Engineering and Technology Journal e-ISSN: 2456-3358

Volume 10 Issue 02 February-2025, Page No.- 3780-3786

DOI: 10.47191/etj/v10i02.07, I.F. – 8.482

Multivariate Statistical Interpretation of Airborne Diseases using Principal Components Analysis

Paul Boye[1], Yao Yevenyo Ziggah[2]

[1]Faculty of Engineering, Department of Mathematical Sciences, University of Mines and Technology, Tarkwa, Ghana

[2]Faculty of Geosciences and Environmental Studies, Department of Geomatic Engineering, University of Mines and Technology, Tarkwa, Ghana

ABSTRACT: Accurate and timely determination of relationships among communicable diseases is crucial in taking precautionary measures to control and prevent the transmission of infectious diseases. This will inform the government to make important policies that will help provide effective patient treatments. This study employed the principal components analysis (PCA) method to explore and interpret the relationship among airborne communicable diseases: Tuberculosis (TB), chickenpox, and measles based on a medical dataset obtained from Bibiani Government Hospital, Ghana. The Kaiser criterion was employed to determine a suitable number of principal components (PCs) to feature in the statistical analysis. A scree plot diagram was also used to affirm the number of PCs needed in the analysis. Projection of diseases on PC planes was also used to interpret the relationships among the diseases. From the study, statistical results revealed that the first principal component (PC1), second principal component (PC2), and third principal component (PC3) performed significantly well in the disease interpretation by explaining a total variation of 46.70994%, 30.61631%, and 22.67376, respectively of the useful information in the dataset. There were also marked strong correlations among the diseases concerning PCs. Due to the limited number of diseases considered, this study will serve as preliminary investigation on the use of PCA as a versatile and promising multivariate statistical technique that can be relied upon by public health experts and policymakers to interpret relationships among diseases and make informed decisions very well.

KEYWORDS: Airborne communicable diseases, dimensionality reduction, principal component analysis

1. INTRODUCTION

Communicable diseases have often been described as infectious and highly contagious. Exploring the relationship among communicable diseases helps in precautionary measures, and effective treatment at an early stage is helpful for patients. Over the years, several coordinated efforts by scientists to curb the marked spread of such infectious diseases have been explored. In this study, three of the most notable communicable diseases that remain the leading cause of death in Ghana, namely Tuberculosis (TB), chicken pox, and measles have been studied. The burden of these diseases, if not combated, will increase economic suffering, compound an already fragile healthcare infrastructure, and render the livelihood of the citizenry poor (Mettle *et al.*, 2020; Whittaker *et al*., 2019; de-Graft Aikins *et al.*, 2012).

From the literature, researchers have frequently applied the multivariate statistical method as the most popular tool to analyse multivariate datasets since it reveals the important features of the dataset (Salem & Hussein, 2019; Ahmad *et al.*, 2018). In this study, principal components analysis (PCA) as an example of a multivariate statistical method was employed to explore and interpret the relationship among the three communicable diseases. This was achieved by reducing the dataset dimensionality without much loss of information.

The main advantages of using the PCA in mathematical analogy and epidemiology include its suitability for visualisation, its ability to capture variation in complex datasets, low noise sensitivity, decreased requirements for capacity and memory, and increased efficiency given the processes taking place in a smaller dimension (Mohammed *et al.*, 2016; Karamizadeh *et al.*, 2013).

