

Performance Driven Outlier Detection in Health-Care Data: A Hybrid Approach Using Dual-Feature Optimization and Segmentation Techniques

Ankita Roy^{1*} , and Atul Garg² 

¹Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India; ankita1001cs.phd20@chitkara.edu.in

²Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India; atul.garg@chitkara.edu.in

*Correspondence: ankita1001cs.phd20@chitkara.edu.in

ABSTRACT- The healthcare sector is a domain where the implementation of human-centered design approaches and concepts can significantly impact well-being and patient care. Delivering superior medical care necessitates a profound comprehension of an individual's desires, encounters, and interests. This study examined the quantitative evaluation and utilization of MRI scans for preoperative conditions of the brain, lungs, and heart. However, identifying these intricate compositions is a formidable challenge. Traditional diagnostic methods are laborious and rely heavily on the clinical expertise of radiologists. This research proposes a non-invasive automatic diagnosis system for diseases utilizing hybrid deep learning approaches, specifically LSTM & PSO (Long Short-Term Memory & Particle Swarm Optimization), to improve the efficiency of outlier detection. Initially, images are initialized to generate standardized images. Next, an improved histogram equalization technique is used to improve the low-contrast MRI images. Finally, segmentation is performed using heat map and contouring techniques based on outlier detection. The findings indicate that the approach developed in this study is capable of classifying distinct groups of brain, lungs, and heart diseases through MRI images. The process involves extracting several physiognomies of the tumor and subsequently selecting the most suitable features through a combination of optimal feature selection techniques, namely PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor). The results demonstrate that the proposed approach exhibits superior productivity and effectiveness, and achieved an accuracy of 98.58%, 98.47%, and 98.58% on brain, lungs, and heart MRI images respectively, offering enhanced reliability and healthcare solution inclusively.

Keywords: Brain tumors, Diseases, Deep learning, Health, Optimization, Outliers.

ARTICLE INFORMATION

Author(s): Ankita Roy, and Atul Garg;

Received: 25/10/2024; **Accepted:** 06/05/2025; **Published:** 30/05/2025;

E- ISSN: 2347-470X;

Paper Id: IJEER 2510-23;

Citation: 10.37391/ijeer.130207

Webpage-link:

<https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130207.html>



Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

The application of computer vision medical care facilitates the exploration of hidden structures within healthcare datasets, which can subsequently be utilized for clinical evaluation. This work involves the detection and extraction of critical, sensitive, and irreplaceable information from brain, lungs, and heart MRI images from extensive databases. The healthcare sector encompasses a significant volume of individual patient information along with a variety of disease findings. Despite their extensive availability at the time, actual health information is marked by heterogeneity, substantial volume, and broad distribution. Consequently, a methodical approach to data collection is necessary. Medical offices are progressively

implementing a digital information control system to efficiently and consistently manage information about patients.

A brain tumor is a malignancy considered by the proliferation of anomalous cells within the tissues of the human brain. Medical images obtained due to the complex nature of the brain's structure and the potential for gliomas to be mistaken for benign headaches, various sources, and modalities often display significant inconsistencies due to differences in imaging techniques, equipment variations, and patient-specific factors [1]. Early detection of gliomas using brain medical images is challenging due to traditional artificial diagnostic approaches for gliomas consistently demonstrate low levels of accuracy and efficiency, requiring radiologists to engage in extensive reading tasks daily, which requires a significant amount of time and can lead to fatigue and inappropriate screening. Nonetheless, the organization of multidisciplinary studies necessitates the integration of modern technology with traditional evaluation for automated diagnostics. Because of this, the healthcare industry depends on technologies like artificial intelligence and imaging for medicine. Diagnostic imaging encompasses the technology employed to generate visual depictions of the body's inside. In recent years, it has been extensively utilized in the healthcare sector. It plays a crucial role in contemporary medicine and is utilized in nearly every facet of patient care, including diagnosis, treatment, and surgery. It assists clinicians in accurately identifying and delineating illness progressions [2].

Cardiovascular disease is the leading cause of death across all age demographics in modern civilization. Therefore, it is essential to improve the criteria for predicting heart attacks through diverse DL techniques in clinical settings. The clinical diagnosis of cardiac disease mostly depends on existing information from associated pathogenic occurrences [3]. Lung illnesses are medically categorized ailments that compromise lung functionality. The pathological condition of the lung is generally associated with distinct indications and symptoms. An inherent dysfunction of the lungs accelerates the advancement of the diseases. Conventional diagnostics methods emphasize manual symptom assessment for diagnosing pulmonary diseases, with doctors determining subsequent treatment options based on evaluated disease characteristics.

Magnetic Resonance Imaging (MRI) is employed to capture cardiac images and is a widely utilized and dependable technique for assessing hypokinesia to formulate a computational approach for recognizing medical disorders through magnetic resonance images. The fundamental idea of MRI involves generating a strong magnetic field, utilizing radiofrequency pulses, and employing a computer system to visualize images of organs, tissues, bones, and other internal structures, so enabling physicians to diagnose and identify specific diseases. Similar to X-rays, MRI operates without the use of ionizing radiation. In instances of global hypokinesia, cardiac MRI utilizes a potent electromagnetic field, electromagnetic radiation, and a computer program to provide a comprehensive image of the heart. MRI is a safe and dependable technique for diagnosing cardiac disease, assessing heart shape, function, and internal examination of the heart. Many individuals succumb to heart attacks, necessitating the advancement of computer methods to anticipate the disease using MRI data obtained during patient scans. Consequently, our methodology has devised a computational technique that predicts disease from cardiac MRI images.

Heterogeneous statistics marked by numerous properties with null or absent values, and incoherent information originating from multiple avenues, are prevalent characteristics of real-life data sets. A multitude of data preparation techniques has been established to tackle issues such as chaotic information, regaining of lost and inconsistent information, and the removal of redundant characteristics. These tactics seek to augment the dataset's quality and optimize the efficiency of ensuing processes. To reduce the manpower needed and improve patient longevity, implementing an automated system is crucial.

Feature extraction is a crucial aspect of medical image analysis, involving a complex procedure that identifies meaningful patterns and distinct characteristics from sophisticated medical images. In this field, it is essential to extract robust and accurate features from high-dimensional image data using modern imaging techniques to ensure proper analysis, its diagnosis as well as prognosis [4]. To enhance precision, preprocessing and segmentation techniques have been implemented [5]. This work identifies four distinct categories of brain tumors that can be recognized. This research introduces a hybrid model that combines PCA, PSO, and LSTM networks to detect outliers in multi-class brain tumor data from MRI scans. This work

employs histogram equalization to improve the local contrast of a specified location followed by CLAHE (Contrast Limited Adaptive Histogram Equalization) technique which is a contrast restriction method to avoid excessive amplification of noise and artifacts present in MRI images of the patients. The analysis also includes the evaluation of the impact of different segmentation techniques, such as heatmap and contouring, for the detection of outliers in MRI images. This work further aims to compare the reliability, convergence, and performance of the PSO-LSTM approach with a feature selection technique in addressing large-scale healthcare challenges.

The workflow is as follows: *section 2* describes the Literature Review followed by the Methodology in *section 3*, *section 4* describes the Proposed Work, *section 5* discusses the Results and Discussion followed by the Conclusion and Future Scope in *section 6*.

2. LITERATURE REVIEW

It has been discovered that different approaches use different kinds of data sets to develop robust and efficient techniques for the diagnosis of disease and understanding the contributory aspect, this literature review gives an overview of the recent development in this area, emphasizing different approaches to data sources and finding that collectively contribute to the largest scope of this significant research area.

By applying image augmentation techniques, a deep CNN model for classifying brain tumors is developed; achieving an overall accuracy of 98.09% and 96.35% on the training and validation sets [2]. However, the lengthy training period without the use of transfer learning was an inconvenience, making the technique less realistic for large amounts of data. For investigating medical images, feature extraction is the crucial step of a complex prediction system that recognizes and distinguishes medical images. To develop a robust image processing technique, the extraction of the most significant and relevant features from the MRI images is required for effective evaluation and diagnosis of brain tumor diseases [6]. Different statistical approaches like morphological operations, LBP, wavelet transformation, frequency and spatial domain transformation as well as deep learning approaches are required for feature extraction [4]. Feature optimization plays a significant role in improving the quality of the feature extraction process from the MRI images. These statistical techniques entail the extraction of the best feature space vectors for the prediction system [3].

