Machine learning on thyroid disease: a review

Kwang-Sig Lee1,*, Hyuntae Park2,∗

1AI Center, Korea University College of Medicine, 02841 Seoul, Republic of Korea
2Department of Obstetrics & Gynecology, Korea University College of Medicine, 02841 Seoul, Republic of Korea
*Correspondence: ecophy@hanmail.net (Kwang-Sig Lee); cyberpelvis@gmail.com (Hyuntae Park)

Abstract

This study reviews the recent progress of machine learning for the early diagnosis of thyroid disease. Based on the results of this review, different machine learning methods would be appropriate for different types of data for the early diagnosis of thyroid disease: (1) the random forest and gradient boosting in the case of numeric data; (2) the random forest in the case of genomic data; (3) the random forest and the ensemble in the case of radiomic data; and (4) the random forest in the case of ultrasound data. Their performance measures varied within 64.3–99.5 for accuracy, 66.8–90.1 for sensitivity, 61.8–85.5 for specificity, and 64.0–96.9 for the area under the receiver operating characteristic curve. According to the findings of this review, indeed, the following attributes would be important variables for the early diagnosis of thyroid disease: clinical stage, marital status, histological type, age, nerve injury symptom, economic income, surgery type [the quality of life 3 months after thyroid cancer surgery]; tumor diameter, symptoms, extrathyroidal extension [the local recurrence of differentiated thyroid carcinoma]; RNA features including ADD3-AS1 (downregulation), MIR100HG (downregulation), FAM95C (downregulation), MORC2-AS1 (downregulation), LINC00506 (downregulation), ST7-AS1 (downregulation), LOC339059 (downregulation), MIR181A2HG (upregulation), FAM181A-AS1 (downregulation), BLACAT1 (upregulation), hsa-miR-9-5p (downregulation), hsa-miR-146b-3p (upregulation), hsa-miR-199b-5p (downregulation), hsa-miR-4709-3p (upregulation), hsa-miR-34a-5p (downregulation), hsa-miR-214-3p (downregulation) [papillary thyroid carcinoma]; gut microbiota RNA features such as veillonella, paraprevotella, neisseria, rheinheimera [hypothyroidism]; and ultrasound features, i.e., wreath-shaped feature, microcalcification, strain ratio [the malignancy of thyroid nodules].

Keywords: thyroid; early diagnosis; machine learning; random forest; review

1. Introduction

The thyroid gland is an endocrine gland creating thyroid hormone. It is shaped like a butterfly and positioned in the front of the neck. Thyroid hormone involves the regulation of metabolism and various problems can occur in the gland. It can create either too little or too much hormone (hypothyroidism or hyperthyroidism). The former condition causes fatigue, weight gain and intolerance to cold temperature, whereas the latter leads to anxiety, weight loss and sensitivity to heat. Also, malignant cells can develop there (thyroid cancer) [1,2]. These disorders, thyroid disease, has been a leading cause of disease burden in the world [3–6]. The number of individuals with thyroid disease is estimated to be 200 million in the world [3], whereas the incidence and mortality of thyroid cancer registered rapid growths of 169% and 87% during 1990–2017, i.e., 95,030 and 22,070 to 255,490 and 41,240, respectively [4]. Hypothyroidism is reported to cause significant disease burden and direct, morbidity and mortality cost, as well [5,6]. It has various risk factors and many of them are still unknown. Its diagnosis and prognosis are considered to be quite challenging given that its symptoms are very similar with other diseases such as depression [1–3]. It is not surprising that there exists a high degree of variation among clinical experts in terms of its diagnosis and prognosis. In this context, more research is to be done on this important topic. Recently, on the other hand, the terms “deep learning”, “machine learning” and “artificial intelligence” have attracted great attention all over the globe. For instance, their Google trends recorded ten-fold expansions from 10 to 100 during 2013–2018. Artificial intelligence can be defined as “the capability of a machine to imitate intelligent human behavior” (the Merriam-Webster dictionary). The definition of machine learning can be a division of artificial intelligence to “extract knowledge from large amounts of data” [7].

