International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Hypothyroidism Disease Diagnosis by Using Machine Learning Algorithms

Awad Bin Naeem[1], Biswaranjan Senapati[2], Alok Singh Chauhan*[3], Mukta Makhija[4], Arpita Singh[5], Meghna Gupta[6], Pradeep Kumar Tiwari[7], Wael M. F. Abdel-Rehim[8]

Submitted: 29/04/2023 **Revised**: 25/06/2023 **Accepted**: 08/07/2023

Abstract*:* Hypothyroidism is recognized as one of the most dangerous medical disorders in the world, requiring pricey therapy. This kind of research supports the implementation, development, and assessment of clinical decision support systems, with accruable diagnosis as its most important component. Increasing the accuracy of ML algorithms is essential for the development of high-performance computer-aided diagnostic systems. The purpose of this research was to show how ensemble approaches performed in a medical data set, which might be utilized to produce more accurate diagnoses and so enhance the health index. Three algorithms were employed in this investigation. It delivers a substantial result by comparing several algorithms, culminating in the conclusion of the study and the attainment of its major purpose. The purpose model was evaluated utilizing secondary hypothyroid data, according to the experimental findings. In recent years, many academics in the healthcare industry have presented several sorts of mining algorithms. The technique may not be suited for many applications due to its difficulties in detecting acceptable data types. The accuracy of the SVM machine-learning classifier is 84.72% in diagnostic patients' hypothyroidism symptoms in this research.

Keywords: *Medical Data Set, Hypothyroidism Symptoms, SVM, KNN, Naive Bayes*

1. Introduction

Diseases that fall during the disorder of Harmon due to the thyroid. It is a type of endocrine system [1]. It causes low heart rate, high temperature, dry skin, Anxiety, constipation, poor muscle tone, feeling cold, neck swelling and slow resting pulse rate [2]. The symptoms of Hyperthyroidism are dry skin, weight loss, nervousness, high heart rate, and irregular stomach movements. Millions of people have thyroid disorders, especially women [3]. Most of them have undiagnosed thyroid disease, which becomes a major problem. It was shown fatigue and depression to weight gain leading to a thyroid problem. It produced two types of large thyroid hormones, (T4) and (T3). In a supervised machine learning algorithm classification of the algorithm is a very important step. This algorithm needs very large data sets. These data sets are consisting of many features. The main function of the thyroid is to regulate the rate of metabolism. The IoT and AI are used to connect patients and medical staff so that information about diseases may be obtained. They require personal information like your age, gender, and the outcomes of your medical tests and treatments. Using algorithms, we apply various algorithms and determine which has the biggest influence on the outcomes of the diagnostic. In the existing medical knowledge network, we explain the relationships between things, but this is insufficient to solve problems.

Additionally, it has been noted that due to improperly timed examinations, thyroid illness is only discovered and treated when it is already chronic. The study of data is an interdisciplinary field that draws conclusions from complex, unstructured, and large datasets using scientific methods, procedures, algorithms, and systems and that significantly contributes and provides predictive analyses to a variety of application domains. Data science, in its broadest sense, is the study of data, including its origins, significance, and potential application as inputs and outputs for IT projects. Data mining is a method for extracting patterns and other significant information from large data sets. Data mining techniques have rapidly improved over the past few decades, transforming unusable data into knowledge. The data mining methods employed in these areas can be divided into two groups: either they use machine learning algorithms to define the target dataset, or they don't. The four fundamental steps in data mining include setting objectives, acquiring data and preparing it, employing data mining techniques, and evaluating the results. A KNN, decision trees, and AI networks are among the most widely used data mining methods.

