Big Data Analytics Challenges and Opportunities in Heart Disease Recognition: Novel Dimensionality Reduction with Classification Approach

Dr. Anupa Sinha[1], Yalakala Dinesh Kumar[2]

[1]*Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India*
[2]*Research Scholar, Department of CS & IT, Kalinga University, Raipur, India*

KEYWORDS

Big Data, Challenges, Modified Principal Component Analysis, Modified Deep Convolutional Neural Network (MDCNN)

ABSTRACT

Due to the current technological growth, a number of strategies have been developed, and more are being developed to eliminate problems that arise in many fields. Big Data techniques are employed to effectively stored health data due to the continual and massive volume of data created by the human body. Furthermore, the most important procedure is the classification of health data since it must be carried out precisely in order to diagnose cardiac disease early. The database images are various in size to reduce the dimension the Modified Principal Component Analysis (MPCA) Algorithm is used. the proposed MPCA algorithm is act as a feature selection model to pick features. One of the best and most effective techniques for classifying medical data is the Modified Deep Convolutional Neural Network (MDCNN). It has been shown to work for a variety of hospitalized patients. Consequently, the simulation results show that this proposal enhances classification accuracy in experimental research for the detection of heart ailment. Hence, the proposed method leads to an efficient usage of the resources and cost reduction. This approach assists the physician in taking suitable decision for giving a better treatment at right moment.

1. Introduction

The systematic provision of medical treatment to individuals or communities in order to ensure proper health is known as health care. Healthcare systems [1] are designed and formulated based on needs in a given community. The resources available differ from place to place, and it is also important that in a healthcare system with limited resources, the community health goals should be achieved in an efficient way [8]. A non-compromising quality healthcare system is must for gaining health equality for all, irrespective of individual social status and also for promoting a sustained good health, preventing and controlling diseases morbidity and mortality. Quality healthcare services can be achieved by inculcating digital healthcare technologies, which include artificial intelligence, nano technology, virtual reality, use of robots and robotic machines, 3D printing etc. They have a huge impact in future on health care. In today's active society people with various work profile also mandate health monitoring system through routine checkups every 3 months / 6 months in a year. This has given rise to many diagnostic centers with door services. By this, huge amount of data is accumulated at the back end in the cloud or in the physical servers. Technology has penetrated into health care to the extent tomorrow robot's / smart devices will replace doctor's for pre diagnosis. Today many smart models are available in society which do not take the support of doctors / nurses [2].

Health care is an application to computing areas like Pattern recognition, machine learning, Artificial Intelligence and Data Sciences. Earlier to these buzz words in technology, health care data was getting stored in devices which were a combination of electrical, electronic and mechanical concepts. These were being used to capture images and monitor patients' clinical readings. Big Data has already drawn a remarkable attention for current researcher's frontiers [4]. More attention is given to the big data because data collection has become much cheaper now. Data is growing tremendously as it is generated by low-priced several information-sensors like mobile devices, wireless sensor networks, cameras, etc. "Big Data and Hadoop" employ an important role in analyzing huge quantity of data in a real-time. The health checkup data is the most imperative dataset in research field [19].

Big Data

"Big Data: it's not the data". In mid 1990s, the word "big data" is emanated in the lunch-table conversation at Silicon Graphics Inc. It became widespread in 2011 [15]. Over the years, the definitions of big data is evolving leading to the confusion in the mind of researchers. 7Vs of BD as shown in

Figure 1.

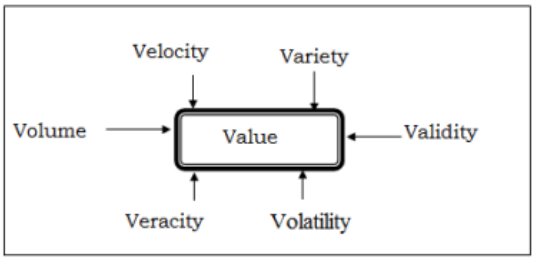


Figure 1. 7 Vs of Big Data

The goal is to advance current business intelligence (BI) trends that are more profound. Big data analytics has several advantages for businesses, as shown in Figure 2, including a significant increase in operational efficiency, improved customer service, the discovery of new and diverse items in a cutthroat market, and more.

