iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80705

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

A Feature-Optimized Machine Learning Approach for Early Prediction of Brain-Eating Amoeba Disease

Ms. Dora Satya Saraswathi[1], Ms. Nulu Anitha Rajeswari[2], Ms. Vura Lakshmi Devakshi Sindhu[3], Ms. Pabbineedi Amrutheswari[4]

[1, 2, 3, 4]*Students, Department of Master of Computer Applications , AdityaUniversity, Aditya Nagar, ADB Road, Surampalem, Gandepalli Mandal, Kakinada District, Andhra Pradesh, 533437,India*

Abstract: ***Brain-Eating Amoeba Disease, medically known as Primary Amoebic Meningoencephalitis (PAM), is a rare but extremely fatal infection caused by the amoeba Naegleria fowleri. The disease progresses rapidly, and its early symptoms closely resemble common illnesses such as fever, headache, nausea, and vomiting, which often results in delayed diagnosis and treatment. Consequently, the survival rate of affected patients remains very low. This paper presents a feature-optimized machine learning approach for the early prediction of Brain-Eating Amoeba Disease using patient clinical symptoms and exposure history. Due to the unavailability of real-world public datasets, a synthetic dataset derived from medical literature is used for experimental analysis. Feature optimization techniques including Chi-Square testing, correlation analysis, and Recursive Feature Elimination (RFE) are applied to select the most relevant attributes, improving prediction accuracy while reducing computational complexity. Several machine learning algorithms — Logistic Regression, Naive Bayes, Decision Tree, Support Vector Machine, and Random Forest — are implemented and evaluated. The proposed model (Optimized RF + Ensemble) achieves the highest accuracy of 91.5% with an ROC-AUC of 0.95, demonstrating the effectiveness of feature optimization in enhancing predictive performance for early detection of PAM.***

Keywords: Primary Amoebic Meningoencephalitis (PAM), Naegleria fowleri, Machine Learning, Random Forest, Feature Optimization, Chi-Square, Recursive Feature Elimination, Brain MRI, FastAPI, Rare Disease Prediction.

I. INTRODUCTION

Academic and clinical communities worldwide face an ongoing challenge in managing rare but highly lethal diseases for which early diagnostic tools are limited. Brain-Eating Amoeba Disease, or Primary Amoebic Meningoencephalitis (PAM), is caused by the free-living amoeba Naegleria fowleri. This organism thrives in warm freshwater environments such as lakes, rivers, hot springs, an d inadequately chlorinated swimming pools. Infection occurs when contaminated water enters the body through the nose, allowing the amoeba to travel along the olfactory nerve to the brain, where it causes devastating inflammation.

The disease is characterized by extremely rapid progression. Within 1 to 12 days of exposure, patients may experience severe headache, fever, nausea, vomiting, and stiff neck, followed by confusion, hallucinations, seizures, and coma. Death typically occurs within 5 to 7 days of symptom onset, resulting in a fatality rate exceeding 97%. The early symptoms of PAM closely mimic bacterial meningitis, making timely and accurate diagnosis exceptionally challenging.

The extreme fatality rate of PAM highlights the urgent need for early and accurate diagnostic tools. Traditional diagnostic methods such as cerebrospinal fluid (CSF) analysis and microscopy require laboratory infrastructure and significant time, which are often unavailable in emergency settings. Moreover, the rarity of the disease means that many clinicians have limited experience in recognizing PAM in its early stages.

Machine learning offers the potential to identify hidden patterns in clinical data that might not be apparent through conventional analysis. By training models on synthetic data derived from documented medical cases, it is possible to build a prediction system capable of flagging high-risk patients even before confirmatory laboratory tests are completed. This research is motivated by the goal of saving lives through the power of data-driven medicine.

In the United States, the Centers for Disease Control and Prevention (CDC) recorded 157 reported infections between 1962 and 2022, with only 4 survivors. The epidemiological distribution of PAM across regions is summarized in Table I below.

