GMR
The Original

BENCHMARKING MACHINE LEARNING CLASSIFIERS FOR HOSPITAL READMISSION PREDICTION IN DIABETES CARE

Amit Kumar Gupta[1] Arun Choudhary[2]

[1] Computer Science & Engineering, HRIT University, Ghaziabad (U.P.), India, 240151005@hrituniversity.edu.in
[2]Dean (Industrial Relations), HRIT University, Ghaziabad (U.P.), India, deanir@hrituniversity.edu.in

ABSTRACT:
Hospital readmission is a serious problem in healthcare systems as it leads to higher treatment costs, resource use and patient morbidity. Identifying patients who are likely to be readmitted can help guide prompt action by clinical staff and enhance health outcomes. In this study, a comprehensive benchmarking analysis of supervised machine learning classifiers for diabetic patients is presented. The data pre-processing steps comprised of missing value handling, drop of identifier attributes, categorical feature encoding, and binary target generation. In order to overcome class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was used only for the training set, and the original test set was kept imbalanced to evaluate the algorithm in an unbiased manner. The performance of eight machine learning classifiers- Logistic Regression, Decision Tree, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, XGBoost and LightGBM was assessed. The experimental results show the superiority of ensemble learning approaches over the conventional classifiers in predicting hospital readmission. Extra Trees with 86.54% classification accuracy obtained the best ROC-AUC 0.6079 among models and with the training time of 1.87 s, it shows a balance between the predictive power and computational efficiency. Random Forest had the highest precision (21.58%), AdaBoost had the highest recall (33.42%) and F1 score (0.1919), Logistic Regression had the highest balanced accuracy (53.35%) and XGBoost had the highest Matthews correlation coefficient (0.0759). All the findings suggest that no one classifier was superior in all the evaluation measures, which illustrates the need for multi-metric evaluation in designing predictive models for imbalanced healthcare datasets. This benchmark study offers an experimental framework and hands-on guidance on selecting appropriate machine learning models for hospital-readmission predictions in diabetes management.

KEYWORDS: Diabetes Management, Machine Learning, Electronic Health Records, SMOTE Technique, LightGBM, Benchmarking, Clinical Decision Support.

1. INTRODUCTION

High prevalence of hospital readmission is now a major problem in the health sector as it has a negative impact on patient health, costs on healthcare resources and also on the healthcare expenses. Many readmissions, especially those within 30 days of the initial discharge, could be prevented with appropriate post discharge care, timely clinical interventions, and effective patient monitoring. Therefore, the ability to predict hospital readmission is amongst the important research works in the medical decision support systems and medical analytics. Diabetic patients often need multiple hospitalizations because of the progression of the disease, comorbid conditions, side effects of their medications, and sub-optimal glycemic control. The frequent admissions have a negative effect on the quality of life for patients, hospital efficiency and treatment cost. Thus, detecting patients at risk of readmission prior to discharge can help healthcare providers to fine-tune intervention strategies and enhance patient management over the long haul. Electronic Health Records (EHRs), has huge amounts of clinical data which are now available to be used for predictive analytics. These records include useful demographic data, lab tests, medication lists, diagnoses, admission information, and treatment-related variables, which combine to offer a full profile of patient health. These diverse data sources have now been made possible by AI/ML to predict clinical outcomes, disease diagnosis, patient deterioration, mortality risk, and hospital readmission. In the field of healthcare, machine learning methods have shown great potential for automatically detecting intricate nonlinear relationships in vast clinical datasets. These complex interactions are hard to capture using traditional statistical modelling, while machine learning models like Decision Trees, Random Forest, Gradient Boosting, XGBoost, LightGBM and Extra Trees can model data with higher dimensionality more accurately. Ensemble learning methods are gaining much attention in recent years because of their resistance to overfitting, low error rates and better generalization performance in a wide variety of healthcare applications. Although there has been an increasing number of studies that have examined hospital readmission prediction, there are some challenges that remain unaddressed. First, clinical data is usually very skewed with readmitted patients being a small percentage of the population. This imbalance may lead to overly optimistic classification performance figures and inaccurate identification of the patients with genuine risk of readmission. Therefore, the use of accuracy alone can be misleading to assess model performance. For imbalanced health datasets, classifiers need to be evaluated using more informative metrics. One of the challenges of previous research is the lack of uniformity in comparing machine learning algorithms. Comparing the various investigations is challenging, as many only explored one

or two predictive models or utilize different pre-processing. Moreover, if the problem of class imbalance and data leakage is not handled properly during model building, then the model performance estimates will be excessively optimistic. It is therefore necessary to have a reproducible benchmarking framework, based on a shared data set, a consistent preprocessing pipeline, standardized evaluation metrics and identical experimental settings to compare the performance of classifiers objectively. In this work, we address these challenges and provide a thorough benchmarking analysis of supervised machine learning classifiers for predicting 30-day hospital readmission for diabetic patients. Experimental evaluation was done on publicly available Diabetes 130-US Hospitals dataset which contains 101,766 patient records. A systematic preprocessing pipeline was developed, which involved the treatment of missing values, the elimination of the identifier attributes, the encoding of categorical features and the creation of binary targets. The Synthetic Minority Oversampling Technique (SMOTE) was used only on the training data set to overcome class imbalance, and the original imbalanced testing set was kept to evaluate the performance of the model set for testing. A comprehensive assessment of predictive capability and effectiveness of computation was conducted by assessing them with multiple performance metrics for eight widely used supervised machine learning classifiers.

This study has contributed to the literature in the following ways:

- A large real world diabetes hospital readmission dataset is used to create a benchmarking framework using various supervised machine learning classifiers.
- Systematic preprocessing strategy including handling of missing values, categorical features, class imbalance treatment (SMOTE) is implemented without data leakage.
- The relative analysis reveals the strengths and weaknesses of the most widely used machine learning algorithms for hospital readmission prediction, and offers practical insights into the selection of the appropriate predictive models in diabetes care.

The remaining section of paper is organized as follows. Section 2 summarizes studies on hospital readmission prediction using machine learning techniques in recent years. The dataset, data pre-processing, machine learning classifiers and evaluation metrics are discussed in Section 3. The experimental results and comparison are given in Section 4. In the last section, the paper is summarized and directions for future research are proposed.

