Journal of Modern Technology and Engineering
Vol.7, No.2, 2022, pp.134-145

DIAGNOSIS OF THYROID DISEASE USING MACHINE LEARNING TECHNIQUES

Ege Savcı[1], Fidan Nuriyeva[1,2*]

[1]Dokuz Eylul University, Department of Computer Science, Izmir, Turkey
[2]Institute of Control Systems of ANAS, Baku, Azerbaijan

Abstract. Thyroid disease is quite common in the world. In this respect, it is very important that the diagnosis of this disease is fast and accurate. It is possible and a good option to detect this disease with machine learning techniques, which is usually diagnosed with various laboratory tests and imaging tests. There are two types of thyroid disease (hyperthyroid and hypothyroid) in the target class, as well as healthy patients, on the dataset containing a total of 7000 lines of analysis and information about the people undergoing the tests. In this multi-classification problem, it is aimed to determine whether the patients are hyperthyroid, hypothyroid or healthy. Since there were unbalanced distributions on the dataset and overfitting occurred, various operations were required. These included feature extraction by correlation, undersampling and oversampling. In addition, the parameters affecting the result were found and the unimportant ones were removed. After these processes, the algorithms ran on these datasets and it was seen that the best result was feature extraction with correlation. Support vector machines, artificial neural networks, logistic regression, k-Nearest Neighbors and decision trees (random forest algorithm) were used for prediction. The most successful among them were artificial neural networks, k-nearest neighbors and support vector machines.

Keywords: Thyroid disease, multiclassification, neural network, random forest, support vector machines, k-Nearest Neighbors, feature extraction, oversampling, undersampling.
AMS Subject Classification: 68T01.
Corresponding author: Fidan, Nuriyeva, Dokuz Eylul University, Department of Computer Science, Izmir, Turkey, e-mail: *nuriyevafidan@gmail.com*
Received: 22 March 2022; Revised: 14 June 2022; Accepted: 2 July 2022; Published: 6 September 2022.

1 Introduction

In this paper we are going to discuss different machine learning algorithms for detecting thyroid diseases. Thyroid disease is widely common in the world, especially on women. It is estimated that approximately worldwide about 200 million people have a thyroid disorder (Jonklaas et al., 2014).

The thyroid gland is a tiny organ in the front of the neck that produces thyroid hormones. When your thyroid isn't functioning properly, it can have a negative impact on your entire body. Hyperthyroidism is a condition that occurs when your body produces too much thyroid hormone. Hypothyroidism is a condition in which your body produces too little thyroid hormone. Both illnesses are dangerous and require medical attention. This is why the false diagnoses might be severe, and need to be perfectly examined by experts.

With recent development on machine learning and data mining, it is not unusual for experts to get help by these algorithms and machines. Our aim is to help experts through their examination process with provided data. Hence, we need to create models for specific goal and make the data meaningful for them. Since it is matter of health, sensitivity of the results is very

important. We aim to predict whether the patient has thyroid disease or not, therefore we value every little development on our outputs.

The purpose of this study is also comparing the algorithms' accuracy while predicting the class of diagnose. In order to increase the efficiency of the corresponding classifier, data preparation and reducing features will be important, also we aimed that showing the most essential part of creating models, preparing the data, for the best result. Accordingly, we've implemented feature selection with correlation, mutual information, and oversampling/undersampling over data to tune our model for the better results.

All algorithms and prediction models has been implemented with Python. The reason behind this is Python being one of the most used and wanted programming language for machine learning. With its packages such as Numpy, Sci-kit Learn, Pandas and Keras makes it easier to deploy a model and tune in however you need.

2 Related Work

Generally, diagnosis of diseases and cancers are very popular in machine learning, because of potential data amount in healthcare makes it suitable for prediction algorithms. There are several important researches about this topic. One of the most related work is Korhan (2016). In this work, they have used the same data within this work. However, they approached the problem only with Support Vector Machines (SVM). They have prepared the data with feature selection and optimization. The results are given with their accuracy, sensitivity and specifity.

The other work on thyroid diagnosis is Adak & Yumusak (2016). They've used Neural Networks with PSO (Particle Swarm Optimization) and trained neural networks with it. Also, results are optimized with genetic algorithms. For the results, they've compared them with back propagation method. Again, only one approach used in this work for the diagnosis, which was Artificial Neural Network.

