Thyroid Function Classification using Multi-layer Perceptron

Elsha M. Siochi

I. Introduction

According to The Philippine Thyroid Diseases Study (PhilTiDeS 1) conducted in 2008 [1], 8.53% of the 4,897 tested non-pregnant Filipino adults had thyroid function abnormalities. These abnormalities can be classified into hypothyroidism and hyperthyroidism, with subclinical (mild) and overt severity. The thyroid is a small butterfly-shaped gland responsible for secreting T3 (triiodothyronine) and T4 (thyroxine) [2]. These hormones are responsible for regulating how the body uses energy, known as metabolism. The thyroid is normally instructed by the pituitary gland to release the two thyroid hormones, by secreting TSH (thyroid-stimulating hormone). In a similar fashion, the hypothalamus tells the pituitary gland to secrete TSH by releasing TRH (thyrotropin-releasing hormone) [3]. The relationship between the glands are shown in Fig. I.

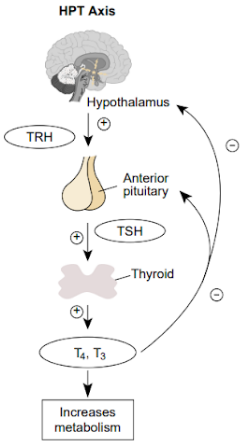

Fig. 1. Hypothalamus-Pituitary-Thyroid Axis, image from [4]

On one hand, an under active thyroid gland which produces insufficient hormones results in hypothyroidism. This may cause the patient to gain weight, become depressed and forgetful, feel weakness, and in extreme cases cause heart failure and fatal *myxedema* coma. On the other hand, an over active thyroid which produces excessive hormones results in hyperthyroidism. This may cause the patient to lose weight, develop irregular heartbeat called *arrhythmia*, experience muscle weakness and excessive sweating, and in extreme forms develop thyroid storm which is fatal [2].

The initial tests used to screen thyroid function abnormalities are TSH and free T4 (thyroxine) tests. This helps determine whether the abnormality stems from the thyroid, pituitary, or hypothalamus. An important note to remember is pregnancy may elevate the normal levels of thyroid hormones, as the body produces an excess to support the baby. In this case, thyroid problems are harder to diagnose since hormones fluctuate and symptoms of pregnancy and thyroid abnormality may overlap. Depending on the results of the preliminary tests, other forms of screening may be required such as T3, binding protein tests, and more [3].

As the country lacks medical professionals, automating a diagnosis could be beneficial to ease the load of endocrinologists. Given a dataset containing quantitative and categorical attributes relevant to hypo- and hyperthyroidism, this research aims to experiment on a new artificial neural network (ANN) model to automate the classification of the thyroid function into normal, hypothyroidism, and hyperthyroidism. This does not intend to replace the doctor's diagnosis, rather act as a preliminary assessment to be confirmed still by the professional.

A. Inputs

There are 21 attributes in total in the dataset donated by Peter Turney, of which only eight were used in the model. There were plenty of studies which used nearly all attributes. The researcher opted to hand-pick only a few of those, to experiment on how the classifier will perform even with less features. The eight attributes used were (1) age, (2) sex, (3) pregnant, (4) Thyroid stimulating hormone or TSH value, (5) Triiodothyronine or T3 value, (6) Total Thyroxine or TT4 value, (7) Total Thyroxine Uptake or T4U value, and (8) Free Thyroxine Index or FTI value. The latter five values are results from blood tests done to evaluate the thyroid function of an individual. The categorical data like sex (male or female) and pregnant (true or false) were already encoded as numerical values (0 or 1) and continuous values were encoded within 0 to 1. Table 2 shows a summary of the features used, as found in the dataset.

