

A Robust Multi-Class Classification Model for Pharma Healthcare Data Using Artificial Intelligence Techniques

Mr. Anoop Sunarty,

Assistant Professor, Department of Computer Sciences and Applications

Mandsaur University, Mandsaur

anoop.sunarty@meu.edu.in

Corresponding Author: *Anoop Sunarty (anoop.sunarty@meu.edu.in)

Received: 8 September 2026 | **Revised:** 13 September 2026 | **Accepted:** 28 September 2026

Abstract—Artificial Intelligence (AI) is rapidly becoming an integral part of healthcare systems across the globe. This is because AI can process large volumes of medical data and make more informed clinical decisions. In the healthcare sector, machine learning and deep learning algorithms are used in various applications, such as intelligent diagnostic systems, medical image analysis, pharmaceutical data processing, and illness prediction. This study presents a MIMIC-IV Dataset based Smart Classification Model for Pharma Healthcare Data. Different datasets preprocessing techniques have been used in the experimental analysis in order to improve the data quality and model performance. These methods comprised missing value imputation, normalization, one-hot encoding, outlier removal and data balancing using SMOTE. XGBoost achieved an impressive 96% accuracy (ACC), 94.6% precision (PRE), 91.6% recall (REC), and 97.7% F1-score (F1) out of all the models used. The outcomes of the proposed classification model based on XGBoost are reliable, accurate and efficient and looks extremely suitable for smart systems in the healthcare and medicine field as well as advanced systems in clinical decision making.

Keywords—Personalized Medicine, AI Ethics, Drug Discovery, Convolutional Neural Networks, Machine Learning, Pharmacological Research, Artificial Intelligence.

I. INTRODUCTION

Drug classification is a key concept in the pharmaceutical and health care industry because the classification is important for precise categorization of drugs in respect of patient condition and medical needs [1]. Traditionally, drug classification relied heavily on expert knowledge and manual analysis, which was often time-consuming and prone to human error. With the proliferation of healthcare data and its rapid evolution, automated intelligent systems are playing a crucial role in improving the PRE and productivity of drug decision making processes [2][3]. Healthcare information is generated from medical centers, clinical consultations, laboratory reports, and patient records and carrying valuable information which can be utilized for effective healthcare management and personalized treatment strategies [4][5].

The pharmaceutical industry is dedicated to advancing innovation and new technologies to address global healthcare issues and improve patient outcomes [6]. The development of new drugs, biologics, and personalized treatment solutions is a major research effort in manufacturing, packaging, safety assessment, and toxicity prediction [7]. Furthermore, the healthcare industry has to deal with issues like staffing shortages, skill deficits, drug side effects and rising patient expectations, where technology driven solutions are essential for efficient management [8][9]. Healthcare systems have been switched to Electronic Health Records (EHRs) [10] and wearable devices, genomics databases and diagnostic imaging have generated huge healthcare datasets which can be used for predictive and preventive healthcare applications [11][12].

Artificial Intelligence (AI) has emerged as a game-changer in the pharmaceutical sciences and healthcare analytics. Drug discovery, disease prediction, pharmacovigilance, clinical trial optimization, medical image, and personalized medicine are among the areas where AI methods, such as machine learning, deep learning, and natural language processing, have had a major impact [13][14][15][16]. These systems can leverage artificial intelligence to uncover patterns in vast biomedical datasets, forecast drug effectiveness, track side effects, and aid in making informed decisions during clinical practice. Furthermore, AI-driven predictive models can help to streamline healthcare resources by providing more accurate patient diagnosis, treatment recommendations, and live patient monitoring [17][18][19].

ML and DL can be used to design intelligent classification models for analyzing complex healthcare data [20]. The data is often high-dimensional and can have non-linear relationships among variables, making the use of traditional statistical and biostatistical analysis challenging in many medical applications [21]. These models are not directly applicable to learning meaningful features and derive patterns from structured and unstructured healthcare data automatically like ML and DL. Multi-class classification models are also important in pharmaceutical healthcare applications, where they can be used to classify multiple

classes of drugs, disease conditions or treatment recommendations using patient health attributes and clinical parameters. In real-world scenarios, particularly within healthcare, these classification methods can struggle with data imbalances, scalability, feature complexities, and prediction consistency when handling large datasets. These clever approaches have proved very effective for medical classification, prediction and decision making problems [22].

Hence, this paper suggests a robust multi-class classification model for pharma healthcare data using advanced Artificial Intelligence techniques that improves the classification ACC, the reliability of the model based on the pharma healthcare data [23], computational efficiency, and intelligent healthcare decision support systems. The proposed framework is intended to process complex pharmaceutical and healthcare data that able to detect meaningful patterns and enhance the predictability in various healthcare classes [24].