The most crucial part of research is data analysis. The

[1] *Department of Computer Science, NCB&E, Multan, Pakistan*
[2] *Doctor in Computer and Data Science Parker Hannifin Corp, USA*
[3] *School of Computing Science & Engineering, Galgotias University, Greater Noida, India*
[4] *Integrated Academy of Management and Technology, Ghaziabad, India*
[5] *Integrated Academy of Management and Technology, Ghaziabad, India*
[6] *Department of Computer Applications, ABES Engineering College, Ghaziabad, India*
[7] *Dayanand Academy of Management Studies, Kanpur Nagar, India*
[8] *Faculty of Computers and Information, Suez University, Suez, Egypt*
* Corresponding Author Email: awadbinnaeem@gmail.com

information gathered is summarized using this technique. It involves applying logic and analysis to data to spot patterns, trends, and relationships [4]. Although there are other types of data analysis, we will just address predictive analysis [5]. Machine learning has been widely used in bio-medical datasets [6]. In this area, algorithms may be taught to reason like people, learning from a large number of datasets and making choices based on each person's comprehension and capacity for reasoning [7]. Data mining is defined as a more extensive withdrawal from enormous data sets that may contain hidden, possibly crucial information. SVM, MLPNN, DTF, and NBC, among other data mining and ML techniques, are used [8]. They have demonstrated effective outcomes in the prediction and categorization of datasets in the field of bioengineering [9].

Since original data is untidy and frequently confused when used, pre-processing methods for original data should be taken before applying data mining algorithms [10]. Additionally, handling missing data and removing cluttered, out-of-place, and anomalous information are necessities [11]. The valued qualities to describe the data are then decided using the data diminution [12]. The two steps of data mining techniques are as follows. The model is trained using the input and the anticipated output in the first step [13]. The second step, known as the test or validation phase, is where the model's effectiveness is assessed [14, 15]. The current study's objective is to calculate the ratio between the number of hypothyroidism cases and the number of units in the population at risk, as well as a means of improving public health by reshaping policy assessments by identifying the elements that pose a threat to infection and pursuing preventive healthcare goals by critically analysing various risk factors for disease transmission. Additionally, we will compare model findings and construct models that can be deployed in the Rapid Miner tool to test prediction accuracy. In our study different Algorithms has apply to solve the problem and find the best solution. In this study, a machine learning algorithm was used for hypothyroid disease. In the computer field, it will help the basic drawbacks of Hypothyroid and reduce the patient illness factors. A completely automated technique has been suggested to handle this problem, removing the danger of human mistakes and shortening the time required to establish the severity of the illness. The proposed approach will benefit medical practitioners and researchers by making hypothyroidism disease diagnosis and management easier.

Reduction in overfitting

1. Increase Precision
2. Shortened training period
3. To find the symptoms and some factors of hypothyroid using a machine learning framework.
4. To be able and help the patient, medical officials and medical companies about that with finding factors.
5. By applying different techniques and methodologies to find more hidden features of ML and DL.

Different related methodologies are approved for diagnosing thyroid disorders. To find the correct thyroid disorder overall methodologies provided decent benefit. Due to hormonal changes, the level of the thyroid was also fluctuating. Therefore, it is an issue for classification which needs to address to clear these problems. At an early stage, it will help the medical practitioner to diagnose the ailment of the thyroid. Especially, it gives better knowledge of thyroid cancer to avoid complexities.

The structure of this work is as follows: The methodology of the investigation is presented in Section 2. Section 3 explains the results. Section 4 finally describes the conclusion and future work.

2. Methodology

We used the Rapid Miner tool using KNN, Naïve Bayes and SVM. Our study consists of three steps. In the first step, we collect the related data of hypothyroid in Kaggle. Second, we compare the classification of the model. In the third, we finalized the result.

1. Add the Dataset
2. Data techniques
3. Comparative Analysis of Algorithms
4. Result

2.1 Data Collection

All the data was gathered from the Kaggle dataset which consists of 3371 patient data [15]. The sample size was 371 patients. The dataset consists of 33 attributes shown in table 1.

