

THYROID PREDICTION USING MACHINE LEARNING

MUNUKUNTLA KALYANI¹ PROF.M. SADANANDAM²

¹ PG Student, Department of Computer Science and Engineering , Kakatiya University College of Engineering and Technology Warangal -506009, (TS).

²Professor, Department of Computer Science and Engineering ,Kakatiya University, Warangal -506009, (TS).

[¹kalyanimunukuntla46@gmail.com](mailto:kalyanimunukuntla46@gmail.com), [²msadanandam@kakatiya.ac.in](mailto:msadanandam@kakatiya.ac.in)

ABSTRACT

A challenging assumption in medical research, thyroid illness is a key source of formation in medical diagnostics and in the prediction of onset. One of the most vital organs in our body is the thyroid gland. Thyroid hormone releases are responsible for regulating metabolism. The two most prevalent thyroid disorders, hyperthyroidism and hypothyroidism, cause the secretion of thyroid hormones, which control the body's metabolic rate. Techniques for data purification were used to prepare the data so that analytics could be performed to determine the likelihood that a patient will develop thyroid disease. This work addresses the analysis and classification models that are being utilized in thyroid illness based on the data acquired from the dataset taken from the UCI machine learning repository. Machine learning is a critical component in the process of disease prediction. Ensuring a solid knowledge foundation that can get ingrained in and be used as a hybrid model for complicated learning activities, such prognostic and medical diagnostic tasks, is crucial. Additionally, we offered many machine-learning methods and diagnosis in this research for thyroid prevention. Support vector machines (SVM), K-NN, Decision Trees, and machine learning algorithms were used to forecast the estimated probability of a patient developing thyroid illness.

Index Terms: medical research, hyperthyroidism, machine-learning, thyroid illness.

INTRODUCTION

1. THYROID AND THYROID GLANDS:

The medical field makes advantage of advances in computational biology. It made it possible to gather patient data that had been saved in order to anticipate medical diseases. Various sophisticated prediction algorithms are accessible for the early detection and diagnosis of the illness. Despite the abundance of data sets in the medical information system, no intelligent algorithms exist that can quickly analyse diseases. In the process of creating a prediction model, machine learning algorithms become more important in resolving complicated and nonlinear issues. Any illness prediction model must prioritise the elements that may be chosen from many datasets and used as accurately as feasible for a classification in healthy patients. If not, a healthy patient could get needless therapy as a consequence of misclassification. Therefore, it is of the utmost

cardinality to forecast any illness in combination with thyroid disease. An endocrine gland located in the neck is the thyroid gland. It grows behind the Adam's apple in the thinner area of the human neck, where it helps secrete thyroid hormones, which in turn affect metabolism and protein synthesis. Thyroid hormones have a variety of roles in regulating the body's metabolism, including heart rate and rate of calorie burning. The thyroid gland's production of thyroid hormones aids in controlling the body's metabolism. Levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3) are the two active thyroid hormones that make up the thyroid glands. These hormones are essential for controlling body temperature throughout manufacture, thorough building, and oversight. In particular, the thyroid glands are normally responsible for producing two kinds of active hormones: thyroxin (T4) and triiodothyronine (T3). These hormones play a critical

role in controlling the amount of protein in the body, regulating body temperature, and carrying and transferring energy throughout the whole body. Iodine is thought to be the primary component of the thyroid glands for these two thyroid hormones, or T3 and T4, and it is disrupted in a few particular issues, some of which are quite common. Both an excess or insufficient amount of thyroid hormones may lead to hyperthyroidism and hypothyroidism. Underactive thyroid glands and hyperthyroidism have several causes. There are many different types of drugs. Ionising radiation, persistent thyroid discomfort, iodine deficit, and enzyme deficiencies that result in hypothyroidism are risks associated with thyroid surgery.