Attribute	Type
age	Continuous
sex	Binary, 0 or 1
pregnant	Binary, 0 or 1
TSH	Continuous
T3	Continuous
TT4	Continuous
T4U	Continuous
FTI	Continuous

TABLE 2
Features used in machine learning model

Submitted to Professor Jaderick Pabico, in fulfillment of the requirements for CMSC 191: Introduction to Neural Networks

Afterwards, MinMax Normalization was performed separately on 90% train and 10% test split to scale the continuous values within 0 to 1. Normalization is done to represent different features within the same scale, which improves stability of the model [5]. For example, even out the importance of a feature whose values range from 1,000 to 10,000 and a feature which goes from 0 to 10. The training data was first applied with Min-Max scaling. Then, the minimum and maximum variables learned after fitting the training data, were also used to normalize the test dataset. The scaling was done separately to avoid the final test dataset, which acts as the unseen data, from affecting the tuning of the parameters of the model. The formula for MinMax normalization is done independently per feature and is given by:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where
x' is the normalized value
x is the current value
x_{min} is the minimum value within that feature
x_{max} is the maximum value within that feature

B. Outputs

The classification output in the dataset are 1, 2, and 3 corresponding to hyperthyroidism, hypothyroidism, and normal respectively. Since these are dependent variables, no normalization was needed to be applied. However, to arrive at these classifications in the output layer, softmax activation function was used. Three output nodes were used to represent the classes, each containing the probability that the instance belonged to that class. More details about the activation function can be found in the ANN discussion.

II. Review of Literature

Multiple studies were done using the thyroid disease classification dataset [6] from University of California, Irvine (UCI) Machine learning repository, with different attributes and approaches used. The dataset contains smaller separate datasets whose classifications vary. The following discussion will revolve around the studies and core concepts which considered in crafting the proposed solution.

A. Previous Studies

In a 1998 study by Zhang and Berardi [7] on neural networks and thyroid functions, they conducted experiments that had shown neural networks performed better compared to logistic regression. The authors used a fully-connected three-layer feed forward neural network, a multi-layer perceptron (MLP); citing studies which deem one hidden layer as sufficient for the problem at hand. The network consisted of ten hidden nodes, and three binary output nodes corresponding to three classifications (normal, hypothyroid, hyperthyroid). As an example, a hypothyroid case will have 0-1-0. The authors used all 21 features of the dataset, and performed a stratified four-fold cross validation on the 7200 instances in the dataset then used accuracy as the primary metric to assess models. The average accuracy obtained was 98.55%. The authors recommended selecting only some of the attributes to be used in the future neural networks. They also emphasized the usefulness of k-fold cross validation in designing models able to classify in a variety of sampling situations.

A related study was conducted in 2013 [8], where the researchers focused on the finer-grained classification of hypothyroid disease which are: compensated hypothyroid, primary hypothyroid, secondary hypothyroid, and negative. The particular data is also available in the thyroid disease dataset, and has 29 attributes in particular. The study compared various classifiers implemented in Weka such as Bayesian net, MLP, Radial Basis Function (RBF) networks, Decision stump, C4.5 Decision trees, Classification and Regression Tree (CART), Reduced Error Pruning (REP) tree. In all classifiers, a six-fold cross validation was performed. In contrast to the earlier study, the MLP classifier did not perform well when compared against mentioned decision tree algorithms. The latter garnered higher accuracy, precision, recall, and F-measure reaching around 99% versus the former (MLP) with 93% to 94% score in all metrics.

Another previous study focuses on thyroid disease diagnosis using a variety of neural networks [9]. The author used a different dataset, again from UCI, with only five features resulting from thyroid hormone-related blood tests. The MLP design contain two hidden layers each with 50 neurons, and an output layer with three binary output nodes, and a sigmoid activation function. The MLP performed second best with an accuracy of 92.96% compared to the Probabilistic Neural Network (PNN) which has 94.81%, in a ten-fold cross validation.

B. Background of the Proposed Solution

Taking into consideration the mentioned literature, this study proposes the use of an MLP paired with grid search and stratified five-fold cross validation to choose a suitable set of parameters.

1) Multi-layer Perceptron: Neural networks are a class of models whose connections take after the brain. These networks communicate by passing information from one layer to another, much like the passing of electrical signals between the neurons of the brain. When the signal moves in a single direction per pass, with no output of a layer looping to itself, it is categorized as a feedforward neural network. A multi-layer perceptron (MLP) is a form of feedforward neural network, which are fully connected from input to hidden to output layers. Figure 2 is an example of an MLP with four hidden layers.

However, the output of a single forward pass may not be enough to correctly classify the input [10]. The loss or cost function models the disparity between the predicted and actual classification, at least with respect to the training set. Through backpropagation algorithm, the weights of the neurons are readjusted to move towards minimizing the cost.