Another paper who used same dataset within this paper, they've managed to used Naive Bayes, Decision Trees, Neural Networks and Radial Basis Function Network with the help of a KNIME software. They also compared the results adding and removing specific attributes. And concluded that the best approach was Decision Trees among the others (Ionita & Ionia, 2016).

Decision Tree used for diagnosis Thyoid Disease, is Al-muwaffaq & Bozkus (2016). PCA (Principle Component Analysis) used for feature selection, which be used in this work too. They've also developed an app for real life use cases, taking inputs of new patients as parameters. The accuracy is very high with their model.

In this paper they have compared three algorithms, logistic regression, decision trees and k-Nearest Neighbors (k-NN). They've taken the data from the Graven Institute in Australia. The dataset contains 5 attributes and 215 instances. They used information gaining for decision tree and evaluate the best model among many. For k-NN, they've used Euclidean distance as they evaluate the model. As result, they've achieved 81% accuracy for logistic regression, 87% for decision tree and 96% for k-NN algorithm. In general, the paper describes in detail the data preparation, training, and testing, as well as a step-by-step discussion of each of the strategies employed and a comparison of the methods' accuracy in prediction Chaubey et al. (2021).

Another paper Kousarrizi et al. (2012), which gathered their dataset from UCI machine learning repository with 215 instances and 5 attributes. Addition to this dataset, they worked on a real dataset gathered by the Intelligent System Laboratory of K.N.Toosi University of Technology from Imam Khomeini hospital. Pattern recognition was used to extract or choose a feature set for usage in the pre-processing stage. For feature selection, it's been used Sequential forward selection and sequential backward selection. Diagnosis has been predicted by Support Vector Machines with 3- and 10-fold cross validation with the accuracy of 98%.

3 Materials and Methods

3.1 Dataset

The data set used for this research downloaded from university of California of Irvin (UCI) repository site. It consists of 3772 training instances, 3428 testing instances and 22 attributes with total 7200 instances and 3 classes. Dataset describes the main characteristics of the thyroid data set and its attributes. Each measurement consists of 21 values – 15 binary and 6 are continuous. Each of the measurement vectors is assigned one of three classes, which correspond to hyperthyroidism, hypothyroidism or normal thyroid function.

Feature information of dataset is given below.

Table 1: Feature information of dataset

Attribute Name	Value Type	Clarification
Age	number	1, 10, 20, 50, ...
Sex	1, 0	1=m, 0=f
On thyroxine	1, 0	1=yes, 0=no
Query on thyroxine	1, 0	1=yes, 0=no
On anti thyroid medication	1, 0	1=yes, 0=no
Sick	1, 0	1=yes, 0=no
Pregnant	1, 0	1=yes, 0=no
Thyroid Surgery	1, 0	1=yes, 0=no
IT3T Treatment	1, 0	1=yes, 0=no
Query hypothyroid	1, 0	1=yes, 0=no
Query hyperthyroid	1, 0	1=yes, 0=no
Lithium	1, 0	1=yes, 0=no
Goitre	1, 0	1=yes, 0=no
Tumor	1, 0	1=yes, 0=no
Hypopituitary	1, 0	1=yes, 0=no
Psych	1, 0	1=yes, 0=no
TSH	Analysis ratio	Numeric Value
T3	Analysis ratio	Numeric Value
TT4	Analysis ratio	Numeric Value
T4U	Analysis ratio	Numeric Value
FTI	Analysis ratio	Numeric Value
Output Class	1, 2, 3	1=normal, 2=hypothyroid, 3=hyperthyroid

3.2 Materials and Methodology

In this research, all implementations have been made with Python 3.6.5. The reason behind this is Python being very strong and trending programming language among machine learning. It is easy to process data and creating models. With packages you can tune your models with high detail.

Among the machine learning techniques, several classification algorithms have been used. Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN) and k-Nearest Neighbors algorithms are mostly used in classification problems, also developed for such classification problems.

One of the essential things in our research is the target has 3 different classes.That means binary classification algorithms would not fit into this problem. For this, we've implemented multi-class classification since we have 3 different target classes (non-sick, hyperthyroid, hypothyroid).

Since the dataset is very wide, it needs to be processed to reduce redundancy and possible overfitting problems. Also, target class distribution is highly unbalanced. There are 6666 non-sick, 368 hypothyroid and 166 hyperthyroid instances.