The four classifications of thyroid problems that the patients will be placed in are as follows:

i. Hypothyroid: Many bodily processes slow down when the thyroid gland produces insufficient amounts of hormones. Increased TSH, Decreased FT4, Weight Gain, Reduced Appetite, Slow Pulse and Fatigue, and Decreased Metabolism are among the symptoms.

ii. Hyperthyroidism: The thyroid gland produces more hormones than the body needs, which causes the body to operate more quickly. Decreased TSH, elevated FT4, weight loss, increased appetite, elevated pulse, sweating, and elevated metabolism are among the symptoms.

iii. Euthyroid: Patients with euthyroidism (a thyroid gland that functions normally) who experience transient sickness produce proportionally less hormones from their thyroid gland. FT4 and FT3 increases are among the symptoms.

iv. Euthyroid (negative): A thyroid gland that is functioning normally. Accurate calculations of the performance metric are made using the confusion matrix. The experimental data came from the Waikato Environment for Knowledge Analysis, or WEKA.

II.LITERATURE SURVEY

Using the backpropagation algorithm, Prerana, Parveen Sehgal, and Khushboo Taneja presented a method for identifying thyroid disease. Artificial Neural Networks, or ANNs, are trained using various training datasets and are built by identifying preliminary thyroid predictions via the backpropagation of error. MATLAB was used to generate the experimental outcomes. Using the support vector machines model, Ling Chen, Xue Li, Quan Z. Sheng, and Wen-Chih Peng suggested a three-stage expert system for the detection of thyroid illness. Ammulu & Venugopal gathered the

information from the UCI library and used the random forest technique to forecast the hypothyroid condition. The confusion matrix is used to determine the performance measure with precision. To categorize thyroid patients, Shankar and Lakshman suggested an optimum feature selection strategy using a multi-kernel support vector machine model. Using the Thyroid Disease Dataset, Irina Ionita suggested comparing many categorization models, including multi-layer perception, Naive Bayes, Radial Basis Function Network, and Decision Tree. To categorize thyroid patients, Kulkarni and Karunakar devised the MFHLSCNN (Modified Fuzzy Hyper Line Segment Clustering Neural Network) method. An intelligent method using artificial neural networks to identify thyroid disorders in expectant mothers was the concept put out by Vinod and Vimal. The Thyroid Disease Type Diagnostics (TDTD) framework was developed by Ahmed and Sumarni. It assists medical

professionals in the diagnosis of thyroid problems and the cleansing of medical data. To categorize thyroid patients, Pandey Tiwari, A. Shrivasa, and A. K. Sharma suggested an ensemble model with feature selection. Termites used neural networks to classify thyroid problems and offered a comparative analysis of many machine-learning algorithms.

Numerous machine learning techniques, such as random forest, decision tree, naïve Bayes, SVM, and ANN, are widely used in prognostic and recurring illness issues. Diseases linked to heart disease, diabetes, Parkinson's, hypertension, the Ebola virus (EV), forecasting and diagnosis, R-NA sequencing data analysis, and biomedical imaging allocation are just a few of the activities that are included. Nevertheless, developing a medical diagnosis and a machine learning-based illness prediction system is a challenging undertaking. To train the machine,

several fundamental difficulties must be addressed, such as data collection, compilation, and grouping.

Large biological data sets over a deep continuance are needed for the real activity difficulties, but they are almost nonexistent.

A technique for early thyroid illness identification via a neural network's backpropagation algorithm. An inaccuracy that is being utilized for previous illness predictions is delicately established by ANN via backpropagation. The influence of artificial neural networks (ANNs) is being trained using empirical data and testing methods that validate the usage of data that was not included in the training process. ANN shows an advanced neural network that can replace earlier illness forecasts and ends with excellent agreement with the preliminary data. The four classification models—Naive Bayes, Decision Tree, Multilayer Perceptron, and Radial Basis Function Network—

were examined and contrasted by the writers.

All of the categorization models exhibit astounding accuracy, as shown by the conclusion. The Decision Tree model outperforms the other approaches to categorization. In this work, 29 dataset attributes—known as Chi-Square features—are mandated and enforced as part of a feature selection technique. The datasets are filtered by applying unsupervised coated filters to the attributes to convert continuous values into nominal values, thereby reducing the 29 attributes to 10.

III. EXISTING SYSTEM

Much effort has been done in the last few years to identify the many thyroid disorders. Numerous writers have used a range of data mining methods. It has been shown by the authors that they have a sufficient method and level of confidence to identify thyroid-related disorders via their work, which makes use of several datasets and algorithms connected to future work that will be necessary to achieve more efficient and superior outcomes. The purpose of the study

is to explain different data mining procedures and statistical features that have gained popularity in recent years for the certain interpretation of thyroid illnesses by different authors in order to achieve different possibilities and for different approaches.

