Performance Evaluation of AI Models for Multi-Class Brain Disease Prediction

Aakriti Khanna , Shubham Kumar , Saruchi Kukkar
Department of Computer Science and Engineering Chandigarh University , Department of Computer Science and Engineering, Chandigarh University , Department of Computer Science and Engineering, Chandigarh University
aakritikhanna1720@gmail.com , shuh5225@gmail.com, ganpati.saruchi@gmail.com

Abstract.

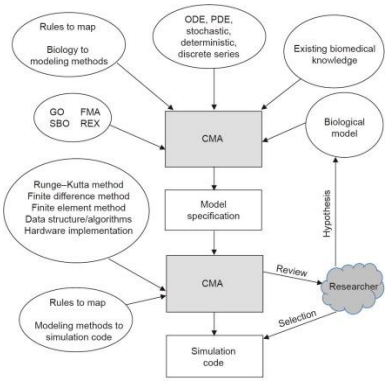
Machine learning and artificial intelligence are revolutionizing medical research, development and care. Open-source platforms, collaborative projects, and information exchange forums have democratized AI and ML, enabling healthcare innovation and knowledge sharing. The proposed work is divided into three parts that collect the data sets, train different models, and compare the results. We have taken five brain diseases dataset that include depression, brain stroke, Alzheimer, Parkison disease, and Epilespy. Also we have trained six different MLmodels that are random forest that is RF, support vector machine that is SVM , linear discriminant analysis that is LDA, NB that is Naive Bayes , Gradient Boost that is XGboost and logistic regression that is LR. The highest accuracy for logistic regression is 90% for parkisons disease , 97% for XGboost model of depression , 97% for for random forest of epilepsy , 87% for LDA model of Parkisons disease, 96\% for naive bayes of epilepsy and 97% for SVM models of epilepsy disease. We have also used the heatmaps and bar graphs to show the comparison between them all.

Keywords. Artificial Intelligence, Machine Learning, LR model , Naive Bayes model, XGboost model , SVM model, LDA model , Random forest model , Brain diseases

1. Introduction

AI can improve healthcare efficiency, accessibility, and cost. Figure 1 shows how AI with proper regulation can improve equity, care, and society's access to new technology, cures, and pharmaceuticals. This ensures timely resource availability, reducing waste and improving care. AI may solve healthcare's biggest problems: cost, efficiency, and quality. AI-powered automation and optimization can cut costs and simplify patient scheduling, billing, and EHR management. AI changes pharmaceutical R&D, pharmacokinetics, assessment, production, marketing approval, and pharmacovigilance. AI assists patient stratification, digital twins, clinical trial simulations, and pharmacokinetic dose recommendations. Simplifying research, streamlining processes, and avoiding clinical trial failures saves time and money for safe and effective pharmaceuticals. Personalized medicine diagnoses, predicts, and analyzes patients using enormous healthcare data. Imaging, drug combinations, population-wide patient health outcomes, genetic or family data, medical papers natural language analysis [3]. Researchers and medics want to reduce worldwide illness mortality. Due to expanding healthcare data from varied and incompatible sources, predictive analytic models are becoming more relevant in medicine. General database storage cannot manage massive healthcare historical and streaming data. Medical diagnostics fix issues. Observational diagnosis is disease. Diseases have chemical and data anomalies. ML algorithms can learn categories or anticipate unknown circumstances from data[6]. Many old brain data-processing processes work. ML detects Alzheimer's, dementia, schizophrenia, MS, cancer, viral, and degenerative diseases in brain data. Sectioning and identifying sick tissue and brain structure. Diagnosis, surgery, postoperative examination, and chemo/radiotherapy require tissue identification. AI-powered virtual assistants and chatbots offer individualized health advice, prescription reminders, and real-time support. Despite progress, healthcare AI/ML integration is difficult. To improve healthcare using AI and ML, we must address ethics, data protection, regulatory compliance, and the digital divide.Fig. 1.