3.3 Pre-Processing

The "curse of dimensionality", which refers to a number of issues that arise while examining data in a high-dimensional domain, has greatly increased the computational burden. This computational burden can also lead to false results by running algorithms or models. So, analyzing data in high-dimensional space requires good amount of time for pre-processing. Despite it is costly for both time and resource, it is a must for any high-dimensional data. Because that computational burden's cost is much higher than pre-processing's.

Data processing becomes significantly more difficult in high-dimensional space because data becomes sparser and a large number of samples are required to train models. One of the most successful tools for reducing dimensionality and addressing the problem mentioned above is feature selection. It has been widely used as an effective preprocessing tool in a variety of applications, including data mining, machine learning and pattern recognition.

A vast number of attributes and limited available samples are common characteristics of biological data learning. Filtering out insignificant features prior to model fitting, for example, by discarding the ones that are least associated with the result, is a typical method. Mutual information (MI) is a commonly used measure of association in this context (Guyon & Elisseeff, 2003). It has also been shown to favor variables with more categories. It is also closely related to the Gini index and it has been proven also to be biased in favor of variables with more categories (Achard et al., 2005).

Feature selection, in general, yields a feature subset, with the purpose of obtaining the most informative subset by selecting crucial traits and eliminating unnecessary aspects from the original feature set. During the selection process, candidate traits and target classes are invariant, but the relationship between them is not. It will be a constantly changing value as more characteristics are added to the subset. As a result, determining the relationships between candidate features, selected features, and categories during the selection process is difficult. Classifier independent Filter, classifier-dependent Wrapper, and Embedded techniques are three types of supervised algorithms that deal with diverse interactions between data and classifiers.

3.4 Scaling Data

Feature scaling is a technique for normalizing a set of independent variables or data components. It is also known as data normalization in data processing and is usually done during the data preprocessing step. This strategy is necessary for accurate prediction and outcomes. When one of the columns has a very high value in comparison to the others, the impact of that column will be significantly greater than the impact of the other low-valued columns. Not every dataset requires normalization. It is required only when features have different ranges.

Some examples of algorithms where feature scaling matters are,

1. k-Nearest Neighbors with a Euclidean distance measure

2. k-means

3. Logistic Regression, SVMs, Neural Networks

4. Tree based models are not distance based models and can handle varying ranges of features. Hence, Scaling is not required while modelling trees.

The data has been normalized by min-max normalization. One of the most popular methods of data normalization is min-max normalization. This technique converts its minimum value to 0, and maximum to 1. All between values transforming to decimal values between 0 and 1. In this way all data values are more balanced for the model.

3.5 Correlation Feature Selection

In this case, we've aimed to find the specific columns that has correlated with target or with each other. It is better to remove high correlated features within columns, because it causes overfit or underfit with low correlation. However, if specific column has high correlation with the target, it needs to be included in the training data, since it is the significant column effecting the target column. Correlation is a measure of the linear association between two or more variables. The higher the correlation, the more linearly associated the variables are. Correlation is often a useful property. If two variables are correlated, we can predict one from the other. Therefore, we generally look for features or predictor variables that are highly correlated with the target, particularly for linear machine learning models. However, if two predictor variables are highly correlated among themselves, they provide in essence, redundant information about the target because with just one of them we can make accurate predictions on the target. The second predictor does not add additional information. Therefore, to make good machine learning models, in general will look for variables that are highly correlated with the target yet uncorrelated among themselves. In other words, we want the predictors to be correlated with the target but the predictors should not be correlated among themselves. In fact, the center hypothesis of correlation feature selection is that a good set of features, this is, a good set of predictive variables, contains variables that are highly correlated with the class or the target but uncorrelated with each other. Correlated features do not affect classification accuracy per se. The problem arises when we have datasets with loads of features and in extreme cases more features than examples. This is commonly known as the cause of dimensionality. If 2 numerical features are correlated, then one doesn't add much additional information over the other feature. So, if the numbers of features are high, removing correlated features is a good approach to reduce the feature space without losing information (Wei et al., 2020). One thing though, is that correlated features do affect more than interpretability and in particular, in linear models. If two variables are correlated, the model will fit coefficients to both variables that somehow capture the correlation. Therefore, these will be misleading on the true importance of each of the individual features. And this is also true for ensemble tree models, if two features are correlated, tree methods will assign roughly the same importance to both but half of the importance they would assign if we had only one of the correlated features in the dataset. The data's correlation heat maps are below.

