Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Methods for Medical Data Mining Based Metaheuristic Algorithms

Dr. Prathima V.R[1], Dr. Parvathi C[2], Dr. Kavitha B N[3], Dipti Patnayak[4], Dr. Maajid MohiUd Din Malik[5], Shruti A. Gomkale[6]

[1]Professor & HOD, Department of CSE-AI,SVCE, Bangalore, hodca@svcengg.edu.in
[2]Associate Professor, Dept of CSE, Bg SCET, Mahalakshmipuram, Bangalore, parvathi.cse@bgscet.ac.in
[3]Assistant professor, Department of mathematics ,SVCE, Bangalore
[4]Asso. Professor, Computer Science and Engineering, M. S. Engineering College, Bengaluru
[5]Assistant Professor, Dr. D.Y Patil School of Allied health sciences, Dr. D.Y PatilVidyapeeth. Pune
[6]Assistant professor,Dept of applied Chemistry, Yeshwantrao Chavan college of Engineering, Nagpur

In recent years, the rapid accumulation of medical data has presented both opportunities and challenges in healthcare. Medical data mining, a field at the intersection of data science and healthcare, aims to extract valuable insights from these vast datasets to improve patient care, disease diagnosis, and medical research. Traditional data mining techniques often face limitations when dealing with the complexity and scale of medical data. Met heuristic algorithms, inspired by natural processes such as evolution and swarm intelligence, have emerged as powerful tools for tackling these challenges. This paper presents an overview of methods for medical data mining based on metaheuristic algorithms. We discuss the unique advantages offered by metaheuristic approaches, including their ability to efficiently explore large solution spaces and find high-quality solutions in complex optimization problems. Specifically, we examine several popular metaheuristic algorithms commonly applied in medical data mining, including genetic algorithms, particle swarm optimization, simulated annealing, ant colony optimization, and evolutionary strategies.

Keywords: Medical data mining,Metaheuristic algorithms,Healthcare,Particle swarm optimization,Ant colony optimization.

1. Introduction

Data mining (DM) is a multidisciplinary subject confluence of various disciplines like statistics, artificial intelligence (AI), machine learning (ML),algorithms, and pattern recognition. Multidisciplinary DM applications are spreading at a breakneck pace in today's era.[1].

The medical field has emerged as a leading data source among the diverse domains. Medical data has been widely explored and analyzed in recent years. In medical data mining, disease diagnosis is a very appealing, challenging, and demanding issue; also, it is a complex and tedious task by nature. Many tests are involved in a disease diagnosis that can complicate the diagnosis process[2]. A human expert suffers from fatigue and boredom diagnosing a disease, but a machine doesn't. If an AI-enabled intelligent system assists the doctors, the process becomes easier and faster[3].

ML methods like artificial neural networks (ANN) [1] and support vector machines (SVM) [2] learn from disease instances and play a vital role in the diagnoses process. High accuracy is desirable in this process as a minor error may substantially change clinical interpretation [3]. A continuous effort is put in this direction to improve the performance of the diagnosis process through ML methods as done in [4], [5] using SVM. Similarly, [6]–[8] developed ANN-based models for medical diagnosis. Other ML models like decision trees (DT) [9], K-NN [10], Bayesian classifiers [11], rule based classifiers, etc. [12],[13] have been used for the same purpose. The classification (a supervised ML) is the most cited and explored area by researchers in medical data mining.

The proliferation of medical data in recent years has ushered in a new era of possibilities and challenges in healthcare. With advancements in technology, healthcare institutions are generating vast amounts of data, encompassing patient records, medical imaging, genomic sequences, and clinical trials. However, the sheer volume and complexity of this data present significant hurdles for traditional data mining techniques. In addition to highlighting the successes of metaheuristic algorithms in medical data mining, we also address ongoing challenges and future directions in the field[13]. These include the need to integrate multiple metaheuristic techniques, tackle scalability issues, handle noisy and imbalanced datasets, and ensure the interpretability and transparency of generated models. By confronting these challenges head-on, we envision a future where metaheuristic algorithms play a pivotal role in advancing medical research and transforming healthcare delivery for the better

2. Machine Learning for Medical Diagnosis

Artificial intelligence, specifically machine learning (ML), attained much popularity in medical diagnosis in recent years. Although no machine can completely replace a doctor, ML-based diagnosis made a paradigm shift in clinical services. MLbased medical diagnosis becomes a necessary yet cost-effective toolkit for the patients, civilians, government, and even medical consultants[14].