Table 1. Attributes of Data Set

SR. No	Attributes	Parts
1	Patient ID	
2	age	
3	sex	
4	sick	
5	Iodine deficiency	
6	radiation therapy	
7	Medications	
8	pregnant	

9	goitre	Factors
10	tumour	
11	hypoituitary	
12	TSH measured	
14	t3 measured	
16	tt4 measured	
18	t4U measured	
19	ttI measured	
20	TBG measured	
22	weight gain	Symptoms
23	fatigue	
24	brain tog	
25	Low pulse rate	
27	feeling cold	
28	Constipation	
29	Poor muscle tone	
31	Sugar Level increase	
32	Neck swelling	
33	Slow resting pulse rate	

2.2 Population of Sampling

Classification is used to classify data into predefined labels class. It is defining the dependent variable. It takes 371 sample data from the Kaggle dataset.

2.2.1 K-Nearest Neighbour (KNN)

One of the most fundamental ML algorithms based on supervised learning. Assuming that the new case/data and the present cases are comparable, it assigns the new instance to the classification that is closest to the current categories. The system categorizes new data points based on similarity to all previously stored data. Being a non-parametric approach, it makes no assumptions about the underlying data. KNN categorizes new data into a category that is highly similar to the existing data by merely storing the data during the training phase. Consider the scenario when we are confused about the species in a snapshot since it resembles both cats and dogs. As a result, because the KNN approach is based on a similarity measure, we may apply it to this identification. Our KNN model will decide which group the images belong to cat or dog, by comparing the attributes of the new data set to those in the photographs of cats and dogs.

2.2.2 Naïve Bayes

The Naive Bayesian classifier is one of these probabilistic statistical classifiers. Assumptions that the existence of certain features in a dataset is independent of the existence of other features are referred to as "naive". The "naive" hypothesis reduces the complexity of the computation to a straightforward probability multiplication. Due to the algorithm's simplicity, handling datasets with many dimensions was made simple.

The following equation shows how the technique produces posterior probability P (c| x) from the probabilities of x, y, and P (x | c):

X | C = P (C | X) P (C) / P (X) (X).

2.2.3 Support Vector Machine (SVM)

The SVM is utilized to categories the data. The main step of the process is sorting and categorizing things. A large number of data will all contain identical information, and we can use it to identify patterns. It takes a scientific or algorithmic approach to problem management in general. By selecting the maximum and excess data that can discriminate between the given values and confusing values, the SVM technique eliminates insufficient elements. A supervised binary machine learning classifier called a support vector machine (SVM) is most frequently used for categorization and making decisions about outliers. SVM has been used to successfully resolve a variety of classification problems, including text classification, image processing, object recognition, etc. The SVM functions as:

- To categories the data, map it to a high-dimensional feature space.
- Use a decision boundary to categories the data.
- Modify the data to create a decision boundary on a hyperplane.
- To get the ideal hyperplane based on the support vectors, determine the distance (Margin) between the points and the hyperplane using a vector. The hyperplane has the largest margin

2.3 Data Mining Software

RAPID MINER is an open-source collection of machine learning algorithms and data processing tools. To explore if there is a way to speed up and enhance the interpretation of the hypothyroid data set, the RAPID MINER data mining program is utilized. The data has to go through some preprocessing before being put into RAPID MINER. For usage in RAPID MINER, a copy of the Excel data set prepared for statistical analysis was made, and it was then converted to the CSV file format. The CSV file extension allows for initial analysis before turning the trial results into an ARFF RAPID MINER data file and saving them. After the preprocessing stage, the data mining platform

offered a variety of data interpretation possibilities, such as classification and cluster association techniques. Since the researchers only need a tiny fraction of the hypothyroid data set's results and values to be missing, no filtering was necessary. The initial screen provides the researchers with a set of data they need, and using current statistical techniques, it took a long time to complete. Nave Bayes, SVM, and Logistic regression were applied to classify the hepatitis patients using the complete hypothyroid data set. By categorizing the training set of data and observing the instances that were correctly classified, as well as by applying NB to the test set and observing the cases that were correctly and mistakenly recognized, the model was constructed. By contrasting them, you can determine which is more accurate.