Overall, the columns we're going to use mainly for our model are TSH, T3, TT4, T4U, and FTI, along with categorical variables sex and On-thyroxine.

In Figure 3 below, there is the FTI and TSH values' correlation graph. There can be seen negative correlation and the line between two variables. As it has been mentioned before, the correlation between two variable is -0.5.

That means there is an inverse proportion between variables, when one is high the other one is lower.

3.6 Mutual Information

Mutual Information measures non-linear relationships between two random variables. It also shows how much knowledge can be gleaned from one random variable by viewing another random variable.

It is intertwined with the idea of entropy. This is due to the fact that it can also be referred to as the reduction of uncertainty of a random variable when another is known.

As a result, a high mutual information value denotes a significant reduction in uncertainty, whereas a low value denotes a minor reduction. If the mutual information is 0, the two random variables are considered independent.

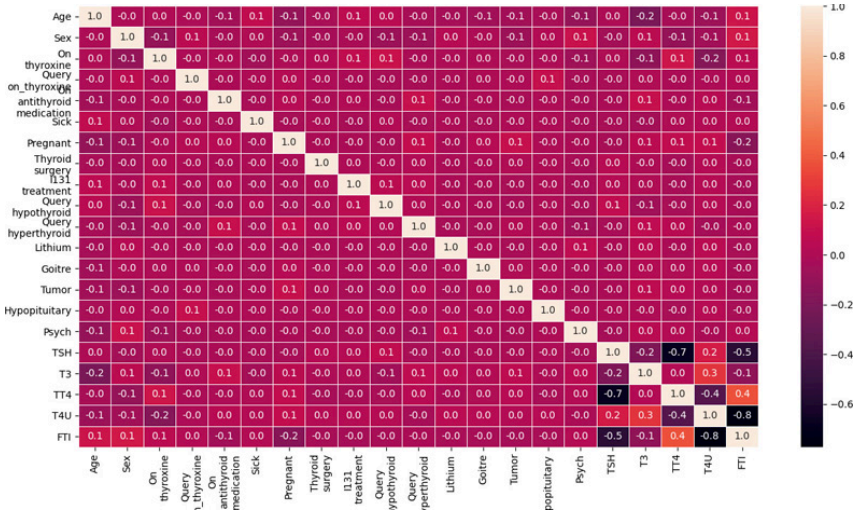

Figure 1: Correlation heatmap of dataset

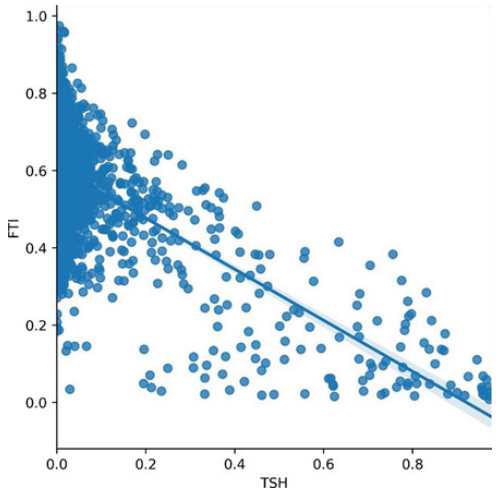

Figure 2: Correlation graph between FTI and TSH values

Mutual information is especially guides us to select important features. As you can see below (Figure 3), we've managed to find mutual information with target class. We've aimed the both hypothyroid and hyperthyroid as a comparing class.

