https://doi.org/10.46344/JBINO.2025.v14i02.05

MULTIVARIATE APPROACH TO INVESTIGATE RISK FACTORS PREDISPOSING UNDER-FIVE TO SEVERE BRONCHO-PNEUMONIA

Ahmed Sule Shehu*, Bilkisu Maijama'a * and Monday Osagie Adenomon *

*Department of Statistics, Nasarawa State University, Keffi, Nasarawa State

maijamaab@nsuk.edu.ng

ABSTRACT

Severe broncho-pneumonia is a leading cause of morbidity and mortality among under-five children worldwide. Identifying risk factors is crucial for targeted interventions. This work applies Principal Component Analysisand Discriminant Analysis to identify therisk factors and predict the prevalence of Broncho-Pneumonia status in under-fivechildren. The data used in this study were collected from health institutions in Minna, Niger state. Predictorvariables which are well-recognized for characterizing broncho-pneumonia were considered, these includes; Weight of Baby at birth,Weightof Baby after a month, Genderof Baby,Age of Mother, Occupation ofMother, Socioeconomic Status, Weather Condition, Nutritional Status, Genetics, Attitudinal Family Background.According to the outcome of other related researchers who conducted comparable studies shows thatthe study predicted the BPn status of new infants using the discriminant model with more than 80% correctly classified. The new datasets were used to validate the model. Researchers also disclosed that baby's weight at birth is best at discriminating between the two groups, since it has the least value of Wilk's Lambda compare to other predictor variables.

Keywords: Investigate Risk Factors, Broncho-Pneumonia, Principal Component Analysis, Predisposing and Discriminant Analysis.

1. INTRODUCTION

Multivariate analysis is concerned with the interrelationships among several variables. The data may be metrical, categorical, or a mixture of the two. Multivariate data may be, first, summarized by looking at the pair-wise associations. Beyond that, the different methods available are designed to explore and elucidate different features of the data. The article briefly summarizes the scope and purpose of the following methods: cluster analysis, multidimensional scaling, principal components analysis, latent class analysis, latent profile analysis, latent trait analysis, factor analysis, regression analysis, discriminant analysis, path analysis, correspondence analysis, multilevel analysis, and structural equation analysis (Batholomew, 2010).

In clinical situations, the status of a patient is assessed by the presence or absence of a disease. There are many factors to consider which may or may not correlate with the incidence of the disease. There has been numerous retrospective medical research studies published each year that review past medical records and charts of former patients to help determine some of the risk factors (or causing agents) of diseases that are of interest. Finding the risk factors and the potential risk factors can help to prevent the development of the disease. All of the diseases nearly all of the risk factors considered are categorical variables (variables taking on two or more possible values).

Data analytics is all about looking at various factors to see how they impact certain situations and outcomes. However, if dealing with data that is more than two variables, the Multivariate analysis would be best used. For instance, in marketing, you might verify at how the variable "money spend on advertising" impacts the variable "number of sales." (Emily, 2023).

Multivariate Analysis of Covariance (MANCOVA) is a combination of the MANOVA model with a multivariate regression model with one or more predictors. The inclusion of these covariates leads to noise reduction in the sense that the variance associated with the covariates is removed from the error variance, thus providing a more powerful test for the difference between the groups. (Bryan, 2023).

Michael *et al*, 2022 explain the Principal Component Analysis (PCA) as a multivariate statistical method that combines the in-formation from several variables observed on the same subjects into fewer variables, called principal components (PCs). "Information" is measured by the total variance of the original variables, and the PCs optimally account for the major part of that variance. The PCs have geometric properties that allow for an intuitive and structured interpretation of the main features inherent in a complex multivariate dataset.

Principal Component Analysis (PCA) is a powerful technique used in data analysis, particularly for reducing the dimensionality of datasets while preserving crucial information. It does this by transforming the original variables into a set of new, uncorrelated variables called principal components. Key aspects of PCA are; Dimensionality

Reduction, Data Exploration and Visualization, Linear Transformation, Feature Selection, Data Compression, Clustering and Classification, Matrix Requirements, Eigenvalues and Eigenvectors, Number of Components etc.

PCA is a method of transforming a set of correlated variables into a smaller set of uncorrelated variables, called principal components (PCs). Each Principal Component represents a linear combination of the original variable, weighted by their contribution to the variance of the data. The first PC captures the most variance, the second PC captures the most variance among the remaining variables, and so on. (Brennan, 2024).

Principal Component Analysis (PCA) computation is basically an eigenvalue-eigenvector problem, solvable using either correlation or covariance matrices. If multiple variables measure the same underlying construct, researchers often seek to consolidate these variables into a smaller set of combined variables, or "super variables," to simplify analysis and reveal latent patterns. (Akash, 2018).

Discriminant analysis (DA) is a multivariate statistical techniques designed to classify members of two or more normal populations into two or more groups based on the observed information to them. It is also concerned with separating distinct sets of objects and with allocating new objects to previously defined groups. (Usman, 2023).

The objectives of discriminant analysis is to; determine the relative importance of each variable (p) in differentiating between groups, identify the prime projection plane to illustrate the relationships and configurations among groups. Another goal is the classification and prediction: Employ linear functions of variables to; assign new observations to the existing groups, also allocate sampling units to their appropriate groups.