Artificial intelligence in medical field



2. Literature Review

The substantial healthcare literature on AI and ML reflects its many effects on patient care, clinical decision-making, and health system optimization. Early and present healthcare AI and ML studies demonstrate its benefits and drawbacks. Early 2000s studies showed that computer-aided diagnosis systems can diagnose numerous diseases, laying the framework for AI and ML medical imaging interpretation. ML method by Smith et al. detects CT lung nodules with 92\% sensitivity and 96\% specificity. AI-driven mammography interpretation for breast cancer diagnosis was proven by Lee et al.[4]. Faster, more accurate, and scalable AI diagnostics beat traditional approaches. A meta-analysis by Johnson et al. evaluated radiologists and AI systems for CT pulmonary angiography embolism detection. Deep learning enhances healthcare AI/ML. Deep learning algorithms, inspired by the brain, excel in photo identification and natural language processing, making them ideal for medical data analysis[1] Deep learning models can improve brain and heart diagnosis and therapy utilizing MRI and CT. AIML and deep learning customize brain,heart health. AI-driven models improve treatment results and side effects using genetics, biomarkers, and clinical history. A personalized healthcare plan could personalize patient care[7].

3. Methodology

3.1. Dataset: Our study used Kaggle's Parkinson's, Alzheimer's, epilepsy, brain stroke, and depression datasets. Logistic Regression (LR), Gradient Boosting (GBoost), SVM, and Random Forest were created and tested using these datasets. Every dataset includes 5000 people's names, genders, and medical histories (heart disease, hypertension, and average glucose). All datasets had binary targets of (0) or (1). All data were.csv. Spreadsheets create CSV files using commas [2]

3.2.Data Preprocessing: Performing categorization quality, consistency, and sustainability preprocessing before model training and testing. BMIs missing were imputed using all entries' mean. One-hot encoded multi-class features like work type and smoking status, while label encoded categorical data. Numbers become 0 and 1. A lightweight, high-performing convolutional neural network, EfficientNet B0, extracted features to improve representation. All extracted features go to machine learning models following pretrained data extractor architecture. Mixed handmade and learning components increase model capture of complex patterns in the hybrid method.

3.3. Disease

3.3.1 Alzheimer's disease: Alzheimer's causes most dementia. Alzheimer's begins with amyloid plaques and neurofibrillary tangles[8]. Brain cell death lowers it gradually. Protein accumulation around brain cells may cause Alzheimer's. Amyloid surrounds brain cells. Tau protein causes brain tangles.

3.3.2. Brain Stroke: Stroke results from brain blood flow stoppage. In an emergency. Brain function requires continual oxygen and nourishment. Even a brief blood shortage can be problematic[4]. Two main factors induce stroke. A clogged brain artery causes ischemic stroke. Hemorrhagic strokes result from brain blood vessel ruptures. Some people experience brief ischemia attacks, which disrupt cerebral blood flow.

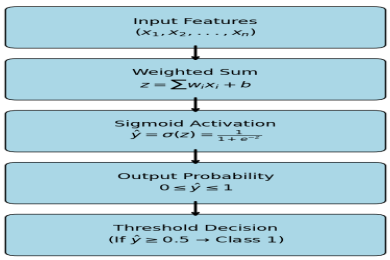
3.3.3.Depression: Depression is common and affects everyone. Loss of interest or long-term depression describe it. This differs from mood swings and daily feelings. No single reason causes depression. It has several causes and triggers.

3.3.4.Parkison Disease: The neurological illness PD impacts movement, mental health, sleep, pain, and more. Over time, PD worsens. Therapies and drugs reduce symptoms but do not cure. Speech issues, tremors, and strong muscle contractions are prevalent. Brain chemical messenger neuron loss causes many Parkinson's symptoms. The messenger is dopamine. Brain activity changes with reduced dopamine. This causes Parkinson's mobility difficulties.

3.3.5.Epilepsy: Epilepsy affects all ages. Globally, 50 million have epilepsy. Epileptics are common in poor countries. Neurological epilepsy causes seizures[5]. Different epilepsy types exist. Some know why. Unknown causes exist elsewhere. Epilepsy is prevalent. The CDC reports 1.2\% of Americans have active epilepsy. All sexes, races, ethnicities, and ages have epilepsy. The symptoms of seizures differ. Seizures knock you out. A few seizure victims stare blankly. Some jerk legs or arms.

