A Statistical Approach for COVID-19 Pandemic Data Analysis

Sunday Adeola Ajagbe[1,2,*], Ismaila Muritala[2,3], Pragasen Mudali[1] and Matthew O Adigun[1]

[1] *University of Zululand, Kwadlangezwa, South Africa*
[2] *Abiola Ajimobi Technical University, Ibadan, Nigeria*
[3] *Osun State University, Osogbo, Nigeria*

Abstract

This study investigated the statistical insight of big data analytics for effective pandemic analysis. The COVID-19 pandemic dataset, containing 5010 negatives and 4990 positives, was obtained from a prominent source known as the Kaggle repository. The dataset was categorized into four classes: COVID-19, Normal, Pneumonia, and Tuberculosis using diverse statistical metrics of image characteristics: mean intensity of a color channel, homogeneity, dissimilarity, correlation, and density. The objective of the statistical analysis was to validate the use of big data analysis for pandemic preparedness, with a working hypothesis that the intensity of the colour channel images would be different ($p < 0.05$) between COVID-19 patients and patients with other health conditions. A statistical package for the Social Sciences version 20 was used for the analysis. The outcomes show that the mean intensity of color channels (Blue, Green, Red) in COVID-19 and Tuberculosis cases was greater in the blue and green channels than in Normal and Pneumonia cases. Pneumonia and normal cases exhibited comparable and reduced mean intensities across all three channels. The Normal condition exhibited the largest mean contrast, whereas Pneumonia displayed the lowest. The mean dissimilarity and homogeneity indicate that tuberculosis demonstrated the highest dissimilarity, signifying greater diversity in pixel intensity. COVID-19 and Pneumonia exhibited comparable homogeneity; however Tuberculosis demonstrated more homogeneity. The mean correlation indicates that Normal images exhibited the highest correlation, whilst Tuberculosis demonstrated the lowest. The mean density study indicates that normal cases demonstrated the highest mean density, whereas tuberculosis exhibited the lowest mean density. In conclusion, our findings established the advantages of statistics for modeling and analyzing big data in the pandemic domain, covers the state-of-the-art for practical analysis.

Keywords

Big data analytics, Pandemic analysis, Statistical analysis, Machine learning

1. Introduction

Big Data Analytics have become common in many areas, from research to practice, and even in everyday life. It is a prominent tool in pandemic preparedness and control as an aspect of the fourth industrial revolution (4IR) [1]. COVID-19 was first found in Wuhan, China, in December 2019 and has spread worldwide. Chinese and global epidemiological studies show person-to-person transfer [2, 3]. Sneezing and coughing spread harmful COVID-19. COVID-19 was spread by touching infected surfaces or hands and then their lips, nose, or eyes. Statisticians must prove their outcomes are not random [4, 5]. Failure prevention and planning are often driven by statistics [6, 7]. Failures in pandemic and infectious diseases control can lead to economic loss and loss of life.

Recently, Mwamnyange et al., [8] proposed a big data analytics platform for childhood infectious illness surveillance and response. The system helps healthcare professionals track, monitor, and analyze infectious illness reports via social media to prevent and control child-related diseases. The proposed technique was validated using use-case scenarios and performance comparisons [9]. However, statistical insights might be gained from data analytics for effective infectious diseases detection, but it was not

ICAIW 2025: Workshops at the 8th International Conference on Applied Informatics 2025, October 8–11, 2025, Ben Guerir, Morocco

[*] Corresponding author.

saajagbe@pgschool.lautech.edu.ng (S. Ajagbe); MuritalaI@unizulu.ac.za (I. Muritala); MudaliP@unizulu.ac.za (P. Mudali); adigunm@unizulu.ac.za (M. Adigun)
0000-0002-7010-5540 (S. Ajagbe); 0000-0002-9254-4077 (I. Muritala); 0000-0003-4720-1726 (P. Mudali); 0000-0001-6256-5865 (M. Adigun)

considered. To provide some concrete examples, artificial intelligence (AI) was used to analyze incident cases on the website. The AI Incident database (2021) compiles news articles from several sources. We discovered that 72 of the 126 occurrences that have been reported thus far may be connected to dependability problems. Based on the 72 reliability-related incidents, despite the ubiquity, it is also observed that out of those 72 occurrences, 29 incidences resulted in fatalities or serious injuries, demonstrating that dependability problems may cause significant damage [3, 10]. From a different angle, public trust is necessary for the widespread use of AI technology.