A. Motivation and Contribution

The increasing need for intelligent and accurate healthcare data classification systems in the context of pharmaceutical research and clinical decision-making can serve as a motivating factor for this study. As medical and pharmaceutical information grows quickly, manual and statistical techniques are unable to process complex data efficiently and effectively in the healthcare sector. AI-driven machine learning and DL algorithms offer more advanced features for recognizing complex patterns and enhancing predictive healthcare outcomes. Hence, the objectives of this research are to design an AI-based classification model which can improve the accurate prediction of healthcare, efficient and intelligent decision support applications. This research offers several key contributions as listed below:

- Utilized healthcare datasets from the MIMIC-IV database and Kaggle repositories containing patient clinical and pharmaceutical records.
- Compared several categorization models, such as XGBoost, RF, LR, NBs, LSTM, and SVM.
- Utilized the XGBoost model to attain a 96% classification ACC while simultaneously enhancing REC, PRE, and F1 values.
- Developed a smart AI-based classification framework for pharmaceutical healthcare data analysis.
- Demonstrated that the proposed framework provides reliable and efficient healthcare data classification for intelligent pharmaceutical and clinical decision-support systems.

The motivation behind this study is the rising need for intelligent and credible healthcare data classification systems to enhance pharmaceutical analysis and clinical choice making. Existing healthcare prediction approaches often face challenges in handling complex and high-dimensional medical datasets efficiently. The novelty of this work is the development of a smart classification framework that integrates multiple machine learning and deep learning models for the comprehensive analysis of healthcare data. The proposed XGBoost framework offers superior predictive power, lightning speed feature learning, and robust classification ability for intelligent healthcare applications in the pharmaceutical sector.

B. Organization of the Paper

This paper is structured as follows: Section II reviews related studies on AI-based pharmaceutical healthcare data classification. The methodology used and the experimental setup are detailed in Section III. These classifications are reported and commented on in section IV. Lastly, section V presents the conclusions, limitations and future directions of the research.

II. LITERATURE REVIEW

This research was supported with a comprehensive review of the latest research on Smart Classification Models for Pharma Healthcare Data. The models and main findings from previous studies are summarized in the proposed models, key findings and challenges in Table I.

Goddilla and Sivasangari (2025) looks at how these two machine learning methods can be used together to predict disease symptoms, with the goal of assisting early detection and enhancing healthcare outcomes. The performance, pros and cons of regression and classification approaches to symptom prediction are discussed and assessed. Developed predictive models that were able to capture the complexities of healthcare data by using a comprehensive dataset with medical records, demographic data, and clinical variables. Ensemble methods and feature engineering techniques were employed to improve the ACC of the model, and the best model, which was the RF, showed an ACC of 91.84% [25].

G. A et al. (2024) suggested a new approach to data analysis and prediction that combines Blockchain technology with artificial intelligence and machine learning techniques like SVM, K-means, and LSTM. This would improve healthcare organizations' pharmacovigilance processes. Safe and transparent data transfer is enabled at every stage of a drug's development by combining blockchain technology with ML algorithms. This includes early discovery as well as post-marketing surveillance. The proposed system improved patient outcomes and drug safety, where SVM achieved 85% ACC, K-means 90%, and LSTM 94% [26].

Hameed Shaker and Khalaf (2023) The goal is to come up with new ways to predict how people feel based on reviews written by patients in order to improve data analysis. A feature extraction method called TF-IDF is then employed. According to the results of the experiments, the RFC outperforms all other models found in the literature in terms of PRE, REC, F1, and ACC, hitting 93% [27].

Habib et al. (2022) provided a model that integrated ML with unsupervised learning. In order to formulate and validate the study hypothesis, the consolidation approach was useful in establishing a foundation. An 8-cluster solution was produced by the self-organizing map. Classification tasks were built from clustered data to further validate the posterior-probability multi-class SVMs performance. The elements that contribute to depression were identified as a result of the assumptions made by importance sampling. An improved classification performance of 91.16% was achieved after using the suggested model [28].

Bardalai et al. (2022) evaluated healthcare network features, contrasted machine learning-based prediction models and algorithms, and highlighted the challenge of accurately predicting smart healthcare performance metrics. established a model to predict the flow of two separate types of healthcare data using the tenets of Healthcare 4.0: data from

sensors used for remote monitoring, including measurement of oxygen saturation (SpO2) and blood pressure (BP), and audio/video data used for remote consultations. While NB provides superior ACC at 72% in forecasting the throughput, K-means has the greatest performance among the algorithms with 75.9% sensitivity and 72.68% ACC [29].

Ben Ali Kaddour and Abdulaziz (2021) utilized image processing algorithms to develop digital pathology, which aids pathologists in making more accurate diagnoses with less

room for human mistake. With the advent of digital healthcare, computer vision and AI methods like DL and ML algorithms were able to work together. An AI program that can automatically generate medical reports, secure patient data, and conduct pathological analysis for various diseases from medical photos was the goal of the project. With an F1 of 0.8332 and an ACC of 90.47%, the model used CNNs and morphological feature extraction to classify images of various types [30].

TABLE I. RECENT STUDIES ON SMART CLASSIFICATION MODEL FOR PHARMA HEALTHCARE DATA USING MACHINE LEARNING TECHNIQUES.

