PCOS - Detection and Analysis

Advaith Chandra Srivastav
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India
advaith.srivastav@gmail.com

Ayan Samanta
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India
ayan.samanta2020@vitstudent.ac.in

Eshana Mohan
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India
eshanamohan@gmail.com

Hrishita Panjetha
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India
hrishita.panjetha@gmail.com

Sasikala R
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India

***Abstract*— The objective of the Polycystic Ovary Syndrome (PCOS) Detection and Analysis project is to find a solution to the ongoing problem of polycystic ovary syndrome (PCOS) by employing data visualization in the form of graphs and machine learning algorithms to diagnose and investigate the condition. The polycystic ovary syndrome (PCOS) is a complicated medical condition that affects the reproductive system of women and can lead to congenital abnormalities as well as other health complications. It is difficult to diagnose and treat PCOS because its underlying cause is still unknown, even though there are medications available to treat the condition. By employing graphs and algorithms for machine learning, the project will analyze data in an effort to recognize patterns, with the ultimate goal of enhancing our understanding of PCOS and reducing the risk of further injury to those who are afflicted. In this way, the project hopes to come up with an efficient solution for the diagnosis and treatment of polycystic ovary syndrome (PCOS).**

Keywords – PCOS, Data Visualization, TensorFlow, Plotly

I. INTRODUCTION

PCOS is an ongoing health problem that has a negative impact on the reproductive system of women and has been linked to the development of congenital abnormalities, making it a cause for concern for future generations.

Despite the fact that there are treatments for this condition, it is still quite common, and its underlying cause is still a mystery. PCOS has repercussions that go beyond reproductive health and can have an effect on a person's height, weight, and body mass index (BMI).

Data visualisation using graphs can be a useful tool in identifying and analysing PCOS, which can help affected individuals avoid further injury to themselves. In addition to making use of graphs, we will also be applying three-ml algorithms in order to evaluate the results of these models and further our comprehension of this complicated condition.

Despite the availability of several treatments, Polycystic Ovary Syndrome (PCOS) is a pervasive and persistent health problem. Due to the involvement of the ovaries, which are the primary reproductive organs, PCOS is a high-risk syndrome that may result in birth defects in future generations.

Although the underlying aetiology of PCOS is unknown, the condition's complications are a cause for concern. In addition to having negative effects on the female reproductive system, PCOS can also alter height, weight, and BMI.

Using data visualisation through graphs can be an excellent method for recognising and studying PCOS, so preventing further harm to those affected. The harmful effects of PCOS can be minimised with early detection and intervention.

In addition, to improve the accuracy of PCOS diagnosis and analysis, 3 ml algorithms will be implemented and their outputs evaluated. Algorithms for machine learning have the ability to enhance our understanding of the situation and reveal patterns that may not be discernible using conventional techniques.

This combination of data visualisation and machine learning algorithms provides a holistic approach to PCOS identification and analysis, resulting in a more accurate and efficient method of addressing this complex health issue.

II. MOTIVATION

A. Technical Motivation

Data visualisation has improved information dissemination and prompted faster action. Data visualisation simplifies complicated material for visual learners. Data visualisation tools are attractive and effective sales tools that convey information in an engaging and dynamic manner.

Technology has changed data presentation. Beautiful and useful visualisations of complex data can now be created from boring data. Data visualisation tools may customise graphs and charts in minutes. Visual representations speed information processing, improving productivity and results.

The 3 ML algorithms improve complex data visualisation. These algorithms can reveal hidden patterns and trends, helping us make better decisions. Data visualisation and machine learning algorithms improve data analysis and decision-making productivity and results..

B. Economical Motivation

There's no single test to specifically diagnose polycystic ovary syndrome (PCOS). Your healthcare provider is likely to start with a discussion of your symptoms, medications, and any other medical conditions. Your provider also may ask about your menstrual periods and any weight changes. The procedure is usually painless and should not take long. If you experience pain on entry or are uncomfortable at any time you should inform the person doing the scan (nurse, ultra sonographer, or physician) immediately. Doctors opt for this

method instead of an external ultrasound because of the nature of PCOS. To prevent PCOS a person may have to spend a lot of money to have the perfect diet, prevent any heavy exposure to pollution by using an air purifier, and so on. This all can be reduced using data visualisation and can be economically helpful to the person.