Big data analytics are needed to manage infectious diseases. Massive data sets require statistics to comprehend, analyze, and predict infectious disease patterns. AI technology is still growing and has difficulties. Challenges include dependability, safety, durability, and trustworthiness. Dependability is important because AI systems must be trusted. The system may work as intended. Resource allocation and public health crisis response depend on these criteria. Examine reliability, strategy, and failures. Recently developed AI dependability statistical analysis, especially for statisticians, highlights statistical research concerns. The authors discussed AI dependability's out-of-distribution detection, training set influence, adversarial attacks, model correctness, and uncertainty quantification. A few examples show study options. A recent study explored AI reliability evaluation, data gathering, and test preparation, and how to improve system designs for AI dependability. Since data is vital for reliability assessment [10]. Data on infectious diseases are initially analyzed using descriptive statistics and AI applications. The key variables, like mean intensity of a color channel, standard error of the mean (SEM), mean correlation, mean homogeneity, mean dissimilarity, mean correlation, and mean density, are crucial to pandemic analysis. Statisticians can map some variables to the geographic and demographic spread of pandemic diseases using big data sources search patterns.

The objective of this study is to investigate the feasibility of effective pandemic investigation using statistical analysis. Specifically, the statistical package for Social Sciences was used for the analysis: mean intensity of a Color channel, SEM, mean correlation, mean homogeneity, mean dissimilarity, mean correlation, and mean density.

This study is structured as follows. Section 2 provides a literature review to emphasize the necessity of the present investigation. Section 3 delineates the approach, attributes of the dataset, and the research instrument. Section 4 presents a comprehensive analysis of statistical insights on big data analytics for the effective detection of infectious diseases. Section 5 finishes the research and proposes future endeavors for the implementation of effective pandemic control utilizing statistical tools.

2. Literature Review

Big Data Analytics is crucial for advancing the Fourth Industrial Revolution and tackling the challenges presented by the pandemic. The Fourth Industrial Revolution is a significant transformation characterized by the convergence of physical, digital, and biological technologies. Currently, Big Data propels substantial progress in automation and robotics, enabling these systems to analyze operational data for enhanced efficiency and flexibility. It also enables the IoT, where interconnected devices across diverse platforms generate vast data that, when analyzed, can improve operational efficiencies and predictive maintenance capabilities. Moreover, in smart manufacturing, Big Data enhances the optimization of production processes, minimizes downtime, and enables real-time customization of output. During the COVID-19 pandemic, Big Data has proven its essential significance across multiple crucial sectors. It has enabled effective epidemiological surveillance by analyzing data from several sources, including travel records, social media, and government reports, to track the virus's spread [11].

In pandemic management, analytics have enhanced hospital resource allocation, forecasted patient outcomes, and tailored treatment options. Furthermore, the rapid progress and distribution of vaccinations have been expedited by Big Data, which has enabled the analysis of massive clinical trial data. It has also impacted public policy, aiding governments in developing specific lockdown measures and informed reopening strategies based on real-time data concerning infection rates and public sentiment. The use of Big Data in addressing COVID-19 and advancing Fourth Industrial Revolution technologies has accelerated the adoption of digital solutions, including telemedicine, remote work, and online education, which are vital to the current digital transformation. This integration highlights the essential role of Big Data in crisis management and as a core component of digital and industrial advancement in the Fourth Industrial Revolution, demonstrating its capacity for comprehensive, real-time decision-making across public health, economic, and industrial sectors [4, 11].

Hong et al. [11] explored statistical methodologies for assessing AI reliability, particularly for autonomous systems. The authors introduce the "SMART" statistical framework to evaluate the robustness of AI models used in various domains, including healthcare. The study highlights how AI reliability decreases in out-of-distribution settings and suggests that statistical models outperform deep learning in long-term failure prediction. However, it lacks real-world infectious disease applications and validation on epidemiological datasets.