Six common machine learning algorithms are the decision tree, the naïve Bayesian predictor, the random forest, the support vector machine, the artificial neural network, and the deep neural network (deep learning). A decision tree has three components: an intermediate node (a test on an independent variable), a branch (an outcome of the test) and a terminal node (a value of the dependent variable). A naïve Bayesian predictor makes an early diagnosis based on Bayes’ theorem, which states that the probability of the dependent variable given certain values of independent variables comes from the probabilities of the independent variables given a certain value of the dependent variable. A random forest is a collection of many decision trees with a majority vote on the dependent variable (“bootstrap ag-
regression”). Let us take a random forest with 1000 decision trees as an example. Here, the algorithm samples 1000 training sets with replacements, trains 1000 decision trees with the 1000 training sets, makes 1000 predictions with the 1000 decision trees, and takes a majority vote on the dependent variable. A support vector machine originates a line or space called a “hyperplane” (a collection of “support vectors”). The hyperplane divides data with the greatest distance between different sub-groups [7].

An artificial neural network is a network of “neurons”, i.e., information units combined through weights. Usually, the artificial neural network has one input layer, one, two or three intermediate layers and one output layer. Neurons in a previous layer connect with “weights” in the next layer and these weights represent the strengths of connections between neurons in a previous layer and their next-layer counterparts. This process starts from the input layer, continues through intermediate layers and ends in the output layer (feedforward operation). Then, learning happens: these weights are accommodated based on how much they contributed to the loss, a difference between the actual and predicted final outputs. This process starts from the output layer, continues through intermediate layers and ends in the input layer (backpropagation operation). The two operations are replicated until a certain expectation is met regarding the accurate diagnosis of the dependent variable. In other words, the performance of the artificial neural network improves as long as its learning continues. Finally, a deep neural network is an artificial neural network with a large number of intermediate layers, e.g., 5, 10 or even 1000. The deep neural network is called “deep learning” given that learning “deepens” through numerous intermediate layers [8].

Traditional research considers a limited scope of predictors for the early diagnosis of disease, whereas adopting logistic regression with an unrealistic assumption of ceteris paribus, i.e., “all the other variables staying constant”. In this context, emerging literature uses artificial intelligence for the early diagnosis of disease, e.g., arrhythmia [8], birth outcome [9–14], cancer [15–19], comorbidity [20–22], menopause [23] and temporomandibular disease [24,25]. It does not require unrealistic assumptions of “all the other variables staying constant” while managing to analyze which predictors are more important for the early diagnosis of the dependent variable. The purpose of this study is to review the recent progress of machine learning for the early diagnosis of thyroid disease.

2. Materials and methods

Twenty original studies were selected for review out of 33 original studies in PubMed with the search terms “thyroid” (title) and “random forest” (abstract). The inclusion criteria of this review were: (1) the intervention(s) of the decision tree, the naïve Bayesian predictor, the random forest, the support vector machine and/or the artificial neural network; (2) the outcome(s) of accuracy and/or the area under the receiver operating characteristic curve for the early diagnosis of thyroid disease; (3) the publication year of 2020 or later; and (4) the publication language of English. The following summary measures were adopted: machine learning methods, sample size, data type, performance measures and important attributes (predictors). Here, accuracy can be defined as the proportion of correct predictions over all observations, while the area under the receiver operating characteristic curve (AUC) can be defined as the area under the plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The exclusion criteria of this review was that thyroid disease is an independent variable (attribute) instead of the dependent variable.

