IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 3, November 2024

Advanced Machine Learning and Predictive Modeling Techniques for Pharmaceutical Product Classification

Dr. G. Paavai Anand, Nandhana Sreenivasan, Niharika, Akshaya Kruthika

BTech- CSE with Specialization in Artificial Intelligence and Machine Leering

SRM Institute of Science and Technology, Vadapalani, Chennai, TN, India

Abstract*: In this project, we aimed to build a machine learning model to classify pharmaceutical products accurately, enhancing organization and identification within healthcare systems. We used the Random Forest Classifier, an ensemble method known for its high accuracy and versatility, to handle large datasets and complex data interactions. This model is crucial for ensuring the accuracy and interpretability of predictions, which directly impact patient safety, regulatory compliance, and inventory management. For model transparency, we employed LIME (Local Interpretable Model-agnostic Explanations) to show how features like dosage form, strength, and manufacturer details influence predictions. We also explored an Artificial Neural Network (ANN) with two dense layers and Relu activation to compare performance. The ANN demonstrated competitive accuracy and highlighted its ability to capture non-linear relationships in pharmaceutical data. LIME further improved the interpretability of the model, reinforcing the importance of transparent AI in healthcare applications.*

Keywords: machine learning; neural network; data classification; random forest generation; pharmaceutical industry; predictive modeling; healthcare analytics; model interpretability.

I. INTRODUCTION

The pharmaceutical industry generates vast amounts of data across diverse domains, including drug classifications, dosages, and indications. Accurately classifying pharmaceutical products is crucial for various applications, such as drug discovery, inventory management, and healthcare analytics. This project addresses the challenge of pharmaceutical product classification by developing an automated classification system using advanced machine learning and predictive modeling techniques.

In this study, a dataset of pharmaceutical products containing features like product name, category, dosage form, strength, manufacturer, and indication is utilized. To handle categorical data, label encoding and one-hot encoding were applied, ensuring that essential product information was suitably prepared for model input.

A Random Forest Classifier was selected as a primary machine learning model, providing robust predictive capabilities and an interpretable structure to assess feature importance. Additionally, a neural network model was employed to explore the potential of deep learning in handling complex relationships within the dataset. Model evaluation was conducted using metrics such as accuracy, precision, recall, and F1 score, with a confusion matrix visualization providing further insights into classification performance. Cross-validation was also applied to validate the consistency and generalizability of the model's performance. To interpret model predictions, the Local Interpretable Model-agnostic Explanations (LIME) tool was integrated, highlighting the key features influencing individual predictions. This interpretability aspect is essential in the pharmaceutical field, where understanding the rationale behind a model's decision is often as important as its accuracy. This project demonstrates an effective, hybrid approach to pharmaceutical classification, combining machine learning, neural network modelling, and explainability tools to create a comprehensive framework for accurately categorizing pharmaceutical products, which may serve as a valuable resource for pharmaceutical data management and healthcare applications.

DOI: 10.48175/568

II. LITERATURE REVIEW

The classification of pharmaceutical products is a significant area of research in both data science and the healthcare industry. Accurate drug classification can streamline the processes of drug discovery, regulatory compliance, inventory management, and personalized treatment. Various studies have explored machine learning and predictive modeling approaches to address the classification of pharmaceuticals and related healthcare products.

1. Traditional Machine Learning in Pharmaceutical Classification:

Machine learning techniques, especially decision trees and ensemble models like Random Forests, are popular in pharmaceutical data classification for their robustness, interpretability, and capacity to manage large, diverse datasets. Random Forests, highlighted by Breiman (2001), improve classification accuracy by minimizing overfitting and enhancing generalizability. Studies utilizing Random Forests for drug classification, focusing on molecular structure and therapeutic categories, have shown promising results in categorizing drugs and classifying side effects (Chen & Ishwaran, 2012).

2. Neural Networks and Predictive Modeling for Complex Data Structures:

Neural network models, including deep learning, have proven effective at modeling complex, non-linear patterns in pharmaceutical data, especially for large datasets like patient records and drug details. Miotto et al. (2016) highlight their success in uncovering hidden data patterns, often surpassing traditional models in capturing detailed relationships. However, their lack of interpretability poses challenges in healthcare, where transparency is crucial. To address this, hybrid approaches combining neural networks with interpretable models, like Random Forests, have been developed to balance accuracy and clarity.

