Diagnosis of Sleep Apnea Using Machine Learning Utilizing Pre-Screening Questionnaires

Alena Mullee, Courtney Patterson, Gerakios Sam

[#] *Computer Science Department, Southern Connecticut State University*
501 Crescent St, New Haven, CT 06515 United States(US)
[1] mulleea1@southernct.edu
[2] pattersonc9@southernct.edu [3] samg1@southernct.edu

***Abstract*—Sleep apnea is a widespread sleeping disorder estimated to affect 20% of all American adults. The diagnosis of sleep apnea involves extensive sleep studies that measure various variables while a patient is sleeping. Sleep studies are recommended when patients show symptoms that are linked to sleep apnea. Various pre-screening questionnaires aid in this process. These studies usually are expensive, and many patients claim them invasive. This study analyzes the questionnaires and pre-screening questions using decision tree and neural network methods to the most effective questionnaires and questions. Machine learning models were trained using sets of data from different questionnaires and evaluated on their effectiveness. After multiple trials with different data sets, almost all questionnaires had similar results with very low standard deviations. The Zung Depression Scale Questionnaire boasted the highest conclusion values at 87% correctly classified and determined to be the best questionnaire.**

I. Introduction

Sleep apnea is a common sleeping disorder that can be diagnosed in anyone but is more commonly found in older men that are overweight [1]. A person with sleep apnea will stop breathing several times during their sleeping patterns and possibly hundreds of times during the night[2]. When left untreated, sleep apnea can lead to sleep deprivation and numerous issues associated with the condition, such as work-related accidents and motor vehicle collisions. In extreme cases, untreated sleep apnea causes serious health problems, including diabetes, hypertension, stroke, cardiomyopathy, heart failure, and heart attacks[1].

It is believed that one in fifteen adults in the United States has the disorder. An estimated 2% of children in the United States suffer from the disorder brought on by enlarged tonsils or adenoids[3] .A telephone poll done in 2005 by the National Sleep Foundation was analyzed to determine if a respondent was at risk for Obstructive sleep apnea. Of the 1,506 respondents, 26% met the berlin questionnaire criteria for high risk. Separately, 57% of the respondents were at high risk for obesity as well [4].

There are two types of sleep apnea: obstructive and central. Obstructive sleep apnea is the more common and occurs as “repetitive episodes of complete or partial upper airway blockage during sleep” [1]. This is known as an apneic episode. According to the Cleveland Clinic, the diaphragm and chest muscles work harder during an apneic episode as the pressure increases to open the airway. Breathing usually resumes with a loud gasp or body jerk. These episodes can interfere with sound sleep, reduce oxygen flow to vital organs, and cause heart rhythm irregularities [1]. Central sleep apnea is not as common but just as dangerous. A person suffering from central sleep apnea will find that the brain does not signal the muscles to breathe due to instability in the respiratory control center, thus making the problem related to the central nervous system [1]. This paper will focus on obstructive sleep apnea.

Sleep apnea symptoms often go unnoticed by the afflicted individual but will be realized by their partner or family. The most common symptoms of obstructive sleep apnea are snoring, restless sleep, night sweats, frequent nighttime urination, awakening sudden while gasping or choking, fatigue or sleepiness when awake, headaches, sexual dysfunction, cognitive degradation or dissonance, and dry mouth or sore throat upon waking up[1].

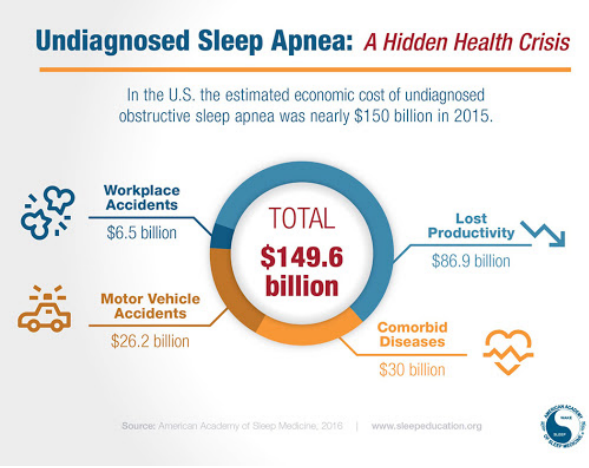

Fig. 1. Cost of Accidents Caused by Undiagnosed Sleep Apnea[5]

A. Motivation

An in-lab sleep study test can be costly. They usually range from 500 USD to more than 3000 USD. Health insurance may cover a portion of the cost, but sometimes the portion

insurance will pay is small and leave the patient to pay out of pocket for the rest [6]. Patients also report that testing itself is invasive and uncomfortable. The testing room is a large, monitored room, so the patient feels exposed, and sometimes the room is also cold. The patient is also required to be hooked up to various machines via sensory pads that make it hard for them to get comfortable. The pads' location requires them to maintain one position for the study, which might not be the position they are comfortable falling asleep. Patients report that the machines' sound can be loud, especially to patients that are unused to noise when they are trying to fall asleep, thus making it difficult to fall asleep. They are required to do this for 6-10 hours, making for a highly uncomfortable evening.

It may take over a year for the patient to see a doctor to have the study requested. The average wait time for a sleep study after that initial meeting is another six months[2], and a single test can take between seven and ten hours, though on average, it takes 8 hours [7]. It may also take up to a month for the patient to get the test results back as one test can generate up to 1000 pages of data for the sleep specialist to analyze[2] . Travel time is also an issue for the patient, as these studies are done in specialized clinics with a limited number of beds in them[2].

B. Goals

The purpose of this project is to more accurately predict the presence of sleep apnea in patients based on pre-screening questionnaires. Due to polysomnography testing's invasiveness and cost, predicting sleep disorders before testing will significantly help both patients and doctors. With the high time cost of the overnight polysomnography testing, a reliable model to predict a sleep disorder can decrease false test orders and decrease wait times for patients who need the tests. The trained models will also isolate the most critical variables in predicting the presence of sleep apnea, which can help create more accurate questionnaires in the future. These factors combined would make it easier for patients to be diagnosed with higher success rates and seen in clinic faster for their condition.