Author	Proposed Work	Dataset	Results	Limitations & Future Work
Goddilla and Sivasangari (2025)	Applied combined machine learning techniques for disease symptom prediction using regression and classification models with ensemble learning and feature engineering.	Medical records, demographic data, and clinical variables dataset	Random Forest achieved 91.84% accuracy.	Limited generalization across diverse healthcare datasets; future work can focus on real-time clinical deployment and advanced deep learning models.
G. A et al. (2024)	Integrated AI-based ML algorithms (SVM, K-means, LSTM) with Blockchain for pharmacovigilance and ADR prediction.	Pharmacovigilance and ADR healthcare dataset	SVM achieved 85%, K-means 90%, and LSTM 94% prediction accuracy.	Blockchain integration increases computational complexity, future work may improve scalability and security optimization.
Hameed, Shaker and Khalaf (2023)	Developed sentiment prediction models using patient reviews and TF-IDF feature extraction with machine learning classifiers.	Patient review and sentiment analysis dataset	Random Forest Classifier achieved 93% accuracy with improved Precision, Recall, and F1-score.	Performance depends on text quality and language diversity, future work can explore transformer-based NLP models.
Habib et al. (2022)	Proposed a hybrid unsupervised and supervised ML framework using Self-Organizing Maps and multi-class SVM for depression-related classification.	Clustered healthcare and depression-related dataset	Achieved 91.16% classification accuracy.	Limited dataset size and clustering complexity, future studies can use larger datasets and deep learning methods.
Bardalai et al. (2022)	Designed a Healthcare 4.0 throughput prediction model using ML algorithms for telemonitoring and teleconsultation data.	BP, SpO2 sensor, audio, and video healthcare data	K-means achieved 72.68% accuracy and 75.9% sensitivity; Naïve Bayes achieved 72% precision.	Lower prediction accuracy for complex healthcare environments, future work may include hybrid AI optimization models.
Ben Ali Kaddour and Abdulaziz (2021)	Proposed AI-based digital pathology using CNN architectures and morphological feature extraction for multi-disease medical image classification.	Medical image pathology dataset	Achieved 90.47% accuracy and 0.8332 F1-score.	Limited to specific pathology datasets, future work can improve generalization across multiple diseases and integrate explainable AI.

Research Gaps: The existing literature on healthcare and pharmaceutical data classification has shown good results with ML and DL methods, but most current models still fail to achieve good generalization ability, prediction ACC, and to be able to deal with the complex and high dimensionality of healthcare and pharmaceutical data. There are some already developed solutions that are applicable only to a specific set of diseases, medical images, or to medical text analysis, thus restricting their use in multiple healthcare contexts in the pharmaceutical sector. Furthermore, the model performance and reliability are negatively affected by challenges such as class imbalance, computational complexity, and poor feature optimization. Therefore, there is a need for an effective and powerful AI based healthcare classification system that can improve the prediction ACC, process complex healthcare data and can be used in intelligent clinical decision making system etc.

III. RESEARCH METHODOLOGY

The proposed methodology is based on the Kaggle's MIMIC-IV dataset which is used to develop the Smart Classification Model for Pharma Healthcare Data. The acquired data undergoes a series of pre-processing processes to improve its quality and distribution of classes. These steps include removing outliers, imputation of missing values, one-hot encoding, normalization of z-scores, and data balancing using SMOTE. Classification of the processed data is then carried out using the XGBoost algorithm. ACC, PRE, REC, F1, confusion matrix, and ROC curve analysis are then used

to evaluate the model's performance. The proposed FL of the Smart Classification Model for Pharma Healthcare Data is shown in Fig.1.

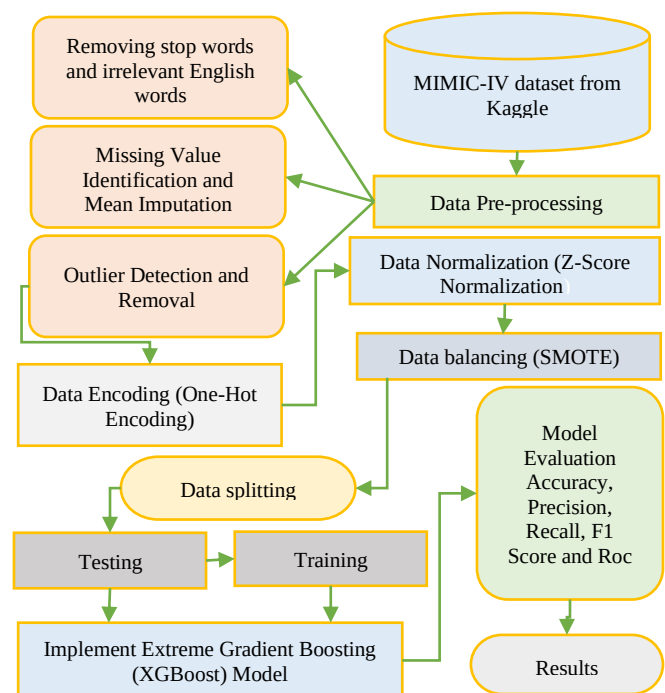


Fig. 1. Proposed Flowchart for Smart Classification Model for Pharma Healthcare Data

