SELECTIVE FEATURE BASED THYROID DISEASE CLASSIFICATION USING DEEP LEARNING

ABHISHEK HR[1]
Dept. of CSE
MANGALAM COLLEGE OF ENGINEERING ,
Kerala, India

ANKITHA MURALI[2]
Dept. of CSE
MANGALAM COLLEGE OF ENGINEERING ,
Kerala, India

BENDEK JACOB MATHEWS A[3]
Dept. of CSE
MANGALAM COLLEGE OF ENGINEERING ,
Kerala, India

ANOOP T SASIDHARAN [4]
Dept. of CSE
MANGALAM COLLEGE OF ENGINEERING ,
Kerala, India

***Abstract*—** There are many thyroid diseases affecting people all over the world. Many diseases affect the thyroid gland, like hypothyroidism, hyperthyroidism, and thyroid cancer. Thyroid inefficiency can cause severe symptoms in patients. Effective classification and machine learning plays a significant role in the timely detection of thyroid diseases. This timely classification will indeed affect the timely treatment of the patients. In this study, a novel end-to-end knowledge-driven classification framework is presented. The main goal of this study is to classify thyroid disease into three categories: hyperthyroidism, hypothyroidism and normal and obtained classification results are used for the diagnosis purposes. Existing approaches often target binary classification, the used datasets are small in size and results validation is not done properly. Also, existing approaches focus less on feature engineering and model optimization. To overcome the limitations present in existing models, this work mainly concentrates on feature engineering and model optimization for deep learning. For getting better accuracy extra tree classifier based selected features are used for feature selection along with random forest classifier. As demonstrated by the results, the proposed system achieves relevant performances in terms of qualitative metrics for the thyroid nodule classification task, thus resulting in a great asset when formulating a diagnosis. K-fold validation technique along with F1 score corroborate the superior performance of the proposed

I. INTRODUCTION

Thyroid disease is becoming more common in recent years. One of the most important functions of the thyroid gland is to regulate metabolism. Thyroid gland irregularities can result in a variety of abnormalities, the most common of which are hyperthyroidism and hypothyroidism. Every year, a large number of people are diagnosed with thyroid diseases like hypothyroidism and hyperthyroidism. Thyroid hormones such as levothyroxine (T4) and triiodothyronine (T3) are produced by the thyroid gland, and a lack of thyroid hormones can result in hypothyroidism or hyperthyroidism. Many approaches to detecting thyroid disease diagnosis are proposed in the literature. A proactive thyroid disease prediction is required to properly treat the patient at the appropriate time, saving human lives and medical costs. Machine learning and deep learning techniques are being used to predict thyroid diagnosis in the early stages and classify thyroid disease types such as hypothyroidism, hyperthyroidism, and others as a result of technological advancements in data processing and computation. The healthcare domain benefited from leveraging technology in many healthcare areas for human well-being due to advancements in technologies such as data mining, big data, image and video processing, and parallel computing. Data mining-based health care applications may include disease detection and diagnosis, virus outbreak prediction, drug discovery and testing, health care data management, and patient personalized medicine recommendations, among other things.

Health care professionals strive to identify diseases in their early stages so that proper treatment can be provided to patients and the disease can be cured in a short period of time and with minimal expenditure. Thyroid disease is one of the diseases that affects a large number of people worldwide. Thyroid disease affects 20 million Americans, according to the world's leading professional association (American thyroid association). A thyroid condition affects 12% of the US population at some point in their lives. These figures show that thyroid disease should not be taken lightly. The use of advanced technologies to improve health care practices for detecting and preventing thyroid diseases is highly desired. Existing research works primarily focus on binary classification problems in which subjects are classified as thyroid patients or healthy subjects, with only a few multiclass-based detection works. Even for those, there are three categories to consider: normal, hypothyroidism, and hyperthyroidism. For the most part, the emphasis is on the optimization of machine learning and deep learning models, while the feature selection component is under-studied or completely ignored in the context of a thyroid disease problem. Despite the high accuracy reporting approaches, such approaches are tested on samples with a sample size of less than 1000, and the results are not validated. The classification in terms of patient status, such as treatment condition, health condition, and general health issues, is desired in order to effectively predict and treat the patient's thyroid condition. Furthermore, no performance comparison of machine learning and deep learning models is performed. This research aims to address these issues and makes the following contributions:

• A novel machine learning-based thyroid disease prediction approach focusing on the multiclass problem is proposed. Unlike previous studies that focused on the binary or three-class problem, this study takes a five-class disease prediction problem into account.