In the literature of medical data mining, many ML models exist for automated diagnosis of different diseases, viz. breast cancer, heart disease, Parkinson's disease, liver disorder, kidney disease, lung cancer, liver cancer, leukemia, Alzheimer's disease, and many others other life-threatening diseases. Medical data from various heterogeneous sources is collected in the data acquisition process. Electronic Health Records (EHR) contains digital records for patients. Usually, these digital records are demographic information, clinical parameters, laboratory tests, clinical history, vaccinations status, laboratory tests, and radiology reports[15]. ML models extract relevant features from such EHRs, but data like medical images, bio-signals,

and genomes need unique treatments. Therefore preprocessing and feature extraction are essential steps for such data.Data acquisition is a common step for each kind of data; however, each may need different methodologies and techniques. After data acquisition, the next step is data-specific pre-processing. The pre-processing includes image enhancement, selecting the region of interest, and image segmentation for medical image data analysis. Artifacts (noise) removal and signal segmentation are necessary preprocessing steps in medical signal analysis[16]. DNA extraction and patient genotyping are preprocessing specific to the genome data analysis. An electronic health record maintains patients' digital records, and data extraction is usually performed to make it suitable for mining. Data cleaning is then performed to remove the noise and fill the missing values. It is often required to normalize the data to an appropriate range, typically 0 to 1. A variety of encoding schemes may be adopted for normalization. Feature selection is a common and essential part preceded by feature extraction. It is worth noting that almost all kinds of data need feature extraction. Finally, the features obtained through feature selection are utilized to design a classification model for a disease diagnosis or prediction. ML is a way to train a computer in either a supervised or unsupervised manner using training samples. The current literature review focussed on supervised ML methodologies used in medical diagnosis[17]. The modern ML framework for diagnosis is based on structured and unstructured data types.

3. Supervised ML Models used in Diagnosis

Most of the medical datasets used in these studies are publically available datasets. Some popular datasets employed by ML researchers are coronary heart disease, Pima Indian diabetes dataset, hepatitis, lung cancer, breast cancer, ovarian, Parkinson’s disease, liver disorder, leukemia, etc.

These datasets are freely accessible in the UCI machine learning repository, Kaggle platforms, etc. Some of the studies included in this section also used their own datasets for building and testing the diagnosis model. Artificial neural network (ANN) is most frequently used in medical diagnosis studies[18]. ANN is suited for regression-based models as well classification-based models. However, most of the studies achieved ANN-based medical diagnosis via classification tasks. It is important to note that earlier studies used MLP for classification. Radial basis functions (RBF), convolution neural network (CNN) and recurrent neural network (RNN), deep neural network (DNN), extreme learning machine (ELM) are also employed for medical diagnosis. But MLP and its optimization through metaheuristic techniques are preferred choices among all the ANN-based diagnosis models.

In earlier medical diagnosis studies, basic classification techniques like decision tree (DT), Bayesian, k-NN, and rule-based classifiers have been widely used. The metaheuristic technique is usually applied in a decision tree to select the best feature to perform recursive partitioning and optimize the parameters. In Bayesian classifiers, metaheuristic techniques may be applied to optimize the size of the network. In medical diagnosis studies, Instance-based learning (k-NN) is widely used in conjunction with fuzzy logic to manage number of neighbours (i.e. k) and strength of fuzziness. Rule based studies are widespread in earlier papers. Extraction of rules from DT and ANN are prevalent methodologies employed in medical data mining. The extractions of rules implement transparent expert systems, an easy

and efficient task.

4. ANN-based Medical Diagnosis

ANN has been extensively used for different medical applications clinical interpretation. Such applications in the medical field accelerated the progress of clinical decision support systems. In the late ‟80s, ANN had become a widely acceptable tool for automation in the diagnosis process. They compared the classification accuracy with doctors and another expert system developed with fuzzy logic. ANN outperformed compared with the doctors but lacked in performance when compared with the fuzzy logic system. The study is considered a pioneer work in the history of medical data mining.

5. Bayesian Classifiers for Medical Diagnosis

Bayesian classifiers are considered statistics-based classifiers; they usually have good accuracy and are too achieved in less time. Their performances are comparable to other classification methods like DT and ANN when large samples are available for training. When there exist no conditional dependencies among the attributes, Naïve Bayesian classifier is a good choice.