Another study highlights AI's role in infectious disease monitoring and projection, emphasizing data-driven approaches for disease surveillance. It demonstrates high sensitivity (91%) in predicting disease outbreaks through AI models but suffers from a lack of statistical validation, poor interpretability, and limited disease-specific analysis [12].

Fei et al. [13] also provided a broad overview of big data analytics applied to healthcare, particularly during COVID-19. It discusses statistical inference challenges in high-dimensional data, highlighting the effectiveness of Bayesian models in long-term trend prediction. However, it has a limited focus on disease classification and weak integration of statistical methods in real-time applications.

Mwamnyange et al. [8] developed a big data analytics framework for tracking childhood infectious diseases using a modified MapReduce algorithm. While it enhances early outbreak detection through social media analytics and reduces processing time for large epidemiological datasets, it lacks statistical validation of detection accuracy and focuses more on data processing efficiency than disease classification. Olaboye et al. [9] presented a paper on policy and technical framework for using big data analytics in epidemic forecasting. It improves forecasting accuracy with multi-source data integration, but has limited classification models for specific diseases and underutilizes machine learning for outbreak detection.

Panah et al. [14] investigated the integration of AI and big data analysis with public health infrastructure for early detection of infectious disease outbreaks. The authors propose an AI-driven data integration framework leveraging social media, wearable technology, and traditional clinical databases. While the paper emphasizes AI's potential in public health surveillance, it lacks a detailed statistical validation framework and real-time adaptive learning methodologies.

Adegoke et al. [15] presented a paper that provides a comprehensive review of data analytics models used for disease outbreak prediction. It evaluates traditional statistical methods such as time-series forecasting and Bayesian inference, alongside AI-based models. The study highlights the importance of integrating epidemiological data with environmental and social media data for robust predictions. However, it does not propose a standardized statistical framework for evaluating model performance across different data sources.

Piontti et al. [16] explored computational models for infectious disease forecasting, focusing on data science methodologies such as agent-based modeling and network theory. It provides insights into how big data can enhance disease spread predictions, but lacks emphasis on statistical uncertainty quantification and model generalizability across varying outbreak conditions.

Zhou et al. [17] investigated an integrated health big data system in China for infectious disease prevention and control to detect dengue fever, tuberculosis (TB), and vaccination gaps in migrant children. The system outperformed traditional surveillance methods by identifying more cases of TB and dengue and more children with incomplete vaccinations.

Michael and Krishnan [18] studied how big data analytics affects healthcare prediction insights. The study refines predictive models for disease progression, risk assessment, and individualized treatment using EHRs, medical imaging, genomic data, and wearable sensors. Case studies show it reduces hospital remissions and optimizes healthcare resource management. AI-driven healthcare analytics is lauded, but real-time flexible models and scalability across healthcare systems are not.

Table 1 contains a summary of the review of the related studies on big data analytics for pandemic classification.

Table 1
Summary of the review of the related studies on big data analytics for pandemic classification.

Ref.	Methodology	Dataset Type	Results	Limitations
Hong et al.,[11]	SMART framework, survival analysis, hazard functions	Simulated AI reliability datasets	AI reliability decreases in out-of-distribution; statistical models perform better in failure prediction	No epidemiological validation, not focused on disease detection
Wong et al.,[12]	AI-based ML (SVM, RF, Neural Networks), social media and genomic data fusion	Syndromic surveillance, search trends, clinical data	91% sensitivity in outbreak prediction	Lacks statistical validation, poor model interpretability
Fei et al., [13]	High-dimensional statistical models (LASSO, Bayesian Inference), computational epidemiology	COVID-19 epidemiological data, EHRs	Bayesian models outperform AI in long-term trend prediction	Limited disease classification focus, weak integration of statistical methods
Mwamnyange et al., [8]	MapReduce, Hadoop-based data integration	Clinical records, social media, surveillance data	Faster epidemiological data processing, early outbreak detection	Lacks statistical validation, prioritizes data processing over classification
Panah et al., [14]	AI-driven data integration, public health surveillance, wearable data fusion	Social media, wearable tech, clinical databases	Enhanced real-time disease tracking	Lacks statistical validation, no adaptive learning methodology
Adegoke et al., [15]	Investigation of statistical and AI models (Bayesian inference, time-series, ML)	Epidemiological, environmental, social media data	Identifies gaps in predictive analytics models	No standardized framework for model comparison
Piontti et al., [16]	Agent-based modelling, network theory, computational epidemiology	Global epidemiological and mobility data	Insights into disease spread via big data	Lacks statistical uncertainty quantification, limited model generalizability
Zhou et al., [17]	Investigates an integrated health big data system for detecting and preventing infectious diseases.	Empirical study with comparative analysis.	Health data from clinics, hospitals, and government records in China.	AI-based screening, big data analytics, disease detection algorithms.
Michael & Krishnan, [18]	Investigates the role of data analytics in predictive healthcare, focused on risk assessment and treatment personalization.	Case study analysis with machine learning models applied to patient data.	EHRs, medical imaging, genomic sequencing, wearable sensor data from City Hospital.	ML predictive modeling, statistical analysis, Apache Hadoop, Spark,Python, Tableau.

