An ANN based Screener for the early diagnose of Polycystic Ovarian Syndrome in adolescent and young women

Dr.V.Krishnaveni¹, Dr.C.Deepa², R.S.Cindhu³, Dr.K.G.Santhinga⁴

¹Associate Professor in Computer Science, Kongu Arts and Science College (Autonomous), Erode, Tamilnadu– 638107
²Assistant Professor & Head in Biotechnology, Kongu Arts and Science College (Autonomous), Erode, Tamilnadu– 638107
³Assistant Professor in Biotechnology, Kongu Arts and Science College (Autonomous), Erode, Tamilnadu– 638107
⁴Associate Professor in Computer Science, Kongu Arts and Science College (Autonomous), Erode, Tamilnadu– 638107

Abstract

Polycystic Ovarian Syndrome (PCOS) is the most common cause of menstrual dysfunction and hyperandrogenism. PCOS is recognized as a heterogeneous disorder that results in overproduction of androgens, primarily from the ovaries and leads to anovulation, hirsutism, and insulin resistance. PCOS diagnosis is challenging for the physicians because of the varying diagnostic criteria and inconsistency of the patients’ complaints. It is imperative for health care providers to diagnose PCOS early to decrease risks for comorbidities such as obesity, hypertension, dyslipidemia, cancer, and Type 2 diabetes. Hence, a validated diagnostic screening Tool would be very helpful in assisting physicians in making an early diagnosis of PCOS. The objective of this paper is to build a PCOS Screener, based on the Artificial Neural Network, a classification model with the support of a Questionnaire based diagnostic tool designed for this purpose, especially for adolescent and young women, which would assist the physicians in early diagnosing the PCOS and would save considerable time in examining the patients and hence would reduce the delay in diagnosing the risk of PCOS.

Keywords: PCOS, heterogeneous disorder, PCOS Screener, Questionnaire, Validation, Artificial Neural Network

1. Introduction

Polycystic ovarian syndrome (PCOS) is the most common endocrine disorder and cause of anovulatory infertility in childbearing age women (Tang et al., 2006). This disease is complex and the exact physiology is unclear (Garad, Teede, & Moran, 2011). What is known about this condition is that hormone imbalance is the underlying problem. Hyperandrogenism and insulin resistance contributes to the etiology process of PCOS (Garad, Teede, & Moran, 2011). Women with PCOS can present with polycystic ovaries, but it is not necessary for this diagnosis (Boyle & Teede, 2012). This population of women may have a plethora of symptoms and
findings related to their condition. According to Madnani, Khan, Chauhan, Parmar (2013), the following signs and symptoms are common for women of reproductive age with PCOS: metrorrhrea or amenorrhea, acne, irregular menses, hirsutism, alopecia. Additional symptoms included metabolic syndrome, obesity, insulin resistance, acanthosisnigricans, Type 2 diabetes, dyslipidemias, hypertension, non-alcoholicliver disease, and obstructive sleep apnea (Madnani et al., 2013).

The organization of paper is as follows: section 2 describes the Related Works; section 3 discusses The Need and the Design of a Diagnostic Tool. Evaluation of the Diagnostic Tool and the Results of the Diagnosis are present in section 4; Section 5 describes the Building of ANN based Classification Model – PCOS Screener; the description on the Validation of the Classification Model is found in section 6 and finally section 7 is the conclusion.

2. Related Works

The features to be considered for building the Tool were observed from the following works in the domainrelated to PCOS.

Moran et al. reported PCOS affects 5-10% of women who are in the reproductive age group in [1]. According to Sirmans& Pate (2014) 30% of the PCOS population will experience normal menses[2]. Several studies have suggested that hypertension is more prevalent in the PCOS population compared to the general population [3]. According to [4], hypertension represented 45% of the PCOS population. A common factor in PCOS is obesity. Obesity seems to pose a higher risk for hypertension in the PCOS population. Women with PCOS may lack a healthy vasculature secondary to a decrease in nocturnal blood pressure [3] and experience elevations in their mean arterial and ambulatory systolic pressures [5]. Often times PCOS is undiagnosed in women. This means that this population of women is uneducated regarding their condition, possible comorbidities, and treatment options. It is imperative for health care providers to diagnose PCOS early to decrease risks for comorbidities. Understanding some of the PCOS related comorbidities such as obesity, hypertension, dyslipidemia, cancer, and Type 2 diabetes, will help this population live a balanced and healthier life[6].