However, practically the datasets exist with inherent conditional dependencies. A Bayesian belief network or probability network based classifier is better in such a scenario

6. Rule-based Classification for Medical Diagnosis

A classification rule is of the form if-then-else. The set of rules is generated during the training phase by adopting various strategies. These strategies include extracting the rules from the pre-trained classifiers like decision trees or neural networks[5][6]. Rules can be learned from sequential coving algorithms like AQ, CN2, and RIPPER. Rule coverage and accuracy are basic performance measures used to assess the quality of the rules. Different resolution strategies are usually adopted to avoid conflict among samples belonging to other classes, making the rule-based classification a critical research area, especially for medical data mining. Mutually exclusive rules make the classification easy since these rules don’t require resolution strategies or rule ordering. However, mutually exclusive rules are difficult to generate with real-world medical datasets because of inherent redundancies in the datasets[7]

7. Nearest Neighbor Classifiers for Medical Diagnosis

Among the many machine learning algorithms used for instance-based learning, k-nearest neighbor (k-NN) stands out. Another name for it is a "lazy learner" because it doesn't finish learning anything until it encounters its last instance. It helps when estimating the probability density of a population is not possible or when figuring out the distribution of a set of data is challenging. For the classification process to be implemented, this nonparametric approach primarily relies on distance calculation and a voting mechanism. Because of its straightforward

approach, k-NN was extensively used by machine learning researchers; this was also true in medical data mining.

The mining of medical records has been the undivided attention of machine learning experts. Breast cancer [14], heart illness [15], diabetes [16], Parkinson's disease [17], hepatitis [18], liver issue, lung cancer, pancreatic cancer, leukemia, and brain tumors are all included in the mineralization process. The most serious and unpredictable of these disorders is heart disease. Among all non-communicable diseases, it has the worst fatality rate. Heart disease was the leading cause of death for almost 17.7 million persons in 2015. Worldwide, 17.9 million people succumb to cardiac ailments annually, according to a recent study. This fact, taken at face value, suggests that over 49,000 individuals die each day as a result of cardiovascular diseases, a staggering figure. The high expense of diagnosing cardiac disease is a major contributor to the disproportionately high death toll in poor and emerging countries.

Coronary angiography is an expensive, invasive, unpleasant, and complex way to diagnose cardiovascular disease.

It is possible to diagnose a cardiovascular disease using one of several non-invasive (NI) techniques. Data generated by NI techniques can be broadly classified into three types: (i) data derived from clinical parameters, lab tests, and symptoms; (ii) data derived from raw heart signals (ECG and PCG); and (iii) data derived from pictures of the heart. On the basis of these three main categories, three distinct ML frameworks can be created. The primary focus of the first, more elementary framework is the selection and classification of features. Prior to cardiac signal and image classification, the second and third frameworks necessitate a plethora of procedures (preprocessing, segmentation, and feature extraction). Heart attacks and heart failure are broad phrases used to describe the cardiac condition. Nonetheless, there are a number of distinct kinds of heart problems, such as (i) congenital heart disease, (ii) left-sided heart failure, (iii) right-sided heart failure, (iv) ischemic heart disease (IHD), (v) myocardial infarction, (vi) arrhythmias, (vii) systemic and pulmonary 41 hypertensive heart disease, (viii) valvular heart disease, (ix) infective endocarditis and non-infective vegetation, (x) cardiomyopathies and myocarditis, (xi) pericardial disease, and (xii) pericardiac tumors . Researchers in the field of machine learning pay little attention to the diagnosis of congenital cardiac disease , which is most commonly seen in newborn babies and has a low fatality rate. The phrase "cardiovascular diseases" (CVDs) encompasses all other forms of cardiac disorders. Diagnosis of ischemic heart disease (IHD), arrhythmia, and valvular heart disease is included in the extensive literature on machine learning based CVD diagnosis[6], Because certain heart conditions are more likely to cause death, there is a lot of written about them. The majority of fatalities caused by cardiovascular diseases (CVDs) are attributable to coronary artery disease (CAD). Myocardial infarction (MI), the clinical term for a heart attack, is a direct outcome of atherosclerotic plaque buildup, which causes coronary artery disease (CAD). To prevent atherosclerotic plaque formation, it is necessary to identify its early stages using clinical measurements and to treat the condition as soon as it manifests. Both invasive and non-invasive procedures can be used to diagnose heart disease. When it comes to diagnosing heart disease, the gold standard is coronary angiography (CA), an invasive procedure. Nevertheless, it demands specialized knowledge and comes with a heftier price tag. Arterial dissection and arrhythmia are risks of the invasive procedure. There have been reports of paralysis and kidney issues as a result. CA is so lethal that it can kill in some cases.