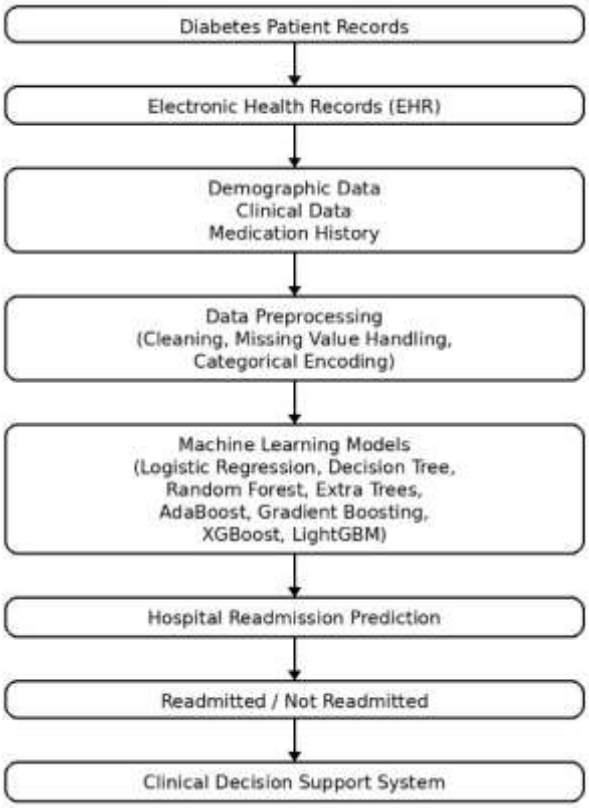

Figure 1. General framework for machine learning-based hospital readmission prediction using electronic health records.

2. RELATED WORK

AI has transformed modern healthcare, with their contributions to clinical decision support, disease diagnosis, patient monitoring, medical image analysis, and outcome prediction. With the swift accumulation of healthcare data from various sources like electronic health records, wearable devices, diagnostic imaging technologies, and laboratory analyses, researchers have been driven to create smart predictive models that can assist healthcare practitioners in making timely and accurate clinical decisions. As a result, machine learning has become one of the most promising technologies in the

field of health care quality improvement, reducing diagnostic errors, and enabling personalized medicine [1]-[7]. Machine learning methods have shown their applicability in a wide range of healthcare fields, in recent studies. To monitor a patient's health in an IoT environment with the aim of detecting diseases, Verma et al. [1] presented an effective healthcare monitoring system based on intelligent prediction model and secure communication. They emphasized the critical significance of integrating healthcare monitoring systems with machine learning methods, emphasizing the potential for enhancing diagnostic accuracy and supporting remote patient care. In the same manner, Vidushi et al. [2] explored the prediction of Alzheimer's disease by applying the supervised learning algorithms and they showed that machine learning models could be useful for clinicians in determining neurological disorders from clinical data. Rajak et al. [3] conducted a comparative analysis of various classification algorithms based on machine learning techniques for predicting student performance, highlighting the need for multiple classifiers to be benchmarked instead of a single predictive model. In the healthcare sector, Rajak et al. [4] compared various supervised machine learning classifiers for predicting TB and showed that alternative classifiers have different strengths and weaknesses for different measures. It is concluded from these studies that comparative evaluation is a key element in choosing a predictive model for healthcare applications. Medical image analysis is one of the fastest growing fields of artificial intelligence in medicine. In particular, brain tumor diagnosis has been a topic of great interest as accurate segmentation of the regions in a tumor directly affects treatment planning and prognosis. Prastawa [8] was one of the first to propose an automatic brain tumor segmentation framework using outlier detection techniques, laying the groundwork for future deep learning models. With the advent of convolutional neural networks (CNNs), researchers were now able to create fully convolutional architectures with the ability to learn hierarchical image representations directly from magnetic resonance imaging (MRI) data. Zhao et al. [9] added fully convolutional neural networks and conditional random fields to boost segmentation accuracy, while Li et al. [10] designed an end-to-end segmentation system for further boosting segmentation performance. In recent years, there have been major advances in deep learning techniques for brain tumor analysis and several review articles summarizing these advances and outlining the growing use of federated learning, attention mechanisms, and transformer models in medical image segmentation [11]-[14]. Automated medical image segmentation has been greatly enhanced with the advent of encoder-decoder architectures like U-Net. Several modified U-Net variants have been suggested for better segmentation performance by adding attention mechanisms, EfficientNet encoders, residual connections and multi-scale feature extraction to reduce the computational complexity [15]-[24]. The investigations show that the state-of-the-art deep learning models can be used to learn very discriminative image features, without the need for manual feature engineering. In addition, recent review papers highlight that model generalization, explainability and computational efficiency still constitute relevant research challenges in medical image analysis, despite the amazing success of deep learning approaches in this field [10]-[24]. Automated analysis of chest X-rays is another rapidly emerging field of AI in healthcare. Pneumonia detection, tuberculosis detection, and chest X-ray-based detection of pulmonary disease are just some of the many applications deep learning techniques have found for detecting disease from medical images. Transfer learning was used successfully for pneumonia detection by Chouhan et al. [25] and Vision Transformer (ViT) was introduced by Dosovitskiy et al. [26] and had a significant impact on the development of medical image analysis models using transformer. Later research has investigated how to leverage convolutional vision transformer, attention mechanisms, and hybrid deep learning architectures for chest disease detection, and all these studies have found that these methods are more effective at predicting the disease than conventional CNN models for chest disease classification [27]-[30]. With the publication of large public benchmark datasets like ChestX-Ray8, automated diagnosis of thoracic diseases has been boosted further by allowing objective comparison of machine learning and deep learning algorithms [31]. Recent studies still have shown the efficacy of transformer-based models for chest radiograph analysis and their promise for clinical use [32]-[41]. EEG-based healthcare applications are also a field of remarkable success by machine learning. EEG signals are extensively used in brain-computer interface, diagnosis of neurological disorders, motor imagery classification, emotion recognition, and assessment of cognitive status. In recent years, CNN, transformer-based architectures, transfer learning and domain adaptation approaches have been proposed to enhance the performance of cross-subject EEG classification [42]-[50]. Bagchi and Bathula [42] proposed EEG-ConvTransformer, a CNN based approach with transformer structures for visual stimulus classification, while Zhong et al. [43] proposed a multi-source domain generalization framework for the classification of motor imagery EEG. One of the most popular compact CNN models used for EEG signal analysis is EEGNet presented by Lawhern et al. [44] The recent transfer learning frameworks have also made an effort to overcome the inter-subject variability and improve the generalization of the model to different EEG datasets [46], [50].

In the healthcare sector, AI systems designed to make clinical decisions must also be interpretable, as this is an additional crucial component of the field. Explainable Artificial Intelligence (XAI) methods try to make the model predictions interpretable and gives an explanation about the features, so as to give the clinicians more confidence in the AI supported decision support systems. Ali et al. [45] provided an extensive review of the methods for XAI and highlighted the need for transparency, fairness, and accountability to ensure trustworthy health care AI systems. Likewise, the study by Sturm et al. [47] explored explainable deep learning models for EEG classification and showed how explanatory power could help yield clinically relevant insights as well as prediction performance. The results of these studies point towards the need for future healthcare applications to focus on model interpretability and predictive performance to enable clinical use. Despite the achievement of many healthcare-related applications of machine learning, the selection of the most suitable algorithm is a difficult one, due to the fact that the predictive power depends upon the characteristic of the data, the representation of the features and the distribution of the classes. Comparative benchmarking studies are thus very useful to objectively compare different classifiers in similar experimental environments. Recent studies have shown an increasing preference to use multiple complementary performance measures to gain a balanced understanding of the behavior of classifiers, such as accuracy, precision, recall, F1-score, ROC-AUC, and computational efficiency [3, 4, 34].