Pneumonia is defined as an acute inflammation of the Lungs' parenchymal structure. It is a major public health problem and the leading cause of morbidity and mortality in under-five children especially in developing countries. (Biruk, 2020).

Anthony, 2010 described Pneumonia as an illness, usually caused by infection, in which the lungs become inflamed and congested, reducing oxygen exchange and leading to cough and breathlessness. It affects individuals of all ages but occurs most frequently in children and the elder.

According to Cologne, 2021. Pneumonia can be identify as an inflammation of the air sacs in the lungs (alveoli) and the surrounding tissue. It often leads to a sudden high fever, the feeling that you are very unwell, a cough and shortness of breath.

The broncho-pneumonia pattern has been associated with hospital acquired pneumonia, and with specific organisms' *staphylococcus aureus, klebsiella pneumoniae subsp and pseudomonas aeruginosa*.In bacterial pneumonia, invasion of the lung parenchyma by bacteria produces an inflammatory immune response. This response leads to a filling of the alveolar sacs with exudates. The loss of air space and its

replacement with fluid is called consolidation (Janelle, 2019).

Bronchial pneumonia is among the top causes of death in infants. Despite being one of the largest killer of children under five, this disease receives disproportionately limited global funding and resources. With 1.8 million child fatalities annually, the disease deserves far greater global investment and action. The disturbing disparity between the disease's devastating impact and the relatively meager resources allocated to combat it demands change. (Global Action Plan for Prevention and Control of Pneumonia, 2009). This disease causes about 15% of majority of deaths in under age 5 children worldwide, while, 2% of which are new-born (Janelle and Rachel, 2017). Bronchial-pneumonia disproportionately affects infants due to their immature respiratory immune system, making them more susceptible to infection.Bronchopneumonia is a typically triggered by bacterial lung infections, most notably;*Streptococcus pneumoniae* and *Haemophilus influenza type b (Hib)*. Viral and fungal lung infections can also cause pneumonia (Aaron, 2018).
Low Birth Weight (LBW) is described as a birth weight of a live born infant of less than 2,500g (5 pounds 8 ounces) regardless of gestational age. Subcategories include; Very Low Birth Weight (VLBW) which is less than 1500g (3 pounds 5 ounces) and ExtremelyLowBirth Weight (ELBW) which is less than 1000g (2 pounds 3 ounces). Normal Weight at term of delivery is 2500g - 4200g (5 pounds 8 ounces – 9 pounds 4 ounces). Most normal babies weigh 5.5 pounds by 37 weeks of gestation. (Krasevec, 2022)

In strict terms, the application of statistical techniques to biological and medical data is called Biostatistics. Generally speaking, bio-statistical methods are relevant in virtually every branch of applied medicine, pharmacy, nutrition and public health. They come into play either when we have a medical theory to test or when we have a relationship in mind that has some importance for medical decision or policy analysis in public health. Bio-statistical methods in medicine are more or less empirical analysis using data to test a theory or to estimate a relationship in medicine, pharmacy, public health and other areas.

2. Statement of the Problem

Baby weight less than 2.5kg is considered as Low Birth Weight (LBW). It remains a significant public health problem in both developed and developing countries. These infants with LBW encounter greater neonatal morbidity and mortality and significantly higher rates of physical and mental handicaps later in life (Pope, 2010). Taking the infants population globally, the proportion of babies with a LBW is an indicator of a multifaceted public-health problem that includes the sex of an infant as well as the birth weight and weight four weeks after birth. Also the mother's age and mother's occupation are important variables that could predict the pneumonia status in infant in the study. Baby with bronchopneumonia may have trouble breathing because their airways are constricted, due to inflammation, their lungs may not get enough air. Therefore, the main problem which comes up in this particular study is how to identify the risk variables and develop linear discriminant

models that is capable of predicting the Broncho-Pneumonia (BPn) status of the infant using the critical variables as predictor variables. However, since the model comprise classification, it is in the interest of the researcher to classify some infants as normal of Broncho-Pneumonia (BPn) patients using the developed model. Hence, a suitable prediction model will be constructed to satisfy the best methods of validation as well as diagnostics of statistical decisions.

The aim of this study is to investigate the critical factorspredisposing under-fives to severe Broncho-Pneumonia

3. LITERATURE REVIEW

Acute Respiratory Tract Infections (ARIs), primarily pneumonia are a leading cause of illness and death among infant in developing countries.ARIs encompass infections affecting any part of the respiratory system, including: Upper respiratory tract (i.e. Nose, Throat, and Larynx). Lower respiratory tract (such as trachea, bronchi, and lungs). Related structures (Para-nasal sinuses, Middle ear and Pleural cavity) Bipin *et al*, 2011. According to Wonodi *et al*, (2012) that carried out research on the Pneumonia Etiology Research for Child Health (PERCH) to assess the risk factors for severe pneumonia in hospitalized children at 7 sites. They identified relevant risk factors by literature review and iterative expert consultation. The variables identified are; Demography, Birth milestone, Nutritional status, Treatment interventions, Maternal characteristics and Family environment. They aimed to standardize questions at all sites, but significant variation in the economic, cultural, and geographic characteristics of sites made it difficult to obtain this objective.