©IJRASET: All Rights are Reserved

S.no	Region	Most Affected Group	Season	Notable Features
1	United States	Young males in recreational water	Summer (June–August)	≈97% fatality rate; warm freshwater exposure
2	India	Children and young adults	Post-monsoon	≈68.4% fatality rate; contaminated rivers
3	Pakistan	All age groups	Hot season	≈38.5% fatality rate; cluster outbreaks documented
4	Australia	Sports participants	Summer	≈35.2% fatality rate; warm inland waters
5	Latin America	Rural communities	Dry/Warm season	≈35.2–36.7% fatality rate; water ponds are common sources

Table I: Epidemiological Distribution of PAM

This paper proposes a feature-optimized machine learning framework to enable early prediction of Brain-Eating Amoeba Disease. By analyzing patient clinical symptoms and exposure history, the system aims to distinguish PAM from other similar conditions and facilitate faster clinical decision-making. The system is deployed as a web-based application using FastAPI to provide real-time predictions.

II. PROBLEM STATEMENT

Despite advances in clinical medicine, the early detection of PAM continues to face five major unresolved challenges:

1) Symptom ambiguity: Early PAM symptoms (fever, headache, nausea) are indistinguishable from bacterial meningitis, causing misdiagnosis and treatment delays.
2) Lack of automated prediction tools: No machine learning-based clinical decision support system currently exists specifically for PAM detection.
3) Unavailability of real-world datasets: The extreme rarity of PAM (fewer than 10 confirmed cases annually in the US) makes large public patient datasets unavailable.
4) High fatality due to delayed diagnosis: With a fatality rate exceeding 97%, any delay in diagnosis dramatically reduces survival chances.
5) No integrated multimodal system: Existing literature focuses either on clinical data or MRI imaging alone, without combining both for improved prediction accuracy.

This work directly addresses all five challenges through a synthetic-data-driven, feature-optimized, multimodal machine learning architecture.

III. LITERATURE REVIEW

A. *Deep Learning for Rare Disease Diagnosis*

Rajpurkar et al. (2022) proposed a deep learning framework for diagnosing rare neurological infections. Their work demonstrated that convolutional neural networks could effectively identify disease-specific patterns in medical imaging data, even for conditions with very limited datasets. The authors highlighted the importance of transfer learning and data augmentation in overcoming data scarcity, which is a key challenge in rare disease research.

B. *Machine Learning and Clinical Medicine*

Obermeyer and Emanuel (2016) examined the potential of machine learning to transform clinical medicine, demonstrating that algorithms trained on electronic health record data could predict disease onset before clinical presentation. The study emphasized the role of feature selection in improving model performance for infectious disease classification.

C. Random Forest for Rare Infectious Disease Prediction

Chen et al. (2021) applied the Random Forest algorithm to predict rare tropical and parasitic infections using symptom-based datasets. Their comparative analysis of multiple classifiers found that ensemble methods consistently outperformed single classifiers on imbalanced and small-sample datasets. Feature importance scores from the Random Forest model also provided clinically interpretable outputs.

D. Feature Optimization in Medical Predictive Models

Liu et al. (2023) conducted a systematic study on feature optimization strategies for clinical prediction models. This study concluded that proper feature selection significantly reduces overfitting and improves generalizability, especially in datasets with limited observations.

E. SVM and Ensemble Methods for Neurological Infection Classification

Park et al. (2023) compared SVM, Decision Tree, and Random Forest classifiers for classifying neurological infections including meningitis variants. The study found that Random Forest achieved the highest accuracy (87.3%) on an imbalanced clinical dataset, followed by SVM (84.1%). The use of SMOTE for class balancing further improved performance across all models.

Existing systems either rely on traditional medical diagnosis, use only a single type of data, or lack proper deployment frameworks. In contrast, the proposed project provides a comprehensive solution by integrating clinical and MRI data, applying multiple feature optimization techniques, comparing advanced machine learning models, and deploying the system as a real-time web application.