A. Data Gathering and Analysis

The dataset was collected from the MIMIC-IV database and Kaggle healthcare repositories, containing patient clinical records, diagnoses, treatments, and clinical summaries from 2008–2019. The dataset consisted of 999 records with two main columns, and visualization techniques such as bar plots and heatmaps were used for data analysis below Figure 1 illustrates the proposed methodology for smart classification of pharmaceutical healthcare data using an intelligent predictive framework:

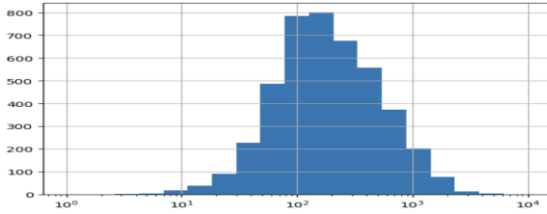


Fig. 2. Power Distribution of Number of Patients Records

Fig 2 presents the provided histogram displays the power distribution of patient records, utilizing a logarithmic scale on the x-axis ranging from 10^0 to 10^4 to represent the volume of records. This distribution is approximately symmetric and bell-shaped on this log scale, with the maximum amount of data peaking around 100 to 200 patient records. Records occur at a predictable frequency towards the lower end and towards the higher end (around 1 and approaching 10,000) and have a typical log-normal characteristic.

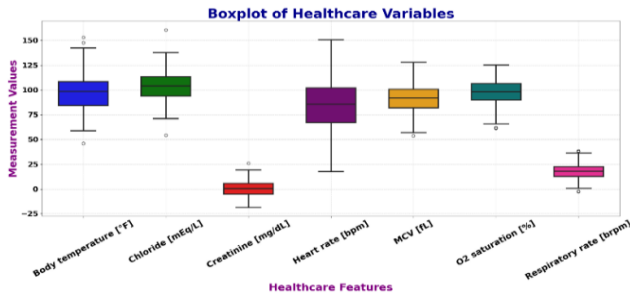


Fig. 3. Boxplot of the examined Healthcare variables

Fig. 3 displays the distribution and variation of seven different variables related to the health care system, ranging from a single measurement scale and demonstrating the central tendencies, interquartile ranges, and outliers of each variable. The features such as body temp, chloride, HR, O2 sat, MCV have high baselines with midpoints between 75 and 110, with HR being the widest and most volatile. By contrast, the numeric ranges of the other two, creatinine and respiratory rate, are much smaller, with one interesting case of a negative value for creatinine, possibly an artifact of the data pre-processing or a scaled/normalized transformation of the features.

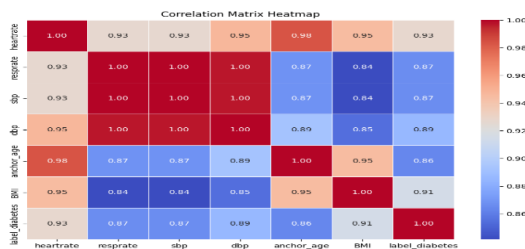


Fig. 4. Correlation Matrix Heatmap

Fig 4 presents the correlation matrix heatmap demonstrates extremely strong positive correlations among all clinical and demographic variables, with Pearson coefficients ranging from 0.84 to 1.00. High correlations are seen between the features respiratory rate, SyBP, DiBP, and other features like heart rate, BMI, anchor age and diabetes label are also highly correlated, which shows that the features are related.

B. Data Pre-Processing

The MIMIC-IV database was used for data concatenation, cleaning, and data pre-processing. Preprocessing steps included stop words removal, irrelevant English words removal, missing value imputation by mean, outlier detection and elimination, and data labelling and normalization. The main pre-processing tasks are described briefly as follows:

- **Removing stop words and irrelevant English words:** This phase involved removing phrases such "patient," "mg," "doctors," "B," "case," "always," and "getting" that contributed little or nothing to the analysis, in order to make the data easier to understand and improve the ACC.
- **Missing Value Identification and Mean Imputation:** Identified and replaced low rate of missing values in healthcare features to maintain data consistency by applying Mean Imputation technique.
- **Outlier Detection and Removal:** Identified the outliers by using the Boxplot and IQR method and removed the noisy data points which are extreme to obtain the accurate model and better prediction capability.

C. Data Encoding (One-Hot Encoding)

The dataset was prepared for ML analysis by converting categorical healthcare variables to numerical format using the One-Hot Encoding approach. This resulted in conversion of each category to binary to facilitate model compatibility and prediction performance.