This study investigates four feature engineering approaches to assess their efficacy for the problem at hand. Forward feature selection (FFS), backward feature elimination (BFE), bidirectional feature elimination (BiDFE), and machine

learning-based feature selection with an extra tree classifier are all included.

• For experiments, machine learning models such as random forest (RF), logistic regression, and support vector machine (SVM) are chosen based on their reported performance for disease prediction. In addition, three deep learning models are used: convolutional neural network, long short-term memory (LSTM) network, and CNN LSTM. In addition to accuracy, precision, recall, and F1 score, performance is measured using the confusion matrix, 10-fold cross validation, and standard deviation. The rest of this article is structured as follows. Section 2 discusses cutting-edge research to detect and classify thyroid diseases. Section 3 discusses the proposed methodology for addressing the problem of thyroid disease prediction. This section also includes the methods for selecting features, the machine learning techniques used in the article, and the dataset description considered for this study. Section 4 describes our study's experimental results and comparisons with prior art studies. Section 5 brings the article to a close with our contributions.

II. RELATED WORK

2.1 Thyroid Disease Classification Using Machine Learning Algorithms Key Aspects

Machine learning algorithms are used in the rapid and early diagnosis of thyroid diseases and other diseases, as they now hold a significant position in the health field and assist us in diagnosing and classifying diseases. As a result, they have been able to collect a large amount of data on thyroid diseases,whichweareusinginourstudyondiseaseclassification.ThedatathatIused in this study is a set of data taken from external hospitals and laboratories specialized in analyzing and diagnosing diseases, and the sample taken from the data is the data of people and the type of data taken related to thyroid disease, where data were taken on 1250 people divided into males and females. The data were collected over a period of one to four months, with the primary goal of classifying thyroid disease using machine learning algorithms. Gender, age, T3 (triiodothyronine), T4 (thyroid hormone), TSH(thyroid stimulating hormone), and a variety of other characteristics are included in this data. As the data obtained consists of 17 variables or attributes, all of which were considered in our study (id, age,gender, query thyroxine, on antithyroid medication, sick, pregnant, thyroid surgery, query hypothyroid,queryhyperthyroid,TSH,TSH,T3M,T3,T3,T4,Category).

Here two steps are going to be considered.

1. DataPre-processing

2. Deep Learning-based algorithms

Merits:EvaluationismadeondifferentmachinelearningalgorithmsandDecisiontreecouldobtainedbestaccuracy.

Demerits:Still There Exist Limitations On Attributes.

2.2 Performance Analysis of Machine Learning Algorithms for thyroid Disease Key Concepts Thyroid disease is caused by an abnormal growth of thyroid gland.Thyroid Disorder Occurs When this gland produces an abnormal amount to hormones; The two main types of thyroid disorder are hypothyroidism(inactive thyroid gland)and hyperthyroidism (hyperactive thyroid gland). To detect and diagnose thyroid disease, this study proposes the use of efficient classifiers based on machine learning algorithms in terms of accuracy and other performance evaluation metrics. This study examines various classifiers such as K-nearest neighbor (KNN), NaveBayes, support vector machine, decision tree, and logistic regression implemented with or without feature selection techniques. Thyroid information was obtained from DHQ Teaching Hospital in DeraGhaziKhan,Pakistan. Thyroid data set was distinct from other existing studies because it included three additional features: pulse rate, BMI, and blood pressure. The experiment consisted of three iterations;the first iteration did not use feature selection, while the second and third used an L1-, L2-based feature selection technique. The experiment was evaluated and analyzed, and many factors such as accuracy,precision, and receiver operating curve with area under curve were considered. The results showed that classifiers that used L1-based feature selection achieved higher overall accuracy (Naive Bayes 100%,logistic regression 100%, and KNN 97.84%) than classifiers that did not use feature selection or L2-based feature selection technique.

