11-24-2023

Intelligent biomedical image classification in a big data architecture using metaheuristic optimization and gradient approximation

Laila Almutairi
Majmaah University

Ahed Abugabah
Zayed University, ahed.abugabah@zu.ac.ae

Hesham Alhumyani
Taif University

Ahmed A. Mohamed
Majmaah University; Assiut University

Wireless Networks
https://doi.org/10.1007/s11276-023-03573-5

Intelligent biomedical image classification in a big data architecture using metaheuristic optimization and gradient approximation

Laila Almutairi[1] · Ahed Abugabah[2] · Hesham Alhumyani[3] · Ahmed A. Mohamed[1,4]

Accepted: 10 October 2023

Abstract
Medical imaging has experienced significant development in contemporary medicine and can now record a variety of biomedical pictures from patients to test and analyze the illness and its severity. Computer vision and artificial intelligence may outperform human diagnostic ability and uncover hidden information in biomedical images. In healthcare applications, fast prediction and reliability are of the utmost importance parameters to assure the timely detection of disease. The existing systems have poor classification accuracy, and higher computation time and the system complexity is higher. Low-quality images might impact the processing method, leading to subpar results. Furthermore, extensive preprocessing techniques are necessary for achieving accurate outcomes. Image contrast is one of the most essential visual parameters. Insufficient contrast may present many challenges for computer vision techniques. Traditional contrast adjustment techniques may not be adequate for many applications. Occasionally, these technologies create photos that lack crucial information. The primary contribution of this work is designing a Big Data Architecture (BDA) to improve the dependability of medical systems by producing real-time warnings and making precise forecasts about patient health conditions. A BDA-based Bio-Medical Image Classification (BDA-BMIC) system is designed to detect the illness of patients using Metaheuristic Optimization (Genetic Algorithm) and Gradient Approximation to improve the biomedical image classification process. Extensive tests are conducted on publicly accessible datasets to demonstrate that the suggested retrieval and categorization methods are superior to the current methods. The suggested BDA-BMIC system has average detection accuracy of 94.6% and a sensitivity of 97.3% in the simulation analysis.

Keywords Biomedical image classification · Metaheuristic optimization · Gradient approximation · Big data architecture

✉ Laila Almutairi
L.Almutairi@mu.edu.sa

Ahed Abugabah
ahed.abugabah@zu.ac.ae

Hesham Alhumyani
h.alhumyani@tu.edu.sa

Ahmed A. Mohamed
amohamed@mu.edu.sa

[1] Department of Computer Engineering, College of Computer and Information Sciences, Majmaah University, Majmaah 11952, Saudi Arabia

[2] College of Technological Innovation, Zayed University, Abu Dhabi Campus, Abu Dhabi, UAE

[3] Department of Computer Engineering, College of Computers and Information Technology, Taif Uni-versity, Taif, P.O. Box 11099, 21944, Saudi Arabia

[4] Department of Information Technology, Faculty of Computer and Information, Assiut University, Assiut 71515, Egypt

Published online: 24 November 2023

1 Introduction to biomedical image classification

Biomedical image processing is crucial to computer-assisted diagnosis [1]. The modern medical care sector relies heavily on computer-aided diagnostic technologies. The field of biomedical image processing is crucial to the advancement of computer-assisted diagnostics in healthcare. Enhancing images in this way helps doctors see more detail and make more informed diagnoses.Early illness diagnosis is aided by biomedical image processing algorithms' ability to detect and pinpoint lesions or anomalies.

It is able to extract quantitative elements from images, giving clinicians more data with which to make objective diagnoses. Aiding radiologists in their quest for greater diagnostic precision and efficiency, computer-aided diagnosis (CAD) systems are made possible by biomedical image processing. Surgical procedures and interventional guiding are aided by processed photos that provide visualizations, measurements, and 3D reconstructions.

Automatic systems are efficient enough to uncover a great deal of concealed information and handle vast quantities of data in an acceptable period. The use of automated systems for biological image processing has a number of benefits, including the following:

Automatic systems can analyses huge numbers of biological pictures swiftly and consistently, allowing the analysis of vast datasets that would be too time-consuming or labor-intensive to analyse manually. Automatic systems are consistent and free of the subjective biases that can creep into manual analysis. They offer standardized, objective metrics that guarantee repeatability and credibility in biological image analysis.

High-throughput Analysis: High-throughput analysis is made possible by automatic systems, which allow for the rapid screening and analysis of a huge number of biological pictures. This capacity is especially helpful in large-scale investigations, such as population surveys.

Humans may not be able to immediately recognize hidden patterns or relationships within biological picture data, but automatic systems can. Using data mining and machine learning algorithms, it may be possible to find previously undetected small traits or relationships that will aid in medical diagnosis and research.

Automatic solutions are easily scalable to accommodate expanding datasets and can interface with existing data processing tools or systems. Scalability and integration allow for in-depth analysis, data fusion, and the investigation of intricate connections between different types of clinical and omics data and biological pictures.

Automatic systems utilize the power of automation to improve biomedical image processing in a number of ways, including efficiency, objectivity, scalability, and discovery potential, which in turn opens up new avenues for increasing medical research, diagnosis, and patient care.

Conventional human-based diagnostic techniques are time-consuming and prone to a variety of mistakes. Therefore, biological image analysis is included in automated healthcare care platforms. Occasionally, computer vision may outperform human eyesight [2]. Computer vision has the potential to outperform human vision in a number of ways when used to the analysis of biomedical images: When compared to human observers, computer vision algorithms are more accurate and precise in spotting and interpreting tiny details and anomalies in biomedical images. Consistency and Reproducibility: Unlike traditional diagnostic methods, computer vision systems provide both consistency and reproducibility in their findings. Processing Speed: Computer vision algorithms can analyses huge numbers of biomedical pictures quickly, allowing for timely analysis and diagnosis.

Computer vision offers quantitative analysis, which can help with decision making and illness monitoring by giving objective measurements, exact quantifications, and numerical data. Computer vision algorithms may learn from large amounts of labelled biomedical image data, allowing for pattern detection and the identification of complicated correlations and patterns that may not be obvious to the human eye. Compared to human vision, computer vision is more accurate, reliable, fast, objective, and able to extract useful information from biological images because it makes use of cutting-edge computational algorithms and machine learning. This has the potential to improve patient care by allowing for more precise diagnosis and more deliberate treatment planning.

Nonetheless, for any image processing technique, picture quality is of paramount importance. If the picture quality is inadequate, the result of the image assessment method cannot achieve the required precision. Occasionally, the processing techniques require an excessive amount of time, rendering them inappropriate for real-world applications [3]. Time constraints are a common barrier for biomedical image analysis processing techniques, limiting their usefulness in clinical settings. Because of the time and effort needed to execute their sophisticated algorithms and computations, many processing techniques are prohibitively inefficient.

High-resolution pictures and 3D scans are only two examples of the types of biomedical datasets that can be quite large, requiring extensive data loading, preprocessing, and analysis, which can add significant time to the overall processing time. Imperative for Real-Time Processing There are times in medicine, such as during surgery or emergency conditions, when quick decisions must be made, necessitating real-time or near real-time analysis.

Time-sensitive situations may be less amenable to the use of these methods if they require too much processing time. Extensive processing times might interrupt the workflow or cause delays in patient care if the analysis results are not accessible in a timely manner, which is why integrating processing processes into existing clinical workflows can be hard. Scalability and efficiency: Acceptable processing speeds are essential for practical application when scaling up processing techniques to deal with larger datasets or incorporating them into high-throughput environments.

The medical application of processing techniques can be greatly improved by addressing these time-related challenges. To address these obstacles and pave the way for these methods to be used in practical medical settings, researchers are investigating ways to enhance algorithm efficiency, implement hardware acceleration, engage in parallel processing, and build real-time processing frameworks.

In the healthcare business, diagnostic precision is of the utmost importance. Incorrect diagnosis may cause physicians to delay initiating appropriate treatment. Errors in diagnosis may sometimes result in improper therapy, which can be severe [4]. Inappropriate or lack of therapy may increase the fatality rate and, in certain cases, result in organ loss. Thus, it is usually required that automated procedures generate reliable findings and extract relevant data within a certain time frame. Hence, sophisticated analytic algorithms alone are insufficient to provide high-quality findings [5]. Techniques for image preprocessing must be robust enough to improve picture quality and facilitate the following phases of processing. Biomedical pictures are very noise-sensitive [6]. Since the quality of the images has a direct bearing on the reliability of the assessment techniques, it is well acknowledged that it is of paramount importance in biomedical image analysis.

Clarity of images allows for accurate study and interpretation of anatomical structures, lesions, or abnormalities by both human doctors and computerized algorithms.

Detection Image quality has a direct effect on the detecting systems' sensitivity. Low resolution or artifacts in the image can be interpreted incorrectly, leading to missed or misinterpreted findings that throw off the accuracy of the evaluation. Image quality has an effect on the precision with which quantitative metrics and feature extraction are extracted. Well-defined boundaries, accurate pixel values, and low noise all contribute to precise measurements and all are affected by image quality. Inconsistencies in ratings are minimized when there is a consistent level of quality across photographs. The accuracy of an analysis can be negatively impacted by inconsistencies in quality, such as those caused by differences in contrast or lighting. Quality photos are essential for achieving repeatable results. If the same image was acquired in varied quality, the results may not be comparable or repeatable.

Accurate assessments in biological image analysis rely on good picture quality since it facilitates clear visualization, increases detection sensitivity, enhances quantitative measures, maintains consistency and dependability, and encourages reproducibility of results.

Furthermore, these images contain a significant amount of information that must be carefully maintained to provide the required accuracy and security. Hence, image augmentation techniques must consider the effectiveness of the pictures. Adjusting contrast is one of the most important steps in picture improvement. Modifying the intensity range enhances the image's transparency and allows for the discovery of previously unknown patterns and data. Optimization of contrast may make a picture more easily decipherable by both computer vision systems and humans [7]. Certain approaches can maximize an image's contrasts, but as an undesirable side effect, they alter the image's fundamental architecture.

The use of big data technology in healthcare analysis may result in improved application performance [5]. Big data pertains to large data sources that incorporate the following features: volume, which refers to large amounts of information; speed, which indicates that data is generated quickly; variety, which emphasizes that information comes in various formats; and, eventually, factuality, which indicates that information comes from reliable sources.

Volatility is another property of large data sets. It shows data flow rate fluctuations. Due to the information's repeated peaks and valleys, speed does not offer a consistent explanation [8]. Another crucial component of big data is its difficulty, which emerges from the reality that big data is frequently produced by a multitude of sources, necessitating the execution of numerous operations on the data, such as recognizing connections and sanitizing and transforming information flowing from various origins.

In the setting of medical care, many medical sources create large amounts of data, such as biomedical imaging, lab test results, physician handwritten notes, and health state metrics that enable real-time patient medical surveillance. In contrast to its vast quantity and variety, medical data moves rapidly. Hence, big data methods provide enormous prospects for improving the efficacy of medical systems [9].