Due to the above-mentioned advantages of the PCA, researchers have used it widely in many fields of study. For example, Shilaskar & Ghatol (2013) investigated two feature extraction techniques PCA and factor analysis (FA) on a medical dataset for heart disease classification. The techniques maintained the integrity of the dataset by improving the diagnosis performance. On the other hand, Yadav & Jat (2020) used the dimensionality reduction technique for chronic disease prediction. Their results showed a significant improvement in the prediction accuracy. Meghraoui *et al.* (2016) also applied PCA in the medical field to select the main voice principal components (PCs) of a person infected with Parkinson's disease. Their results gave very promising prediction accuracy. In hydrology, Sharma *et al.* (2015) applied PCA to dimensionless geomorphic parameters to group the parameters under different components based on significant correlations. Results

revealed that some of the parameters were strongly correlated with the components, but texture ratio and hypsometric integral do not show a correlation with any of the components. Chen et al. (2020) used artificial neural networks (ANN) and PCA methods to build a cost prediction model in aviation. In the PCA, the eigenvalues of the first two PCs indicated that the PCs had the strongest interpretation of the original variable information and were retained as cost-influencing variables to train the ANN model.

From the enumerated review, as PCA has proven to be a versatile and reliable method, it would be prudent to adopt it and explore its frontiers further to analyse and interpret the airborne communicable disease dataset obtained from a Government Hospital in Ghana. The communicable diseases considered in this study are Tuberculosis (TB), chickenpox, and measles. Hence, the application of PCA will help in interpreting the interrelationship among the three communicable diseases. Furthermore, the control and prevention of the transmission of infectious diseases which is a public health priority depend on the early detection of the pathogens (viruses, bacteria, fungi, and protists) causing these diseases. In that regard, this study provides the solution to the urgent action that is needed to curb the rising rates of these communicable diseases in low- and middle-income countries to reduce the resulting social and economic burdens by producing not only timely results, but accurate statistical interpretations (Janati *et al.*, 2015).

Finally, the findings of this research will inform the government to make important policies including programs funding for the control and prevention of these communicable diseases.

2. STATISTICAL METHOD

2.1 Principal Components Analysis

From the literature, it is difficult and complicated to determine the process causing the airborne spread of infectious diseases and improve the influencing factors and transmission routes of these diseases (Li *et al.*, 2007).

PCA is a statistical technique for simplifying datasets by using an orthogonal transformation to transform correlated variables into a set of uncorrelated variables called PCs. The PCs sequentially capture the maximum variability among the original variables and ensure minimum loss of information (Konatè *et al.*, 2015; Muhammad *et al.*, 2019; Ngo & Turbow, 2019; Luo *et al.*, 2020). To show all the variables involved in the analysis, PCA was applied to the airborne communicable diseases considered in this study.

Consider n objects and p variables ($n \times p$ matrix) observed on the disease dataset **X**. The PCA stems new variables as a weighted linear transformation of the variables of **X** into a new set with uncorrelated desirable properties with each other so that their relations with another variable can be explored more easily. By this process, the bulk of information in the dataset is taken care of by progressively computing the PC that accounts for the combined variability in the dataset.

To compute the PC, **X** was decomposed into vectors and a matrix using the singular value decomposition (SVD) theory (see Equation (1)). SVD is one of the most useful results in matrix theory as it provides a solution in exactly the form that is required for a biplot. This theory was employed to decompose **X** in the following manner (Orumie & and Ogbonna, 2019; Dash *et al.*, 2014; Cherry, 1996):

$$\mathbf{X} = \mathbf{U}\mathbf{D}_{\lambda}\mathbf{V}^{T} \quad (1)$$

where $\mathbf{U}$ = left singular vector, $\mathbf{V}$ = right singular vector, $\mathbf{U}\mathbf{U}^{T} = \mathbf{I}_{n}$ (square matrix, *n* dimension), and $\mathbf{V}^{T}\mathbf{V} = \mathbf{I}_{p}$. (square matrix, *p* dimension). T is the transpose of the matrix and **D** is a diagonal matrix with singular values λ_{τ}. Consequently, Equation (2) presents the correlation of standardized variables of the sample dataset as follows:

$$\mathrm{Cor}(\mathbf{X}) = \mathbf{V}\Delta\mathbf{V}^{T} \quad (2)$$

where $\Delta = diag(\sigma)$ is a diagonal matrix having $\sigma_1^{\,2} \geq \sigma_2^{\,2} \geq \ldots \geq \sigma_p^{\,2} \geq 0$ so that $(n-1)\sigma_{\nu}^{2} = \lambda_{\nu}^{2}$ for $\nu = 1, 2, \ldots, p.$ Studies have shown that the PC score, **J,** is the projection of **X** onto the orthonormal basis of $\mathbf{V}$ (see Equation (3)):

$$\mathbf{J} = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{D} \quad (3)$$