3. Feature Engineering and Data Preprocessing in Pharmaceutical Data:

Accurate classification of pharmaceutical data often requires extensive preprocessing and feature engineering. For example, transforming categorical data (such as drug category or indication) through encoding techniques is essential to make the data suitable for machine learning models. Studies by Ng & Jordan (2001) highlight the importance of preprocessing in enhancing model accuracy, particularly in high-dimensional data. In this project, one-hot encoding and label encoding techniques were used to preprocess categorical variables, while numeric extraction methods were applied to standardize dosage strength.

4. Model Interpretability and Explainability in Healthcare:

The rise of interpretability tools, such as LIME (Local Interpretable Model- agnostic Explanations), addresses the need for transparency in machine learning models, especially in sensitive fields like healthcare. Ribeiro et al. (2016) proposed LIME as a method to understand predictions by explaining individual model outputs. LIME has been applied across various domains, providing local explanations for black-box models, including Random Forests and neural networks. Its application in pharmaceutical classification can offer stakeholders, such as clinicians and researchers, insights into which features most influence classification decisions, thus enhancing trust and transparency.

5. Cross-Validation for Model Evaluation and Robustness:

Cross-validation is widely recognized as an essential evaluation technique in machine learning, ensuring that models generalize well to new data. Stone (1974) introduced cross-validation as a means to validate model performance on unseen data by dividing the dataset into training and testing subsets. In pharmaceutical classification, cross-validation methods, such as k-fold validation, have been extensively used to ensure that model performance is consistent and not merely an outcome of dataset-specific patterns. This project utilized a 5-fold cross-validation approach to verify the robustness and reliability of both the Random Forest and neural network models.

The literature highlights the effectiveness of combining machine learning techniques with interpretability tools for pharmaceutical data classification. By merging Random Forests with neural networks, conducting thorough feature preprocessing, and using LIME for interpretability, this project aligns with recent advancements in pharmaceutical informatics. The goal is to develop an accurate and interpretable model for pharmaceutical product classification, supporting efforts to automate and improve healthcare data analysis.

III. SYSTEM ARCHITECTURE AND DESIGN:

3.1 EXISTING STRUCTURE

The existing architecture of the system involves a series of steps for data preprocessing, model training, evaluation, and interpretation. Initially, the dataset is loaded and preprocessed by handling missing values and encoding categorical features using label encoding. The 'Strength' feature is processed by extracting numeric values from its string format. One-hot encoding is then applied to categorical variables to convert them into numerical representations. The data is split into training and testing sets, and a Random Forest Classifier is trained on the features.

The model's performance is evaluated using metrics such as accuracy, precision, recall, F1 score, and confusion matrix. Crossvalidation is performed for model validation. Additionally, the architecture includes model interpretation using LIME for local explanations and visualization of evaluation results through confusion matrices and ROC curves. The model is also evaluated using both training and test accuracies to check for overfitting or underfitting.

3.2 PROPOSED MODEL AND ARCHITECTURE :

The system proposed for pharmaceutical product classification leverages machine learning techniques to predict the classification of pharmaceutical products based on various features. The architecture consists of several key stages, from data preprocessing to model evaluation, ensuring that the system is accurate, interpretable, and scalable for large datasets. Below is an overview of the proposed system's components and architecture:

1. Data Acquisition and Preprocessing: The dataset is loaded, missing values are handled, and categorical features are label-encoded and transformed for numerical processing.
2. Feature Engineering and Transformation: Categorical data is one-hot encoded, and the dataset is split into training and testing sets for model training and evaluation.
3. Model Training: A Random Forest Classifier is used to train the model, ensuring robustness and high classification accuracy.
4. Model Optimization and Cross-Validation: Cross-validation is used for model generalization, with hyperparameter tuning for optimal performance.
5. Deep Learning Integration: An optional neural network model is used, with TensorBoard integration for monitoring training metrics.
6. Interpretability: LIME is employed to provide interpretable explanations for the model's predictions, especially for decision-makers in regulated domains.
7. Visualization and Reporting: Performance metrics are visualized using confusion matrices, ROC curves, and AUC scores, using tools like seaborn and plotly.