This research study contributes by proposing a big data infrastructure for smart healthcare comprised of many components capable of collecting, analyzing, and analyzing voluminous data in real time and via batch methods. This study illustrates the possibilities of using big data analysis in the medical industry to extract meaningful information from very important information.

Metaheuristic Optimization (MO) (such as Genetic Algorithm) is used to feature selection and Gradient Approximation (GA) to improve the biomedical image classification results. The experimental outcomes show the system's effectiveness.

The remaining sections of the paper are listed as follows: The second part provides context for the biomedical categorization models. The design and discussion of the proposed BDA-based Biomedical Image Classification (BDA-BMIC) system are presented in Sect. 3. In Sect. 4, the software analysis and system outcomes are detailed. Section 5 exhibits the system's conclusion and conclusions.

2 Background to the biomedical classification models

Biomedical image classification seeks to identify the greatest number of features (characteristics) from the original database while maintaining classification results. The problem's difficulty rises geometrically with the number of characteristics in the datasets. Thus, Metaheuristic Search (MHS) methods have been used to enhance the acquired result and reduce computing time for huge tasks.

AIFM-CRC [10] is a revolutionary Artificial Intelligence (AI) based fusion model for the detection and categorization of Colorectal Cancer (CRC) illness. As a preprocessing step, the current AIFM-CRC model predominantly employs Gaussian filtration for noise reduction and contrast improvement. Scale-Invariant Feature Transform (SIFT) based handmade features and Inception v4-based deep characteristics are merged in a fusion-based attribute extraction procedure.

The research presents a unique fusion of convolutional neural networks to construct a more effective and economical classification for biomedical pictures that incorporates shallow and deeper layer information from the suggested deep neural network structure [11]. It was discovered that shallower layers gave more specific local characteristics that could differentiate between illnesses within the same classification.

The research presents a new deep feature extracting and categorization approach dubbed Diagonal Bilinear Interpolated Deeper Residual Networks (DBI-DRSN) [12] for biological pictures. The DBI-DRSN approach combines a balancing of information or features using the Directional Bilinear Interpolation pretreatment model and classification of the characteristics using the Deep Residual Networks model's fine-tuning.

The research presents a Synergistically Deep Learning (SDL) model to solve this problem by using several Deep Convolutional Neural Networks (DCNNs) concurrently and allowing them to learn from each other [13]. The learned picture representations of each pair of DCNNs are combined as the source for a fully linked synergic network that predicts whether a pair of image pixels belongs to the same category.

The research presents a novel mixed convolutional and recurring deep neuronal network for image categorization of breast cancer histopathology [14]. Based on the deeper multilevel feature extraction of the histopathology input images, the technique combines the benefits of convolutional and recurring neural networks while preserving the short- and long-term spatial connections between fragments.

This study shows an ensemble deep-learning technique [15] for the classification of cancerous breast pictures. Based on pre-trained VGG16 and VGG19 designs, the research educated four distinct models. The review discusses automated picture segmentation using deep learning techniques in the field of diagnostic imaging [3]. Recent advances in machine learning, especially those linked to it, are becoming useful for identifying and quantifying trends in medical image data.

CheXGCN [16] is a revolutionary label co-occurrence learning system that actively explores the relationships between diseases for the multi-label breast X-ray picture categorization challenge. Image Feature Embedding (IFE) and Labeling Co-Occurrence Learning (LCL) are the two parts that make up CheXGCN.

The objective of the current work is to develop effective deep-learning algorithms, trained on chest X-ray images, for quick COVID-19 patient assessment [17]. The research developed Artificial Intelligence based categorization and other important contagious illnesses using datasets of adult patients.

This study proposes a dilated Convolutional Neural Network (CNN) model that is constructed by substituting the convolutional kernels of regular CNN with dilated convolutional units, which is then evaluated on the Mnist handwriting digital identification data set [18]. Secondly, to address the issue of detail loss in the expanded CNN models, the Hybrid Dilated CNN (HDC) is constructed by layering dilated convolutional kernels with varying dilation speeds.

3 Proposed BDA-based biomedical image classification system

In recent years, the use of Machine Learning (ML) technologies with Support Vector Machine (SVM), particularly Deep Learning (DL) with CNN, for biological image categorization research has gained popularity. The primary

aim of medical picture categorization is to compute which portions of the person are infected with the illness, not only to obtain high efficiency. Based on the preceding section, there are two categorization techniques in the suggested workflow: one for a medium-sized database and the other for a large database. In this connection, SVM and DL are then utilized accordingly.

The system process for biological image categorization is shown in Fig. 1. As seen in the flowchart, the categorization process consists of two fundamental processes. In the first stage, ML (SVM or CNN) techniques are used to construct a classification structure based on the tagged biomedical pictures of two fundamental processes. The following are some of the ways in which machine learning techniques are applied to the process of classifying tagged biomedical images:

Machine learning algorithms learn from a database of labeled medical images called training data. Images are manually annotated with their classes and categories to provide training data with essential ground truth.

Learning Relevant qualities for Each Category With the use of tagged photos, machine learning algorithms may learn the pertinent qualities associated with each category. The labeled photos are analyzed by the algorithms, and features that distinguish between classes are extracted.

Parameters of the machine learning model are fine-tuned with the help of the labeled photos. Training the model involves making small, incremental changes to the model's internal parameters in order to reduce the discrepancy between the projected classes and the actual labels of the tagged images.

Images that have been annotated are used to test the accuracy of the categorization system and measure improvement over time. Classification results may be trusted because the trained model is evaluated on an independent collection of labeled pictures using several metrics like as accuracy, precision, recall, etc.

A more universal system of categorization can be built with the help of tagged photos. The model is strengthened and improved in its ability to categorize unseen, untagged biomedical images by including a wide variety of tagged photos representing distinct variants and complexities within each category.

Machine learning algorithms may learn and optimize the classification structure with the use of labeled training data obtained from tagged biomedical images. In this way, new biomedical images can be reliably and accurately classified through feature learning, model optimization, performance evaluation, and generalization of the classification structure.

Machine learning techniques are used to construct a classification structure based on the tagged biomedical picture. After the classification system has been created, any unlabeled biomedical pictures may be provided to it to determine the group to which they belong. Figure 1 depicts a high-level system method for classifying biological images, which includes the following steps: collecting the image, performing any necessary pre-processing, extracting features, training a classification model, making predictions, and assessing the model's performance. These procedures improve the overall efficiency of categorizing biological images and allow for more precise classifications to be made. There are two main steps involved when classifying biological images:

Feature extraction is the procedure of identifying and separating out the elements that are most important in a biological image. These properties are excellent in capturing distinguishing physical traits between organisms, such as texture, shape, or colour. In order to quantify the information contained in an image, feature extraction methods including edge detection, texture analysis, and shape descriptors are typically used.

When the characteristics have been extracted, the next step is to classify the images into the appropriate

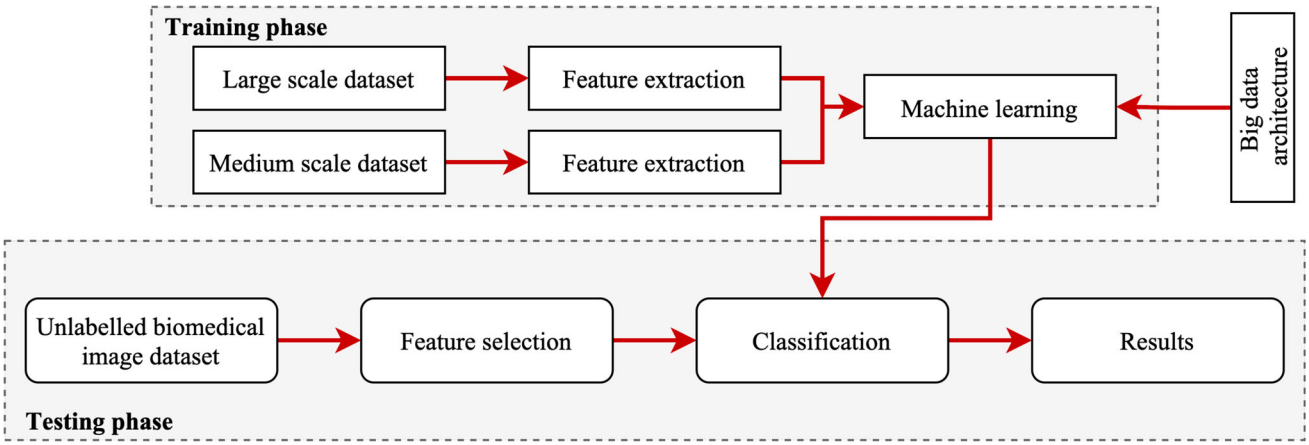


Fig. 1 Workflow of the proposed BDA-BMIC system

categories. Classification algorithms are used for this purpose, and they gain knowledge from the extracted features and the labelled training data. Using the discovered correlations and patterns in the feature space, the categorization model assigns items to their appropriate classes.

The technique of classifying biological images allows for the automatic identification and grouping of images based on typical properties by combining feature extraction and classification. This makes it easier to analyse, organize, and interpret massive amounts of biological imaging data, which in turn improves our knowledge and understanding across many fields of biology.

The CNN structure for biomedical picture categorisation is shown in Fig. 2. The system consists of a convolutional layer, Recurrent Learning Unit (ReLU) level, a pooling level, fully connected level to find the final results using output classes [19]. The training step of categorization involves presenting the information from the training database (labelled biological pictures in this example), extracting characteristics, and training the system by translating the source to the predicted result. Using the gradient descent method, the system can now learn. The objective of the BDA-BMIC system is to determine the network weighting factors that minimize the difference between the actual and predicted outcomes. The Back Pressure (BP) algorithm enables the system to estimate how significantly the weights of bottom-layer networks must be modified by the BDA. Typically, the training process comprises three key steps: retrieval of labelled biomedical imaging datasets, extraction of features, and the machine learning method.

Figure 3 depicts the architecture of the suggested BDA-BMIC system for the biomedical classification process using Genetic Algorithm (GA) and Metaheuristic Optimisation (MO). This research uses photos of aberrant brain tumours captured in actual environments [20]. This study employs 450 photos from four distinct aberrant groups, including metastasis, glioma, astrocytoma, and meningioma. These photos are grayscale images measuring 256 by 256 pixels. A comprehensive collection of features is retrieved from these photos. This study employs fourteen textural properties taken from the Gray Level Difference Matrix (GLDM). Depending on the study's methodology and goals, there may be a number of different ways to capture all of the relevant features from images. However, in the field of biological image analysis, feature extraction often necessitates multiple stages:

Noise reduction, image enhancement, and normalization are all examples of preprocessing techniques that could be applied to the photographs before they are used.