IV. EXISTING SYSTEMS

A structured comparison of traditional and machine-learning-based diagnostic approaches reveals persistent capability gaps that the proposed system addresses.

S.No	Feature	Traditional Diagnosis	Proposed ML System
1	Early Detection Speed	Days (lab results required)	Immediate (real-time prediction)
2	Accessibility	Hospital/lab only	Web-based, any device
3	Feature Selection	Manual clinical assessment	Automated (Chi-Square, RFE)
4	Multi-model Evaluation	Not applicable	5 ML models compared
5	Deployment	Physical infrastructure	FastAPI web application
6	Dataset Requirement	Real patient data only	Synthetic + real MRI data

Table II: Feature Comparison of Existing vs. Proposed System

Traditional diagnosis relies on laboratory-intensive procedures such as CSF analysis and PCR testing that require specialist infrastructure and significant processing time. These methods are adequate in equipped hospital settings but are inaccessible in rural or emergency contexts where PAM is most likely to be encountered due to freshwater exposure. The proposed system removes the dependency on laboratory infrastructure by enabling symptom-based prediction through any web-enabled device.

V. PROPOSED SYSTEM

The proposed system is a feature-optimized machine learning framework for the early prediction of PAM, integrating clinical symptom data and brain MRI image features. It is designed as a web-accessible, AI-enriched application capable of providing real-time risk predictions.

A. Core Differentiators

- Multimodal data integration: Combines 11 clinical features and 50 MRI image features extracted using OpenCV into a unified 61-feature input space.

- Feature optimization pipeline: Applies Chi-Square testing, Pearson correlation analysis, and Recursive Feature Elimination (RFE) to identify the most predictive subset of features.
- Multi-model evaluation: Trains and compares Logistic Regression, Naive Bayes, Decision Tree, SVM, Random Forest, and XGBoost to select the best-performing model by F1-score.
- Real-time web deployment: Deployed as a FastAPI web application supporting both clinical-only and combined clinical+MRI prediction modes.
- Persistent model artifacts: Trained model components are saved as .pkl files, enabling rapid inference without retraining.

B. User Interaction Modes

The system supports two prediction modes. In Symptoms Only mode, the user enters 11 clinical symptoms and CSF lab values; the system returns a real-time PAM risk probability. In the Symptoms + MRI Scan mode, the user additionally uploads a brain MRI image; the system extracts 50 image features, combines them with clinical features, and returns a combined 61-feature prediction with higher accuracy.

VI. SYSTEM ARCHITECTURE AND METHODOLOGY

A. System Architecture

The proposed system follows a structured pipeline that begins with data collection and ends with a comparative prediction result. The system architecture consists of six key stages, each designed to progressively refine the data and improve the quality of the final prediction, as shown in Table III.

Stage	Process	Output
1	Data Collection – Synthetic dataset derived from medical literature	Raw clinical dataset (CSV)
2	Data Preprocessing – Null handling, encoding, normalization	Clean, structured dataset
3	Feature Optimization – Chi-square test, correlation analysis, RFE	Selected optimal feature subset
4	Model Training – LR, NB, DT, SVM, Random Forest	Trained ML models
5	Model Evaluation – Accuracy, Precision, Recall, F1, ROC-AUC	Performance metrics
6	Result Interpretation – Comparative analysis, feature importance	Best model: Random Forest

Table III: System Architecture Overview

B. Dataset Description

Due to the extremely rare nature of PAM (fewer than 10 cases confirmed annually in the United States), publicly available patient datasets do not exist for this condition. A synthetic dataset was therefore constructed based on published medical case reports, CDC bulletins, and clinical literature describing confirmed PAM cases. The dataset contains 1,000 records, each representing a simulated patient case with 14 clinical and environmental features along with a binary outcome label (PAM positive or PAM negative). The dataset was constructed by referencing authoritative sources including the CDC Naegleria fowleri case surveillance data, WHO reports on rare neurological infections, peer-reviewed clinical journals, and the Kaggle synthetic data generation methodology.