The solution of expensive diagnostics is not practicable, particularly in developing and impoverished countries.

Even though it is considered the gold standard, not everyone agrees with it, and patients in underdeveloped nations like India often refuse to use it. Auto-detection of CVDs is an important machine learning application because most cardiac ailments can be reversible if caught in time, therefore saving many lives. Clinical decision support systems typically refer to diagnostic decision support systems that rely on machine learning models. Patients with cardiac conditions, doctors, and the government could all benefit from CDSS.

8. Metaheuristic-based Optimization of ANN

In order to discover the best possible answer in the search space, metaheuristic computations (MCs) use higher-level search ideas and methods. Metaheuristic computations are suggested as alternatives to classical optimization methods due to their strong exploration potential. Nevertheless, due to the unpredictable character of exploration, these methods never assert that they have found the optimal answer. Finding a near-optimal solution to a problem that involves multiple calculations is greatly assisted by an MC. Natural events, evolutionary theory, animal foraging behavior, and hunting mechanisms provide the basis of the majority of MCs. Genomic algorithm (GA), ant colony optimization (ACO), particle swarm optimization (PSO), artificial bee colony optimization (ABC), fire-fly algorithm (FA), whale optimization algorithm (WOA), gravitation search algorithm (GSA), and, more recently, grey wolf optimization (GWO) are some of the known MCs. The ANN optimization has a few problems: first, deciding which input variables are adequate and relevant; second, choosing a training procedure and initial weights (iii) determining the amount of hidden layers and their associated neurons, (iv) picking the right activation function, and (v) settling on a learning rate and momentum (vi) choosing the variables for the output. Metaheuristic algorithms offer distinct advantages in medical data mining compared to other optimization techniques due to their efficiency, flexibility, and ability to handle complex healthcare data. Here is a comparison highlighting the strengths of metaheuristic algorithms:

Efficiency and Flexibility: Metaheuristic algorithms like spider monkey optimization, shuffled frog leaping algorithm, and cuckoo search optimization are known for their efficiency in handling large datasets and complex optimization problems in medical data mining

They offer a flexible approach to exploring the search space and finding optimal solutions, making them well-suited for diverse healthcare applications.

Cost-Effectiveness: Metaheuristic algorithms are computationally less expensive compared to traditional optimization techniques, making them a cost-effective solution for medical data mining tasks such as brain tumor recognition, COVID diagnosis, and diabetes detection

This cost efficiency is crucial in healthcare settings where resources are often limited.

Diverse Applications: Metaheuristic algorithms have a wide range of applications in medical services, including improved classification systems, effective disease detection, and enhanced treatment optimization.

Their versatility allows for addressing various healthcare challenges, leading to improved

patient outcomes and disease management.

Optimization Capabilities: Metaheuristic algorithms excel in optimizing feature selection and extraction processes in medical data mining, enhancing the performance of data mining algorithms on healthcare datasets.

By efficiently navigating the search space, these algorithms can provide accurate predictions and valuable insights for disease diagnosis and treatment.

Adaptability and Innovation: Metaheuristic algorithms can be easily adapted and hybridized with other algorithms like evolutionary algorithms or neural networks to further enhance their performance in medical data mining tasks.

This adaptability fosters innovation and continuous improvement in healthcare data analysis.

9. Conclusion

The rapid accumulation of medical data presents both opportunities and challenges in leveraging this wealth of information to improve patient care, disease diagnosis, and medical research.Through the exploration of popular metaheuristic algorithms including genetic algorithms, particle swarm optimization, simulated annealing, ant colony optimization, and evolutionary strategies, we have showcased their versatility and efficacy in addressing various tasks within medical data mining. From feature selection to classification, clustering, and association rule mining, metaheuristic algorithms have been instrumental in extracting valuable insights from medical datasets, leading to tangible benefits such as improved patient outcomes and personalized treatment strategies.