Images are analyzed to locate and isolate regions of interest (ROIs) that contain the structures or abnormalities of interest. This process aids in identifying and isolating key regions before feature extraction.

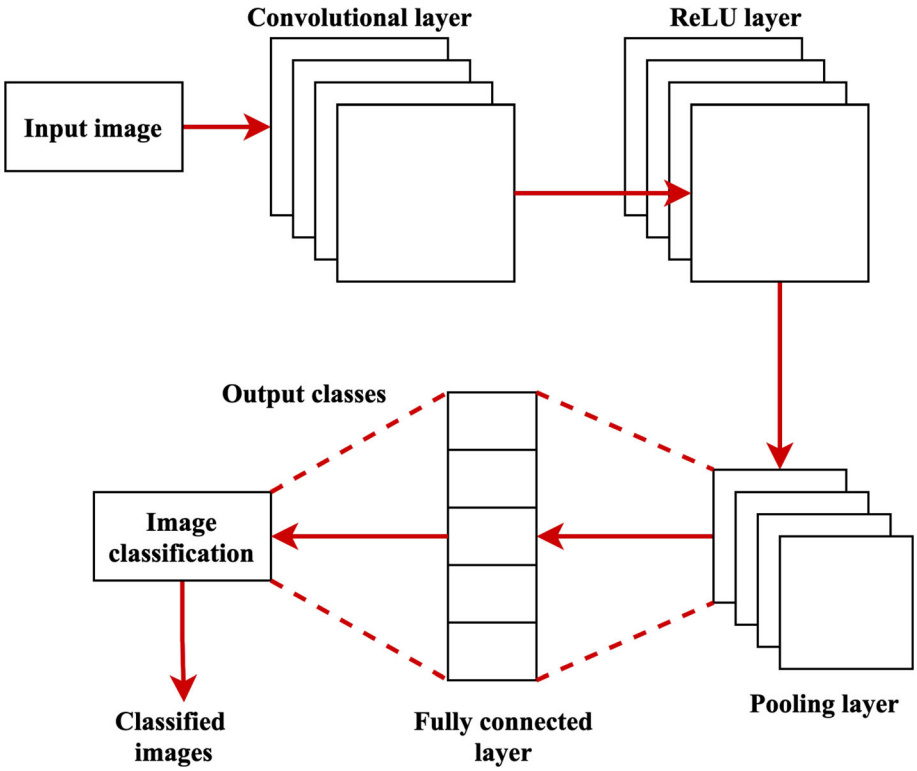


Fig. 2 CNN structure for biomedical picture categorisation

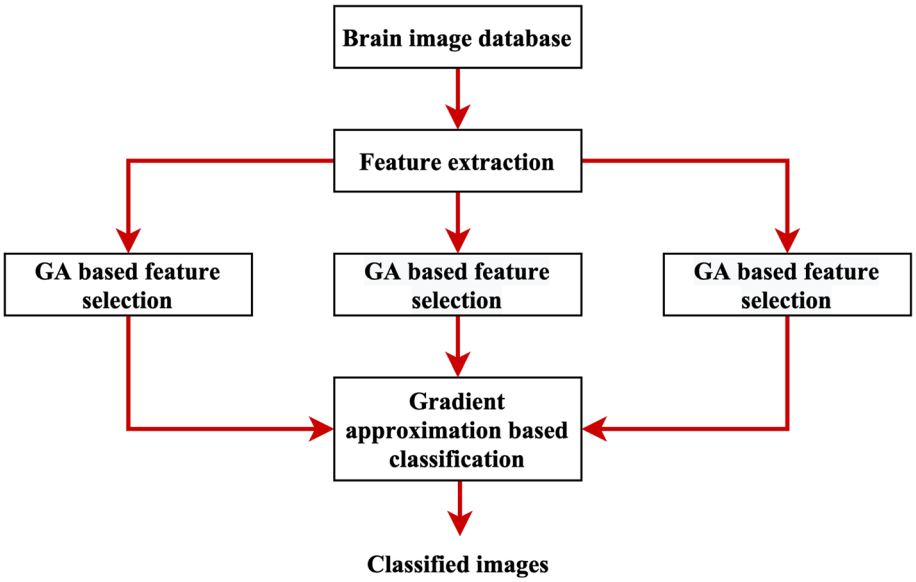


Fig. 3 GA and MO-based biomedical classification process of the proposed BDA-BMIC system

The segmented ROIs then undergo feature extraction, during which a number of distinct methods are used. These methods may make use of either learnt or hand-made components. Features that have been hand-crafted are mathematical representations that have already been defined, while features that have been learnt are the result of deep learning models or other machine learning techniques. Images' unique shapes, textures, intensities, and statistical aspects are all captured by these attributes.

The Gray Level Difference Matrix (GLDM) is a tool for texture analysis that provides a quantitative description of the spatial variation in pixel brightness across a picture. The incidence frequencies of individual gray level deviations are used by GLDM to derive a number of statistical metrics. There are many different types of statistical measurements that can be used. Understanding an image's textural qualities is made possible by GLDM, which records data about the variations in pixel intensities. These GLDM-derived textural properties can serve as features for biological picture categorization and analysis.

The optimal collection of characteristics is then determined using the suggested GA techniques. Each customized GA method will result in distinct characteristics. This study employs a Back Propagation Neural (BPN) system as a classification to verify the effectiveness of the suggested GA approaches. In terms of reliability, responsiveness, and effectiveness metrics are evaluated. The BDA-BMIC architecture has multiple benefits for simplifying biomedical classification:

Metaheuristic Optimization (MO) and Genetic Algorithm (GA) Combination: These optimization strategies, when applied to the classification process, can improve its precision and productivity. If you need help choosing which features to employ for classification, the GA can lend a hand. The MO can be used to fine-tune the model's settings, leading to better results from the classification procedures.

It is possible that the BDA-BMIC system's architecture was developed to efficiently process massive biomedical datasets through the use of parallel processing and scalability. Distributed computing and parallel processing techniques could be used to speed up the categorization process by processing and analyzing photos more efficiently.

Biomedical classification, specifically images of atypical brain tumors obtained in their native environments, the unique aspects of which may include:

Atypical brain tumors may have sizes and shapes that are out of the ordinary for brain tissue. The tumour's malignant or benign status may be inferred from these characteristics.

Atypical brain tumors may have areas of higher contrast on imaging scans than the surrounding healthy tissue. There may be useful clues for categorization in the existence of these contrast variations.

Tissue Texture and Density Tumorous regions may have a different tissue texture or density than the rest of the brain. Different approaches to texture analysis can capture these variations, which can then be utilized as characteristics to classify objects.

Intra-tumor heterogeneity is commonly seen in atypical brain tumors, with various parts of the tumor displaying distinct cellular features. Observing these spatial differences may help doctors identify subtypes of cancer.

Depending on the study's methodology, picture modalities, and analytic procedures, the actual identifying features used may be different from one study to the next. Figure 3 is described in more detail, along with some

insights into the characteristics of the used pictures, in the corresponding journal or source, which should be consulted for further information about the specific study and its findings.

The BDA-BMIC system focuses mainly on feature extraction and classification process of the BDA-BMIC system and the details are discussed in the subsection.

3.1 Feature extraction

The extraction of characteristics is a crucial step in a picture-based automated categorization system. It is employed to capture the important distinguishing characteristics of pictures from several categorisations, that assists the classification in making correct classifications. Fourteen textural characteristics were retrieved from the photos in this study. These pictures are derived from the input picture using the GLDM. Characteristics derived using GLDM are referred to as statistical characteristics of a higher order. The formula for estimating GLDM is expressed in Eq. (1).

$$R(x,y,l,\theta) = \sum_{i=0}^{N}\sum_{j=0}^{N} \delta(x - p(i,j))\delta(y - p(i,j) + d\theta) \quad (1)$$

$p(i,j)$ denoted the pixel intensity, $p(i,j) + d\theta$ denoted the neighboring pixel's luminance value at a given distance and direction angle. The deviation is expressed δ, and the angle variation is expressed $d\theta$. The deviation between the pixels is expressed in Eq. (2).

$$\delta(i-j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases} \quad (2)$$

In the formula above, i and j represent the intensities of two distinct pixels. The variable l represents the movement, whereas the parameter θ represents the direction angle. The movement and degrees 0, 45, 90, and 135 are studied for the use of this approach. The values of the matrices are summed across the four directions. The formation of a full matrix that represents the probability of paired occurrences of pixel brightness values. To build the probability vector, the variables are further standardized.

3.2 Feature selection

The selection of features is the most important process in a healthcare image categorization system. The existence of irrelevant characteristics affects the system's overall precision [21]. This attribute selection phase avoids this particular issue and increases the method's success. Bio-inspired optimizing approaches are frequently used to select characteristics in healthcare images. Genetic algorithms are the first bio-inspired evolutionary computation strategy.

The process that the genetic algorithm goes through can be seen in Fig. 4. The process of population size initialization begins after the extraction of the input parameters derived from the selected characteristics. To discover the most recent version of the population feature, the fitness function is calculated. Repeating the procedure is necessary to get the desired result.

The stages included in the modified genetic algorithm are shown below:

Stage 1: Initialization of the community and chromosomal representation of each individual Each gene is denoted by 12 bits representing a distinct characteristic. The starting population size is 20 individuals with random genomes.

Step 2: Prediction of each participant's fitness level using an optimal solution. This study uses the fitness value provided by Eq. (2).

$$F = (k_1 * k_3) + k_2 * \left\{\frac{N_f - L_c}{N_f}\right\} \quad (2)$$

k_3 is expressed as the classification accuracy, and N_f is expressed as the total number of features. The CNN weights are expressed k_1 and k_2, and the length of the chromosome is expressed L_c. The weights of the CNN model are expressed in Eq. (3).

$$k_1 \in \{0,1\} \text{ and } k_2 = 1 - k_1 \quad (3)$$

The categorisation result is assessed using the chromosomal characteristics, with 1. Every member will provide distinct values. The CNN weights are expressedk_1 and k_2

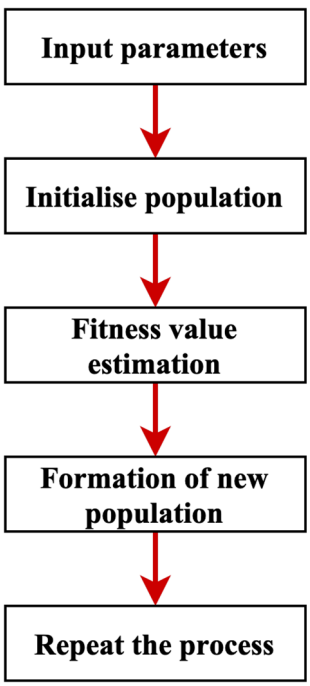


Fig. 4 Workflow of the genetic algorithm

Step 3: Elimination of the poorest individuals and production of new members via reproductive operations like crossover and mutation.