2) Grid Search: Grid search is the process of conducting an exhaustive search through different combination of parameters and their values, to find the optimal model. Since the studies

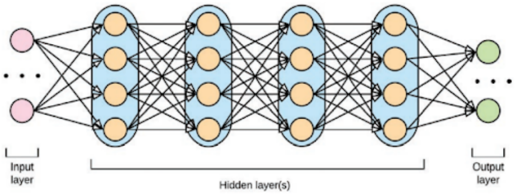

Fig. 2. A multi-layer perceptron model with four hidden layers, image from [10]

do not pose general recommended parameters, the researcher opt to perform grid search on limited parameter values.

3) Stratified K-fold Cross Validation: The stratified k-fold cross validation technique attempts to maintain the percentage of classes in each data split. This is especially valuable in the current study, which is characterized by an imbalanced dataset. Suppose a random split created has zero representation of either the hypothyroid or hyperthyroid cases, which are the minority, the validation score for the said split will be misleading. Stratified k-fold cross validation ensures representation and sample variability.

III. Methodology

In building the architecture and evaluating the results, the libraries used are: scikit-learn, imbalanced-learn, numpy, and pandas.

The data was split into a 90:10 stratified ratio, where 90% was used to tune the parameters, and 10% was used to test the final model. Grid Search was performed to tune the hyperparameters namely: hidden layer, learning algorithm, number of epochs, and activation function for the hidden nodes. Stratified five-fold cross validation was performed during this stage. Shown in Table 3 is the hyperparameter space used to search for the optimal model. There were a total of 2 x 3 x 2 x 4 = 48 models trained from the combination of parameters used in grid search.

Model attribute	Possible configurations
Hidden layer(s)	One layer with 12 nodes Two layers with 12 and 8 nodes respectively
Activation function	Sigmoid Tanh Rectified linear unit (ReLu)
Learning algorithm	Stochastic gradient descent Adam
Epochs	200 300 400 500

TABLE 3

Hyperparameter space specified for Grid Search

The other parameters were left at their default values in *scikit-learn version 1.0.2 MLPClassifier* in the duration of the search. Afterwards, the best model was chosen through the geometric mean score (G-Mean), and tested on the hold out test set. The details of the architecture found in the ANN Architecture section is based on the best-performing model.

A. ANN Architecture

The best-performing model from the Grid Search is a two-hidden layer MLP, and its parameters are summarized in Table 4.

Model attribute	Configuration
Hidden layer(s)	Two layers with 12 and 8 nodes respectively
Activation function	Rectified linear unit (ReLu)
Learning algorithm	Adam
Epochs	500
Initial learning rate	0.001
Alpha	0.0001
Beta 1	0.9
Beta 2	0.999
Epsilon	1e-8

TABLE 4

Configuration of the MLP model

Rectified Linear Unit (ReLU) is the activation function for the hidden layer of the best-performing model. It is said to diminish the effect of the vanishing gradient problem [10], wherein the gradient becomes too small for any weight update to occur. This is encountered in gradient-based learning algorithms, which were used in this study. It can be represented as:

$$f(x) = max(0, x)$$

where x is an input value

Softmax activation was used in the output layer, to assign probabilities that the instance is in either of the classes. In essence, three output nodes were used to represent the probabilities for each class. The probabilities in the resulting vector should sum to 1, and the output node with the highest probability is regarded as the final classification. The softmax function is written as [10]

$$\sigma(\hat{y}(k))_i = \frac{e^{\hat{y}(k)_i}}{\sum_{j=1}^{K} e^{\hat{y}(k)_j}}$$

where
$i = \{1, ..., K\}$classes and
$\sigma(\hat{y}(k))_i)$ results to a k-sized vector of estimated probabilities of the instance belonging to each class. The final classification can be obtained through $argmax$.

The Adam algorithm was used to solve for the optimal weights. It is an optimized version of the widely used Stochastic gradient descent. According to its proponents, it is computationally-efficient and requires little memory, thus recommended for studies with a large number of data, parameters, or both. [11].

B. Dataset

The thyroid dataset used was sourced from the University of California at Irvine's Machine Learning Repository. The original thyroid dataset contains smaller databases. In particular, what was used here was the collection containing 3772 training data and 3428 testing data donated by Peter Turney which were also derived from Ross Quinlan's dataset. It is different from the others because all attributes in this were already normalized. It also has three classifications namely: normal, hypothyroidism, and hyperthyroidism only, as opposed to others with more specific diagnosis. The dataset used was already normalized. Thus, no further standardization was done.