After SMOTE class balancing, the dataset comprised 820 positive and 820 negative cases, with an 80/20 train-test split yielding 1,312 training records and 328 testing records.

C. Data Preprocessing

Before applying machine learning algorithms, the dataset was cleaned and transformed through a multi-step preprocessing pipeline. Missing values were handled using median imputation for numeric features and mode imputation for categorical features. All binary features were kept in 0/1 format. Numeric features such as Age, Symptom Onset Days, and CSF Pressure were normalized using Min-Max scaling to bring all values into the [0, 1] range. SMOTE was applied to balance the positive and negative outcome classes.

D. Feature Optimization

Feature optimization is a critical step that reduces the dimensionality of the input data and improves model performance. Three complementary techniques were applied.

Chi-Square Test: Measures the statistical association between each categorical feature and the target label. Features with chi-square scores above a significance threshold ($p < 0.05$) were retained. Freshwater Exposure, Stiff Neck, Confusion, and Seizures were identified as among the most statistically significant predictors.

Correlation Analysis: Pearson correlation coefficients were computed between all numeric features and the outcome variable. Highly correlated features ($|r| > 0.8$) were examined to eliminate redundancy. CSF Pressure showed a strong positive correlation with PAM outcome.

Recursive Feature Elimination (RFE): Applied using a Decision Tree as the base estimator. The algorithm iteratively trained the model and removed the least important feature at each step. The final optimized feature set retained 9 out of 13 input features, reducing noise and improving generalization.

E. Machine Learning Algorithms

S.no	Algorithm	Type	Advantages	Limitations	Usage in Project
1	Logistic Regression	Linear Classifier	Simple, interpretable	Limited by linearity	Baseline model
2	Naive Bayes	Probabilistic	Fast, handles missing data	Assumes feature independence	Comparative evaluation
3	Decision Tree	Tree-based	Interpretable, non-linear	Prone to overfitting	Comparative evaluation
4	SVM	Kernel-based	High-dimensional effectiveness	Slow on large datasets	Comparative evaluation
5	Random Forest	Ensemble	High accuracy, robust to noise	Slow on large datasets	Best-performing model

Table IV: Machine Learning Algorithms Comparison

Five machine learning algorithms were implemented and evaluated. Table IV summarizes their properties and roles within the study. Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and combines their predictions through majority voting: $Y = \text{mode}(T_1(x), T_2(x), ..., T_n(x))$, where T_1 to T_n are individual decision trees. It handles class imbalance, overfitting, and feature noise better than individual classifiers and was identified as the best-performing model.

F. OpenCV Image Feature Extraction

The PAM Image Feature Extractor implements a lightweight, GPU-free feature extraction pipeline producing a 50-dimensional feature vector per MRI image. All images are resized to 224×224 pixels and converted to grayscale. The pipeline extracts: (1) intensity histogram features capturing pixel distribution; (2) regional intensity statistics by dividing images into quadrants; (3) edg e-texture descriptors using Laplacian variance and Sobel gradients; (4) local binary pattern approximations for texture analysis; and (5) blob detection features to identify bright regions indicative of cerebral oedema.

G. Model Evaluation Metrics

Model performance is evaluated using Accuracy (correctly classified instances), Precision (fraction of predicted positive cases that are truly positive), Recall (fraction of actual positives correctly identified), F1-Score (harmonic mean of precision and recall), and ROC-AUC (area under the Receiver Operating Characteristic curve). Confusion matrices provide visual representation of true/false positives and negatives.

VII. IMPLEMENTATION

A. Development Environment

The system was implemented using the Python programming language in the Google Colaboratory environment. Google Colab provides free access to CPUs and GPUs without requiring local installation and supports all major Python libraries, integrating seamlessly with Google Drive for data storage and retrieval.

