GMR
The Original

A SYSTEMATIC REVIEW OF MACHINE LEARNING ALGORITHMS FOR DISEASE OUTBREAK PREDICTION IN PUBLIC HEALTH

Shilpy Singh, Wamika Goyal, Dr. G. Subash Chandrabose, Dr. Praveen Priyaranjan Nayak, Dr.Shanmugapandian, Dr. Parag Amin, Dr. M N Nachappa

Assistant Professor, Department of Biotechnology and Microbiology, Noida International University, Greater Noida, Uttar Pradesh, India. shilpy.singh@niu.edu.in , 0000-0003-2274-9090
Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India. wamika.goyal.orp@chitkara.edu.in , https://orcid.org/0009-0004-8729-7464
Department of Community Medicine, Aarupadai Veedu Medical College and Hospital, Puducherry Vinayaka Mission Research Foundation(DU)India. subash.stat@gmail.com , Subash.gandhi@avmc.edu.in , https://orcid.org/0000-0002-5867-7255
Associate Professor, Department of Electronics and Communication Engineering, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India, Email Id- praveennayak@soa.ac.in , Orcid Id- 0000-0003-1726-1605
PROFESSOR, Department of Pharmacy, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India, Email Id- drshanmugapandiyan.pharmacy@sathyabama.ac.in, Orcid Id- https://orcid.org/0000-0002-1432-8221
Professor, ISME, ATLAS SkillTech University, Mumbai, India, Email Id- parag.amin@atlasuniversity.edu.in , Orcid Id- 0009-0005-0146-1815
Professor, Department of CS & IT, Jain (Deemed-to-be University), Bangalore, Karnataka, India, Email Id- mn.nachappa@jainuniversity.ac.in , Orcid id- 0000-0002-4007-5504

ABSTRACT

In this digital age, because of the speedily increasing advances in science and technology, vast amounts of health care data are produced from health care applications with varying technologies such as embedded systems, intelligent health devices, and computers. Machine learning algorithms are being considered effective technologies in the health care sector, which can be employed effectively in the early detection of disease by extracting significant patterns from the data. Over the past few years, the problem of chronic disease has become a worldwide challenge. It requires hours for a doctor to detect a chronic disease effectively. Machine learning algorithms can be integrated with feature selection algorithms to avoid these limitations. An effective integration of machine learning and feature selection-based models can assist doctors in predicting the risk of chronic diseases at an early stage. The chapter deals with the various problems which still haunt traditional methods and the motivational goals towards overcoming these problems by way of the suggested research work.

Keywords: *health, medicine, disease, machine learning.*

INTRODUCTION

Machine learning is a semi-automated process where useful information and knowledge are mined from data. This is carried out by algorithms, which the computer employs to examine the data in order to generate meaningful patterns and knowledge. To guarantee the success of the process, many wise decisions might have to be made. Machine learning encompasses artificial intelligence, which is a method that instructs computers to make predictions based on data [1]. This technique enables computers to learn without the need for intricate programming. Machine learning can be thought of as a learning system that can distinguish between spam and non-spam emails, putting the former in a spam folder and the latter in a non-