Yakubu et al. (2019) conducted a study to investigate the impact of significant risk factors on infant bronchopneumonia. The researchers employed statistical modeling to identify the most parsimonious model with the fewest parameters. The study utilized a random sample of 433 births from two tertiary health institutions in north-central Nigeria. The sample consisted of mothers recruited from the specific hospitals/health institutions. The study's results provide valuable insights into the risk factors associated with infant bronchopneumonia in north-central Nigeria, informingstrategiesforpreventiona ndintervention.

Anna *et al* (2014), they used PCA method in reduction of the number of studied variables with the maintenance of as much information as possible, and also as a first step in analyzing data from IVF (in vitro fertilization). The next step and main purpose of the analysis was to create models that predict pregnancy.

Mahdi *et al* (2020), in their research on cancer diseases using discriminant method found that Gender has the highest discriminating power, whereas the Grade variable has the least discriminatory power. It was discovered that, the probability of correct classification reached 92%, followed by the second group with brain cancer, where the probability of correct classification was 64%. Finally, the first group with bladder cancer had the lowest probability of correct classification. They concluded that, increasing the

sample size has a significant impact on the correct classification of observations.Bronchopneumonia disproportionately affects infants due to their immature respiratory immune system. According to Aaron (2018), the primary causes of bronchopneumonia include;Streptococcus pneumoniae and Haemophilus influenza type b (Hib). Viral and fungal lung infections can also cause pneumonia

Beki (2012), employed advanced statistical techniques to predict the incidence of Bronchopulmonary Dysplasia (BPD) among infants. The study utilized; Discriminant Analysis to identify significant predictors of BPD and Binary Logistic Regression to model the probability of BPD occurrence. This study aimed to develop a predictive model for BPD, a chronic lung disease affecting premature infants. The research analyzed specify sample size. The researcher used three possible predictor variables i.e. weight at birth, weight four weeks later and gender and built a discriminant model that is capable of tracking Broncho-Pulmonary Dysplasia (BPD) infants.

Principal component analysis (PCA) and logistic regression can help determine relevant risk factors and identify LN patients at high risk of hypothyroidism; as such, these tools may prove useful in managing this disease. Our PCA–logistic regression analysis results demonstrated that serum creatinine, blood urea nitrogen, blood uric acid, total protein, albumin, and anti-ribonucleoprotein antibody were important clinical variables for LN patients with hypothyroidism. The area under the curve of this model was 0.855 (Ting, 2020). Danbaba *et al* (2013), carried out a research on low birth weight using logistic regression analysis to determine the prevalence of Low Birth Weight (LBW) and some of its risk factors in maternity hospitals in Wushishi Local Government of Niger State. Data from a sample of 200 live births were collected in the hospital from June – September 2011. The data were collected by obtaining the mother's age at birth, mother's weight at birth, mother's education level, mother's occupation, gestational age, birth interval, twin or singleton birth and parity. The study fitted the logistic regression model to the data.

Nahida *et al* (2020), in their research proposes a fusion of Deep Convolutional Neural Network Model with Principal Component Analysis (PCA) feature extraction model and Logistic Regression (LR) classifiers for the diagnosis of pneumonia from chest X-ray images. In this study, fine-tuned pre-trained CheXNet model is used as Convolutional Neural Network (CNN) model on standard pneumonia dataset collected from Guangzhou Women and Children's Medical Center, Guangzhou.

Clement 2022, carried out a research on the use of machine learning systems to detect respiratory diseases via non-invasive measures such as physical and laboratory parameters is gaining momentum and has been proposed to decrease diagnostic uncertainty associated with bacterial pneumonia. Hence, the researchers conducted several experiments using eight machine learning models to predict pneumonia based on biomarkers, laboratory

parameters, and physical features. Vitmalkumar *et al,* (2011) conducted a study to determine the prevalent risk factors of Nephropathy in type-2 diabetic patients. A tertiary hospital was used for the study aimed to build a binary logistic model for predicting Nephropathy status among type-2 diabetic patients using age, sex, socio-economic status, and duration of Nephropathy history as covariates.

Vishwa *et al* (2015) stated that discriminant analysis and classification are multivariate techniques concern with separating district sets of objects (or observations) and with allocating new objects (or observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a classificatory procedure, it is often employed on a one-time basis in order to investigate observed differences when casual relationships are not well understood.

Shibuya *et al,* (2013) investigated the use of high magnification bronchovideoscopy combined with narrow band imaging (NBI) for the detailed examination of Angiogenic Squamous Dysplasia (ASD). This was carried out in relation to bronchial vascular patterns with abnormal mucosal fluorescence in heavy smokers at high risk for lung cancer. Forty eight patients with sputum cytology specimens suspicious or positive for malignancy were entered into the study.

Fernandez *et al,* (2010) used a discriminant analysis to investigate whether FT – Raman spectroscopy as spectroscopic fingerprint techniques combined with some chemo metric tools can be used as a rapid and reliable method for the discrimination of honey according to their sources.