Apridonidze et al. (2005) reported obesity in 67% of women with PCOS[4]. Obesity seems to have increased over the years, which has had a big impact on the development of chronic conditions such as metabolic syndrome, coronary heart disease and type 2 diabetes [4]. Reproductive health is significantly impacted by obesity due to increased body weight that contributes to ovulatory infertility. Overweight and obesity are commonly seen in PCOS women [4]. Having excess body weight can exacerbate symptoms of PCOS such as hyperandrogenism, menstrual problems, infertility, insulin resistance, dyslipidemia, increased risk of metabolic syndrome, impaired glucose tolerance, and type 2 diabetes [5]. Cardiovascular risk factors are significant among the PCOS population [5]. The risk varies according to the levels of LDL, HDL, triglycerides, and total cholesterol. Atherosclerosis has been reported to occur at higher rates in women with PCOS [5]. It has been noted that early coronary and other vascular diseases has been documented in the PCOS population by different techniques [5].

Dyslipidemia is strongly associated with the PCOS population (Apridonidze et al., 2005). According to data collected in a study by Apridonidze et al. (2005), 35% of the PCOS population had elevated lipid levels[4]. Triglycerides and the low-density lipoproteins (LDL) were elevated above the normal limits and the high-density lipoproteins (HDL) were decreased [4]. Diamanti-Kandarakis, Papavassiliou, Kandarakis, and Chrousos, (2007), the National Cholesterol Education Program (NCEP) guidelines stated that approximately 70% of PCOS
patients exhibit abnormal serum lipid levels [7]. An unfavorable lipid profile that consists of increased LDL, decreased HDL and increased total cholesterol and triglyceride levels are associated with the elevated androgen and insulin levels found in women with PCOS [8].

Most PCOS patients have a family history of PCOS. It is estimated that PCOS patients that have family members with a metabolic disorder, have approximately a 2.7 higher chance of having dyslipidemia than non PCOS patients.

Type 2 Diabetes and Impaired glucose tolerance is also prevalent in the PCOS population. These patients have a 5 to 10-fold increased risk of developing Type 2 diabetes (Hoyt & Schmidt, 2004). It has been noted that there was a high prevalence of first degree relatives with type 2 diabetes, which confirms family history as an important risk factor (Hoyt & Schmidt, 2004).

3. The Need and the Design of a Diagnostic Tool

3.1 The Need

The PCOS diagnosis is found to be challenging for health care providers because of the varying diagnostic criteria and variance in patients’ complaints. Over the years, different diagnostic criteria have been developed, a) National Institutes of Health (NIH) 1990 which includes the following elements and both criteria needed for diagnosis: 1. chronic anovulation and 2. Clinical and/or biochemical signs of hyperandrogenism (with exclusion of other etiologies) (National institutes of health, 2012); b) Rotterdam 2003 which includes the following criteria and two of three criteria needed for diagnosis: 1. Oligo and/or anovulation 2. Clinical and/or biochemical signs of hyperandrogenism 3. Polycystic ovaries (National institutes of health, 2012); c) Androgen Excess & PCOS Society (AE-PCOS) 2006, which includes the following criteria and (both criteria needed for diagnosis: 1. Clinical and/or biochemical signs of hyperandrogenism 2. Ovarian dysfunction (oligo-anovulation and/or polycystic ovarian amorphology) (National institutes of health, 2012).

The PCOS criteria have their own strengths and limitations. They can cause confusion among health care professionals, which could delay the progress in diagnosing the PCOS and the ability to collaborate with women to address and manage their PCOS health related issues. (National institutes of health, 2012). In order for this population to live a balanced life with PCOS, early detection is necessary.

3.2 The Design of the Diagnostic Tool

The objective of the paper is to propose a PCOS diagnostic screener as a tool to diagnose probable PCOS patients. The tool has been designed to classify the patients into three categories such as the patients with Low Risk, Moderate Risk and High Risk.

3.2.1 Methods and Materials

A Stratified random sampling method was used to collect the data from the Adolescent and young girl, with probable PCOS symptoms, in and around Erode, Gobichettipalayam, Coimbatore, Tirupur and Udumalpet after explaining the purpose of the study and getting their consent and cooperation. The study protocol was approved by the Ethics Committee of Erode Cancer Center, Erode, Tamilnadu, India.

The data was collected by using a structured questionnaire on PCOS built using the knowledge gained from the Literature study and the Physicians. Out of 450 samples collected for this study, 300 samples have
been selected for building the prediction model among them 116 have been classified with High risk of having PCOS; 92 with moderate risk; and 92 with low risk.