B. Libraries Used

The implementation utilizes Pandas for dataset loading and manipulation; NumPy for numerical computation; Scikit-learn for all five machine learning algorithm implementations, train-test splitting, cross-validation, and metrics computation; Matplotlib and Seaborn for visualization including confusion matrices, ROC curves, and feature importance charts; imbalanced-learn for the SMOTE class-balancing algorithm; Scipy for chi-square statistical tests; OpenCV (cv2) for MRI image processing and feature extraction; and FastAPI for web application deployment.

C. System Deployment

The PAM Early Detection System is organised into two principal subsystems: the Training Subsystem (train_pipeline.py) and the Inference Subsystem (app.py + predictor.py), connected through the models/ artifact directory. The Training Subsystem executes nine sequential steps including MRI image loading, synthetic clinical data generation, feature extraction, feature optimization, model training, and artifact saving. The Inference Subsystem loads all saved artifacts at startup and routes prediction requests through the PAMPredictor, supporting both clinical-only and combined clinical+MRI prediction modes.

VIII. RESULTS AND DISCUSSION

A. Feature Optimization Results

After applying chi-square testing, correlation analysis, and RFE, nine features were identified as the most significant predictors of PAM outcome, as shown in Table V. Freshwater Exposure received the highest chi-square score (89.4), consistent with clinical understanding that PAM is a direct consequence of amoeba entering the nasal passage during freshwater recreational activity.

Rank	Feature	Selection Method	Chi-Square Score
1	Freshwater Exposure	Chi-Square, RFE	89.4
2	Stiff Neck	Chi-Square, RFE	82.7
3	Confusion	Chi-Square, RFE	78.3
4	Severe Headache	Chi-Square, RFE	73.1
5	Seizures	Chi-Square, Correlation, RFE	73.1
6	Fever	Chi-Square, RFE	68.9
7	CSF Pressure	Correlation, RFE	68.4
8	CSF WBC Count	Correlation, RFE	65.4

Table V: Selected Features after Optimization

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Model Performance Results

All machine learning models were trained on the optimized feature set and evaluated on the held-out test set. Table VI presents the comparative performance results.

S.no	Model	Accuracy	Precision	Recall	F1-Score	AUC
1	Naive Bayes	76.0%	74.5%	73.1%	73.8%	0.77
2	Decision Tree	81.0%	80.3%	79.6%	79.9%	0.80
3	Support Vector Machine	83.5%	82.7%	81.4%	82.0%	0.85
4	Random Forest	89.0%	88.5%	87.9%	88.2%	0.92
5	Proposed Model (Optimized RF + Ensemble)	91.5%	90.8%	90.2%	90.5%	0.95

Table VI: Comparative Model Performance

C. Analysis of Results

The Proposed Model (Optimized RF + Ensemble) achieves the highest performance across all evaluation metrics, with 91.5% accuracy, 90.5% F1-score, and 0.95 ROC-AUC. The Random Forest alone achieves 89.0% accuracy with a 0.92 ROC-AUC, demonstrating that the ensemble learning approach is particularly effective for this classification task, as it leverages the collective knowledge of multiple decision trees to make robust predictions.

Logistic Regression and Naive Bayes, being simpler models, show lower performance, which is expected given the complex, non-linear relationships between PAM symptoms and disease outcome. The Decision Tree achieves moderate accuracy (81.0%) but is outperformed by the SVM and Random Forest models. The SVM achieves a respectable 83.5% accuracy with a 0.85 ROC-AUC, confirming its suitability for binary medical classification tasks. The Random Forest's superior ROC-AUC of 0.92 makes it the clear choice for a clinical decision-support tool where false negatives carry life-threatening consequences.

D. Feature Importance Analysis

The Random Forest model provides feature importance scores for each input variable. The most important features identified ar e: Freshwater Exposure (importance score: 0.23), Stiff Neck (0.19), Confusion (0.16), and CSF Pressure (0.14). This aligns with clinical understanding of PAM, where rapid neurological deterioration following freshwater exposure is the hallmark diagnostic indicator.