Evans, (2014): on the analysis of BPD and prediction for extremely premature infants used Support Vector Machine (SVM) algorithm implemented in LIBSVM. The results was compared with others gathered in previous work. Fourteen different risk factor parameters were considered and due to the high computational complexity only 3375 random combinations were analyzed. Usman et al. (2012) carried out a statistical analysis utilizing NCSS and GESS 2007 software to investigate the relationship between crimes against persons and properties. The study employed Scree plot to determine the optimal number of factors and Loading plot to visualize factor loadings. From the outcome, three components have been retained and also indicate that correlation existbetween crimes against persons and crime agaist properties. . Mobil*et al,* (2010) used both principal component Analysis (PCA) and Partial Least Square– Discriminant Analysis (PLS-DA) in the analysis and interpretation of the Raman spectra collected from microorganism of different species recorded in the spectral range of 2000 to 200 cm^{-1}. To develop a classification rule, the researcher used PLS-DA in a Leave-One-Out Cross Validation (LOOCV) method for the calibration and validation of a classification model. Pepke *et al,* (2011) used a supervised micro-calcifications based on Fisher's Linear discriminant analysis by methodological approach of breast density which allow them to identify micro-calcifications even in difficult cases (i.e when there is not high

contrast between the micro- calcification and the sound.

Emin *et al,* (2014) conducted a research to find out the relationship between diabetic nephropathy and Visceral Adipose Tissue (VAT). The Neutrophil-to-Lymphocyte Ratio (NLR) and Platelet-To-Lymphocyte Ratio (PLR) are simple, inexpensive, and useful markers to determine inflammation. The research aim was to investigate the association between diabetic nephropathy, NLR, and PLR as inflammatory markers. The methods used were a cross-sectional study involving 200 diabetic patients.

4. RESEAERCH METHODOLOGY

In classification design, the researcher is not interested in a mere collection of facts but model would be used to classify the Pneumonia status of an infant whose status is not known earlier.

However, the major statistical components form the basis of the research design which includes both the sampling plan and the modeling procedures. The sampling plan is the methodology used for selecting the sample from the population. The modeling procedure is the algorithms or formulae used for obtaining model of population values from the sample data and for estimating the reliability of these model.

4.1 Technique for Data Analysis and Model Specification:

In this study, Principal Component Analysisand Discriminant Analysis would be used as Data Analysis Technique**,** The data will be categorized as; infant weight < 1.00 (kg) coded as 0, between 1.00 – 1.99 (kg) as 1 and ≥ 2.0 (kg) as 2 for weight at birth. Weight four weeks after birth coded as 0 for < 2.0 weight, 1 for weight between 2.00 – 2.99 (kg) and 2 coded for weight ≥ 3.00 (kg). However, the gender will be labeled as 0 for the male infant and 1 for female infant. Mother's age takes 0 for age ≤ 18, 1 for the range of 19-29 age and ≥ 30 coded as 2. In Mother's occupation, House wife is coded 0, Civil servant as 1 and 2 for Business mother. Health Status will be coded as 0 for Healthy and 1 for Unhealthy.

The predictor variables outlined above have been authenticated as reliable indicators of pneumonia in infants. Clinical observations and empirical evidence from medical practice suggest that these factors display substantial variability between infants with and without pneumonia.

In this research, we proposed to use mixed research method, i.e primary and secondary source data that would be carefully and technically extracted directly from the individual client's medical folder from the randomly selected health institutions within Minna, Niger state. The baby's weight at birth (kg), weight of baby after a month (kg), baby's gender, Age ofMother, Occupationof Mother, Socioeconomic Status, Weather Condition, Nutritional Status, Genetics, Attitudinal Family Background and other related variables will be collected and tabulated as independent variables.

It will be necessary to outline the framework of the major components involved in the sampling design and data collection procedures adopted in this research.

4.2 Model Specification:

4.2.1 Principal Component Analysis(PCA)

Principal Component Analysis (PCA) will be employed as a dimensionality reduction technique to transform a set of correlated variables into a new set of linearly uncorrelated variables, known as principal components, through an orthogonal transformation. This process will; identify patterns and relationships within the data, reduce multidimensional data to lower dimensions and reveal the underlying variance-covariance structure

PCA computes an orthogonal set of principal components, maximizing variance explanation. The components are ordered by eigenvalues, ensuring the first few capture the most variability. Unlike Factor Analysis, PCA yields a deterministic solution, unaffected by rotational transformations, apart from sign ambiguity.

Let $X = (X1, X2, \ldots, Xp)$ be a vector of p random variables. The goal of Principal Component Transformation (PCT) is to identify a smaller set of derived variables, say k (k < p), that capture most of the information contained in the variance of the original variables.

Suppose the random vector $X' = X_1, X_2, X_3, ---, X_p$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \lambda_p \geq 0$ consider the linear combinations:

$$Y_j = \alpha_j' X = \alpha_{j1}X_1 + \alpha_{j2}X_2 + \alpha_{j3}X_3 + \cdots + \alpha_{jp}X_p = \sum_{k=1}^{p} \alpha_{jk}X_p \quad (1)$$

Such that $j = 1, 2, \ldots, p$ are the indices of the elements of X and $\alpha_{j1}, \alpha_{j2}, \alpha_{j3}, \ldots, \alpha_{jp}$ are the components of α_j vector of p^{th} term.