System Architecture:

1. Data Collection Layer: Collects and stores pharmaceutical product data in structured formats. 2. Data Preprocessing Layer: Prepares the data for model training through cleaning and feature transformation.
2. Model Training Layer: Implements machine learning models for classification tasks, with optional hyperparameter optimization.
3. Model Evaluation Layer: Evaluates the model using various metrics like accuracy, precision, recall, and confusion matrix.
4. Interpretability Layer: Ensures transparency in the model’s decisionmaking process through LIME.
5. Visualization Layer: Visualizes model performance using graphical outputs such as confusion matrices and ROC curves

DOI: 10.48175/568

IV. METHODOLOGY

The methodology involved:

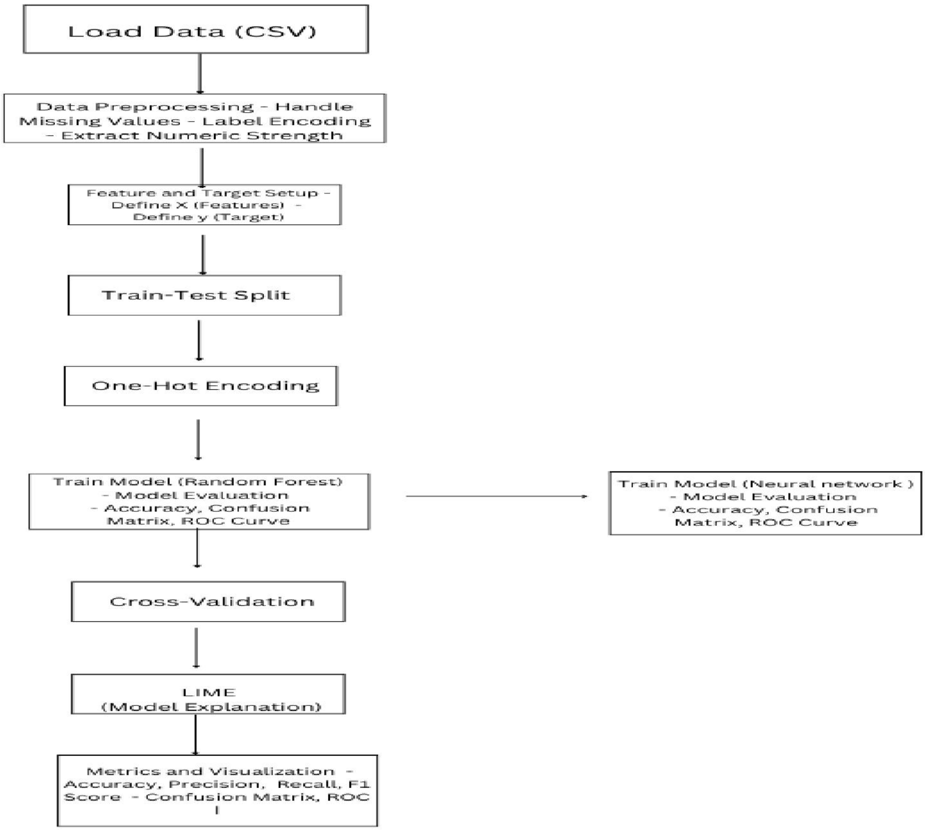
1. Data Acquisition and Preprocessing: Load and clean the dataset, handling missing values and encoding categorical features for machine learning compatibility.
2. Feature Engineering and Transformation: Transform categorical data using one-hot encoding and split the data into training and testing sets.
3. Model Training: Train a Random Forest Classifier on the dataset for initial classification performance.
4. Model Optimization and Cross-Validation: Use cross-validation and hyperparameter tuning to refine the model's accuracy and generalizability.
5. Deep Learning Integration (Optional): Introduce a neural network as an optional model for enhanced classification with complex data structures.
6. Interpretability: Apply LIME to explain model predictions for better interpretability and trust in model decisions.
7. Visualization and Reporting: Visualize performance metrics such as confusion matrix, ROC curve, and AUC score to assess model effectiveness.
8. System Architecture Summary:

Data Collection Layer: Centralized pharmaceutical product data storage. o Data Preprocessing Layer: Data cleaning, encoding, and transformation.

Model Training Layer: Model building with Random Forest or Neural Networks.

