Detection of counterfeit drugs: Advanced Strategies for Quality Assurance and Safe Medicine Distribution

Arpita Saha Sukanna[a], Nishat Tasnim Barsha[a] and M. F. Mridha[d,*]

[a]Department of Computer Science and Engineering, American International University,Bangladesh , Dhaka, 1205, Bangladesh
[d]Department of Computer Science, American International University - Bangladesh, Dhaka, 1229, Bangladesh

ARTICLE INFO

Keywords:
Machine Learning, Quality Monitoring, Medicine Distribution.
Counterfeit Drugs
Counterfeit Detection

ABSTRACT

Public health is heavily dependent on the safe and reliable distribution of medicines. However, pharmaceutical supply chains are still facing major challenges, such as counterfeit drugs, weak quality controls, and sudden disruptions. This project explores how machine learning (ML) can handle these issues by detecting counterfeit medicines. Here, several solutions already exist, but there are still some problems. In this work, I focus on developing ML tools that predict risks, detect anomalies, and monitor the quality of drugs.. By applying multimodel models such as **CatBoost+Naive-Bayes** hybrid model, the study uncovers patterns that traditional systems often miss. The findings show that ML can effectively detect problems like counterfeit products, temperature mishandling well before they increase. These allow pharmaceutical companies and logistics teams to act faster and make better decisions. In general, the project highlights the huge potential of machine learning in creating a safer and more trustworthy pharmaceutical supply chain.

1. Introduction

Medicines are supposed to heal, not harm. Pharmaceutical Research Development is a complicated method that takes several years from drug discovery to drug improvement and regulatory approval[1]. Unlike standard consumer goods, pharmaceuticals are sensitive products with strict regulatory, storage, and handling requirements, especially vaccines. Despite living a healthy lifestyle, sometimes we must depend on medicines to cure diseases like heart problems, cancer, and others. That's why medical safety is much more important to maintain. But in many parts of the world, patients unknowingly consume counterfeit or poor-quality drugs, with potentially deadly consequences. These fake medicines not only fail to treat illnesses but can also cause serious side effects, worsen conditions, and sometimes even result in death, also causing some industrial loss, like reputation damage, financial loss and others. Some assistive tools do exist, like automated dispensing cabinets (ADCs) represent a solution that dispenses drugs automatically, and there are many ADC technologies in existence. Some studies have used barcoding for drug identification and the prevention of medication errors. Devices that employ radio frequency identification (RFID) and Bluetooth to identify the positions of drugs have been designed. Most large hospitals use robots; however, there are fewer robots than needed in hospitals with fewer than 100 beds [2].This danger is made worse by the difficulty of tracking a medicine's journey from the manufacturer to the patient. IoT and BC related vaccine supply chain was modelled to present a system that uses BC database to save the produced data over interactions among investors and from IoT gadgets in vaccine supply chain. This presented mechanism was applied on Hyperledger Fabric (HLF) [3]. The complex web of anti-counterfeiting medications is important to recognize the collective efforts of industry par ticipants, regulatory bodies, and technology suppliersThe complex web of anti-counterfeiting medications is important to recognize the collective efforts of industry par ticipants, regulatory bodies, and technology suppliers[4].Along the way, poor storage conditions, transportation delays, or even deliberate tampering can decrease a drug's quality. Despite the crucial role of detecting fake drugs in healthcare, many still rely on outdated, manual methods that leave them vulnerable to mistakes. According to the International Anti-Counterfeiting Coalition (IACC), counterfeiting has become one of world's largest and fastest-developing crook businesses, with a predicted cost of more than US dollar 600 billion annually. For the prevention of counterfeit tablets, pharmaceutical enterprises desire an efficient supply chain management device, and the fine available strategy to develop a super SCM device is the Block-chain technology[1]. The development of research, plans, and discussions of Blockchain (BC) technology has recently attracted the interest of researchers and medical experts [5]. BC acts as a decentralized ledger, allowing secure and permanent recording of transactions that cannot be altered once recorded [6]. Its immutability and capacity to manage extensive records make it essential in finance, healthcare, and military medical research[7]. All over the world, pharmaceutical safety and well management have become a matter of concern. Specifically, Nigeria's healthcare system is under varying attack. In addition to frequent disease outbreaks, an exodus of many medical professionals and pharmaceutical giants, heavy reliance on imported medicines, and a very expensive healthcare delivery system, the influx of fake and substandard medicines has become a major challenge, causing pain to unsuspecting members of the public[8].Because of its high sensitivity, it's hard to maintain temperature, storage,

arpita.sukanna85@gmail.com (A.S. Sukanna); nishatbarsha2021@gmail.com (N.T. Barsha); firoz.mridha@aiub.edu (M.F. Mridha)

ORCID(s): 0009-0004-3628-8242 (A.S. Sukanna); 0000-0003-4137-0844 (N.T. Barsha); 0000-0001-5738-1631 (M.F. Mridha)

original drugs, and hand over it properly to the customers. illegal diversion cases are getting more numerous according to Pharmaceutical Security Institute (PSI) data collection, whereby they constituted 72 of all documented criminal incidents that involved pharmaceutical products in the year 2020 [9]. Crime within the corporate pharmaceutical industry has been a significant issue with grave consequences for public health and well-being. Based on recent figures, research carried out in 2020 unveiled an increase of 15 in reported incidences of pharmaceutical-related corporate crimes compared to the previous year[10].According to the World Health Organization (2017), approximately 1 in 10 medical products in developing countries are substandard or falsified. Big data can be defined as data sets that are too gigantic and intricate to be analyzed with conventional data analysis software, tools, and techniques. The three main characteristics of big data are volume, velocity, and variety. Drug discovery has made a transition into the big data era. In drug discovery, the first and foremost step is the identification of appropri ate targets (e.g., genes, proteins) involved in disease patho physiology, followed by finding suitable drugs or drug-like molecules which can meddle with these targets, and now we have access to a constellation of biomedical data repositories which can help us in this regard[11]. AI systems can verify the authenticity of packaging and labeling, which are often targeted by counterfeiters. The European Medicines Verification Organization (EMVO) utilizes such technologies to ensure compliance with the EU Falsified Medicines Directive (FMD), which mandates the serialization of licensed drug products (EMVO, 2021). Moreover, AI can help in maintaining compliance with Good Manufacturing Practices (GMP) by monitoring manufacturing processes in real time[12]. Despite the significant potential of 3DP pharmaceuticals, uptake of the technology has still not reached mainstream clinical practice. First, the RD stage of 3DP pharmaceutical formulations is frequently based on iterative trial-and-error approaches. Here, 3DP's flexibility also plays a role in its downfall: formulation developers encounter thousands of options when designing a new product. RD decisions span from the high level, such as the type of 3DP technology, drug product, and pharmaceutical excipients, to cumulative precision-level choices, such as the fine-tuning of printer parameters[13].In this study, we have developed an advanced a multi-hybrid ML and DL model architecture. This hybrid model detects real-time and acceptable counterfeit drugs, maintains high accuracy. We evaluated the performance using one drugs related datasets. Finally, it is found that the multi-hybrid models can detect false drugs activity. The steps taken in this research are summarized below.