C. Impact

This project serves to have the most significant impact on patients and doctors but will aid researchers and insurance companies as well. Patients should see reduced wait times for testing, out-of-pocket spending, and overall stress around unnecessary testing. Doctors will see more accurate test results and reduced wait time for procedures and results to serve their patients better. The researchers that administer the tests will see the need for some unnecessary procedures removed, which will increase the time they can spend with each patient and allow them to generate more accurate testing results more quickly. Insurance companies will benefit from removing unnecessary expenditure on unneeded procedures and problems with undiagnosed sleep apnea.

II. Machine Learning

Machine learning is applying artificial intelligence, which automatically provides digital computerized systems to learn and self-improve from experience without explicit programming. It is the data analysis method that works by automating the analytical model building. Machine learning is the branch of artificial intelligence formed on the idea that the systems can automatically study and learn from the data, establish patterns, and then make appropriate decisions with minimum or no human intervention[8]. The entire process begins with the system making observations of the data to establish any possible patterns and then develop better decisions based on the discovery. The principal aim is to give the systems the powerful ability to automatically observe, learn and act accordingly without any human assistance.

The algorithms used in machine learning are divided into two classes: supervised and unsupervised. Supervised machine learning requires the programmer to use a training data set, explain and specify the input, and, consequently, the expected corresponding output. The algorithm used in learning can then compare its output with the intended output, determining any possible errors to make any necessary modifications in the model[9]. An unsupervised machine learning model is where there is no prior classification or labeling of the information used in training. Here, the system, not figuring out the appropriate output, analyses the given data and then draws inferences for describing the hidden structures in the unlabeled data[10]. Machine learning gives computerized systems the powerful ability to analyze massive quantities of data. It ensures the delivery of fast and accurate results to establish more good chances or dangerous risks. Combining cognitive machine learning and artificial intelligence could make the systems more effective and ensure quality and profitable results.

III. Decision Trees

Machine learning algorithms classify the data to understand and interpret it more objectively. The classification process involves two steps: the first step is learning, where a model is designed based on the and prediction steps. The second step is the prediction step, where the model predicts the response for the given data set. The simplest form of classification algorithms is the decision tree. Decision Trees are defined as supervised machine learning models where data is separated according to a specific set parameter. Two aspects describe a decision tree: the nodes and leaves. The nodes of any decision tree are the questions asked. For example, "What is the type of car?" and "What is Color?". The leaves are the outcomes or decisions determined after analysis, like "Mercedes" and "Blue." For regular operations in research, especially in decision analysis, decision trees are the more popular method. It helps in identifying the best strategy for reaching a specific goal.

A. Decision Tree Algorithms

Decision trees fall into two main types; the Classification trees, otherwise known as Yes/No types, and the Regression trees, whose outcome is a continuous variable, like the numbers 1,2,3. Various algorithms are used along with the construct decision trees. The most common and the best algorithm is the Iterative Dichotomiser 3 Algorithm. The Iterative Dichotomiser 3 algorithm constructs decision trees applying a 'greedy search' technique, top-down based, not giving space for any backtracks[11]. 'Greedy' algorithm affects decisions that seem to be the most appropriate at any given instance. According to Rokach and Maimon[12], the Iterative Dichotomiser 3 algorithm is executed in the following steps: it commences with a root node from an original set (S). It iterates through the least used attribute of the set S and then computes the Entropy (H) and the Information gain (IG) of the same attribute, executing this on each iteration. That which has the least Entropy or the most significant information gain is then selected. S then gets broken down by the chosen attribute to return a subset of the data. This process then continuously recurs every subset, sticking to only those previously selected attributes. According to Sheppard[11], Entropy of this sort is Shannon Entropy. It is defined as the measure of randomness or uncertainty in information processing. Information gain measures the Entropy's relative change. Other algorithms used include the C4.5, Classification and Regression Tree, Multivariate Adaptive Regression Splines, and the Chi-Square Automatic Interaction Detection.

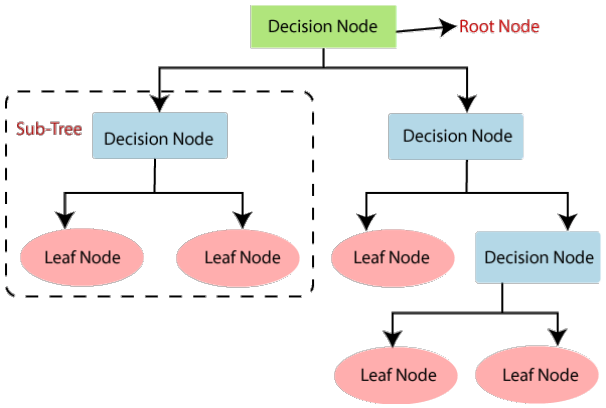

Fig. 2. Decision Tree Layout

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \quad (1)$$

IV. Artificial Neural Networks

Artificial intelligence is a quickly growing digital and computing branch aimed at creating intelligent electronic devices and computer systems. These systems can perform complex tasks that would traditionally require human intelligence, with multidisciplinary science approaches bringing change across various sectors of technology. Human and rational thinking and acting act as the fundamental approaches that can describe this highly advanced field. Neural networks are a branch of artificial intelligence. They have been built to imitate the human brain's behavior by allowing computer systems to analyze and recognize patterns and, with this, get to solve common problems in machine learning and deep learning. Also known as artificial neural networks, they are inspired by the human brain, trying to simulate how the biological neurons signal one another. They depend on training data to learn and have their accuracy improved over time[13].

Algorithm 1 C4.5 Algorithm Pseudo Code

procedure C4.5 ALGORITHM
  Check for base case
  For each feature f
  Find Information Gain 'G' by splitting based on f
  Assume f_{best} is the attribute with the best gain 'G'
  if f_{best} = found **then**
    Create Decision Node
    Re-cure on the sub-lists and add children nodes
    Repeat until all features are used
  else
    Stop (Best Tree Found)

Neural networks interpret sensory data through advanced machine perception, raw input clustering, and labeling. The patterns recognized by neural networks are primarily numerical and contained in vectors. All real-world data, including images, time series, images, text, or sound, have to be converted into numerals. Additionally, neural networks provide significant aid in clustering and classification. It is easiest to think of them as a layer on top of stored and regularly managed data. With neural networks, it is possible to group unlabeled data depending on their similarities, like inputs. The neural networks then classify this data when they get a dataset that has been appropriately labeled to train on[13]. They also have the power and ability to extract features passed to other algorithms for classification and clustering. Because of this reason, neural networks can be thought of as components of expansive machine learning applications which involve powerful algorithms for conducting classification, regression, and reinforcement learning.