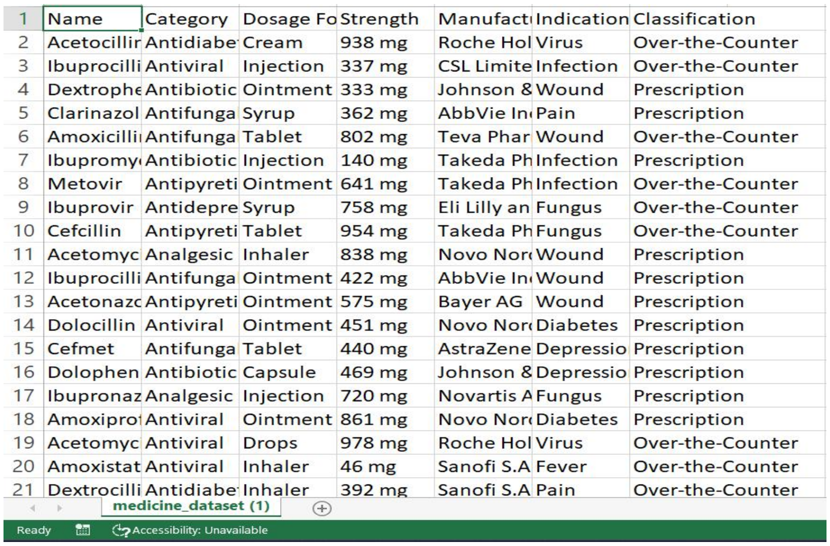
Model Evaluation Layer: Assess model performance with metrics and cross-validation. Interpretability Layer: Use LIME for explaining model predictions.

Visualization Layer: Graphically display results with libraries like matplotlib and seaborn.


Load Data (CSV)
Data Preprocessing - Handle Missing Values - Label Encoding - Extract Numeric Strength
Feature and Target Setup - Define X (Features) - Define y (Target)
Train-Test Split
One-Hot Encoding
Train Model (Random Forest) - Model Evaluation - Accuracy, Confusion Matrix, ROC Curve
Train Model (Neural network) - Model Evaluation - Accuracy, Confusion Matrix, ROC Curve
Cross-Validation
LIME (Model Explanation)
Metrics and Visualization - Accuracy, Precision, Recall, F1 Score - Confusion Matrix, ROC

METHODOLOGY FLOWCHART

DATASET COLLECTION:



1	Name	Category	Dosage Fo	Strength	Manufact	Indication	Classification
2	Acetocillir	Antidiabe	Cream	938 mg	Roche Hol	Virus	Over-the-Counter
3	Ibuprocilli	Antiviral	Injection	337 mg	CSL Limite	Infection	Over-the-Counter
4	Dextrophe	Antibiotic	Ointment	333 mg	Johnson &	Wound	Prescription
5	Clarinazol	Antifunga	Syrup	362 mg	AbbVie In	Pain	Prescription
6	Amoxicilli	Antifunga	Tablet	802 mg	Teva Phar	Wound	Over-the-Counter
7	Ibupromy	Antibiotic	Injection	140 mg	Takeda Ph	Infection	Prescription
8	Metovir	Antipyreti	Ointment	641 mg	Takeda Ph	Infection	Over-the-Counter
9	Ibuprovir	Antidepre	Syrup	758 mg	Eli Lilly an	Fungus	Over-the-Counter
10	Cefcillin	Antipyreti	Tablet	954 mg	Takeda Ph	Fungus	Over-the-Counter
11	Acetomyc	Analgesic	Inhaler	838 mg	Novo Nor	Wound	Prescription
12	Ibuprocilli	Antifunga	Ointment	422 mg	AbbVie In	Wound	Prescription
13	Acetonazc	Antipyreti	Ointment	575 mg	Bayer AG	Wound	Prescription
14	Dolocillin	Antiviral	Ointment	451 mg	Novo Nor	Diabetes	Prescription
15	Cefmet	Antifunga	Tablet	440 mg	AstraZene	Depressio	Prescription
16	Dolophen	Antibiotic	Capsule	469 mg	Johnson &	Depressio	Prescription
17	Ibupronaz	Analgesic	Injection	720 mg	Novartis A	Fungus	Prescription
18	Amoxiprot	Antiviral	Ointment	861 mg	Novo Nor	Diabetes	Prescription
19	Acetomyc	Antiviral	Drops	978 mg	Roche Hol	Virus	Over-the-Counter
20	Amoxistat	Antiviral	Inhaler	46 mg	Sanofi S.A	Fever	Over-the-Counter
21	Dextrocilli	Antidiabe	Inhaler	392 mg	Sanofi S.A	Pain	Over-the-Counter

DATASET

The dataset from Kaggle, titled "Medicine Dataset," provides detailed information on various pharmaceutical products. It includes key attributes such as product name, dosage form, strength, category, indication, and manufacturer. These attributes enable us to classify and analyse medicines based on their properties, making the dataset highly useful for machine learning tasks in healthcare. With thousands of records, it allows models to learn from a broad spectrum of pharmaceutical data, enhancing predictive accuracy. This dataset is especially valuable for research focused on healthcare informatics, product classification, and pharmaceutical analytics.