One of the challenges mentioned in the study of Zhang et al. [7] is the imbalanced sample data for each classification, with normal amounting to 92.6% of the data. However, the data can be considered an adequate representation of the true rate of cases in comparison to the PhilTiDeS 1 study, where there are indeed more normal cases than ones with increased or diminished functioning thyroids. Instead of performing resampling techniques, additional metrics were used to better estimate the performance on the minority classes.

In partitioning the data, first, the 7200 samples were split into 90:10 stratified ratio of train data (6480 samples) for tuning the parameter, and test data (720 samples) for testing the final model.

Second, the 6480 samples were used in the stratified five-fold cross validation. This sampling technique was chosen to achieve a more balanced and varied evaluation while searching for the optimal model. Since there are five folds, on each run, the model was able to train on 5184 instances and test on the remaining validation fold of 1296 instances. All the while aiming to maintain the 2.3% hypothyroid, 5.1%, hyperthyroid, and 92.6%, normal thyroid case representation in each fold.

Finally, the held out test set of 720 samples was used to evaluate the performance of the optimal model, with the percentage of classifications still maintained.

C. Quality Assurance

Given that 92.6% of the data contains normal classification, an accuracy of 92.6% could mean correctly classifying only the normal thyroid function instances due to its prevalence in the training, which defeats the purpose of automated disease detection. To ameliorate this, metrics such as G-mean, precision, recall, and F1-measure were also used to consider the minority classes which exhibit thyroid dysfunction.

In tuning the parameters to get the optimal model, G-mean, Accuracy, and Macro F1-measure were monitored, however G-mean is the basis for the ranking of the models. The descriptions and formulas for each metric are listed below.

1) Accuracy: The accuracy is the rate of correct positive and negative predictions.

$$Acc = \frac{TP + TN}{N}$$

where

TP is the number of true positives

TN is the number of correctly detected true negatives

N is the total number of instances of that class, or TN + TP + FN or false negatives + FP or false positives

2) Precision: The precision measures the proportion of predicted positives that are actually correct. For example, a high precision in hypothyroid class means most or all of the instances predicted as hypothyroid are indeed hypothyroid cases.

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

3) Recall or Sensitivity: The recall measures the proportion of actual positives predicted correctly. In the study and in medicine in general, it is ideal to have a high recall with respect to the disease-ridden instances. For example, a high recall in classifying hypothyroidism meant no to low false negatives, and instances are not misclassified as hyperthyroidism or normal. This ensures that thyroid dysfunction cases are properly diagnosed and are given necessary attention.

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

4) F1-score: The F1-score is a variant of F-beta score where beta is 1. It measure the trade-off between precision and recall, by computing the harmonic mean between the two [12]. A beta of 1 indicates that precision and recall are equally important in the formula. In the parameter-tuning stage, the F1-scores per class were computed and averaged to get Macro F1.

$$F1 = \frac{(\beta^2 + 1) \times R \times P}{R \times \beta^2 \times P}$$

5) G-mean score: The geometric mean is given by the root of the product of recall and specificity, in binary classifications. However in the imbalanced-learn library, geometric mean for multi-class cases is the higher root of the product of the class-wise recall scores. Since the study uses an imbalanced dataset and majority of the instances are normal, G-mean was utilized as it can represent the performance of the model in minority classes.

$$Sp = \frac{TN}{TN + FP}$$

$$G = \sqrt{R \times Sp} \text{ for binary classification}$$

$$G = \sqrt[n]{R_1 \times R_2 \times R_3 ... \times R_n} \text{ for multi-class classification}$$

where

n is the number of classes

R is the recall for a class

Sp is the specificity for a class, which measures proportion of actual negatives predicted correctly

IV. Results

The precision, recall and F1 scores of the best-performing model, when tested against the hold out test dataset is summarized in Table 5. It obtained an accuracy of 97.92%, greater than the baseline 92.6% from only classifying the normal cases correctly. Precision in the minority classes is only satisfactory, it means there are cases mistyped as hypo- or hyperthyroid. However, the recall rate is high among classes, which is vital for proper diagnosis and treatment. This means important cases such as hypo- and hyperthyroid cases are diagnosed correctly, and not misclassified as normal.