Merits: Processing missing values and cleaning unnecessary data can potentially improve accuracy of overall result

Demerits: Feature selection is not applied on attributes or in every iteration.

2.3 Thyroid Disease Treatment prediction with machine learning approach Key Concepts

The thyroid is an endocrine gland located in the anterior region of the neck, and its primary function isto produce thyroid hormones, which are essential to our overall health. Its possible dysfunction can result in either insufficient or excessive thyroid hormone production. As a result of one or more dwellings forming inside the thyroid, it can become inflamed or swollen. Some of these nodules may harbor malignant tumors. One Of the most commonly used treatments is sodium levothyroxine(LT4), a synthetic thyroid hormone used to treat thyroid disorders and diseases. Predictions about treatment can be useful for assisting endocrinologists in their work and improving the quality of life for patients. There are numerous studies in the literature that focus on the prediction of thyroid diseases based on the trend of people's hormonal parameters. This study, on the other hand, aims to forecast theLT4 treatment trend for hypothyroid patients. To that end, a dedicated dataset containing medical information about patients being treated at Naples' "AOU Federico II" hospital was created. Because The Clinical History of each patient is available over time, it was possible to predict the course of each patient's treatment in order to understand whether it should be increased or decreased based on the trend of the hormonal parameters and other attributes considered. We used various machine learning algorithms to carry out this research.We specifically compared the results of different classifiers.The Performance of the various algorithms is good, especially in the case of the Extra-Tree Classifier, which has an accuracy of 84%.

Merits : The result is compared with 10 different classifiers and used extra tree classifiers to improve accuracy

Demerits: The sizeof the dataset very less

2.4 Intelligent Diagnosis Of Thyroid Ultrasound Imaging Using An Ensemble Of DeepLearning Methods KeyConcepts

Thyroid disorders are currently prevalent in the global population, necessitating the development of alternative methods for improving the diagnosis process. Materials andProcedures:To accomplish this, we created an ensemble method that combines two deep learning models,onebasedon convolutional neural networks and the other on transfer learning. We created an efficient end-to-end trained model with five convolutional layers for the first model, called 5-CNN, and repurposed,optimized, and trained the pre-trained VGG-19 architecture for the second model. We trained and validated our models using an ultrasound image dataset that included four types of thyroidal images: autoimmune,nodular,micronodular,and normal.
Results: The ensemble CNN-VGG method outperformed the 5-CNN and VGG-19 models, yielding excellent results: 97.35% overall test accuracy with an overall specificity of 98.43%, sensitivity of 95.75%, positive and negative predictive value of 95.41% and 98.05%, respectively. Each receiver operating characteristic curve's micro average area was 0.96. Two physicians, an endocrinologist and a pediatrician, also validated the findings. Conclusions: We proposed a deep learning study to assist physicians in diagnosis of ultrasound thyroid images.
Merits : Excellent results were obtained by the ensemble CNN-VGG method, which outperformed the5-CNNandVGG-19 models

III. METHODOLOGY

A. Proposed System III

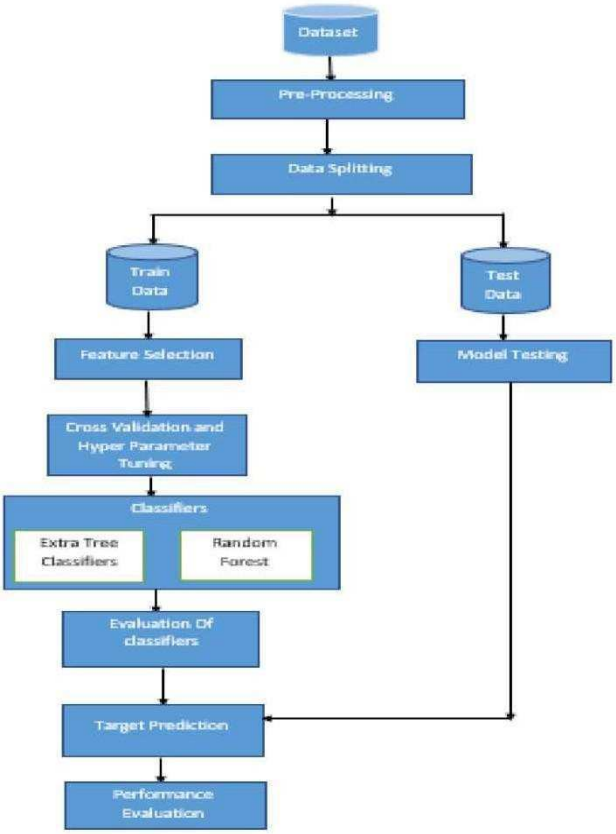
The proposed system employs a novel end-to-end knowledge-driven classification framework. The primary goal of this research is to divide thyroid disease into three categories. Rather than using common approaches such as binary classification techniques, this paper focuses on feature engineering and model optimization techniques. Extra tree classifier techniques improve accuracy, and random forest classification is used for classification.