DISADVANTAGES OF EXISTING SYSTEM:

The results are not accurate with naive bayes algorithm with WEKA Tool.

IV PROPOSED SYSTEM:

The information mining task of identifying a function from named training data is known as supervised learning. An arrangement of prepared drawings made up the teaching materials. Every case in controlled adaptation consists of two parts: the intended output value (also known as the supervisory flag) and an information input object (usually referred to as a vector). An indirect function that may be used to map fresh pictures is generated by a supervised learning computation that examines

the training data [17]. A perfect enhancement will consider the computation required to determine the class names for situations that have not yet been encountered. This necessitates using computation to compile training data and concealed conditions in a way that makes "sense."

ADVANTAGES OF PROPOSED SYSTEM:

- Time Consumption is less.
- More accurate results are to be seen.

V. SYSTEM DESIGN

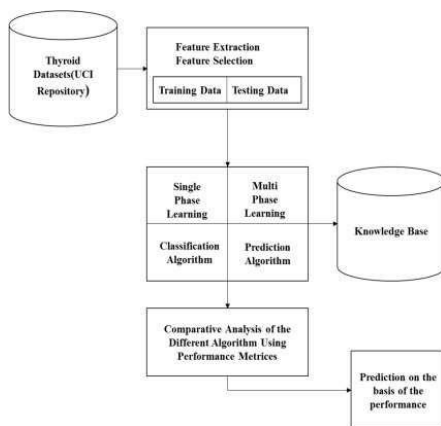


Fig1: Architecture of system.

i).Data Set Description:

The data taken end from

UCI repository undergoes preprocessing where missing values and not number constraints are checked using the masking method. If the missing value or Not a Number (NaN) values are present it is replaced by the mean value of the column.

By analyzing the above research work it is found that frequently used medical attributes to perform experimental work for the diagnosis of thyroid diseases are given below in below table no.1. Among these attributes almost every researcher has selected attributes to perform work for thyroid disease diagnosis.

Table 1:- Dataset Attributes

Attributes	Description
Age	In years
Sex	Male or female
TSH	Thyroid-Stimulating Hormone
T3	Triiodothyronine
TBG	Thyroid binding globulin
T4U	Thyroxin utilization rate
TT4	Total Thyroxin
FTI	Free Thyroxin Index

ii).Feature Extraction

Feature extraction is the process of converting the original data into a dataset that has a minimal number of variables, containing only discriminatory information. It reduces the amount of input data by distilling its representative descriptive attributes. Principal Component Analysis (PCA) for Feature Extraction: Principal Component analysis or PCA is a procedure which utilizes a certain number of transformation procedures to transform a dataset of closely related variables into a set of variables that are uncorrelated known as principal components. The data is transformed in a way such that the first principal component has the greatest variance which implies that it accounts for the

maximum amount of variability in the data. PCA is utilized as a tool for data analysis and for making models for prediction. It reveals the internal details of the data and gives an explanation for the variance in the data. The total variance is the sum of variances of all principal components. The fraction of variance explained by PCA is the ratio between that principal component and the total variance. So the variances of all principal components are divided by the total variance.

iii). Support Vector Machines

SVM (Support vector machine) is a classifier which works by separating classes through a hyper plane. The input to the algorithm is a set of labeled training data (supervised learning) and the output is a graph separating new instances of data into the classes through an optimal hyper plane. The hyper plane is basically a line separating a plane into 2 parts, each class lies on either side of the line. It can be utilized for both regression and classification problems, although it is mostly used for classification problems. Every data point is plotted in an dimensional space , 'n'

is the total no. of features and the value of each feature is the value of that particular coordinate on the graph Support vectors are essentially the coordinators of individual observations and a Support Vector Machine is the model that best separates the 2 classes of support vectors machine.