3. Results

3.1 Summary of review

The summary of review is shown in Tables 1,2 (Ref. [26–45]). The tables have five summary measures, i.e., machine learning methods, sample size, data type, performance measures, important attributes, and whether the variable importance of the random forest is reported (VI-Yes 1). Based on the results of this review, different machine learning methods would be appropriate for different types of data for the early diagnosis of thyroid disease: (1) the random forest and gradient boosting in the case of numeric data; (2) the random forest in the case of genomic data; (3) the random forest and the ensemble in the case of radiomic data; and (4) the random forest in the case of ultrasound data. Their performance measures varied within 64.3–99.5 for accuracy, 66.8–90.1 for sensitivity, 61.8–85.5 for specificity, and 64.0–96.9 for the AUC (Table 1). According to the findings of this review, indeed, the following attributes would be important variables for the early diagnosis of thyroid disease: clinical stage, marital status, histological type, age, nerve injury symptom, economic income, surgery type [the quality of life 3 months after thyroid cancer surgery]; tumor diameter, symptoms, extrathyroidal extension [the local recurrence of differentiated thyroid carcinoma]; RNA features including ADD3-AS1 (downregulation), MIR100HG (downregulation), FAM95C (downregulation), MORC2-AS1 (downregulation), LINC00506 (downregulation), ST7-AS1 (downregulation), LOC339059 (downregulation), MIR181A2HG (upregulation), FAM181A-AS1 (downregulation), LBX2-AS1 (upregulation), BLACAT1 (upregulation), hsa-miR-9-5p (downregulation), hsa-miR-146b-3p (upregulation), hsa-miR-199b-5p (downregulation), hsa-miR-4709-3p (upregulation), hsa-miR-34a-5p (upregulation), hsa-miR-214-3p (downregulation) [papillary thyroid carcinoma]; gut microbiota RNA features such as veillonella, paraprevotella, neisseria, rheinheimera [hypothyroidism]; and ultrasound features, i.e., wreath-shaped feature, micro-calcification, strain ratio [the malignancy of thyroid nodules] (Table 2). However, machine learning is a...
data-driven method and more study is to be done for greater external validity.

3.2 Summary of selected studies

The summary of selected studies is presented in this section. The aim of a recent study \[27\] was to adopt machine learning and numeric data for predicting the quality of life three months after thyroid surgery. Data came from 286 participants and the attributes were European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Version 3 responses. The accuracy of the random forest for the validation set was 89.7. Based on random forest variable importance, clinical stage, marital status, histological type, age, nerve injury symptom, economic income and surgery type were the most important variables for predicting the quality of life three months after thyroid surgery. Likewise, the purpose of recent research \[30\] was to employ machine learning and numeric data for predicting the local recurrence of differentiated thyroid carcinoma. The accur-
<table>
<thead>
<tr>
<th>ID</th>
<th>Class [attributes]</th>
<th>Important attributes</th>
<th>VI-Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>16 PTC vs. 31 Papillary Micro Carcinoma [Raman Intensity]</td>
<td>Clinical Stage, Marital Status, Histological Type, Age, Nerve Injury Symptom, Economic Income, Surgery Type</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>Quality of Life 3 Months after Thyroid Cancer Surgery [European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Version 3]</td>
<td>Opioids</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Thyroid Complications after Receiving Programmed Cell Death 1/Programmed Cell Death Ligand 1 Inhibitors</td>
<td>Tumor Diameter, Symptoms, Extrathyroidal Extension</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Malignancy of Indeterminate Thyroid Nodules after Fine Needle Aspiration Biopsy [Diagnostic Pathology Features]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Local Recurrence of Differentiated Thyroid Carcinoma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Malignancy of Indeterminate Thyroid Nodules after Fine Needle Aspiration Cytology [Diagnostic Pathology Features]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Local Recurrence of PTC [Serum Thyroglobulin/Antithyroglobulin Features]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Malignancy of Thyroid Nodules [Image Features Collected Based on Scale-Invariant Feature Transformation and CNN]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>PTC [703 RNA Features from Cancer Genome Atlas Data]</td>
<td>FAM181A-AS1 (Downregulation), LBX2-AS1 (Upregulation), BLACAT1 (Upregulation); Group 2 Including hsa-miR-9-5p (Downregulation), hsa-miR-146b-3p (Upregulation), hsa-miR-199b-5p (Downregulation), hsa-miR-4709-3p (Upregulation), hsa-miR-34a-5p (Upregulation), hsa-miR-214-3p (Downregulation)</td>
<td>1</td>
</tr>
</tbody>
</table>
### Table 2. Continued.