spam folder. It is used several times in many fields, such as medical field diagnosis to detect tumors, auto-driving cars, computational biology like DNA sequencing, analysis in the stock market, recognizing faces, detecting motion, aerospace, and predictive maintenance in manufacturing and drug detection. Machine learning is very useful when there is a shortage of skilled professionals for the detection of the phenomenon. It maintains accuracy of the predicted outcome [16]. Human accuracy may not be accurate like in reading the disease's images [17]. Usually, machine learning is a better option because of its advantages [2]. It is more accurate than a human's result and accuracy when it is data driven. Several times humans are not able to show what they know. Machine learning does not need a human expert. It's an automatic method to search for hypotheses explaining data. Machine learning is the process of extracting knowledge from data in a semi-automatic manner. Computers that apply algorithms to the data are used in machine learning to generate the required knowledge. To ensure the process's success, numerous wise choices may need to be made. Machine learning includes artificial intelligence, which is a technique that teaches computers to make predictions from data [1]. This technique enables computers to learn without the need for intricate programming. Machine learning can be thought of as a learning system that can distinguish between spam and non-spam emails, putting the former in a spam folder and the latter in a non-spam folder. For example, Disease prediction in supervised classification shows that a person is unhealthy or healthy [9][10]. Supervised learning can be accomplished in two steps. The first one is training the machine learning model with existing labeled data (data labeled with the outcome which has values "0" for a healthy or "1" is a non-healthy person) to teach the model with higher accuracy [21]. The second one is expecting the trained model to predict with high accuracy with the true outcome from new data without human interference[3] [19] .This is the purpose of supervised learning to build models that generalize. Using information from prior patients, such as age, height, weight, blood pressure, etc., supervised machine learning can also be used to forecast any disease within a certain time range. It can also predict frequent responses, like changes in temperature and fluctuations in power demand using regression techniques. Unsupervised learning is the learning task that shows how to represent data in the best way and it does not conclude the right or wrong answer. Applications for unsupervised learning like gene sequence analysis, object recognition, and market research. For instance, researching the buying habits of consumers on Amazon or at any mall may result in the formation of several groups, including women, children, students, and so on. In such cases, determining the number of clusters, allocating individuals to respective groups, and describing each cluster precisely is a major challenge. In addition, scarce resources and technical skills provide further hurdles to adoption, especially in low- and middle-income nations. To transcend these obstacles and realize the full potential of machine learning (ML) for public health surveillance, strategic investment in infrastructure, training, and legislative frameworks is necessary [11] [13-15]. If these challenges are addressed, ML-based systems can develop into mature and capable tools that complement conventional approaches, with ethical and equitable application [18].

Objectives

- This work discusses the disease outbreak prediction transformation capability of machine learning (ML) through utilising heterogeneous dynamic data sources. ML solutions bring a paradigm revolution in epidemic control, beyond the constraints of existing surveillance systems plagued by delayed reaction and poor integration of data. ML models improve forecast accuracy and timeliness by analyzing massive datasets, such as social media activity, climatic data, and electronic health records (EHRs). This allows for early intervention and resource management.

Research question

- Which machine learning algorithms are most suited for public health disease outbreak prediction, and how do they stack up in terms of precision, comprehensibility, and practicality?
- Which machine learning algorithms are most effective for disease outbreak prediction in public health?
- How do various machine learning models stack up in terms of epidemic forecasting accuracy, sensitivity, and specificity?

Methodology

When dealing with complicated issues including a lot of variables and sophisticated data, machine learning is utilized. This implies that the only way to deal with handwritten rules, face recognition, and speech recognition is through machine learning. Additionally, it manages tasks with ever-changing rules, such fraud detection from transaction records. It is also a solution for data types that are constantly changing and require program adaptation, such automated trading, forecasting energy use, and identifying purchasing trends. One of the most effective tools for controlling disease epidemics is predictive analytics. Predictive models can spot trends that indicate possible epidemics by examining both historical and current data, allowing for prompt actions [4]. For example, Google Flu Trends, though imperfect, demonstrated the potential of using search engine data to track influenza. Historically, disease surveillance relied on statistical models and epidemiological methods to predict and manage outbreaks. These decades-old systems constitute the core of public health initiatives. Traditional disease surveillance systems rely on statistical models, such as regression analysis and compartmental models (e.g., SIR models), to provide estimates of disease transmission dynamics and forecast trends of outbreaks [6].