Then, $Var(Y_j) = \alpha_j' \Sigma a_j$

$j = 1, 2, \ldots, p$ (2)

$Cov(Y_j, Y_k) = \alpha_j' \Sigma a_k \quad j = 1, 2, \ldots, p$ (3)

The principal components are those uncorrelated linear combinations $Y_1, Y_2, Y_3, \ldots, Y_p$ whose variances in (2) are as large as possible. In finding the PCs we concentrate on the variances. The first step is to look for a linear combination $\alpha' X$ with maximum variance, so that

$$\alpha_1' X = \alpha_{11}X_1 + \alpha_{12}X_2 + \alpha_{13}X_3 + \cdots + \alpha_{1p}X_p = \sum_{k=1}^{p} \alpha_{1k}X_p \quad (4)$$

Then, we look for the linear combination $\alpha_2' X$ uncorrelated with $\alpha_1' X$ having maximum variance and so on, hence at the k^{th} stage a linear combination $\alpha_k' X$ is found that has maximum variance subject to being uncorrelated with $\alpha_1' X$, $\alpha_2' X$, $\alpha_3' X$, …, $\alpha_{k-1}' X$. The k^{th} derived variable $\alpha_k' X$ is the k^{th} principal component. Up to p principal components can be found, but we would hope to stop after the q^{th} stage as $(q \leq p)$, i.e. when almost all of the variation in X would have been explained by q^{th} principal components.

Note:

The variance of the i^{th} principal component is equivalent to its correspondingeigenvalue, λ_i

$V(Y_j) = \alpha_j' \Sigma a_j = \lambda_j$

$j = 1, 2, 3, \ldots, p$ (5)

The total variance in data set is equal to the total variance of principal components

$$\sigma_{11} + \sigma_{22} + \sigma_{33} + \cdots + \sigma_{pp} = \sum_{j=1}^{p} Var(X_j)$$

$$= \lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_p = \sum_{j=1}^{p} Var(Y_j) \quad (6)$$

Data standardization will be applied to normalize variables to a common scale, achieving zero mean $(\mu = 0)$ and unit standard deviation$(\sigma = 1)$. For a random vector $X' = [X_1, X_2, X_3, \ldots, X_p]$ the corresponding standardized variables are

$z = \left[Z = \frac{(X_j - \mu_j)}{\sqrt{\sigma_j}}\right]$ for $j = 1, 2, 3, \ldots, p$ in matrix notation (7)

$$Z = \left(\theta^{1/2}\right)^{-1}(X - \mu), \quad (8)$$

Where $\theta^{1/2}$ is the diagonal standard deviation matrix and it's a unit.

Where, $E(Z) = 0$ and $Cov(Z) = \rho$.

The PCs of Z can be obtained from eigenvectors of the correlation matrix ρ of X. All our previous properties for X are applied for the Z, so that the notation Y_j refers to the j^{th} PC and (λ_j, a_j) refers to the eigenvalue – eigenvector pair. However, the quantities derived from Σ are not the same from those derived from ρ (Richard and Dean, 2001).

The j^{th} PC of the standard variables $Z' = [z_1, z_2, z_3, \ldots, z_p]$ with $cov(Z) = \rho$, is given by

$$Y_j = \alpha_j' Z = a'\left(\theta^{1/2}\right)^{-1}(X - \mu) \quad (9)$$

So that $\sum_{j=1}^{p} V(X_j) = \sum_{j=1}^{p} Z = \rho for\ j = 1, 2, 3, \ldots, p$ (10)

Thus, (λ_1, α_1), (λ_2, α_2), (λ_3, α_3), …, (λ_p, α_p) are the eigenvalue - eigenvector pairs for ρ-diamensional space with $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \lambda_p \geq 0$.

4.2.2 Interpretation of Principal Components Analysis (PCA) Results:

The eigenvector $\alpha_j = \alpha_1, \alpha_2, \alpha_3, \ldots \alpha_p$quantifies the contribution of each measured variable to a given Principal Component (PC). When all loadings are positive, the first PC represents a weighted average of the variables, generally interpreted as an overall crime rate index. Subsequent components with positive and negative coefficients can be considered as distinct crime type indicators.

The score represents the relevance of a Principal Component (PC) to a specific data point. New observationsY_{ij} are generated by substituting the original variablesX_{ij} onto the set of the first qPCs.

As a result, we have

$$Y_{ij} = \alpha_{j1}' X_{i1} + \alpha_{j2}' X_{i2} + \alpha_{j3}' X_{i3} + \cdots + \alpha_{jp}' X_{ip}$$

$$i = 1, 2, 3, \ldots, p \qquad j = 1, 2, 3, \ldots, p$$

The plot of the first two or three PCs against each other enhances visual interpretation.

The proportion of variance will be used to tell us the PC that best explained the

original variables and the measure of how well the first q PCs of Z explain the variation is given by:

$$\phi_q = \frac{\sum_{j=1}^{q}\lambda_j}{p} = \frac{\sum_{j=1}^{q}V(z_j)}{p} \quad (11)$$

The cumulative proportion of variance accounted for by each PC provides a quantitative criterion for component retention. A Scree plot graphically illustrates the critical point, signifying the optimal number of components.