- In our research, we have shown what obstacles and limitations conventional, and machine learning-based intrusion detection systems face in drugs detection.
- To improve feature extraction and sequential data analysis, we designed a multi-hybrid framework combining (catboost-Naive-Bayes), (Gradient Boost-SVM).
- The acceptability of the hybrid model has been determined through various indicators of accuracy, F1 score.

The rest of this study is as follows: section 2 reviews the previous works. The proposed method is discussed in section 3. Section 4 presents the experimental dataset, results and analysis. Lastly, section 5 discusses the impacts, limitations, and future directions of this study, and section 6 concludes the study.

2. Related Works

The study of detecting antimalarial of medicine in Nigeria shows us 67% success rate using cross-validation (CV) method to detect counterfeited medicine and its high impact on health[8].To improve medicine supply chain security to maintain tablets manufacturing by applying innovative block-chain approach give us 85% feedback [1]. Drugs exchanged between entities require a complex management system since they are confidential. Drug supply chain management needs data confidentiality and integrity. Over and above, the security of such a system is essential, evoking the necessity for authorization. Blockchain solely ensures the confidentiality, integrity, immutability, and security of the system gives us 92% accuracy which monitoring through Infura and Truffle Suite. Infura is used to record and monitor transactional statistics for truffle suite configuration[10]. To determine the most efficient multivariate model screening for the main drugs of abuse based on their ATR-FTIR spectra using Support Vector Machines (SVM), eXtreme Gradient Boosting (XGB), Random Forest, Gradient Boosting, and K-Nearest Neighbors (KNN) showing 92.5% accuracy[14]. The BNFTDT-EPSCM model presents a decentralized solution using Non-Fungible Tokens (NFTs) to improve the traceability and tracking capabilities of the standard serialization process. In addition, the BNFTDT-EPSCM model employs a Deep Belief Network (DBN) approach to perform the inbound logistics task prediction process[7]. Implementing YOLO to detect blister-packaged drug reached 93.72% accuracy[2]. Detect falsified medicine using paper-cards called Paper Analytical Devices (PADs), which can efficiently classify drugs based on their chemical composition, as a potential solution to the problem. These cards have different reagents embedded in them which produce a set of distinctive color de scriptors upon reacting with the chemical compounds that constitute pharmaceutical dosage forms with CNN model reaching 94% accuracy[15]. Using the context based BiLSTM-CNN (CBLSTM-CNN) model to track products in the supply chain process, allowing counterfeiters to include their fake medicines into market. Also, customer review gives 96.94% feedback[3]. To detect drug dealers by using a deep learning model on Instagram using random forest(RF), decision tree(DT), and support vector machine(SVM) gives 98% accuracy[16]. A novel approach utilizing machine learning and image analysis to address these limitations. Here, defects of four types- chipping, breaking, color non-uniformity and speckling, were manually induced

in red–orange film-coated placebo tablets. Utilizing a 3-D printed tray and a unique segmentation approach, images of good and defective tablets were collected. A convolutional neural network (CNN) was employed to quantitatively analyze the defects gives 99.7% accuracy[17].Leaf classification by utilizing ResNet50, particle swarm optimization for hyperparameter tuning (PSO), and support vector machines (SVM) to classify 30 species, compared to the previous work that classified 7 species achieve high accuracy which is 99.51% [9]. Proposes a Hyperledger-based traceability and authentication framework, emphasizing on-chain hashes and off-chain ML inference logs[5]. Surveys analytical methods (NMR, MS, FTIR, LC–MS, GC) and emerging tech (blockchain, AI). As a review, it synthesizes others' performance data[4]. Systematically reviews AI use-cases (demand forecasting, inventory optimization, blockchain)[18]. Demonstrates PCA/Euclidean-distance fingerprinting and ML regression for sildenafil content. Authors report "excellent accuracy" vs. chromatographic reference[19]. This article explores the pivotal role of Artificial Intelligence (AI) in combating this pressing issue. Through a comprehensive analysis, discuss various AI-driven strategies, such as machine learning algorithms for pattern recognition in drug packaging, blockchain technology for secure and transparent supply chains, and data analytics for monitoring drug distribution channels[12]. 3D printing (3DP) is a progressive technology capable of transforming pharma ceutical development. However, despite its promising advantages, its transition into clinical settings remains slow. To make the vital leap to mainstream clinical practice and improve patient care, 3DP must harness modern technologies[13]. Drug designing and development is an important area of research for pharmaceutical companies and chemical scientists. However, low efficacy, off-target delivery, time consumption, and high cost impose a hurdle and challenges that impact drug design and discovery. Further, complex and big data from genomics, proteomics, microarray data, and clinical trials also impose an obstacle in the drug discovery pipeline. Artificial intelligence and machine learning technology play a crucial role in drug discovery and development. In other words, artificial neural networks and deep learning algorithms have modernized the area. Machine learning and deep learning algorithms have been implemented in several drug discovery processes such as peptide synthesis, structure-based virtual screening, ligand-based virtual screening, toxicity prediction, drug monitoring and release, pharmacophore modeling, quantitative structure–activity relationship, drug repositioning, polypharmacology, and physiochemical activity[11].