D. Data Normalization (Z-Score Normalization)

Data normalization is a preprocessing method that helps models run more smoothly by bringing all of the data into a consistent distribution. Applying z-score normalization ensures that feature values centered around the average with unit variance. This method adjusts the data to have a mean of 0 and a standard deviation of 1, thereby reducing outliers. Equation (1) defines the z-score normalization.

$$E' = \frac{E - \bar{M}}{\sigma_M} \quad (1)$$

Where E' and E are new and old for every data point, $\bar{M}$ is the mean, and σ_M is the standard deviation

E. Data Balancing (SMOTE)

Data balancing involves changing the distribution of classes to enhance the performance of the ML model. This is done when there is an imbalance in the dataset, such as when one class contains far less samples than the others. When one class is significantly underrepresented in a dataset, SMOTE is a popular data augmentation technique for bringing the two classes back into parity. Applying preexisting instances of the minority class to generate new data for that class is key to SMOTE. As a basic method, the algorithm selects a minority class instance by considering its k closest neighbors in the feature space.

F. Data Splitting

A stratified 80:20 train-test split was used. By utilizing this stratified split, the original dataset's proportionate class distributions were maintained in both the training and testing groups.

G. Proposed Extreme Gradient Boosting (XGBoost) Model

The Smart Classification strategy for Healthcare Data is suggested to use XGBoost, a supervised ML-based boosting strategy. Utilizing an ensemble learning technique that incorporates many decision trees, XGBoost enhances the PRE and efficacy of forecasts. The model can also minimize a loss function, which represents the difference between the actual and expected goal values, in regression tasks. The XGBoost regression model is mathematically described by Equation (2):

$$y = f(x) \quad (2)$$

the XGBoost model that takes a vector of input features (such square footage, number of bedrooms, etc.) and utilizes them to predict a variable y , where y is the expected property price, and functions from x to y . The XGBoost algorithm builds a network of decision trees that have been trained to minimize the MSE loss function in order to determine $f(x)$. The model takes an average of all the DT results to get the final projection. The XGBoost regression model in its whole appears to be Equation (3)

$$y = \sum_{k=1}^K f_k(x) \quad (3)$$

where K is the total number of DTs in the ensemble and $f_k(x)$ is the prediction of the K th DT. All trees learn their leaf values during training, and their predictions are a weighted average of those values. The XGBoost model makes a prediction for each input x by summing the predictions of all the ensemble DTs.

H. Evaluation Metrics

A number of classification measures were used to judge how well the proposed model worked. First, a confusion matrix was created to show the categorization findings. This matrix included the number of correct and incorrect guesses for each class. Valuable metrics including TP, FP, TN, and FN were extracted from this matrix. Important evaluation metrics like as ACC, PRE, REC, and F1 were computed using these data, as explained below:

Accuracy: The trained model's ACC is defined as the percentage of occurrences it properly predicts out of all the examples in the dataset. Equation (4) provides a mathematical representation of this

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Precision: The "PRE" stands for the proportion of correctly predicted positive events to the total number of positive events. It is defined in Equation (5) and indicates how well the classifier can detect positive classifications

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall: The REC of a model is the proportion of correctly predicted positive cases to the total number of positive examples. In mathematics, it is expressed as Equation (6) and it assesses the model's capacity to detect all pertinent positive instances

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F1 score: A balanced measure between REC and PRE, the F1 is the harmonic mean of the two. A higher score suggests greater categorization ability; it runs from 0 to 1. Its mathematical representation is given by Equation (7)

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Receiver Operating Characteristic Curve (ROC): The ROC curve is a graph that shows how the TPR and the FPR change as the classification threshold number changes. Among other names, TPR might mean sensitivity or REC, whereas FPR is determined by dividing specificity by 1.

IV. RESULTS AND DISCUSSION

The setup was implemented using an Intel Core i7-13620H CPU and a graphics card from NVIDIA - an RTX 4060Ti with 24GB of GDDR6 memory, which is suitable for efficient computation and model training. The data was managed using PostgreSQL 13 and Python 3.9 was used as the programming language. This entire framework has been created and deployed on the Windows 10 operating system with the PyCharm 2020 Community Edition platform. The MIMIC-IV healthcare dataset was used to train the proposed XGBoost and other important performance metrics (ACC, PRE, REC and F1) were used to evaluate the proposed XGBoost model as presented in Table II. The results from the experiments indicate that the model yielded an accuracy of 96% in the classification of the healthcare data, hence the model is highly accurate with low prediction error. Moreover, its model PRE was 94.6% and REC was 91.6%, showing its high efficiency for accurately recognizing the positive health care cases and for searching the relevant clinical cases. The balanced, reliable and robust performance of the proposed XGBoost framework for intelligent pharmaceutical healthcare classification and clinical decision-support applications is further validated by the F1 of 97.7%.

TABLE II. CLASSIFICATION RESULTS FOR PHARMA HEALTHCARE DATA

Matrix	Extreme Gradient Boosting (XGBoost) Model
Accuracy	96
Precision	94.6
Recall	91.6
F1-score	97.7



Fig. 5. Confusion Matrix for the Xgboost Model

The confusion matrix of the XGBoost model is shown in Fig. 5, which has the ability to predict the binary classification with high accuracy, with 98 negative samples correctly classified, and 92 positive samples correctly classified. The classification error was very low (only two false positives and eight false negatives) with excellent prediction power. Matrix diagonal values that are most often used to indicate the XGBoost model's reliability and usefulness in healthcare data classification and diagnosis prediction.