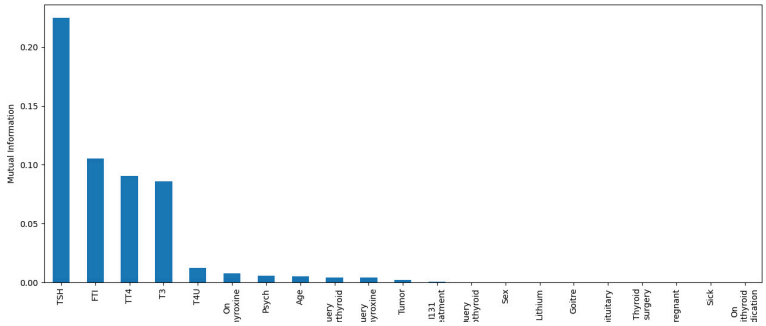

Figure 3: Mutual Information Graph

3.7 Feature Importance

One of the most extensively used data analysis approaches is feature importance, which is utilized for two main goals. One of them is to choose features for predictive models, removing the least predictive features to simplify and potentially increase the model's generality, and to apply business, medical, or other insights, such as customer-valued product characteristics or treatment contributions, to business, medical, or other insights. The likelihood of reaching that node to compute feature importance is weighted by the decrease in node impurity.

The node probability is calculated by dividing the number of samples that reach the node by the total number of samples (Breiman et al., 2017).

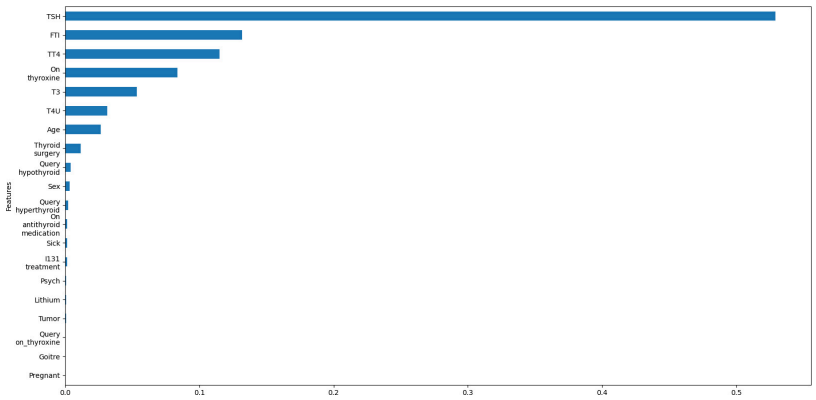

Figure 4: Feature Importance Order

In our data, we've found the most characteristically important features with random classifier technique. Results are also similar with correlated features that we've discovered. Our data with most important features as above (Figure 4).

Importance of features above: Most important features by random classifier tree are TSH, FTI, TT4, On-thyroxine, T3 and T4U. Again, these are the most correlated features.

In machine learning, outliers are as important as correlated features. In this case, we can remove those outliers as well as we use featured variables. According to our feature importance list, we can remove the least significant values. These are the values that close to 0 feature importance or the least correlated features.

According to these outputs, we can remove the most unrelated field for our target class to maintain the accuracy for our algorithms. In this case, we've removed Goitre, Pregnant, Query On-Thyroxine fields.

3.8 Over and Under Sampling

In machine learning, predicted values depends on model's ability to capture characteristics of different classes. In some cases, data has been skewed towards positive or negative class. If this skewness has not been normalized or optimized, it may result to overfitting or underfitting while predicting classes. The issues that come from data imbalance are handled by two resampling procedures, namely oversampling and undersampling, from a data-centric approach. Both resampling strategies add data preprocessing expenses to the equation, which might be overpowering in the case of very large-scale training data (Dey et al., 2001).

Practically, this process is taking skewed part and optimize it for the other values', means that number of values in that target class equalized with each other.

Oversampling does not result in the loss of information because all samples from the minority and majority classes are kept. However, because to the higher number of training examples, oversampling generates more training time throughout the learning process. Furthermore, using

a complicated oversampling strategy results in substantial computational expenses during data preprocessing.

Undersampling has been presented as a useful way to improve a classifier's sensitivity. However, this strategy may result in the omission of potentially useful data that is critical to the learning process (Ertekin, 2013).

In our data most target values consist of one class, which it corresponds as the "normal". Other two classes are 2 and 3, hypo and hyper thyroid respectfully.

Here is the data before and after sampling.

Table 2: Compare of Sampled Data

Class	**RawData**	**Undersampled Data**	**Oversampled Data**
1	6666	368	6666
2	368	350	5000
3	166	166	3000

Oversampling can be defined as adding more copies to the minority class. Oversampling can be a good choice when there is not many data to work with. However, tend to result in data that is non-smooth, has boundaries or small features. One way to avoid this pitfall is to combine undersampling and boosting. You might also want to manually resample or repair any holes in the data algorithmically.