3.4 Classifiers

3.4.1.Logistic regression (LR): Linear regression is the accurate correspondence between dependent and independent variables with few or more variables[6]. Since linear regression accurately predicts the change in dependent class variables to independent class variables. The algorithm of LR is shown in figure 2.



3.4.2.Gradient boost (GBoost): Gradient boosting (GBoost) creates powerful prediction models from weak learners like decision trees. Simple models predict first. Improved accuracy is achieved by training new trees each iteration to forecast the aggregate ensemble's residual errors.

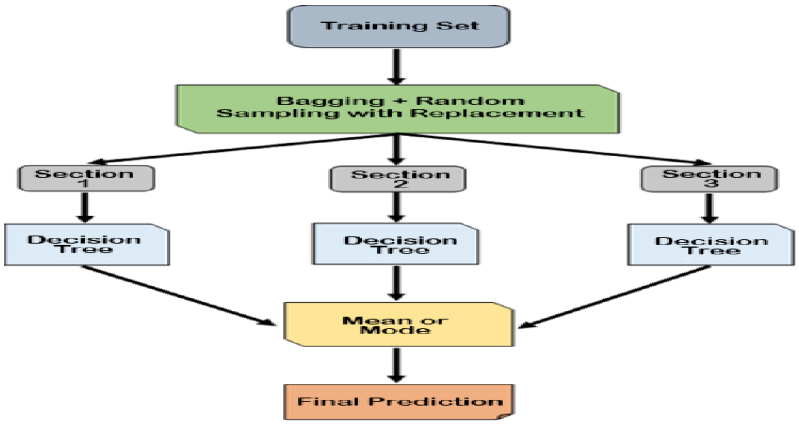
3.4.3.Naive Bayes: Naive Hayes classifies statistically and supervisedly. Bayesian classifier. A value is assumed. Unrelated feature any other feature. Before probability, likelihood evaluated posterior probability[9]. The method for highest posterior probability varmeter estimate. The method is easy. enough training data to estimate parameters .; Required categorization. Training time and classification drops.

3.4.4. Linear Discriminant Analysis (LDA): A popular natural language processing generative probabilistic topic model is Linear Discriminant Analysis. LDA uses word distributions to identify corpus document latent themes. The "bag-of-words" assumption ignores document word order and emphasizes frequency and co-occurrence. LDAs are created by choosing a subject distribution, topic, and word from the topic's word distribution for each word.

3.4.5.Support vector machines (SVM): SVM classifies data points uniquely using an N-dimensional hyperplane. Supervised machine learning for classification, regression, outliers. Margin maximisation aids new-data model generalization. SVM handle linear and nonlinear data. For linearly separable data, model will find out the straight line

3.4.6. Random forest (RF)

Random forests use randomized data and several separate decision trees (DT). These trees grow throughout training. compiles the output of the decision tree. Voting methods determine the prediction of the algorithm. This paradigm requires DTs to choose stroke sign indictations or absences. Forecast comes out. Working Random Forest chooses the most-voted class as shown in figure 3.



3.5. Validation

3.5.1.Accuracy: Accuracy (the percentage of correctly identified examples among all predictions) can be misleading in imbalanced datasets.

Accuracy = Total positive value/Total Value

3.5.2.Precision: Precision measures positive case prediction accuracy.

Precision = True positive / Total positive

3.5.3.F1 Score: The harmonic mean of accuracy and recall knowns as the F1 score, adjusts for imbalanced classes or large false positives and negatives.

F1 score = 2* PV *RV /PV+RV , PV = Precision Value , RV = Recall Value

3.5.4. Recall : Recall (or sensitivity) measures how many positive cases the model correctly discovered.

Recall = Original real detection value / Original real detection value + False wrong detection value

3.5.5. Confusion Matrix : Verified real positives, verified real negatives, verified fake positives, and verified fake negatives are listed in the confusion matrix to detail prediction outcomes.