C. Environmental Motivation

Polycystic ovary syndrome (PCOS) is a common endocrine disorder affecting women of reproductive age. Environmental factors such as endocrine-disrupting chemicals (EDCs) and obesity, as you mentioned, have been linked to the development of PCOS. Individuals should limit their exposure to EDCs by reducing their use of products containing such chemicals and limiting their exposure to pollution. Additionally, eating a healthy diet and consulting with an environmental healthcare expert can help reduce the risk of developing PCOS.

Our project will employ three different machine learning methods to aid in the detection of PCOS. These methods will aid in identifying patterns and characteristics shared by people with PCOS. We hope that by using these techniques, we will be able to provide a more accurate and reliable method of detecting PCOS, which will help reduce stress and pressure on individuals suffering from the condition. Overall, our project aims to improve health outcomes for women with PCOS by combining lifestyle changes and cutting-edge technology.

D. Demographic feasibility

There are many demographic factors that can affect PCOS, they mostly are age, ethnicity, geographic location, education level, etc. PCOS mostly becomes very serious because the women who are affected are generally unaware of it. This can lead to it becoming a widespread problem and so to prevent this our project can come in handy. Since our project uses data visualisation and analysis to identify the PCOS and it is a piece of software we can implement it easily and then make it accessible for people in rural areas or people who are living in geographically dangerous areas or ones which are tough to reach.

III. Literature Review

A. Current aspects of polycystic ovary syndrome

A literature review [1] by Victor Hugo Lopes DE Andrade 1, Ana Maria Oliveira Ferreira DA Mata 1, Rafael Soares Borges 2, Danylo Rafhael Costa-Silva 2, Luana Mota Martins 2 3, Paulo Michel Pinheiro Ferreira 1 3, Lívio César Cunha-Nunes 1 3, Benedito Borges DA Silvaheads unless they are unavoidable. The purpose of this article was to present a review of the literature focusing on polycystic ovaries, including its pathogenesis, clinical manifestations, diagnosis, and therapeutic aspects, as well as its association with cardiovascular and arterial hypertensive disorders. Because the aetiology of PCOS is unknown, treatment is limited to managing signs and symptoms. More research is needed to understand the pathophysiology of PCOS and the development of high blood pressure in women with the disorder.

B. Polycystic Ovarian Syndrome(PCOS): A literature review of the polycystic ovary syndrome.

J.Desai, B. Buysee, J.D. Albano [2] PCOS is associated with hormonal disruption and psychological consequences, resulting in a lower health-related quality of life (HRQOL). In Ovary Syndrome 1998, the Polycystic Questionnaire was (PCOSQ) developed as the only patient validated PCOS specific outcome reported (PRO). This review examines the use of PCOSQ in conjunction with other QoL instruments in recent research. Despite well documented patient reported outcomes, some aspects of PCOS are not adequately assessed by the PCOSQ. To obtain a more comprehensive for patient view assessing HRQoL, the PCOSQ can be used in general instruments conjunction with other outcomes overcome key HRQoL indicators not captured in the PCOSQ. Note that the equation is centered using a center tab stop.

C. Metformin and Polycystic Ovary Syndrome: A Literature Review

[3] Khalid A.AwartaniMD, FRCSC(Fellow)Anth ony P.CheungMBBS, MPH, FRACOG, FRCSCZ(Assistant Professor and Medical Director, IVF Program). This review identified 23 prospective studies on metformin's effects on PCOS. Only a qualitative assessment of the data was possible due to the heterogeneity of the published reports. A review of the literature confirms that metformin can help reduce insulin resistance in some PCOS women. Other beneficial biochemical effects include lower free testosterone levels and higher sex hormone-binding globulin levels (SHBG). Metformin may improve menstrual regularity, resulting in spontaneous ovulation, as well as the ovarian response to traditional ovulation-induction. However, there is little evidence to support the use of metformin to aid in weight loss, improve serum lipids, or treat hirsutism. More research is needed to determine metformin's longterm effectiveness, who will benefit from metformin treatment, and the optimal duration of metformin therapy.