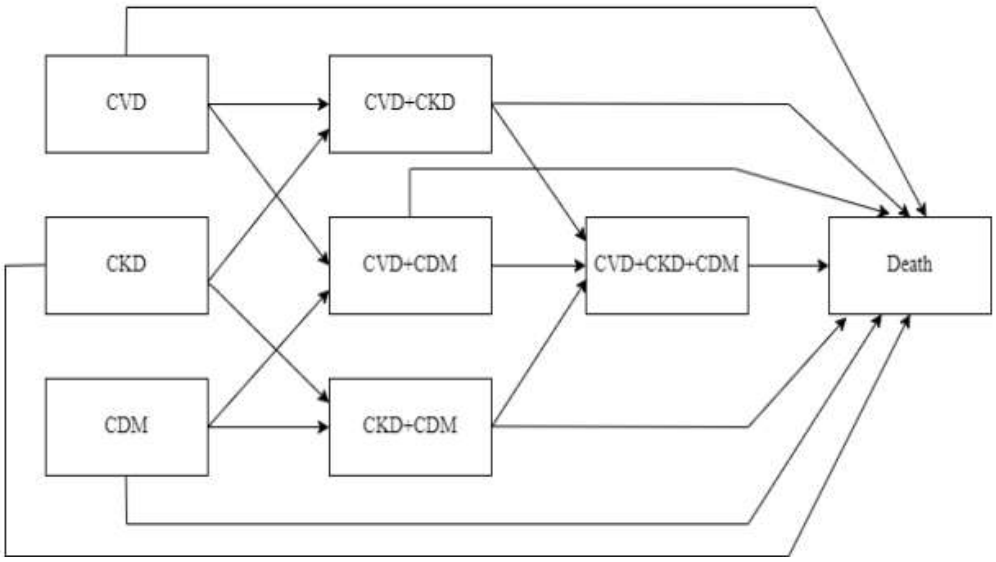

Figure 1: Progressive Relationship between Chronic Diseases

- Evaluation metrics are necessary for measuring a model's predictive accuracy and its efficacy in predicting outbreaks. Some of the regularly used metrics are:
- Precision: Estimates the proportion of correctly predicted positive cases out of all predicted positive cases. High precision shows fewer false positives.
- Sensitivity (Recall): Measures the ability of the model to differentiate true positives from all other positives. High recall suggests fewer false negatives.

Experimental analysis

Unreliable reporting, gaps in datasets, and source data biases can compromise predictive reliability. For instance, regional variations in electronic health record (EHR) documentation may introduce biased analysis to lower predictive model generalizability. Data integration is another issue because merge heterogeneous data sources like social media, EHRs, and climate data require sophisticated preprocessing tools to make them compatible. Even when sophisticated frameworks offer improved harmonization, complexities of data cleaning and harmonization continue to limit universal adoption. Privacy is also a complicating factor in ML use since how sensitive health information is treated by surveillance systems raises concerns about data ownership, consent, and abuse.