Fig 3.1 : Flow Of the proposed methodology
Work flow of selective feature based thyroid prediction between various entities are shown in figure.
3.1
Figure 3.2 shows the architecture and flow of the proposed approach for thyroid disease prediction. The data set contains several thyroid-related disease records as well as numerous target classes. Because the samples for target classes are insufficient to train models, we chose only those target classes with samples greater than 250, yielding five target classes. We performed data balancing after selecting the target classes for experiments. Because the normal class samples were 6771 in total, which was more than the other target class samples, we randomly selected only 400 samples for the normal class to balance the dataset. It is followed by the feature selection process, which employs a variety of feature selection techniques. Several machine learning and deep learning models are used in experiments with an 80-20 train-test split. The performance of deep learning models on the used dataset is also evaluated using each feature selection technique. We compared deep learning models to machine learning models such as LSTM, CNN, and CNN-LSTM. As shown in Table 15, these models are used with cutting-edge architectures. Deep Learning Models Can Have A Variety of Layers, dropout layer positions, neurons, and activation functions. Each model is trained usingthe‘categorical_crossentropy’lossfunction,whilethe‘Adam ’optimizerisused.Themodelsaretrainedwitha 16bathand100epochsareusedfortraining. In general, the proposed approach presents a promising approach to thyroid disease classification using a novel end-to-end knowledge-driven classification framework. This approach addresses the limitations of existing models by focusing on feature engineering and model optimization, resulting in better accuracy and performance. The use of extra tree classifiers and random forest classifiers for feature selection and classification is a novel and effective approach. The incorporation of k-fold validation technique and F1 score provides a robust evaluation of the model's performance. Overall, the above study has significant implications for the timely and accurate diagnosis of thyroid diseases, which can greatly improve patient outcomes. Further research and validation of your approach can lead to its implementation in clinical settings, providing healthcare professionals with a valuable tool for diagnosing and treating thyroid diseases.

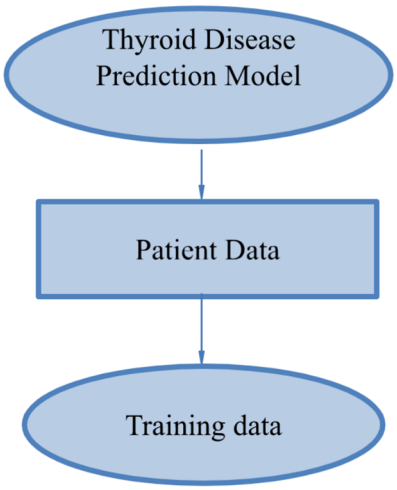
B. Proposed Architecture

The proposed system architecture is a centralized system that requires a high-performance computing environment with sufficient memory and storage capacity to process the large volume of patient records. The system primarily consists of functional components such as data collection, preprocessing, feature extraction, model training, and prediction, which are tightly integrated to perform automated disease predictions. The architecture utilizes open-source machine learning libraries and frameworks such as TensorFlow and Scikit-Learn to support the implementation of various machine learning algorithms. The system is designed to be highly scalable and can accommodate future updates and modifications to the dataset or model parameters.