3.5.6 Receiver Operating Characteristic Curve : The ROC curve demonstrates how well the model detects classes by comparing recall and false real detection rates across the limiting value . More AUCs mean better performance.

True Positive Rate (TPR): Real detection / Real detection + Fake detection.

The fake real detection rate is the sum of the fake real detection and original fake detection. ROC Curve Algorithm (Binary Classification)

3.5 Results

3.5.1 Alzheimer's disease: We have trained six machine learning approaches using the Alzheimer dataset in the form of csv file and their performance indicators were presented through bar graphs, a comparison table and heatmaps. Gradient Boosting exhibited the highest performance metrics, obtaining a good accuracy value of 95.58\%, with precision of 95. 61\%, recall score of 95.58\%, and an F1 score of 95.59\%. Random Forest classifier also performed well, with an accuracy of 93.49\%. Moderate performance was observed with linear discriminant analysis (81.16\%), logistic regression (81.63\%), and support vector machine (83.02\%). Although Naive Bayes recorded the lowest metrics, it still achieved a respectable accuracy of 76. 98\%, which is acceptable to some existing models. The results obtained after training all the models are shown in Table 1. The correlation and effectiveness of each model across the four evaluation metrics were effectively visualized using the heat map and the bar graph, which is shown in Figures 4 and 5.

Model	Accuracy	Precision	F1 Score	Recall
SVM	0.83	0.82	0.83	0.82
LR	0.81	8.81	0.81	0.81
LDA	0.81	0.81	0.81	0.81
RF	0.93	0.93	0.93	0.93

Naive bayes	0.76	0.77	0.76	0.77
Gradient boost	0.95	0.95	0.95	0.95

3.5.2. Brain Stroke: We enhanced performance measurements with brain stroke datasets and modified machine learning models. Gradient Boosting model outperformed others with high metrics and 90.51\% accuracy during evaluation. Linear Discriminant Analysis and Random Forest models produced balanced results with 80.64\% and 88.16\% accuracy. The dataset's class imbalance hurt Naive Bayes. The heatmap, bar graphs, and comparison table below highlight these models' performance differences.

Model	Accuracy	Precision	Recall	F1 Score
LR	0.84	0.14	0.44	0.21
SVM	0.86	0.11	0.28	0.16
RF	0.88	0.14	0.28	0.18
Naive bayes	0.39	0.06	0.88	0.12
LD	0.85	0.19	0.44	0.26
Gradient boost	0.90	0.20	0.32	0.24

3.5.3.Depression: The depression dataset showed the best accuracy of 99\% in logistic regression, suggesting faultless categorization. Gradient boost achieved amazing 97.5\% accuracy, followed by SVM (96.25\%) and Random Forest (95\%). Improved accuracy of 92.05\% is achieved with Naïve Bayes. The figure 08 shows heatmaps of all models , Table 3 shows comparison tables, and figure 09 shows bar graphs of all the six trained models results . These results are further used in comparision with all other diseases. The bar graph shows the comparision between all the trained models results. Futher these results are compared with other four diseases result.

Model	Accuracy	Precision	Recall	F1 Score
LR	1.00	1.00	1.00	1.00
SVM	0.96	0.95	1.00	0.97
RF	0.95	0.94	1.00	0.97
Naive bayes	0.92	0.92	1.00	0.95
XGradient boost	0.97	0.97	1.00	0.98

3.5.4.Epilepsy: Random Forest , gradient-boosting and SVM perform well after model training. Random Forest achieved the highest performance metrics with 97\% accuracy. SVM and Gradient Boosting obtained excellent accuracy of 97\% and 96.65\%. Naive Bayes was balanced, while LR and LDA were more precise. Machine learning algorithms had the top metrics. The accuracy , precision , F1 score and recall value obtained after training model is shown in Table 04. Along with this their heatmap is shown in figure 10 and figure 11 represents the bar graph of all models.