D. A Literature Review on The Rising Phenomenon PCOS

[4] Vikas B1, Sarangi, Manaswini Chilla, K. Santosh Bhargav and B S Anuhya. Data mining techniques and algorithms such as clustering, classification, SVM, and the Naive Bayes algorithm have previously been used to predict heart disease and liver problems. They have also been used in the diagnosis and prognosis of breast cancer, as well as in the predictive analysis of diabetic treatment, with notable results. In this view, Data Mining can be used to work on past patient records, analyse the data, and identify general trends, as well as possible treatment solutions, to aid in the diagnosis and analysis of PCOS. The authors of this paper presented a study on PCOS symptoms and treatment. PCOS research is still in its early stages and much remains to discovered. Thus, the authors attempted investigate the various applications of data mining techniques in the medical field, which could later be extended to PCOS research.

E. Polycystic Ovary Syndrome: Important Underrecognized Cardiometabolic Risk Factor in Reproductive-Age Women

[5] Dinka Pavicic Baldani,Lana Skrgatic, and Roya Ougouag. Because PCOS appears to be dominated by metabolic consequences, both because of the condition and as a vector for further complications, such as DM2, CVD, and exacerbation of the syndrome's reproductive features hirsutism and an/oligoovulation) it is clear that research on the metabolic and cardiometabolic features of PCOS is required. The current review is a contribution to this larger effort. The sections that follow go over the cardiometabolic aspects of PCOS, their potential causes, associated risks, and potential screening measures. Many of the metabolic abnormalities seen in PCOS are exacerbated by the presence of obesity.

However, some of these metabolic disturbances occur even in lean women with PCOS and are thus rightfully recognised as being inherent to PCOS. This paper examines the intrinsic factors that cause these metabolic disturbances. Obesity and other Metabolic abnormalities are also discussed in detail. PCOS patients' metabolic disturbances cause chronic low-grade inflammation and cardiovascular impairments, which increase their risk of developing cardiovascular disease. Despite the fact that many studies have shown an increase in surrogate biomarkers of cardiovascular disease in PCOS women, it is still unclear to what extent and magnitude the increase causes more frequent and earlier events.

IV. Overview of the Proposed Work

A. Aspects of Data Visualization

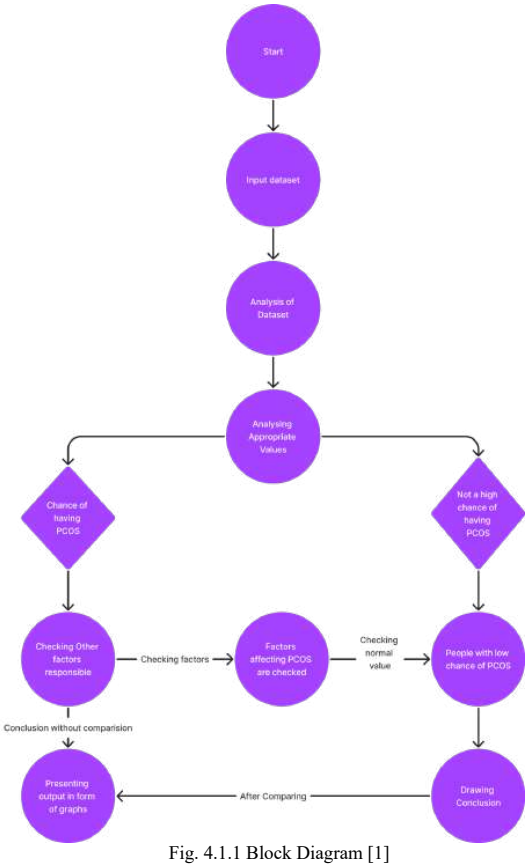

Fig. 4.1.1 Block Diagram [1]

- Import the necessary libraries for image processing and TensorFlow for the model.
- Define a function that takes an image path as an argument.
- Load the image from the given path and resize it to a 224x224 shape
- Load the image from the given path and resize it to a 224x224 shape
- We will check the predicted class of the image and print whether it is “Not Affected” or “Affected”.
- Finally, we will display the input image using matplotlib.