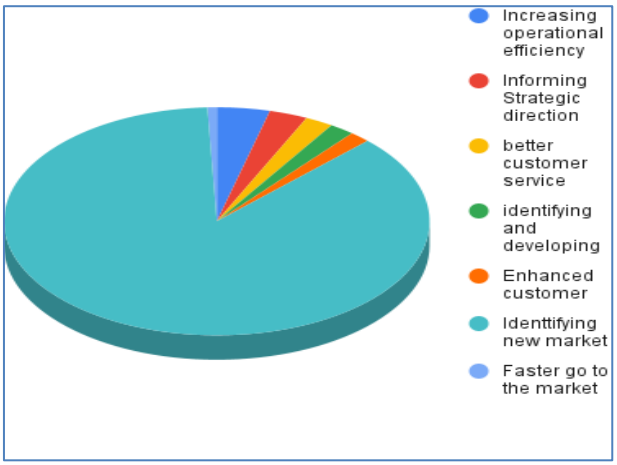


Figure 2. Big Data Analytics Opportunities

There are multiple sections to the paper. The pertinent study conducted by several different researchers is reviewed in the section II that follows. A description of the suggested system's framework is given in part III. Section IV provides a description of the experimental outcome and remarks. Section V provides a conclusion to the suggested task.

Literature Review

To improve clinical care, produce new knowledge, and streamline public health surveillance, Roberta Pastorino et al. [5] give an overview of best practise projects relating to Big Data analytics in the public health and oncology sectors. Sebastian Salas et al. [15] examined many concerns linked to health care big data, such as privacy and data protection, availability of information, consistency of data, interoperability, management, and governance, and offered a balanced framework to handle the issues. Clemens Scott Kruse's evaluation [7] set out to outline the drawbacks of big data analytics as well as the opportunities it offers the healthcare industry. Mumtaz Karatas et al.'s main objective [16] is to give readers a review of articles that touch on Industry 4.0, Big Data, and healthcare operations, together with insights into the future. Based on our analysis, BD is one of the key Industry 4.0 technologies available to the healthcare industry. The review of literature by Nishita Mehtaa and Anil Pandit [9] tries to determine the breadth of Big Data in health care, as well as its applications and hurdles in adoption. Establishing (big) information management, creating a knowledge transfer culture, implementing security measures, and training key individuals to apply Big Data analytics are some of the hurdles [10]. They also want to figure out how to overcome the difficulties. The goal of the study by Rakesh Raja et al. [20] is to assess the potential applications and adoption hurdles of big data analytics in healthcare, as well as the extent of its use [17]. The essay also covers the massive

amounts of data generated by different healthcare systems, its features, the numerous challenges associated with managing big data, and the ways in which big data analytics helps to acquire insightful knowledge from these data sets. In summary, the essay reviews the literature on big data in healthcare and lays the groundwork for future studies in the field. The impact of big data in healthcare and the several Hadoop technologies available to handle it are discussed by Sunil Kumar and Maninder Singh [11]. We also look at the conceptual architecture of clinical decision support systems, EHRs, text/imagery, and the history of data collection across several departments in big data analytics for healthcare.

Big Data Challenges in Health Care

Big data comprises of a large and complicated set of data that might be structured or unstructured. Data is continually created from a variety of sources, including research institutes, GPS data, and medical records. Health care data is also big data which is characterized by these 4 v's as depicted below.

Volume [11]: Health care data is voluminous in nature which debates the volume of characteristic of big data. For instance, pulse monitoring system records data every second for a patient in ICU. This may happen in days or months or years. Analyzing this data is only possible by transforming data from one space to another.

Variety [11]: The data format of same attribute may vary as per the device manufactures and sometimes, it is also possible during conversion of unstructured to structured data may generate variety of data posing challenge to analysis [3].

Velocity [11]: The rate at which data is generated may also go incomparable as they are time sensitive, time can also be considered as integral part of the data. Many times, data is plotted on a x-y plot, which time is one of the axes to monitor time progression in patient.

Veracity [11]: It relates to the information gained from various dataset's trustworthiness, dependability, noise, and uncertainty. Unstructured data is, in general, very changeable and frequently incomplete

In healthcare, big data analytics faces a number of issues, including data storage, adequate data collecting, and summarization. The structuring of the resulting data after it has been extracted from several layers is also a difficult undertaking [12]. Authentication, archiving, safeguarding, information retrieval, administration, and representation are some of the sub-processes. In addition, security standards and preservation mechanisms should be used to verify the quality of information at each level

- Maintaining a longitudinal correlation between the patient and the records.
- In order to make use of unstructured data, understanding of data is important.
- Analyzing the substance of the missing health data.
- Imaging data has a high dimensionality difficulty.
- Maintaining patient identification privacy and security.
- Analyzing genetic data is also a difficult undertaking that necessitates additional storage and processing resources.
- Several sensors are used to record the patient's behavior and social interactions.