Studies have further shown that the major aim of PCA is to reduce the size of multivariate matrices like **X** and the complexity of the interrelationship among the variables to a relatively smaller number of linear combinations of them, which are referred to as Principal Components (PCs) without much loss of information. Suppose $k \leq p,$ Equation (4) presents the elementwise model.

$$x_{\tau\nu} = \sum_{t=1}^{k} u_{\tau t}\lambda_{t}v_{t\nu} + \varepsilon_{\tau\nu} \quad (4)$$

Here, $u_{\tau t}$ = matrix $\mathbf{U}$ elements and $v_{t\nu}$ = matrix $\mathbf{V}$ elements. $\varepsilon_{\tau\nu}$ = residual terms for $\tau = 1, 2, \ldots, n$ and $\nu = 1, 2, \ldots, p.$ Hence, the explained variability is given in Equation (5) as follows:

$$\text{Percentage Variance} = \left(\frac{\sum_{\nu=1}^{k} \lambda_{\nu}}{\sum_{\nu=1}^{p} \lambda_{\nu}} \right) 100\% \quad (5)$$

3. DATA DESCRIPTION

In this study medical dataset that spans from January 2013 to December 2018 was obtained from a Government Hospital in Ghana, West Africa. The acquired dataset contains only the number of recorded cases of tuberculosis, chicken pox and

measles over the stipulated time frame. It is important to mention that the respective subjects were not considered in the data analysis. Table 1 shows the descriptive statistics of the dataset.

Table 1: Descriptive Statistics of Dataset

	Statistics			
Disease	**Mean**	**Standard Deviation**	**Minimum Value**	**Maximum Value**
Tuberculosis	18.6667	8.6056	0	37
Chicken pox	31.6667	14.6883	5	63
Measles	3.25	2.8864	0	13

4. NUMERICAL APPLICATION

In this study, the authors performed correlation analysis on airborne diseases to find the strength and direction of the linear relationship among them. Table 2 shows the correlation matrix of TB, chickenpox, and measles. According to Hunter *et al.* (2020), correlations are analysed by using the following scale: 0 corresponds to no linear relationship, 0 to 0.3 or 0 to -0.3 corresponds to a weak linear relationship, 0.3 to 0.7 or -0.3 to -0.7 corresponds to a moderate relationship and 0.7 to 1.0 or -0.7 to -1.0 corresponds to a strong linear relationship. From the correlation results in Table 2, it can be observed that the symptoms related to TB and chickenpox problems have a moderate negative correlation of -0.3158. This indicates that TB and chickenpox patients are more likely to transmit their diseases to each other. It can also be seen that the symptoms related to TB and measles had a weak negative correlation value of -0.1078. This means that TB and measles patients are less likely to transmit their diseases to each other. Similar results can be said about the symptoms related to chickenpox and measles with a weak positive correlation of 0.1536. This suggests that patients with these two categories of diseases cannot spread them among themselves (Wang *et al.*, 2020).