A flexdashboard in R Markdown is made, which comprises two tabs, each with two graphs.

The first graph depicts the association between PCOS (yes or no), age (in years), and colour (to distinguish between the two groups). A jitter plot is used to display the data, with each point representing a distinct individual. Each group's mean is shown by a red dot.

The second graph depicts the link between PCOS (yes or no), weight (in kilogrammes), and colour (to distinguish between the two groups). A jitter plot is also used to depict the data, and the red dot shows the mean of each group.

The final graph depicts the link between PCOS (yes or no), height (in centimetres), and colour (to distinguish between the two groups). A jitter plot is used to display the data, and the red dot shows the mean of each group.

The first figure depicts a pie chart of blood group distribution among PCOS patients. The pie chart is divided into eight portions, one for each blood group, and the chart shows the percentage of each blood group among PCOS patients. Each section's colours are described using hexadecimal values, and a legend is supplied.

The second plot depicts a pie chart of the percentage of obese women who have PCOS. The chart is divided into two sections: one for obese women with PCOS and one for obese women who do not have PCOS. The percentage of obese women affected by PCOS is shown in the chart, along with a legend.

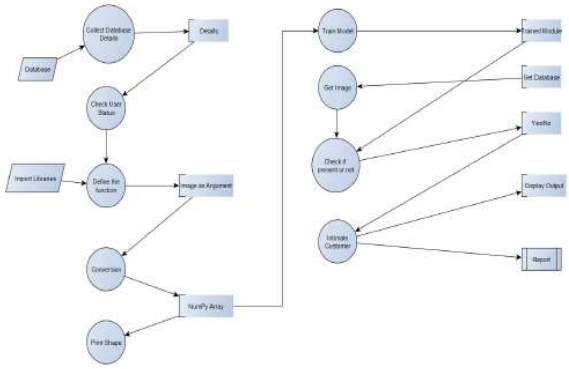

Fig. 4.2.1 Data Flow Diagram [2]

B. Aspects of Machine Learning

- Import the necessary libraries for image processing and TensorFlow for the model.
- Define the function that takes an image path as an argument
- Load the image from the given path and resize it to a 224x224 shape.
- Convert the image to a numpy array and normalize the pixel values between 0 and 1.
- Create a numpy array from the image array.
- Print the shape of the numpy array.
- Used the trained model to predict the class of the image.

- Check the predicted class of image and print whether it is “Not Affected” or “Affected”.
- Display the input image using matplotlib.

This code defines a convolutional neural network (CNN) model named "model4" using the Keras API with a TensorFlow backend. The model is designed to perform image classification on input images of size 224x224 with 3 color channels (RGB).

The model architecture consists of three convolutional layers followed by max-pooling layers to reduce the spatial dimensions of the feature maps, a flatten layer to convert the feature maps into a 1D vector, and a dense layer with softmax activation for classification. The first convolutional layer has 12 filters of size 5x5, with a "valid" padding mode to not pad the input image, and an activation function of Rectified Linear Unit (ReLU). The input shape of the layer is (224,224,3), which corresponds to the size and number of channels of the input image. The first max-pooling layer reduces the spatial dimensions of the feature maps by a factor of 4 using a 4x4 pooling window.

The second convolutional layer has 10 filters of size 5x5, with a "valid" padding mode and ReLU activation. The second max-pooling layer again reduces the spatial dimensions of the feature maps by a factor of 4 using a 4x4 pooling window.

The third convolutional layer has 8 filters of size 3x3, with a "valid" padding mode and ReLU activation. The third max-pooling layer further reduces the spatial dimensions of the feature maps by a factor of 3 using a 3x3 pooling window.

The flatten layer converts the 3D tensor of feature maps into a 1D vector for input to the dense layer.

The dense layer has 2 output units, corresponding to the two classes of images that the model is designed to classify. The softmax activation function ensures that the outputs are normalized and can be interpreted as probabilities.