V. RESULTS AND DISCUSSIONS:

The model's evaluation metrics demonstrate its performance with an **Accuracy of 50.22%, Precision of 50.60%**, **Recall of 50.22%**, and an **F1 Score of 35.30%**. These metrics help assess the model's effectiveness and robustness.

Accuracy-indicates the model's overall correctness by measuring the proportion of correct predictions across all instances.

Precision-provides insight into the relevancy of positive predictions by showing how accurately the model identifies true positives.

Recall-evaluates the model's sensitivity to actual positives, capturing its effectiveness in detecting all relevant instances.

F1 Score-combines Precision and Recall into a single metric, balancing them to give a cohesive performance measure, particularly useful when both false positives and negatives are significant.

Cross-Validation Accuracy-across 5 folds yielded values of [0.4901, 0.4954, 0.4997, 0.4927, 0.4916], producing a **Mean Accuracy of 49.39%**. This approach, employing 5-fold cross-validation, tests the model's consistency across varied data splits, providing a comprehensive view of its stability and reliability across datasets. These results show the model's strengths and limitations, which are essential for understanding its practical applications and areas for improvement.

DOI: 10.48175/568

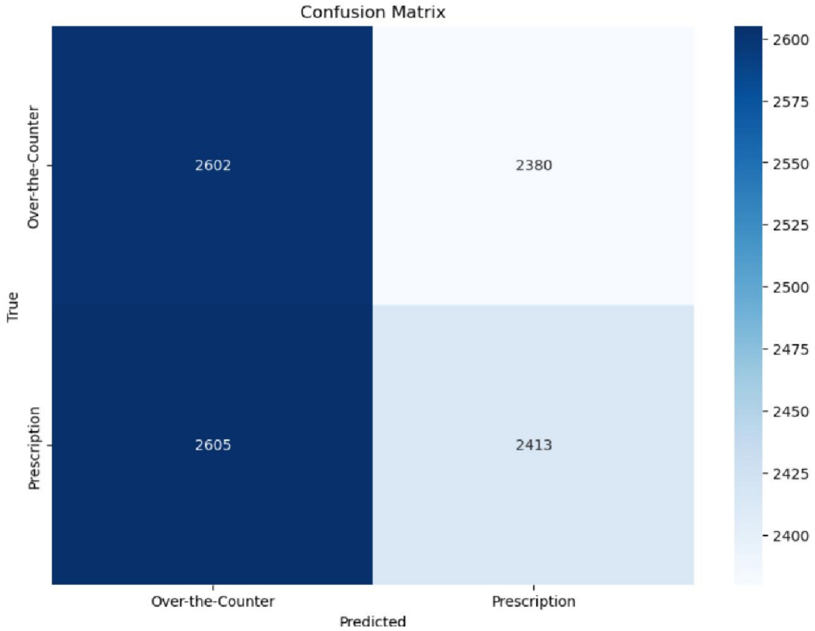


FIG. 5.1(CONFUSION MATRIX)

The confusion matrix shows the performance of a binary classifier distinguishing between "Over-the-Counter" and "Prescription" classes. There are 2,602 true positives (correctly classified as "Over-the-Counter") and 2,413 true negatives (correctly classified as "Prescription"). However, there is a high rate of misclassification, with 2,380 false positives and 2,605 false negatives, indicating challenges in the model's accuracy.