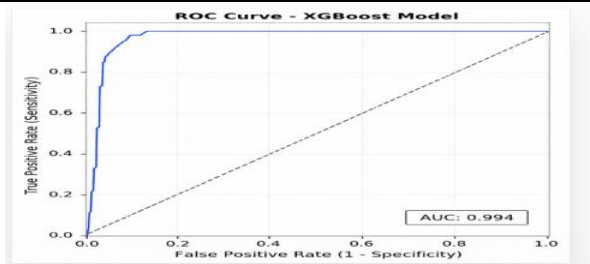


Fig. 6. ROC Curve for Xgboost Model

The classification performance of XGBoost model is shown in Figure 6 with the TPR (Sensitivity) and the FPR (1 – Specificity). The curve is situated near the upper left corner, which demonstrates a good discrimination power between classes and a very good prediction power. The XGBoost model's AUC score of 0.994 shows that it is very accurate and dependable in classification tasks.

A. Comparative Analysis

The suggested model's usefulness is judged by seeing how accurate it is compared to other models in Table III. Different models such as XGBoost, RF, LR, NB, LSTM and SVM models were tested and evaluated with the help of different metrics like F1, REC, ACC and PRE. The accuracy of XGBoost for the classification of the pharma healthcare data is 96%, PRE of 94.6%, REC of 91.6%, and F1 of 97.7% which is better when compared to other models. RF, LR, NB and LSTM had lower predictive ability, while SVM had comparatively good results but were still not as good as XGBoost. The results show that the proposed XGBoost model is better at healthcare data classification for AI healthcare apps using pharmaceutical data in terms of reliability, balance, and efficiency.

TABLE III. COMPARISON OF DIFFERENT MACHINE LEARNING AND DEEP LEARNING MODELS FOR CLASSIFICATION MODEL FOR PHARMA HEALTHCARE DATA

Model	Accuracy	Precision	Recall	F1-score
RF[31]	68.5	69	66.7	66.9
LR[32]	72	88	72	78
NB[32]	73.8	79.4	87.9	72
LSTM[33]	72.04	65.53	64.08	63.38
SVM[34]	91.4	91.8	90.8	91.2
XGBoost	96	94.6	91.6	97.7

The suggested XGBoost model did a better job of classifying pharma healthcare data than other machine learning and deep learning models. It was found that its ensemble boosting mechanism was effective in dealing with complex healthcare features, the reduction of prediction errors and the overall ACC of classification. The model's high F1, REC, and PRE scores indicate that it reliably and accurately predicts healthcare outcomes. These results indicate that the proposed XGBoost framework is a good option for the smart drug classification system and healthcare decision support applications.

V. CONCLUSION AND FUTURE STUDY

Machine learning and deep learning have emerged as viable approaches to address complex challenges in smart healthcare systems over the past few years. These challenges involve drug discovery, disease diagnosis, decision making for physicians, and data analysis in healthcare. This research proposed a comprehensive classification for Pharma Healthcare Data using different ML and DL models in order

to find the optimum forecasting method. Different classification models like XGBoost, RF, LR, NB and LSTM were put to experimental evaluation. At 96%, the XGBoost model outperformed all of the others, including RF (68.5%), LR (72%), NB (73.8%), LSTM (72.04%), and SVM (91.4%), according to the data. XGBoost has several advantages over the other models, such as ensemble boosting, feature learning, and it has lower risk of overfitting yet still predicts well. It can easily process high dimensional healthcare data. Hence, it is concluded that, the classification system developed using XGBoost is an efficient, reliable, scalable solution for healthcare data analysis and intelligent healthcare decision making applications in the pharmaceutical field. However, the disadvantages are that it is based on a relatively small dataset and structured clinical information. Larger multi-source datasets, sophisticated deep learning models, and explainable AI methods should be explored in the future for improved ACCs of predictions and for real-time, informed healthcare decisions.

- **Funding:** This research received no external funding.
- **Institutional Review Board Statement:** Not applicable.
- **Informed Consent Statement:** Not applicable.
- **Data Availability Statement:** The data supporting the findings of this study are available from the corresponding author upon reasonable request.
- **Conflicts of Interest:** The author declares no conflict of

How to cite this article:

A. Sunarty, "A Robust Multi-Class Classification Model for Pharma Healthcare Data Using Artificial Intelligence Techniques," *J. Glob. Res. Electron. Commun.*, vol. 2, no. 9, pp. 9–15, Sep. 2026. Doi.XXX