A. Neural Networks Algorithm

Artificial Neural Network's algorithms are founded on a radial basis functionality, which can be effectively used for strategic courses. Being inspired by the human brains functioning, the artificial neural networks get trained on various situations and data, and they automatically adjust like the human brain. Within the Neural networks, there are input, output, and hidden layers. The main task of neural networks is to transform the input into desired output units. The flow of information through neural networks occurs in two ways. The first technique is the feedforward networks, where signal transmission is only in one direction without any loops, for example, towards the output layer. The feedforward method,

which is also extensively used in recognizing patterns with a single input and output layer, can have multiple or no hidden layers and has two standard methods: at the time of training and at the time of operation.

A supplement method for the feedforward mechanism, referred to as the multi-layer perceptron, is denoted as MLP and comprises three layers: the input, output, and hidden layers. They are mainly used for pattern recognition, classification, approximation, and prediction. The multi-layer perceptrons are developed for approximation of any continuous functions and can also solve linearly inseparable problems. On receipt of information, the input layer passes it to the hidden layer, the computational engine. Later, it is moved to the output layer that carries out classification and prediction[14].

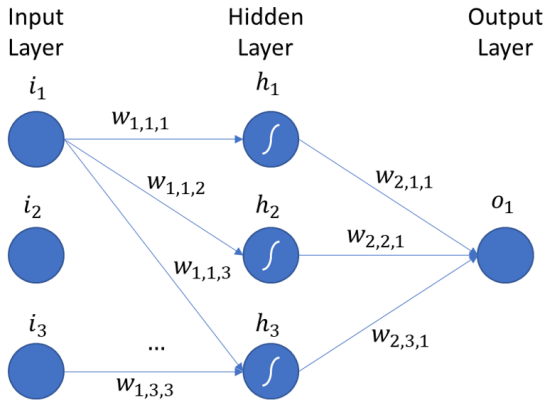

Fig. 3. Neural Network Architecture

The second method is the feedback method, in which interactive networks can use their internal state or memory to compute input sequences [14]. Signals can travel both ways with loops.

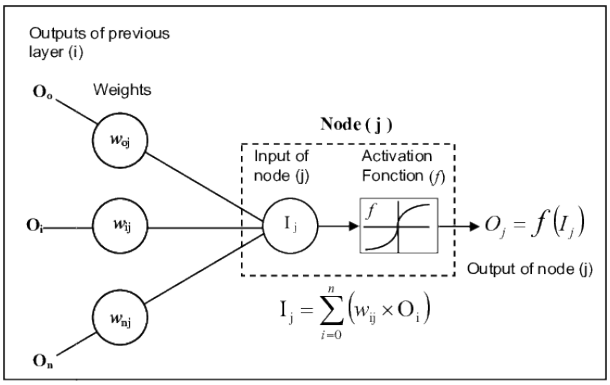

Fig. 4. Perceptron Model

The algorithms work on three main layers, which form the underlying basis for neural networks, with each layer having its responsibility. In the neural networks, neurons are used in processing data, discern patterns into objects that naked eyes cannot see. The layers will take a set of data as input, extract its features, and then transport it to the next layer. Since the networks are founded on a layered model, each layer acts as the input to the next layer. The purpose of this is to receive the last layer's job, process it, and move it to the next to come out with the required output. The hidden layers conduct all the execution processes[14].

V. Data Set

This project uses the Wisconsin Sleep Cohort (WSC) study data set. This study ran from 2000 to 2015 and had 1123 subjects in the age range of 27 to 85. This study was born when the clinical interest in sleep-disordered breathing and other sleep disorders rose in the 1990s, along with the need for more studies to examine those disorders' outcomes and burdens. The sampling frame for the study is the payroll files of Wisconsin State employees in the year 1988. This sample includes a wide range of demographics. All employees also had the same access to healthcare, which reduced any biases related to that factor. A survey was sent to this sampling frame which questioned sociodemographics, lifestyle, health habits, and sleep characteristics. Responders were categorized as high risk or low risk depending on the answers. Subjects were recruited from those responders with roughly 1.5:1 weighting of high: low risk [15].

The Wisconsin Sleep Cohort study focuses on the ongoing causes, consequences, and history of sleep disorders, focusing on sleep apnea. The data set was chosen for this project as it includes a significant amount of pre-screening data for each of the patients through various questionnaires.

The questionnaires this project will be focusing on will be the Zung Depression Scale, the Epworth Sleepiness Scale, and Sleep Disturbance. Sets of data from each separate questionnaire will each train a model, and the model's effectiveness at diagnosing sleep apnea off of the initial pre-screening data was evaluated. This information will more accurately prescribe further testing through sleep studies such as polysomnography if needed.

A. Zung Depression Scale

The Zung Depression Scale questionnaire is a self-assessing scale to assess depression concerning arousal response during sleep and changes with the disorder's treatment. This scale is used in this study since it is short and concise, and sleep disorders and depression can be closely linked [16]. This questionnaire has the subject rate their well-being on a scale of one to four on various questions such as when they feel the best during the day, how their sleep quality is, and their levels of restlessness.

Variable	Description
zung1_scored	I feel down-hearted, blue and sad.
zung2_scored	Morning is when I feel the best.
zung3_scored	I have crying spells or feel like it.
zung4_scored	I have trouble sleeping through the night.
zung5_scoredd	I eat as much as I used to.
zung6_scored	I enjoy looking at, talking to and being with attractive women/men.
zung7_scored	I notice that I am losing weight.
zung8_scored	I have trouble with constipation.
zung9_scored	My heart beats faster than usual.
zung10_scored	I get tired for no reason.
zung11_scored	My mind is as clear as it used to be.
zung12_scored	I find it easy to do the things I used to do.
zung13_scored	I am restless and can not keep still.
zung14_scored	I feel hopeful about the future.
zung15_scored	I am more irritable than usual.
zung16_scored	I find it easy to make decisions.
zung17_scored	I feel that I am useful and needed.
zung18_scored	My life is pretty full.
zung19_scored	I feel that others would be better off if I were dead.
zung20_scored	I still enjoy the things I used to do.
zung_index	Standardized Zung Depression Scale Total Score
zung_score	Zung Depression Scale Total Score

TABLE I
ZUNG DEPRESSION SCALE VARIABLES

B. Epworth Sleepiness Scale

The Epworth Sleepiness Scale is a self-administered questionnaire that measures the general levels of daytime sleepiness. The subjects will rate the chances that they will doze off or fall asleep when in eight different situations. These ratings help distinguish whether a patient may have a sleeping disorder such as sleep apnea, idiopathic hypersomnia, or narcolepsy when added together [17].