Benchmarking studies also increase the reproducibility by providing standardization in preprocessing pipelines, uniform validation procedures, and transparent experimental protocols, which will enable fair comparisons of different predictive models. In this regard, the present study conducts a thorough benchmarking comparison of eight popular supervised machine learning classifiers for forecasting 30-day hospital readmission for diabetic patients. To overcome class imbalance in the training data, we used the Synthetic Minority Over-sampling Technique (SMOTE), and evaluate the models by using the original independent test set to provide unbiased evaluation. Classifiers are compared to each other by multiple metrics. The proposed benchmarking framework brings in a repeatable experimental methodology and provides practical guidance on what machine learning algorithms can be used for predictive healthcare analytics.

3. MATERIALS AND METHODS

3.1 Dataset

Publicly available dataset of electronic health records for the Diabetes 130 Hospitals dataset, which has ten years of records from 130-hospitals in the US. The data comprises demographic details, laboratory results, medications, hospital admissions, diagnosis information, and hospital readmission status of diabetic patients. The initial dataset consisted of 101,766 patient records having 50 attributes. Three of the identifier attributes, which were not used in predictive modeling, were eliminated in the pre-processing stage, leaving 47 predictor variables. The attribute was the actual readmitted attribute and was converted to a binary classification problem, with patients who were readmitted within 30 days classified as positive and the rest classified as negative. This dataset was highly skewed with about 88.84% of the records in the non-readmission class, and 11.16% of records in the readmission class. This imbalance drove the need to adopt an oversampling approach in model development.

3.2 Data Preprocessing

A data processing pipeline was in place to enhance the data quality and consistency across the whole experimental study. At first, all the symbols representing missing values (the "?") were substituted with unknown values. Appropriate pre-processing was performed on features that contained missing data (weight, payer_code, medical_specialty, race, diag_1, diag_2, diag_3, A1Cresult, max_glu_serum). Patient_nbr and Original readmitted attribute were removed from the data as they did not give any meaningful predictive information. Afterwards, all categorical variables were converted to numerical values applying label encoding for using them with machine learning algorithms. After preprocessing, the data was split into training and testing sets in the ratio of 80 and 20, maintaining the same class distribution of each set. The training set contained 81,412 patient records and the independent test set contained 20,354 patient records.

3.3 Handling Class Imbalance

Hospital readmission prediction is an imbalanced classification problem since the number of patients who are readmitted within 30 days is small. Directly trained machine learning models are prone to overfitting to the majority class and have suboptimal performance in identifying high risk patients. To tackle this problem, SMOTE algorithm is used only on the training data set. The main idea of SMOTE is to create synthetic minority class samples by using the existing minorities' samples instead of duplicating them, so as to achieve a balanced distribution of the minority and majority classes. The independent test set was not used for SMOTE as it did not affect the class distribution during the evaluation process, thus preventing data leakage. Following oversampling, a balanced training set was obtained with 144,652 records, with 72,326 records per class.

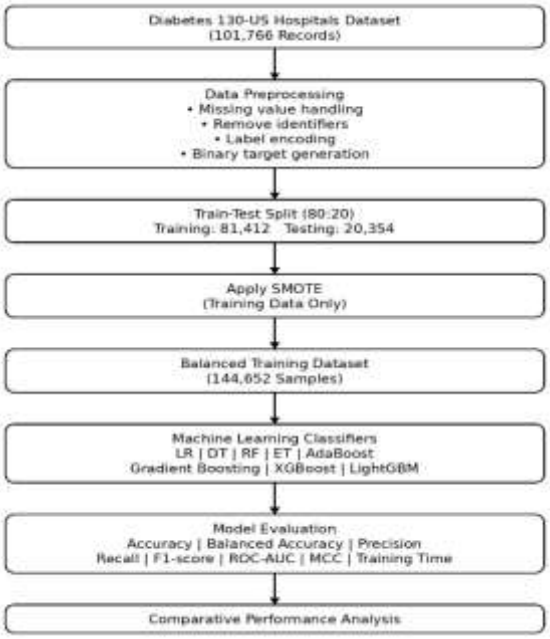

Figure 2. Workflow of the proposed benchmarking framework for hospital readmission prediction using machine learning.

3.4 Machine Learning Classifiers

Eight machine learning classifiers were chosen to be compared, all supervised, and of different learning paradigm. Logistic Regression (LR): A linear classification model that is used as a baseline due to its simplicity, computational efficiency, and interpretability. Decision Tree (DT): A recursive partitioning based on tree structure in the feature space to create decision rules for patient classification. Random Forest (RF): Ensemble learning technique which is formed by aggregating multiple decision trees using the bootstrap aggregation strategy to increase the stability of the prediction and mitigate overfitting. Ensemble classifiers with additional randomness during the construction of trees: Extra Trees (ET). AdaBoost (AB): Boosting algorithm that iteratively adds weak learners, putting more focus on misclassified samples. - Gradient Boosting (GB): A method of ensemble learning that sequentially builds-up decision trees to reduce the prediction error. Extreme Gradient Boosting (XGBoost): An optimized version of gradient boosting that includes regularization and fast parallel processing. Histogram-based gradient boosting algorithm for scalable and efficient learning on large datasets: Light Gradient Boosting Machine (LightGBM). Experimental reproducibility was achieved by implementing all classifiers with a fixed random seed using machine learning libraries that are widely used.

3.5 Experimental Setup

The experiments have been carried out with Python programming language in the Kaggle Notebook platform. Commonly-used machine learning libraries such as Pandas, NumPy, Scikit-learn, LightGBM, XGBoost, and Imbalanced-learn were used for data preprocessing and implementation of the model. After implementing SMOTE, the training data set was balanced and each classifier was trained with the balanced training dataset. The performance of models was quantified on the original independent test set to measure the generalization ability of the model in realistic clinical settings. These complementary metrics give a better idea of the predictive performance than classification accuracy only, in the era of imbalanced healthcare datasets.