3. Materials and Methods

The decision-making process is significantly enhanced by the application of statistical methods, even in instances involving extensive data sets, sometimes referred to as big data [1, 10]. To implement data processing techniques in certain sectors, it is essential to delineate the many types of data, encompassing volume, variety, and velocity. Distinct types of data might be produced from multiple sources, and it is essential to create systems capable of managing the characteristics of data. Figure 1 illustrates the established architecture for big data analytics aimed at the efficient classification of infectious diseases by statistical methodologies. The developed architecture depicts a complete classification and detection procedure of the presence of COVID-19 using a synthetic dataset. It consists of three phases: (i) Pandemic dataset acquisition and source phases, (ii) Data preparation phase, and (iii) Statistical analysis of COVID-19 pandemic data diagnosis.

3.1. Pandemic Dataset Acquisition and Source

The COVID-19 pandemic dataset utilized in this investigation was obtained from a prominent source known as the Kaggle database *(https://www.kaggle.com/datasets/rishanmascarenhas/covid19-temperatureoxygenpulse-rate)*. The distribution of the pandemic datasets sample in this analysis which had values ranging from 0 to 9. The details include the ID, Oxygen, Pulse Rate, Temperature, and results $(+/-)$. In total, the dataset contains 5010 negative and 4990 positive examples that support replicability, making it large enough to be classified as big data. Figure 2 shows the distribution of the datasets by classes and evidence of balanced classes using explorative data analysis. There is no ethical concern regarding this dataset because it was sourced from opensource and has been deanonymized to protect privacy and security.

3.2. Data Preparation

Framing the research as a classification problem: Let $S = \{(x_1, y_1), (x_2, y_2), ...(x_n, y_n)\}$ constitute the collection of training cases of size d. $Y = \{y_1, y_2..., y_n\}$ be the set of labels where x_i is a feature with corresponding y_i label, and Y is a set of features according to classification task definition. The preliminary phase of this data analysis and categorization commenced with the data preparation; the dataset suffers from missing data and duplicated data. This was succeeded by a crucial step of eliminating non-relevant data from the converted raw dataset. Duplication and missing values were addressed using an imputation technique, wherein the meaning of the respective column was calculated and used to replace the missing values. The issue was resolved through the identification of essential aspects, conducted in accordance with [1, 3].

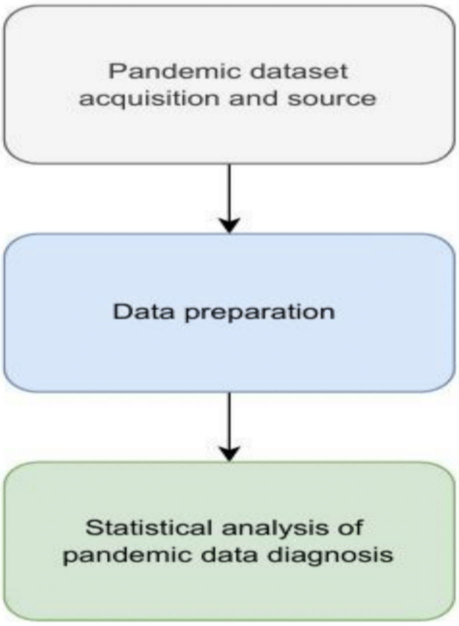
Pandemic dataset acquisition and source
Data preparation
Statistical analysis of pandemic data diagnosis

Figure 1: A block diagram for statistical big data analytics for pandemic classification.