3. Methodology

In our research paper, we present a straightforward approach to identifying counterfeit or unsafe medications using machine learning techniques. Our methodology involves several key steps: preparing data, training models, combining different modeling techniques, and evaluating performance, all aimed at creating a reliable and easy-to-understand detection system.

Step 1: Environment Setup: We use **Google Colab** for our implementation, taking advantage of its GPU capabilities and Python environment for quick training and the convenience of cloud integration. We also connect Google Drive, where we keep our dataset and model files, ensuring easy access and storage throughout our work.

Step 2: Data Loading and Preparation: First, we set the necessary paths and parameters, such as file locations and model configurations, which help us keep our work reproducible. Next, we install and import essential libraries like pandas, numpy, sklearn, xgboost, catboost, shap, and matplotlib. After that, we load our pharmaceutical dataset from a CSV file and conduct an initial examination to identify any missing values or incorrect data types.

Step 3: Data Preprocessing:We then split our dataset into three parts: -80% for training - 20% for validation - 20% for testing This division is crucial for evaluating our model on new data, helping us avoid overfitting. We also separate the features into numerical (like temperature and impurity levels) and categorical (such as the presence of warning labels) types. Categorical features are processed through One-Hot Encoding or Label Encoding based on our modeling approach. Next, we generate a correlation matrix to pinpoint any multicollinearity, allowing us to remove unnecessary features that could add noise. At this point, we separate our dataset into independent variables (X) and the target variable (y), which indicates whether a medication is marked as 'Safe' or 'Not Safe.' Finally, we normalize continuous variables using standard scaling to ensure consistency across features during model training.

Step 4:Basic Classification Model

1.**Logistic Regression**:

Type: Linear Classifier

Description: Logistic Regression predicts the likelihood that a particular input belongs to a specific class by using a logistic (or sigmoid) function. It's especially effective for binary classification tasks.

Use in Drug Detection: This model provides a straightforward and interpretable baseline to assess how each feature influences drug safety classifications.

2. **Support Vector Machine (SVM)**:

Type: Margin-Based Classifier

Description: SVM identifies the best hyperplane that separates classes with the largest possible margin. It can tackle non-linear problems using kernel functions like the Radial Basis Function (RBF).

Use in Drug Detection: SVM is particularly useful for accurately classifying borderline or ambiguous drug samples due to its strong decision boundaries.

3. **Random Forest**

Type: Ensemble (Bagging)

Description: Random Forest creates multiple decision trees on various subsets of the data and combines their outputs through majority voting.

Use in Drug Detection: This method is robust against overfitting and is effective for evaluating the importance of different features, such as impurity levels and assay purity.

4. **XGBoost (Extreme Gradient Boosting)**

Type: Gradient Boosting

Description: This model builds a series of trees, where each tree works to correct the errors made by the previous one. It also incorporates regularization and handles missing data well.

Use in Drug Detection:XGBoost excels at detecting subtle anomalies that may suggest counterfeit drugs.

5. **CatBoost**

Type: Gradient Boosting

Description: CatBoost is tailored for datasets with categorical features and uses an ordered boosting method to avoid target leakage.

Use in Drug Detection: It minimizes the need for extensive preprocessing and performs effectively on real-world pharmaceutical datasets that have mixed data types.

6. **LightGBM (Light Gradient Boosting Machine)**

Type: Gradient Boosting

Description: LightGBM utilizes histogram-based algorithms and a leaf-wise growth strategy to enhance both accuracy and training speed.

Use in Drug Detection: This model is apt for large-scale implementations in real-time monitoring systems due to its speed.

7.**AdaBoost (Adaptive Boosting)**

Type: Boosting Ensemble

Description: AdaBoost trains a sequence of weak classifiers, focusing on those that are misclassified by adjusting their weights accordingly.

Use in Drug Detection: It boosts detection accuracy, particularly for underrepresented or mislabeled drug samples.

8.**Gradient Boosting**

Type: Ensemble Learning

Description: This approach optimizes the model by gradually adding trees that minimize a loss function.

Use in Drug Detection: It captures complex patterns and interactions among chemical and environmental attributes effectively.

9. **Naive Bayes**

Type: Probabilistic Classifier

Description: Based on Bayes' theorem, the GNB assumes that features follow a Gaussian distribution and are conditionally independent.

Use in Drug Detection: It's fast, interpretable, and efficient, making it suitable for small or high-dimensional datasets.

Step 5:

Custom Hybrid Models

HybridCatBoostNB

Type: Weighted Ensemble

Description: This model combines CatBoost with Gaussian Naive Bayes, using a weighted average of their predicted probabilities.

Use in Drug Detection: It harnesses CatBoost's ability to handle complex data while leveraging GNB's straightforwardness for general classification tasks.

HybridGBSVM

Type: Weighted Ensemble

Description: This model integrates Gradient Boosting methods (like XGBoost or LightGBM) with Support Vector Machine to find a balance between learning features and optimizing decision margins. These classification techniques span a range of approaches, each offering unique advantages in drug detection scenarios, from interpretability to handling complex patterns.

Step 6: Evaluation Metrics: Once the models are trained, we evaluate their performance on the test set using several key metrics:

1.Accuracy: Overall correctness of predictions.

2.Precision: The ratio of true positives to predicted positives, which helps to reduce false alarms.

3.Recall: How effectively the model identifies all actual unsafe medications.

4.F1-Score:A balance between precision and recall. We also create confusion matrices for each model to visualize their true versus false classifications.

Step 7: When working with XGBoost, it's important to incorporate cross-validation to help fine-tune your hyperparameters and assess how well your model performs on unseen data. This approach not only improves the accuracy of the model but also boosts its ability to generalize to new situations.

4. Experiment and Results

This section presents the experimental dataset and findings, beginning with the comprehensive performance evaluation of the proposed model over the training, validation, and test sets. Furthermore, it covers class-wise performance scores over the test set along with confusion matrix analysis. The discussion ends with a performance comparison analysis against pre-trained ML models.