REFERENCES

- [1] A. Warriar, "iPaaS Solutions for Healthcare Enterprise Integration: Cloud-Native Integration Platforms for Multi-System Orchestration," *Int. J. Lead. Res. Publ.*, vol. 3, no. 1, pp. 1–9, 2022, doi: 10.70528/ijlrp.v3.i1.1770.
- [2] K.-K. Mak and M. R. Pichika, "Artificial intelligence in drug development: present status and future prospects," *Drug Discov. Today*, vol. 24, no. 3, pp. 773–780, Mar. 2019, doi: 10.1016/j.drudis.2018.11.014.
- [3] V. Chavda *et al.*, "Conventional and Novel Diagnostic Tools for the Diagnosis of Emerging SARS-CoV-2 Variants," *Vaccines*, vol. 11, no. 2, p. 374, Feb. 2023, doi: 10.3390/vaccines11020374.
- [4] R. Sharma, A. Shishodia, A. Gunasekaran, H. Min, and Z. H. Munim, "The role of artificial intelligence in supply chain management: mapping the territory," *Int. J. Prod. Res.*, vol. 60, no. 24, pp. 7527–7550, Dec. 2022, doi: 10.1080/00207543.2022.2029611.
- [5] M. R. Anand and A. K. S., "Temporal Fusion Transformer Forecasting and MILP Prescriptive Optimization for Hospital Pharmacy Supply Chain Orchestration," in *2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Nov. 2025, pp. 1206–1213, Nov. doi: 10.1109/ICECA66444.2025.11382695.
- [6] S. Dharmavaram and P. Bhanushali, "Machine Intelligence-Driven Forecasting for ED Triage and Dynamic Hospital Patient Routing," Feb. 20, 2026, doi: 10.64898/2026.02.18.26346566.
- [7] A. L. Grilo and A. Mantalaris, "The Increasingly Human and Profitable Monoclonal Antibody Market," *Trends Biotechnol.*, vol. 37, no. 1, pp. 9–16, Jan. 2019, doi: 10.1016/j.tibtech.2018.05.014.
- [8] S. Colombo, X. Zeng, H. Ragelle, and C. Foged, "Complexity in the therapeutic delivery of RNAi medicines: an analytical challenge," *Expert Opin. Drug Deliv.*, vol. 11, no. 9, pp. 1481–1495, Sep. 2014, doi: 10.1517/17425247.2014.927439.
- [9] P. Mishra, "Design Of Intelligent Healthcare It Infrastructure