Utilization of embedding and wrapper techniques can allow the fast exploration of feature vector space by making it easier to recognize discriminative features that influence the diagnosis process accurately. Other techniques like PCA and LDA can also be used for dimensionality reductions of the features extracted from different machine learning techniques to address the issues of high dimensionality inherited from MRI image data sets [3]. Irrespective of the substantial development of deep learning approaches for the automatic detection of outliers, different challenges need to be addressed in the field of healthcare. It is significantly important to address the various

limitations that influence the performance of disease diagnosis systems. By addressing these shortcomings, the development of an accurate and reliable outlier's diagnosis system in the field of health care is assured. To discourse these research areas that have not been sufficiently discovered the present work developed a new innovative system by using hybrid deep learning and optimization techniques. The major objective of this work is to progress an inclusive context that can effectively recognize outliers in MRI images of the brain, lungs, and heart diseases. To accomplish this task the proposed approach influences the unification of deep learning and optimization techniques to analyze MRI images of the patient. Specifically, the hybrid method enables a better-organized recognition system to improve predictive analysis of the outliers present in MRI images of the patients. After analyzing the literature and gaps, the following contributions have been incorporated in the work;

- This Work employs histogram equalization to improve the local contrast of a specified location followed by CLAHE technique which is a contrast restriction method to avoid excessive amplification of noise and artifacts present in MRI images of the heart, lungs, and brain diseases. The analysis also includes the evaluation of the impact of different segmentation techniques, such as heatmap and contouring, on the detection of outliers in MRI images.
- Statistical technique like PCA and t-SNE is used in this article to perform dimensionality reduction by transforming the feature vector space into most relevant and optimized feature space. These techniques represent the complex structure of the data by uncovering patterns and clusters that are not easily visible in high-dimensionality feature space.
- For the adjustment of hyper-parameters and optimization of the deep learning model, a hybrid model using PSO and LSTM techniques followed by the implementation of dimensionality reduction techniques is utilized for sequential data modeling to address the data complexity and improve the classification outcomes of different classes of brain tumor.
- This work further aims to compare the reliability, convergence, and performance of the PSO-LSTM approach with a feature selection technique in addressing large-scale healthcare challenges.

3. METHODOLOGY

The prospects of different diseases are dependent upon the specific type and magnitude of the severity. Machine learning exhibits immense potential for this endeavor and is capable of effectively categorizing various tumor locations with a high level of accuracy within a brief timeframe. Nevertheless, the process of training a model can be time-consuming and resource-intensive. Optimal feature selection is an effective strategy for minimizing the training duration of a model when resources are constrained. This article employed hybrid optimization techniques to classify different disease locations present in MRI images. The sequence of steps includes employing histogram equalization followed by CLAHE which is used for restricting contrast and then the segmentation

techniques such as heatmaps and contouring are employed. For reducing the number of dimensions PCA and t-SNE are used with the hybrid LSTM+PSO Model. The flow diagram of the proposed methodology is demonstrated in *figure 1*.

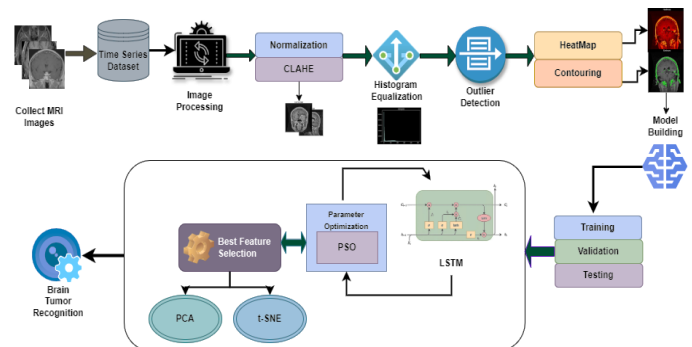


Figure 1. Work Flow of Proposed Model

3.1. Dataset

The data obtained from Kaggle [7-9] was utilized to generate predictions for cardiac MRI scans. The collection includes pictures of thirty individuals, with each category containing unique groupings of cardiac scans. The categories comprise 13,540 photos of healthy subjects and 22,855 images depicting hyperkinesias [7]. A brain tumor is a conglomeration of proliferative cells within the cerebral cortex. The dataset of the publication consists of 1,311 MRI images of the human brain, divided into four categories: glioma with 300 images, meningioma with 306, no tumor with 405, and pituitary with 300 samples [8]. Out of 1,181 photographs, 944 were designated for training and 237 for validation, following an 80/20 split. A subset of 130 photographs was selected from a total of 1,311 for testing purposes. The lung cancer dataset from the National Center for Cancer Diseases was collected at the aforementioned specialty hospitals over three months in 2019. The dataset includes CT scans of individuals diagnosed with lung cancer at different stages, as well as individuals without the condition. The collection has 1,190 images that precisely illustrate CT scans from 110 distinct instances [9]. The provided examples are classified into three distinct categories: normal, benign, and malignant. Of the overall occurrences, 40 are classified as malignant, 15 as benign, and 55 as normal. *Figure 2* illustrates MRI scans acquired from multiple sources.

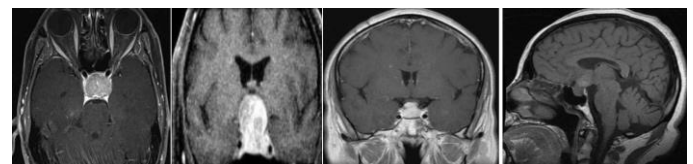


Figure 2. MRI Images Collected

3.2. Data-Augmentation

In deep learning, the availability of a large amount of data is a prerequisite, so data augmentation is employed to produce a significant number of trainable datasets from a restricted sample size. Overfitting occurs when a limited number of samples are used as training data. Data augmentation is the optimal technique

for mitigating overfitting in application to image data. Multiple data augmentation techniques are employed to augment the number of pictures in MRI data sets. It enhances the capability of deep learning models by using data augmentation features. Hence, seven factors have been included to augment the training and testing phases of the MRI datasets. As represented in *figure 3* the outcome of the augmentation function of Random Flip (Horizontal), Random Translation (HeightFactor=0.1, WidthFactor=0.1), Random Contrast (+50%), Random Rotation (200 and 300), Random Zoom (HeightFactor=0.3, WidthFactor=0.3), and brightness (+50%) are the standards provided [10].

3.3. Histogram Equalization & CLAHE

In this study, the pre-processing technique acknowledged as CLAHE has been applied. CLAHE includes a mechanism to limit contrast, preventing excessive noise and artifact amplification [11]. This ensures that the enhanced image retains the local statistical properties of the original image, thus preserving essential information, as shown in *figure 4*. The histogram equalization curve and the MRI image sample are illustrated in *figure 5*. A histogram is generated to represent the frequency distribution of brain tumors, lungs, and cardiovascular diseases MRI data. MRI images are processed to extract relevant intensity values, which are then plotted to visualize their distribution [12]. Histograms are valuable for understanding the contrast and intensity distribution in MRI images, aiding in the differentiation between healthy tissue and tumor regions.