Many organizations and individuals can benefit greatly from Big Data concept. Information science, data modelling, cloud services, machine learning, statistical modelling, and pattern recognition are all examples of big data applications. The appropriate and effective integration of such technology and analytical models aids in forecasting of future deviations [18]. However, since most big data tools are available as open source online, they might serve as an optional key for intruders and attackers. The benefits of big data applications also can expose a number of security and privacy concerns. When it comes to patient privacy and security, there needs to be a mechanism in place that assures patient

medical and personal information is secured and the security continues when used in analytical modelling. As a conclusion, the primary focus will be on open research challenges in medical big data analytics, with researchers paying specific attention to relevant key points. The research objectives in this research are:

- The proposed framework is used for detection and prediction of heart disease.
- This frame work builds an inexpensive device for detecting the early-stage cardiac problems [6]. It also detects Endothelium dysfunction and Arthroscleroses.
- To test and validate the performance of proposed framework using various parameters like total time, accuracy, etc.
- To propose a Dimensionality Reduction (DR) based deep learning framework which consists of features like preprocessing and classification etc.
- To study and compared existing machine learning algorithms for data analytics

2. Methodology

In this research work a detailed framework has been proposed to encapsulate the gap in healthcare analytics to data acquisition. Data collected from various devices and various diagnostic centres are collected in a time sorted manner to understand the sequence of occurrences. The data passes through a process of clean up to avoid redundant and missing data, A detailed framework is depicted pictorially to indent the purpose of this research work in the figure 3.

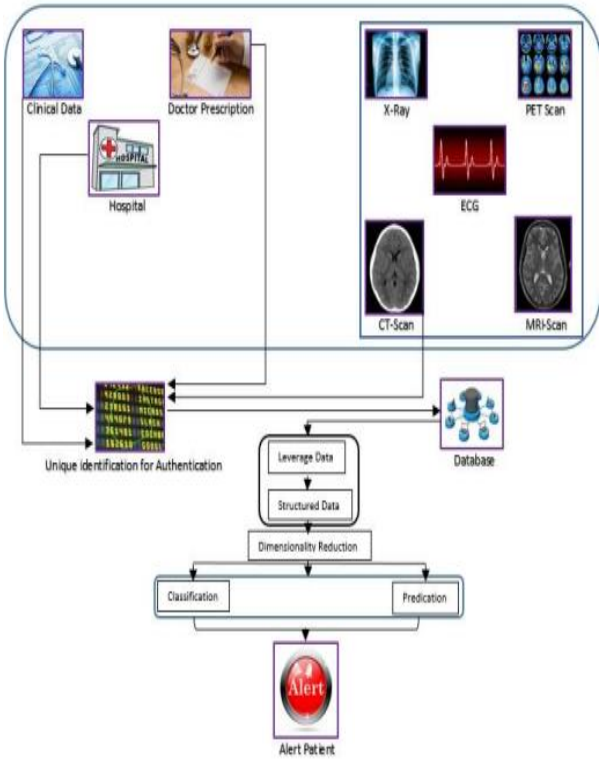


Figure 3. Proposed framework

Initially the images are obtained from the database. Then the imagers are cleaned for further step. The DR Classifiers are the algorithms utilized in deep learning techniques.

The data has several dimensions and levels of complexity. Consequently, a technique that may reduce the complexity of data is required. The aim of complexity reduction is to minimise classification errors. One kind of complexity reduction that is accomplished using the Modified Principal Component

Analysis (MPCA) algorithm is called dimensional reduction. By leveraging previously acquired eigenvectors and eigenvalues to retrieve the attributes of the dataset, multidimensional reductions with MPCA can reduce dimensional complexity [13].

The linear transform is expressed as:

$$Y = TX \qquad (1)$$

X and Y represent the original and modified vectors, respectively, when T is the transform matrix. The following expression can be used to solve the transformation matrix T:

$$(\lambda I - S)U = 0 \quad (2)$$

Where I, S, U, and λ are the square matrix, the covariance matrix, and the eigenvalues.