- Using Graph Theory, Network Analysis, And Artificial Intelligence,” *Int. J. Appl. Math.*, vol. 38, no. 12s, pp. 2267–2280, Dec. 2025, doi: 10.12732/ijam.v38i12s.1547.
- [10] A. Warriar, “Machine Learning-Enhanced Data Quality Validation for Electronic Health Record Integration,” *Int. J. Sci. Technol.*, vol. 11, no. 1, pp. 1–8, Feb. 2020, doi: 10.71097/IJSAT.v11.i1.8387.
- [11] R. Müller and Junghanns, “Nanocrystal technology, drug delivery and clinical applications,” *Int. J. Nanomedicine*, p. 295, Oct. 2008, doi: 10.2147/IJN.S595.
- [12] J. A. Kachhia, “Healthcare Predictive Analytics based on Machine Learning Techniques for Identifying Cardiovascular Risks Screening,” *Int. J. Curr. Eng. Technol.*, vol. 13, no. 6, pp. 635–342, Mar. 2026, doi: 10.14741/ijcet/v.13.6.17.
- [13] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, “Artificial intelligence in drug discovery and development,” *Drug Discov. Today*, vol. 26, no. 1, pp. 80–93, Jan. 2021, doi: 10.1016/j.drudis.2020.10.010.
- [14] N. Sapoval *et al.*, “Current progress and open challenges for applying deep learning across the biosciences,” *Nat. Commun.*, vol. 13, no. 1, p. 1728, Apr. 2022, doi: 10.1038/s41467-022-29268-7.
- [15] R. Snehamrutha, “Patient Engagement Strategies in Community Pharmacies and their Effect on Vaccination Uptake and Medication Synchronizations,” *ESP J. Eng. Technol. Adv.*, vol. 3, no. 3, pp. 163–173, 2023, doi: 10.56472/25832646/JETA-V3I7P120.
- [16] S. Mukherjee, “An Effective System for Medical Image Diagnosis Using Deep Convolutional Networks (CNNs) in Healthcare Sector,” in *2026 14th International Symposium on Digital Forensics and Security (ISDFS)*, Boston, MA, USA: IEEE, Mar. 2026, pp. 01–06, doi: 10.1109/ISDFS69419.2026.11459010.
- [17] H. Kim, E. Kim, I. Lee, B. Bae, M. Park, and H. Nam, “Artificial Intelligence in Drug Discovery: A Comprehensive Review of Data-driven and Machine Learning Approaches,” *Biotechnol. Bioprocess Eng.*, vol. 25, no. 6, pp. 895–930, Dec. 2020, doi: 10.1007/s12257-020-0049-y.
- [18] Y. You *et al.*, “Artificial intelligence in cancer target identification and drug discovery,” *Signal Transduct. Target. Ther.*, vol. 7, no. 1, p. 156, May 2022, doi: 10.1038/s41392-022-00994-0.
- [19] S. Mahmud, “AI And Data Analytics For Enhancing Home Healthcare: Optimizing Patient Outcomes And Resource Allocation,” *Front. Appl. Eng. Technol.*, vol. 2, no. 01, pp. 78–100, February, Jan. 2025, doi: 10.70937/faet.v2i01.61.
- [20] P. Kumar, “Leveraging Generative AI for Automated Data Standardization and Interoperability in Healthcare,” in *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, Dec. 2025, pp. 99–104, doi: 10.1109/ICAAIC64647.2025.11330217.
- [21] S. Golriz Khatami, S. Mubeen, V. S. Bharadhwaj, A. T. Kodamullil, M. Hofmann-Apitius, and D. Domingo-Fernández, “Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures,” *npj Syst. Biol. Appl.*, vol. 7, no. 1, p. 40, Oct. 2021, doi: 10.1038/s41540-021-00199-1.
- [22] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, “Machine learning approaches to drug response prediction: challenges and recent progress,” *npj Precis. Oncol.*, vol. 4, no. 1, p. 19, Jun. 2020, doi: 10.1038/s41698-020-0122-1.
- [23] P. Kumar, “Edge Computing and IoT for Real-Time Healthcare Data Processing and Integration,” in *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India: IEEE, 2025, pp. 105–110, December, doi: 10.1109/ICAAIC64647.2025.11331211.
- [24] M. R. Anand, “Enhancing Pharmaceutical Supply Chains with Densenet-121 and 1D CNN Integration,” in *2025 Global Conference in Emerging Technology (GINOTECH)*, PUNE, India: IEEE, May 2025, pp. 1–7, July, doi: 10.1109/GINOTECH63460.2025.11076754.
- [25] N. R. Goddilla and A. Sivasangari, “Optimizing Symptom Prediction in Healthcare: An Integrative Machine Learning Approach to Regression and Classification Models,” in *2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, IEEE, Jan. 2025, pp. 1–6, doi: 10.1109/IITCEE64140.2025.10915417.
- [26] S. G. A. G. S. J. V. J. and M. K., “A Novel Pharmacovigilance Strategy for Detecting Adverse Drug Reactions in Healthcare Using Machine Learning and Blockchain,” in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, IEEE, Aug. 2024, pp. 763–767, doi: 10.1109/ICoICI62503.2024.10696339.
- [27] M. A.-A. A. Hameed, K. Shaker, and H. A. Khalaf, “Sentiment Classification of Drug Reviews Using Machine Learning Techniques,” in *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, IEEE, Jan. 2023, pp. 457–462, doi: 10.1109/DeSE58274.2023.10099735.
- [28] M. Habib, Z. Wang, S. Qiu, H. Zhao, and A. S. Murthy, “Machine Learning Based Healthcare System for Investigating the Association Between Depression and Quality of Life,” *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 5, pp. 2008–2019, May 2022, doi: 10.1109/JBHI.2022.3140433.
- [29] P. Bardalai, H. Neog, P. E. Dutta, N. Medhi, and S. K. Deka, “Throughput Prediction in Smart Healthcare Network using Machine Learning Approaches,” in *2022 IEEE 19th India Council International Conference (INDICON)*, IEEE, Nov. 2022, pp. 1–6, doi: 10.1109/INDICON56171.2022.10040160.
- [30] A. Ben Ali Kaddour and N. Abdulaziz, “Artificial Intelligence Pathologist: The use of Artificial Intelligence in Digital Healthcare,” in *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, IEEE, Dec. 2021, pp. 31–36, doi: 10.1109/GCAIoT53516.2021.9693090.
- [31] H. Huynh, T. Cao, and H. Tran, “Machine Learning-Based Diabetes Classification Using Vital Signs and Clinical Information from the MIMIC-IV Dataset,” *Sci. J. Comput. Sci.*, vol. 2, no. 2, pp. 148–159, Mar. 2026, doi: 10.64539/sjcs.v2i2.2026.439.
- [32] H. Chen *et al.*, “Machine learning prediction and interpretability analysis of high-risk chest pain: a study from the MIMIC-IV database,” *Front. Physiol.*, vol. 16, no. June, pp. 1–17, 2025, doi: 10.3389/fphys.2025.1594277.
- [33] A. Gabitashvili and P. Kellmeyer, “Predicting Length of Stay in Neurological ICU Patients Using Classical Machine Learning and Neural Network Models: A Benchmark Study on MIMIC-IV,” 2025, *arXiv*, doi: 10.48550/arXiv.2505.17929.
- [34] Y. She, L. Zhou, and Y. Li, “Interpretable machine learning models for predicting 90-day death in patients in the intensive care unit with epilepsy,” *Seizure*, vol. 114, no. November 2023, pp. 23–32, 2024, doi: 10.1016/j.seizure.2023.11.017.