Table 2: Correlation Matrix of Airborne Diseases

Disease	**TB**	**Chickenpox**	**Measles**
Tuberculosis	1		
Chicken pox	-0.3158	1	
Measles	-0.1078	0.1536	1

The criterion utilised in this study to determine the number of PCs in PCA modeling is the Kaiser criterion (Kaiser Criterion) (Costa *et al.*, 2014; Kanyongo, 2005). This criterion could retain and interpret any PCs with an eigenvalue greater than one. Table 3 clearly shows the share of overall variability explained by each PC. From Table 3, the first principal component (PC1), the second principal component (PC2), and the third principal component (PC3) had eigenvalues of 1.401298, 0.918489 and 0.680213, respectively. In this respect, only PC1 satisfied the Kaiser criterion. However, PC2 and PC3 with eigenvalues less than one were retained in this study to help in revealing the relationship among the communicable diseases. This is because the presented study is based only on the interpretation of the interrelationship among TB, chicken pox and measles. From Table 3, it is obvious that PC1, PC2 and PC3 could explain 46.70994%, 30.61631% and 22.67376% of the airborne disease dataset. The cumulative sum of the useful information derived from the disease dataset is 100%. This means that all the useful information about the disease dataset was used for the interpretation of the relationship among the diseases without any information being lost.

Table 3: PCA results on the medical dataset

Principal Component (PC)	**Eigenvalue**	**% of Total Variance**	**Cumulative %**
1	1.401298	46.70994	46.7099
2	0.918489	30.61631	77.3262
3	0.680213	22.67376	100.0000

Figure 1 is a scree plot employed in the study to further show the PCs' ability to explain the variation in the communicable disease dataset. The figure clearly shows each PC's share of the total variation. The PC loadings (Table 4) show the nature of the correlation between the PCs and communicable diseases (TB, chickenpox, and measles). The interpretation of the PCs is based on finding which communicable diseases are most strongly correlated with each PC. Here a correlation above 0.3 is deemed important.