4.2.3 Concept of Discriminant Analysis

Assuming there are two multivariate normal populations with equal variance-covariance matrices, $N(\mu_1,\Sigma)$and $N(\mu_2,\Sigma)$ where $\mu_1(i=1,2) = (\mu_1, \mu_2, - - -, \mu_p)$ is the vector of means of the i[th] population and Σ is the variance-covariance matrices of the two populations. The probability density function of ith population is given as follow:

$$P_i(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(X-\mu_i)'\Sigma^{-1}(X-\mu_i)\right]$$

The ratio of the densities of two multivariate normal populations is given below (Usman (2023):

$$\frac{P_1(X)}{P_2(X)} = \frac{exp\left[-\frac{1}{2}(X-\mu_1)'\Sigma^{-1}(X-\mu_1)\right]}{exp\left[-\frac{1}{2}(X-\mu_2)'\Sigma^{-1}(X-\mu_2)\right]} \geq k$$

$$exp\left[-\frac{1}{2}\{(X-\mu_1)'\Sigma^{-1}(X-\mu_1) - (X-\mu_2)'\Sigma^{-1}(X-\mu_2)\}\right] \geq k \quad (13)$$

By taking the natural logarithms of equation (13) above; which is monotone increasing we have:

$$-\frac{1}{2}\{(X-\mu_1)'\Sigma^{-1}(X-\mu_1) - (X-\mu_2)'\Sigma^{-1}(X-\mu_2)\} \geq log\ k$$

The second term of (14) above is the Mahalonobis square distance between $N(\mu_1,\Sigma)$ and $N(\mu_2,\Sigma)$. For k suitably chosen (which of course can be one and then log k will be zero), the left hand side of equation (14), can be expanded and reposition to get the following equation:

$$X'\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \quad (15)$$

The first expression of equation (15) above is the well known as Fisher's linear discriminant function which is linear in the component of the observation vector.

Let

$$\bar{X}_i = \begin{pmatrix} \bar{x}_{i1} \\ \bar{x}_{i2} \\ ,, \\ ,, \\ ,, \\ \bar{x}_{ip} \end{pmatrix} \quad (16)$$

Where $\bar{X}_i$ represent the sample mean vector (12)
Affected).

Let $\bar{x}i1, \ldots, ip$ represent the individual mean

For instance;

$$\bar{X}_{i1} = \frac{1}{k}\sum_{i-1}^{n} X_{i1} \quad (17)$$

Where X_{11} is the mean of the first variable in
first
variable in group two, k is the number of the c
Particular group (Usman, 2012). (3.2)

The sample variance-covariance matrix is give

$$S_i = \begin{pmatrix} S_{11}S_{12} & \ldots & S_{1p} \\ S_{21}S_{22} & \ldots & S_{2p} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ S_{p1}S_{p2} & \ldots & S_{pp} \end{pmatrix} \quad (18)$$

(14)

Where S_i denote variance-covariance matrix, fo

S_{ii} denotes an individual variance and S_{ip} denotes an individual covariance for $p = 1, 2,$

The illustrations are given below,

$S_{ij} = \frac{1}{k_i}\sum_{i-1}^{n_i}(X_{ij} - \bar{X}_{.i})^2$ (general variance)

$S_{11} = \frac{1}{k}\sum_{i-1}^{n}(X_{i1} - \bar{X}_1)^2$ (an individual variance)

$S_{12} = \frac{1}{k}\sum_{i-1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$ (an individual covariance)

Let π_1 denote group one (Unaffected infant) and π_2 denote group two (Affected infant)

The Euclidean distance for the Unaffected infants (π_1) is;

$\hat{l}_1 = X_1^{'} S_p^{-1}(\bar{X}_1 - X_2^{-})$ (22)

and Euclidean distance for the Affected infants (π_2) is;

$\hat{l}_2 = X_2^{'} S_p^{-1}(\bar{X}_1 - X_2^{-})$ (23)

Where S_p denotes the pooled variance matrix

The mean Euclidean distance used in this study for the two groups is given as;

$\hat{M} = \frac{1}{2}(\hat{l}_1 + \hat{l}_2)$

(24)

And the Discriminant function is calculated by

$\hat{Y} = X^{'} S_p^{-1}(X_1 - X_2)$ (25)

Therefore, the classification rule is that;

$if \quad \hat{Y} \geq \hat{M} \quad classified\ as\ group\ one\ (\pi_1)$

and

$if \quad \hat{Y} < \hat{M} \quad classified\ as\ group\ two\ (\pi_2)$

Where $\hat{Y}$ denote the Discriminant function, and $\hat{M}$ denote the mean Euclidean distance for Unaffected and Affected (BPn) groups (19)

$X^{'} = (x_1 x_2)$ (26) (20)

$S_p = \frac{n_1 s_1 + n_2 s_2}{n_1 + n_2}$ (27) (21)

If $n_1 \neq n_2$, equation (27) will be used but if $n_1 =$ variance S_p above becomes:

$S_p = \frac{s_1 + s_2}{2}$ (28)

Where $S_1 and\ S_2$ are the respective sample varian two groups, and $n_1 and\ n_2$ are the sample size of t

The Fisher's Linear Discriminant model to be used is:

$$Y_{HS} = \beta_0 + \beta_i X_i + \varepsilon (29)$$

Where;

$Y_{HS}\ denote response probability (Health Status)$

$\beta_0\ is\ the\ expected\ value\ of\ Y\ when\ the\ X^{'}s\ are\ set\ as$

$\beta_i\ is\ the\ regression\ coefficient\ for\ each\ correspon$

$X_i\ is\ the\ predictor\ or\ independent\ variables$

$\varepsilon\ is\ the\ error\ of\ the\ predictor.$

In this study, '0' will be used to represent "Unaffected" and '1' to represents "Affected (BPn)". The mean of the dichotomous random variable Y_{HS}, designated by P_{HS}, is the proportion of times that the pneumonia takes the value '0'. Equivalently;

$P_{HS} = P(HS = 0) = P(\text{Unaffected})$ (30)

5.