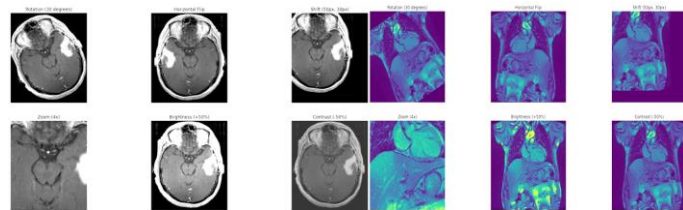


Figure 3. Augmented Images

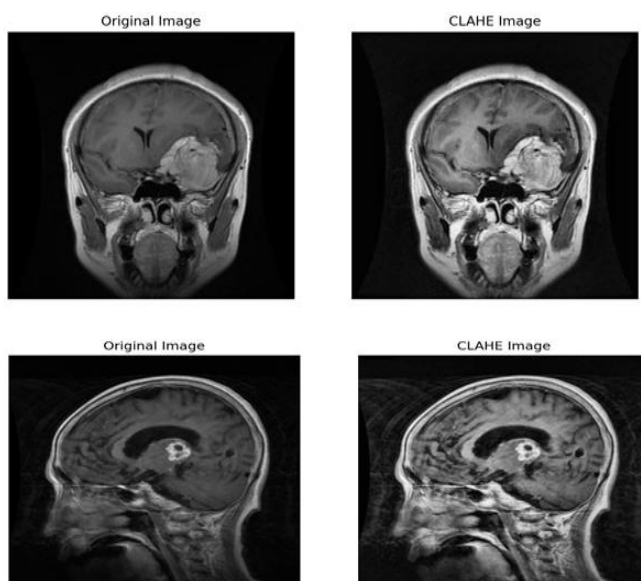


Figure 4. Pre-Preprocessed Images using CLAHE

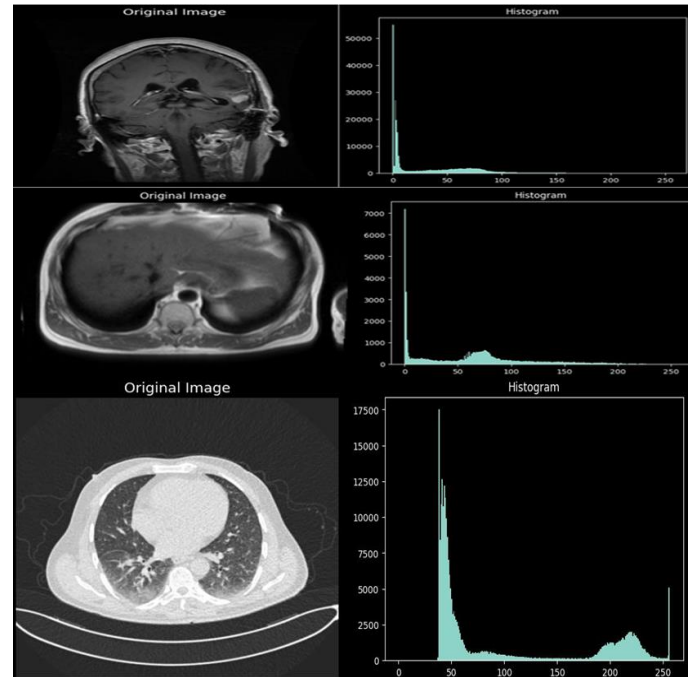


Figure 5. Histogram Equalization Obtained from MRI Image Sample

3.4. Segmentation

Active contours are widely used in image processing to define seamless shapes within images and generate enclosed contours for specific regions. Their primary purpose is to detect irregular structures, such as those found in CT and MRI scans. Active contours are methods used to extract deformable models or structures from an image by applying restraints and services for segmentation, resulting in a parametric curve or contour. Contours are essential for analyzing shapes and detecting and recognizing objects [13]. In MRI analysis, heatmaps are used to highlight significant regions, such as tumors, lesions, or other anomalies [14]. These visual representations help understand the model's decision-making process, aiding in clinical diagnosis and research, as shown in *figure 6*.

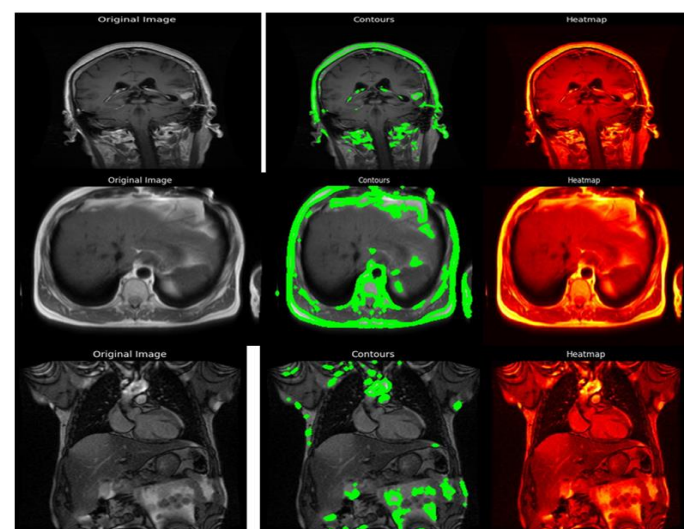


Figure 6. Heatmap and Contouring Image Generated

4. PROPOSED WORK

The brain MRI images were categorized into four classes and then divided into training, validation, and testing subsets. The images undergo intensity normalization using the contrast-constrained adaptive histogram equalization approach. The MRI images are further segmented using contouring and heatmap techniques. The hybrid model employs LSTM, PSO, t-SNE, and PCA approaches. PCA was employed for feature optimization. The chosen characteristics were further enhanced through the utilization of the PCA and t-SNE techniques to decrease dimensionality, eliminate noise, and enhance the effectiveness and efficiency of subsequent algorithms. The workflow of the suggested technique is described below.

4.1. Optimal Feature Selection

As datasets become larger and more diverse, it is crucial to decrease their size to accommodate system memory for subsequent research. Feature selection strategies aim to decrease the quantity of input variables by eradicating irrelevant or redundant features. Insignificant, superfluous, and duplicate features can impede an algorithm, resulting in decreased learning performance, accuracy, and computing cost.

4.1.1. Principal Component Analysis

PCA is used as an effective means for extracting the most prominent characteristics from datasets with a large number of dimensions, thereby uncovering the basic trends and associations that are innate to the data. PCA does this by converting a group of variables that are correlated with each other into an entirely new set of variables that are not correlated, which are called principal components. The linear combinations of the original variables are organized in declining order of relevance [15]. The main goal of PCA is to internment the uppermost amount of variance in the data using the first principal component, while the following components capture progressively less quantities of variance. By expressing the data about these primary components, it becomes feasible to efficiently decrease the dataset's complexity while preserving the most important information.

4.1.2. t-Distributed Stochastic Neighbor

t-SNE embedding is a method used to transform data with high dimensions into an easily understandable format, usually in 2D or 3D, to enhance interpretation. Stochastic Neighbor Embedding begins by transforming the Euclidean distances, which measure distances between each data point in a high-dimensional space, into conditioned likelihoods that indicate the similarities between them [16]. The degree of similarity between datapoint x_i and data-point x_j can be measured by the conditional probability, $p_{j|i}$. The conditional probability $p_{j|i}$ can be stated mathematically as follows in equation (1):

$$\dot{p}_{y|x} = \frac{\exp\left(-\frac{\|z_x - z_y\|^2}{2\sigma_i^2}\right)}{\sum_{x \neq y} \exp\left(-\frac{\|z_x - z_y\|^2}{2\sigma_i^2}\right)} \quad (1)$$

It is feasible to calculate a conditional probability in equation (2), denoted by $\rho_{y|x}$, for the equivalent low-order dimensional value t_x and t_y of the higher order dimensions values z_x and z_y .

$$\rho_{y|x} = \frac{\exp(-\|t_x - t_y\|^2)}{\sum_{x \neq y} \exp(-\|t_x - t_y\|^2)} \quad (2)$$

It should be noted that $\dot{\rho}_{y|x}$ and $\rho_{y|x}$ are assigned a value of zero, as our goal is solely to represent bilateral similarities.

SNE uses a gradient descent method to reduce the total divergences for all data points, which measures the minimization of the sum of differences of conditional likelihoods. Absolute entropy, also known as Kullback-Leibler divergence [17], quantifies the additional length of the message per data point that needs to be transmitted when using an optimal code designed for a different (incorrect) distribution Q , as opposed to employing an optimal code based on the actual distribution P which represents the additional entropy in equation (3).

$$D_{KL}(M||N) = \sum_{z \in Z} \dot{\rho}(z) \log \frac{1}{\dot{\rho}(z)} - \sum_{x \in X} \dot{\rho}(z) \log \frac{1}{\dot{\rho}(z)} = \tilde{H}(M, N) - \tilde{H}(M) \quad (3)$$

The cross-entropy of M and N , denoted as $\tilde{H}(M, N)$, is a degree of the difference between the probability distributions of likelihood M and N . The entropy of P , denoted as $H(P)$, is a measure of the uncertainty or randomness in the distribution M .

The entropy of this distribution increases as the standard deviation (σ_i) upsurges. t-SNE utilizes a binary search algorithm to find the magnitude of σ_i that generates a $\dot{\rho}_{z_i}$ with a predetermined perplexity, as given by the user. Perplexity is a measure that is used to quantify the level of uncertainty or confusion in a given situation or problem as shown in equations (4) & (5).

$$\text{Perpl}(\dot{\rho}_z) = 2^{\tilde{H}(\dot{\rho}_i)} \quad (4)$$

The term $\tilde{H}(\dot{\rho}_z)$ is the Shannon entropy of P_i expressed in bits.

$$\tilde{H}(\dot{\rho}_z) = - \sum_j \dot{\rho}_{j|i} \log_2 \dot{\rho}_{j|i} \quad (5)$$

Perplexity can be understood as a continuous metric that quantifies the effective quantity of neighboring points. t-SNE, a non-linear reduction of dimension strategy, identifies observable clusters in the data by analyzing similarities of data points with several attributes [18].