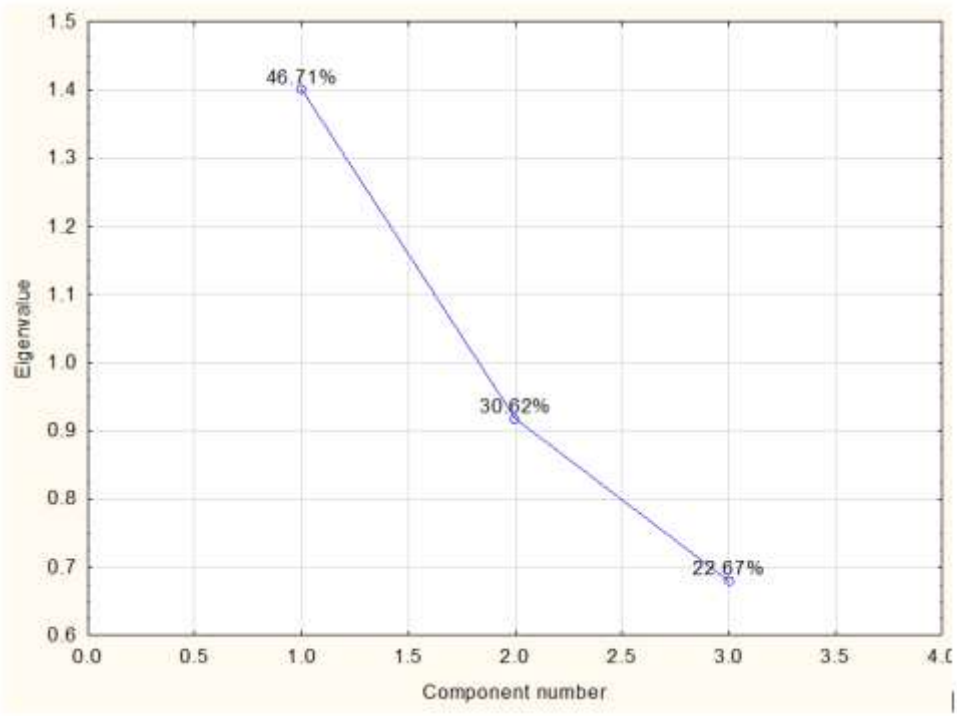

Figure 1: Scree Plot of the three PCs

Figures 2 to 4 show the projection of the diseases on the PC plane to confirm further the correlation among the communicable diseases. This visualisation helps to show the most correlated diseases to each component. From these figures, diseases pointing in the same direction in a quadrant on the PC plane suggest that they correlate positively and vary together. This means an increase in the transmission rate of one of the diseases tends to increase the transmission rate of the corresponding disease and vice versa. On the other hand, diseases pointing in the opposite sense on the PC plane correlate negatively. This also suggests that an increase in the transmission rate of a disease will cause the transmission rate of the corresponding disease to decrease and vice versa. Figure 2 is an example of such negatively correlated diseases for TB and chickenpox. Similar results can be said about Figures 3 and 4 for TB and measles, and chickenpox and measles, respectively.

However, diseases presented on the PC plane in an orthogonal way are uncorrelated. The uncorrelated diseases can be seen in Figure 2 for TB and measles as well as measles and chickenpox. Similar uncorrelation was observed for TB and chickenpox as well as chickenpox and measles in Figure 3. In Figure 4, TB and chickenpox as well as TB and measles showed uncorrelation. Thie interpretation for this uncorrelation is that an increase or decrease in the transmission rate of one of the diseases has no influence on the behaviour of the other disease.