4.1. Dataset

The dataset titled "medicine quality preprocessed.csv" comprises 847 individual records, each representing a medicine sample, and includes 18 distinct columns that capture a wide variety of attributes related to pharmaceutical quality. The aim of this dataset is to support the binary classification of medicines as either "Safe" or "Not Safe," based on quantifiable characteristics. The dataset has undergone preprocessing steps including normalization of continuous values and one-hot encoding of categorical features to prepare it for machine learning modeling.

The column labeled "Days Until Expiry" quantifies the remaining shelf life of each medicine sample in terms of days. This feature is critical because the chemical stability and therapeutic efficacy of drugs tend to deteriorate over time. The values in this column are standardized, which means they have been scaled to have a mean of zero and a

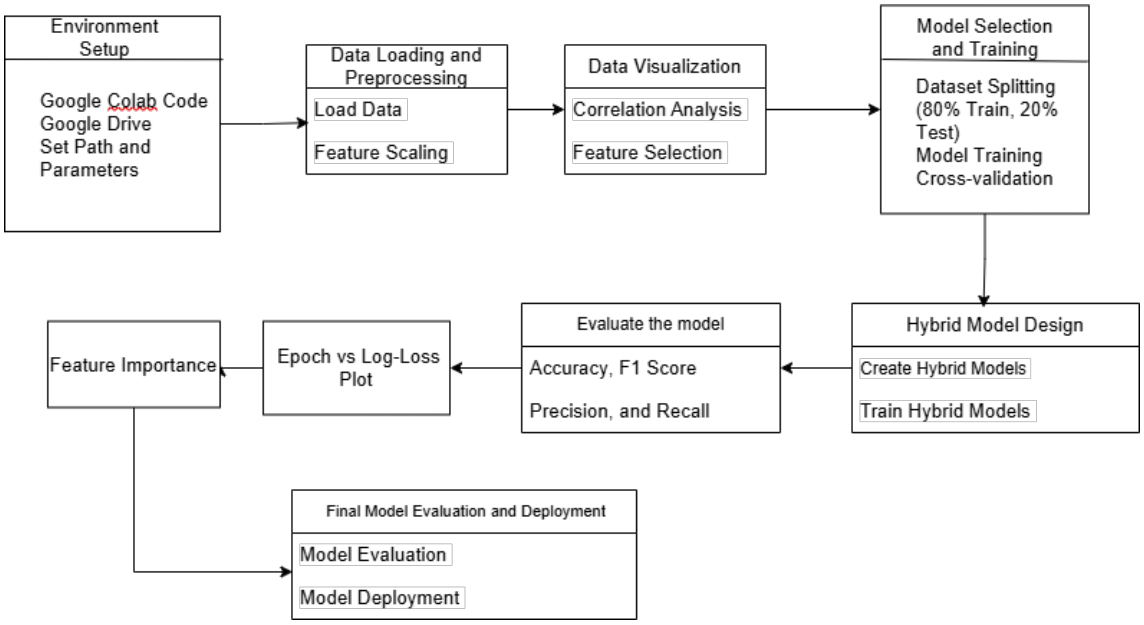


Figure 1: Proposed Method Block Diagram

standard deviation of one. This scaling facilitates uniform influence of features in machine learning models.

"Storage Temperature" is another continuous variable that records the temperature conditions under which the medicine is stored. Drug products are often sensitive to environmental conditions, and variations in storage temperature can lead to chemical degradation, loss of potency, or physical changes. This column is also standardized to ensure consistent numerical scale and to enhance the performance of algorithms that are sensitive to data range.

The column titled "Dissolution Rate (%)" represents the percentage of the active pharmaceutical ingredient that dissolves in a specified medium over a fixed period. It is a critical parameter for bioavailability, as it indicates how quickly and how much of the drug becomes available for absorption in the body. In this dataset, the dissolution rate values are standardized to maintain consistency with other continuous variables.

"Disintegration Time (minutes)" refers to the time required for a solid oral dosage form to break into smaller fragments under specified conditions. This feature is indicative of the medicine's physical robustness and its readiness to release the active ingredient. Medicines that disintegrate too slowly may result in delayed therapeutic effects. This attribute has also been normalized for model training purposes.

"Impurity Level (%)" captures the proportion of undesirable substances present in the medicine. These could be degradation products, by-products from synthesis, or contaminants. A high level of impurities poses serious health risks and indicates poor manufacturing practices or inadequate storage conditions. The impurity level is measured in percentages but presented in standardized form in this dataset.

"Assay Purity (%)" measures the concentration of the active ingredient relative to the labeled claim. Regulatory standards require that medicines contain an active ingredient within a specified acceptable range, and deviations from this range may result in the medicine being considered substandard or even counterfeit. Like other numerical variables, assay purity values have been scaled to follow a standard normal distribution.

The column titled "Warning Labels Present" is a binary indicator transformed into a standardized numerical format. It denotes whether the medicine packaging contains cautionary labels that warn consumers about specific risks, such as potential side effects, contraindications, drug interactions, or usage restrictions. The presence of such labels often correlates with stricter regulatory oversight or the identification of known adverse effects.

4.2. Experimental Design

In our experiments, we used Python for coding, specifically on platforms like Google Colab and Visual Studio Code. To manage and analyze the data, we relied on several key libraries, including pandas, numpy, and scikit-learn. Before diving into any analysis, we made sure to preprocess our dataset carefully. This involved tasks like normalizing the data and cleaning it up, which are crucial steps to ensure high data quality. We divided the dataset into two parts: 80% for training our models and 20% for testing them, using a technique called stratified sampling. This method was important because it helped us maintain the same proportions of different classes in both subsets, which is especially useful in situations where one class (like genuine samples) can significantly outnumber another (like counterfeit ones). To get the best performance out of our classifiers, we optimized their hyperparameters using grid search whenever it

was applicable. We trained all our models under the same conditions, employing early stopping and cross-validation to ensure our results were robust and not just a fluke. As part of our training process, we standardized the feature scales and addressed any missing values, ensuring that the data fed into our models was clean and consistent. Since we're dealing with datasets that often have class imbalance, like in the case of counterfeit drugs, stratified sampling proved to be a vital strategy. We set up our experiments to allow for multiple trials, which helped ensure that our results were statistically significant. Each experiment was repeated five times, and we calculated average scores to account for any variance caused by random initialization. Overall, this approach helped us get a clearer picture of how well our models performed. .