Variable	Description
ess	Total score
ep1	Chance of dozing while sitting and reading
ep2	Chance of dozing while watching TV
ep3	Chance of dozing while sitting, inactive in a public place
ep4	Chance of dozing as a passenger in a car for an hour without a break
ep5	Chance of dozing while lying down to rest in the afternoon
ep6	Chance of dozing while sitting and talking to someone
ep7	Chance of dozing while sitting quietly after lunch without alcohol
sleepiness	Excessive daytime sleepiness
ps_eds	Excessive daytime sleepiness

TABLE II
EPWORTH SLEEPINESS SCALE VARIABLES

C. Sleep Disturbance

The Sleep Disturbance Questionnaire has the subjects rate how often certain events will happen to them. Zero represents an event that never occurs. The frequency increases with each value up to four, representing an event occurring 16 to 30 times a month. The questionnaire focuses on events such as the subject having trouble falling asleep, waking up repeatedly through the night, and excessive daytime sleepiness. The symptoms this questionnaire focuses on can lead to various sleep disorder diagnoses, provided they occur often.

Variable	Description
eval_general	Self-reported satisfaction with sleep
ps_diff	Difficulty falling asleep
ps_backsleep	Difficulty falling back to sleep after waking up during the night
ps_wakerepeat	Wake up frequently during the night
ps_tooearly	Wake up too early
ps_notrested	Not rested regardless of sleep amount
ps_wakeup	Difficulty waking up in the morning
ps_nightmare	Frequency of having nightmares or disturbing dreams
anyinsomnia	Indicator for insomnia symptoms
ninsomnia	Count of frequent insomnia symptoms
ninsomniadays	Total days per month having any insomnia symptoms

TABLE III
SLEEP DISTURBANCE VARIABLES

VI. CLASSIFICATION SOFTWARE

The classification software used to process, train and test the models was WEKA 3, a data mining and machine learning software. Weka comes with many preprocessing tools referred to as filters. A few filters and processes built-in to the software were employed to handle the data and create more optimal results. Weka's built-in j48 decision tree algorithm was used to build the decision tree models. The essential feature of this algorithm is its use of pruning to delete unneeded tree branches when building the model. For the neural network models, the MultilayerPerceptron algorithm was used. This model uses neurons trained with the backpropagation learning algorithm. Weka possesses visualization capabilities, which are utilized to provide visualizations for the decision trees and neural network models. Weka also has an error visualizer that is used to see where any errors may have taken place in each model's classification [18].

A. Data Set Processing

The questionnaire variables for each separate questionnaire were parsed from the complete data set and separated into different files to prepare the data set for model training and testing. Each questionnaire file was saved as a .csv file for easy use with Weka. Each data set was then parsed through in order to delete any tuples containing empty data.

Due to the low number of positive sleep apnea patients in the study as seen in Table IV, a number of tuples were deleted to bring the positive and negative apnea counts to more even count. The updated positive and negative sleep apnea counts can be seen in Table V.

	Y	N
Zung Depression Scale	392	2170
Epworth Sleepiness Scale	380	2121
Sleep Disturbance	395	2174
Average	391.4	2161.4

TABLE IV
ORIGINAL DATA SET TOTAL APNEA DIAGNOSIS COUNTS

VII. MODEL EVALUATION

The models trained off of the questionnaires will be evaluated based on WEKA's evaluation module's various statistical

	Y	N
Zung Depression Scale	392	407
Epworth Sleepiness Scale	380	420
Sleep Disturbance	395	404
Average	389	410

TABLE V
SMALLER DATA SET TOTAL APNEA DIAGNOSIS COUNTS

parameters. These are all evaluations based on the rate of true positives (TP), false negatives (FN), false positives (FP), and true negatives(TN) appearing in the test data, which are tabulated in a confusion matrix.

Medical tests should aim to have low amounts of FNs and FPs. A metric that can summarize and simplify the confusion matrix and take this into account is the F^{β} score. This measurement takes into account sensitivity and precision and uses parameter β to balance sensitivity and precision [19].

For this study, the most crucial quantity is the true positives, which represent when the model correctly predicts the presence of sleep apnea. False positives are where the model would predict that sleep apnea is present and recommend further testing when sleep apnea is not present, making the further testing arbitrary. False negatives represent where the model predicted that no sleep apnea was present when the subject did have sleep apnea. The model would not recommend further testing with false negatives, and the sleep apnea would go undiagnosed and untreated as a result. True negatives represent when the model correctly predicts no sleep apnea present in the subject, and no further testing is recommended.

The metrics of accuracy, precision, and sensitivity are measured from those quantities. Accuracy measures the percentage of all the correct predictions by measures of the TPs and TNs. It will measure the number of correct classifications, but accuracy does not consider the rates of FNs and FPs, which are more common. Sensitivity will measure the number of known positives that are classified. Precision is an important metric that will measure both TPs and FPs and calculates the predicted positives that are correctly classified.

A. Precision

Equation 2 represents the equation for the Precision metric. This metric informs the user of the proportion of prediction positives that is truly positive. In medical terms, this measures the percentage of people with a positive diagnostic test who have the disease. Lastly, how often is the model correct, when the result of the test is yes?