Figure 7 depicts the outcome of feature selection using PCA and t-SNE on an MRI image.

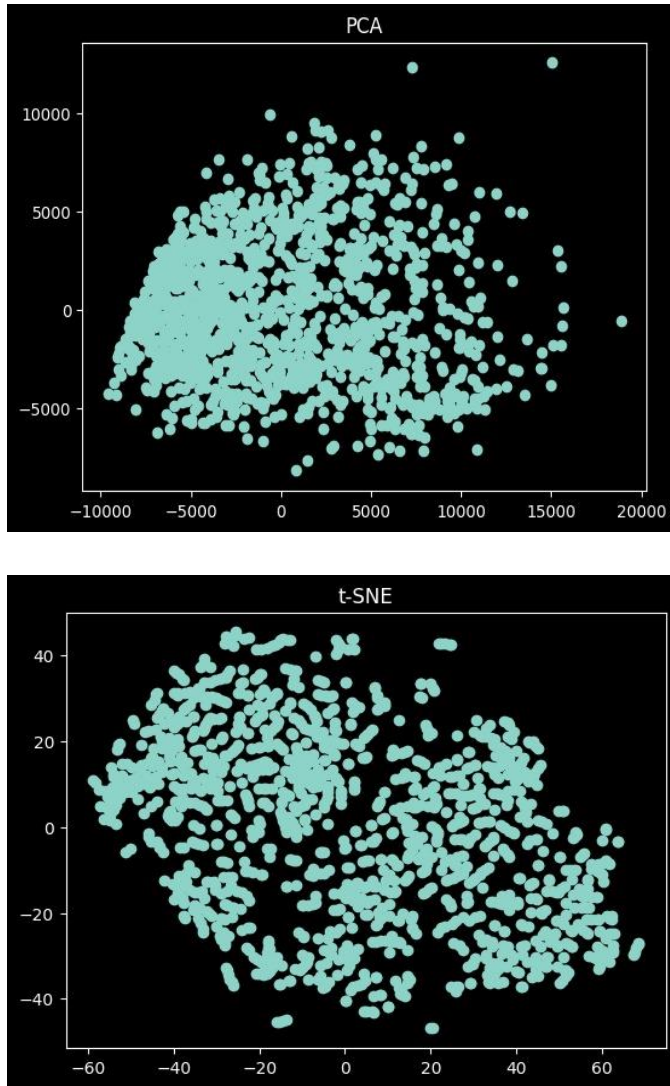


Figure 7. Feature Selection from PCA and t-SNE on MRI image

4.1.3. Particle Swarm Optimization

The utilization of a particle swarm approach to improve non-linear behaviors by drawing inspiration from the migration patterns of birds and consumption patterns [19]. The approach considers individuals as particles inside a search space that has several dimensions. The PSO algorithm is initialized with random parameters and subsequently adjusted after every iteration to determine the optimal selection. Every possible solution, referred to as a particle, is depicted as a point in the multi-dimensional solution space. Particles move towards the predefined velocity for optimal solution and PSO unifies the local as well as global search approaches while balancing the exploration. M-Dimensional space particles X belong to the $\check{D}$, group which includes d_1, d_2, d_k elements representing the $dx = [dx_1, dx_2, dx_N]$ vector space by each element. The following qualities currently belong to the x th particle from equation (6) – (9):

$$\check{D}_x = (d_{x1}^t, d_{x2}^t, \dots, d_{xN}^t)^T \quad (6)$$

$$V_x = (v_{x1}^t, v_{x2}^t, \dots, v_{xN}^t)^T \quad (7)$$

$$P_i = (p_{x1}^t, p_{x2}^t, \dots, p_{xN}^t)^T \quad (8)$$

$$P_g = (p_{g1}^t, p_{g2}^t, \dots, p_{gN}^t)^T \quad (9)$$

d_x^t is the current position, v_x^t current velocity and p_{xt} the best position in the Historical events of the particles and shows the ideal location of the entire population of swarms duration as shown in equation (6) – (7). The position and the velocity change with each iteration based on the equation (10) and (11).

$$v_x^{t+1} = wv_x^t + \alpha_1\beta_1^t(p_x^t - d_x^t) + \alpha_2\beta_2^t(p_g^t - d_x^t) \quad (10)$$

$$d_x^{t+1} = d_x^t + v_x^{t+1} \quad (11)$$

v_x^{t+1} Is the updated velocity and d_x^{t+1} is updated position after each iteration. α_1, α_2 are the constant and β_1, β_2 arbitrary number of the range 0-1.

4.1.4. LSTM Algorithm

LSTM consists of different gates like forget gate memory cell and output gate [20]. These gates regulate the storage space, information, and transition states of the cell. The mathematical expression of these units is expressed in equations (12) – (18).

$$\text{Forget gate: } \varphi_f = \sigma(\omega_f[\chi_t, \lambda_{t-1}] + \beta_f) \quad (12)$$

$$\text{Input gate: } \varphi_i = \sigma(\omega_i[\chi_t, \lambda_{t-1}] + \beta_i) \quad (13)$$

$$\text{Output gate: } \varphi_o = \sigma(\omega_o[\chi_t, \lambda_{t-1}] + \beta_o) \quad (14)$$

$$\text{Cell state: } \zeta_t = \tanh(\omega_c[\chi_t, \lambda_{t-1}] + \beta_c) \quad (15)$$

$$\zeta_t = \varphi_f * \varphi_i + \varphi_o * \zeta_{t-1} \quad (16)$$

$$\text{Output value: } \lambda_t = \varphi_o * \zeta_t \quad (17)$$

$$\text{Where, } \sigma(\chi) = \frac{1}{1+e^{-\chi}}, \tanh = \frac{1-e^{-2\chi}}{1+e^{-2\chi}} \quad (18)$$

The hyperbolic tangent function is in the cell state as activation values whereas, sigmoid function is used as activation function of the gate units of LSTM which controls the amount of data obtained. The current time χ_t input and the previous output λ_{t-1} , update the cells state while transitioning from $t-1$ to t time. Using Gate φ_o and updated cell state ζ_t , output λ_t is computed.

4.2. Hybrid Optimization Technique

As datasets become larger and more diverse, it is crucial to decrease their size in order to accommodate in system memory for subsequent research. Feature selection strategies aim to decrease the quantity of input variables by eradicating irrelevant or redundant features. Insignificant, superfluous, and duplicate features can impede an algorithm, resulting in decreased learning performance, accuracy, and computing cost.

4.2.1. Principal Component Analysis

PCA is used as an effective means for extracting the most prominent characteristics from datasets with a large number of dimensions, thereby uncovering the basic trends and associations that are innate to the data. PCA does this by converting a group of variables that are correlated with each other into an entirely new set of variables that are not correlated, which are called principal components. The linear combinations of the original variables are organized in declining order of relevance [15]. The main goal of PCA is to internment the uppermost amount of variance in the data using the first principal component, while the following components capture progressively fewer quantities of variance. By expressing the data about these primary components, it becomes feasible to efficiently decrease the dataset's complexity while preserving the most important information.

4.2.2. t-Distributed Stochastic Neighbor

The arrangement of PCA/t-NSE and PSO with LSTM provides the benefit of both approaches to discover complex data evaluation and optimizing challenges. The working of the proposed approach is depicted in *algorithm 1* which takes input parameters as the new fitness function and Prev fitness as the entire population. Additionally, the algorithm requires the total number of particles as input, denoted as 'Max'. Upon full execution, the algorithm yields the chosen features as its output. The purpose of the fitness function is to assess the level of fitness of the particles. For each iteration, the algorithm continuously compares the fitness of the current particle with the fit particle in the entire population until the optimal characteristics are identified. The position and velocity values are updated based on *equations (10) and (11)*. The entire procedure is iterated until the most desirable characteristics have been attained. *Figure 8* is the flowchart that outlines the primary methodology employed in this research.