3.3. Statistical analysis of pandemic data diagnosis

The statistical data analysis using one-way analysis of variance. Means were separated using Tukey. Normality test was carried out using the Shapiro-Wilk test, and homogeneity of variances was verified with Levene's test implemented in SPSS v 21.0 (IBM Corp., Armonk, NY, USA). This was accomplished using Intel(R) Core (TM) i7-4600U CPU running from 2.10GHz to 2.70GHz on a Windows 10 professional with 8GB RAM. The statistical analysis was to validate the use of big data analysis for pandemic preparedness, with a working hypothesis that the intensity of the colour channel images would be different $(p < 0.05)$ between COVID-19 patients and patients with other health conditions.

In statistics, statistical significance denotes the likelihood that the outcomes of a study or experiment are attributable to factors other than random chance. Various metrics and tests are typically employed to ascertain the statistical significance of a result. This research utilized primary criteria for statistical significance contingent upon the specific type of pandemic dataset examined. Statistical significance is a crucial concept; nonetheless, it is essential to recognize that statistical significance does not equate to practical relevance. Always consider the context of the findings and additional metrics, such as effect size, when evaluating results. The equations shown below describe these metrics and express them statistically, using mean intensity of a Color channel, SEM, mean correlation, mean homogeneity, mean dissimilarity, mean correlation, mean correlation and mean density. The detailed description of statistical metrics used, with accompanying formulas are presented in this section.

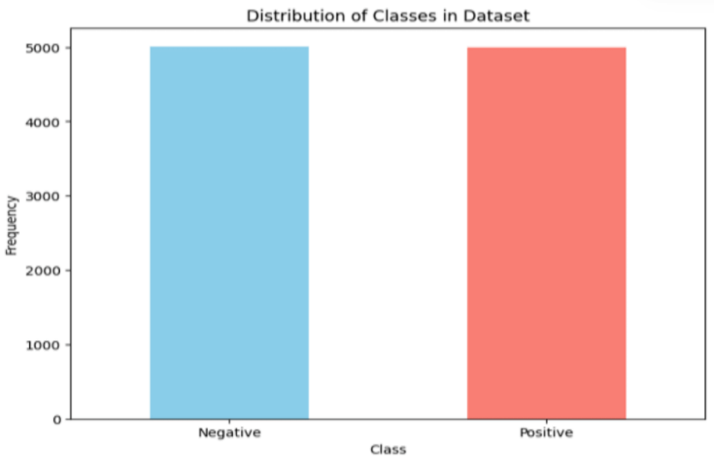

Figure 2: Datasets class distribution.

Equation 1 shows the average intensity of a color channel in the pandemic classification.

$$I_{mean} = \frac{1}{N}\sum_{j=1}^{N} Ii \tag{1}$$

where Ii is the intensity of the pixel in the color channel, and N is the total number of pixels.
Equation 2 shows the Standard Error of the Mean (SEM) for pandemic classification.

$$SEM = \frac{\sigma}{\sqrt{N}} \tag{2}$$

where σ is the standard deviation of pixel intensities and N is the number of pixels for pandemic classification.

Equation 3 shows the Gray Level Co-occurrence Matrix (GLCM), a statistical method for examining texture that considers the spatial relationship of pixels. Mean Contrast (from Gray Level Co-occurrence Matrix- GLCM)

$$Contrast = \sum_{i,j} P(i,j)(i-j)^2 \tag{3}$$

where $P(i,j)$ is the probability of intensity pairs in the GLCM for pandemic classification.
Equation 4 shows Mean Homogeneity;

$$Homogeneity = \sum_{i,j} \frac{P(i,j)}{1+|i-j|} \tag{4}$$

Equation 5 shows Dissimilarity

$$Dissimilarity = \sum_{i,j} P(i,j)|i-j| \tag{5}$$

Equation 6 shows Correlation

$$Correlation = \frac{\sum_i (i-\mu_i)(i-\mu_j)P(i,j)|}{\sigma_i \sigma_j} \tag{6}$$

where μ and σ are the means and standard deviations of intensity levels for pandemic classification.
Equation 7 shows Mean Density (assuming it refers to pixel density in a binary image)

$$Density = \frac{\sum Ibinary}{N} \tag{7}$$

where $Ibinary$ represents pixel values in a binary image for pandemic analysis.