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

B. Sensitivity

Equation 3 represents the equation for the Sensitivity. This metric informs the user of the proportion of positives that are correctly classified. In medical terms, the proportion of patients that have both the disease and a positive result in testing. That is: when the diagnosis is positive, how often the model predicts yes.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

C. F-Measure (F1 Score)

Equation 4 represents the equation for the F-Measure metric. This metric maintains the balance between the recall and precision for the classifier because it's difficult to compare two models if there wasn't any balance. This is a weighted average of the recall and precision values.

$$F_1 = \frac{2(Precision)(Sensitivity)}{Precision + Sensitivity} \tag{4}$$

D. Confusion Matrix

The Confusion Matrix is a table that represents the performance of the algorithm [19]. The predictions are summarized and displayed by class. Figure 5, illustrates a confusion matrix model.

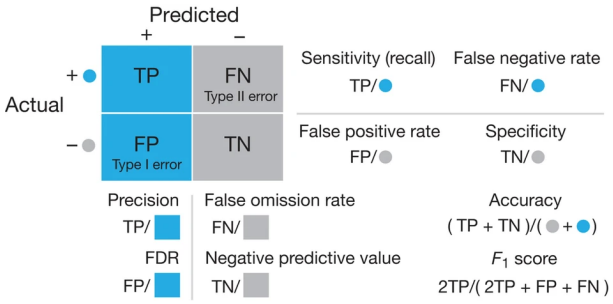

Fig. 5. Confusion Matrix Quantities and Metrics [19]

E. Matthews Correlation Coefficient

One final parameter that can represent the full confusion matrix is the Matthews Correlation Coefficient (MCC):

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

The MCC ranges from -1 to 1. At -1, the classification is always wrong, at 0, the classification is seemingly random, and at 1 the classification is always correct [19]. The MCC is an easy way to evaluate the general effectiveness of each model at a glance. The strengths and weaknesses of each model are evaluated by using these various values and metrics.

VIII. EXPERIMENTAL RESULTS

For each questionnaire, a machine learning model was trained and tested using the j48 decision tree and multilayer perceptron neural network model. Ten-fold cross-validation was used to train and test the models. The confusion matrix values for true positives(TPs), false negatives(FNs), false positives(FPs), and true negatives(TNs) were recorded for each test, as well as the Weka-provided value for the correctly-classified percentage. For each questionnaire the sensitivity,

precision, accuracy, F-measure, and MCC are calculated and recorded. The confusion matrix and calculation results are all recorded in Table IX. Lastly, the standard deviation was calculated for each of these metrics. The standard deviation measures the variation between the metrics of the different questionnaires.

	Predicted Yes	Predicted No
Actual Yes	331	48
Actual No	61	359

Fig. 6. Zung Depression Scale Decision Tree Confusion Matrix

	Predicted Yes	Predicted No
Actual Yes	330	41
Actual No	62	366

Fig. 7. Zung Depression Scale Neural Network Confusion Matrix

	Predicted Yes	Predicted No
Actual Yes	298	87
Actual No	62	333

Fig. 8. Epworth Sleepiness Scale Decision Tree Confusion Matrix

	Predicted Yes	Predicted No
Actual Yes	195	54
Actual No	185	366

Fig. 9. Epworth Sleepiness Scale Neural Network Confusion Matrix

A. Analysis

When examining the accuracy of these models, it can be seen that they all possess similar rates of correct classifications. With a standard deviation of 0.074 the differences between the models is negligible. The models with the highest accuracy is the Zung Depression Scale neural network model with an accuracy of 0.871 and the Zung Depression Scale decision tree model with an accuracy of 0.871. The Zung Depression Scale models also boast the highest values for precision and sensitivity, those metrics also boast low standard deviations of 0.067 and 0.075 respectively. Due to the low standard deviation, the performance of the models in relation to those metrics are all similar.

When examining the MCC values for the models, the lowest value is the Epworth Sleepiness Scale MCC with a value of 0.415 and the highest value is the Zung Depression Scale neural network with an MCC of 0.743. The MCC provides a quick metric by which to measure the general effectiveness of the models. With values ranging from roughly 0.4 to 0.75, the models can be judged as performing well with results that skew towards making correctly classifications the majority of the time. The closer the MCC is to zero the more random its classification is, with an MCC value of 0.743 for its neural network model and an MCC value of 0.727 for its decision tree model the performance of the Zung Depression Scale models are the closest to always correct classification.

	Predicted Yes	Predicted No
Actual Yes	231	54
Actual No	164	350

Fig. 10. Sleep Disturbance Decision Tree Confusion Matrix

	Predicted Yes	Predicted No
Actual Yes	241	68
Actual No	154	336

Fig. 11. Sleep Disturbance Neural Network Confusion Matrix

The effectiveness of each model is also judged by their confusion matrix values. In medical testing the amount of false negatives is an important metric. In these models a false negative reflects the model predicting that no sleep apnea is present in the patient when there is sleep apnea present. The models all show false negative counts in the range of 41 to 87 with a standard deviation of 16.488. The Epworth Sleepiness Scale decision tree model possesses the highest amount of false negatives with a count of 87(Figure 8). This is a rate of 11% of the classifications resulting in a false negative. The models with the lowest amount of false negatives are the Zung Depression Scale models with a count of 48 for its decision tree models(Figure 6) and a count of 41 for its neural network model(Figure 7). Despite the higher rate of false negatives in its decision tree model, the Epworth Sleepiness Scale neural network has a low count of false negatives with a total of 54 (Figure 9).

The amount of false positives in these models ranges from

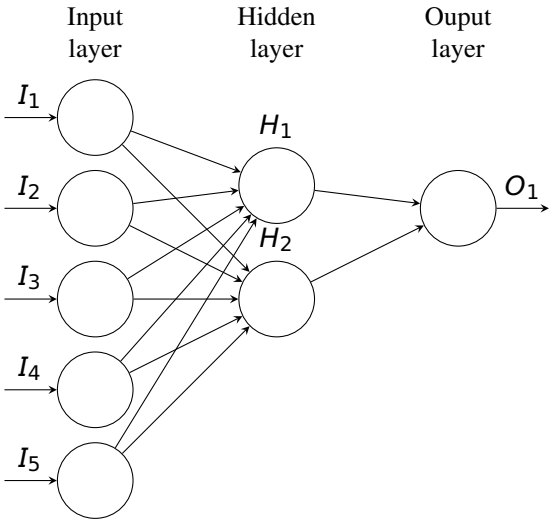

Fig. 12. Example ANN Architecture

TABLE VI
EPWORTH SLEEPINESS SCALE NEURAL NETWORK INPUTS AND OUTPUTS

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
Inputs	ess	ep1	ep2	ep3	ep4	ep5	ep6	ep7	sleepiness	ps_eds
Output	apnea									