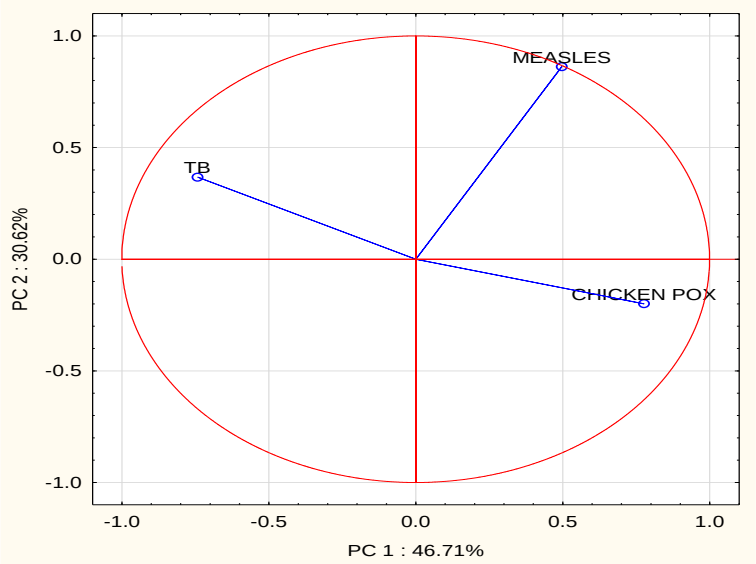

Figure 2: Projection of the disease variables on the PC1 versus PC2

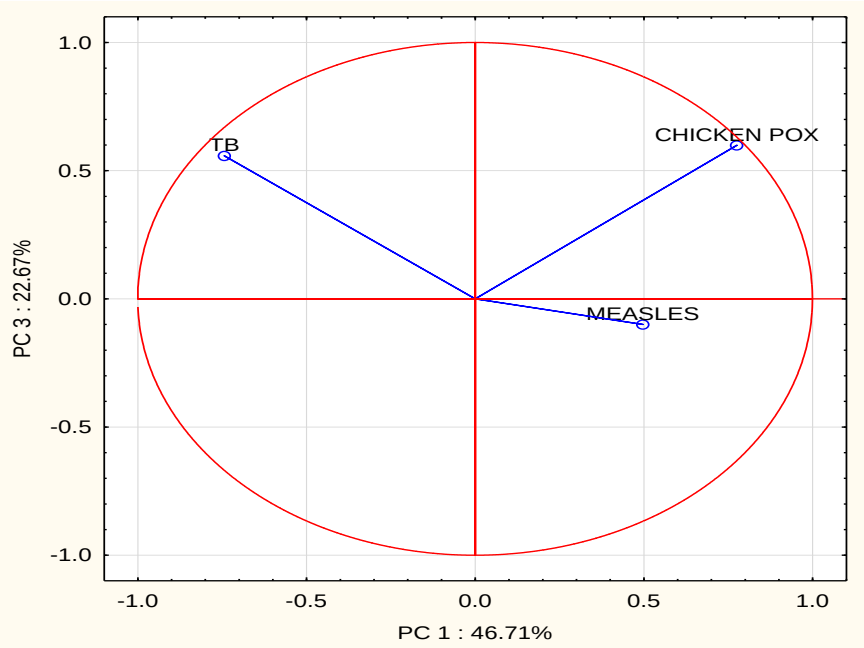

Figure 3: Projection of the disease variables on the PC1 versus PC3

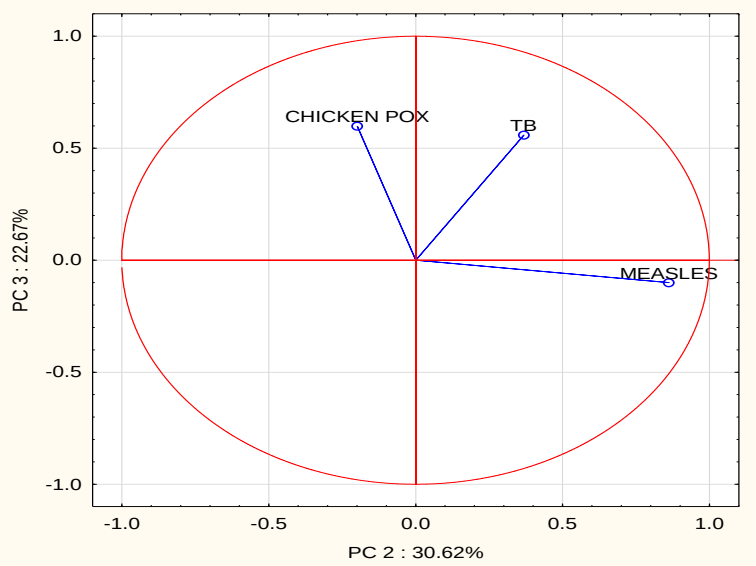

Figure 4: Projection of the disease variables on the PC2 versus PC3

The PCs can be interpreted using PC loadings to understand the significance of each communicable diseases to PC1, PC2 and PC3. From Table 4, PC1 is strongly correlated with TB, and chicken pox, while measles moderately correlates to PC1. Thus, the correlation in PC1 increased with increasing chickenpox and measles but decreased with TB. This suggests that these three diseases do not vary together. Since chickenpox and measles vary together, an increase in the transmission rate of any one of them will cause the transmission rate of the corresponding disease to increase and vice versa. However, the transmission rate of TB will decrease in this instance and vice versa. Similar results can be said about PC2 that it is strongly correlated with measles and TB. This means that if the transmission rate of measles is increased, then that of TB will also increase, and vice versa. However, the transmission rate of chickenpox will decrease and vice versa. It can also be seen that PC3 is strongly correlated with chickenpox and TB. Thus, an increase in the transmission rate of chickenpox will cause an increase in the transmission rate of TB, but there will be a decrease in the transmission rate of measles and vice versa.

Table 4: Principal Component Loadings

Variable	PC 1	PC 2	PC 3
TB	-0.743791	0.367868	0.558075
Chickenpox	0.775508	-0.199358	0.599036
Measles	0.496650	0.862217	-0.099599