The transformation matrix T' was written by following equations:

$$Y' = T'X \qquad (3)$$

It is evident from equations 3 that there is a discrepancy in the transform matrix and, eventually, in the sample that was utilised to determine the covariance matrix.

Here MDCNN classifier is used for classification tasks [14]. In this way, every supplied feature is received by the MDCNN classifier. Every input has weights associated to it that are randomly assigned values. The input parameter and the linear function of all connected input layers are added by the hidden nodes in the subsequent hidden layer. To get the desired outcome, the backpropagation approach is improved by using random edge weights. This is how optimisation works. The activation technique is then used to send the output of this layer to the next layer.

The provided feature values and their corresponding weights are in equations (4) and (5):

$$G_i = \{G_1 . G_2 \ldots . G_n\} \text{ (4)}$$

$$R_i = \{R_1 . R_2 \ldots . R_n\} \text{ 5)}$$

F stands for the input value, $G_1 . G_2 \ldots . G_n$, which indicates n selected features, and Wi stands for Fi i's weight value, which defines the n matching values of $R_1 . R_2 \ldots . R_n$.

Then represent the provided feature values and their corresponding weights are in equations (6).

$$AF_i = C_{i \sum_{i=1}^{n} F_i W_i} \quad (6)$$

Here, the exponential of Fi is specified by Ci, and the activation function Afi is specified. The aforementioned equations are applied to every layer of the MDCNN. Lastly, sum the values of all the input signals to calculate the output values for the output layer neurons.

$$R_i = B_i + \sum(o_i W_j) \qquad (7)$$

where Oi is the layer value of the previous output layer, Bi represents the hidden values of the layer, and Ri represents the output values of the previous equation.

$$Wc_i = \alpha \delta_i(F_i) \qquad (8)$$

Error diffusion, momentum, and weight correction are represented across the network by Wc_i, and error is represented by δ_i.

Performance Measures

The statistical methods are Sensitivity, Specificity, Accuracy, Precision, and Border error.

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (9)$$

$$Specificity = \frac{TN}{TN+FP} \times 100\% \quad (10)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad (11)$$

$$Border\ Error = \frac{FP+FN}{TP+FN} \times 100\% \quad (12)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100 \quad (13)$$

3. Results and discussion

This section will analyze the proposed MPCA based MDCNN model in terms of recall, accuracy, sensitivity, specificity, and precision. Furthermore, datasets with differing numbers of diabetes and heart disease cases are used to validate the results. A computer equipped with a 1 TB hard drive for file storage and 16 GB of RAM was used to evaluate the suggested approach.

Table 1 overall analysis

Algorithm	Precision	Recall
RNN	90.3	80.3
RBN	89.2	84.4
proposed	**98.4**	**98.7**

Table 1 and Figure 4 compare the recall and precision of the proposed MDCNN-based heart disease prediction and diagnostic system with the existing RNN and RBN systems.

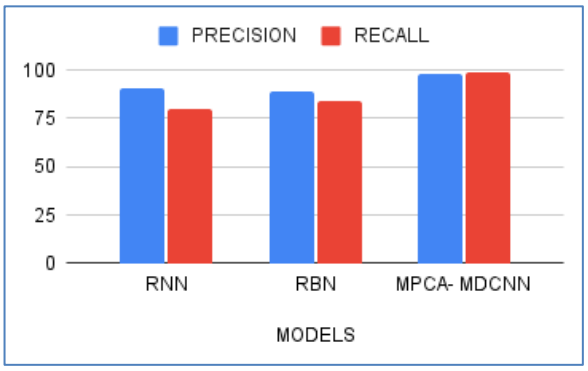
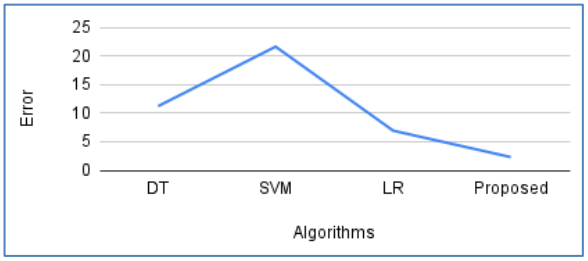
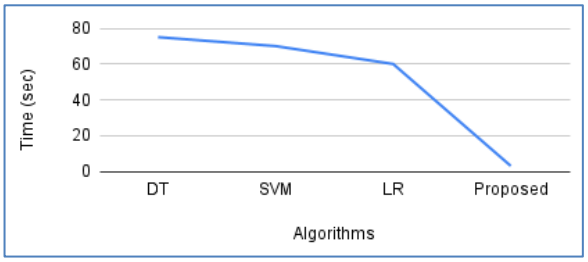


Figure 4. Performance analysis of various



(a)



(b)

Figure 5. comparison of error rates and time

Figure 5 displays the error rate and temporal complexity of the suggested and current methods.