4. Result and Analysis

4.1. Results

This section presents the results attained from the statistical analysis of the pandemic dataset. The task is conceptualized based on the statistical tools called for a detailed analysis of the COVID-19 pandemic dataset. A statistical analysis was conducted to categorize infectious diseases, including COVID-19. A model was developed utilizing data from Kaggle, encompassing steps of data gathering, preparation, analysis via SPSS, and essential statistical measures (intensity, SEM, contrast, homogeneity, dissimilarity, correlation, and density). The findings are articulated through sophisticated statistical metrics, including the mean intensity of a color channel, the standard error of the mean (SEM), mean correlation, mean homogeneity, mean dissimilarity, repeated mean correlation, and mean density.

4.2. Analysis of big data analytics for pandemic classification.

Statistics on the pandemic dataset in this study are discussed in this subsection. It delineates the statistical analysis of four pandemic data scenarios/conditions. Data is divided into COVID-19, Normal, Pneumonia, and Tuberculosis. For each occurrence, it includes statistical measurements of image attributes, including the mean and standard error. It is a clear presentation of results with mean values and standard errors; Table 2 is the comparison of the blue channels in images across different health conditions, and Table 3 presents the comparison of green channels in the image obtained for different health conditions. While Table 4 compares the red channel in the image obtained for different health conditions, Table 5 compares the contrast of the image obtained for different health conditions. The focus of Table 6 is the comparison of dissimilarity of the image obtained for different health conditions, and Table 7 shows the comparison of homogeneity of the image obtained for different health conditions. Table 8 compares the correlation of the images obtained for different health conditions, and finally, Table 9 compares the correlation of the images obtained for different health conditions. In COVID-19 and TB cases, blue and green channels have a higher mean intensity than Normal and Pneumonia channels. All three channels show similar and reduced mean intensities for pneumonia and normal patients.

Table 2
Comparison of the blue channels in images across different health conditions.

Case	Sample size	Mean intensity ± SEM	p-value	95% CI
COVID-19	554	134.87 ± 0.94^a		
Normal	1779	122.15 ± 0.33^b	< 0.001	124.29-125.24
Pneumonia	4061	122.91 ± 1.05^b		
Tuberculosis	729	133.77 ± 1.05^a		

a,b: indicate significant difference (p<0.05); CI: confidence interval.

Table 3
Comparison of green channels in the image obtained for different health conditions.

Case	Sample size	Mean intensity ± SEM	p-value	95% CI
COVID-19	554	134.59 ± 0.94^a		
Normal	1779	122.15 ± 0.33^c	< 0.001	123.80-124.74
Pneumonia	4061	122.91 ± 0.31^c		
Tuberculosis	729	129.22 ± 1.02^b		

a,b,c: indicate significant difference (p<0.05); CI: confidence interval.

Table 4
Comparison of red channel in the image obtained for different health conditions

Case	Sample size	Mean intensity ± SEM	p-value	95% CI
COVID-19	554	134.29 ± 0.96^a		
Normal	1779	122.15 ± 0.33^b	< 0.001	122.49-123.46
Pneumonia	4061	122.91 ± 0.31^b		
Tuberculosis	729	116.79 ± 1.15^c		

a,b,c: indicate significant difference (p<0.05); CI: confidence interval.

Table 5
Comparison of contrast of the image obtained for different health conditions

Case	Sample size	Mean intensity ± SEM	p-value	95% CI
COVID-19	554	134.36 ± 4.37^b		
Normal	1779	163.81 ± 1.08^a	< 0.001	129.87-132.76
Pneumonia	4061	121.02 ± 0.86^c		
Tuberculosis	729	107.79 ± 2.34^d		

a,b,c,d: indicate significant difference (p<0.05); CI: confidence interval.