TABLE VII
SLEEP DISTURBANCE NEURAL NETWORK INPUTS AND OUTPUTS

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
Inputs	eval_general	ps_diff	ps_backsleep	ps_wakerepeat	ps_tooearly	ps_notrested	ps_wakeup	ps_nightmare
Output	apnea							

	I_9	I_{10}	I_{11}
Inputs	anyinsomnia	ninsomnia	ninsomniadays
Output	apnea		

TABLE VIII
ZUNG DEPRESSION SCALE NEURAL NETWORK INPUTS AND OUTPUTS

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
Inputs	zung1_scored	zung2_scored	zung3_scored	zung4_scored	zung5_scored	zung6_scored	zung7_scored	zung8_scored
Output	apnea							
	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}
Inputs	zung9_scored	zung10_scored	zung11_scored	zung12_scored	zung13_scored	zung14_scored	zung15_scored	zung16_scored
Output	apnea							

	I_{17}	I_{18}	I_{19}	I_{20}	I_{21}	I_{22}
Inputs	zung17_scored	zung18_scored	zung19_scored	zung20_scored	zung_index	zung_score
Output	apnea					

61 to 185 with a standard deviation of 55.825. A false positives represents when the model predicts the presence of sleep apnea when there is no sleep apnea present. While this is not as significant in medical testing as false negatives, a false positive will result in further testing when none is needed and will nullify the goal of these models in reducing the amount of unneeded testing. The Epworth Sleepiness Scale neural network model has the highest amount of false negatives with 185 (Figure 9), this is 23% of the full data set that was tested for this model. The Sleep Disturbance questionnaire also possesses high false negative models with a count of 164 (Figure 10) for its decision tree model and a count of 154 (Figure 11) for its neural network model. The Zung Depression Scale boasts the lowest count of false positives in its model, possessing a total of 61 false positives in its decision tree model (Figure 6) and 62 in its neural network model (Figure 7).

B. Conclusions

Through looking at the various metrics provided, it can be concluded that the Zung Depression Scale models shows the highest performance. With the lowest count of false negatives and a high rate of accuracy as well as a high MCC value, the models perform well in relation to the metrics important in medical diagnoses. The Zung Depression Scale questionnaire focuses on the self-rated severity of depression symptoms that the subjects feel. As discussed earlier, sleep disordered breathing is often linked to various mental disorders such as depression[1]. The effectiveness of the models trained on this questionnaire further shows the strong correlation between the two disorders. The Zung Depression Scale questionnaire also possess the highest number of variables between all the questionnaires. It created the most expansive decision tree model (Figure 16) and neural network (Table VIII) of the questionnaires as a result of this. This abundance of variables present in this questionnaire may also be contributing to the higher rate of success it boasts.

Despite the high performance of the Zung Depression Scale models, the other models also possess high accuracy, sensitivity and precision ratings and are successful in diagnosing sleep apnea and suggesting further testing relatively reliably. The MCC of each model being roughly 0.5 shows that their classifications are by no means random though the effectiveness of each does vary as seen in the confusion matrix of each model. With low standard deviations across the board for all metrics, each model can be judged as being effective.

C. Limitations

An important aspect to consider in this study is the relatively low number of positive sleep apnea patients in the study. As seen in Table IV, the number of positive sleep apnea diagnoses present in the data set is roughly 18% of the total data points. This skewed proportion is why each questionnaire was parsed down and relation tuples were deleted in order to bring the count of positive and apnea classifications to a more even amount. The apnea counts for this smaller table can be seen in Table V. If there were a higher amount of positive apnea classifications present in the data set and it could be expaned with more data, the models would most likely have an increased effectiveness.

Another factor in this data set is the variable number of visits and tests that each subject underwent. Each subject had

	Correctly Classified (%)	TP	FN	FP	TN	Precision	Sensitivity	Accuracy	F-Measure	MCC
Zung Depression Scale										
j48	86.358	331	48	61	359	0.864	0.864	0.864	0.864	0.727
MultilayerPerceptron	87.109	330	41	62	366	0.872	0.871	0.871	0.871	0.743
Epworth Sleepiness Scale										
j48	78.875	298	87	82	333	0.789	0.789	0.789	0.789	0.577
MultilayerPerceptron	70.125	195	54	185	366	0.721	0.701	0.701	0.690	0.415
Sleep Disturbance										
j48	72.716	231	54	164	350	0.745	0.727	0.727	0.721	0.471
MultilayerPerceptron	72.215	241	68	154	336	0.732	0.722	0.722	0.719	0.454

TABLE IX
MODEL TESTING RESULTS AND CALCULATIONS

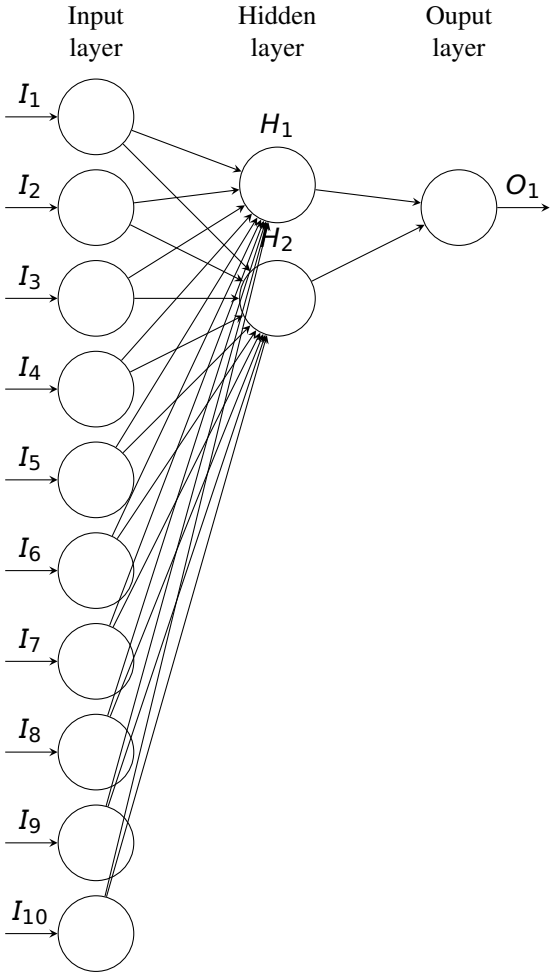

Fig. 13. ANN Architecture for Epworth Sleepiness Scale

between one to four visits each, and each time would fill out all questionnaires. Though their questionnaire answers may change, their sleep apnea diagnosis would not. These multiple visits may have skewed the data as well.