<table>
<thead>
<tr>
<th>ID</th>
<th>Class [attributes]</th>
<th>Important attributes</th>
<th>VI-Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>[35]</td>
<td>Papillary Thyroid Cancer [Magnetic Resonance Imaging Features]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[37]</td>
<td></td>
<td>46 Follicular Thyroid Carcinoma vs. 46 Follicular Adenoma [70 DNA Methylation Haplotype Blocks]</td>
<td></td>
</tr>
<tr>
<td>[38]</td>
<td></td>
<td>Aggressive Extrathyroidal Extension PTC [Magnetic Resonance Imaging Features]</td>
<td></td>
</tr>
<tr>
<td>[39]</td>
<td></td>
<td>Thyroid Peroxidase Activity/Thyroid Hormone Receptor Modulation [Molecular Features from ToxCast Data]</td>
<td></td>
</tr>
<tr>
<td>[40]</td>
<td></td>
<td>Malignancy of Thyroid Nodules [Gray-Level Run-Length Matrix Run-Length Nonuniformity, Maximum Standard Unit Value]</td>
<td></td>
</tr>
<tr>
<td>[42]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[43]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[44]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[45]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PTC, papillary thyroid carcinoma.
racy range of logistic regression, the decision tree and the random forest was 84.7–89.7. According to random forest variable importance, tumor diameter, symptoms and extrathyroidal extension were the most important variables for predicting the local recurrence of differentiated thyroid carcinoma. The results of these studies demonstrate that a combination of machine learning and numeric data is expected to have great utility for predicting the quality of life after thyroid surgery or local recurrence of thyroid cancer.

In a similar vein, a combination of machine learning and genomic data would make great contribution for the early diagnosis of thyroid disease. The aim of a recent study [34] was to use machine learning and genomic data for the early diagnosis of papillary thyroid carcinoma. The source of data was 506 participants enrolled in Cancer Genome Atlas data and their 703 RNA features served as the attributes of this study. Among these attributes, two groups were selected as the most important variables in terms of random forest variable importance for the early diagnosis of papillary thyroid carcinoma: Group 1 including ADD3-AS1 (downregulation), MIR100HG (downregulation), FAM95C (downregulation), MORC2-AS1 (downregulation), LINC00506 (downregulation), ST7-AS1 (downregulation), LOC339059 (downregulation), MIR181A2HG (upregulation), FAM181A-AS1 (downregulation), LBX2-AS1 (upregulation), BLACAT1 (upregulation); Group 2 including hsa-miR-9-5p (downregulation), hsa-miR-146b-3p (upregulation), hsa-miR-199b-5p (downregulation), hsa-miR-4709-3p (upregulation), hsa-miR-34a-5p (upregulation), hsa-miR-214-3p (downregulation). The accuracy of machine learning based on Group 1/Group 2 was decision tree 92.2/87.0, random forest 99.2/99.5, and support vector machine 99.0/98.3.

In a similar context, the purpose of recent research [43] was to adopt machine learning and genomic data for the early diagnosis of hypothyroidism. The sample size of this study was 92 and the attributes of this study were gut microbiota RNA features. Among these features, veillonella, paraprevotella, neisseria and rheinheimera ranked the top in terms of random forest variable importance. Finally, a recent study [44] demonstrates that machine learning together with ultrasound data would provide effective non-invasive decision support systems for predicting the malignancy of thyroid nodules. Data came from 177 thyroid nodules and the following 10 attributes were considered: size, shape, margins, micro-calciﬁcation, composition, the echogenicity of the solid portion, halo sign, vascularity, the color scale scoring system of real-time elastography and strain ratio. The random forest showed the best performance in terms of accuracy and the AUC: logistic regression 84.2/92.8, random forest 86.0/93.4, support vector machine 84.8/92.3, gradient boosting 83.7/92.6, and artificial neural network 84.8/90.8. Among the ten attributes, wreath-shaped feature, micro-calciﬁcation and strain ratio were the most important variables in terms of random forest variable importance for predicting the malignancy of thyroid nodules.