4 Algorithms

4.1 Support Vector Machine

SVM is a widely used in classification of high-dimensional data. It is applied by training on data. Any method used in convex optimization can be chosen for training. Basically, it allows separating a data set that cannot be linearly separated in low dimensions by moving it to a higher dimension with the help of a plane.

Support vector machines (SVMs) were created with binary classification in mind. It's still a work in progress to figure out how to make it work for multiclass classification. Numerous methods have been developed, in which a multiclass classifier is often built by integrating several binary classifiers (Hsu & Lin, 2002). The purpose is to transfer data points to a high-dimensional space so that classes can be separated mutually linearly. This is known as a One-to-One technique, in which the multiclass problem is broken down into several binary classification problems. Each pair of classes has a binary classifier (Mircia et al., 2010). Another approach one can use is One-to-Rest. The breakdown is set to a binary classifier per class in that manner Chen et al. (2012).

We've implemented one-to-one approach with Python in our research. 10 folds been used in SVM algorithm with 30% training and 70% test data separation.

4.2 Multinomial Logistic Regression

Multinomial Logistic Regression is a subset of logistic regression that can handle multiclass classification problems. To support multi-class classification issues natively, this process requires updating the loss function to cross-entropy loss and to a multinomial probability distribution, predict the probability distribution (Sultana & Jilani, 2018). This approach is using in different fields such as image classification, mail classification etc.

When modifying model weights during training, cross-entropy loss with one- vs-all method is used. The objective is to minimize the loss; the lower the loss, the better the model. A perfect

model has zero cross-entropy loss. Most of the time, it's used for multi-class and multi-label classifications.

In this paper, we used one-vs-rest scheme for multi-class classification and cross-entropy loss.The number of bits necessary to represent or transmit the average event from one distribution compared to another is calculated using cross entropy, which is based on the entropy theory.

4.3 Random Forest

Random Forest is a decision tree algorithm that uses a divide-and-conquer strategy to construct-decision trees from a randomly partitioned dataset. A group of decision tree classifiers is referred to as the forest. An attribute selection indicator such as information gain, gain ratio, or Gini index is used to generate individual decision trees for each attribute. Each tree is built using a different random sample. The most popular class is chosen as the final result when each tree votes on a categorization challenge.

Each forest tree votes with a single unit, allocating each input to the most likelyclass label. It's a fast, noise-resistant algorithm with a successful ensemble for detecting non-linear patterns in data. It is capable of handling both numerical and category data with ease (Titapiccolo et al., 2013). One of the most significant advantages of Random Forest is that it does not suffer from overfitting, even as the forest grows larger.

In the case of regression, the final result is the average of all tree outputs, and it is both simpler and more powerful than other non-linear classification techniques. Data has been trained with 20 trees, entropy as the loss function method. It's been tried with 5, 10 and 20 folds.

4.4 Artificial Neural Network

Artificial neural networks are systems that can learn events through examples, derive and discover new information using the information they have learned, and thus respond to the effects of the environment similar to those of humans with the knowledge, experience and experiences they have gained (Haciefendioglu, 2012).

Two hidden layers and an output layer make up the artificial neural network model. The neurons in the buried layers are 50 and 20 respectively. Rectified Linear Unit is selected for activation in these layers' functions. Since the problem is multiclass classification, the output layer consists of 3 neurons and the activation function is chosen as softmax. The dataset is trained for 200 epochs.

4.5 k-Nearest Neighbors

The k-nearest neighbors (k-NN) algorithm is a data categorization method that uses the data points closest to it to estimate the likelihood that a data point belongs to one of two categories. To solve classification and regression problems, the supervised machine learning algorithm k-nearest neighbor is applied. The k-NN algorithm assigns a class to a new observation and is one of the most extensively used classification algorithms in a range of disciplines (Guney & Atasoy, 2003).

There are four ways to calculate the distance measure between the data point and its nearest neighbor: Hamming distance, Manhattan distance, Euclidean distance, and Minkowski distance. Out of the three, Euclidean distance is the most commonly used distance function or metric. Which is used in this research too.

k-NN doesn't work well if there are too many features. That is why, feature selection is a must for this problem. Because the dataset itself has high dimension. For this algorithm dataset reduced to 5 columns which were the most correlated with the target class.

k-NN has been applied with $k = 2$ to $k = 8$, with step of 1. The result graph has been shown below (Figure 5). As you can see below since k increases the accuracy has dropped as well.