3. Results & Discussions

In the overall study of hypothyroid, we analyzed that different models were used to find a better result. In our study, we will find which Algorithm provides us with the highest accuracy with the help of measured data and dataset parameter factors (Patient of sick, iodine deficiency and goitre) shown in fig 1.

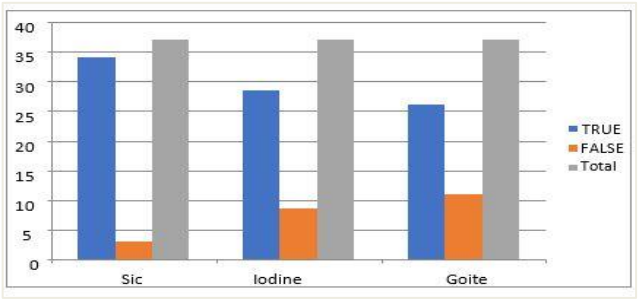

Fig. 1. Patients of sick, iodine deficiency & goiter

The diagnosis at an early stage can find due to hormonal changes. However, with time various features in the dataset difficult to find the level of hypothyroid. We have a better approach to solving this issue by applying real dataset factors (Patient of sick, iodine deficiency and goitre) that are shown in table 2. It will help Practitioner as well patient to avoid the risk associated with Hypothyroid.

Table 2. Patients of Sick, Iodine Deficiency & Goitre

Factor	True	False	Total
Sick	340	30	370
Iodine deficiency	285	85	370
Goiter	260	110	370

We have a better Technique to solve this issue by utilizing real dataset symptoms (patient weight Gain, exhaustion and brain fog) that are displayed in fig 2. It will assist Practitioner as well patient to avoid hazards linked with Hypothyroid.

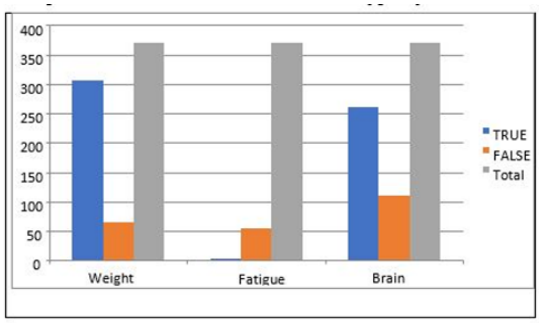

Fig. 2. Patients with weight gain, fatigue & brain fog

Unfortunately, with time many characteristics in the dataset difficult to detect the amount of hypothyroid. With real dataset symptoms, we can solve this issue more effectively (patient weight gain, fatigue and brain fog) that are presented in table 3. It will enable Practitioner as well patient to avoid danger related to Hypothyroid.

Table 3. Patients with weight gain, fatigue & brain fog

Symptoms	True	False	Total
Weight Gain	305	65	370
Fatigue	315	55	370
Brain Fog	260	110	370

One of the most severe medical disorders in the world, hypothyroidism requires expensive treatment. The accruable diagnosis is the most crucial element in this sort of study, which supports the creation, implementation, and assessment of clinical decision support systems. The accuracy of the KNN machine-learning classifier is 79.25% which is shown in table 4 for applying our data set to a diagnosis system.

Table 4. KNN Classifier

accuracy: 79.31% +/- 2.22% (mikro: 79.25%)

	true Medications	true 1	true 0	class precision
pred. Medications	0	0	0	0.00%
pred. 1	1	1039	148	87.46%
pred. 0	2	131	38	22.22%
class recall	0.00%	88.80%	20.43%	

One of the most serious medical conditions in the world, hypothyroidism requires expensive treatment. This kind of research encourages the implementation, creation, and evaluation of clinical decision support systems., with the accruable diagnosis being the most significant component. In table 5, the accuracy of the NB is shown to be 74.77% when our data set is applied to a diagnostic system.