C. Aspects of Regression

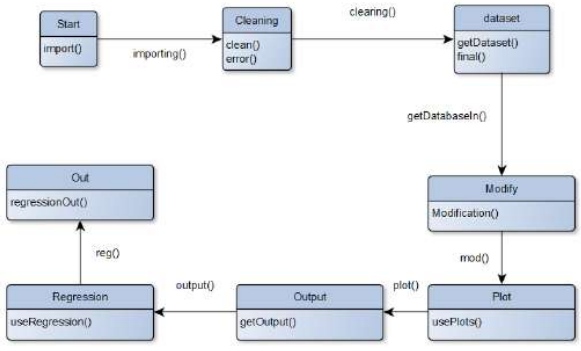

Fig. 5.1.1 State Transition Diagram of Regression [3]

Using the sklearn library, the fit method was used to train a LogisticRegression model using the training data (X train and y train). The model.score(X train, y train) command assessed the model's accuracy by measuring the proportion of correctly identified training data instances. Before evaluating the model with unknown test data, it is possible to determine whether the model overfits or underfits the training data. If a model's accuracy is high on training data but low on test data, it has been overfit and does not generalise well. Testing the model on both training and test datasets determines its ability to generalise to new data and produce accurate predictions. This is essential for the development of machine learning models that can predict new data.

V. Result and Discusssions

A. Aspects of Data Visualization

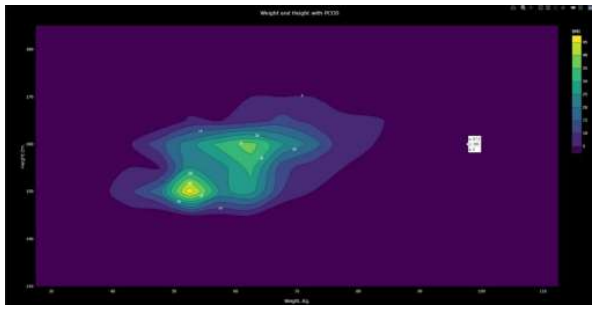

Fig. 5.1.1 Counter Diagram [4]

The first graph is a scatter plot that compares the attributes of age, weight, and height of people with PCOS, showing that weight increases as x values increase.

The second graph is a violin plot that combines a box plot and a kernel density plot to show that women with PCOS tend to have larger hip and waist sizes compared to those without PCOS.

The third graph is a distplot that displays the patterns of weight, haemoglobin levels, respiratory rate, and pulse rate, allowing us to compare their correlations and respective means, medians, and modes, with PCOS in mind.

The general heatmap of weight, height, and BMI levels with people having PCOS shows user behavior on specific attributes, with counter lines. [5.1.1]

Under-looked reports of PCOS patients, such as BP systolic, BP diastolic, and more, are often overlooked. We can compare them using medians, means, and sums.

The Progesterone Serum Levels (PSG Levels) are directly proportional to waist and hip inch sizes, as shown in the 3D scatterplot that compares them.

The 3D bubble chart compares the height, weight, and age of PCOS patients simultaneously, using blood groups as the color difference and the size of the bubble to depict BMI levels.

The parallel coordinate plot connects individual attributes or observations, linking weight, height, and BMI of PCOS patients to find trends among them, with lighter colors representing older patients.

The 3D line chart shows the relation between the density of PCOS patients with respect to blood groups and blood haemoglobin levels, establishing the probability of PCOS for different blood groups with respect to their haemoglobin levels, with BG-18 (AB -ve) having the highest density of PCOS patients.

The interactive table allows us to view and sort the dataset according to our needs, making it easier to find specific data, such as women with PCOS who weigh over 60 KGs.