XPECTED OUTCOME

This study's expected outcome is anticipated to align with the findings of previous researchers who conducted similar work, though with some variations in data usage. Specifically, our research will utilize a broader dataset, building

upon the existing body of work. For context, the expected outcomes from comparable studies are summarized below;

Anna *et al* (2014), they used PCA method in reduction of the number of studied variables with the maintenance of as much information as possible, and also as a first step in analyzing data from IVF (in vitro fertilization). The next step and main purpose of the analysis was to create models that predict pregnancy. Therefore, 805 different types of IVF cycles were analyzed and pregnancy was correctly classified in 61-80% of cases for different analyzed groups in models obtained

Beki (2012) used discriminant analysis and binary logistic regression to trace the prevalence of Broncho-Pulmonary Dysplasia (BPD) among children. The researcher used three possible predictor variables i.e. weight at birth, weight four weeks later and gender and built a discriminant model that is capable of tracking Broncho-Pulmonary Dysplasia (BPD) infants.The study predicted the BPD status of five new infants using the discriminant model in which all the five new cases were correctly predicted. The discriminant model built had a perfect classification of five new cases in Kaduna while it has misclassification of one of five new cases in Sokoto.

Mahdi *et al* (2020), in their research on cancer diseases using discriminant method found that Gender has the highest discriminating power, whereas the Grade variable has the least discriminatory power. Similarly, Behavior has the highest discriminatory power, whereas the Government has the least biased power. It became clear that the third group (those with breast cancer) had the highest probability of the correct classification. It was discovered that, the probability of correct classification reached 92%, followed by the second group with brain cancer, where the probability of correct classification was 64%. Finally, the first group with bladder cancer had the lowest probability of correct classification. They concluded that, increasing the sample size has a significant impact on the correct classification of observations.

Nahida *et al* (2020), in their research proposes a fusion of Deep Convolutional Neural Network Model with Principal Component Analysis (PCA) feature extraction model and Logistic Regression (LR) classifiers for the diagnosis of pneumonia from chest X-ray images. In this study, fine-tuned pre-trained CheXNet model is used as Convolutional Neural Network (CNN) model on standard pneumonia dataset collected from Guangzhou Women and Children's Medical Center, Guangzhou. The proposed model is capable of detecting pneumonia with an accuracy which outperforms the existing methods from 0.8% to 21.9% approx. Comparison with existing models and methods reveal that the proposed model delivers superior results than others according to precision.

References

Aaron K. (2018) via https://www.medicalnewstoday.com/articles/3231 67.php Accessed on 25th January, 2019

Akash D., (2018): The Mathematical Behind PCA from Raw Data to

Principal Components. Published in Towards Data Science.

Anderson, R. E., Hair, J. F., Black, W. C. and Babin, B. J (2018): Multivariate Data Analysis (8 ed.). Cengage Learning: London.

Anna, J. M., Dorota, J., Dorota, C., Teresa W., Brian, A., Robert, M. (2014): The Use of Principal Component Analysis and Logistic Regression in Prediction of Infertility Treatment Outcome. Studies in Logic, Grammerand Rhet oric 39 (52).

Anthony J, Brooks W. A, Malik J. S, Douglas H and kim M. E (2010): *Pneumonia research to reduce childhood mortality in the developing world.*

Barthlomerv, D.J (2010): Analysis and Interpretation of Multivariate Data. An International Encyclopedia of (Third Edition).

Beki, (2012). *The Use of Discriminant Model and Logistic Regression for Tracking the Incidenceof Broncho-Pulmonary Dysplasia among Infants.* A dissertation submitted to Usman Danfodio University, Sokoto Nigeria.

Bipin, P; Nitiben, T; Sonaliya, KN (2011). A study on prevalence of Acute Respiratory-tract Infections (ARI) in under five children in urban and rural communities of ahmedabad district,Gujarat. *National. J. of Comm. Med*. Vol 2 Issue 2.

Biruk,B., Melaku B., Ayeligu, M., Mesfin W. and Mdla, A (2020): Prevalence of Pneumonia and its Associated Factors Among Under-Fire Children an East Africa: A Systematic Review and Meta-Analysis: Published May 2020.

Brennan Whitfield (2024): A Step-by-Step Explamate of PCA.

Bryan kaller, Dobrin marchev (2023): Quantitative Research and Educational Measurement. International encyclopedia of education (fourth edition)*by FT-Raman spectroscopy and Chenometrics*. Journal on Biotechnol Agron Soc Environ ; 1 Vol 15(1) pages 75- 84.

Clement, Y. E., Ruoqi, M., Emmanuel, K. D., Clement, A., Ruiping, Q., Yongjun, W., Lijun, M., and Yanbin, W. (2022): Machine learning-assisted prediction of pneumonia based on non-invasive measures. Front Public Health; 10:938801. Published online.