Step 4: Using the same fitness component, estimate the fitness value of new individuals in the community.

Step 5: Continue the procedure a certain number of times until the ideal output is achieved. The characteristics with a bit value of "1" are allowed, while the characteristics with a bit value of "0" are refused.

Unfortunately, traditional GA has several disadvantages that restrict its practical implementation. Many improved techniques are given in this work to address these disadvantages. These updated strategies are explained in the sections that follow.

3.2.1 Modified GA1 method

The successive proportion of any GA method is highly dependent on the crossover operations and mutation functions, who are acceptable for populating the population with new individuals. Usually, the procedure is done from the populace. Likewise, the choice of crossover sites in the traditional approach is similarly wholly arbitrary. In the suggested GA1 method, necessary adjustments are made to address these shortcomings. The efficiency with which a genetic algorithm (GA) generates novel individuals from the members of an existing population depends critically on the proportion of crossover operations and mutation functions present in the GA, as discussed in the context of the Modified GA1 approach.

The crossover process controls the transfer of genetic information between two individuals, resulting in children having a hybrid of their parents' traits. The rate at which this is sent is based on the proportion of crossover operations. Increasing the proportion may lead to more people exploring the search space and increasing population diversity, both of which may improve convergence.

Also, the mutation function can be used to randomly alter specific chromosomes and so expand the search field. This variation's frequency is based on the proportion of mutation operations. An increase in the proportion can encourage variety and forestall an overly rapid convergence to poor solutions.

Traditionally, genetic crossover locations in GA have been determined using either a uniform crossover algorithm or a set of predefined crossover points. Crossovers occur at predetermined points on the chromosome called fixed crossover spots. During a uniform crossover, genes are crossed over from both parents at random.

Traditional crossover approaches have the drawback of being deterministic and, thus, may miss significant parts of the search area. They can cause the population to converge on local optimum solutions and hinder the search for the global optimum by the algorithm. Adjustments to the crossover method, such as the use of adaptive or variable crossover techniques that permit more versatile and diverse genetic exchanges based on the situation at hand, are required to increase exploration and exploitation of the search area in light of this restriction.

The rationale for the modified strategy is that "the children produced by two powerful parents must also strongly match the grandparents." Because crossover operations need two parents, the child must strongly match the strongest parent. This goal will prevent the unpredictability seen in the usual crossover procedure. The likeness between parents and children is determined by their numerical values. Although 14 characteristics (each bit corresponding to an input characteristic) are used in this study, the suggested method uses fewer characteristics to simplify. Any number of characteristics may be added to the same procedure. By proposing a new approach to crossover operations and the identification of crossover locations, the GA1 method hopes to remedy the shortcomings of the conventional method. The revised strategy's emphasis on offspring's genetic similarity to grandparental generations is central to its justification. The revised approach seeks to both expand the population's horizons and celebrate the strengths of its most successful members (parents).

The GA1 proposal recommends a crossover operation in which a highly fit member of the population is used as a reference, and the population member with the highest resemblance to the reference is chosen. Henceforth, this person shall be addressed as Grandparent. Based on the corresponding positions in the reference parent and the grandparent, the crossover sites are then identified.

The idea is that the successful attributes of the powerful parent can be passed on to the kids by making sure there is a good genetic match between the child and the parent with the most desirable traits. This aids in preserving useful characteristics and speeds up the population's progress toward optimal solutions.

The GA1 method improves the efficiency and efficacy of the algorithm in producing new people with desirable traits by placing a premium on pairing offspring with influential parents. When used during the crossover phase, it aids in retaining desirable features and avoiding the loss of important genetic information. As a result, the algorithm's exploration-exploitation balance is enhanced, and it may be able to better sift through the population for optimal or near-optimal solutions.

The parents for the classification of biomedical image processing using GA are expressed in Eq. (4).

$$\begin{array}{ll} 11010011 & \text{parent 1} \\ 10000101 & \text{parent 2} \end{array} \tag{4}$$

The binary digits are transformed in the first stage. Parent 1 correlates to the numerical value 219, and Parent 2 relates to the numeric value 141 in the preceding example. Parent 1 is the healthiest if the goal is the maximization value, while Parent 2 is the healthiest if the goal value is the minimizing feature. Let's compare the two scenarios.

Case 1 Maximization issue.

Any mother (criterion 1) with a greater number of 1's and (criterion 2) any mother with a large number of 1's on the Most Significant Bit (MSB) side is the most suitable parent [22]. The existence of additional ones results in a greater numerical value. All bits on the right half of the subdivision are on the Least Significant Bit (LSB) side, while all bits on the left half of the subdivision are on the next side. Hence, both the "location" and "quantity" of 1's are taken into consideration in this technique, that shows the suggested approach to its resilience. According to the reasoning employed, the progeny must be similar to Parent 1 more than Parent 2 in the preceding scenario. Hence, the progeny must likewise include more 1s than 0s. For this purpose, a binary-OR procedure is carried out between the two parents. According to the OR data table, three-fourths of the outcomes are 1. This strategy eliminates the need for crossing sites, the primary source of unpredictability for the crossover operators. The OR operation results are expressed in Eq. (5).

$$\begin{array}{ll} 11010011 & \textit{Parent}\ 1 \\ 10000101 & \textit{Parent}\ 2 \\ 11010111 & \textit{offspring} \end{array} \tag{5}$$

The offspring has a numeric value of 215, which is nearer to the value of the parent with the highest fitness. According to the rationale of this technique, this strategy also eliminates the typical way of simultaneously producing two children, which might cause misunderstandings. The presence of the number "1" in the MSB may also confuse. In this situation, a parent will be deemed the most suitable despite having a greater number of "0s" in the remaining spots. In these circumstances, condition 2 must take precedence over criteria 1. Consequently, a progeny is produced devoid of crossover spots, the primary source of unpredictability in reproductive operations.

Case 2 Minimization issue.

Any mother (criterion 1) with a greater number of zeroes and (criterion 2) any mother with a large number of zeroes in the MSB are the most suitable parents [23]. The offspring must match the parent whose numerical value is the smallest. This approach uses a binary AND function to produce the offspring. According to the AND truth tables, three-fourths of the outcomes are 0. The AND operation result is expressed in Eq. (6).

$$\begin{array}{ll} 11010011 & \textit{Parent}\ 1 \\ 10000101 & \textit{Parent}\ 2 \\ 10000001 & \textit{offspring} \end{array} \tag{6}$$

The created offspring has a numeric value of 129, which is nearer to the value of the mother with the highest fitness. Condition 2 must be given more weight than condition 1. Hence, the inherent unpredictability of the traditional crossover approach is avoided entirely. Each parent is joined with the other healthiest parents as a further modification to the suggested GA1 strategy. All parents undergo the same process of crossing over. All the children produced are sorted, and the fittest will be chosen. This strategy thereby eliminates the random sample of parents employed by the usual method. In the end, the ideal chromosome is selected, and characteristics with the bit location "1" are utilized in the categorization procedure.

3.2.2 Modified GA2 approach

In a second way, the offspring are generated using the idea of grey code. With this technique, just one father is used to produce kids. The required information is regarded as being in binary format. This binary information is transformed into grey code, which represents the progeny. The translation from binary to grey code includes two concepts: As grey code is a unit-length code, the differences between binary digits and grey code will be small. The MSB of binary and grey codes will remain unchanged. These two requirements will assure that the created offspring substantially resembles the parent, per the purpose. Imagine an example of procreation using this methodology. Let the fittest parent survive. The parent for the GA2 process is shown in Eq. (7).

$$11010011 \quad \textit{Parent} \tag{7}$$

The offspring are produced using the method:

Stage 1: The first MSB of the mother is preserved in the offspring.

Stage 2: The next bit of the offspring is formed by combining the first bit of the mother with the second bit of the mother. Equation (8) shows the technique for the exclusive addition.

$$\begin{array}{ll} \textit{Inputs} & \textit{Output} \\ 1 \text{ and } 1 & 0 \\ 0 \text{ and } 0 & 0 \\ 0 \text{ and } 1 & 1 \\ 1 \text{ and } 0 & 1 \end{array} \tag{8}$$

Stage 3: The procedure is done with subsequent bits to create the various offspring bits. The offspring produced by the preceding example is expressed in Eq. (9).

$$\begin{array}{ll} 11101001 & Parent \\ 1010011 & offspring1 \\ 0111010 & offspring2 \end{array} \quad (9)$$

Stage 4: Examine the amount of distinct bits and placements in both the mother and off springs.

Stage 5: The procedure is done for all fathers.

Stage 6: The children who have the fewest differences from their parents are chosen to be the next generation. Finding the smallest difference between the mother and the child depends a lot on where the bits are.

Hence, modified GA2 proves to be an alternative to the standard genetic algorithm's crossover procedure. When children are created according to a predefined logic, the unpredictability seen in traditional GA is removed entirely. This minimizes the likelihood that inefficient children will survive. Furthermore, the procedure eliminates in traditional GA because it is performed with all potential fittest parents. Based on the operational metrics, the successive function of the suggested improved GA2 method is analyzed.

3.2.3 Modified GA3 model

An altered mutation operation is used to produce progeny. Traditionally, just one of the fittest parents is chosen for the mutation operation. Furthermore, a bit is drawn at random from the children. To create the offspring, this bit is flipped ('1' for '0' and '0' for '1'). Hence, the difference between parent and offspring is minimal. Nevertheless, this strategy has a hidden disadvantage. The danger is that the typical technique does not emphasize the location of the bit be modified. The location of the bit has a significant effect on the processes of mutation that produce children with comparable characteristics to the fittest mother. Assume the parent for the GA3 process is expressed in Eq. (10).

$$10111010 \quad Parent \quad (10)$$

If the MSB is modified at random, the off springs will vary substantially from the mother. For the preceding data, the off springs are expressed in Eq. (11).

$$00111010 \quad offspring \quad (11)$$

Even if the third bit of MSB is modified, the offspring will be drastically altered. However, if the LSB is altered, the offspring is identical to the mother. So, it is preferable to modify a bit that is located in the centre of the mother. This procedure will produce the best children for the future generation. Yet, the true difficulty is in locating the finest portion of the mother. In the suggested method for offspring production, the following basic stages are employed:

Stage 1: Pick the optimal parent and determine if the issue is lower or higher.

Stage 2: Switch the bit from "0" to "1" for the maximum issue and "1" to "0" for the minimum problem.

Stage 3: It entails dividing the mother into three parts of three to four bits each. The parent may be split into several segments if the size is higher. But if the number of bits in a section is increased, even a single bit change could cause a bigger change.

Stage 4: Select the middle section if the number of sections is odd. In the event of even sections, any section in the centre location may be selected.

Stage 5: Finally, depending on whether the solution is a maximization issue or a minimization issue, the bit to be modified is picked.