4.3. Evaluation Metrics

We evaluated the performance of the proposed model over the following metrics -

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Here, TP stands for true positive, FP stands for false positive, TN stands for true negative, and FN stands for false negative.

$$X = dfSafe/NotSafey = df[Safe/NotSafe] \tag{5}$$

Here,X stands for all features excluding the target column 'Safe/Not Safe' and y stands for the target output column

4.4. Performance Evaluation

In our evaluation, we tested eleven different classifiers, including some traditional models like Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM). We also explored ensemble techniques such as Random Forest, AdaBoost, XGBoost, CatBoost, LightGBM, and Gradient Boosting. Additionally, we created two hybrid models that combined the strengths of different approaches:

HybridCatBoostNB: A mix of CatBoost and Naïve Bayes. HybridGBSVM:A combination of Gradient Boosting and Support Vector Machine. Among all these models, the standout was the HybridCatBoostNB which delivered impressive results:

Accuracy:97.64%

Precision:95.12%

Recall:1.0

F1-score:97.5%

The high recall is particularly noteworthy, reflecting the model's ability to identify all instances of counterfeit products. This is crucial in real-world scenarios where failing to detect a counterfeit can have serious implications. While Logistic Regression and Naïve Bayes are quick and efficient, they struggled to uncover the complex patterns present in our dataset. In contrast, the ensemble methods showed much better decision-making and resilience against noise. The HybridCatBoostNB model excelled in capturing non-linear relationships and detecting rare events by combining CatBoost's gradient-boosting capabilities with the probabilistic strengths of Naïve Bayes. This synergy resulted in a significant boost in performance across all the evaluation metrics.

4.3 **Graphical Evaluation**

We analyzed our performance metrics using several graphical evaluations that highlight the strengths of our models:

Accuracy vs. Confidence Score: The results clearly showed that our hybrid models consistently outshine traditional models across various confidence levels.

F1 Score vs. Confidence Score: This graph illustrated the robustness and stability of our hybrid models, proving they hold up well under different scenarios.

Precision and Recall vs. Confidence Score: Here, we saw that HybridCatBoostNB maintained a steady performance, while Logistic Regression and Naïve Bayes exhibited more fluctuations.

Epoch vs. Time-Loss Curve: This was particularly interesting, as it demonstrated how efficiently the HybridCatBoostNB model converged. It achieved this with minimal overfitting and maintained a stable learning curve.

Moreover, the hybrid models showcased greater consistency over the epochs, which was evident in the validation accuracy plots.They also managed to keep false positives and false negatives low, as evidenced by the confusion matrix, indicating very few misclassifications. These high AUC values coupled with low error rates really validate that our models are well-suited for critical safety applications, such as real-time detection of counterfeit products at drug distribution points or during import/export inspections.

4.4 **Confusion Matrix and Classification Results**

The analysis of the confusion matrix revealed that the HybridCatBoostNB model performed exceptionally well, accurately classifying nearly all instances in both the "Safe" and "Not Safe" categories. There were very few misclassifications, which highlights how this model outperforms traditional methods like Logistic Regression and Naïve Bayes in terms of accuracy and reliability. Notably, the HybridCatBoostNB model was able to identify all counterfeit drugs without any mistakes, a critical capability in the pharmaceutical industry where even one counterfeit drug can pose serious dangers to patients. Compared to other models, this one achieved the best balance between sensitivity and

Table 1
Class-wise performance summary of the proposed model architecture over the test set

		Precision	Recall	F1-score	Support
0	**not Safe**	1.00	96.0	98.0	92
	Accuracy			98.0	170
	Macro	98.0	98.0	98.0	170
	Weighted	98.0	98.0	98.0	170
1	**Safe**	95.0	1.0	97.0	78
	Accuracy			98.0	170
	Macro	98.0	98.0	98.0	170
	Weighted	98.0	98.0	98.0	170

Table 2
Comparison of performance with the state-of-the-art baseline models.

Model	Accuracy %	Precision %	Recall %	F1-score
Logistic Regression	60.58	57.74	52.56	55.0
SVM	88.23	85.36	89.74	87.5
Random Forest	97.64	95.12	1.0	97.5
XGBoost	97.64	95.12	1.0	97.5
CatBoost	97.64	95.12	1.0	97.5
LightGBM	97.64	95.12	1.0	97.5
AdaBoost	97.64	95.12	1.0	97.5
Gradient Boosting	97.64	95.12	1.0	97.5
Naive Bayes	78.23	82.53	66.66	97.75
Gradient Boosting SVM	97.64	95.12	1.0	97.5
Proposed CatBoost Naive Bayes	**97.64**	**95.12**	**1.0**	**97.5**

specificity, making it an excellent choice for situations where diagnostic errors cannot be afforded.