Cologne, 2021; National Library of medicine (National Center for Biotechnology information). Institute for Quality and efficiency in Health care (1 QWIG)

Danbaba A, Audu M, Mahmud A. M (2013). *Investigation of risk factors of low birth*
Danfodio University, Sokoto Nigeria.

Emily Steven, 2023: An Introduction to Multivariate Analysis.

Emin M. A, Levent D, Adelet O, Aysu T. (2014): *Epicardial adipose tissue, netrophil – to-* Epidemiology. Vol. 6:1

Evans, J. M., Dharmar, M., Meierhenry, E., Marcin, J. P., & Raff, G. W. (2014). Association between Down syndrome and in-hospital death among children undergoing surgery for congenital heart disease: a US population-based study. *Circulation: Cardiovascular Quality and Outcomes*, *7*(3), 445-452.

Fernaudez P, Abbas O, Dardenne P, Baeten V. (2010) : *Discrimination of Corsican Honey*

First Trimester of Pregnancy attending an inner-city Antenatal Department in the UK.
Gujarat. *National. J. of Comm. Med*. Vol 2 Issue 2.

https://www.analyticsvidhya. Com/blog/2016/03.

https://www.linkdin.com /advice/0/how-can-principal-component-analysis-best-leveraged-tawst.

Janelle M. (2019): Bronchopneumonia: Symptoms, Risk factors and treatment. Medically reviewed by Gerhard Whitworth, R. January, 2019 *Journal of clinical and experimental medicine.Journal of Royal Society for the Promotion of Health*;125(5): 232 – 238.

Krasevec J, Blencowe H, Coffey C (2022): Study protocol for UNICEF and WHO Estimates of Global, Regional and National Low Birthweight Prevalence for 2000 to 2020. Gates Open Res (2022), 6:80

Mahdi, N. N., Abed, H. T., & Sadik, N. J. (2020). The discriminant analysis in the evaluation of cancers diseases in Iraq. *Int J Adv Sci Eng Inf Technol,* 10(5), 2170-2176.

Michael, G; Patrick, J.F.G; Trevor, H; Alfonso L.D. (2022) Principal component Analysis. Natural Reviews Methods Primers Vol.2 (1): 100.

Mobil P, Landero A .De-Antoni G. Araujo Andraote C, Avila- Donoso H, Moreno J. (2010): *Multivariate Analysis of Raman Spectra applied to microbiology: Discriminant of Microorganism at the species level*. REVISTA MEXICANA DE FESICA 56 (5) 378-385.

Nahida, H., Hasan, M. M., and Rahman, M. M. (2020): Fusion of Deep Convolutional Neural Network with PCA and Logistic Regression for diagnosis of pediatric pneumonia on chest X-Rays. Network Biology, Vol. 10(3):62-76

Pepke-Zaba, J., Delcroix, M., Lang, I., Mayer, E., Jansa, P., Ambroz, D., ... & Simonneau, G. (2011). Chronic thromboembolic pulmonary hypertension (CTEPH) results from an international prospective registry. *Circulation*, *124*(18), 1973-1981.

Pope d. p, Mishra V., Thompson L (2010): *Risk of low birth weight and Stillbirth associated*

Pract;13: 365 – 70.

Shibuya K , Hoshimo H, chiyo M. et al. (2003) *subepithelial vascular patherns in bronchial*

dysplasias using a high magnification bronchovideoscope. Thorax 2003; 58: 989-995

Ting, H., Jiarong L. and Weiru Z. (2020): Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidis m. BMC Medical Research Methodology.

Usman, A. (2012). *Statistical methods for Biometric & Medical research*. Kaduna, Millennium Publisher (New Nigeria Newspaper).

Usman, A. (2023). *Bivariate and MultivariateStatistical Analysis*. 2nd Edition. Published by: Kadpoly Spider Press Limited, Kaduna Nigeria.

Vimalkumar, V. K, R. C Moses, Padmanaban S. (2011): *Binary logistic model for detecting Prevalence & Risk Factor of Nephropathy. In Type 2 Diabetic patients.* International Journal of collaborative Research on internal medicine and public Health 3 (8):598-615.

Vishwa N. M, Madaki U. Y, Vijay V. S and Babagana M. (2015) *Application of Discriminant*
weight using multivariate logistic regression analysis: Journal of humanities, science
withindoor air pollution from solid fuel use in developing countries. Epidemiol Rev. 70-80.

Wonodi, C. B., Deloria-Knoll, M., Feikin, D. R., DeLuca, A. N., Driscoll, A. J., Moïsi, J. C. (2012): Pneumonia Methods Working Group and PERCH Site Investigators.. Evaluation of Risk Factors for Severe Pneumonia in Children: the Pneumonia Etiology Research for Child Health study. *Clinical infectious diseases*, *54*(suppl_2), S124-S131.
www.marchofdimes.org

www.stanfordchildrens.org/standardmedicine. Children's Health.

Yakubu, Y., Ahmed, S. S., Audu, I., & Usman, A. (2019). Binary logistic regression methods for modeling broncho-pneumonia status in infants from tertiary health institutions in north central Nigeria. *Journal of Applied Sciences and Environmental Management*, *23*(8), 1607-1614.