Math, Science, and Health: Analysis on Women Health Issues: An Algorithmic Perspective

Shivani Ramakrishnan[1], Fouzia Fathima A.M[2], Jagatheeswari M[3]
[1,2,3]*Student, CEG, Anna University, Chennai, India*

Abstract **- There has been a lack of research and detailed studies on women's health and diseases that predominantly affect them. In this paper, we aim to examine certain conditions which mainly affect women and identify the factors which play a significant role in their occurrence. For this study, diseases such as Anemia, breast cancer, and Polycystic ovary syndrome (PCOS) were considered. Machine learning models and algorithms such as XGboost and Random Forest Classifier were used to find top causes.**
Model performance and results were examined for each of the diseases, and data interpretation was made with the help of SHAP plots. The results showed that a focused study on women's health yielded less biased and accurate results.

Index Terms **- Predictive Analysis, Cancer, Anemia, PCOS, Machine Learning.**

I.INTRODUCTION

Women frequently spend so much time helping others that they neglect to take care of themselves. It is acceptable to devote time to improving one's health. Because of biological and gender differences, being a man or a woman has a significant impact on health. Women's and girls' health is of particular concern because they face discrimination based on sociocultural factors in many societies. Some of the sociocultural factors that prevent women and girls from receiving quality health care and reaching their full potential include:

- Social norms that limit education and paid employment opportunities; an exclusive focus on women's reproductive roles; and
- The potential or actual experience of physical, sexual, and emotional violence.

Our research aims to identify the factors that are important in the diagnosis/occurrence of diseases such as anemia, breast cancer, and PCOS.

II. BACKGROUND

Gender differences in disease susceptibility, symptoms, and response to treatment exist in many areas of health, and this is especially true when viewed globally. Cancer, reproductive health, maternal health, human immunodeficiency virus (HIV), sexually transmitted infections, violence, mental health, non-communicable diseases, youth, and aging were identified as the top ten issues in women's health by the World Health Organization in 2015. Some of the significant factors that will be discussed in this paper are as follows:

Anemia is a condition in which a person does not have enough healthy red blood cells to transport enough oxygen to the body's tissues due to a low hemoglobin concentration. Iron deficiency is more common in women because iron loss in the blood can occur as a result of heavy menstruation or pregnancy.

Breast cancer, also known as breast malignancy, is one of the most common diseases found in women of all ages, ranging from 20 to 50 years old, and is more common in women over the age of 50. Breast cancer develops when a mass of tissue grows abnormally, resulting in the expansion of malignant cells and the development of acute breast cancer.

Polycystic ovary syndrome is a condition characterized by infrequent, irregular, or prolonged menstrual periods and often elevated levels of male hormone (androgen) and is seen among women of reproductive age [1]. The ovaries form numerous small fluid collections known as follicles and may fail to release eggs on a regular basis.

III. METHODOLOGY

A dataset must undergo many phases before it can be utilized to construct an acceptable machine learning model. The algorithm's input must be filtered and

noise-free. The algorithms' outputs are compared and reviewed, with the objective of employing a machine learning model to detect women-centric health concerns with the greatest accuracy. The Methodology's block diagram is shown in fig.1.

III. MODEL SELECTION

Anemia: - Anemia was studied using data from the National Health and Nutrition Examination Survey (NHANES), a successful program of the National Centre for Health Statistics (NCHS), from 2013 to 2018. It contained various data types and demographics, dietary, examination, laboratory and questionnaire datasets. Anemia prediction was done by analyzing hemoglobin levels, an indicator of iron levels. 'LBXHGB'. Hemoglobin count was used from the laboratory dataset. Recursive feature elimination (RFE) was a method used to select important features from demographic, examination, and laboratory data with around 200 features. Ensemble learning methods can be used for accurate analysis [2].XGBoost, an ensemble method, was used for prediction, and gradient boosting produced promising results for this classification problem. Hyperparameter tuning was performed using OPTUNA, a Bayesian-based hyperparameter optimization framework. Stratified K-fold cross-validation, which is a procedure used to estimate the model performance on new data, was incorporated. A hyperparameter called scaled pos weight was added to address the class imbalance and was designed to tune the algorithm for imbalanced classification problems. Based on the primary evaluation metric-f1 score, the XGboost hyperparameters were tuned in order of importance groups.