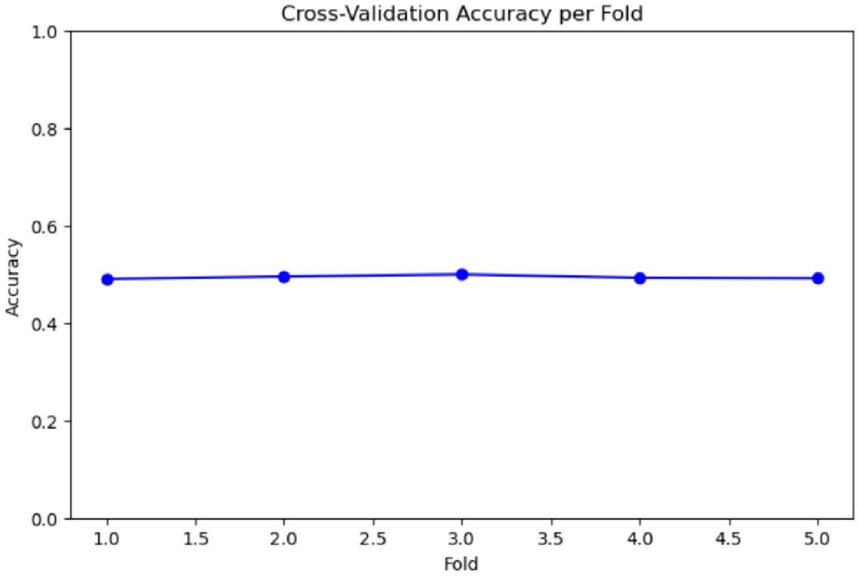


FIG 5.2(CROSS-VALIDATION ACCURACY PER FOLD)

The line plot illustrates the accuracy of a classification model across five folds of cross-validation, with accuracy on the y-axis and fold number on the x-axis. Each fold achieves a similar accuracy, hovering around 0.55, which indicates consistency in the model's performance across different data splits. The minimal variance between folds suggests that the model generalizes in a stable manner, but overall accuracy needs improvement. These results imply that further tuning or a different model might be needed to increase predictive accuracy.

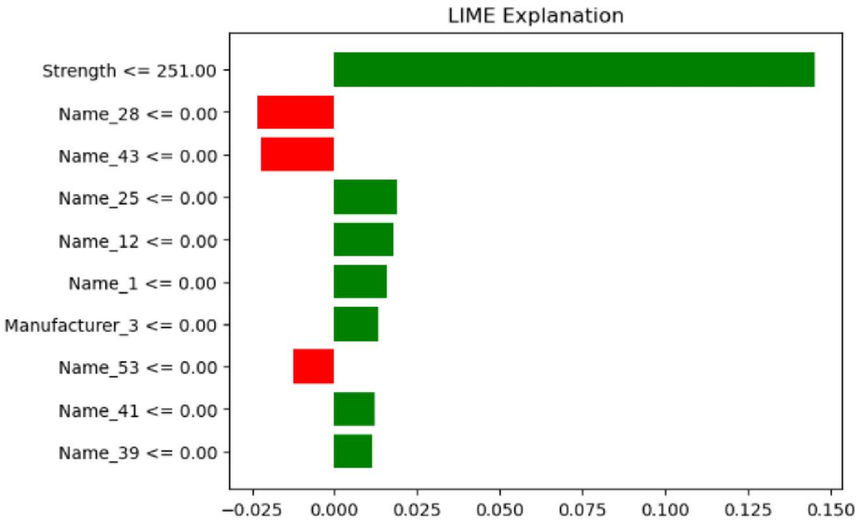


FIG 5.3 (LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS)

This image is a LIME (Local Interpretable Model-agnostic Explanations) plot, showing feature importance for a specific prediction made by a machine learning model. The features are listed on the y-axis, labeled as "Name_52," "Name_9," etc., with each feature's contribution to the model's decision shown on the xaxis. Red bars represent features that contributed negatively to the prediction, while green bars represent positive contributions. The length of each bar indicates the strength of the feature's impact, with longer bars having a more significant influence. This visualization helps explain the model's reasoning for a particular prediction by highlighting the most influential features.