Proposed Algorithm

Input: Set of MRI images and segmented (training dataset), Fitness function as $new_{fitness}$, Total population $Prev_{fitness}$, max is the total particle

Output: Subset of optimal and relevant features and outlier detection

1. Split Train and Test Set
2. Compute the node count LSTM network layer
3. Initial particles $\vec{D}_x$ and V_x randomly chosen
4. Compute the log loss $= -(1/D) \times \sum [\sum (g_{xy} * \log(\hat{p}_{xy}))]$, fitness value
5. While (! Optimal_features) do
6. for (iteration = 1 to max) do
7. if $new_{fitness} > Prev_{fitness}$ then
8. Update $Prev_{fitness} = new_{fitness}$
9. if fitness $new_{fitness} > Max_{fittest}$ then
10. Update $Max_{fittest} = new_{fitness}$
11. end if

12. end if
13. Update velocity $v_x^{t+1} = wv_x^t + \alpha_1\beta_1^t(p_x^t - d_x^t) + \alpha_2\beta_2^t(p_g^t - d_x^t)$
14. Update position $d_x^{t+1} = d_x^t + v_x^{t+1}$
15. Next particle in the population
16. end for
17. Obtain particular extreme and global extremum.
18. Repeat until convergence
19. Perform PCA ($Feature_{best}$)/t-SNE($Feature_{best}$)
20. Contemplate weights as global best parameters
21. end while

4.2.3. Fitness Function

It measures the performance of ML models in carrying out their goals. Common fitness is cross entropy loss, log loss, and hinge loss. In this work, log loss is used as the fitness function also referred to as cross-entropy loss [21]. The objective is to minimize this disparity, which signifies a more accurate alignment between the projected and real results. The formula for log loss is as follows in equation (19):

$$\log \text{ loss} = -(1/D) \times \sum [\sum (g_{xy} * \log(\hat{p}_{xy}))] \quad (19)$$

The function g_{xy} is an indicative expression that takes the value 1 if the value x belongs to the y , and 0 otherwise. The variable $\hat{p}_{xy}$ represents the estimated possibility that x belongs to class y . The model seeks to enhance its predictions and maximize the class probabilities for each input by reducing the log loss.

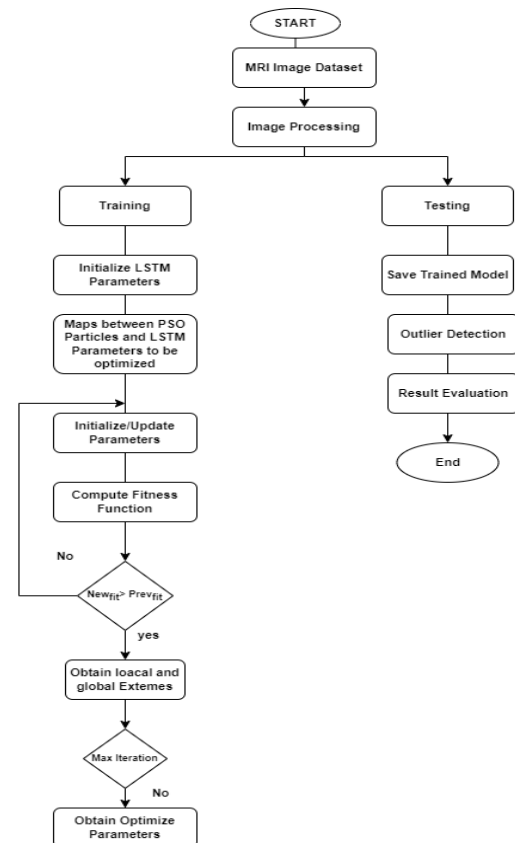


Figure 8. Flowchart of Hybrid LSTM+PSO Model

5. RESULT AND DISCUSSION

Python 3.9 is used as the language used for programming and employed Python for implementing the particle swarm optimization algorithm. Additionally, Tensor-flow is utilized for training and testing the model. In this experiment, batch sizes of either 64 or 512 are used, time steps ranging from 4 to 30, and a variable number of cells between 64 and 512 for the particle swarm algorithm. In addition, learning rate of 0.0001 is specified, and the activation function 'tanh' to construct the model. The number of epochs is set to 30 and implemented early stopping to prevent overfitting of the system. The confusion matrix for the proposed technique is presented in *table 1*. The accuracy for glioma, meningioma, healthy, and pituitary images on the test set is 95.33%, 92.67%, 96.67%, and 98% respectively. Additional performance measures such as recall, F1-score, and precision for the four classes of brain tumors are shown in *table 2*. These results demonstrate that the proposed method performs exceptionally well.

Table 1. Matrix Generated from Test Set

Class	Glioma	Meningioma	No_tumor	Pituitary
Glioma	27	3	1	0
Meningioma	0	24	2	1
No_tumor	2	2	57	0
Pituitary	1	1	0	29

Table 2. Evaluation Parameters of Four Classes of Brain Tumors

Class	Accuracy	Recall	Precision	F1-Score
Glioma	95.33	90	87	89
Meningioma	92.67	80	83	81
No_tumor	96.67	95	97	96
Pituitary	98	97	94	95

The confusion matrix for the proposed technique on different disease types of MRI dataset is presented in *table 3*. The

accuracy for brain, lungs, and heart MRI images on the test set is 98.58%, 98.47%, and 98.58%, respectively. Additional performance measures such as recall, F1-score, and precision for the three classes of disease are shown in *table 4*. These results demonstrate that the proposed method performs exceptionally well using ADAM optimizer.

Table 3. Confusion Matrix of Three Disease Type

Disease Type	Brain	Lungs	Heart
Brain	288	3	3
Lungs	4	295	2
Heart	3	5	311

Table 4. Evaluation Parameters of Three Disease Type

Disease Type	Precision (%)	Accuracy (%)	Recall (%)	F1-Score (%)
Brain	98.58	98	98	98
Lungs	98.47	98	97	98
Heart	98.58	97	98	98

The performance of the proposed model was assessed utilizing the relu activation functions, together with the ADAM optimizer. Graphs are generated for the given MRI dataset of three different diseases at epoch 17. *Figure 9* represents the loss and accuracy curve obtained from training and validation using the proposed model with the Adam optimizer applied to MRI image samples of brain, lungs, and heart diseases. *Figure 10* represents the loss and accuracy curve obtained from training and validation using the proposed model with the Adam optimizer applied to MRI image samples of lung diseases. *Figure 11* represents the loss and accuracy curve obtained from training and validation using the proposed model with the Adam optimizer applied to MRI image samples of heart diseases.