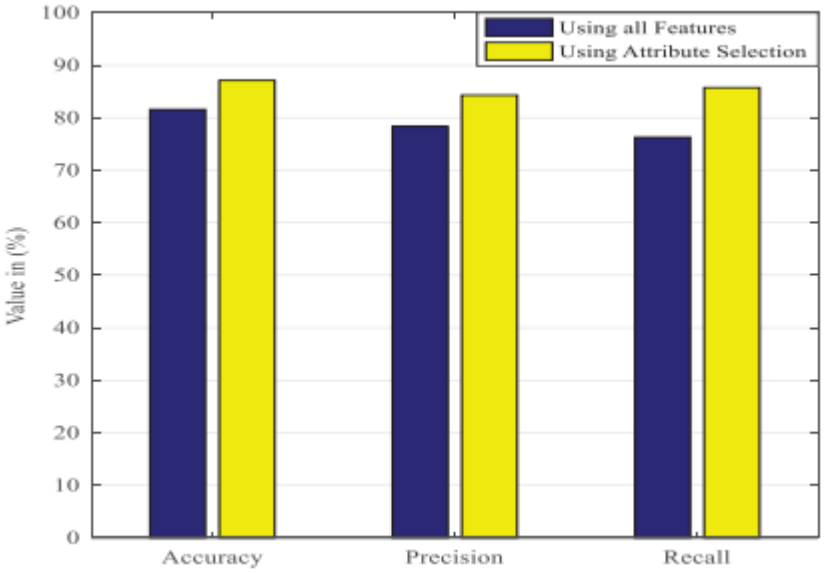

Figure 2: Simulation results of ML model

Balancing successful monitoring and the protection of privacy for individuals is essential, especially when utilizing social media and mobile tracking data. Ethical issues, though, transcend privacy. Machine learning (ML) model bias can entrench health inequities if poor and underserved groups are left out of training datasets. Second, the difficulty of interpreting algorithmic decision-making, known as the "black-box" problem, degrades public health practitioner trust and acceptance. Providing fairness, transparency, and accountability to ML models is necessary to alleviate these ethical issues and encourage public health surveillance use of ML in a broader scale.

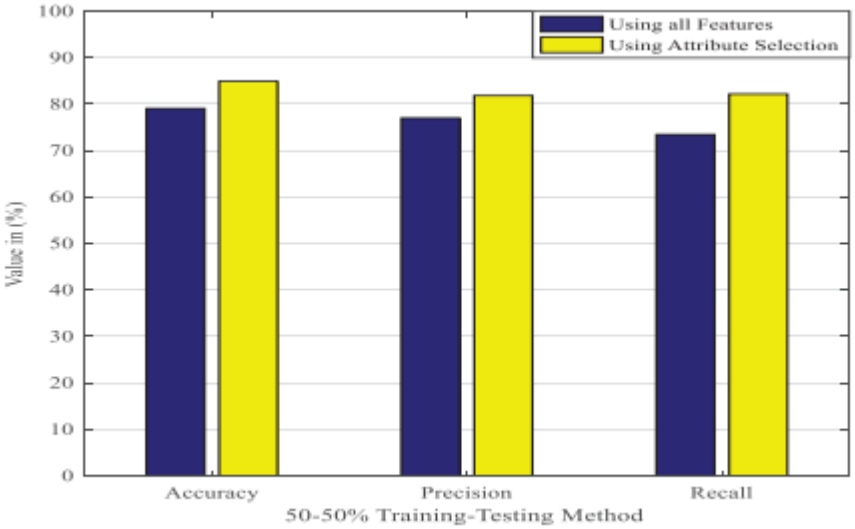

Figure 3: Simulation results of 50-50% training testing

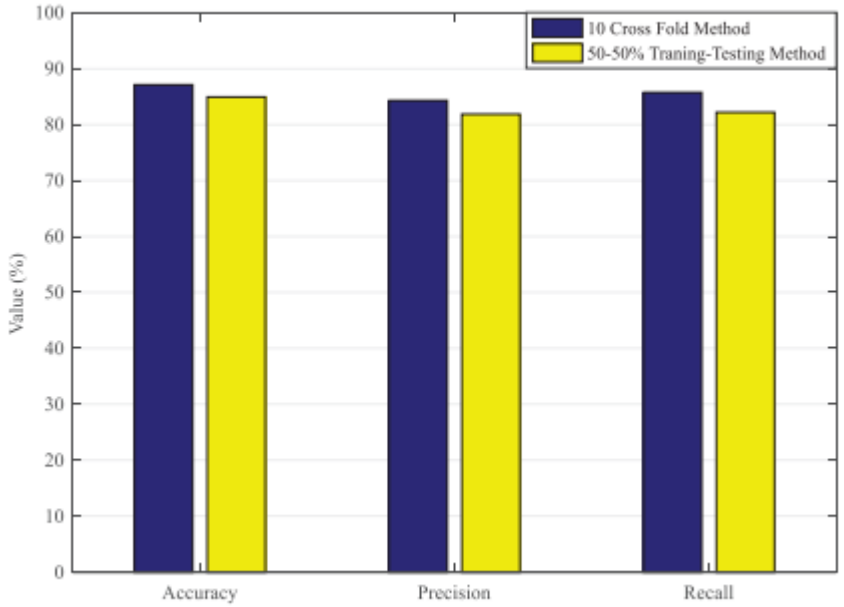

Figure 4 Comparison of simulation results of 50-50% training testing

Machine learning (ML) portends a paradigm shift in infectious disease outbreak forecasting and management. Through examination of enormous, varied datasets and exposing patterns beyond conventional methods, ML has become a critical part of contemporary public health surveillance, changing the way we predict, prevent, and react to infectious disease outbreaks.