Stage 6: For higher problems, the section's bit with the value "0" is selected. If several bits have the value "0," the bit that is placed first on the MSB is selected for reversal. If the segments do not include bits with a "0" value, the section of the centre section is selected, and the same technique is used to invert the bit.

Stage 7: In the event of a lower issue, the bit with the number "1" is selected from the section. If several bits have the value "1," the first bit from the LSB is selected for reversal. If the section does not include several 1, the procedure will be used to choose the bits.

Significant characteristics of the suggested method include:

- The offspring will not vary much from the fittest mother.
- The quantity of the offspring is increased by selecting the "0" bits for the maximization issue.
- The worth of the offspring is decreased by selecting the "1" bits for the minimizing issue.
- This enhanced GA3 technique for offspring production takes into account both the quantity and location of bits. Hence, the disadvantages of traditional mutation operations are addressed by this suggested method.

This paper proposes three feature-based selection approaches based on the basic premise that children should resemble their mothers for improved results. Even if these procedures are intellectually sound, they must be evaluated empirically to confirm their efficacy. In the tests, the collected characteristics from each approach are submitted to a classification independently, and the outcomes are evaluated.

3.3 Gradient approximation (GRA)-based classification

Some modifications are made to the CNN in light of the newly developed GRA method for enhancing categorization accuracy [24]. In this study, the GRA optimizing procedure is used to both maximize the activation function

and the epoch to enhance the categorization procedure. CNN activation functions play a crucial role in enhancing the learning process by adjusting network variables appropriately. So, choosing the proper activating function impacts CNN training. Many typical CNN-activating algorithms are detailed below.

Sigmoid: The sigmoid activating value is dependent on the logistic function, causing the values of the attributes (k) to fall between 0 and 1. The theoretical functionality of the logistic sigmoid is expressed by Eq. (12).

$$F(k) = \frac{1}{1 + \exp(-k)} \tag{12}$$

Tanh: This represents the exponential tangent functions and hence takes on the hyperbola's result. Depending on the proportion of the sine and cosine ratios, the exponential tangent offers certain features (k). The Tanh function is expressed in Eq. (13).

$$F(k) = \frac{\exp(k) - \exp(-k)}{\exp(k) + \exp(-k)} \tag{13}$$

Relu: The non-linear translation uses the real unit (k), which has the value indicated in Eq. (14).

$$F(k) = \begin{cases} k & k > 0 \\ 0 & \text{else} \end{cases} \tag{14}$$

Recurrent ReLU (RReLU): The RReLU variable is employed to describe samples with a negative slope, and expressed in Eq. (15).

$$F(k) = \begin{cases} k & \text{if } k > 0 \\ \propto (1, \theta) & \text{else} \end{cases} \tag{15}$$

Therefore, the parameter $\propto$ is defined as falling between and (0, 1), and the input is denoted k.

Periodic maximization: The interval value is one of the crucial factors evaluated during machine learning training. The iteration value specifies the number of times the complete data set is sent forward and backward during the classification model. In other words, an epoch value of 1 implies a one-time forward and reverse feed in the CNN classification using GRA. Selecting a small number for the epoch may result in a problem with a parameter, while selecting a large amount may increase computing complexity. Changing the epoch value may thus regulate the training duration and increase categorization performance. There are a number of concepts and considerations that must be addressed while discussing machine learning training.

Periodic Maximization's interval value describes how often specific training-related actions or updates take place. It controls how frequently tasks like model parameter changes and evaluations are carried out. The dataset's complexity, the availability of computational resources, and the desired balance between training speed and accuracy all play a role in establishing the interval value.

In machine learning training, forward propagation is when data is fed into a model to generate predictions, and backward propagation is when the gradients are calculated and the model's parameters are updated based on the new data. How many rounds of forward and backward propagation are carried out in a single training cycle is represented by the iteration value. It represents the fineness of the parameter adjustments made during each cycle.

Time stamp for CNN's GRA-based classification system: One whole cycle through the CNN model's training dataset is called an epoch. Each epoch, the model iteratively updates its parameters based on both forward and backward propagation of all training examples in the dataset. If the epoch is set to 1, the model is trained once using the full dataset.

Selecting a small value for the number of epochs used in the model increases the risk that it will fail to accurately represent the underlying patterns in the data, a phenomenon known as underfitting. If the model isn't given enough time to learn complex relationships and adjust its parameters, it may not perform as well and produce less accurate results.

Selecting a large epoch value might cause overfitting, in which the model becomes overly specific to the training data and fails to accurately predict new data. To the detriment of its ability to learn from novel data, the model may become overly dependent on its training set. Additional computing resources and time are needed to finish the training process when the epoch value is increased.

To minimize overfitting or undue computing overhead, while still giving enough training iterations for the model to acquire meaningful patterns, is the optimal tradeoff when deciding on an epoch value. Finding the best epoch value for a given problem and data collection usually needs some trial and error, as well as validation using independent test data.

3.4 Big data architecture

Volume (the amount of data generated) is one of the characteristics of large data. To implement the categorization process in a big data architecture, the research must validate this rule for the database supplied to the training stage of the process. Support vector machines (SVMs) and convolutional neural networks (CNNs) are two examples of ML techniques used to build a classification structure in the first stage of the categorization process.

Support vector machines (SVMs) are trained to determine a cut-off value for a set of features that provides the largest possible gap between classes. SVM determines the most important training data points, or support vectors, for

classification through supervised training on labeled data with known categories. The trained SVM model can be applied to data that has not yet been observed to classify it into one of the feature spaces.

In other words, CNNs were developed to do picture classification tasks. Convolutional layers, which are a part of CNNs, are used to extract hierarchical characteristics from input images. Next, fully linked layers receive the learnt features and use them to categorize data. CNNs are optimized for classification accuracy by being trained on labelled data and having their internal parameters adjusted via back propagation.

A classification structure is built using labeled data in both the SVM and CNN techniques. A decision boundary is found using SVM, and hierarchical features are extracted using CNN. An efficient and successful method of categorizing biological photos, these models allow for the automatic classification of new, unlabeled images based on the patterns and features established during the training phase.

Taking into account the size of the database on which categorization may be performed, the research can use SVM or DL as described in the preceding section. Many efficiency measures, including sensitivity, precision, selectivity, and F-score, may be used to assess the effectiveness of the network. The sensitivity of a classification model is the percentage of properly recognized positive instances; the specificity of a classification algorithm is the proportion of successfully recognized true negatives; and the accuracy of a classification model is the total percentage of right classifiers. The Spark architecture is one of the most effective large data processing platforms.

Figure 5 depicts the big data architecture that is used for the categorization of biological images. The design includes both master and slave nodes, and the slave nodes are responsible for storing the biological data in the cloud. The noise that was originally present in the biomedical image is removed after accessing it via the Hadoop Distributed File System (HDFS). First, machine learning is used to figure out how to group things, and then a model is made.

HDFS noise removal from biomedical pictures can affect classification. It affects classification:

Improved accuracy: Biomedical picture noise can distort and impair classification algorithms. Eliminating noise improves image quality and classification accuracy. Biomedical applications require precise classification for proper diagnosis and analysis.

Enhancing feature extraction: Noise can obscure biological image features, making classification systems struggle to extract essential information. The HDFS improves feature extraction by decreasing noise, letting classification algorithms focus on visual patterns and structures. This improves classification and model robustness.

Standardization and consistency: Biomedical images can have varied noise levels and origins. Noise removal may be standardized across photos using the HDFS. This uniformly removes noise, making image comparison and classification easier. Pre-processing consistency ensures fairness and dependability in classification.

Scalability and distributed processing: HDFS handles big data processing workloads across a cluster of devices. Large biomedical picture databases require distributed storage and processing. The HDFS parallelizes noise reduction across cluster nodes, enabling effective image processing of enormous volumes. In real-time or near-real-time analysis, this scalability speeds up classification.

Integration with other big data tools: Apache Spark and Hadoop Map Reduce are utilized with HDFS. These data analysis and machine learning technologies improve categorisation. Machine learning algorithms can train and classify pre-processed images after noise removal.

The HDFS eliminates noise from biological images, improving accuracy, feature extraction, standardization, scalability, and integration with other big data technologies. These parameters optimize categorization, improving biomedical image analysis findings.

Apache Spark has emerged as one of the most effective platforms for dispersed computing systems utilized in Big Data (BD) situations. Spark provides a consistent and comprehensive architecture to handle the many needs for big data analysis with a range of datasets from various sources (batching, real-time streams). Spark offers a wide variety of application programming interfaces (APIs) and packages that expand its capacity for big data processing. Among these APIs are RDDs, DataFrames, and Datasets, each of which offers a unique degree of abstraction and customization. Datasets can undergo complicated transformations, aggregations, and calculations thanks to these APIs. Machine learning (MLlib), graph processing (GraphX), and structured query language (Spark SQL) are all part of Spark's library ecosystem. To further simplify and speed up the development process, these libraries include pre-built functions and algorithms for numerous data analysis tasks.

The Spark architecture solves problems that arise when analyzing large amounts of data from many sources using several datasets. Spark provides an all-inclusive solution for effective and scalable big data analysis across different datasets by giving a distributed computing platform, support for several data storage systems, interaction with batch and streaming processing, and a rich collection of APIs and libraries.

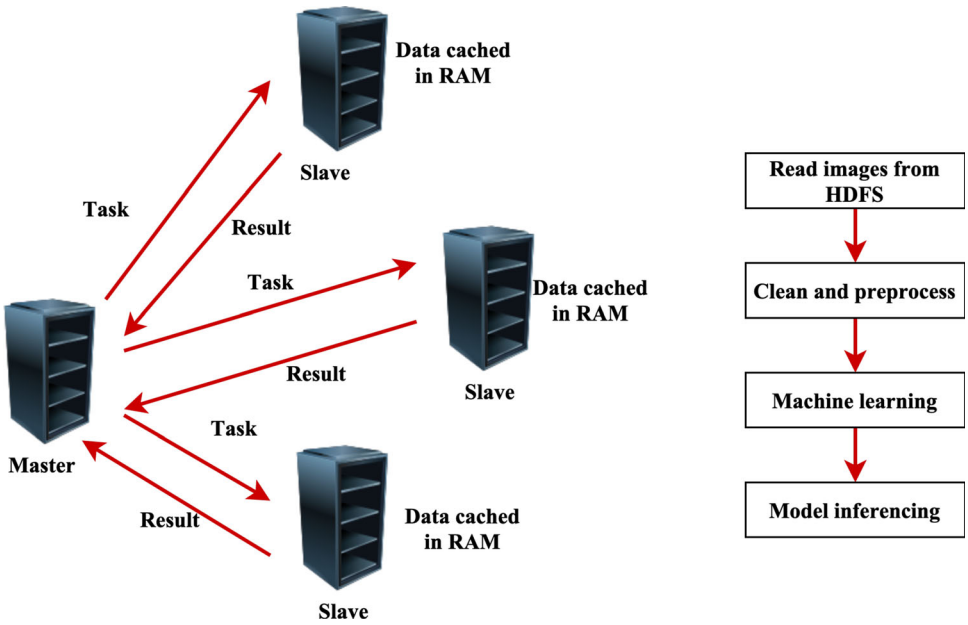


Fig. 5 The big data architecture for biomedical image classification

According to its designers, the Spark architecture was developed to circumvent the shortcomings of Hadoop. Spark architecture is more efficient than Hadoop in several instances. With features like in-memory information storage and real-time analysis, the efficiency may be quicker than that of conventional BD solutions. In the Spark architecture, the primary application (drivers) manages numerous slaves (employees) and gathers their results, while the slaves' nodes receive data segments (blockers) from a dispersed file structure, run various calculations, and write the result to disc. Spark, like Hadoop, is built which attempts to quickly and transparently process data via a group of computers. Spark also enables Structured Querry Language (SQL) queries, streaming information, pattern recognition, and graph processing of the data in addition to the map and reduce functions. On occasion, Spark allows the system to simultaneously develop and run our method on many clusters. Figure 5 depicts the information processing options in four terminals, with master and slave nodes specified. The master directs and assigns tasks to the slave. Depending on the size of the database, the system can select more or fewer than three slaves. Computation time is decreased when more slaves are employed. Usually, this pipeline comprises picture import, preparation, model training, and inference. Due to numerous factors, Apache Spark has become one of the most effective distributed computing platforms for Big Data. Its prominence and advantages over Hadoop are due to several factors:

Speed and performance: Spark processes data in memory, decreasing disk I/O operations. Spark processes data faster than Hadoop due to its in-memory processing capacity. Spark is ideal for complicated analytical workloads due to its ability to cache and reuse data.

Compared to Hadoop's Map Reduce programming model, Spark's high-level API makes it easier to use and more versatile. Developers can use their preferred language with Spark, which supports Scala, Java, Python, and R. Spark's comprehensive set of libraries for SQL, streaming, machine learning, and graph processing makes it a viable platform for Big Data applications.

Unified processing framework: Spark provides a single framework for batch processing, interactive queries, streaming data, and machine learning. Spark lets developers create complicated workflows and pipelines that seamlessly integrate these processing operations. Hadoop requires more integration because it uses Map Reduce for batch processing and Hive or Pig for data querying.

Data resilience and fault tolerance: Spark has built-in fault tolerance techniques for distributed computing. Resilient distributed datasets (RDDs) may analyze data in parallel and are fault-tolerant. RDDs automatically recover from faults and efficiently recomputed lost data partitions. Spark applications are resilient to hardware and network failures due to their resilience.

Interactive analytics and real-time processing: Spark's in-memory processing allows users to explore and analyze big datasets interactively in near-real time. Spark Streaming, Spark's streaming module, supports real-time processing and integrates streaming data with batch and interactive processing to create a unified data processing platform.

Spark works seamlessly with other Big Data technologies and ecosystems. Hadoop Distributed File System

(HDFS), Apache Hive, Apache HBase, and more can be read by it. Spark can be implemented alongside Hadoop clusters to improve processing while using Hadoop's infrastructure and data sources.

Apache Spark's speed, ease of use, versatility, unified processing architecture, fault tolerance, interactive analytics, and integration capabilities make it a popular choice for distributed computing in Big Data situations. Spark is a strong tool for data processing, analytics, and machine learning in large-scale applications due to its ability to efficiently handle different workloads.

Figure 5 shows a large data architecture for biomedical image classification to classify biological organism photos. This architecture simplifies classification. Some ways it does this:

Scalability: The architecture handles big data sets. Big data technology and distributed computing allow it to efficiently handle and analyse huge image data, which is critical for biological images.

Data storage: A distributed file system stores and manages massive image data. This simplifies data retrieval and scalability as the dataset grows.

Parallel processing: The architecture uses distributed computing to parallelize image classification workloads. It can accelerate image categorization by dividing the workload across numerous nodes, enabling real-time or near-real-time analysis.

Machine learning algorithms: The architecture uses machine learning methods to classify images. These algorithms, trained on labelled datasets, automatically learn patterns and characteristics from photos to classify them accurately.

This architecture's master and slave nodes' functions:

The master node coordinates architecture. It oversees classification. It performs:

- Scheduling slave node jobs.
- Image classification on slave nodes.
- Gathering slave node results.
- Managing workflow and synchronization.

Slave nodes: Master nodes assign image classification jobs to slave nodes. They classify visual data using pre-trained machine learning models, returning predictions to the master node. For faster classification, slave nodes process many images simultaneously.

Figure 5's big data architecture uses distributed computing, data storage, and machine learning techniques to classify biological creature photos efficiently and scalable. Master and slave nodes coordinate and execute categorization.

3.5 Testing stage

In the testing stage, the unlabeled biomedical picture dataset's characteristic arrays are used as input. On the foundation of the categorization models and its categorization criteria, a classifier determines to which class or group the characteristic vector corresponds. The classifier is essential in classification because it decides to which set of categories or groups that a specific characteristic vector belongs. It makes predictions using the vector's features by applying learned decision rules or algorithms. The classifier makes use of the training data to discover regularities and connections between the characteristic vectors and their respective classes. It compares the vector representation to the categories to determine which one is more applicable. The decision-making process of the classifier is informed by methods like closest neighbour, support vector machines, decision trees, and deep neural networks, allowing for precise classification of the characteristic vector.

The testing process consists of four major steps: acquisition of unlabeled biomedical images, extraction of features, classification architecture, and predictions. Getting the features is done the same way in the testing stage as it was in the learning stage. Characteristic arrays of unlabeled biomedical images are passed to the categorization model throughout the testing phase of the categorization process. The arrays are processed by the model, which then makes predictions about the image types based on those predictions. Extracted features representing vital visual information from the photos are stored in the identifying arrays. The model is trained on labeled data and then used to categorize photos that have not been labeled. Comparison of the predicted categories with the ground truth labels, if available, is used to assess the model's accuracy and performance.

The four main steps in biomedical image categorization testing are the capture of unlabeled images, feature extraction, classification architecture design, and prediction:

Image Capture: The testing process begins with the capture of unlabelled biomedical pictures utilizing X-ray, MRI, CT scans, or microscope. Categorization system inputs are these images.

Feature Extraction: Relevant characteristics are extracted from unlabelled photos. Texture, shape, intensity, and spatial information are these features. Wavelet transforms, edge detection, and texture analysis are feature extraction approaches.

A classification architecture or model is built once features are extracted. Image categorization is based on this architecture. It may entail choosing or designing machine

learning methods like support vector machines, convolutional neural networks, or decision trees. Labeled data is used to train the architecture to associate extracted characteristics with image categories.

Prediction Making: After the classification architecture is built and trained, it predicts unlabelled photos. The classification model uses the unlabelled images' retrieved features to classify them. Based on visual feature similarity and learnt representations, the model predicts the best category. Unlabelled biomedical pictures are categorized by the predictions.

These four procedures help organize and analyse unlabelled picture datasets by testing and evaluating biomedical image categorization algorithms.

The testing stage uses the same method as the learning stage for feature extraction. However, the feature extraction method is simply applied to unlabeled images to extract relevant characteristics instead of being trained. Images' texture, shape, and intensity are captured through extracted features. The feature extraction strategy is consistent across both stages, maintaining image representation consistency and allowing the classification model to generate accurate predictions based on learnt feature patterns.

3.5.1 Biomedical image labelling

The labelled biomedical image database is compared to the system fit using the unidentified biomedical image database.

During categorization testing, consider the following elements and challenges:

Generalization: The categorization model must generalize well to unknown data. Avoiding over fitting or under fitting, the model should perform well on new and diverse unlabeled images.

Evaluation Metrics: Classification model performance must be assessed using proper evaluation metrics. Accuracy, precision, recall, and F1 score can assess the model's performance and suggest improvements.

Unbalanced or biased datasets might affect model performance. To ensure fair and accurate classification results, account for class imbalance and resolve dataset biases.

Feature Relevance: The retrieved features must be related to the image attributes. For accurate categorization, feature extraction approaches must capture the most discriminative and informative characteristics of images.

Interpretability: Another factor is classification model interpretability. In biomedical scenarios where explain ability and openness are vital, understanding and interpreting the models decision-making process may be important.

The testing phase can produce accurate and trustworthy findings for biomedical image categorization by addressing these elements and solving the accompanying challenges.

3.5.2 Classification

Training a classification on the retrieved characteristics A classifier's objective is to differentiate between photos of the known category and those of alien classes. Hence, a classifier must learn to recognize out-of-class (alien) pictures. During the following level of predictions, the SVM and DL classifications are employed for validation. Ensure picture dataset quality and representativeness when acquiring unlabelled photos. Avoid biases in feature extraction by considering image resolution, noise, and artifacts.

Selecting feature extraction methods that capture meaningful visual information is crucial for correct feature extraction. Biomedical picture texture, shape, and intensity should be considered when selecting approaches.

Over fitting and under fitting must be addressed to make classification architecture-based predictions. Regularization, cross-validation, and hyper parameter adjustment can increase model generalization.

Class imbalance in unlabeled datasets might make predictions harder. Oversampling, under sampling, and class weighting can correct this imbalance and prevent majority class prejudice.

Finally, biomedical applications require classification architecture interpretability and transparency. Explainable AI and interpretability methodologies can improve confidence and comprehension of healthcare models' decision-making processes.

3.5.3 Forecasting

The pre-prediction step of the process dynamically predicts the category to which a picture belongs. Now, the system can examine the mean efficiency of predictions for both SVM and DL. Yet, the research demonstrates that, for big datasets, the performance of the DL classifier is typically superior to that of the classifications. Thus, for a medium-sized dataset, the support vector machine classifier outperforms the deep learning classification.

4 Simulation analysis and performance evaluation

The complete assessment of the suggested BDA-BMIC system was performed using MATLAB 2019a, and for the experimental, a dataset compiled from various samples was evaluated for assessment. The experiments were performed

up to a maximum of 25 times, with an overall population of 10. The HDFS data retrieval system was used to get data from the datasets, and its effectiveness was judged by comparing it to several criteria [25].

The steps involved in simulating the genetic algorithm are outlined in Table 1. The table contains an expression for each of the many factors, including the number of generations, the crossover percentage, the mutation percentage, the survival percentage, the number of grey levels in the biological picture, and the selection model. The BDA-BMIC system makes use of several genetic algorithms to locate the best feature selection process and makes use of the gradient approximation process optimization method to classify the data. Matlab software is used to analyze the findings.

The various classifiers, such as Multi-Layer Perception (MLP), Elaboration Likelihood Model (ELM), CNN, and DNN, as well as the recommended BDA-BMIC system, are assessed, and Fig. 6 plots the findings for accuracy and precision. Improved accuracy and efficiency in the classification of biological images are two outcomes of using ML techniques throughout the categorization process.