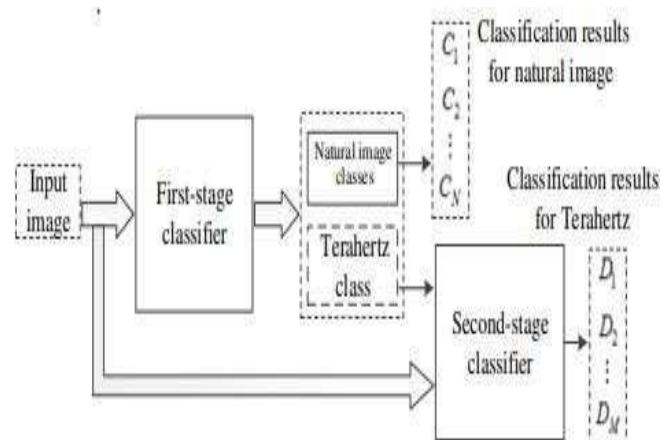
a). Advantages of SVM

- It performs well with a clear margin of separation.
 - Converts low dimensional spaces to high dimensional spaces.
 - It uses memory efficiently.
- ii. Disadvantages of SVM
- Requires higher training time hence does not work on larger datasets.
 - Doesn't perform well with overlapping target classes. The accuracy of the SVM technique improved significantly from 63.9% to 92.92% after applying feature selection. The SVM technique gives the highest accuracy (92.92%).

VI. MODULE DESCRIPTION:

- ❖ Step 1: Define the objective of the Problem Statement At this step, we must understand what exactly needs

to be predicted. In our case, the

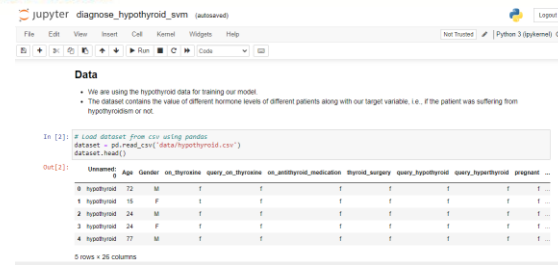


objective is to predict the possibility of rain by studying weather conditions. At this stage, it is also essential to take mental notes on what kind of data can be used to solve this problem or the type of approach you must follow to get to the solution.

- ❖ Step 2: Data Gathering Once you know the types of data that is required, you must understand how you can derive this data. Data collection can be done manually or by web scraping. However, if you're a beginner and you're just looking to learn Machine Learning you don't have to worry about getting the data. There are 1000s of data resources on the web, you can just download the data set and get going.

- ❖ . Step 3: Data Preparation The data you collected is almost never in the right format. You will encounter a lot of inconsistencies in the data set such as missing values, redundant variables, duplicate values, etc. Removing such inconsistencies is very essential because they might lead to wrongful computations and predictions. Therefore, at this stage, you scan the data set for any inconsistencies and you fix them then and there.
- ❖ Step 4: Exploratory Data Analysis Grab your detective glasses because this stage is all about diving deep into data and finding all the hidden data mysteries. EDA or Exploratory Data Analysis is the brainstorming stage of Machine Learning. Data Exploration involves understanding the patterns and trends in the data. At this stage, all the useful insights are drawn and correlations between the variables are understood. For example, in the case of predicting rainfall, we know that there is a strong possibility of rain if the temperature has fallen low. Such correlations must be understood and mapped at this stage.
- ❖ Step 5: Building a Machine Learning Model All the insights and patterns derived during Data Exploration are used to build the Machine Learning Model. This stage always begins by splitting the data set into two parts, training data, and testing data. The training data will be used to build and analyze the model. The logic of the model is based on the Machine Learning Algorithm that is being implemented. Choosing the right algorithm depends on the type of problem you're trying to solve, the data set and the level of complexity of the problem. In the upcoming sections, we will discuss the different types of problems that can be solved by using Machine Learning.
- ❖ Step 6: Model Evaluation & Optimization After building a model by using the training data set, it is finally time to put the model to a test. The testing data set is used to check the efficiency of the model and how accurately it can predict the outcome. Once the accuracy is calculated, any further improvements in the model

can be implemented at this stage. Methods like parameter tuning and cross-validation can be used to improve the performance of the model.



Data

- We are using the hypothyroid data for training our model.
- The dataset contains the value of different hormone levels of different patients along with our target variable, i.e. if the patient was suffering from hypothyroidism or not.