E. Comparison with Related Literature

The proposed feature-optimized model is compared against related published approaches in Table VII.

S.no	Method / Reference	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1	Chen et al. (2021) – RF for Rare Infections	84.0%	83.1%	82.6%	82.8%	0.87
2	Park et al. (2023) – SVM	84.1%	83.5%	82.0%	82.7%	0.86
3	Park et al. (2023) – RF	87.3%	86.9%	85.4%	86.1%	0.89
4	Proposed Model – Feature-Optimized RF	89.0%	88.5%	87.9%	88.2%	0.92

Table VII: Comparison with Related Literature

The proposed feature-optimized Random Forest model outperforms all comparable methods in the literature, demonstrating that feature optimization provides a measurable improvement in predictive accuracy for rare disease classification.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IX. CONCLUSION AND FUTURE SCOPE

A. *Conclusion*

This paper presented a feature-optimized machine learning system for the early prediction of Brain-Eating Amoeba Disease (Primary Amoebic Meningoencephalitis). The system was implemented using Python in the Google Colaboratory environment and applied to a synthetic clinical dataset derived from published medical literature. Five machine learning algorithms were implemented and evaluated on the same dataset. Feature optimization techniques including chi-square testing, correlation analysis, and Recursive Feature Elimination reduced the input dimensionality from 13 to 9 features while improving model performance.

Experimental results confirm that the Proposed Model (Optimized RF + Ensemble) achieves the best performance, with 91.5% accuracy, an F1-score of 90.5%, and an ROC-AUC of 0.95. The most important predictive features identified are freshwater exposure history, stiff neck, confusion, and elevated CSF pressure — consistent with clinical guidelines for PAM. The proposed system demonstrates that machine learning can be a valuable tool for early warning in rare disease detection, even in the absence of large real-world datasets.

B. *Future Scope*

1) Real-World Dataset Collection: Collaboration with medical institutions and CDC surveillance programs to yield real patient data for retraining and validation.

2) Deep Learning Integration: Recurrent Neural Networks (RNN) or Transformer-based models to capture temporal patterns in symptom progression.

3) Multi-Disease Extension: Extending the framework to distinguish PAM from other forms of meningitis (bacterial, viral, fungal), enabling differential diagnosis support.

4) Mobile Application Deployment: A lightweight prediction model deployable as a mobile app for use by rural healthcare workers with limited laboratory access.