Improved Accuracy: ML techniques, like deep learning models or ensemble methods, can uncover previously unseen links and patterns in the picture data, allowing for more precise classification. These algorithms are able to pick up on nuance and differentiate between closely related categories, leading to more accurate classifications.

Relevant features can be automatically learned from the image data using ML methods, specifically deep learning algorithms like CNNs. This eliminates the need for human intervention in feature engineering, which both lowers the possibility of bias and increases accuracy by guaranteeing that only the most informative characteristics are used in categorization.

Both large-scale datasets and previously unknown data are manageable for ML methods. Once the models have been trained, they can categorize new biological images with great accuracy, speeding up and improving the efficiency with which complex and ever-changing image collections may be sorted.

Table 1 The simulation parameters

Parameter	Value
Number of generations	120
Cross over percentage	75
Mutation percentage	10
Survival percentage	60
Number of grey levels	255
Selection type	Gradient approximation

To better adapt to new image attributes or categories, ML models can adjust their parameters depending on fresh labeled data, a process known as "adaptive learning." This flexibility guarantees that the classification procedure will continue to be effective even if datasets change over time.

With the right optimizations and hardware accelerations, ML methods can swiftly process massive amounts of biological pictures. This allows for real-time or near-real-time analysis, improving classification efficiency and making it suited for time-sensitive medical applications.

Improved accuracy via automated feature learning, scalability, flexibility, and efficiency in terms of speed and processing power are only few of the ways in which ML techniques contribute to the classification of biological pictures. These developments help make biomedical classification systems more accurate and efficient.

The BDA and genetic algorithm both contribute to the production of higher simulation results when employing the recommended BDA-BMIC system. The gradient approximation method is responsible for obtaining the best possible outcomes from the input characteristics. Because of this, the results of the biological image categorization are better and the calculation error is even less. Hyper parameters like learning rate, regularization parameters, and network architecture can all affect how well an MLP performs. The performance of MLP can suffer if its hyper parameters are not fine-tuned for a given dataset and task. These considerations are hypothetical and largely reliant on the particular setting, dataset, and evaluation employed. To more accurately analyse MLP's performance in comparison to other studied methods, it would be necessary to conduct a full analysis of the comparative performance and detailed evaluation results.

Ensemble learning techniques, which combine many classifiers to increase predicted accuracy and resilience, may be used by the BDA-BMIC system. Combining the results of multiple classifiers allows it to take in more data and maybe perform better.

In order to find the most useful characteristics for classification, the BDA-BMIC system may use efficient feature selection techniques. Reduces noise and improves the model's capacity to distinguish between classes by zeroing in on the most important and discriminative characteristics.

To effectively store, process, and analyse massive amounts of biomedical image data, the BDA-BMIC system relies on the big data architecture. Scalability, fault-tolerance, parallel processing, and data integration all help to boost performance, accelerate analysis, and refine decision-making for biomedical image categorization.

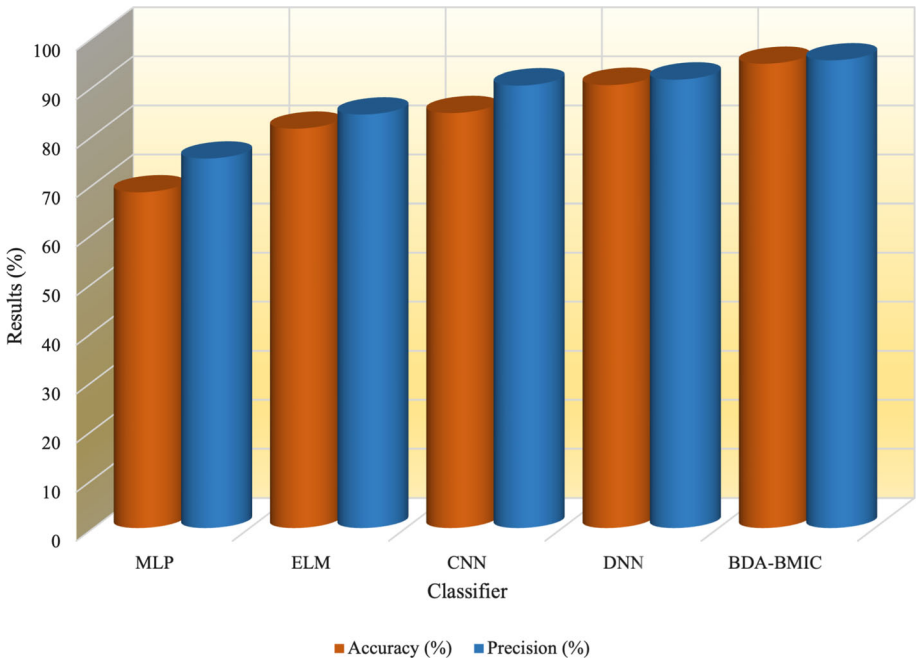


Fig. 6 Simulation performance evaluation of the different classifiers

The simulation performance of the various classifiers, including recall and F-score assessment, is analyzed, and the average results are displayed in Fig. 7. The BDA-BMIC system produces superior results when compared to those obtained using the GA, BDA, and gradient approximation methods. The BDA-BMIC system has a recall rate of 94.8% and an F score of 93.9% for the biomedical image classification procedure, both of which significantly improve the ability to identify illness from samples. The BDA-BMIC system shows an overall improvement of 17.4% and 13.8% in the F score and recall, respectively.

The examination of the specificity and sensitivity of the various classifiers, such as MLP, ELM, CNN, and DNN, as well as the suggested BDA-BMIC system, is presented in Fig. 8. The MLP procedure produces the worst outcomes compared to those of the other methods, while the BDA-BMIC system produces the best results. The BDA-BMIC system used a genetic algorithm for feature selection and gradient approximation for classification optimization outcomes. Big data architecture was employed for data storage and retrieval via HDFS. By evaluating the

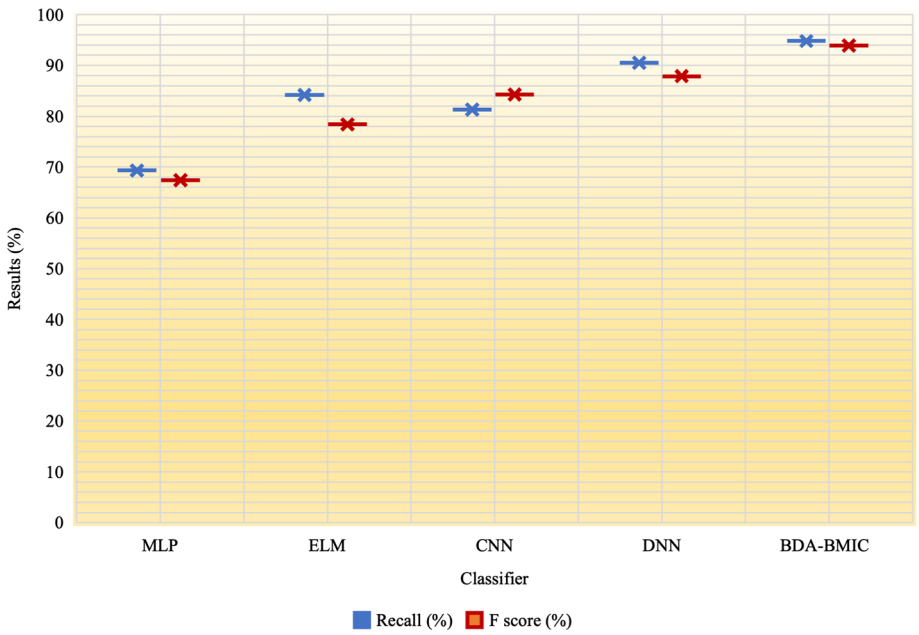


Fig. 7 Recall and F score evaluation of the different classifiers

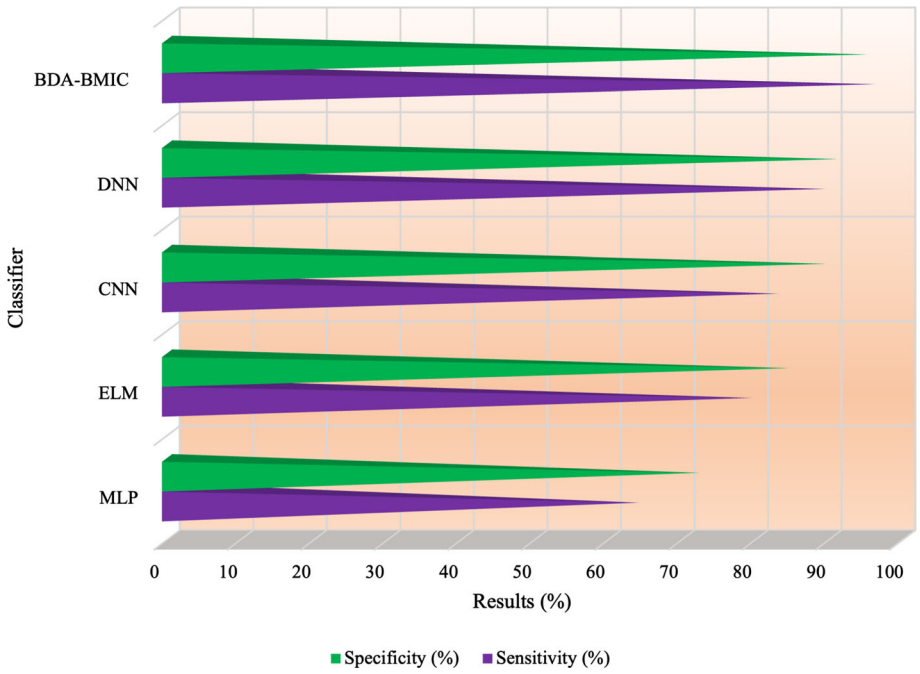


Fig. 8 Specificity and sensitivity analysis of the different classifiers

classifier's results on the test dataset, we can calculate the measures of specificity and sensitivity. Insights into the classifiers' ability to differentiate between positive and negative examples, as well as an evaluation of their overall performance on the classification task, are provided by these metrics.

Many different classifiers' Medical Cost Ratios (MCR), False Discovery Rates (FDR), and True Negative Rates (TNR) are analyzed, and the findings are displayed in Fig. 9. The BDA-BMIC system that has been suggested has a reduced error rate in comparison to other classifiers. The findings ensure that the biomedical image analysis performed on the dataset has the best accuracy possible for disease identification. Big data architecture, evolutionary algorithms for feature selection, and gradient approximation for classification optimization outcomes are used to improve the findings and bring them closer to optimality. The MCR, FDR, and TNR findings for the BDA-BMIC system are, respectively, 4.6%, 4.2%, and 5.3%. The BDA-BMIC system efficiently optimizes classification outcomes by combining the evolutionary algorithm for feature selection with gradient approximation methods. Reduced dimensionality and enhanced classification accuracy are two benefits of using a genetic algorithm to identify the most relevant and discriminative characteristics for the system. By facilitating effective parameter optimization, gradient approximation improves the classification model's overall performance and efficiency.