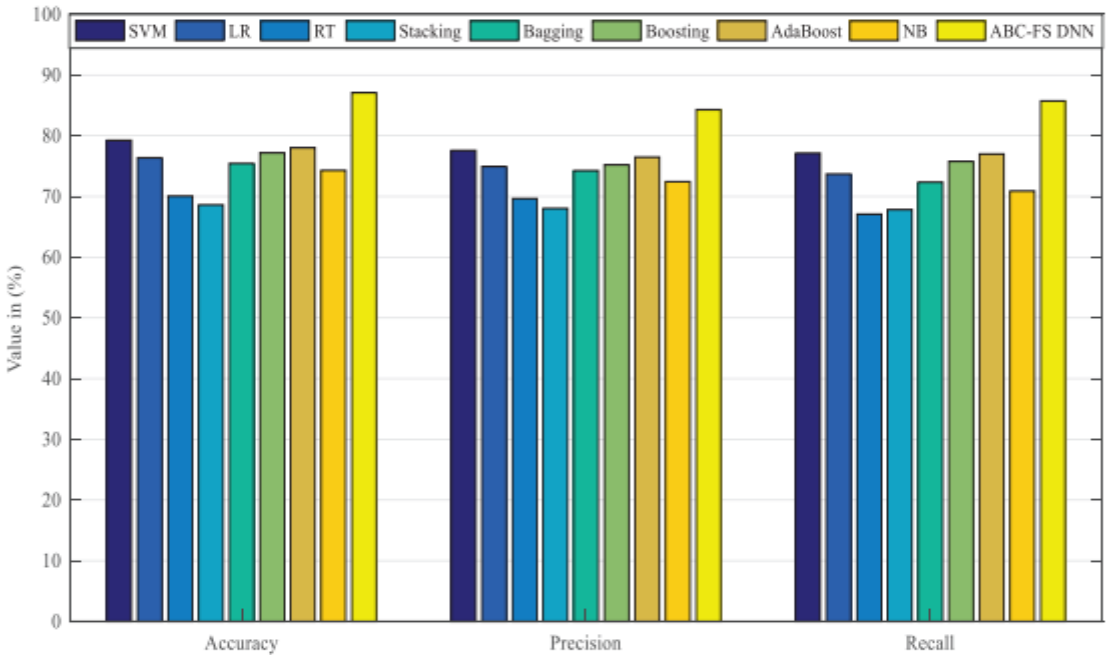

Figure 5 Simulation results various ML models

Machine learning (ML) makes it possible to identify illness hotspots early on and create responsive plans dynamically through the integration of real-time information sources, such as genomic sequence, social media, environmental considerations, and electronic health records (EHRs).