3.6 Workflow of the Proposed Benchmarking Framework

The proposed Benchmarking Framework is shown in the following diagram. The general procedure followed in this study is shown in figure 2. This process starts with data collection from the Diabetes 130-US Hospitals dataset which is publicly available. Data preprocessing steps involve missing value treatment, identification of attribute identifiers removal, encoding of categorical features, and stratified train-test splitting. SMOTE is then only used on the training data to achieve class balance. A balanced training set is then created for training eight supervised machine learning classifiers. Finally, the trained models are tested on the original independent test set, and their predictions are analyzed with several evaluation metrics and by graphical visualization.

4. RESULTS AND DISCUSSION

4.1 Experimental Results

The Synthetic Minority Over-sampling Technique (SMOTE method was used to tackle this class imbalance issue exclusively in the training data, and the imbalanced test data set was not used to balance the data set, but kept aside for performance evaluation. This approach ensures that data leakage is avoided and gives a realistic view of the model's generalization power to unseen clinical data. Accuracy, Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), Matthews Correlation Coefficient (MCC) and Training Time were used as the classifiers' evaluation metrics. Results of all the classifiers are shown in Table 1.

Table 1. Performance comparison of machine learning classifiers for hospital readmission prediction.

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1-score	ROC-AUC	MCC	Training Time (s)
Extra Trees	0.8654	0.5204	0.2125	0.0762	0.1122	0.5903	0.0654	9.41
Random Forest	0.8623	0.5242	0.2158	0.0889	0.1260	0.6057	0.0727	10.27
LightGBM	0.8554	0.5274	0.2078	0.1052	0.1397	0.6079	0.0748	1.87
XGBoost	0.8525	0.5289	0.2053	0.1123	0.1452	0.6038	0.0759	1.68
Gradient Boosting	0.8091	0.5270	0.1577	0.1638	0.1607	0.5680	0.0530	27.88
Decision Tree	0.7521	0.5294	0.1422	0.2426	0.1793	0.5294	0.0471	1.95
Logistic Regression	0.7051	0.5335	0.1378	0.3126	0.1913	0.5451	0.0485	0.48
AdaBoost	0.6860	0.5322	0.1346	0.3342	0.1919	0.5489	0.0453	5.95

The experimental outcomes show that depending on evaluation metrics, the best classification performance is obtained by different classifiers. Extra Trees (86.54%) had the best classification accuracy, which meant that it was the most accurate in its classification of patients. But the classification accuracy was not sufficient to assess the predictive performance as the distribution of the independent test set was still imbalanced. LightGBM has the highest ROC-AUC (0.6079), showing better discriminative power than the other classifiers under all decision thresholds to distinguish patients who are readmitted and those who are not. Moreover, the model only took 1.87 seconds to train, making it efficient to train on large-scale clinical datasets. The highest precision (21.58%) was obtained by Random Forest, which means it predicted higher number of positive values as compared to the other classifiers. By contrast, AdaBoost performed best when it came to recall (33.42%) and F1-score (0.1919) and therefore was better at identifying who were at risk of hospital readmission.

Logistic Regression outperformed other models with the highest Balanced Accuracy (53.35%) and XGBoost recorded the highest Matthews Correlation Coefficient (0.0759) which means that the model had the best agreement with the actual class label. Gradient Boosting took the longest time to train (27.88 seconds), but still did not perform as well as LightGBM, Random Forest, and XGBoost. Likewise, the Decision Tree classifier had a relatively poor ability to predict hospital readmissions on most of the assessment measures, underscoring the benefits of ensemble learning models in predicting hospital readmissions. In summary, the benchmark results show that tree-based ensemble learning methods can outperform the standard machine learning classifiers in this clinical prediction problem in general. Extra Trees had the best classification performance and LightGBM showed the best discriminative performance based on ROC-AUC and it can be used in clinical decision support systems with reduced computation cost, which makes it a suitable model.

4.2 Analysis of Classification Performance

The classification accuracy of the machine learning classifiers used is compared in figure 3. The ensemble learning methods (Extra Trees, Random Forest, LightGBM and XGBoost) often outperformed the single tree and linear classifiers in terms of classification accuracy. The highest accuracy (86.54%) was achieved by Extra Trees, while Logistic Regression and AdaBoost gave relatively poor accuracy. The ROC-AUC comparison of the classifiers evaluated is shown in figure 4. The result of Extra Trees classification accuracy is best, but LightGBM has the best ROC-AUC (0.6079) which shows that LightGBM has the highest discrimination ability in each classification threshold. The result shows that classifiers with a slightly lower overall accuracy can have a higher accuracy in the aspect of ranking the patients for the purpose of identifying the high risk patient. The results of the comparison of F1-scores are shown in figure 5. AdaBoost had the best F1 value, as it had relatively high recall rates, while the ensemble of tree-like classifiers had relatively high classification accuracy and precision, with comparatively low F1 values on an imbalanced test set. The computational efficiency of the classifiers evaluated is compared in Figure 6. Logistic Regression, XGBoost and LightGBM took 0.48 seconds, 1.68 seconds and 1.87 seconds to train respectively. LightGBM showed the best performance in both terms of computational efficiency and predictive performance among all the ensemble learning methods mentioned, with Gradient Boosting taking the longest training time of 27.88 seconds. The ROC-AUC comparison of the classifiers evaluated is provided in figure 4. Extra Trees had the maximum classification accuracy, but LightGBM had the best ROC-AUC (0.6079), which means it was more discriminative over a range of classification thresholds. This observation shows that classifiers with a slightly lower overall accuracy can actually give a better ranking performance to identify high-risk patients.

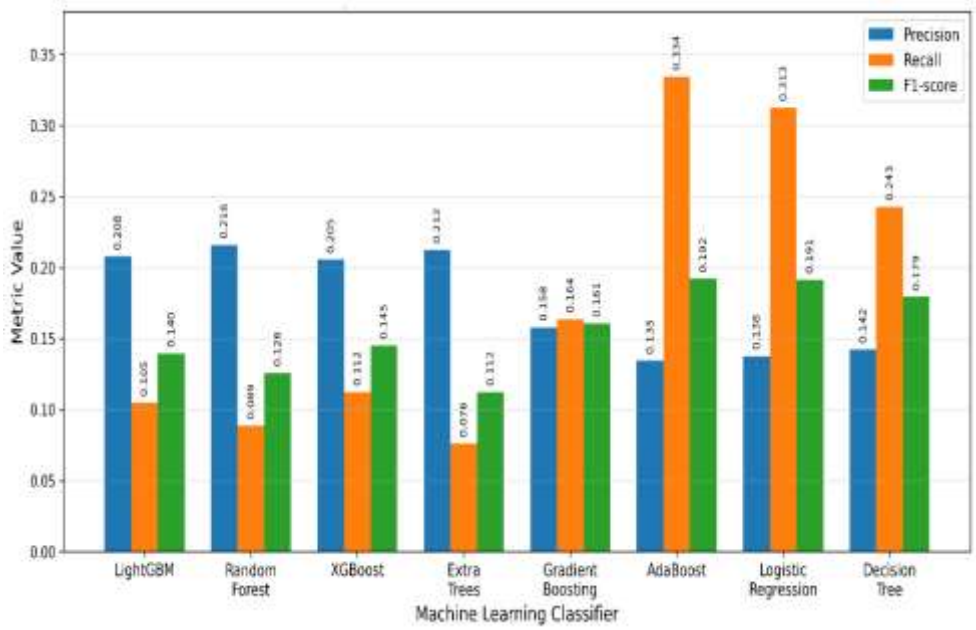

Figure 3. Comparisons of precision, recall and F1-Score of the machine learning classifiers evaluated for predicting hospital readmission.