Table 2 Performance Of Various Classification Algorithms.

Algorithm	DT	SVM	LR	BOVW based MDCNN
Sensitivity	92.34	85.322	97.42	98.78
specificity	83.34	84.4218	92.04	98.95
Accuracy	82.48	73.324	93.04	97.7

Figure 6 displays the average classification analysis of the proposed MPCA based MDCNN model utilising the relevant heart disease dataset. The graph shows that the MPCA-based MDCNN model performed better than other comparison techniques, with an average accuracy of 97.7%, specificity of 96.95%, and sensitivity of 97.78%. The SVM and DT models exhibit a moderate degree of accuracy over LR.

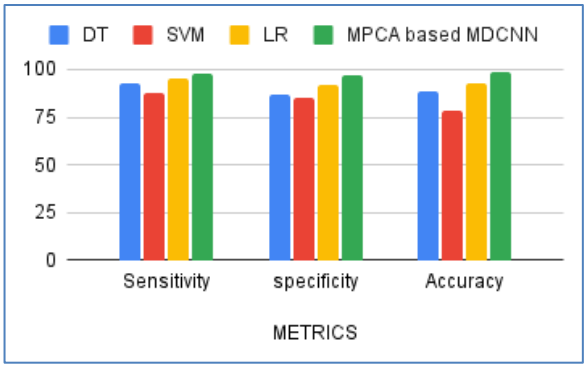


Figure 6. The average classifier results

Concurrently, the LR and suggested MPCA based MDCNN models yielded comparable but marginally accurate results. But in terms of classification accuracy and performance, the suggested MPCA based MDCNN model that was given performed better than the competitors. At 97.7%, the MPCA based MDCNN model demonstrated a high degree of accuracy. To reduce the use of humans as a resource in hospitals and clinics, technology innovation is required. This demand has spurred a slew of technology researchers to create clever algorithms that recognize, classify, and predict health-care data events. The low computing complexity of the SVM – machine learning algorithm gives it an advantage over today's most popular deep architectures. We are currently studying the integration of the MPCA with deep learning on mobile devices, as well as testing multiple models for improved heart disease prediction accuracy.

4. Conclusion and future scope

One of the leading causes of death is heart disease, which requires an accurate diagnosis early on in order to receive treatment. Deep learning algorithms are utilized in disease diagnosis. In this sense, the suggested methodology also makes use of a deep learning algorithm to anticipate cardiac disease before

it becomes worse. The suggested method involves the big data analytics with pre-processing the input data before feeding it into feature selection to choose the relevant feature. In this technique, optimal unsupervised feature selection algorithm is employed for identifying the relevant feature. Unsupervised feature selection is used because, by learning the knowledge, it solves problems and classifies the data without the need for labels. When choosing features for the heart disease model's diagnostic outcome, the MPCA performs better when the suggested MDCNN is applied. The effectiveness of the suggested BOVW based MDCNN model was confirmed with medical data. The study found that the proposed Modified was able to predict cardiac disease with a maximum accuracy of 97.7 percent and diagnose it with a maximum precision of 98.1 percent.

Reference

[1] Rumsfeld, John S., Karen E. Joynt, and Thomas M. Maddox. "Big data analytics to improve cardiovascular care: promise and challenges." *Nature Reviews Cardiology* 13, no. 6 (2016): 350-359.

[2] Rehman, Arshia, Saeeda Naz, and Imran Razzak. "Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities." *Multimedia Systems* 28, no. 4 (2022): 1339-1371.

[3] Yashir Ahamed, M., Lalthlamuanpuii, R., Chetia, B., Lallawmawmi, & Lalngaizuali. (2023). Usage of Medical Library Resources: A Study in the Regional Institute of Medical Sciences, Imphal. *Indian Journal of Information Sources and Services*, *13*(2), 1–6.