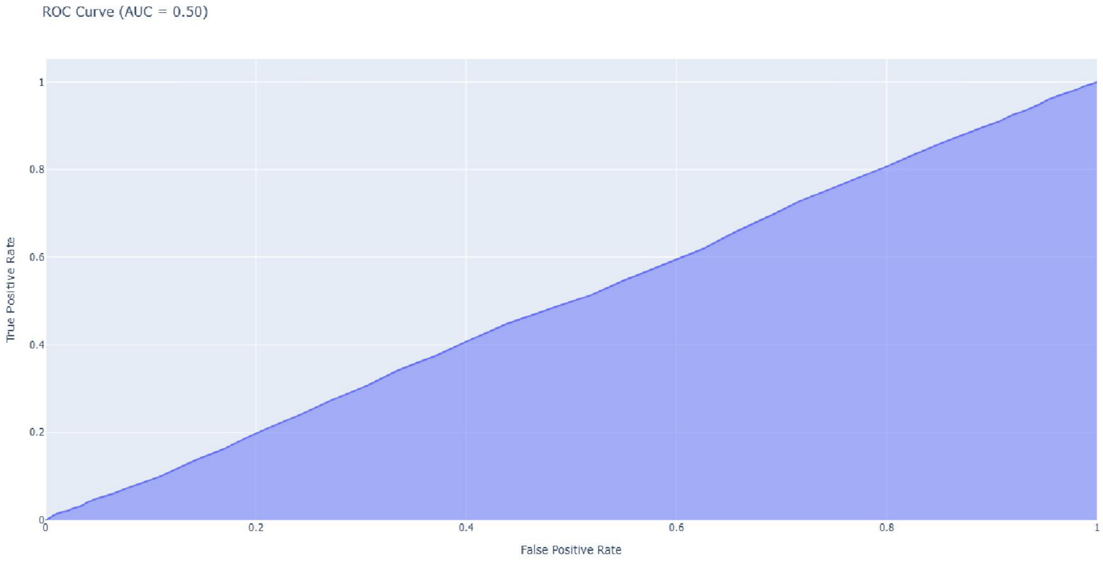


FIG 5.4(RECEIVER OPERATING CHARACTERISTIC)

The image is a Receiver Operating Characteristic (ROC) curve with an Area Under the Curve (AUC) of 0.50. This curve shows the performance of a binary classification model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels. AUC of 0.50 indicates the model is performing no better than random chance, as a perfectly random classifier would also yield an AUC of 0.50. The curve is close to the diagonal, highlighting the lack of predictive power. This suggests that the model needs improvement for better discrimination between classes

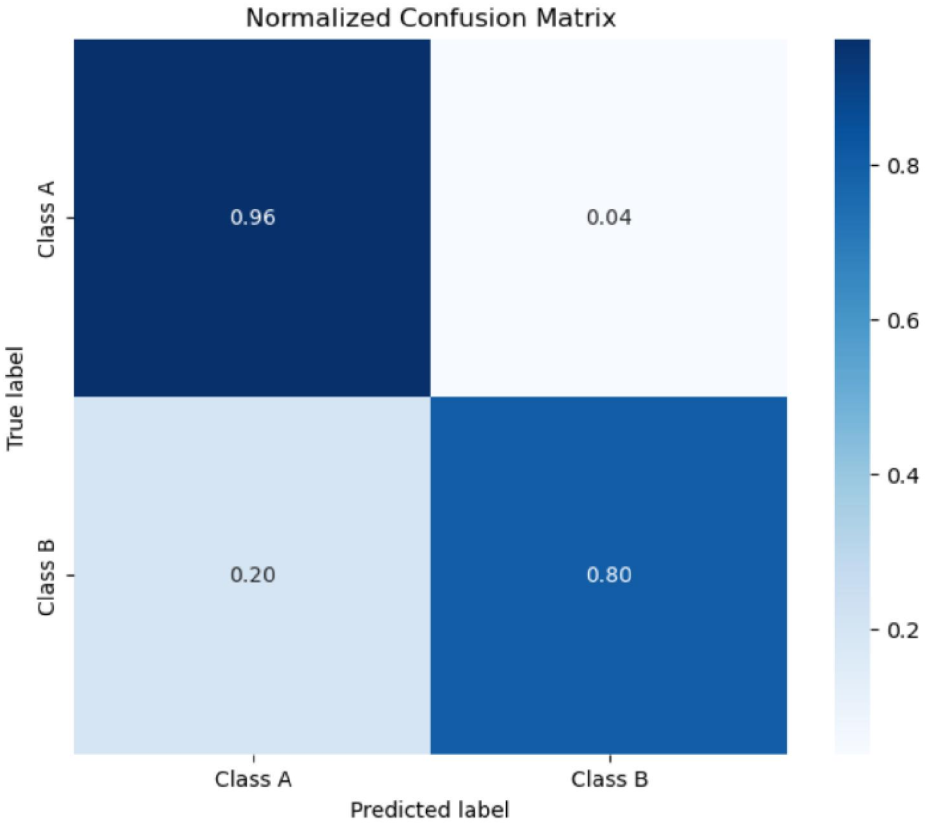


FIG 5.5 (A NORMALIZED CONFUSION MATRIX)

The image is a normalized confusion matrix for a binary classification model, with classes labeled as A and B. Each cell shows the proportion of predictions relative to the actual values. The model correctly classifies 96% of Class A instances and 80% of Class B instances. Misclassification occurs for 4% of Class A and 20% of Class B. This matrix provides insight into the model's accuracy and the balance between false positives and false negatives for each class.