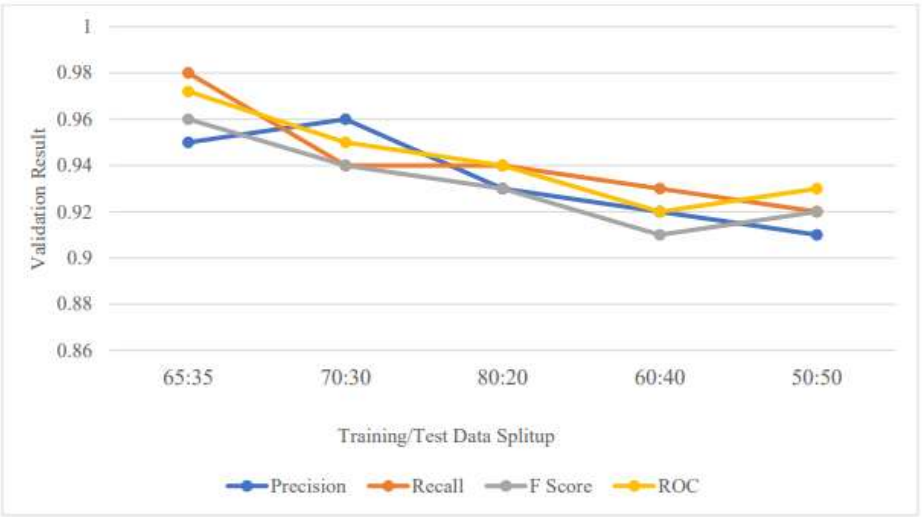

Figure 6: Validation result

This is timely and accurate information that informs health systems so they can invest resources optimally, prevent the spread of diseases, and even save lives. The strength of ML is not just in its capacity to process large volumes of data, but also in its potential to enable proactive, data-informed decision-making in public health.

Conclusion

Machine learning (ML) is superior not only in predicting future results but also in responding to new challenges. During the COVID-19 crisis, ML models were invaluable in tracking the spread of the epidemic, forecasting healthcare demand, and assessing the effectiveness of treatment. For example, models predicting hospitalizations and ICU utilization helped healthcare systems plan for peaks and reduce fatalities. In the same way, the integration of environmental and mobility data within ML models has been crucial for forecasting epidemics of diseases such as malaria, dengue, and cholera, where the dynamics of transmission are dominated by climate and human mobility. In addition to its immediate benefits, ML encourages an anticipatory public health approach. Surveillance has in the past been reactive, but the ability of ML to examine real-time data allows for the possibility of responding in advance. This is especially critical in pandemic prevention, where delay can lead to a sharp increase in cases. Through the use of predictive analytics, public health officials can move from response to prevention, a paradigm shift in global health security. This is an approach that allows for targeted interventions, reduces response times, and saves lives in the long term.

References

[1] Ajagbe, Sunday Adeola, and Matthew O. Adigun. “Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review.” Multimedia Tools and Applications 83, no. 2 (2024): 5893-5927.

[2] Keshavamurthy, Ravikiran, Samuel Dixon, Karl T. Pazdernik, and Lauren E. Charles. “Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches.” One Health 15 (2022): 100439.

[3] Rahman, Syed Ziaur, R. Senthil, Venkadeshan Ramalingam, and R. Gopal. “Predicting Infectious Disease Outbreaks with Machine Learning and Epidemiological Data.” Journal of Advanced Zoology 44, no. S4 (2023): 110-121.

[4] Ekundayo, Foluke. “Using machine learning to predict disease outbreaks and enhance public health surveillance.” (2024).

[5] Santangelo, Omar Enzo, Vito Gentile, Stefano Pizzo, Domiziana Giordano, and Fabrizio Cedrone. “Machine learning and prediction of infectious diseases: a systematic review.” Machine Learning and Knowledge Extraction 5, no. 1 (2023): 175-198.

[6] Mahesh, Batta. “Machine learning algorithms-a review.” International Journal of Science and Research (IJSR).[Internet] 9, no. 1 (2020): 381-386.

[7] Gupta, Aakansha, and Rahul Katarya. “Social media based surveillance systems for healthcare using machine learning: a systematic review.” Journal of biomedical informatics 108 (2020): 103500.

[8] Saleem, Farrukh, Abdullah Saad Al-Malaise Al-Ghamdi, Madini O. Alassafi, and Saad Abdulla AlGhamdi. “Machine learning, deep learning, and mathematical models to analyze forecasting and epidemiology of COVID-19: a systematic literature review.” International journal of environmental research and public health 19, no. 9 (2022): 5099.

[9] Papalou, A. (2023). Proposed Information System towards Computerized Technological Application – Recommendation for the Acquisition, Implementation, and Support of a Health Information System. International Journal of Communication and Computer Technologies, 8(2), 1-4.