3.2.2 Study Instruments

A Questionnaire, which played a vital role for the design of the screener, with 15 Questions was used to collect data from the patients. The patients were classified as adolescent if their age lie within the range [10 to 19] and young if their age lie within the range [20 to 24]. The samples of unmarried young women were taken into account. The prominent factors of PCOS were identified from the literature and were included as the components of the questionnaire as suggested by Physicians.

---

**Screening Questionnaire for diagnosis of Polycystic Ovary Syndrome (PCOS)**

for the UGC Sponsored Minor Research Project

**Principal Investigator:**

Dr. V. Kirthikavani, Associate Professor in Computer Science,
Kongu Arts and Science College (Autonomous), Erode, Tamil Nadu
Mobile : 7898212301

---

**Dear Respondent,**

**Good day!**

Polycystic Ovarian Syndrome (PCOS) is a condition which leads to growth of ovarian cysts. It is a prevalent endocrine disorder in child-bearing women, which can lead to infertility, dangerous illness like Type 2 Diabetes, Gestational Diabetes, Cardiovascular and Cerebrovascular dysfunction and higher risk of mood and anxiety disorders which can cause depression. This questionnaire has been designed as a survey and screening tool for diagnosis of PCOS in young women.

Through this questionnaire, the researcher will be able to assist the physician in diagnosing the presence of PCOS and in recommending appropriate treatments. Thank you very much for your time and response.

<table>
<thead>
<tr>
<th>Name (optional):</th>
<th>Age:</th>
<th>Class:</th>
</tr>
</thead>
<tbody>
<tr>
<td>School/College:</td>
<td>Place:</td>
<td></td>
</tr>
</tbody>
</table>

**(a) SYMPTOMS**

**(i) Weight gain**

1. How much do you crave for carbohydrates and sugar?
   - LOW
   - MEDIUM
   - HIGH

2. Do you feel extremely hungry, irritable, sleepy, or fatigued after eating sweets?
   - NEVER
   - SOMETIME
   - OFTEN

3. What kind of personality you are?
   - SLIM
   - NORMAL
   - OBESE

4. How long you have continuous weight gain in spite of diet control/physical exercise? (In years)
   - Not applicable
   - < 2
   - 2 to 5
   - > 5

5. What is the size of your waistline (in inches)?
   - < 30
   - 30 TO 35
   - > 35

**(ii) Acne (Red Pimples on the face)**

6. Have you undergone medical treatment for Acne?
   - NO
   - YES

---

87
This section deals with the evaluation of the Diagnostic Tool and results of the diagnosis.

**Figure 3.1** The Questionnaire Used

4. Evaluation of the Diagnostic Tool and the Results of the Diagnosis

This section deals with the evaluation of the Diagnostic Tool and results of the diagnosis.
4.1 Evaluation

- **Questions with 3/4 Options**
  
  Each question was assigned a score value based on the Nature of the Option Chosen. The options highly associated with the prevalence of PCOS were assigned with the highest value whereas the options at the other extreme were assigned with the lowest value. The Options in the middle were assigned with middle range values.

- **Questions with Yes/No Options**
  
  Highest Scores were assigned to Yes options as they were highly associated with PCOS positive.

- **Questions with Multiple Choices**
  
  A Ferriman Gallwey Index chart tool was used to help assess for hirsutism. Ferriman Gallwey Index is used for clinical assessment of hair growth in women and to score the degree of excess male pattern body hair. Each choice is assigned with values and they were added up to indicate the severity of hirsutism.

4.2. Results of the Diagnosis

If the total score is between [0 to 35], the patient was classified under Low Risk category and if the total score is between [36 to 70], the patient was classified under Moderate Risk category and if the total score is between [71 to 100], the patient was classified under High Risk category.

Once a patient has been classified under moderate or high risk category, the appropriate lab studies would be suggested and the diagnostic criteria (NIH 1990, Rotterdam 2003, or AE-PCOS Society 2006) can be applied to diagnose the patient.

Once a patient has been classified under Low or high risk category with abnormal symptoms in physical appearance, they would be advised to go for lifestyle changes such as being active, eating well and maximising emotional health.

5. Building of ANN based Classification Model – PCOS Screener

5.1 Training Dataset

One of the main objectives of the study is to build a classification model. The building of a prediction model requires a training dataset with sufficient samples. The Diagnostic Tool developed acted as the main source of generating the samples.