Table 6
Comparison of dissimilarity of the image obtained for different health conditions.

Case	Sample size	Mean intensity ± SEM	p-value	95% CI
COVID-19	554	0.97 ± 0.00^b		
Normal	1779	0.97 ± 0.00^b	< 0.001	0.978-0.979
Pneumonia	4061	0.98 ± 0.01^a		
Tuberculosis	729	0.97 ± 0.00^b		

a,b: indicate significant difference (p<0.05); CI: confidence interval.

Table 7
Comparison of the homogeneity of the image obtained for different health conditions.

Case	Sample size	Mean intensity ± SEM	p-value	95% CI
COVID-19	554	0.26 ± 0.00^c		
Normal	1779	0.23 ± 0.00^d	< 0.001	0.269-0.272
Pneumonia	4061	0.27 ± 0.00^b		
Tuberculosis	729	0.38 ± 0.00^a		

a,b,c,d: indicate significant difference (p<0.05); CI: confidence interval.

Table 8
Comparison of the correlation of the images obtained for different health conditions.

Case	Sample size	Mean intensity ± SEM	p-value	95% CI
COVID-19	554	11.62 ± 0.03^b		
Normal	1779	12.06 ± 0.01^a	< 0.001	11.61-11.64
Pneumonia	4061	11.58 ± 0.01^b		
Tuberculosis	729	10.78 ± 0.03^c		

a,b,c: indicate significant difference (p<0.05); CI: confidence interval.

The blue, green, and red channel images from COVID-19 patients showed significantly $(p < 0.001)$ the highest intensity, compared to patients with different health conditions as indicated in Table 2, 3 and 4, respectively. Mean contrast of the image from Covid-19 patients (134.36 ± 4.37) was significantly lowered $(p < 0.001)$ compared with mean contrast of the image from the normal patients (163.81 ± 1.08) in Table 5.

Additionally, the mean dissimilarity and homogeneity indicate that Pneumonia and tuberculosis demonstrate the highest dissimilarity (0.98 ± 0.01) and homogeneity (0.38 ± 0.00), respectively, signifying greater diversity in pixel intensity as shown in Table 6 and 7. COVID-19 and Pneumonia

Table 9
Comparison of the correlation of the images obtained for different health conditions

Case	Sample size	Mean intensity ± SEM	p-value	95% CI
COVID-19	554	10.64 ± 0.34^c		
Normal	1779	19.20 ± 0.12^a	< 0.001	13.48-13.78
Pneumonia	4061	12.16 ± 0.08^b		
Tuberculosis	729	10.48 ± 0.21^c		

a,b,c: indicate significant difference (p<0.05); CI: confidence interval.

exhibit comparable homogeneity ($0.26 - 0.27$; Table 7). The mean correlation indicates that Normal images exhibit the highest correlation (12.06 ± 0.01) as shown in Table 8. The mean density study indicates that normal cases demonstrate the highest mean density (19.20 ± 0.12; Table 9). Meanwhile, the mean contrast, dissimilarities, homogeneity, correlation and density were lower $(p < 0.05)$ for COVID-19 patients than normal patients.

5. Conclusion Recommendation

The statistical data analysis has changed infectious and pandemic illness research and management. Pattern identification and analysis of infectious and pandemic illness epidemics have underutilized statistical approaches. The global pandemic data study lacks statistical analysis. For pandemic analysis, state-of-the-art statistical methods like mean intensity of a color channel, SEM, mean correlation, homogeneity, dissimilarity, correlation, and density provide depth of finding for informed decision-making. This research shows that pandemics are straightforward to control and minimize mortality if statistical methods are used to analyze the pandemic dataset to advise doctors, policymakers, and other stakeholders on disease spread. This study examines statistics, data science, machine learning, and AI in relation to environmental science, natural sciences, medicine, and technology. It shows how statistics may model and analyze huge data in pandemics, covers the state-of-the-art for practical analysis, and shows how to implement methodologies. The paper makes a more substantial contribution to the field of pandemic analytics. Future research may compare results using theoretical machine learning techniques, ML modeling, and validation, to validate the result of this investigation. Multiple data sources enable a more complete picture of disease spread. Maps of COVID-19 hotspots were created using social media, contact tracing apps, and real-time hospital data.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. A. Ajagbe, P. Mudali, M. O. Adigun, Assessing data-driven of discriminative deep learning models in classification task using synthetic pandemic dataset, in: Southern African Conference for Artificial Intelligence Research, Springer, 2024, pp. 282–299.