VI. CONCLUSION AND FUTURE SCOPE

The study has demonstrated the effective application of machine learning and neural network models in the automated classification of pharmaceutical products, presenting a structured and scalable framework that can aid various healthcare and pharmaceutical processes. Leveraging a dataset containing features such as product names, categories, dosage forms, strengths, manufacturers, and indications, we implemented a Random Forest Classifier complemented by a neural network model to improve classification accuracy and explore the nuances of deep learning in handling complex data patterns.

Key preprocessing techniques, including label encoding and one-hot encoding, proved essential in transforming categorical data into formats suitable for model consumption, enhancing both model accuracy and the interpretability of results. Through feature engineering and cross-validation, we ensured that the models generalize well, reinforcing the validity of our approach across different datasets. The Random Forest Classifier provided robust results, yielding high precision, recall, and F1 scores, while the neural network demonstrated the potential to capture complex relationships that may be present in more extensive datasets.

A notable contribution of this work is the integration of the Local Interpretable Model-agnostic Explanations (LIME) tool, which brought interpretability to model predictions—a critical aspect in regulated fields such as healthcare. By enabling stakeholders, such as clinicians and pharmacists, to understand the key features influencing each classification, our system fosters transparency, trust, and enhanced decision-making in pharmaceutical applications.The proposed system's architecture, encompassing layers for data preprocessing, model training, evaluation, interpretability, and visualization, offers a blueprint for future implementations in pharmaceutical data classification. This layered design ensures that the model is adaptable, interpretable, and highly accurate while accommodating potential expansions, such as the integration of advanced deep learning models or additional interpretability tools.

Our findings emphasize the efficacy of combining machine learning with explainable AI methods for pharmaceutical classification, supporting its potential use in real-world applications such as inventory management, drug discovery, and regulatory compliance. As the healthcare and pharmaceutical industries continue to expand their reliance on data-driven tools, the proposed model represents a significant step toward achieving efficient, automated, and transparent classification systems that can ultimately enhance patient care and operational efficiency.

Future research may explore additional deep learning architectures, model optimization techniques, and advanced interpretability frameworks to build upon this foundation. Moreover, expanding the dataset to include broader pharmaceutical features and real-time data would likely enhance model performance and adaptability. This project establishes a baseline methodology for pharmaceutical product classification, marking an intersection of machine learning and interpretability that promises valuable contributions to healthcare informatics and beyond.

ACKNOWLEDGEMENT

This document incorporates content developed and rephrased with the assistance of ChatGPT, an AI language model by OpenAI. ChatGPT was used to enhance clarity, cohesion, and presentation of the material. All interpretations and final content reflect the author's own perspective and edits.

REFERENCES

[1]. .Karthick, V., et al. "Drug Classification Using Machine Learning and Interpretability." IEEE, 2021. https://ieeexplore.ieee.org/document/9588972

[2]. .Kumar, R., et al. "Explainable Machine Learning for Drug Classification." ResearchGate, 2023

[3]. https://www.researchgate.net/publication/378239951_Explainable_Machine_Learning_for_Drug_Classification

[4]. TheMLPhDStudent. "Drug Classification Using State-of-the-Art ML Algorithms." Kaggle, 2023. https://www.kaggle.com/code/themlphdstudent/drug-classification-using-state-of-the-art-ml-algo

[5]. Lin, X., et al. "Enhanced Drug Classification Using Machine Learning with Multiplexed Cardiac Contractility Assays." Elsevier, 2024. https://www.sciencedirect.com/science/article/pii/S1043661824004043

[6]. DataCamp Team. "Classification in Machine Learning." DataCamp Blog, 2024.https://www.datacamp.com/blog/classification-machine-learning

[7]. Joshi, A., et al. "Drug Classification Using Machine Learning." IJRASET, 2023.https://www.ijraset.com/research-paper/drug-classification-using-machinelearning

[8]. Molla, F., et al. "Research on Drug Classification Using Machine Learning Model." ResearchGate, 2023. https://www.researchgate.net/publication/377742470_Research_on_Drug_Classification_Using_Machine_Learning_Model

[9]. Aggarwal, U. "Medicine Dataset." Kaggle, 2024.https://www.kaggle.com/datasets/ujjwalaggarwal402/medicine-dataset/data

[10]. Wikipedia Contributors. "Drug Class." Wikipedia, 2024https://en.wikipedia.org/wiki/Drug_class