A Training Dataset with 300 samples have been used for building the prediction model among them 116 have been classified with High risk of having PCOS; 92 with moderate risk; and 92 with low risk. Each sample in the dataset has been represented with 16 features, where 15 are input features and 01 Class Label as shown in Fig. 5.1. All the 15 input features represent the score values of the associated Questions in the Questionnaire.

The Label takes the value either 0/1/2 depending on the score values representing Low/Moderate/High risk categories.

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>x6</th>
<th>x7</th>
<th>x8</th>
<th>x9</th>
<th>x10</th>
<th>x11</th>
<th>x12</th>
<th>x13</th>
<th>x14</th>
<th>x15</th>
<th>Label</th>
</tr>
</thead>
</table>

Figure. 5.1 The structure of an instance of the dataset

89
5.2 Learning with Artificial Neural Networks

One of the most successful classification models is the Artificial Neural Network. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired the brain. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning largely involves adjustments to the synaptic connections that exist between the neurons.

The following diagram represents the general model of ANN which is inspired by a biological neuron. It is also called Perceptron.

\[
\text{Mathematically, } x_1w_1 + x_2w_2 + x_3w_3 + \ldots + xnwn = \sum xiwi
\]

The activation function is applied \( \sum xiwi \) and the final result \( Y \) is obtained. Based on the value of \( Y \) the class label is determined.

5.3 Learning Process

At every iteration, the inputs are transformed into a predicted value. Based on the difference between the actual value and the predicted value, an error value also called Cost Function is computed and sent back through the system as shown in Fig. 5.3.

Cost Function: One half of the squared difference between actual and output value.

For each layer of the network, the cost function is analyzed and used to adjust the threshold and weights for the next input. The main aim is to minimize the cost function. The lower the cost function, the closer the actual value to the predicted value. In this way, the error keeps becoming marginally lesser in each run as the network learns how to analyze values.
The resulting data is fed back through the entire neural network. The weighted synapses connecting input variables to the neuron are the only thing the user have control over. This procedure described above is known as Back-propagation and is applied continuously through a network until the error value is kept at a minimum.

5.3.1 Activation Function - Sigmoid Activation Function — (Logistic function)
A Sigmoid function is a mathematical function having a characteristic “S”-shaped curve or sigmoid curve as shown in Fig. 5.4 which ranges between 0 and 1, therefore it is used for models where it is required to predict the probability as an output. This function is used in the proposed prediction model and the output Y was interpreted as follows.

If $0 \leq Y \leq 0.35$, label = 0  
If $0.35 < Y \leq 0.70$, label = 1  
If $0.70 < Y \leq 1$, label = 2

5.3.2 Training ANN with Sigmoid Function

Step-1 → Randomly initialize the weights to small numbers close to 0 but not 0.  
Step-2 → Input the first observation of the training dataset in the input layer, each feature in one node.
Step-3 → Forward-Propagation: From left to right, the neurons are activated in a way that the impact of each neuron's activation is limited by the weights. Propagate the activations until getting the predicted value.

Step-4 → Compare the predicted result to the actual result and measure the generated error using Cost function.

Step-5 → Back-Propagation: from right to left, the error is backpropagated. Update the weights according to how much they are responsible for the error. The learning rate decides how much we update weights.

Step-6 → Repeat step-1 to 5 and update the weights after each observation(Reinforcement Learning)

Step-7 → When the whole training set passed through the ANN, that makes and epoch. Redo more epochs.

6. Validation of the Classification Model

The Model was validated by a second sample of 130 patients received from an endocrinology clinic, where 65 of whom had been diagnosed with High risk, 45 with Moderate Risk and 20 with Low Risk. The components considered in the Questionnaire were taken into consideration for the validation process. The performance of the proposed ANN based classifier was compared with the performance of yet another efficient classifier, the Bayesian Classifier, on various parameters.

6.1 Bayesian Classifier

The Bayesian classifier is the most useful and efficient probabilistic learning technique in the field approach machine learning [24]. It is a supervised statistical method used for binary and multiclass classification. This is based on Baye’s theorem named after Thomas Bayes [13]. Moreover, it can be used for solving diagnostic and predictive problems. In this suppose we have an n-dimensional feature space with values

\[ X = \{x_1, x_2, x_3, \ldots, x_n\} \]

For this data we have \(n = 15\).