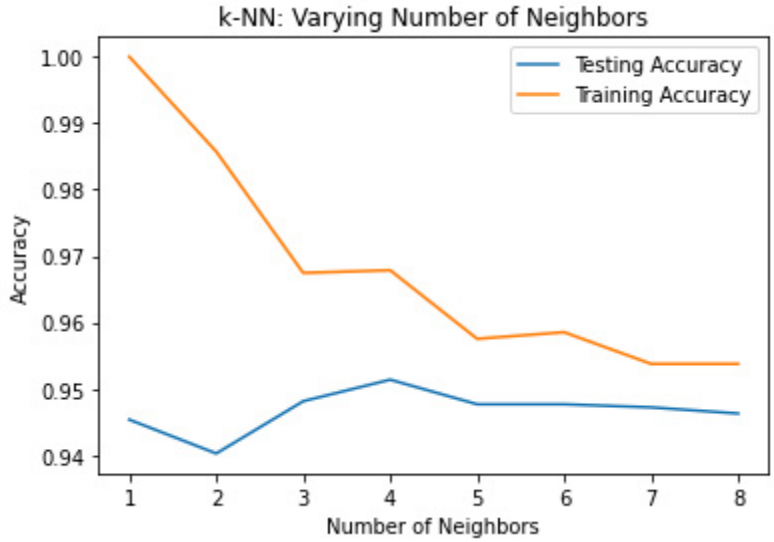

Figure 5: Accuracy graph of k - NN

Graph also explains that the best k value is k = 4. In this way we found the optimal k value for our model.

5 Results and Metrics

With the best result of accuracy is the ANN, SVM and k-Nearest Neighbors algorithms. k-Nearest Neighbors and random forest algorithms highly effected by the undersampled data in a negative way. Especially for random forest tree, there must have been more suitable variables since the categorical variables are imbalanced highly. The fact that these algorithms are widely used for multi-class classification, there has been many tune and parameters in these algorithms. Finding the best one depends on experience and tests on various data.

The "best" choice really depends on the underlying problem and what you are willing to tolerate more, false positives (FP) or false negatives (FN). If it less concerned about producing FN predictions and want to minimize FP predictions, it would probably focus more on achieving high precision scores (compare with the formula). For the opposite case, it would focus more on recall respectively. If it cannot afford such a trade- off between these two, the F1-score is a good measure because it is the harmonic mean of precision and recall.

Accuracy calculated by sum of True Positives and True Negatives divided by sum of True Positive, False Positive, False Negative and True Negative values 1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Results table with algorithms we implemented is below (Table 3).

6 Conclusion

For most of healthcare diagnosis problem, it is important that the data is reliable and has enough dimension for the needed model. Thyroid disease can be diagnosed with specific hormone tests, however being a multiclass classification problem making this diagnose with machine learning algorithms tricky since it is challenging to find which algorithms or techniques will be used. For this purpose, this paper had compared the techniques and algorithms for multi-class classification problem.

Table 3: Result Table by Accuracy

Algorithm	With over-sampling	With under-sampling	With feature extraction
SVM	% 89	% 70	% 94
Multinomial Logistic Regression	% 85	% 68	% 88
ANN	% 90	% 86	% 98
Random Forest	% 62	% 68	% 79
K-nearest neighbors	% 71	% 85	% 95

The most challenging part was lack of dimension in dataset. It is a highly imbalanced and not very big. This made the algorithms overfit easily hence some techniques such as feature extraction, over and under sampling has been applied to prevent overfitting.

To overcome that challenge, we aimed to select the most possible contribute to our prediction class. We've achieved this by finding correlation between variables, extracting mutual information between target class and our variables, and use over/under sampling to gather maximum efficiency. We've been tried these in multiple ways such as with or without over/undersampling or by not normalize the values. All our predictions calculated with the best approach by these methods.