Scalability, redundancy, speed, locality, and compatibility with the rest of the big data ecosystem are all guaranteed by the BDA-BMIC system's use of HDFS. As a result of these enhancements, large-scale biomedical imaging datasets may now be managed and processed with greater performance, dependability, and adaptability.

5 Conclusion and the findings

A BDA-based Bio-Medical Image Classification (BDA-BMIC) system is intended to identify patient sickness by utilizing metaheuristic optimization (MO) (a genetic algorithm) and gradient approximation (GA) to enhance the biomedical image classification procedure. This study reviews prominent healthcare monitoring systems that use big data. Simultaneously, an overview of contemporary methodologies and technology for processing massive amounts of data has been offered. Later, a big data processing system for the medical business was shown; it is controlling the data generated by different healthcare providers. This paper proposes a strategy for optimizing contrast for biomedical picture improvement based on a unique convolutional kernel and metaheuristic algorithms. The acquired findings are both aesthetically and quantitatively encouraging. This approach may be expanded using different metaheuristic methods that can be used in a variety of situations. The produced biomedical images demonstrate the algorithm's efficiency. Contrast-optimized pictures are less susceptible to distortion and are easier to comprehend. Several computer-aided diagnostic systems will be able to examine biological pictures with greater precision. Comprehensive experiments on publicly available datasets are

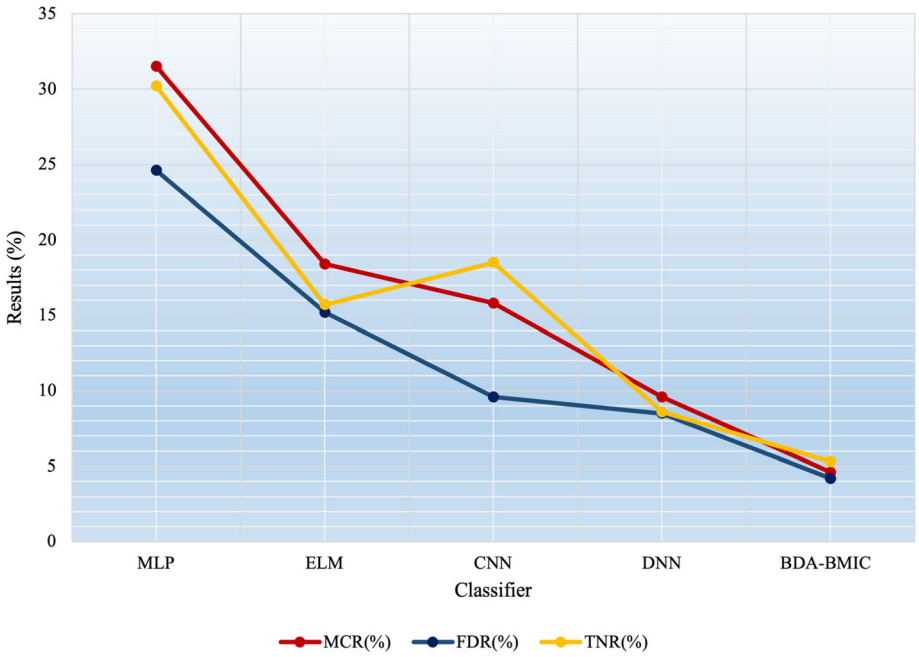


Fig. 9 Experimental result analysis of the different classifiers

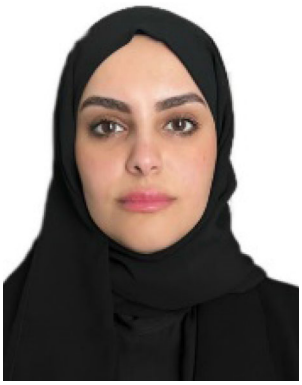
done to show that the proposed retrieval and classification techniques are superior to the present approaches. In the simulation, the recommended BDA-BMIC systems had a mean identification accuracy of 94.6% and a sensitivity of 97.3%. The authors want to construct the conceptual modules and use the expectation-maximization technique to manage missing values in future research.

Acknowledgement The author would like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project Number No. R-2023-830.

References

1. Alnabhan, M., Habboush, A. K., Al-Haija, Q. A., Mohanty, A. K., Pattnaik, S., & Pattanayak, B. K. (2022). Hyper-tuned CNN using EVO technique for efficient biomedical image classification. *Mobile Information Systems, 2022*
2. Habib, G., & Qureshi, S. (2020). Biomedical image classification using CNN by exploiting deep domain transfer learning. *International Journal of Computing and Digital Systems, 10*, 2–11.
3. Haque, I. R. I., & Neubert, J. (2020). Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked, 18*, 100297.
4. Gröhl, J., Schellenberg, M., Dreher, K., & Maier-Hein, L. (2021). Deep learning for biomedical photoacoustic imaging: A review. *Photoacoustics, 22*, 100241.
5. Banerjee, A., Chakraborty, C., Kumar, A., & Biswas, D. (2020). Emerging trends in IoT and big data analytics for biomedical and health care technologies. *Handbook of data science approaches for biomedical engineering*, 121–152.
6. Chen, Y., He, F., Li, H., Zhang, D., & Wu, Y. (2020). A full migration BBO algorithm with enhanced population quality bounds for multimodal biomedical image registration. *Applied Soft Computing, 93*, 106335.
7. Li, C., Chen, G., Zhang, Y., Wu, F., & Wang, Q. (2020). Advanced fluorescence imaging technology in the near-infrared-II window for biomedical applications. *Journal of the American Chemical Society, 142*(35), 14789–14804.
8. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023). MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data, 10*(1), 41.
9. Ben Yedder, H., Cardoen, B., & Hamarneh, G. (2021). Deep learning for biomedical image reconstruction: A survey. *Artificial Intelligence Review, 54*, 215–251.
10. Mansour, R. F., Alfar, N. M., Abdel-Khalek, S., Abdelhaq, M., Saeed, R. A., & Alsaqour, R. (2022). Optimal deep learning-based fusion model for biomedical image classification. *Expert Systems*, *39*(3), e12764.
11. Pang, S., Du, A., Orgun, M. A., & Yu, Z. (2019). A novel fused convolutional neural network for biomedical image classification. *Medical & Biological Engineering & Computing, 57*, 107–121.
12. Assad, M. B., & Kiczales, R. (2020). Deep biomedical image classification using diagonal bilinear interpolation and residual network. *International Journal of Intelligent Networks, 1*, 148–156.
13. Zhang, J., Xie, Y., Wu, Q., & Xia, Y. (2019). Medical image classification using synergic deep learning. *Medical Image Analysis, 54*, 10–19.

14. Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., Rao, X., Zheng, C., & Zhang, F. (2020). Breast cancer histopathological image classification using a hybrid deep neural network. *Methods, 173*, 52–60.
15. Hameed, Z., Zahia, S., Garcia-Zapirain, B., Aguirre, J., & Vanegas, M. (2020). Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors (Basel, Switzerland), 20*(16), 4373.
16. Chen, B., Li, J., Lu, G., Yu, H., & Zhang, D. (2020). Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE Journal of Biomedical and Health Informatics, 24*(8), 2292–2302.
17. Sharma, A., Rani, S., & Gupta, D. (2020). Artificial intelligence-based classification of chest X-ray images into COVID-19 and other infectious diseases. *International Journal of Biomedical Imaging, 2020*, 1–10.
18. Lei, X., Pan, H., & Huang, X. (2019). A dilated CNN model for image classification. *IEEE Access: Practical Innovations, Open Solutions, 7*, 124087–124095.
19. Seo, H., Badiei Khuzani, M., Vasudevan, V., Huang, C., Ren, H., Xiao, R., Jia, X., & Xing, L. (2020). Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. *Medical Physics, 47*(5), e148–e167.
20. Lou, A., Guan, S., & Loew, M. (2023). Cfpnet-m: A lightweight encoder-decoder-based network for multimodal biomedical image real-time segmentation. *Computers in Biology and Medicine, 154*, 106579.
21. Yan, C., Ma, J., Luo, H., & Patel, A. (2019). Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. *Chemometrics and Intelligent Laboratory Systems, 184*, 102–111.
22. Zhou, Y., Yen, G. G., & Yi, Z. (2019). Evolutionary compression of deep neural networks for biomedical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems, 31*(8), 2916–2929.
23. Gupta, T. K., & Raza, K. (2019). Optimization of ANN architecture: A review on nature-inspired techniques. *Machine Learning in bio-signal Analysis and Diagnostic Imaging*, 159–182.
24. Wang, X., Ristaniemi, T., & Cong, F. (2023). Fast learnings of coupled nonnegative Tensor Decomposition using optimal gradient and low-rank approximation. arXiv preprint arXiv:2302.05119.
25. https://www.kaggle.com/datasets/nishanthsalian/genia-biomedical-event-dataset.



Laila Almutairi received the B.S. degree in computer engineering from Kuwait University and the M.S. and Ph.D. degrees in computer and information systems engineering from Tennessee State University. She is currently the Vice Dean of the College of Computer and Information Sciences and an Assistant Professor with the Department of Computer Engineering, Majmaah University, Saudi Arabia. Her research interests include computer networking, network security, cyber security, SDN, and machine and deep learning in cyber security.

Ahed Abugabah is a Professor in Information Systems. He currently works at the College of Technological Innovation at Zayed University. Before joining Zayed University he worked in higher education in Australia where he received his degrees in information systems. His research interests include Information Systems, Applications Machine Learning & Data Mining in Healthcare and healthcare Information Systems.

Hesham Alhumyani received the Ph.D. degree from the University of Connecticut Storrs, USA. He was appointed as the Faculty Dean, in 2019. He is currently working with the Department of Computer Engineering, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia. He has published many research papers in distinctive journals and conferences. His research interests include wireless sensor networks, underwater sensing, the Internet of Things (IoT), and cloud computing.

Ahmed A. Mohamed (aka Ahmed Abo-Bakr Mohamed) is an Assistant Professor at Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Kingdom of Saudi Arabia and on a sabbatical leave from Information Technology Department, Faculty of Computers and Information, Assiut University, Egypt since 2015. He received his B.Sc. degree in 1995 in Electrical Engineering with specialty in Computers and Control, Faculty of Engineering from Assiut University, Assiut, Egypt. In 2006, he was awarded his M.Sc. and Ph.D. degrees from the Computer Science and Engineering Department, School of Engineering, University of Connecticut, USA. He was an Assistant Professor at the Electrical Engineering, Faculty of Engineering, South Valley University between 2007 and 2012. His research interests include artificial intelligence, neural networks, image understanding, data mining, machine learning, and computer vision.