In Bayesian classifier which assigns the given data values into m classes like \(\beta_1, \beta_2, \beta_3, \ldots, \beta_m\). Then according to the greater posterior probability the new data will be classified in any of the classes. That means, \(X\) will be classified to class \(\beta_i\) if

\[ P(\beta_i|X) > P(\beta_j|X) \text{ for all } i \text{ and } j \text{ such that } 1 \leq j \leq m, j \neq i \]

Bayes theorem provides a way of calculating posterior probability \(P(c|x)\) from \(P(c), P(x)\) and \(P(x|c)\) [14][15].

\[ P(c|x) = \frac{P(x|c)P(c)}{P(x)} \]

- \(P(c|x)\) is the posterior probability of class \(c\), target given predictor \(x,\) attributes.
- \(P(c)\) is the prior probability of class.
- \(P(x|c)\) is the likelihood which is the probability of predictor given class.
- \(P(x)\) is the prior probability of predictor.

6.2 Performance Measures

In order to calculate the accuracy and to find out the performance of the proposed Classification model, PCOS screener , a confusion matrix was constructed at first. From that matrix, the true positives, true negatives, sensitivity, specificity and precision were calculated using their specific formula. These parameters were then used to calculate the final accuracy.

Sensitivity
Sensitivity is nothing but, the true positive rate. We can also define it as the fraction of positive tuples that are correctly classified.

\[
\text{Sensitivity} = \frac{\text{Frequency of True Positives}}{\text{Frequency of True Positives} + \text{Frequency of False Negatives}}
\]

**Specificity**

The negative rate which is nothing but the fraction of negative tuples that are correctly classified is called specificity.

\[
\text{Specificity} = \frac{\text{Frequency of True Negatives}}{\text{Frequency of False Positives} + \text{Frequency of True Negatives}}
\]

**Precision**

It is the measure in which the fraction of true positives in contrary to all positive results is calculated.

\[
\text{Precision} = \frac{\text{Frequency of True Positives}}{\text{Frequency of True Positives} + \text{Frequency of False Positives}}
\]

**Accuracy**

The percentage of the test tuples that are properly classified by the classifiers is nothing but the accuracy of the particular algorithm in hand.

\[
\text{Accuracy} = \frac{\text{Frequency of True Positives} + \text{Frequency of True Negatives}}{\text{Frequency of True Positives} + \text{False Positives} + \text{True Negatives} + \text{False Negatives}}
\]

### Table 6.1 Performance Comparison of ANN and Bayesian Classifier

<table>
<thead>
<tr>
<th>Performance Metrics</th>
<th>ANN</th>
<th>Bayesian Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>95.67</td>
<td>92.85</td>
</tr>
<tr>
<td>Specificity</td>
<td>97.17</td>
<td>94.23</td>
</tr>
<tr>
<td>Precision</td>
<td>95.32</td>
<td>93.93</td>
</tr>
<tr>
<td>Accuracy</td>
<td>96.21</td>
<td>94.16</td>
</tr>
</tbody>
</table>

6.3 Result Analysis

It is evident from the table 6.1 that both the classifiers have achieved reasonable Performance. The Bayesian classifier is a supervised classification technique and a good approach for classification technique. It uses simple computation and does not require any random values for initiating the classification process. But, it was observed that the performance degrades as the size of the dataset increases. On the other hand ANN works well when the data set is large. Moreover ANN requires higher computational cost than Bayesian classifier. For the dataset considered, the best result was obtained from ANN classifier (96.21%) as compared to Bayesian classifier (94.16%) and it could be justified that ANN is the best suitable candidate for classifying the dataset that have been considered.
7. Conclusion

Even though PCOS has been identified as the most common endocrinal disease, the authors have observed that there are a very limited number of researches initiated towards building a classification model for the early diagnosis of PCOS and the level of risks associated with the symptoms and other parameters obtained from the patients, especially from the young and adolescent women. An efficient screener for PCOS is the need of the hour as PCOS is found to become prevalent in the young women in our country. The use of Screener of this kind in the clinics would be very helpful to physicians in making a diagnosis of probable PCOS and to treat the patient accordingly. Using the screening diagnostic Tool for probable PCOS would help the physicians be more aware of patients that have PCOS, possibly decrease the cost of ordering various lab studies that may not be needed, and most of all, start early intervention to decrease PCOS related comorbidities. An early and accurate diagnosis of PCOS would allow health care professionals the ability to educate patients early on preventive measures and lifestyle changes that would benefit them. Hence, In this work, a Classification model based on Artificial Neural Network, PCOS Screener, has been suggested. It classified the patients into three categories namely, with low, moderate and high risk of having PCOS. The performance of the model was compared with the Bayesian classifier for the dataset considered and the ANN model outperformed the Bayesian.

Acknowledgement

This work is a part of a Minor Research Project funded by University Grants Commission, New Delhi, India

Reference