Results are showing that when there is an imbalanced data, undersampling is not always a good technique with such small dataset. For same reason, it was not very effective to apply oversampling as well. Since training and test datasets are imbalanced, the best way to prevent this was using oversampling and feature importance together, thus it would reduce the overfit possibility. For preventing this we excluded the non-important features as well as detect the important and not important features. Specific algorithms needed scaling since their structure needs such dataset. Overall, the best way to diagnose thyroid disease was artificial neural networks, however SVM and k-Nearest Neighbors performed well enough with over 95% accuracy. These algorithms are very effective on multi-class classification if they tuned well.

References

Achard, S., Pham, D.T., & Jutten, C. (2005). Criteria based on mutual information minimization for blind source separation in post nonlinear mixtures. *Signal Processing*, 85, 965-974.

Adak, M.F., Yumusak, N. (2016). Diagnosis of Thyroid by Hybrid Parti- cle Swarm Optimization with Artificial Neural Network. *4th International Symposium on Innovative Technologies in Engineering and Science*, Antalya, Turkey, 25-30.

Al-muwaffaq, I., Bozkus, Z. (2016). MLTDD: Use of Machine Learning Tech- niques for Diagnosis of Thyroid Gland Disorder. *The Fourth International Conference on Database and Data Mining*, Dubai, UAE, 10-78.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (2017). *Classification and Regression Trees* (1st Edition). Boca Raton: Routledge.

Chaubey, G., Bisen, D., Arjaria, S., & Yadav, V. (2021). Thyroid disease prediction using machine learning approaches. *National Academy Science Letters, 44*(3), 233-238.

Chen, H.L., Yang, B., Wang, G., Liu, J., Chen, Y.D., & Liu, D.Y. (2012), A three-stage expert system based on support vector machines for thyroid disease diagnosis. *Journal of Medical Systems,* 36, 1953-1963.

Dey, T.K., Giesen, J., Goswami, S., Hudson, J., Wenger, R., & Zhao, W. (2001). Undersampling and oversampling in sample-based shape modeling. In *Proceedings Visualization*, Ohio: IEEE, 83-545.

Ertekin, Ş. (2013). Adaptive oversampling for imbalanced data classification. In *Information Sciences and Systems* 2013 (pp. 261-269). Springer, Cham.

Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selec- tion. Journal Of Machine Learning Research, 3, 1157-1182.

Guney, S., & Atasoy, A. (2012). Multiclass classification of n-butanol concen- trations with k-nearest neighbor algorithm and support vector machine in an electronic nose. *Sensors and Actuators B: Chemical*, 166- 167, 721725. doi:10.1016/j.snb.2012.03.047

Haciefendioglu, Ş. (2012). Diagnosis of glaucoma with machine learning methods. Master's thesis, Selcuk University, Konya.

Hsu, C.W., Lin, C.J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions On Neural Networks*, 13, 415-425.

Ionita, I., Ionia, L. (2016). Applying Data Mining Techniques in Health- care. *Studies in Informatics and Control*, 25, 385-394.

Jonklaas, J., Bianco, A.C., Bauer, A.J., Burman, K.D., Cappola, A.R., Celi, F.S., ... & Sawka, A.M. (2014). Guidelines for the treatment of hypothyroidism: prepared by the american thyroid association task force on thyroid hormone replacement. *Thyroid, 24*(12), 1670-1751.

Kousarrizi, M.N., Seiti, F., & Teshnehlab, M. (2012). An experimental compar- ative study on thyroid disease diagnosis based on feature subset selection and classification. *International Journal of Electrical & Computer Sciences IJECS-IJENS*, 12, 13-20.

Korhan, N. (2016). *Diagnosis Of Thyroid Disease Via Support Vector Machines*. Doctoral dissertation, Istanbul Technical University, Istanbul.

Mircia, E., Imre, S., & Balaci, T. (2010). *Broad Research in Artificial Intelligence and Neuroscience*, (Vol. 1), Bacau: EduSoft.

Sultana, J., Jilani, A.K. (2018). Predicting Breast Cancer Using Logistic Regression And Multi-Class Classifiers In International *Journal of Engineering & Technology*, 7, 22- 26.

Titapiccolo, J.I., Ferrario, M., Cerutti, S., Barbieri, C., Mari, F., Gatti, E., & Signorini, M.G. (2013). Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. *Expert Systems with Applications, 40*(11), 4679-4686.

Wei, G., Zhao, J., Feng, Y., He, A., & Yu, J. (2020). A novel hybrid feature selection method based on dynamic feature importance. *Applied Soft Computing*, 93, 106337.