With regards to grid search with cross validation, the top five model configurations are shown in Table 6. The order is based on their G-mean ranking. Although, the rank of the

Classes	Precision	Recall	F1-score	No. of test instances
1 (Hyperthyroid)	0.8333	0.9375	0.8824	16
2 (Hypothyroid)	0.7955	0.9459	0.8642	37
3 (Normal)	0.9954	0.9820	0.9887	667

TABLE 5

TEST SCORES OF THE FIVE BEST-PERFORMING MODELS

models in terms of G-mean, Marco F1, and Accuracy are actually consistent. Table 7 contains the validation scores and Table 8 contains the training scores. While it cannot be said that overfitting is decidedly removed, the difference between the training and validation scores during cross validation are minimal.

Model rank	Hidden layers	Activation function	Learning algorithm	Epochs
1	Two layers: 12 and 8 nodes	ReLu	Adam	500
2	Two layers: 12 and 8 nodes	ReLu	Adam	400
3	Two layers: 12 and 8 nodes	ReLu	Adam	300
4	Two layers: 12 and 8 nodes	Tanh	Adam	400
5	Two layers: 12 and 8 nodes	Tanh	Adam	500

TABLE 6

CONFIGURATION OF THE FIVE BEST-PERFORMING MODELS

Model rank	G-mean	Macro-F1	Accuracy
1	0.9340	0.8763	0.9708
2	0.9326	0.8747	0.9707
3	0.9112	0.8607	0.9681
4	0.8966	0.8541	0.9648
5	0.8946	0.8534	0.9645

TABLE 7

VALIDATION SCORES OF THE FIVE BEST-PERFORMING MODELS

Model rank	G-mean	Macro-F1	Accuracy
1	0.9468	0.8938	0.9737
2	0.9467	0.8934	0.9736
3	0.9234	0.8759	0.9703
4	0.9015	0.8630	0.9665
5	0.9086	0.8710	0.9679

TABLE 8

TRAIN SCORES OF THE FIVE BEST-PERFORMING MODELS

The recommendations include the conducting of oversampling and undersampling techniques since the dataset is highly imbalanced, and experimenting on the number of hidden layers and units.

REFERENCES

[1] J. Carlos-Raboca, C. Jimeno, S. Kho, A. Andag-Silva, G. Jasul, N. Nicodemus Jr, E. Cunanan, and C. Duante, "The Philippines Thyroid Diseases Study (PhilTiDeS1): Prevalence of Thyroid Disorders Among Adults in the Philippines," *Journal of the ASEAN Federation of Endocrine Societies*, vol. 27.

[2] S. Witemeyer, "Thyroid Disorders: Hypothyroidism and Hyperthyroidism." [Online]. Available: https://coc.unm.edu/common/manual/thyroid_disorders.pdf

[3] M. Armstrong, E. Asuka, and A. Fingeret, *Physiology, Thyroid Function*. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK430685/

[4] S. Hiller-Sturmhöfel and A. Bartke, "The endocrine system: An overview," *Alcohol Health and Research World*, vol. 22, no. 3, p. 153–164, 1998. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6761896/

[5] "Normalization — data preparation and feature engineering for machine learning — google developers." [Online]. Available: https://developers.google.com/machine-learning/data-prep/transform/normalization

[6] J. R. Quinlan, "Induction of decision trees." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Thyroid Disease

[7] P. Zhang and V. Berardi, "An investigation of neural networks in thyroid function diagnosis," *Health Care Management Science*, vol. 1, pp. 29–37, 09 1998.

[8] S. Pandey, R. Miri, and S. R. Tandan, "Diagnosis and classification of hypothyroid disease using data mining techniques," *IJERT*, vol. 2, pp. 3188–3193, 06 2013.

[9] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," *Expert Systems with Applications*, vol. 36, no. 1, pp. 944–949, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417407004903

[10] E. O. Bisong, *Building machine learning and deep learning models on Google Cloud Platform: a comprehensive guide for beginners*. Apress, 2019.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[12] R. Espíndola and N. Ebecken, "On extending f-measure and g-mean metrics to multi-class problems," *Sixth international conference on data mining, text mining and their business applications*, vol. 35, pp. 25–34, 01 2005.