[2] O. Akinlade, E. Vakaj, A. Dridi, S. Tiwari, F. Ortiz-Rodriguez, Semantic segmentation of the lung to examine the effect of covid-19 using unet model, in: International Conference on Applied Machine Learning and Data Analytics, Springer, 2022, pp. 52–63.

[3] O. Ugbomeh, V. Yiye, E. Ibeke, C. P. Ezenkwu, V. Sharma, A. Alkhayyat, Machine learning algorithms for stroke risk prediction leveraging on explainable artificial intelligence techniques (xai), in: 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), volume 1, IEEE, 2024, pp. 1–6.

[4] V. Yiye, O. Ugbomeh, C. P. Ezenkwu, E. Ibeke, V. Sharma, A. Alkhayyat, Investigating key contributors to hospital appointment no-shows using explainable ai, in: 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), volume 1, IEEE, 2024, pp. 1–6.

[5] S. A. Ajagbe, P. Mudali, M. O. Adigun, An empirical assessment of discriminative deep learning models for multiclassification of covid-19 x-rays, 2024.

[6] J. Min, X. Song, S. Zheng, C. B. King, X. Deng, Y. Hong, Applied statistics in the era of artificial intelligence: A review and vision, arXiv preprint arXiv:2412.10331 (2024).

[7] H. Florez, S. Singh, Online dashboard and data analysis approach for assessing covid-19 case and death data, F1000Research 9 (2020) 570.

[8] M. Mwamnyange, E. T. Luhanga, S. R. Thodge, Big data analytics framework for childhood infectious disease surveillance and response system using modified mapreduce algorithm, International Journal of Advanced Computer Science and Applications (IJACSA) (2021).

[9] G. T. Igwama, J. A. Olaboye, C. C. Maha, M. D. Ajegbile, S. Abdul, Big data analytics for epidemic forecasting: Policy frameworks and technical approaches, International Journal of Applied Research in Social Sciences 6 (2024) 1449–1460.

[10] M. N. Sadiku, S. A. Ajayi, J. O. Sadiku, Predictive analytics for supply chain, Int J Trend Res Dev (IJTRD) 12 (2025) 112.

[11] Y. Hong, J. Lian, L. Xu, J. Min, Y. Wang, L. J. Freeman, X. Deng, Statistical perspectives on reliability of artificial intelligence systems, Quality Engineering 35 (2023) 56–78.

[12] Z. S. Wong, J. Zhou, Q. Zhang, Artificial intelligence for infectious disease big data analytics, Infection, disease & health 24 (2019) 44–48.

[13] Z. Fei, Y. Ryeznik, O. Sverdlov, C. W. Tan, W. K. Wong, An overview of healthcare data analytics with applications to the covid-19 pandemic, IEEE Transactions on Big Data 8 (2021) 1463–1480.

[14] H. Rasouli Panah, S. Madanian, J. Yu, Integration of ai and big data analysis with public health systems for infectious disease outbreak detection, 2023.

[15] B. O. Adegoke, T. Odugbose, C. Adeyemi, Data analytics for predicting disease outbreaks: A review of models and tools, International journal of life science research updates [online] 2 (2024) 1–9.

[16] A. Pastore y Piontti, N. Perra, L. Rossi, N. Samay, A. Vespignani, Infectious disease spreading: From data to models, in: Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age, Springer, 2018, pp. 3–10.

[17] X. Zhou, E. W. J. Lee, X. Wang, L. Lin, Z. Xuan, D. Wu, H. Lin, P. Shen, Infectious diseases prevention and control using an integrated health big data system in china, BMC infectious diseases 22 (2022) 344.

[18] H. Michael, S. Krishnan, Big data analytics for big outcomes in healthcare, in: 2019 ASEE Midwest Section Conference, 2020.