Table 5. Naïve Bayes Classifier

accuracy: 74.77%

	true Medica1ions	true 1	true 0	class precision
pred. Medica1ions	0	0	0	0.00%
pred. 1	0	80	13	86.02%
pred. 0	0	15	3	16.67%
class recall	0.00%	84.21%	18.75%	

Hypothyroidism is acknowledged as one of the world's most hazardous medical illnesses, necessitating pricy medication. This sort of research aids in the installation, development, and evaluation of clinical decision support systems, with the accruable diagnosis being the most crucial component. Table 6 shows the accuracy of the SVM for using our data set in a diagnostic system, which is 84.72%.

Table 6. SVM Classifier

accuracy: 84.72% +/- 2.69% (mikro: 84.71%)

	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	201	1114	84.71%
class recall	0.00%	100.00%	

In this study, 3 machine learning algorithms (KNN, SVM and Naive Bayes) are used. It completes the study and fulfils its main objective by comparing several methods and obtaining a noteworthy result. The purpose model was evaluated using secondary hypothyroid data, following the experimental findings. Table 7 clearly shows that SVM shows better accuracy and solves our problem in a better way.

Table 7. Comparison Results of Machine Learning Classifier

Algorithm	**true 1**	**true 0**	**Accuracy**
KNN	88.80%	20.43%	79.31%
Naïve Bayes	84.21%	18.75%	74.77%
SVM	100%	0.00%	84.72%

Fig 3 shows a comparison of show comparison of three machine learning algorism (KNN, Naïve Bayes and SVM) on the base of three variables accuracy, true1 and true 0. Fig 3 clearly shows that SVM has better results as compared to other machine learning algorithms.

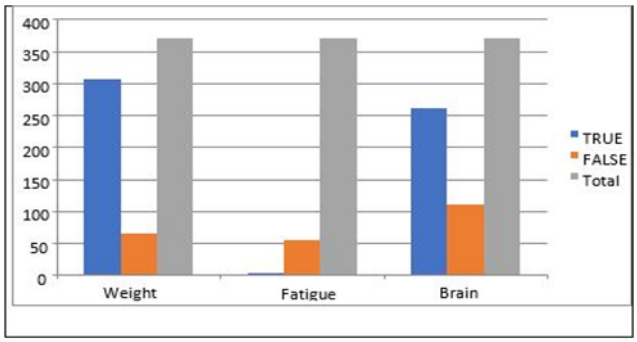

Fig. 3. Comparison Results of Machine Learning Classifier

5. Conclusion and Future work

This research included a summary of the methods utilized in data mining to identify hypothyroidism. Three algorithms were used to show how well ensemble approaches performed on a collection of medical data that may be utilized to enhance the health index and make diagnoses with more accuracy. Although there is no test for hyperthyroid patients, 84.72% of the participants had hypothyroidism. Four interfaces—Explorer, Experiment, Knowledge Flow, and Simple CLI—were used with the Rapid miner tool. A bigger data set might be utilized to create the model, which was built using three classifiers. New models, such as deep learning and fuzzy learning, will be developed to increase accuracy.

Author contributions

Awad bin Naeem 1: Conceptualization, Methodology, Software, Writing-Reviewing. **Biswaranjan Senapati 2:** Data curation, Writing-Original draft preparation, Field study. **Alok Singh Chauhan 3:** Methodology, Visualization, Investigation. **Mukta Makhija 4**: Writing-Reviewing and Editing. **Arpita Singh 5**: Writing-Reviewing and Editing. **Meghna Gupta 6:** Software. **Pradeep Kumar Tiwari 7:** Field study. **Wael M. F. Abdel-Rehim 8**: Software.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] S. S. Islam, M. S. Haque, M. S. U. Miah, T. B. Sarwar, and R. Nugraha, "Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study," PeerJ Computer Science, vol. 8, p. e898, 2022/03/03 2022, doi: 10.7717/peerj-cs.898.

[2] K. Guleria, S. Sharma, S. Kumar, and S. Tiwari, "Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning," Measurement: Sensors, vol. 24, p. 100482, 2022/12/01/ 2022, doi:

https://doi.org/10.1016/j.measen.2022.100482.