4. CONCLUSION

The main contribution of this study is to explore the capability of PCA as a multivariate statistical technique to interpret the relationship between TB, chickenpox, and measles. First, the Pearson correlation analysis revealed the existence of moderate and weak relationships between the diseases. For the PCA results, it was revealed that PC1, PC2 and PC3 could explain the original data by 46.70994%, 30.61631%, and 22.67376% with no loss of information. The dynamics in the disease transmission were revealed from the projection of the diseases onto the PC plane. Here, TB and chicken pox, TB and measles, and chickenpox and measles were found to have a negative correlation suggesting that an increase in the transmission rate of a disease will cause the transmission rate of the corresponding disease to decrease and vice versa. Orthogonal relationships were also observed on the PC plane where TB and measles, measles and chickenpox, and TB and chickenpox showed uncorrelation. This uncorrelation implied that an increase or decrease in the transmission rate of one of the diseases does not influence the behaviour of the other diseases. Because only three airborne diseases were considered, the results presented in this study can only be considered as preliminary work where it is demonstrated that PCA could be a useful technique to understand and interpret the relationships among diseases and their transmission. One limitation of this study is the number of airborne diseases considered. Therefore, for future work, it is recommended that more airborne diseases should be considered for in-depth interpretation of their interrelationships.

ACKNOWLEDGMENT

The authors are thankful to the management of the Government Hospital for providing us with the necessary data for this study.

FUNDING

There was no funding for this research work.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. AHMAD, F. & DAR, W. M. (2018) Classification of alzheimer's disease stages: an approach using PCA-based algorithm. *American Journal of Alzheimer's Disease and Other Dementias*, 33, 433-439.
2. CHEN, X., YI, M. & HUANG, J. (2020) Application of a PCA-ANN-based cost prediction model for general aviation aircraft. *IEEE Access*, 8, 130124-130135, https://doi.org/*10.1109/ACCESS.2020.3008442*.
3. CHERRY, S. (1996) Some comments on singular value decomposition analysis. *American. Meteorological Society*, 10, 1759-1761.
4. COSTA, J. C. G. D., DA-SILVA, P. J. G., ALMEIDA, R. M. V. R. & INFANTOSI, A. F. C. (2014) Validation in principal components analysis applied to EEG data. *Computational and Mathematical Methods in Medicine*, 2014, 1-10. http://dx.doi.org/10.1155/2014/413801.
5. DASH, P., NAYAK, M. & DAS, G. P. (2014) Principal component analysis using singular value decomposition for image compression. *International Journal of Computer Applications*, 93(9), 1-8.
6. de-GRAFT AIKINS, A., ADDO, J., OFEI, F., BOSU, W. K. & AGYMANG, C. (2012) Ghana's burden of chronic non-communicable diseases: future directions in research, practice and policy. *Ghana Medical Journal*, 46(2), 1-4
7. HUNTER, E., Namee, B. M. & Kelleher, J. D. (2020) A model for the spread of infectious diseases in a region. *International Journal of Environmental Research and Public Health*, 17, 3119, 1-19, https://doi.org/10.3390/ijerph17093119.
8. JANATI, A., HOSSEINY, M., GOUYA, M. M., MORADI, G. & GHADERI, E (2015) Communicable Disease Reporting Systems in the world: a systematic review article. *Iran Journal Public Health*, 44(11), 1453-1465.
9. KANYONGO, G. Y. (2005) Determining the correct number of components to extract from a principal components analysis: A Monte Carlo study of the accuracy of the scree plot. *Journal of Modern Applied Statistical Methods*: 4(1), 120-133, https://doi.org/10.22237/jmasm/1114906380.
10. KARAMIZADEH, S., ABDULLAH, S.M., MANAF, A. A. & ZAMANI, M. (2013) An overview of principal component analysis. *Journal of Signal and Information Processing*, 4, 173-175. https://doi.org/10.4236/jsip.2013.43B031.
11. KONATÈ, A. A., PAN, H., MA, H., CAO, X., ZIGGAH, Y. Y., OLOO, M. & KHAN, N. (2015) Application of dimensionality reduction technique

to improve geophysical log data classification in crystalline rocks. *Journal of Petroleum Science and Engineering*, 133, 633-645.