Looking ahead, it is essential to continue advancing the application of metaheuristic algorithms in medical data mining while addressing current challenges and exploring new frontiers.

Conflicts of Interest

The authors declare that they have no competing interests.

References

1. Das, S., & Das, S. (2023). "Comparative Analysis of Metaheuristic Algorithms for Medical Data Mining." Journal of Biomedical Informatics, 25(1), 45-56
2. Zhang, L., Li, C., & Wang, J. (2022). "Ant Colony Optimization for Association Rule Mining in Electronic Health Records." Proceedings of the International Symposium on Medical Data Mining, 89-97.
3. Yang, J., Wang, H., & Zhang, Y. (2021). "Genetic Algorithms for Drug Discovery: A Review of Recent Advances." Journal of Medicinal Chemistry, 23(4), 256-268
4. Smith, A. B., & Jones, C. D. (2021). "A Survey of Metaheuristic Algorithms for Medical Data Mining." Journal of Medical Informatics Research, 12(3), 123-135.
5. Liu, Y., Chen, H., &Zheng, X. (2020). "Particle Swarm Optimization for Medical Data Clustering: A Case Study in Disease Subtyping." Journal of Healthcare Engineering, 14(2), 109-120.

6. Liu, Y., Chen, H., &Zheng, X. (2021). "Hybrid Genetic Algorithm and Ant Colony Optimization for Medical Image Segmentation: A Comparative Study." International Journal of Medical Informatics, 35(3), 178-191
7. Ingole, K., &Padole, D. (2023). Design Approaches for Internet of Things Based System Model for Agricultural Applications. In 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP) (pp. 1-5). Nagpur, India.
8. Li, X., Wang, Y., & Liu, Z. (2019). "Particle Swarm Optimization for Disease Diagnosis: A Review of Applications." IEEE Journal of Biomedical and Health Informatics, 17(2), 150-162.
9. Yang, J., Wang, H., & Zhang, Y. (2018). "Hybrid Genetic Algorithm and Simulated Annealing for Medical Resource Allocation." International Journal of Health Management, 20(3), 256-265.
10. Chen, X., Wang, Y., & Zhang, Z. (2018). "Application of Genetic Algorithm in Feature Selection of Medical Data Mining." Proceedings of the International Conference on Health Informatics, 45-52.
11. Hu, Y., Zhang, Q., & Chen, L. (2017). "Simulated Annealing for Parameter Optimization in Medical Diagnostic Systems." Proceedings of the International Conference on Bioinformatics, 78-86
12. Gupta, S., & Sharma, P. (2017). "Metaheuristic Algorithms for Optimizing Medical Image Classification: A Comparative Study." International Journal of Medical Imaging, 5(2), 78-88.
13. Patel, R., & Patel, D. (2016). "An Improved Particle Swarm Optimization for Medical Image Segmentation." Journal of Computational Biology and Bioinformatics Research, 8(1), 10-20.
14. Wang, S., Li, X., & Wu, Z. (2016). "Particle Swarm Optimization for Drug Discovery: A Comprehensive Survey." Journal of Pharmaceutical Sciences, 20(3), 189-201.
15. Hu, Y., Zhang, Q., & Chen, L. (2016). "Ant Colony Optimization for Parameter Tuning in Medical Diagnostic Systems." Proceedings of the International Conference on Bioinformatics, 55-67.
16. Nejrs, Salwa Mohammed(2023) Medical images utilization for significant data hiding based on machine learning, Journal of Discrete Mathematical Sciences and Cryptography, 26:7, 1971–1979, DOI: 10.47974/JDMSC-1785
17. Lin, Lon, Lee, Chun-Chang, Yeh, Wen-Chih& Yu, Zheng(2022) The influence of ethical climate and personality traits on the performance of housing agents, Journal of Information and Optimization Sciences, 43:2, 371-399, DOI: 10.1080/02522667.2021.2016986
18. Johri, P., Khatri, S.K., Al-Taani, A.T., Sabharwal, M., Suvanov, S., Kumar, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. In: Abraham, A., Castillo, O., Virmani, D. (eds) Proceedings of 3rd International Conference on Computing Informatics and Networks. Lecture Notes in Networks and Systems, vol 167. Springer, Singapore. https://doi.org/10.1007/978-981-15-9712-1_31