[10] Veera Boopathy, E., Peer Mohamed Appa, M.A.Y., Pragadeswaran, S., Karthick Raja, D., Gowtham, M., Kishore, R., Vimalraj, P., & Vissnuvardhan, K. (2024). A Data Driven Approach through IOMT

based Patient Healthcare Monitoring System. Archives for Technical Sciences, 2(31), 9-15. https://doi.org/10.70102/afts.2024.1631.009
[11] Myoa, Z., Pyo, H., & Mon, M. (2023). Leveraging Real-World Evidence in Pharmacovigilance Reporting. Clinical Journal for Medicine, Health and Pharmacy, 1(1), 48-63.
[12] Mehta, V., & Reddy, P. (2024). Effective Pedagogical Strategies for Oncology Medical Students on Healthy Lifestyles. Global Journal of Medical Terminology Research and Informatics, 1(1), 9-15.
[13] Rao, A., & Menon, P. (2024). A Review of Membrane Filtrating Methods for Contaminant/Pollution Removal in Water and Sewage Treatment. Engineering Perspectives in Filtration and Separation, 1(1), 1-6.
[14] Rao, A., & Menon, P. (2024). A Review of Membrane Filtrating Methods for Contaminant/Pollution Removal in Water and Sewage Treatment. Engineering Perspectives in Filtration and Separation, 1(1), 1-6.
[15] Karimov, N., & Sattorova, Z. (2024). A Systematic Review and Bibliometric Analysis of Emerging Technologies for Sustainable Healthcare Management Policies. Global Perspectives in Management, 2(2), 31-40.
[16] Khyade, V. B., & Wanve, H. V. (2018). Statistics as Efficient TOOL of Analysis in the Biomedical Research. International Academic Journal of Science and Engineering, 5(1), 73–84. https://doi.org/10.9756/IAJSE/V5I1/1810007
[17] Jagadeeswaran, Logeswaran, Prasath, S., Thiyagarajan, & Nagarajan. (2022). Machine Learning Model to Detect the Liver Disease. International Academic Journal of Innovative Research, 9(1), 06–12. https://doi.org/10.9756/IAJIR/V9I1/IAJIR0902
[18] Arun Kumar, E., Franklin Pratap, A., & Rathika, S. K. B. (2018). An Efficient Hostel Student Laundry Service. International Journal of Advances in Engineering and Emerging Technology, 9(2), 49–53.
[19] Papalou, A. (2023). Proposed Information System towards Computerized Technological Application – Recommendation for the Acquisition, Implementation, and Support of a Health Information System. International Journal of Communication and Computer Technologies, 8(2), 1-4.
[20] Subermaniam, L., Kuppusamy, M., Isakulova, N., Ghate, A. D., Subrahmanyam, S., & John, B. (2025). A SYSTEMATIC REVIEW OF SUSTAINABILITY MANAGEMENT IN SMALL AND MEDIUM ENTERPRISES TO IMPROVE SUSTAINABLE DEVELOPMENT GOALS. ACTA INNOVATIONS, 38–48. https://doi.org/10.62441/actainnovations.vi.403
[21] Abbood, R. S., & Luaibi, N. M. (2023). Subchronic intraperitoneal toxicity of Sio2NPs on body weight and thyroid gland hormones in female Rats. Bionatura, 8(1), Article 59. https://doi.org/10.21931/RB/CSS/2023.08.01.59
[22] M. Kavitha. (2024). Restoring Wetland Ecosystems Using Native Macrophytes for Improved Water Quality and Aquatic Biodiversity. Journal of Aquatic Ecology and Environmental Sustainability, 1(1), 1–8.
[23] Karpagam, M., Geetha, K., & Rajan, C. (2020). RETRACTED ARTICLE: A modified shuffled frog leaping algorithm for scientific workflow scheduling using clustering techniques: M. Karpagam et al. Soft Computing, 24(1), 637-646..