12. LI, Y., LEUNG, G. M., TANG, J. W., YANG, X., CHAO, C. Y. H. & LI, J. Z. (2007) Role of ventilation in airborne transmission of infectious agents in the built environment – a multidisciplinary systematic review. *Indoor Air*, 17, 2-8. https://doi.org/10.1111_j.1600-0668.2006.00445.x.
13. LUO, Y., WU, J., LU, J., XU, X., LONG, W., YAN, G., TANG, M., Lou, Z., Xu, D., Zhou, P., SI, Q. & ZHENG, X. (2020) Investigation of COVID-19-related symptoms based on factor analysis. *Annals of Palliative Medicine*, 1-8, http://dx.doi.org/10.21037/.
14. MEGHRAOUI, D., BOUDRAA, B., MERAZI-MEKSEN, T. & BOUDRAA, M. (2016) Features dimensionality reduction and multi-dimensional voice processing program to Parkinson disease discrimination. *Proceedings of 2016 4th International Conference on Control Engineering and Information Technology*, 1-5.
15. METTLE, F. O., AFFI, P. O. & TWUMESI, C. (2020) Modeling the transmission dynamics of tuberculosis in the Ashanti region of Ghana. *Interdisciplinary Perspectives on Infectious Diseases*, 2020, 1-16, https://doi.org/10.1155/2020/4513854.
16. MOHAMMED, S., KHALID, A., OSMAN, S. E. & HELALI, R. G. M. (2016) Usage of principal component analysis (PCA) in AI applications. *International Journal of Engineering Research and Technology*, 5 (12), 372-375.
17. MUHAMMAD, M. U., JIADONG, R., Muhammad, N. S., Hussain, M. & Muhammad, I. (2019) Principal component analysis of categorized polytomous variable-based classification of diabetes and other chronic diseases. *International Journal of Environmental Research and Public Health*, 16, (3593), 1-15, https://doi.org/10.3390/ijerph16193593.
18. NGO, A. N. & TURBOW, D. J. (2019) Principal component analysis of morbidity and mortality among the United States homeless population: a systematic review and meta-analysis. *International Archives of Public Health and Community Medicine*, 3(2), 1-9, https://doi.org/10.23937/2643-4512/1710025
19. ORUMIE, U. C. & OGBONNA, O. (2019) Principal component analysis and its derivation from singular value decomposition. *International Journal of Statistics and Probability*, 8(2), 183-191, https://doi.org/10.5539/ijsp.v8n2p183.
20. SALEM, N. & HUSSEIN, S. (2016) Data dimensional reduction and principal components analysis. *Procedia Computer Science*,163, 292–299.
21. SHARMA, S. K., GAJBHIYE, S. & TIGNATH, S. (2015) Application of principal component analysis in grouping geomorphic parameters of a watershed for hydrological modeling. *Applied Water Science*, 5(1), 89-96, https://doi.org/10.1007/s13201-014-0170-1.
22. SHILASKAR, S. & GHATOL, A. (2013) Dimensionality reduction techniques for improved diagnosis of heart disease. *International Journal of Computer Applications*, 61(5), 1-9.
23. WANG, X., PEI, T., LIU, Q., SONG, C., LIU, Y., CHEN, X., MA, J. & ZHANG, Z. (2020) Quantifying the time-lag effects of human mobility on the COVID-19 transmission: a multi-city study in China. *IEEE Access*, 8, 216752- 216761, https://doi.org/*10.1109/ACCESS.2020.3038995*.
24. WHITTAKER, E., LÓPEZ-VARELA, BRODERICK, C. & SEDDOB. J. A. (2019) Examining the complex relationship between tuberculosis and other infectious diseases in children. *Frontiers in Pediatrics*, 7(233),1-23, https://doi.org/10.3389/fped.2019.00233.
25. YADAV, R. & JAT, S. C. (2020) Feature selection and dimensionality reduction methods for chronic disease prediction. *International Journal of Scientific and Technology Research*, 9(4), 2912-2918.