Model	Accuracy	Precision	Recall	F1 Score
LR	0.81	0.98	0.10	0.19
LDA	0.82	0.91	0.13	0.23
RF	0.97	0.95	0.94	0.94
Gradient boost	0.96	0.96	0.86	0.91
Naive bayes	0.96	0.90	0.91	0.91
SVM	0.97	0.96	0.92	0.94

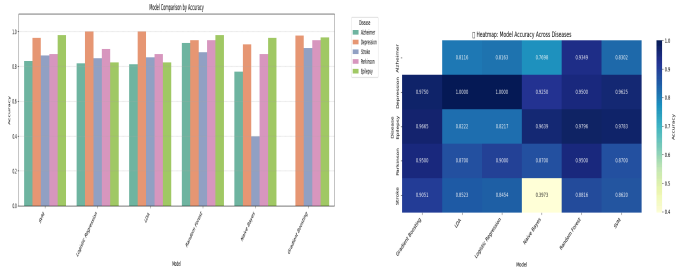
3.5.5.Parkison Disease: The accuracy of Random Forest and XGboost performed is 95\% each . LR with 90\% accuracy was among many other successful models. Naive Bayes and LDA successfully improved accuracy to 87\% apiece. The SVM accuracy was poor at 87\%. The calculated results are shown in Table 05, 12 and 13.

Model	Accuracy	Precision	Recall	F1 Score

SVM	0.87	0.91	0.94	0.92
LR	0.90	0.89	1.00	0.94
RF	0.95	0.94	1.00	0.97
Naive bayes	0.87	0.89	0.97	0.93
Gradient boost	0.95	0.94	1.00	0.97
LDA	0.87	0.89	0.97	0.93

Conclusion

Ensemble techniques outperformed traditional classifiers in our broad examination of machine learning models across datasets and intelligence prediction tasks. Gradient Boosting and Random forest perform best. Every phase had excellent accuracy, precision, recall, and F1 score. Brain stroke gradient classification with balanced parameters showed 95\% accuracy with boosting. In other datasets, LR had perfect or near-perfect metrics, but class imbalance caused low recall scores. Like LDA LR, advanced model evaluation has precision but recall. Logistic Regression and LDA are superior for certain datasets, while Random Forest and Boost are better for medical classification and prediction. AI will help in future for better detection of neuroimaging patterns which will help in early detection of diseases. These findings may support ensemble based model development in reliable healthcare and decision-making systems with dynamic adjustments which can improve outcome of treatment . It is more precise and follows ethical data integration .



Reference

[1] I. Hussain, “Empowering healthcare: AI, ML, and deep learning innovations for brain and heart health,” *Artificial Intelligence and Machine Learning Frontiers*, vol. 1, no. 8, 2024.

[2] A. Khanna, P. Singh, I. Kaur, *et al.*, “Predictive analytics for cardiovascular health: A machine learning approach,” in *Proc. 2024 Int. Conf. Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*, IEEE, 2024, pp. 1–6.

[3] S. Kukkar, “Secure VoIP call on Android platform,” *Global Journal of Computer Science and Technology – E: Network, Web & Security*, vol. 12, no. 12, 2012.

[4] S. Kukkar and J. Singh, “Biomedical mammography image classification by patches-based feature engineering using deep learning with ensemble classifier,” in *AIP Conf. Proc.*, vol. 3072, AIP Publishing, 2024.

[5] S. J. X. Murphy and D. J. Werring, “Stroke: Causes and clinical features,” *Medicine*, vol. 48, no. 9, pp. 561–566, 2020.

[6] S. Saturi, “Review on machine learning techniques for medical data classification and disease diagnosis,” *Regenerative Engineering and Translational Medicine*, vol. 9, no. 2, pp. 141–164, 2023.

[7] J. Singh *et al.*, “Mammography image abnormalities detection and classification by deep learning with extreme learner,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023.

[8] R. D. Thijs, R. Surges, T. J. O’Brien, and J. W. Sander, “Epilepsy in adults,” *The Lancet*, vol. 393, no. 10172, pp. 689–701, 2019.

[9] N. D. Volkow, G. F. Koob, and A. T. McLellan, “Neurobiologic advances from the brain disease model of addiction,” *New England Journal of Medicine*, vol. 374, no. 4, pp. 363–371, 2016.