The comparative analysis of all the machine learning classifiers evaluated in this study is shown in Figure 3. The model with the highest precision was Random Forest (21.58%), providing the most accurate positive predictions when compared to the other models. AdaBoost achieved the best recall score (33.42%) and the highest F1-score (0.1919) showing its best capability in determining patients who were readmitted for 30 days. Logistic Regression also had a relatively high recall but not so high precision. Ensemble models like Extra Trees and LightGBM, on the other hand, focused on the classification accuracy and precision and thus performed worse in terms of recall. The observations highlight the fact that classifiers focus on different aspects of predictive performance, thus calling for the use of a selection of complementary metrics in evaluating healthcare prediction models, rather than classification accuracy alone.

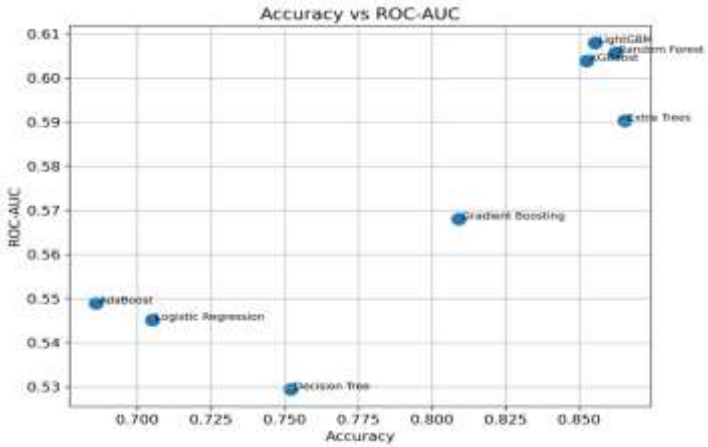

Figure 4. Scatter plot illustrating the relationship between classification accuracy and ROC-AUC for the evaluated machine learning classifiers.

The result of the classification accuracy versus ROC-AUC for classifiers evaluated by the study can be seen in figure 4. The scatter plot shows that the classifiers that improve the accuracy of the classification do not necessarily have the highest discriminative ability. Extra Trees had the best accuracy (86.54%), but had the lowest ROC-AUC (0.5903) among the three. Compared to LightGBM, which had the highest ROC-AUC (0.6079) and competitive classification accuracy (85.54%), LightGBM demonstrated a superior overall analysis performance that also performed well in class discrimination. This distribution of classifiers further demonstrates that the ensemble learning methods consistently are in the upper right part of the plot, which signifies the superiority of ensemble learning methods compared to conventional machine learning algorithms employed to predict hospital readmissions.

4.3 DISCUSSION

The results of this experiment show some of the difficulties of predicting hospital readmission using structured electronic health records. SMOTE was used to balance the training data, but the test set was not changed, meaning that a more realistic assessment of clinical performance was obtained. Thus, high classification accuracy does not necessarily translate into good identification of patients who are at high risk for readmission. Most of the evaluation metrics showed that the tree-based ensemble learning methods consistently emerged as superior to the traditional machine learning algorithms. More specifically, LightGBM exhibited the best ROC-AUC with low computational cost making it a promising model for future practical application in clinical decision-support systems. Different ensemble methods have complementary strengths with Extra Trees having the highest classification accuracy, Random Forest having the highest precision and XGBoost having the highest MCC. The modest performance of precision, recall, and ROC-AUC in all the classifiers examined suggest that the challenge of predicting hospital readmission is not as straightforward as it might seem, as there are many complex interactions among demographic characteristics, laboratory measurements, medication history, comorbidities, and other clinical factors. The results indicate that structured electronic health records might not be enough to fully describe clinical data needed for an accurate prediction. The study's findings could also inspire future research into utilize temporal patient trends, multimodal healthcare information, explainable AI methods, and sophisticated deep learning frameworks to enhance predictive accuracy.

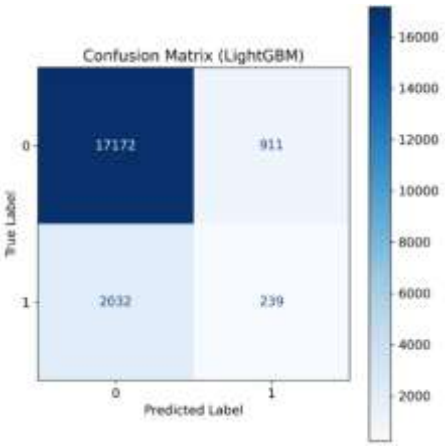

Figure 5. LightGBM Classification Results for the independent Test set on a Confusion Matrix.

The confusion matrix for the LightGBM classifier on the test data set, which is not used for training, is shown in Figure 5. The model correctly identified 17,172 patients who were not readmitted (true-negative group) and 239 patients who were readmitted (true-positive group). There were 911 patients not readmitted who were incorrectly classified as readmitted (false positive) and 2,032 patients readmitted who were misclassified as not readmitted (false negative). The

high rate of false negative cases is due to the challenge of detecting minority class samples in highly imbalanced healthcare datasets. However, the confusion matrix shows that LightGBM performs well on classification of the majority class and achieves the highest ROC-AUC among all the classifiers tested, thus its overall discriminative power is good.

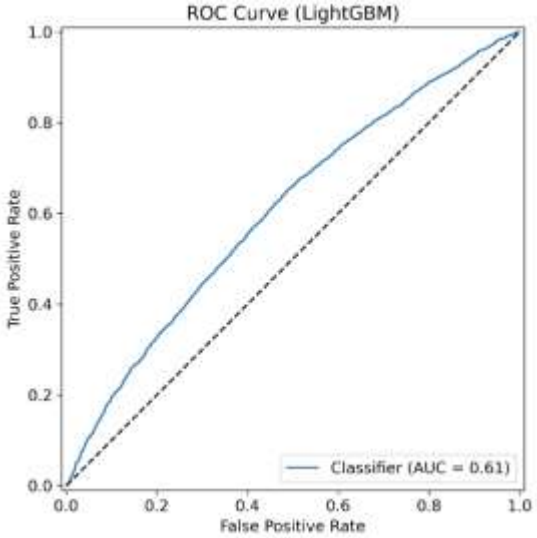

Figure 6. The LightGBM's Receiver Operating Characteristic (ROC) curve on the independent test set.