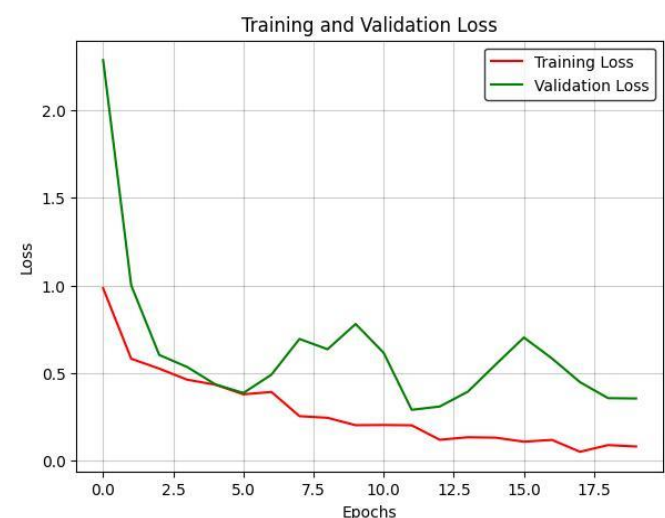


Figure 9. Accuracy and Loss Curves for Brain Diseases

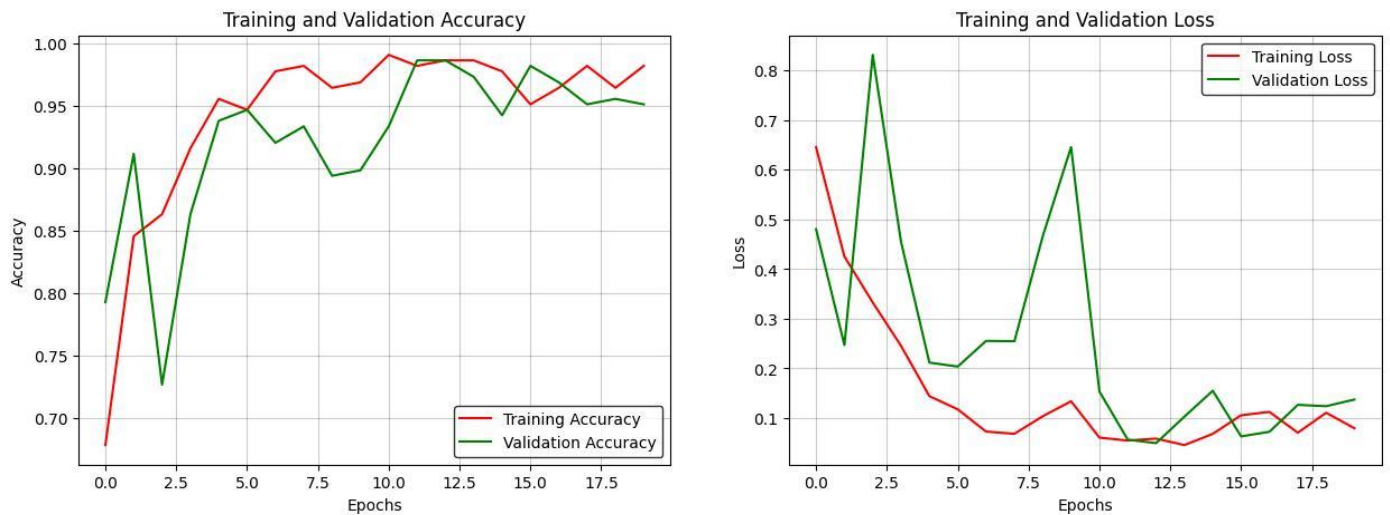


Figure 10. Accuracy and Loss Curves for Lung Diseases

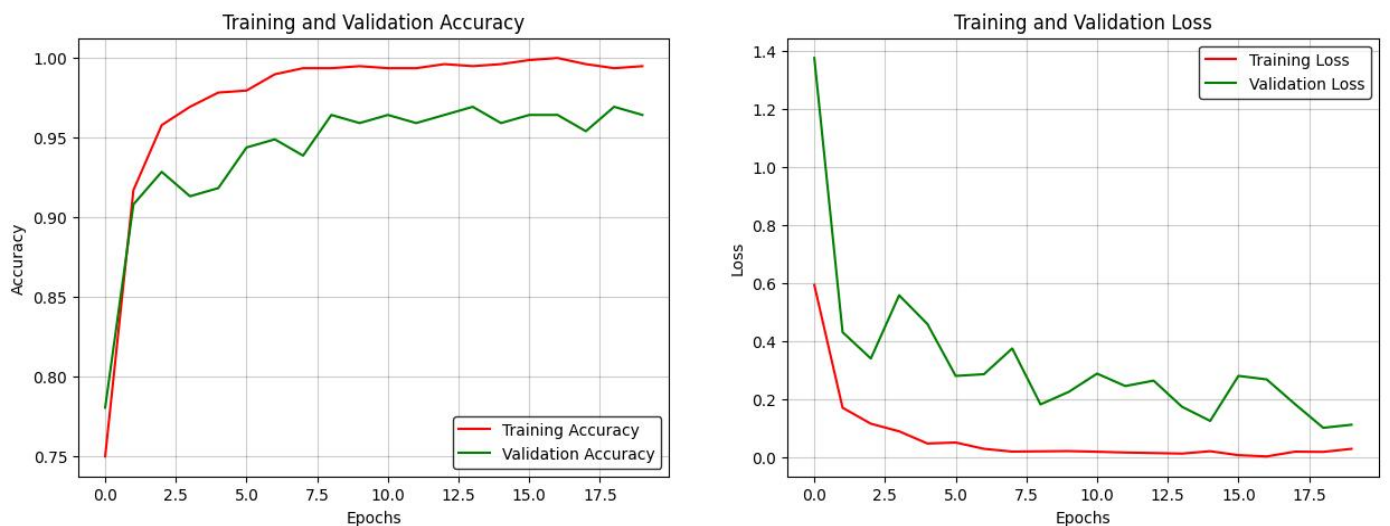


Figure 11. Accuracy and Loss Curves for Heart Diseases

5.1. Statistical Analysis

The different statistical studies were made to thoroughly approve the outcome of the proposed model. Over 30 distinct runs, a paired t-test was used to assess the mean accuracy of the proposed model to baseline methods. The p-values authenticate statistically significant increases in accuracy. Differences in mean accuracy across all four models—LSTM, LSTM+PSO, LSTM+PSO+PCA, LSTM+PSO+t-SNE—were assessed using a one-way ANOVA test. Accuracy measures were considered using 95% CIs. The non-overlapping CIs support even more the dominance of the suggested model. Effect sizes measured the degree of improvement. Applied to paired predictions—Proposed vs. LSTM+PSO—on the test set using Chi-square=12.6, $p=0.0004$. The suggested model significantly reduced misclassifications. *Table 8* illustrates statistical analysis of the proposed model using different statistical test.

Table 5. Statistical Analysis of the Proposed Model

Test	Metric	Proposed Model	Baseline (LSTM)	Significance
t-Test	Accuracy	97.68%	85.42%	$P<0.001$
Anova	F-Stat	42.15	--	$P<0.001$
Confidence Interval	Accuracy Range	97.36-98.00	85.67-87-87.17	Non-overlapping
Cohen's d Effect Size	Effective Size	2.14	--	Large
Chi Square	Misclassification	12.6	--	$P=0.0004$

5.2. Comparative Investigation

The effectiveness of the suggested LSTM+PSO algorithm is compared by adding the two other feature selection algorithms, namely PCA and t-SNE. This comparison aimed to reveal the effectiveness of the recommended algorithm concerning diverse feature selection methods. Multiple evaluation metrics were incorporated to gauge and compare the performance of these algorithms to others. The results presented in *table 6* indicate that the algorithm proposed in this work exhibits significant advantages across three measures.

Table 6. Comparative Analysis

Model	Train (Acc)	Train (Loss)	Val (Acc)	Val (Loss)
LSTM	86.42	0.0361	88.93	0.0312
LSTM+PSO	89.63	0.0213	90.16	0.0211
LSTM+PSO+PCA	96.21	0.0198	93.62	0.0201
LSTM+PSO_t-SNE	97.68	0.0193	96.89	0.0199

The PSO-LSTM model's performance was validated using human MRI images that were categorized into four groups. Throughout the experiment, the PSO-LSTM model incorporating t-NSE provided a maximum accuracy of 97.68%. The analysis suggests that the PSO+LSTM algorithm, when combined with PCA, demonstrates satisfactory performance and increased efficiency, achieving an accuracy of 96.21%. *Figure 12* demonstrates the improved outcomes of PSO & LSTM, specifically in optimal feature selection. The efficiency of the suggested method is depicted in the figure, which showcases the utilization of various evaluation parameters such as specificity, recall, sensitivity, precision, fi-score, and accuracy.

The accuracy of PSO-LSTM is higher than that of LSTM in discriminating between different techniques, as demonstrated in *figure 8*. LSTM has an accuracy of 86.42%, sensitivity of 89.66%, specificity of 85.71%, precision of 86.67%, and fi-score of 88.14%. The LSTM+PSO model, without utilizing optimal feature selection strategies, achieves an accuracy of 89.63%, a sensitivity of 92.31%, a specificity of 85.71%, a precision of 90%, and a fi-score of 91.14%. The LSTM+PSO algorithm, combined with PCA for feature selection, outperformed both the LSTM algorithm and the LSTM+PSO algorithm. It attained an accuracy of 96.21%, a sensitivity of 97.62%, a specificity of 94.12%, a precision of 95.35%, and a fi-score of 96.47%. Finally, the combination of LSTM+PSO with t-SNE yielded the best results among the other algorithms, with an accuracy of 97.68%, sensitivity of 98.08%, specificity of 95.45%, precision of 96.23%, and fi-score of 97.14%. *Figure 13* displays the accuracy and loss curves of four methods on the test set.