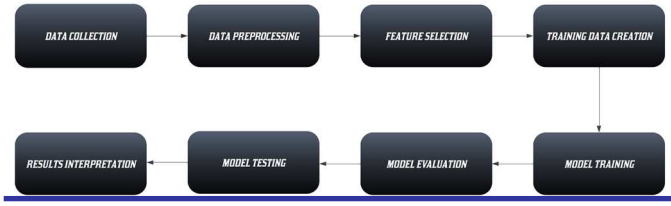
Finally, the system's performance is evaluated by comparing its predictions against the actual target classes of the patient records in the testing dataset. The performance evaluation metrics include confusion matrix, ROC curve, and AUC score, among others. Overall, the system's architecture is designed to provide accurate, automated predictions for different thyroid diseases while ensuring high performance, scalability, and flexibility.



Level 1 DFD



Level 2 DFD



IV. RESULT

In this section, we provide a comprehensive overview of the experiments conducted on thyroid disease prediction utilizing machine learning techniques. We elaborate on the findings obtained through each feature selection approach using both machine learning and deep learning models. The dataset was partitioned into training and testing sets at a ratio of 80:20, with 80% of the data reserved for training the models and 20% for testing. Table 7.1 exhibits the distribution of the target in terms of each target class.

Number of samples for training and test subset.

Target Class	Training	Testing	Total
"_" (0)	325	75	400
F (1)	190	43	233
G (2)	280	79	359
I (3)	271	75	346
K (4)	353	83	436

Table 7.1 Distribution of the target in terms of each target class

Following data splitting, we utilized various machine learning and deep learning models that were trained using the optimal hyperparameter settings. The models were trained using important features selected by feature selection techniques, and their performance was evaluated using 20% test data and 10-fold cross-validation techniques. The evaluation of models was conducted based on several metrics such as accuracy, precision, recall, F1 score, confusion matrix, and standard deviation (SD).

V. FUTURE SCOPE

Here are some possible suggestions for future scope of the above work:

1. Increasing the size and diversity of the dataset: The current study used a small-sized dataset with limited samples for each target class. Future work can explore larger and more diverse datasets to improve the generalizability of the proposed approach.
2. Incorporating more feature selection techniques: While the proposed approach used multiple feature selection

techniques, there are still many other techniques that can be explored, such as genetic algorithms and principal component analysis.

VI. CONCLUSION

Thyroid disease has become a pressing medical issue in recent years, and accurate automatic prediction models are needed to address this problem. However, existing studies have mostly focused on model optimization and feature engineering, while feature selection has received less attention. Additionally, the small size of the dataset used for model evaluation and lack of validation have limited the accuracy of previous approaches. This study addresses these limitations by proposing an approach that uses feature selection in combination with both machine learning and deep learning models. The extra tree classifier-based selected features provide the highest accuracy of 0.99 when used with the RF model, while other feature techniques result in poor performance due to feature reduction, which adversely affects both the deep learning and machine learning models, especially linear models. The lower computational complexity of machine learning models like RF makes them suitable candidates for thyroid disease prediction. The 10-fold cross-validation results validate these findings, and performance comparison with state-of-the-art approaches demonstrates the superiority of the proposed approach. Limitations of the study include feature reduction and the 5-class classification problem, which we intend to address by increasing the number of classes in future work.

VII. ACKNOWLEDGEMENT

The authors would like to thank Principal Vinodh P Vijayan, Neethu Mariya John, H.O.D, Faculty of Computer Science, for their appropriate guidance, valuable assistance and helpful comments during the proofreading process.

REFERENCES

[1] Chaubey, G.;Bisen,D.;Arjaria,S.;Yadav, V. Thyroiddiseasepredictionusingmachine learningapproaches.Natl.Acad. Sci.Lett.2021,44, 233-238. [CrossRef]2.

[2] Ioni;t“a,I.; Ioni;t“a,L.Predictionofthyroiddiseaseusin gdataminingtechniques.BRAINBroadRes .Artif.Intell.Neurosci.2016,7, 115-124.

[3] Webster,A.;Wyatt,S. Health,TechnologyandSociety;Spri nger:Berlin/Heidelberg,Germany,2 020.

[4] Hong, L.; Luo, M.; Wang, R.; Lu, P.; Lu, W.; Lu, L. Big data in health care: Applicationsandchallenges.DataInf. Manag.2018, 2,175-197. [CrossRef]

[5] Association, A.T. General Information/Press Room|American Thyroid Association.Availableonline:https://w ww.thyroid.org/media-main/press-room/(accessedon7April2022).