The Receiver Operating Characteristic (ROC) curve of the LightGBM Classifier on the Independent Test Dataset is shown in figure 6. The model has the highest ROC-AUC value of 0.6079, outperforming all other machine learning classifiers considered in this study. The ROC curve is always above the diagonal line, meaning that the classifier has a better performance than random guesses in distinguishing between readmitted and non-readmitted patients. While the moderate ROC-AUC is expected to be difficult to achieve, it is a natural consequence of the highly imbalanced EHR data. The superior ROC-AUC result achieved by LightGBM also shows its better discrimination power than the other classifiers and justifies its use in the prediction task of health sciences where the ability to rank the risk of the patient is crucial.

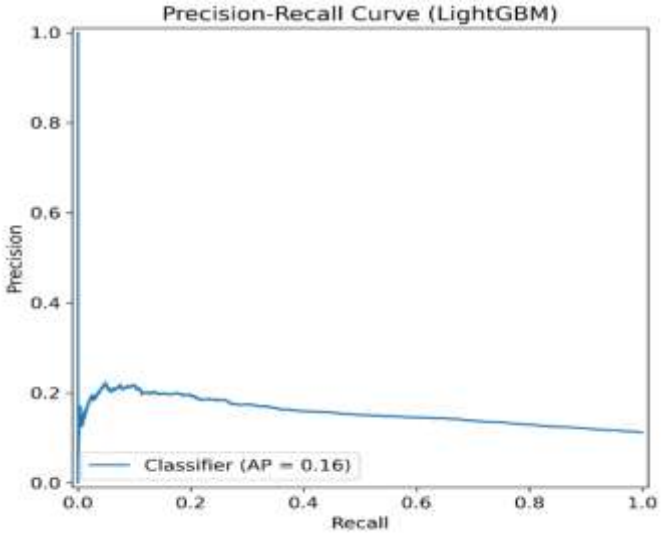

Figure 7. LightGBM (Light GBM) classifier is created for the independent testing set.

In identifying the minority class with a high degree of class imbalance, the model's performance is represented by the Average Precision (AP) score of ~0.16. The curve shows that recall rises while precision falls as more patients are identified, presenting a trade-off between identifying more patients who are readmitted and maintaining prediction precision (Figure 7) . The PR curve shows moderate performance of minority class, but the precision and recall values seen in this curve correspond with those seen in Table 1 for these minority classes. The findings demonstrate that predicting hospital readmission is a complex task and the importance of using the precision-recall graph as well as the traditional ROC-based graph to assess the performance of classifiers.

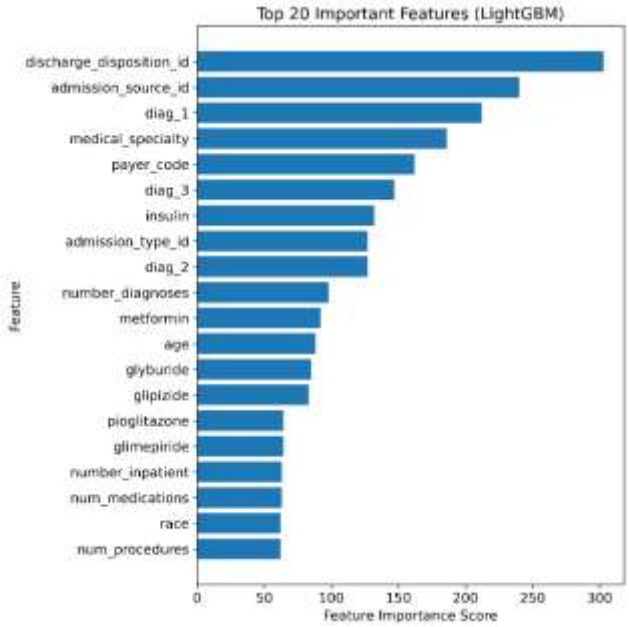

Figure 8. The 20 most significant features found by the LightGBM predictor for predicting hospital readmission.

The top 20 most influential features determined by feature importance scores are shown in Figure 8, obtained by the LightGBM classifier. The predictors that contributed the most to the prediction of the hospital readmission were discharge_disposition_id, admission_source_id, diag_1, and medical_specialty. Other clinical variables like insulin, number_diagnoses, metformin, and age also showed significant predictive role, suggesting that the patients' medication history, diagnosis-related information, and patients' age would be important factors for the readmission risk. The results indicate that some characteristics of the hospital stay, admission data and disease features are important factors for hospital re-admission. Additionally, the feature importance analysis helps in the interpretability of the LightGBM model by identifying the most influential features.

5. CONCLUSION AND FUTURE WORK

This study employed a thorough benchmarking study on eight supervised machine learning classifiers for predicting the 30-day hospital readmission for diabetic patients in the publicly available. A systematical preprocessing pipeline was developed, which involved missing value treatment, categorical feature encoding, removal of identifiers attributes, and class imbalance mitigation using Synthetic Minority Over-sampling Technique (SMOTE. SMOTE was only used on the training set and the original imbalanced test set was used to assess the performance of the models to make them fair. Experimental results showed that ensemble learning algorithms are always better than the conventional machine learning techniques. Extra Trees obtained the best classification accuracy (86.54%), while LightGBM gave the best ROC-AUC (0.6079) with just 1.87 seconds of training time. Random Forest had the highest Precision (21.58%), AdaBoost had the highest Recall (33.42%) and F1-score (0.1919), Logistic Regression had the highest Balanced Accuracy (53.35%) and XGBoost had the highest Matthews Correlation Coefficient (0.0759). These data indicate that none of the classifiers were superior to the others in all the evaluation parameters, underscoring the need to choose predictive models based on markers for the specific clinical goals instead of just on the overall classification accuracy. Another important point about the benchmark results is the difficulty in predicting hospital readmission using a structured electronic health record. While minimizing the imbalanced training set helped the model learn, it only achieved moderate discrimination performance on the original imbalanced test set, suggesting that predicting a patient's high risk of readmission is a difficult clinical prediction task. When working with imbalanced healthcare datasets, complementary evaluation metrics like ROC-AUC, Balanced Accuracy, Precision, Recall, F1-score, and MCC can be used to give a more comprehensive picture of classifier performance than accuracy can. Overall, this work sets a repeatable benchmarking process for the comparison of supervised machine learning algorithms for predicting hospital readmission, and gives insight into the merits and drawbacks of various classification approaches. The proposed approach can be a reference for comparison with other EHR data sets and can help researchers and health care practitioners in choosing the appropriate machine learning models for clinical decision support applications. Future studies could strive to incorporate EAI methods to make the models more interpretable and boost clinicians' confidence in the prediction results. Furthermore, more sophisticated feature engineering, feature selection methods, patient time series/course of treatment, multimodal health data, and deep learning architectures can enhance predictive accuracy. Lastly, external validation with data from various healthcare organizations would enhance the validity, applicability, and robustness of the proposed framework for benchmarking.