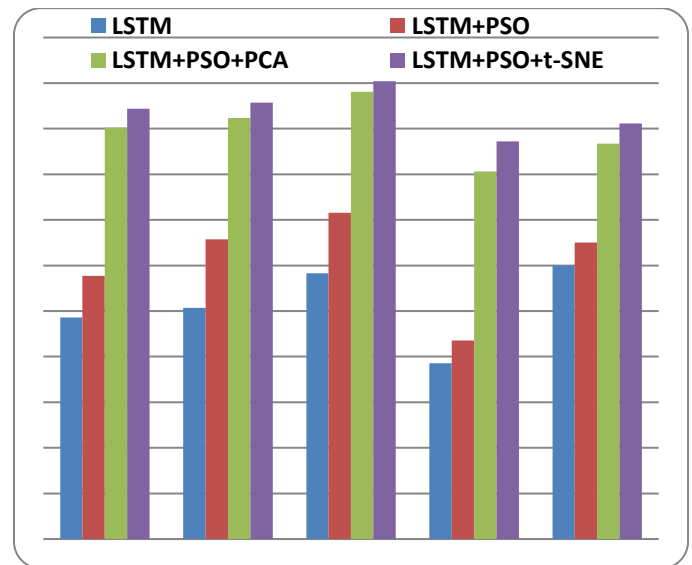


Figure 12. Performance Evaluation using Different Evaluation Metrics

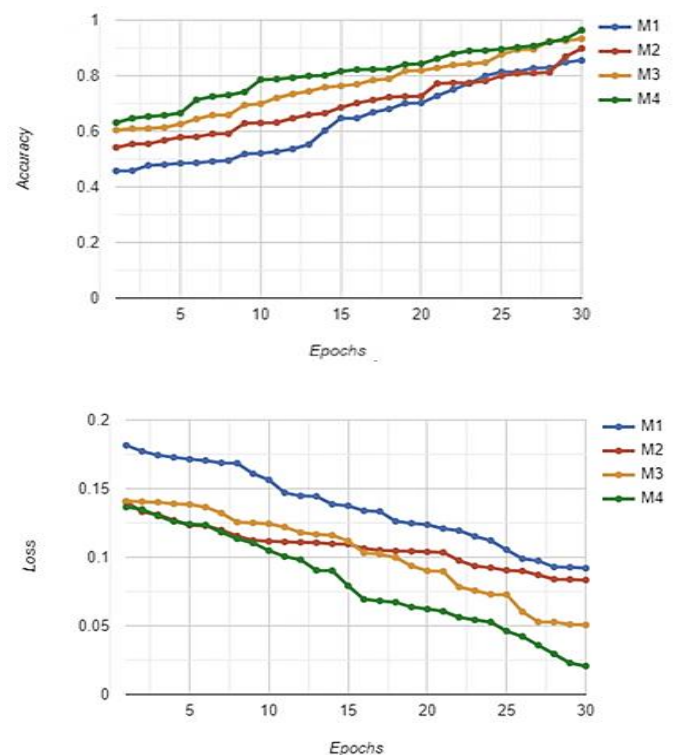


Figure 13. Accuracy and Loss Curve

Heatmaps and contouring techniques are employed to delineate the boundaries of objects or other characteristics to build a parametric curve or contour on MRI images. This allows for the identification of outliers or regions containing malignancies in MRI data. The final output is assessed independently using both methodologies, as demonstrated in *table 7*. *Table 6* shows that contouring approaches have a somewhat higher overall accuracy for detecting outliers compared to heatmap

techniques. Additionally, contouring procedures also result in lower loss. *Figure 14* depicts the accuracy and loss curve obtained through the utilization of LSTM+PSO+t-SNE using a contouring approach.

Table 7. Outcomes of Segmentation Techniques

Segmentation	Overall_Acc	Overall_Loss
Heatmap	95.94	0.0204
Contour	98.62	0.0193

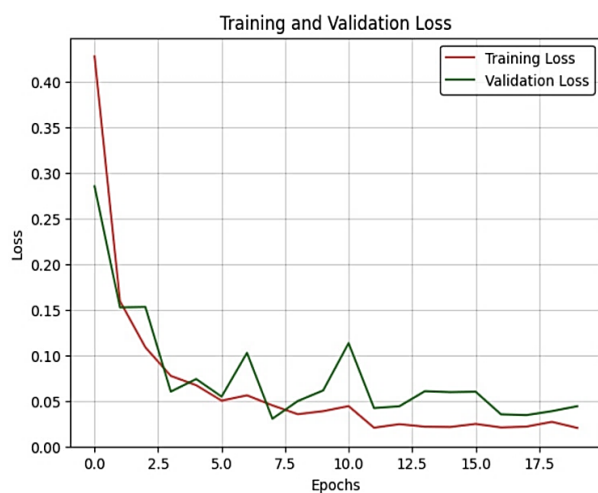


Figure 14. Accuracy and Loss outcomes with Contouring

This work investigated the development of a reliable classification model by employing various combinations of optimization algorithms. Optimization should be performed after image preparation to account for the significant influence of image normalization on outlier identification performance in deep learning classifiers. To mitigate the effects of normalization, CLAHE approaches are employed for the pre-processing of MRI images. The amalgamation of PCA, PSO, and t-SNE with LSTM gives numerous benefits in choosing the best features with the maximum variance from the parameters that are optimized. Initially, it improves the efficiency of optimization by using smaller feature space obtained using PCA and then reducing the redundant and irrelevant features from the given space. Moreover, PSO with dimension reduction techniques like PCA and t-SNE results in better outcomes. The proposed model outperformed the other techniques by demonstrating the originality of the work.

5.3. Comparative Analysis with State-of-Art Techniques

A varied assortment of input MRI scans and techniques has been validated using numerous advanced methods and procedures, resulting in the creation of several unique images and methodologies. Table 8 illustrates discrepancies in sample size, final results, and classification techniques utilized by previous researchers. Numerous diverse models have been trained and tested over time, with most achieving accurate categorization of various diseases using MRI images. Despite the proposed work in a study [25] exhibiting higher outcomes, it was executed on a limited dataset for multi-classification of only lung diseases using X-ray scans.

Furthermore, the research approach might be utilized alongside different techniques for the clinical diagnosis of MRI images. Due to their rapid availability and exceptional effectiveness, scans of the lungs, brain, and heart are especially advantageous for patients in critical conditions. The model can rapidly discern various diseases from MRI pictures. Consequently, it is highly advisable to implement a deep learning model that employs MRI visual data, as it provides a more reliable learning process. Employing healthcare monitors for training deep learning algorithms can improve picture classification and assist physicians in forecasting the prognosis of diagnosed conditions, hence augmenting their capacity to address emerging and prevalent diseases.

Table 8. Comparative Analysis of Proposed with Existing Techniques

Ref	Year	Technique	Dataset	Disease	Accuracy (%)	Optimization	Classification Type
[22]	2021	CNN	33,150 MRI Scans	Heart disease	95	Nil	Binary
[23]	2023	Fuzzy clustering	EMR data	Heart disease and diabetes	98.02	Battle Royale Optimization	Multi

[24]	2023	CNN	224 MRI Scan	Brain	90.1	Nil	Binary
[25]	2024	CNN	47,075 CXR Scan	Lungs	99.82	grid search optimization	Multi
[16]	2024	PSO-LSTM	---	Patient medical data	92.5	PSO	Binay
[26]	2024	SVM, LGB, XGB	599 MRI Scans	Brain	SVM- .989, LGB- 0.992, and XGB- 0.994	PSO + PCA	Multi
--	2024	Proposed Technique	MRI Scan of 22,855 Heart 1181 Brain, 1190 Lungs	Heart Brain, Lungs	98.58 brain, 98.47 lungs, 98.58 Heart	PSO + PCA, t-SNE+ PSO	Multi

5.4. Discussion

This study examined a robust classification model with various combinations of optimization procedures. Given the significant influence of image normalization on the efficacy of a deep learning classifier, meticulous attention must be devoted to the image preprocessing phase before conducting standard radiomic analyses. This work employed the CLAHE intensity normalization approach to mitigate the effects of normalization. Moreover, segmentation techniques for MRI image analysis, such as contouring and heatmaps, are employed to emphasize relevant locations, including tumors, lesions, or other anomalies. Each population-based metaheuristic algorithm consists of two essential phases: feature selection by PCA and t-SNE, and parameter optimization through PSO. A certain parameter or collection of parameters within the algorithm is designated this crucial function. In the context of PSO, the inertia weight serves as that parameter. Consequently, the appropriate modulation of inertia weight throughout each cycle is essential for enhancing classification accuracy. The recommended approach provided efficient results for outlier's detection on MRI images of the different human diseases. Combining feature selection and PSO with LSTM provides the component selection, giving a high range of variation from the optimized parameters, leading to numerous advantages. Primarily, enhances optimization by diminishing the feature space obtained through PCA and t-SNE which eliminate relevant features from the given feature vector space. Moreover, the presence of PSO enables the feature selection to operate with more interpretable data visualization, leading to improved insights into the significant challenges exhibited with the MRI image analysis.