[4] Amin, Rashid, Mohammed A. Al Ghamdi, Sultan H. Almotiri, and Meshrif Alruily. "Healthcare techniques through deep learning: issues, challenges and opportunities." *IEEE Access* 9 (2021): 98523-98541.

[5] Lee, Chonho, Zhaojing Luo, Kee Yuan Ngiam, Meihui Zhang, Kaiping Zheng, Gang Chen, Beng Chin Ooi, and Wei Luen James Yip. "Big healthcare data analytics: Challenges and applications." *Handbook of large-scale distributed computing in smart healthcare* (2017): 11-41.

[6] Lavanya, P., Subba, R.I.V., Selvakumar, V. & Shreesh V Deshpande. (2024). An Intelligent Health Surveillance System: Predictive Modeling of Cardiovascular Parameters through Machine Learning Algorithms Using LoRa Communication and Internet of Medical Things (IoMT). *Journal of Internet Services and Information Security*, *14*(1), 165-179.

[7] Venkatesh, R., C. Balasubramanian, and Madasamy Kaliappan. "Development of big data predictive analytics model for disease prediction using machine learning technique." *Journal of medical systems* 43, no. 8 (2019): 272.

[8] Malathi, K., Shruthi, S.N., Madhumitha, N., Sreelakshmi, S., Sathya, U., & Sangeetha, P.M. (2024). Medical Data Integration and Interoperability through Remote Monitoring of Healthcare Devices. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 15(2), 60-72. https://doi.org/10.58346/JOWUA.2024.I2.005

[9] Harerimana, Gaspard, Beakcheol Jang, Jong Wook Kim, and Hung Kook Park. "Health big data analytics: A technology survey." *Ieee Access* 6 (2018): 65661-65678.

[10] Juma, J., Mdodo, R.M., & Gichoya, D. (2023). Multiplier Design using Machine Learning Alogorithms for Energy Efficiency. Journal of VLSI Circuits and Systems, 5(1), 28-34.

[11] Karatas, Mumtaz, Levent Eriskin, Muhammet Deveci, Dragan Pamucar, and Harish Garg. "Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives." *Expert Systems with Applications* 200 (2022): 116912.

[12] Jelena, T., & Srđan, K. (2023). Smart Mining: Joint Model for Parametrization of Coal Excavation Process Based on Artificial Neural Networks. *Archives for Technical Sciences*, *2*(29), 11-22.

[13] Patil, Siddheshwar V., and Dinesh B. Kulkarni. "Parallel computing approaches for dimensionality reduction in the high-dimensional data." In *Third National Research Symposium on Computing*, p. 25. 2019.

[14] Alonso-Betanzos, Amparo, and Verónica Bolón-Canedo. "Big-data analysis, cluster analysis, and machine-learning approaches." *Sex-specific analysis of cardiovascular function* (2018): 607-626.

[15] Nazir, Shah, Muhammad Nawaz, Awais Adnan, Sara Shahzad, and Shahla Asadi. "Big data features, applications, and

analytics in cardiology—a systematic literature review." *IEEE Access* 7 (2019): 143742-143771.

[16] Yu, Ying, Min Li, Liangliang Liu, Yaohang Li, and Jianxin Wang. "Clinical big data and deep learning: Applications, challenges, and future outlooks." *Big Data Mining and Analytics* 2, no. 4 (2019): 288-305.

[17] Pansambal S.Y., et.al Real-Time Migration Risk Analysis Model for Improved Immigrant Development Using Psychological Factors, Migration Letters, V-20, I-4, PP:33-42, 2023.

[18] Mohapatra, Saumendra Kumar, and Mihir Narayan Mohanty. "Big data classification with iot-based application for e-health care." In *Cognitive big data intelligence with a metaheuristic approach*, pp. 147-172. Academic Press, 2022.

[19] Kutlu, Y., & Camgözlü, Y. (2021). Detection of coronavirus disease (COVID-19) from X-ray images using deep convolutional neural networks. *Natural and Engineering Sciences*, *6*(1), 60-74.

[20] Khan, Sulaiman, Habib Ullah Khan, and Shah Nazir. "Systematic analysis of healthcare big data analytics for efficient care and disease diagnosing." *Scientific Reports* 12, no. 1 (2022): 22377.