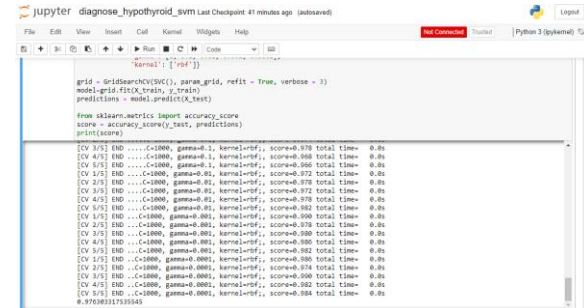
```
In [2]: # Load dataset from csv using pandas
dataset = pd.read_csv('data/hypothyroid.csv')
dataset.head()
```

```
Out[2]:
```

Unnamed: 0	Age	Gender	on_thyroxine	query_on_thyroxine	on_antithyroid_medication	thyroxine_surgery	query_hypothyroid	query_hypothyroid	pregnant
0	hypothyroid	72	M	f	f	f	f	f	f
1	hypothyroid	16	F	f	f	f	f	f	f
2	hypothyroid	24	M	f	f	f	f	f	f
3	hypothyroid	24	F	f	f	f	f	f	f
4	hypothyroid	77	M	f	f	f	f	f	f

0 rows x 10 columns

3.Model Generation screen

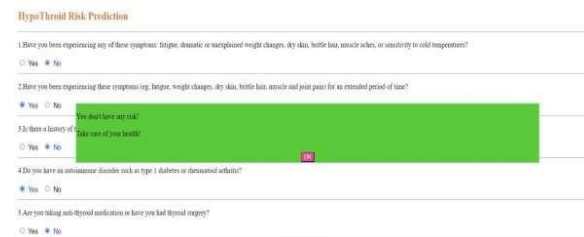


```
grid = GridSearchCV(model, param_grid, refit = True, verbose = 3)
model = grid.fit(X_train, y_train)
predictions = model.predict(X_test)

from sklearn.metrics import accuracy_score
score = accuracy_score(y_test, predictions)
print(score)
```

```
CV 3/5 END ... C=1000, gamma=0.1, kernel=rbf, score=0.570 total time= 0.0s
CV 4/5 END ... C=1000, gamma=0.1, kernel=rbf, score=0.568 total time= 0.0s
CV 5/5 END ... C=1000, gamma=0.1, kernel=rbf, score=0.566 total time= 0.0s
CV 1/5 END ... C=1000, gamma=0.01, kernel=rbf, score=0.572 total time= 0.0s
CV 2/5 END ... C=1000, gamma=0.01, kernel=rbf, score=0.570 total time= 0.0s
CV 3/5 END ... C=1000, gamma=0.01, kernel=rbf, score=0.572 total time= 0.0s
CV 4/5 END ... C=1000, gamma=0.01, kernel=rbf, score=0.570 total time= 0.0s
CV 5/5 END ... C=1000, gamma=0.01, kernel=rbf, score=0.562 total time= 0.0s
CV 1/5 END ... C=1000, gamma=0.001, kernel=rbf, score=0.580 total time= 0.0s
CV 2/5 END ... C=1000, gamma=0.001, kernel=rbf, score=0.580 total time= 0.0s
CV 3/5 END ... C=1000, gamma=0.001, kernel=rbf, score=0.582 total time= 0.0s
CV 4/5 END ... C=1000, gamma=0.001, kernel=rbf, score=0.580 total time= 0.0s
CV 5/5 END ... C=1000, gamma=0.001, kernel=rbf, score=0.580 total time= 0.0s
CV 1/5 END ... C=1000, gamma=0.0001, kernel=rbf, score=0.574 total time= 0.0s
CV 2/5 END ... C=1000, gamma=0.0001, kernel=rbf, score=0.580 total time= 0.0s
CV 3/5 END ... C=1000, gamma=0.0001, kernel=rbf, score=0.582 total time= 0.0s
CV 4/5 END ... C=1000, gamma=0.0001, kernel=rbf, score=0.584 total time= 0.0s
CV 5/5 END ... C=1000, gamma=0.0001, kernel=rbf, score=0.584 total time= 0.0s
0.57080321731545
```

4.Thyroid Prediction Screen



Hypothyroid Risk Prediction

1. Have you been experiencing any of these symptoms: fatigue, dramatic or unexplained weight changes, dry skin, brittle hair, muscle aches, or sensitivity to cold temperatures?