Breast Cancer: - Disease prediction was done using the dataset provided by Wisconsin. It contains the relative sizes of the cell nucleus such as radius, perimeter, symmetry and concavity. The shape of the dataset is (569,33). There are 357 cases of Benign and 212 cases of Malignant. Feature selection was done by using the Principal Component Analysis technique which removed the features that had multicollinearity. Feature scaling was done by using boxplots to normalize the range of data. Ensemble learning can be incorporated to improve the accuracy of breast cancer prediction [3]. XGBoost, an ensemble method, is a machine learning tool that uses distributed gradient boosting to maximize model accuracy. This model had an accuracy of 96.07 percent. The XGBoost Classifier algorithm produced F1 scores that were adequate for both Malignant and Benign cases. SHAP plots were used to identify the top 3 essential features. They were radius_mean, texture_mean, compactness_mean.

PCOS: - Random forests are a type of ensemble learning tool used for classification or regression. Individual decision trees are created for the training set, and classification or regression is generated separately for each tree. The dataset contained all physical and clinical parameters to determine PCOS and infertility-related issues. Data were collected from 10 different hospitals across Kerala, India. Data were preprocessed by removing null values and replacing them with mean or mode based on the variable. All the outliers were removed by analyzing box plots of top features. Random Forest Classifier was used, which has the ability to correct for undesirable properties of decision trees to overfit training data making it a more efficient and reliable model. Women are likely to have PCOS under the following conditions mentioned: Age group between 25 to 35. Weight seems to have a positive correlation with PCOS. Order of Feature Importance: Follicle No. (L), Follicle No. (R), Hair Growth, Cycle Length, Weight Gain and Skin Darkening.

Fig.1: - Methodology Block Diagram

IV. MODEL RESULT

Dataset	*Model Used*	*Metric*
Anemia	XGBoost	F1 score- 0.86
Breast Cancer	XGBoost	Accuracy-0.96
PCOS	Random Forest Classifier	Accuracy - 0.90