REFERENCES

1. Verma, A. K. Gupta, V. Kumar, A. Rajak, S. Kumar, and R. N. Panda, “A Secure Healthcare Monitoring System for Disease Diagnosis in the IoT Environment,” Multimed Tools Appl, vol. 84, no. 7, pp. 3767-3792, Apr. 2024, doi: 10.1007/s11042-024-19131-w.

2. Vidushi, M. Agarwal, A. Rajak, and A. K. Shrivastava, “Alzheimer Disease Prediction Using Learning Algorithms,” in Applied Computational Technologies, vol. 303, B. Iyer, T. Crick, and S.-L. Peng, Eds., in Smart Innovation, Systems and Technologies, vol. 303. , Singapore: Springer Nature Singapore, 2022, pp. 331-339. doi: 10.1007/978-981-19-2719-5_31.
a. Rajak, A. K. Shrivastava, and Vidushi, “Applying and comparing machine learning classification algorithms for predicting the results of students,” Journal of Discrete Mathematical Sciences and Cryptography, vol. 23, no. 2, pp. 419-427, Feb. 2020, doi: 10.1080/09720529.2020.1728895.
3. Rajak, R. N. Panda, A. Kumar, S. Bhardwaj, and D. P. Shrivastava, “Applying several machine learning models to tuberculosis dataset and evaluating the results based on models accuracy,” presented at the 1ST INTERNATIONAL CONFERENCE ON RECENT ADVANCEMENTS IN COMPUTING TECHNOLOGIES & ENGINEERING, Ghaziabad, India, 2024, p. 020035. doi: 10.1063/5.0217830.
4. Vidushi, M. Agarwal, A. Rajak, and A. K. Shrivastava, “Assessment of Optimizers impact on Image Recognition with Convolutional Neural Network to Adversarial Datasets,” J. Phys.: Conf. Ser., vol. 1998, no. 1, p. 012008, Aug. 2021, doi: 10.1088/1742-6596/1998/1/012008.
a. K. Shrivastava, A. Rajak, and S. Bhardwaj, “Detection of Tuberculosis based on Multiple Parameters using ANFIS,” in 2018 3rd International Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH), Ghaziabad, India: IEEE, Nov. 2018, pp. 1-5. doi: 10.1109/CIPECH.2018.8724255.
5. Rajak, “Neural Signatures of Alcoholism Revealed by Event-Related Potential Analysis of Open EEG Data,” JASTT, vol. 7, no. 1, pp. 157-167, Mar. 2026, doi: 10.38094/jastt71333.[1] M. Prastawa, “A brain tumor segmentation framework based on outlier detection,” Medical Image Analysis, vol. 8, no. 3, pp. 275-283, Sep. 2004, doi: 10.1016/j.media.2004.06.007.
6. X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, “A deep learning model integrating FCNNs and CRFs for brain tumor segmentation,” Medical Image Analysis, vol. 43, pp. 98-111, Jan. 2018, doi: 10.1016/j.media.2017.10.002.
7. H. Li, A. Li, and M. Wang, “A novel end-to-end brain tumor segmentation method using improved fully convolutional networks,” Computers in Biology and Medicine, vol. 108, pp. 150-160, May 2019, doi: 10.1016/j.compbiomed.2019.03.014.
8. F. J. Dorfner, J. B. Patel, J. Kalpathy-Cramer, E. R. Gerstner, and C. P. Bridge, “A review of deep learning for brain tumor analysis in MRI,” npj Precis. Onc., vol. 9, no. 1, p. 2, Jan. 2025, doi: 10.1038/s41698-024-00789-2.
9. Md. F. Ahamed et al., “A review on brain tumor segmentation based on deep learning methods with federated learning techniques,” Computerized Medical Imaging and Graphics, vol. 110, p. 102313, Dec. 2023, doi: 10.1016/j.compmedimag.2023.102313.
10. Y. M. A. Mohammed, S. El Garouani, and I. Jellouli, “A survey of methods for brain tumor segmentation-based MRI images,” Journal of Computational Design and Engineering, vol. 10, no. 1, pp. 266-293, Jan. 2023, doi: 10.1093/jcde/qwac141.
11. T. A. Fahim, F. B. Alam, and M. A. Hossain, “Brain tumor detection, classification and segmentation by deep learning models from MRI images: Recent approaches, challenges and future directions,” Array, vol. 28, p. 100571, Dec. 2025, doi: 10.1016/j.array.2025.100571.
12. S. Montaha, S. Azam, A. K. M. Rakibul Haque Rafid, Md. Z. Hasan, and A. Karim, “Brain Tumor Segmentation from 3D MRI Scans Using U-Net,” SN COMPUT. SCI., vol. 4, no. 4, p. 386, May 2023, doi: 10.1007/s42979-023-01854-6.
a. Arora, A. Jayal, M. Gupta, P. Mittal, and S. C. Satapathy, “Brain Tumor Segmentation of MRI Images Using Processed Image Driven U-Net Architecture,” Computers, vol. 10, no. 11, p. 139, Oct. 2021, doi: 10.3390/computers10110139.
13. M. Mostafa, M. Zakariah, and E. A. Aldakheel, “Brain Tumor Segmentation Using Deep Learning on MRI Images,” Diagnostics, vol. 13, no. 9, p. 1562, Apr. 2023, doi: 10.3390/diagnostics13091562.
14. P. R, J. P. P. M, and N. J S, “Brain tumor segmentation using multi-scale attention U-Net with EfficientNetB4 encoder for enhanced MRI analysis,” Sci Rep, vol. 15, no. 1, p. 9914, Mar. 2025, doi: 10.1038/s41598-025-94267-9.
15. H. Byeon et al., “Brain tumor segmentation using neuro-technology enabled intelligence-cascaded U-Net model,” Front. Comput. Neurosci., vol. 18, p. 1391025, Apr. 2024, doi: 10.3389/fncom.2024.1391025.
16. M. Havaei et al., “Brain tumor segmentation with Deep Neural Networks,” Medical Image Analysis, vol. 35, pp. 18-31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.
17. Aumente-Maestro, D. R. González, D. Martínez-Rego, and B. Remeseiro, “BTS U-Net: A data-driven approach to brain tumor segmentation through deep learning,” Biomedical Signal Processing and Control, vol. 104, p. 107490, Jun. 2025, doi: 10.1016/j.bspc.2025.107490.
18. J. Nalepa, M. Marcinkiewicz, and M. Kawulok, “Data Augmentation for Brain-Tumor Segmentation: A Review,” Front. Comput. Neurosci., vol. 13, p. 83, Dec. 2019, doi: 10.3389/fncom.2019.00083.
19. K. Avazov, S. Mirzakhalilov, S. Umirzakova, A. Abdusalomov, and Y. I. Cho, “Dynamic Focus on Tumor Boundaries: A Lightweight U-Net for MRI Brain Tumor Segmentation,” Bioengineering, vol. 11, no. 12, p. 1302, Dec. 2024, doi: 10.3390/bioengineering11121302.
20. K. Kamnitsas et al., “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” Medical Image Analysis, vol. 36, pp. 61-78, Feb. 2017, doi: 10.1016/j.media.2016.10.004.
21. M. Abrar, A. Salam, F. Ullah, F. Ullah, and A. S. Al Ghamdi, “Enhancing brain tumor segmentation using attention based convolutional UNet on MRI images,” Sci Rep, vol. 15, no. 1, p. 36603, Oct. 2025, doi: 10.1038/s41598-025-20329-7.
22. V. Chouhan et al., “A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images,” Applied Sciences, vol. 10, no. 2, p. 559, Jan. 2020, doi: 10.3390/app10020559.

a. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” 2020, arXiv. doi: 10.48550/ARXIV.2010.11929.
23. M. Usman, T. Zia, and A. Tariq, “Analyzing Transfer Learning of Vision Transformers for Interpreting Chest Radiography,” J Digit Imaging, vol. 35, no. 6, pp. 1445-1462, Dec. 2022, doi: 10.1007/s10278-022-00666-z.
24. Oltu, S. Güney, S. E. Yuksel, and B. Dengiz, “Automated classification of chest X-rays: a deep learning approach with attention mechanisms,” BMC Med Imaging, vol. 25, no. 1, p. 71, Mar. 2025, doi: 10.1186/s12880-025-01604-5.
25. O. Abueed, P. Thakkar, W. H. AlAlaween, Y. Wang, and M. T. Khasawneh, “Automatic semantic segmentation in chest X-ray images using deep learning approaches: a literature review,” Neural Comput & Applic, vol. 38, no. 4, p. 70, Feb. 2026, doi: 10.1007/s00521-025-11743-z.
26. Y. Vasilev et al., “Autonomous chest x-ray image classification, capabilities and prospects: rapid evidence assessment,” Front. Digit. Health, vol. 7, p. 1685771, Jan. 2026, doi: 10.3389/fdgth.2025.1685771.
27. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 3462-3471. doi: 10.1109/CVPR.2017.369.
28. H. Wu et al., “CvT: Introducing Convolutions to Vision Transformers,” in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 22-31. doi: 10.1109/ICCV48922.2021.00009.
29. K. D. Reddy and A. Patil, “CXR-MultiTaskNet a unified deep learning framework for joint disease localization and classification in chest radiographs,” Sci Rep, vol. 15, no. 1, p. 32022, Aug. 2025, doi: 10.1038/s41598-025-16669-z.
30. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, “Deep Convolutional Neural Networks in Medical Image Analysis: A Review,” Information, vol. 16, no. 3, p. 195, Mar. 2025, doi: 10.3390/info16030195.
31. P. Lakhani and B. Sundaram, “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks,” Radiology, vol. 284, no. 2, pp. 574-582, Aug. 2017, doi: 10.1148/radiol.2017162326.
32. K. Goswami, R. Kumar, R. Kumar, A. J. Reddy, and S. K. Goswami, “Deep Learning Classification of Tuberculosis Chest X-rays,” Cureus, Jul. 2023, doi: 10.7759/cureus.41583.
33. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 2261-2269. doi: 10.1109/CVPR.2017.243.
34. Ahmad, H. U. Rehman, B. Shah, F. Ali, and I. Hussain, “Dual-model approach for accurate chest disease detection using GViT and swin transformer V2,” Sci Rep, vol. 15, no. 1, p. 31717, Aug. 2025, doi: 10.1038/s41598-025-16422-6.
35. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, “Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning,” Diagnostics, vol. 10, no. 6, p. 417, Jun. 2020, doi: 10.3390/diagnostics10060417.
36. S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, “Efficient pneumonia detection using Vision Transformers on chest X-rays,” Sci Rep, vol. 14, no. 1, p. 2487, Jan. 2024, doi: 10.1038/s41598-024-52703-2.
37. G. Siddharth, A. Ambekar, and N. Jayakumar, “Enhanced CoAtNet based hybrid deep learning architecture for automated tuberculosis detection in human chest X-rays,” BMC Med Imaging, vol. 25, no. 1, p. 379, Sep. 2025, doi: 10.1186/s12880-025-01901-z.
38. S. Bagchi and D. R. Bathula, “EEG-ConvTransformer for single-trial EEG-based visual stimulus classification,” Pattern Recognition, vol. 129, p. 108757, Sep. 2022, doi: 10.1016/j.patcog.2022.108757.
39. X.-C. Zhong et al., “EEG-DG: A Multi-Source Domain Generalization Framework for Motor Imagery EEG Classification,” 2023, arXiv. doi: 10.48550/ARXIV.2311.05415.
40. V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, “EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces,” J. Neural Eng., vol. 15, no. 5, p. 056013, Oct. 2018, doi: 10.1088/1741-2552/aace8c.
41. S. Ali et al., “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” Information Fusion, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.
42. W. Lu, H. Liu, H. Ma, T.-P. Tan, and L. Xia, “Hybrid transfer learning strategy for cross-subject EEG emotion recognition,” Front. Hum. Neurosci., vol. 17, p. 1280241, Nov. 2023, doi: 10.3389/fnhum.2023.1280241.
43. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, “Interpretable deep neural networks for single-trial EEG classification,” Journal of Neuroscience Methods, vol. 274, pp. 141-145, Dec. 2016, doi: 10.1016/j.jneumeth.2016.10.008.
44. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. Muller, “Optimizing Spatial filters for Robust EEG Single-Trial Analysis,” IEEE Signal Process. Mag., vol. 25, no. 1, pp. 41-56, 2008, doi: 10.1109/MSP.2008.4408441.
45. H. Altaheri, G. Muhammad, and M. Alsulaiman, “Physics-Informed Attention Temporal Convolutional Network for EEG-Based Motor Imagery Classification,” IEEE Trans. Ind. Inf., vol. 19, no. 2, pp. 2249-2258, Feb. 2023, doi: 10.1109/TII.2022.3197419.
a. Bleuzé, J. Mattout, and M. Congedo, “Tangent space alignment: Transfer learning for Brain-Computer Interface,” Front. Hum. Neurosci., vol. 16, p. 1049985, Dec. 2022, doi: 10.3389/fnhum.2022.1049985.