The final limitation is the size of the Zung Depression Score questionnaire. With a total of 22 variables in the data set, the models built for the data set are all complex and the visual representations of the models are messy and difficult to read. In an attempt to simplify the models, a simplified Zung Depression Scale model was tested by utilizing Wekas built-in filters and deleting various amounts of variables that affect the model the least. This resulted in a model with less variables and more organized models. However, the simplified model had significantly worse performance than the full model and was abandoned as a result of those poor results.

IX. ACKNOWLEDGEMENTS

This Wisconsin Sleep Cohort Study was supported by the U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (R01HL62252), National Institute on Aging (R01AG036838, R01AG058680), and the National Center for Research Resources (1UL1RR025011). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002). [20][21]

REFERENCES

[1] C. Clinic, "Sleep apnea: Causes, symptoms, tests and treatments," Available at https://my.clevelandclinic.org/health/diseases/8718-sleep-apnea (2021).

[2] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to diagnosis and treatment of patients with suspected sleep apnea," *Am J Respir Crit Care Med.*, Mar. 2004.

[3] S. S. Resource, "Sleep apnea statistics," Available at https://ineedbettersleep.com/faqs/sleep-apnea-statistics/ (2021).

[4] D. M. Hiestand, P. Britz, M. Goldman, and B. Phillips, "Prevalence of symptoms and risk of sleep apnea in the us population," *Chest*, vol. 130, no. 3, pp. 780–786, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0012369215527918

[5] F. . Sullivan, "Economic impact of obstructive sleep apnea," Available at https://aasm.org/advocacy/initiatives/economic-impact-obstructive-sleep-apnea/ (2021).

[6] S. Online, ""home sleep apnea test cost"," Available at https://www.sleepcareonline.com/home-sleep-apnea-test-cost/ (2021).

[7] P. M. P. Health, "Sleep study frequently asked questions," Available at https://www.princetonhcs.org/care-services/sleep-center/sleep-study-frequently-asked-questions (2021).

[8] A. Burkov, *Machine Learning Engineering*. Quebec, CA: True Positive Inc., 2020.

[9] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. New York NY: O'Reilly Media, 2019, vol. 2nd Edition.

[10] L. Moroney, *AI and Machine Learning for Coders: A Programmer's Guide to Artificial Intelligence*, ser. 1st Edition. Sebastopol, CA: O'Reilly Media, 2020.

[11] C. Sheppard, *Tree-Based Machine Learning Algorithms: Decision Trees, Random Forests, and Boosting*, ser. Decision Tree. Createspace Independent Publishing Platform, 2017.

[12] Rokach, Lior, and O. Maimon, *Data Mining With Decision Trees: Theory And Applications (2Nd Edition) (Machine Perception and Artificial Intelligence)*, ser. 2nd Edition. Wspc, 2014.