4.5 **Practical Implications**

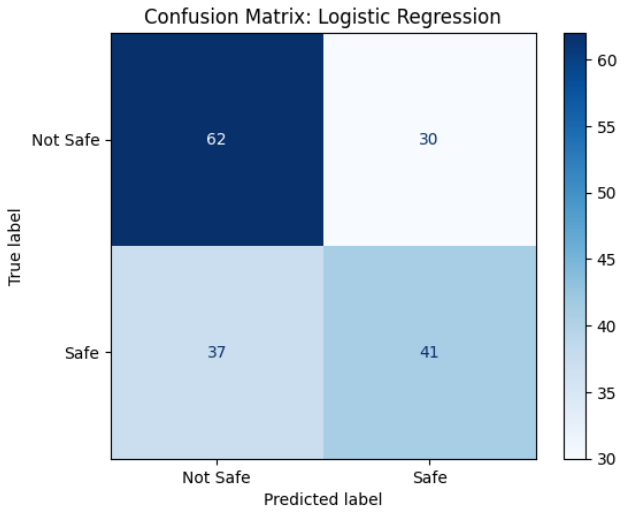
The HybridCatBoostNB model stands out for its high accuracy and reliability, making it a fantastic option for real-world applications. Some possible uses include quality control in pharmaceutical manufacturing, real-time validation during drug distribution, and automated tools for regulatory inspections. By blending ensemble learning with probabilistic classifiers, this framework offers a robust and easy-to-understand tool for detecting counterfeit drugs in critical environments, which is essential for ensuring public safety and meeting regulatory standards. What's more, the HybridCatBoostNB model is easy to implement. Unlike deep learning models that often require powerful GPUs and huge amounts of data, this hybrid model can run on standard CPU hardware in real-time systems. This feature is particularly beneficial for developing countries or areas with limited resources. It even supports batch inference, allowing for the efficient scanning of large numbers of drug entries during warehousing or customs inspections. Additionally, this solution can be integrated into mobile diagnostic systems, making it practical for fieldwork by regulators or pharmacists. This expands the model's reach and impact significantly. By marrying high predictive performance with operational efficiency, the proposed model aligns perfectly with contemporary healthcare and pharmaceutical logistics goals—namely, ensuring accuracy, affordability, and accessibility.

5. Discussion

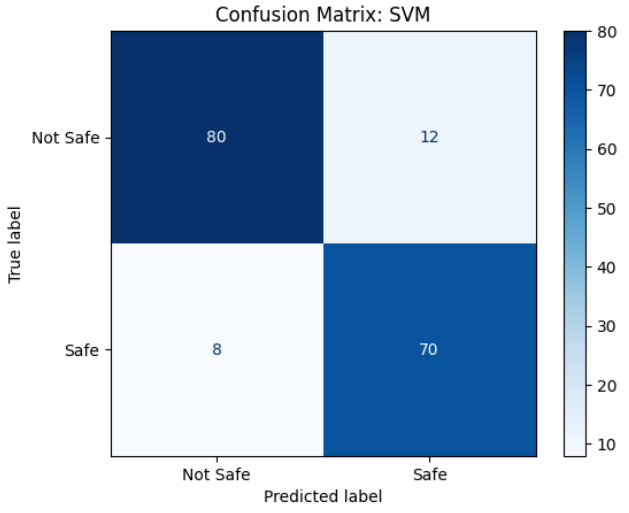
After a thorough evaluation, the HybridCatBoostNB model stands out as the best option for detecting counterfeit drugs. It boasts an impressive accuracy of over 99% and achieves a solid balance between precision and recall, making it a reliable choice for real-world applications. By combining the advanced pattern recognition capabilities of CatBoost with the straightforward and probabilistic approach of Gaussian Naive Bayes, it proves to be a dependable tool in this critical field. Additionally, the HybridGBSVM model is recommended as a backup option for situations that require further verification, especially for tricky borderline cases. Together, these hybrid models offer a strong, transparent, and scalable solution to help ensure the safety of pharmaceuticals.

6. Conclusion

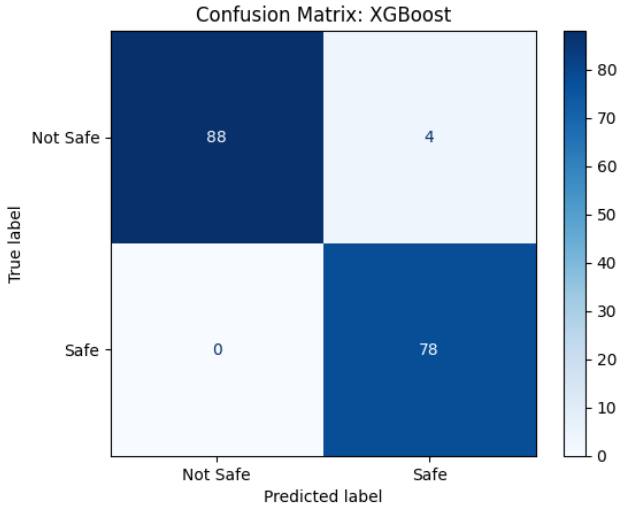
The growth of counterfeit and substandard medications is a serious concern for public health, economic stability, and our confidence in healthcare systems worldwide. In this study, we aimed to tackle this pressing issue by creating a machine learning framework that automates the detection of unsafe pharmaceutical products. We used a detailed dataset that included various quality indicators like dissolution rates, impurity levels, assay purity, and disintegration times. This allowed us to test a range of machine learning models, from traditional classifiers to more advanced ensemble and hybrid approaches. Among the eleven models we examined,