4. Discussion

This study reviewed original studies including the random forest and the four other machine learning methods: The twenty original studies were selected out of 33 original studies in PubMed with the search terms “thyroid” (title) and “random forest” (abstract). This study put more focus on the random forest for two reasons. Firstly, it has the advantage of rigorous cross validation from “bootstrap aggregation”: it is a collection of many decision trees with a majority vote on the dependent variable. For example, a random forest with 1000 decision trees samples 1000 training sets with replacements, trains 1000 decision trees with the 1000 training sets, makes 1000 predictions with the 1000 decision trees, and takes a majority vote on the dependent variable. In other words, the random forest with 1000 decision trees uses rigorous 1000-fold cross validation and this explains why it usually shows the best performance together with boosting and neural network approaches [7,15,17,19]. Secondly, the random forest can analyze which predictors are more important for the early diagnosis of disease [7,15,17,19]. But another method can be more accurate and more appropriate than the random forest in certain circumstances. Little research has been done and more effort is to be made on this topic.

This study reveals that random forest variable importance would vary across different types of data for the early diagnosis of thyroid disease. The following attributes would be important variables in the case of numeric data: (1) clinical stage, marital status, histological type, age, nerve injury symptom, economic income and surgery type for predicting the quality of life 3 months after thyroid cancer surgery; tumor diameter, symptoms and extrathyroidal extension for predicting the local recurrence of differentiated thyroid carcinoma. Likewise, the list of important attributes in the case of genomic data would include: (1) RNA features including ADD3-AS1 (downregulation), MIR100HG (downregulation), FAM95C (downregulation), MORC2-AS1 (downregulation), LINC00506 (downregulation), ST7-AS1 (downregulation), LOC339059 (downregulation), MIR181A2HG (upregulation), FAM181A-AS1 (downregulation), LBX2-AS1 (upregulation), BLACAT1 (upregulation); Group 2 including hsa-miR-9-5p (downregulation), hsa-miR-146b-3p (upregulation), hsa-miR-199b-5p (downregulation), hsa-miR-4709-3p (upregulation), hsa-miR-34a-5p (upregulation), hsa-miR-214-3p (downregulation). The accuracy of machine learning based on Group 1/Group 2 was decision tree 92.2/87.0, random forest 99.2/99.5, and support vector machine 99.0/98.3.
study is to be done for greater external validity. However, the findings above would present useful guidelines on the effective application of random forest variable importance across a variety of data modes for the early diagnosis of thyroid disease in future research.

But current studies on the early diagnosis of thyroid disease based on machine learning has the following limitations. Firstly, many studies adopted cross-sectional data and employing longitudinal data would strengthen the performance of machine learning. Secondly, many studies used data with small sizes in single centers. Using big data (e.g., national health insurance claims data) would make valuable contributions for this area. Thirdly, most studies did not consider possible mediating effects among predictors. Fourthly, some studies reported accuracy or the AUC below 70.0 and these results would not be appropriate as diagnostic tests. Fifthly, binary categories (no, yes) are popular now but they can be refined to multiple categories with more clinical insights. Sixthly, combining different types of machine learning approaches for different types of thyroid data would bring new innovations in many aspects. Finally, it can be noted that this study did not use meta-analysis because different studies would have different diagnostic aims.

5. Conclusions

This article reviewed the recent progress of machine learning for the early diagnosis of thyroid disease. This review demonstrates that machine learning provides an effective, non-invasive decision support system for early diagnosis of thyroid disease.

Abbreviations

AUC, area under the receiver operating characteristic curve.

Author contributions

KSL and HP contributed to research design, data collection, analysis and interpretation, as well as manuscript writing, editing and review. KSL and HP approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This research was supported by the Ministry of Science and ICT of South Korea under the Information Technology Research Center support program supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) (IITP-2018-0-01405).

Conflict of interest

The authors declare no conflict of interest. KSL and HP are serving as the guest editors of this journal. We declare that KSL and HP had no involvement in the peer review of this article and have no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to PA.

References

[16] Lee K, Son S, Park S, Kim ES. Automated detection of colorectal