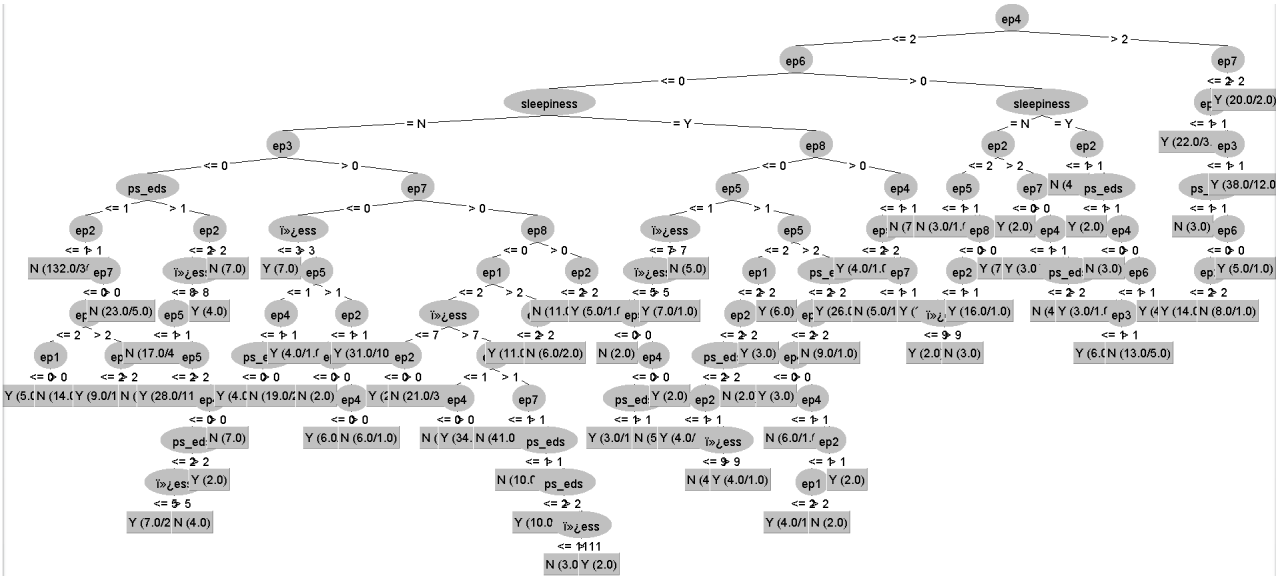

Fig. 14. Epworth Sleepiness Scale Decision Tree Model

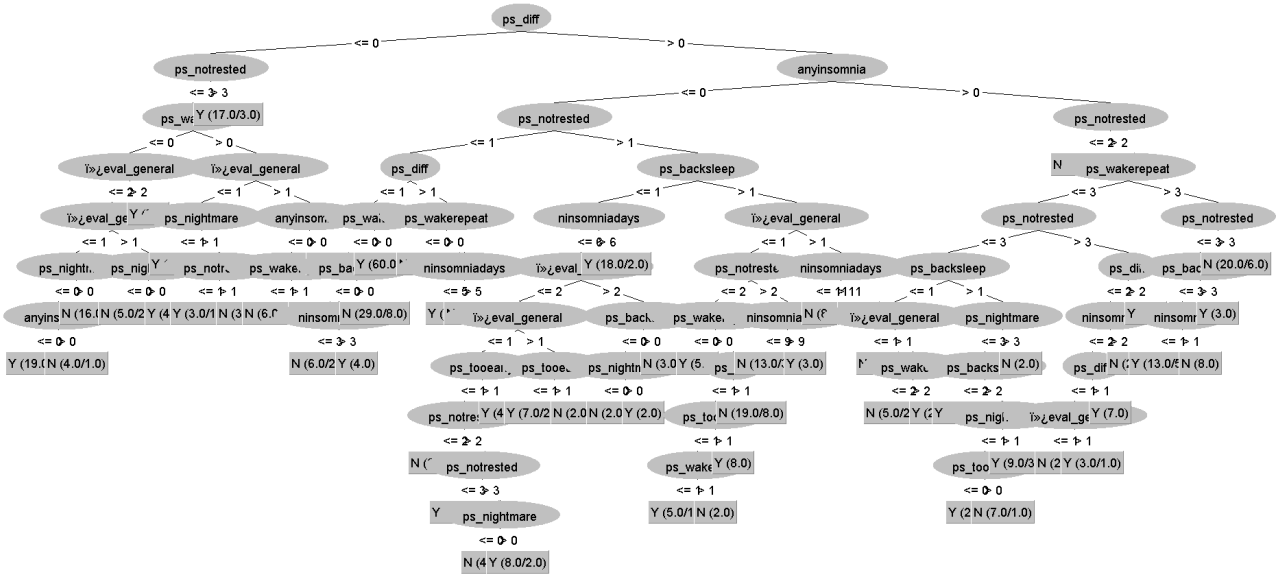

Fig. 15. Sleep Disturbance Decision Tree Model

[13] J. Huang and H. Yau, "Dynamics of deep neural networks and neural tangent hierarchy," *CoRR*, vol. abs/1909.08156, 2019. [Online]. Available: http://arxiv.org/abs/1909.08156

[14] G. R. Yang and X.-J. Wang, "Artificial neural networks for neuroscientists: A primer," *Neuron*, vol. 107, no. 6, p. 1048–1070, Sep 2020. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2020.09.005

[15] T. Young, "Rationale, design and findings from the wisconsin sleep cohort study: Toward understanding the total societal burden of sleep disordered breathing," *Sleep Med Clin*, Mar. 2009.

[16] W. W. Zung, "A self-rating depression scale," *Archives of general psychiatry*, Jan. 1965.

[17] M. W. Johns, "A new method for measuring daytime sleepiness: the epworth sleepiness scale," *Sleep*, Dec. 1991.

[18] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench*, ser. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaugmann, 2016, vol. 4th Edition.

[19] J. Lever, M. Krzywinski, and N. Altman, "Classification evaluation," *Nat Methods*, Jul. 2016.

[20] T. Young, M. Palta, J. Dempsey, P. E. Peppard, F. J. Nieto, and K. M. HLA, "Burden of sleep apnea: rationale, design, and major findings of the wisconsin sleep cohort study," *WMJ: official publixation of the State Medical Society of Wisconsin*, Aug. 2009.

[21] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redlin, "The national sleep research resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, Oct. 2018.

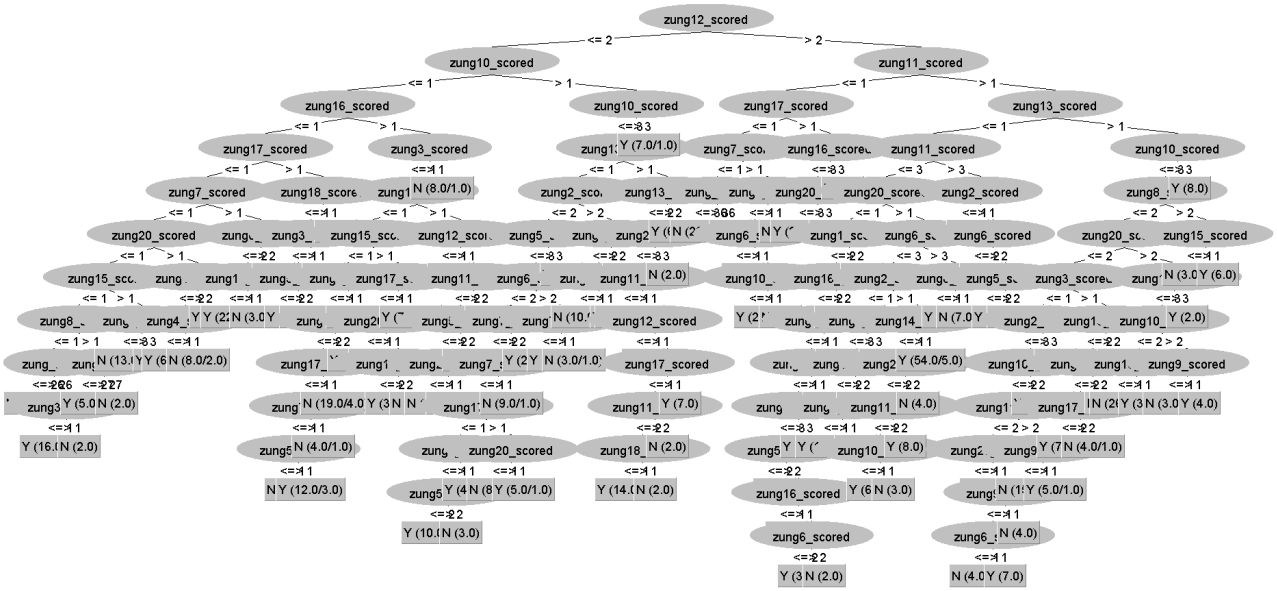

Fig. 16. Zung Depression Scale Decision Tree Model