☒ Yes ☐ No

2. Have you been experiencing these symptoms (eg. fatigue, weight changes, dry skin, brittle hair, muscle and joint pain) for an extended period of time?

☒ Yes ☐ No

3. Do you have a history of "take care of your health"?

☒ Yes ☐ No

4. Do you have an autoimmune disorder such as type 1 diabetes or rheumatoid arthritis?

☒ Yes ☐ No

5. Are you taking anti-thyroid medications or have you had thyroid surgery?

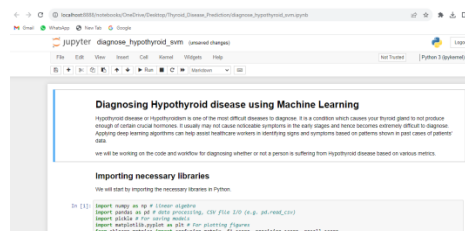
☐ Yes ☒ No

You don't have any risk!

Take care of your health!

VII. RESULT:

1.Import libraries screen



Diagnosing Hypothyroid disease using Machine Learning

Hypothyroid disease or hypothyroidism is one of the most difficult diseases to diagnose. It is a condition which causes your thyroid gland to not produce enough of certain crucial hormones. Usually, this can cause noticeable symptoms in the same stages and hence becomes extremely difficult to diagnose. Applying deep learning algorithms can help assist healthcare workers in identifying signs and symptoms based on patterns shown in past cases of patients with the disease.

We will be working on the code and workflow for diagnosing whether or not a person is suffering from hypothyroid disease based on various metrics.

Importing necessary libraries

We will start by importing the necessary libraries in Python.

```
In [1]: Import numpy, the top N cluster algorithm
Import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
Import sklearn # for machine learning
Import matplotlib.pyplot as plt # for plotting figures
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score
```

2.Data Preprocess screen

VIII. CONCLUSION

We have shown how machine learning may be used to diagnose hypothyroidism in this research. Thyroid illness identification is still a crucial yet challenging problem for statistical classification as well as clinical diagnosis. Extremely imbalanced groups and a multitude of associated patient traits are

used in the diagnosis, which creates a complex interaction between the input elements. However, the flexibility with which the Implemented Machine Learning Model may model intricate patterns of data for diagnosis yields encouraging results. When such issues arise, this information may be further expanded for a wide range of different illness diagnoses.

IX. FUTURE ENHANCEMENT

To enhance my writing, I might use image processing of ultrasonic thyroid scans to forecast cancer and thyroid nodules that are not visible in blood test results. Thyroid disease prediction may encompass all thyroid-related disorders by integrating the two data.

X. REFERENCES

- [1] L. Ozyilmaz and T. Yildirim, "Diagnosis of thyroid disease using artificial neural network methods," in: Proceedings of ICONIP'02 9 th international conference on neural information processing (Singapore: Orchid Country Club, 2002) pp. 2033–2036.
- [2] K. Polat, S. Sahan and S. Gunes, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis," Expert Systems with Applications, vol. 32, 2007, pp. 1141-1147.
- [3] F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab, "Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM," in 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009.
- [4] G. Zhang, L.V. Berardi, "An investigation of neural networks in thyroid function diagnosis," Health Care Management Science, 1998, pp. 29-37. Available: <http://www.endocrineweb.com/thyroid.html>, (Accessed: 7 August 2007).
- [5] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, New York, 2012.
- [6] Obermeyer Z, Emanuel EJ. Predicting the future— big data, machine learning, and clinical medicine. N Engl J Med. 2016; 375:12161219.
- [7] Breiman L. Statistical modelling : the two cultures. Stat Sci. 2001;16:199-231.
- [8] Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new

opportunities, familiar challenges. Clin
Epidemiology. 2017; 9:245- 250

[9] Ghahramani Z. Probabilistic machine
learning and artificial intelligence. Nature.
2015;521: 452-459.

[10] Azimi P, Mohammadi HR, Benzel
EC, Shahzadi S, Azhari S, Montazeri A.
Artificial neural networks in neurosurgery. J
Neural Neurosurg Psychiatry. 2015;
86:251-256.