V. SHAP VALUES

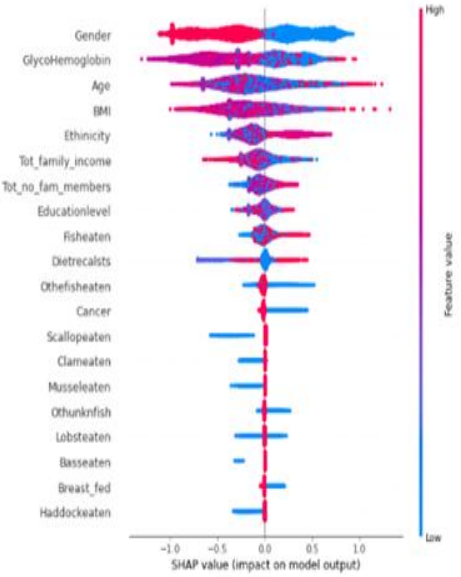

FIG 2:- SHAP PLOT FOR ANEMIA DETECTION USING XGBOOST

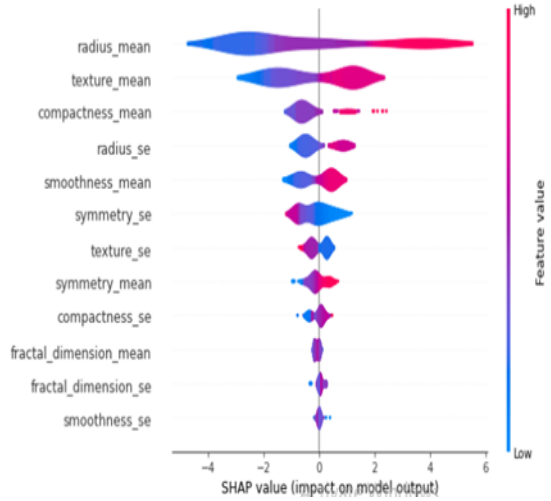

FIG 3: - SHAP PLOT FOR BREAST CANCER USING XGBOOST

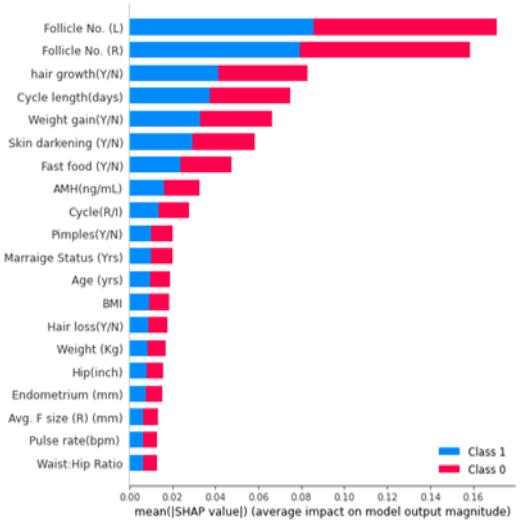

FIG 4: - SHAP PLOT FOR PCOS DETECTION USING RANDOM FOREST CLASSIFIER

VI.CONCLUSION

When diet features are taken into account, a seafood-based diet has a significant impact on Anemia prediction because many of the top features include seafood such as scallops, bass, mussels, and oysters. Anemia is more likely to occur when Glycohemoglobin levels are low. It has also been observed that the lower the weight, the greater the likelihood of Anemia. XGboost proved to be the best performing model for anemia prediction with an f1 score of 0.86.

A person's dietary intake has a significant impact on anemia diagnosis and keeping track of dietary habits regularly, significantly reduces the risk of disease.

Women with PCOS commonly experience excessive weight gain, facial hair growth, acne, hair loss, skin darkening, and irregular periods, leading to infertility in rare cases. Our proposed system aids in the early detection and prediction of PCOS treatment based on an optimal and minimal set of statistically analyzed parameters. The Random Forest Classifier was the most dependable and accurate, with an accuracy of 90%.

The proposed system can be used by both patients and doctors, as a doctor can screen new patients for basic information and prioritize treating patients with PCOS first, followed by patients who do not have PCOS.

Breast cancer is a deadly disease that affects women over the age of 50. An early prognosis is required to reduce disease mortality. Breast cancer usually has

few symptoms, but it can cause lumps and inverted nipples in some people. The study used tumor size and other characteristics and appropriate machine learning methods to find the most accurate model for predicting breast cancer. With a 96.07 percent accuracy, the XGBoost Classifier produced the best results. According to this model, the radius means and compactness mean features contribute the most to cancer prediction.

VII. ACKNOWLEDGMENT

1. Prasoon Kottarathil, Polycystic ovary syndrome (PCOS), 2020, Kaggle.
2. Wisconsin Breast Cancer Dataset by UCI Machine learning repository, Kaggle.
3. National Health and Nutrition Examination Survey by National Center for Health Statistics (NCHS), Kaggle.

REFERENCES

[1] A. Denny, A. Raj, A. Ashok, C. M. Ram and R. George, "i-HOPE: Detection and Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 673-678.

[2] P. T. Dalvi and N. Vernekar, "Anemia detection using ensemble learning techniques and statistical models," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016, pp. 1747-1751.

[3] Dongxiao Gu, Kaixiang Su, Huimin Zhaob(2020): A case-based ensemble learning system for explainable breast cancer recurrence prediction. 'Artificial Intelligence in Medicine,' Volume 107,101858. Elsevier, ISSN 0933-365.