[3] K. salman and E. Sonuç, "Thyroid Disease Classification Using Machine Learning Algorithms," Journal of Physics: Conference Series, vol. 1963, no. 1, p. 012140, 2021/07/01 2021, doi: 10.1088/1742-6596/1963/1/012140.

[4] Y. Li et al., "Serum Raman spectroscopy combined with Deep Neural Network for analysis and rapid screening of hyperthyroidism and hypothyroidism," Photodiagnosis and Photodynamic Therapy, vol. 35, p. 102382, 2021/09/01/ 2021, doi: https://doi.org/10.1016/j.pdpdt.2021.102382.

[5] M. Hosseinzadeh et al., "A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things," The Journal of Supercomputing, vol. 77, no. 4, pp. 3616-3637, 2021/04/01 2021, doi: 10.1007/s11227-020-03404-w.

[6] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, "Thyroid Disease Prediction Using Machine Learning Approaches," National Academy Science Letters, vol. 44, no. 3, pp. 233-238, 2021/06/01 2021, doi: 10.1007/s40009-020-00979-z.

[7] L. Aversano et al., "Thyroid Disease Treatment prediction with machine learning approaches," Procedia Computer Science, vol. 192, pp. 1031-1040, 2021/01/01/ 2021, doi: https://doi.org/10.1016/j.procs.2021.08.106.

[8] H. Abbad Ur Rehman, C.-Y. Lin, Z. Mushtaq, and S.-F. Su, "Performance Analysis of Machine Learning Algorithms for Thyroid Disease," Arabian Journal for Science and Engineering, vol. 46, no. 10, pp. 9437-9449, 2021/10/01 2021, doi: 10.1007/s13369-020-05206-x.

[9] X. Chai, "Diagnosis Method of Thyroid Disease Combining Knowledge Graph and Deep Learning," IEEE Access, vol. 8, pp. 149787-149795, 2020, doi: 10.1109/ACCESS.2020.3016676.

[10] B. Zhang et al., "Machine Learning–Assisted System for Thyroid Nodule Diagnosis," Thyroid, vol. 29, no. 6, pp. 858-867, 2019/06/01 2019, doi: 10.1089/thy.2018.0380.

[11] J. Song et al., "Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules," (in eng), Medicine (Baltimore), vol. 98, no. 15, p. e15133, Apr 2019, doi: 10.1097/md.0000000000015133.

[12] A. H. Shahid, M. P. Singh, R. K. Raj, R. Suman, D. Jawaid, and M. Alam, "A Study on Label TSH, T3, T4U, TT4, FTI in Hyperthyroidism and Hypothyroidism using Machine Learning Techniques," in 2019 International Conference on Communication and Electronics Systems (ICCES), 17-19 July 2019 2019, pp. 930-933, doi: 10.1109/ICCES45898.2019.9002284.

[13] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," BioMed Research International, vol. 2022, p. 9809932, 2022/06/07 2022, doi: 10.1155/2022/9809932.

[14] A. Tyagi, R. Mehra, and A. Saxena, "Interactive Thyroid Disease Prediction System Using Machine Learning Technique," in 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), 20-22 Dec. 2018 2018, pp. 689-693, doi: 10.1109/PDGC.2018.8745910.

[15] M. A. A. R. Asif et al., "Computer Aided Diagnosis of Thyroid Disease Using Machine Learning Algorithms," in 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 17-19 Dec. 2020 2020, pp. 222-225, doi: 10.1109/ICECE51571.2020.9393054.

[16] Mr. Dharmesh Dhabliya, Ms. Ritika Dhabalia. (2014). Object Detection and Sorting using IoT. International Journal of New Practices in Management and Engineering, 3(04), 01 - 04. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/31

[17] Shelar, Y. ., Sharma, P. ., & Rawat, C. S. D. . (2023). An Improved VGG16 and CNN-LSTM Deep Learning Model for Image Forgery Detection. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 73–80. https://doi.org/10.17762/ijritcc.v11i3s.6157