233–238, 2021.

[29] Z. J. Peya, M. K. N. Chumki, K. M. Zaman, "Predictive analysis for thyroid diseases diagnosis using machine learning," in *2021 International conference on science & contemporary technologies (ICSCT)*, 2021, pp. 1–6, IEEE.

[30] L. Royakkers, J. Timmer, L. Kool, R. Van Est, "Societal and ethical issues of digitization," *Ethics and Information Technology*, vol. 20, pp. 127–142, 2018.

[31] T. Jacquemard, C. P. Doherty, M. B. Fitzsimons, "The anatomy of electronic patient record ethics: a framework to guide design, development, implementation, and use," *BMC Medical Ethics*, vol. 22, no. 1, pp. 1–14, 2021.

[32] K. M. Besher, Z. Subah, M. Z. Ali, "Iot sensor initiated healthcare data security," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11977–11982, 2020.

[33] M. E. Shipe, S. A. Deppen, F. Farjah, E. L. Grogan, "Developing prediction models for clinical use using logistic regression: an overview," *Journal of thoracic disease*, vol. 11, no. Suppl 4, p. S574, 2019.

[34] P. Schober, T. R. Vetter, "Logistic regression in medical research," *Anesthesia & Analgesia*, vol. 132, no. 2, pp. 365–366, 2021.

[35] S. Joshi, B. Pandey, N. Joshi, "Comparative analysis of naive bayes and j48 classification algorithms," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 12, pp. 813– 817, 2015.

[36] R. Rajeswari, K. Juliet, D. Aradhana, "Text classification for student data set using naive bayes classifier and knn classifier," *International Journal of Emerging Trends & Technology in Computer Science*, vol. 43, no. 1, pp. 8– 12, 2017.

[37] B. Bustami, "Penerapan algoritma naive bayes untuk mengklasifikasi data nasabah asuransi," *TECHSI- Jurnal Teknik Informatika*, vol. 5, no. 2, 2013.

[38] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, I. Cholissodin, "Optimasi naïve bayes classifier dengan menggunakan particle swarm optimization pada data iris," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 3, p. 180, 2017.

[39] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[40] A. Blum, T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[41] G. P. Redmond, "Thyroid dysfunction and women's reproductive health," *Thyroid*, vol. 14, no. 3, Supplement 1, pp. 5–15, 2004.

[42] R. Sharma, G. K. Mahanti, C. Chakraborty, G. Panda, A. Rath, "An iot and deep learning-based smart healthcare framework for thyroid cancer detection," *ACM Transactions on Internet Technology*, 2023.

[43] X. Zhang, V. C. Lee, J. Rong, J. C. Lee, F. Liu, "Deep convolutional neural networks in thyroid disease detection: a multi-classification comparison by ultrasonography and computed tomography," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106823, 2022.

[44] X. Zhang, V. C. Lee, J. Rong, J. C. Lee, J. Song, F. Liu, "A multi-channel deep convolutional neural network for multi-classifying thyroid diseases," *Computers in Biology and Medicine*, vol. 148, p. 105961, 2022.

[45] H. Lu, M. Wang, W. Zhao, T. Su, J. Yang, "Hyperthyroidism progress prediction with enhanced lstm," in *Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II 4*, 2020, pp. 209–217, Springer.

[46] A. Santillan, R. C. Tomas, R. Bangaoil, R. Lopez, M. H. Gomez, A. Fellizar, A. Lim, L. Abanilla, M. C. Ramos, L. Guevarra, *et al.*, "Discrimination of malignant from benign thyroid lesions through neural networks using ftir signals obtained from tissues," *Analytical and Bioanalytical Chemistry*, vol. 413, pp. 2163–2180, 2021.

Melih Agraz

Melih Agraz graduated from Dokuz Eylül University in Türkiye, majoring in Mathematics in the Department of Education. Following that, he obtained his Master's degree from Dokuz Eylül University in the Department of Statistics. Later, he earned his Ph.D. in Statistics from the Middle East Technical University in Türkiye. Subsequently, he embarked on a Fulbright postdoctoral research position at UC Berkeley in the USA. Afterward, he received an offer from Brown University and Beth Israel Hospital. Currently, he is working as an Assistant Professor in the Department of Statistics at Giresun University in Türkiye and has an affiliation with the Brown University Division of Applied Mathematics.