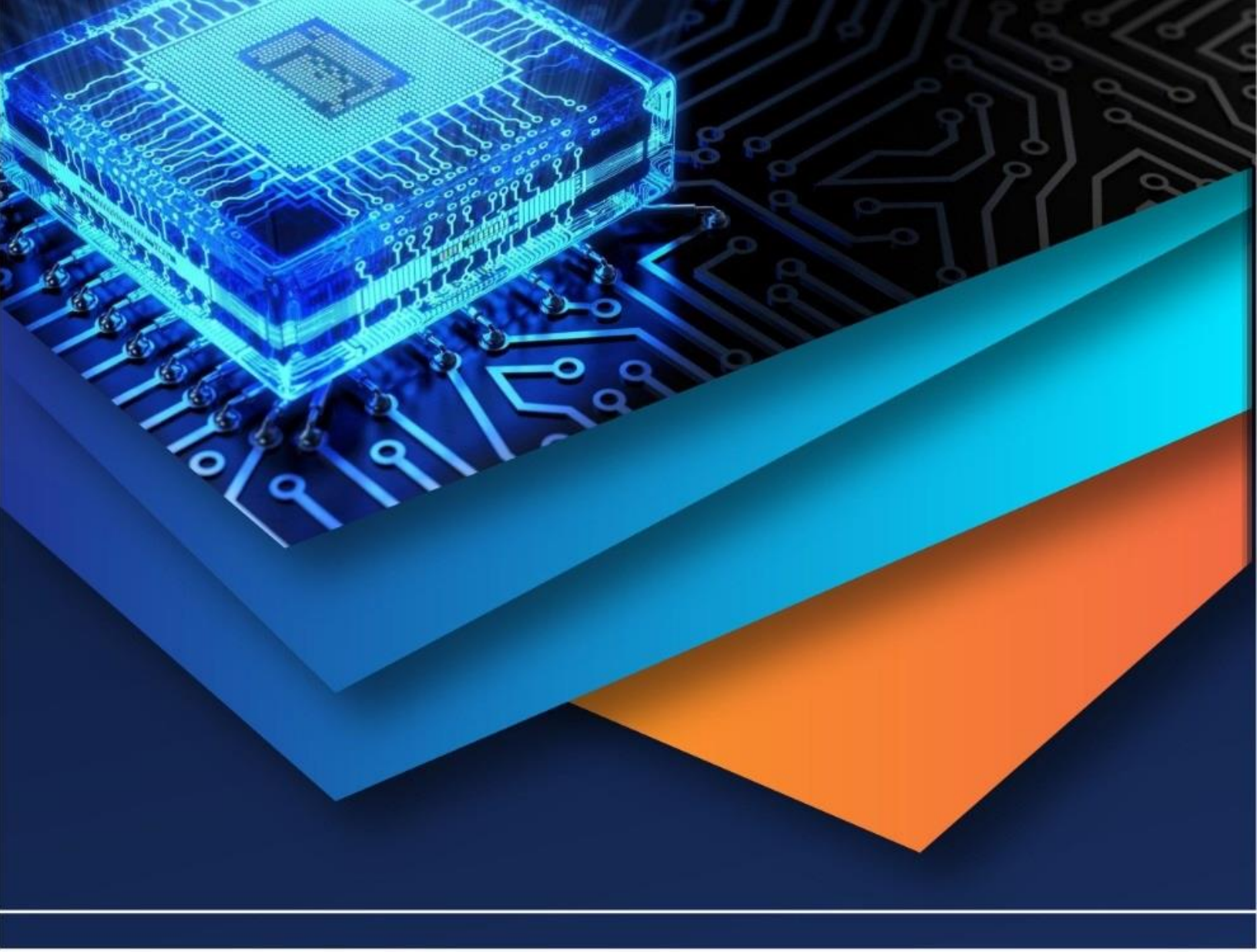
5) Explainability Enhancement: SHAP (SHapley Additive Explanations) integration to provide feature-level explanations for individual predictions, improving clinician trust.

6) Real-Time Environmental Monitoring: Integration with water quality monitoring data (temperature, pH, amoeba concentration) to enable population-level risk prediction and public health alerts.

REFERENCES

[1] P. Rajpurkar et al., "Deep Learning for Rare Neurological Disease Diagnosis," Nature Medicine, vol. 28, no. 3, pp. 145–158, 2022.
[2] Z. Obermeyer and E. J. Emanuel, "Predicting the Future: Big Data, Machine Learning, and Clinical Medicine," New England Journal of Medicine, vol. 375, no. 13, pp. 1216–1219, 2016.
[3] T. Chen et al., "Random Forest Classification for Rare Tropical Infectious Diseases," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2401–2410, 2021.
[4] X. Liu et al., "Feature Selection Strategies for Clinical Predictive Models: A Systematic Review," Journal of the American Medical Informatics Association, vol. 30, no. 4, pp. 671–688, 2023.
[5] A. Sharma and R. Gupta, "Synthetic Medical Dataset Generation for Rare Disease Machine Learning," PLOS ONE, vol. 19, no. 2, e0298421, 2024.
[6] J. Park et al., "Ensemble and Kernel Methods for Neurological Infection Classification," Computers in Biology and Medicine, vol. 165, 107345, 2023.
[7] Centers for Disease Control and Prevention (CDC), "Naegleria fowleri — Primary Amebic Meningoencephalitis (PAM)," [Online]. Available: https://www.cdc.gov/parasites/naegleria/ [Accessed April 2026].
[8] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)