Dimensionality reduction reduces computational complexity, improves visualization, and promotes classification and clustering efficacy. Our work employed a hybrid methodology utilizing feature selection (PSO and t-SNE) for optimization, followed by additional optimization by PSO, which enhanced efficiency, interpretability, and efficacy in addressing the

classification problem. Moreover, our suggested method outperformed all other available approaches across every categorization criterion, demonstrating its originality. This study also has some shortcomings. The limited patient sample size in the experiments may be considered inadequate for robust model validation. The aim is to overcome this constraint by conducting and validating the study using patient data from various institutions across the Indian subcontinent and other continents, which is now in progress. This study employed three types of diseases, which were rather modest in scale, hindering our ability to perform a stratified analysis for each location independently. Imbalanced datasets present considerable difficulty, potentially skewing models in favor of the majority class and undermining the performance of minority classes. The quality of data and preprocessing methods significantly affect model performance, as noisy or missing data might result in inferior outcomes. Ensuring robustness through meticulous preprocessing, managing absent values, and outliers, and rectifying mislabeled occurrences is essential for dependable outcomes.

6. CONCLUSIONS AND FUTURE SCOPE

To accurately recognize the outliers in MRI images, this study recommended a hybrid model using deep learning and dimensionality reduction modules. This work utilizes histogram equalization to enhance the local contrast of a designated area, thereafter applying the CLAHE approach, a contrast-limiting strategy to prevent excessive amplification of noise and artifacts in MRI images of the heart, lungs, and brain disorders. Various data augmentation techniques are utilized to increase the quantity of images in MRI datasets. The analysis encompasses the assessment of various segmentation approaches, including heatmap and contouring, on the identification of outliers in MRI data. The hybrid model utilizes LSTM, PSO, t-SNE, and PCA methodologies. The selected features were further refined using PCA and t-SNE techniques to reduce dimensionality, minimize noise, and improve the efficacy and efficiency of future

algorithms. The outcome of the proposed model revealed that the hybrid model using PSO techniques achieved the maximum efficiency in recognizing the outliers. PCA and t-SNE were used to select the most relevant and optimized features to improve the accuracy of the prediction model to identify the different regions of tumor classes. The proposed work has the potential to identify the three classes of brain, lung and heart diseases and four classes of brain Tumors based on their size. The finding of the existing work reveals that the accuracy of the predicted model is 98%. Moreover, utilization of the segmentation technique for outline detection of outliers in MRI images of the brain results in minimized time and enhanced performance. The suggested model demonstrated exceptional efficiency in carrying out this task. Future studies can investigate alternative time series forecasting techniques that would automatically adapt system settings to minimize the training period.

REFERENCES

- [1] M. Manthe, S. Duffner, and C. Lartzien, 'Whole brain radiomics for clustered federated personalization in brain tumor segmentation', in *Medical Imaging with Deep Learning*, 2024, pp. 957–977. Çinar, A., & Yildirim, M. (2020). Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture. *Medical hypotheses*, 139, 109684.
- [2] A. Çinar and M. Yildirim, 'Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture', *Medical hypotheses*, vol. 139, p. 109684, 2020.
- [3] A. Garg and Others, 'Outlier detection system for cardiovascular disease using time-series data: A comparative analysis', in *2023 IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDEA)*, 2023, pp. 378–382.
- [4] A. Humeau-Heurtier, 'Texture feature extraction methods: A survey', *IEEE access*, vol. 7, pp. 8975–9000, 2019.
- [5] M. M. Ahmed, A. Ali, B. N. Hashmi, W. K. Mashwani, and N. Gündüz, 'Review on Medical Image Segmentation and Feature Extraction Approaches', *Journal of Independent Studies and Research Computing*, vol. 21, no. 2, pp. 5–15, 2023.
- [6] J. He and X. Bi, 'Automatic classification of spinal osteosarcoma and giant cell tumor of bone using optimized DenseNet', *Journal of Bone Oncology*, vol. 46, p. 100606, 2024.
- [7] <https://www.kaggle.com/datasets/danialsharifrazi/cad-cardiac-mri-dataset> assessed on march 2024.
- [8] "Kaggle: Your Home for Data Science." [Online]. Available: <https://www.kaggle.com/datasets/shreyag1103/brain-mri-scans-for-brain-tumor-classification> assessed on march 2024.
- [9] <https://www.kaggle.com/datasets/adityamahimkar/iqothnccd-lung-cancer-dataset> assessed on march 2024.
- [10] H. Naveed, S. Anwar, M. Hayat, K. Javed, and A. Mian, 'Survey: Image mixing and deleting for data augmentation', *Engineering applications of artificial intelligence*, vol. 131, p. 107791, 2024.
- [11] C. Hu, H. Li, T. Ma, C. Zeng, and X. Ji, 'An improved image enhancement algorithm: Radial contrast-limited adaptive histogram equalization', *Multimedia Tools and Applications*, vol. 83, no. 36, pp. 83695–83707, 2024.
- [12] A. Naser, 'A proposed CLCOA Technique Based on CLAHE using Cat Optimized Algorithm for Plants Images Enhancement', *Wasit Journal of Computer and Mathematics Science*, vol. 3, no. 1, pp. 18–27, 2024.
- [13] B. Qian et al., 'Deep contour attention learning for scleral deformation from OCT images', *The Visual Computer*, vol. 41, no. 2, pp. 1155–1170, 2025.
- [14] J. Haferkamp, M. W. Ulmer, and J. F. Ehmke, 'Heatmap-based decision support for repositioning in ride-sharing systems', *Transportation Science*, vol. 58, no. 1, pp. 110–130, 2024.
- [15] I. Pavlů, J. Machalová, R. Tolosana-Delgado, K. Hron, K. Bachmann, and K. G. van den Boogaart, 'Principal component analysis for distributions observed by samples in Bayes spaces', *Mathematical Geosciences*, vol. 56, no. 8, pp. 1641–1669, 2024.
- [16] A. Ala, V. Simic, D. Pamucar, and N. Bacanin, 'Enhancing patient information performance in internet of things-based smart healthcare system: Hybrid artificial intelligence and optimization approaches', *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107889, 2024.
- [17] L. Xia, C. Lee, and J. J. Li, 'Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters', *Nature Communications*, vol. 15, no. 1, p. 1753, 2024.
- [18] J. Huang, F. Wang, L. Qiao, and X. Yang, 'T-distributed stochastic neighbor embedding echo state network with state matrix dimensionality reduction for time series prediction', *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106055, 2023.
- [19] I. Sembiring, S. N. Wahyuni, and E. Sedyono, 'LSTM algorithm optimization for COVID-19 prediction model', *Heliyon*, vol. 10, no. 4, 2024.
- [20] P. Dhiman, V. Kukreja, and A. Kaur, 'Citrus fruits classification and evaluation using deep convolution neural networks: an input layer resizing approach', in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2021, pp. 1–4.
- [21] P. Dhiman et al., 'PFDI: a precise fruit disease identification model based on context data fusion with faster-CNN in edge computing environment', *EURASIP Journal on Advances in Signal Processing*, vol. 2023, no. 1, p. 72, 2023. P. Dhiman, "Citrus Fruits Classification and Evaluation using Deep Convolution Neural Networks: An Input Layer Resizing Approach," pp. 21–24, 2021.
- [22] A. Sharma, R. Kumar, and V. Jaiswal, 'Classification of heart disease from MRI images using convolutional neural network', in *2021 6th international conference on signal processing, computing and control (ISPCC)*, 2021, pp. 358–363.
- [23] F. Yan, H. Huang, W. Pedrycz, and K. Hirota, 'A disease diagnosis system for smart healthcare based on fuzzy clustering and battle royale optimization', *Applied Soft Computing*, vol. 151, p. 111123, 2024.
- [24] Y. Wang et al., 'A brain MRI dataset and baseline evaluations for tumor recurrence prediction after Gamma Knife radiotherapy', *Scientific Data*, vol. 10, no. 1, p. 785, 2023.
- [25] S. Ashwini, J. R. Arunkumar, R. T. Prabu, N. H. Singh, and N. P. Singh, 'Diagnosis and multi-classification of lung diseases in CXR images using optimized deep convolutional neural network', *Soft Computing*, vol. 28, no. 7, pp. 6219–6233, 2024.
- [26] S. Pal, R. P. Singh, and A. Kumar, 'Analysis of Hybrid Feature Optimization Techniques Based on the Classification Accuracy of Brain Tumor Regions Using Machine Learning and Further Evaluation Based on the Institute Test Data', *Journal of Medical Physics*, vol. 49, no. 1, pp. 22–32, 2024.



© 2025 by the Ankita Roy, and Atul Garg. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).