B. Aspects of Machine Learning

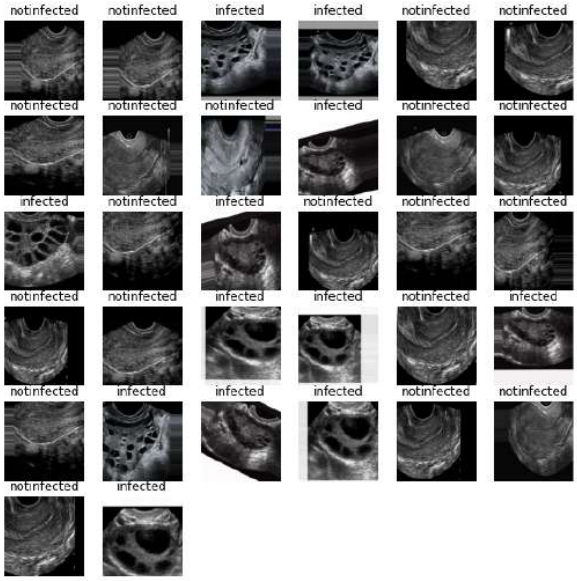

Fig .5.2.1 Affected/Not Affected [4]

This code generates a 6x6 grid of subplots, each displaying an image from the training dataset and its accompanying label. The plt.figure(figsize=(10, 10)) command generates a new 10x10-inch figure. The class names variable is defined as a list of the dataset's class names. The outer for loop iterates through two batches of images and labels from the train ds dataset. The inner for loop iterates for 32 times, representing the 32 photos in each batch.The command axe = plt.subplot(6, 6, I + 1) generates a new subplot on the grid with the row index I / 6, column index i% 6, and subplot number i+1. The image in the subplot is displayed with the plt.imshow(images[i].numpy().astype("uint8") command. The image tensor is converted to a numpy array using the numpy() technique, and the pixel values are converted to 8-bit unsigned integers using astype("uint8"), which is the desired format for showing images with imshow (). The plt.title(class names[int(labels[i])]) command changes the title of the subplot to the image's label name. Lastly, the plt.axis("off") command disables the subplot's axis labels and ticks.

C. Aspects of Regression

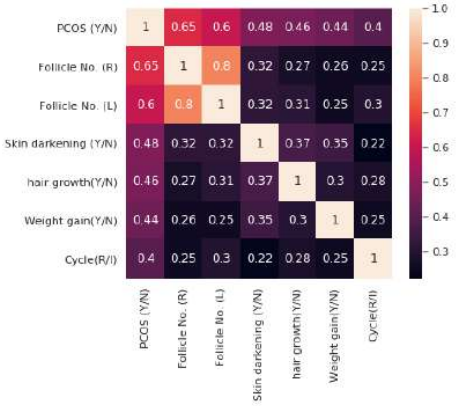

Fig. 6.1.1 Heatmap [4]

This investigation examines the dataset performance of a machine learning model. To visualise dataset correlations, the seaborn library was used to generate a heatmap of the correlation matrix of the "df noinf dataframe". Using plt.figure(figsize=(6,5)) and matplotlib.pyplot with a (width, height) tuple in inches, the heatmap was displayed.

On a dataset, the machine learning model was then evaluated. "model.score(X test,y test)" displayed the model score for test data. X test and y test represent the characteristics and target values, respectively, of the test data. model is a sklearn-trained model.

A confusion matrix was constructed using "sklearn.metrics.confusion" matrix to evaluate the machine learning model. The perplexity matrix was generated using the test data's actual target values (y test) and the model's predicted target values (y pred). The total number of correct classifications was calculated by adding the diagonal elements of the confusion matrix (p right=cm[0][0]+cm[1][1]) and the total number of incorrect classifications was calculated by adding the off-diagonal elements (p wrong=cm[0][1]+cm[1][0]). Published were analyses. print(f"Right classification: p right") and print(f"Wrong classification: p wrong") printed the number of instances of test data that were correctly and incorrectly categorised, respectively. The confusion matrix was displayed via cm tables.

This study demonstrates how the machine learning model performed on the evaluation dataset and can inform decisions regarding model refinement or optimisation.

VI. Conclusion

In conclusion, the above aspects demonstrate how to build and train different CNN models for image classification using TensorFlow and Keras. The other method is by using Data Visualization. The results show that the simplest model (model4) with the fewest parameters can achieve high accuracy when augmented training data is used. However, the effectiveness of a model can also depend on the specific problem and dataset, and it is recommended to experiment with different architectures and hyperparameters to find the best model for a given task. Hopefully, this project can help the Detection and Analysis of PCOS easier.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.