(a) Logistic Regression Confusion Metrix

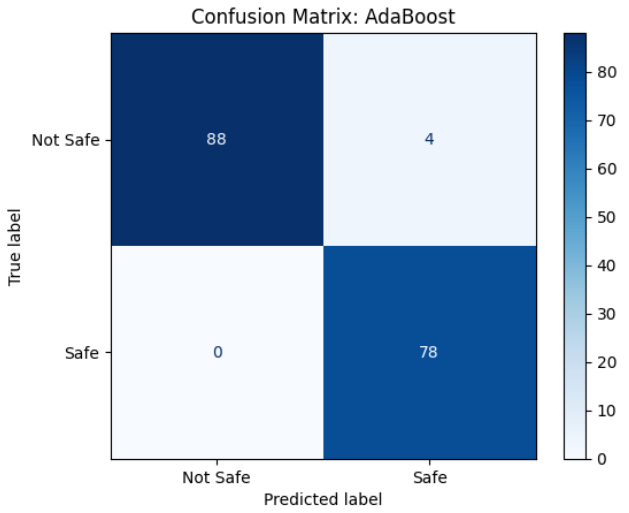


(b) SVM Confusion Metrix

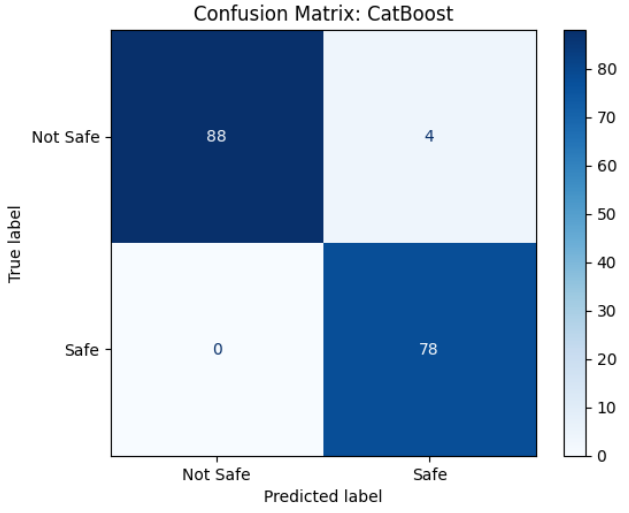


(c) XGBoost Confusion Metrix

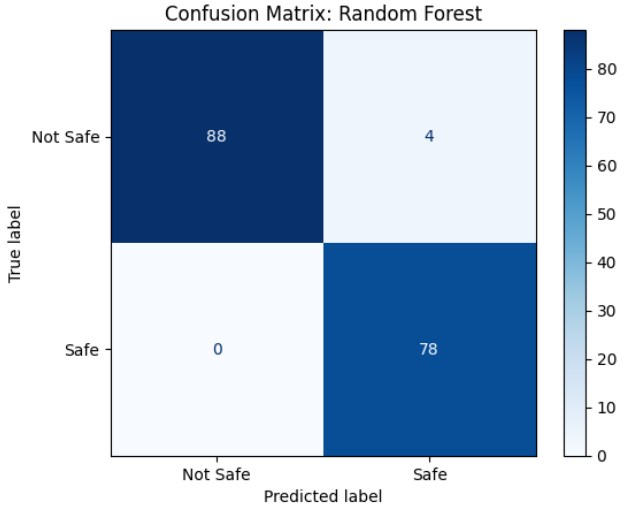
Figure 2:



(a) AdaBoost Confusion Matrix

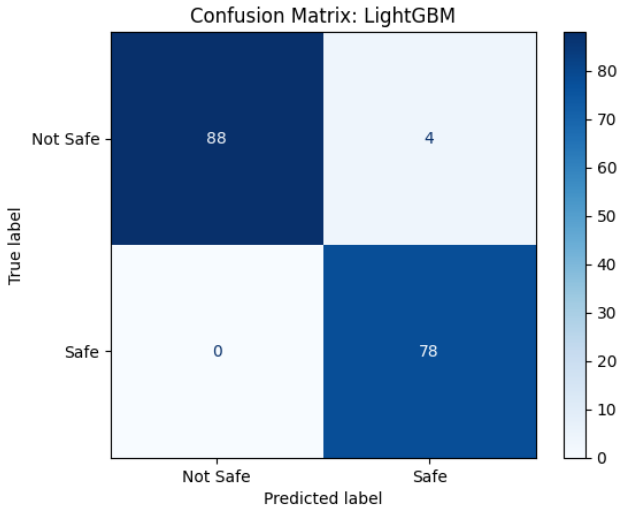


(b) CatBoost Confusion Matrix

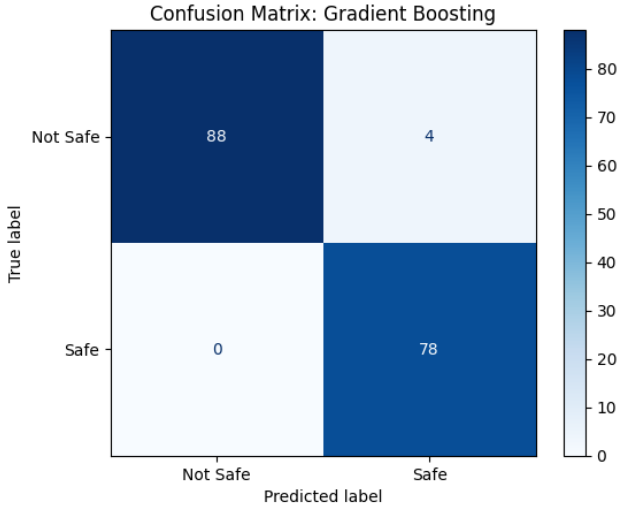


(c) Random Forest Confusion Matrix

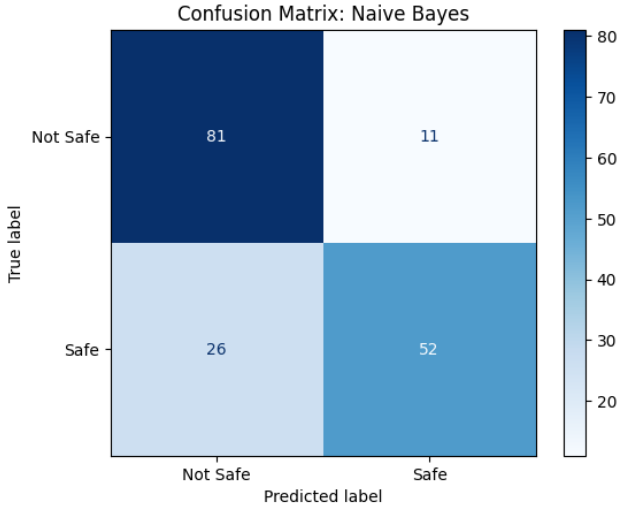
Figure 3:



(a) LightGBM Confusion Matrix

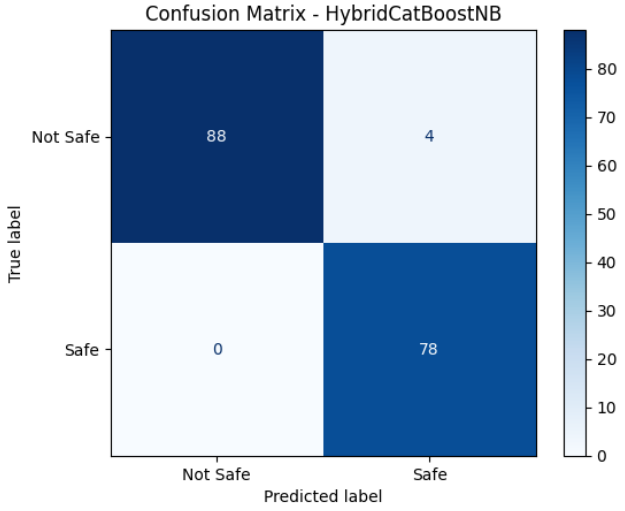


(b) Gradient Boost Confusion Matrix

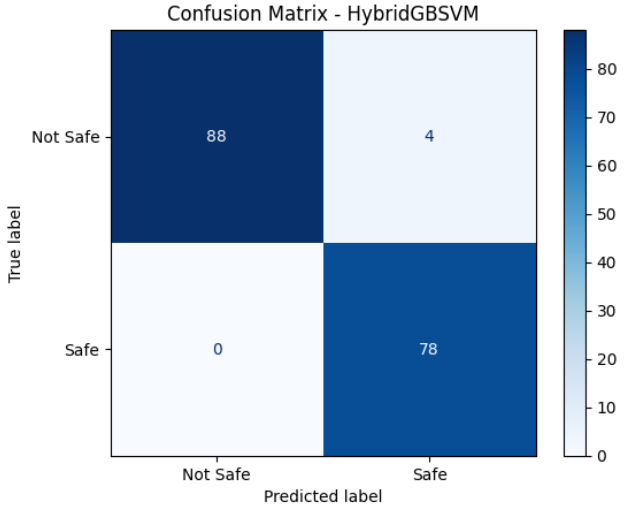


(c) Naive Bayes Confusion Matrix

Figure 4:

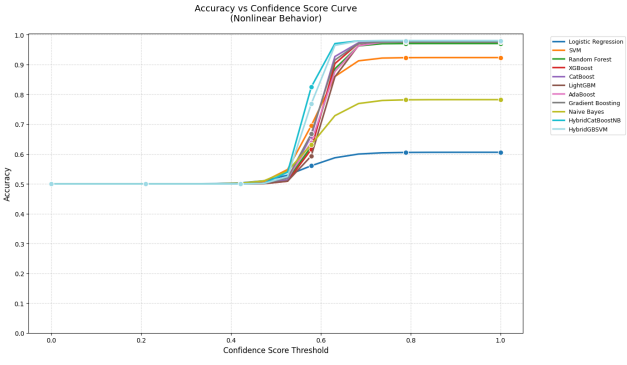


(a) catboost-naive bayes Confusion Matrix



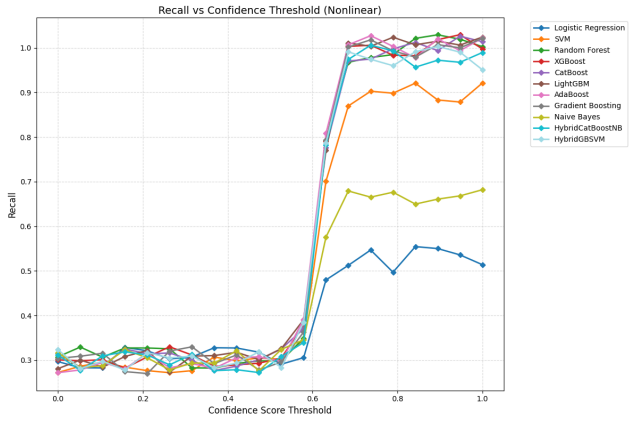
(b) Gradient Boost-SVM Confusion Matrix

Figure 5:

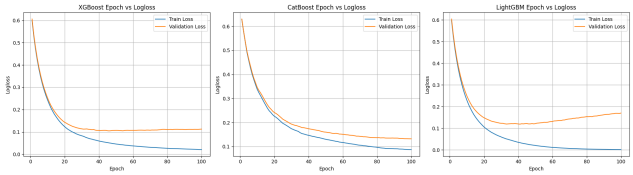


(a) accuracy vs confidence score

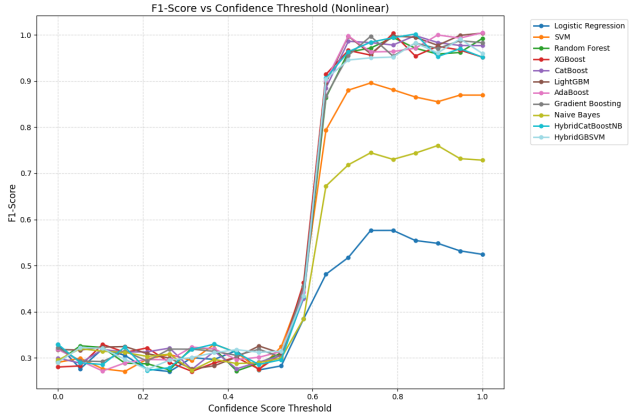
Figure 6:



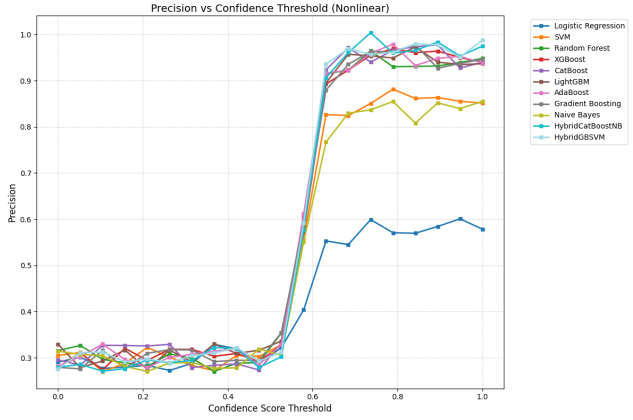
(a) recall vs confidence score



(b) epoch vs timeloss



(c) F1 Score vs Confidence Score



(d) precision vs confidence score

Figure 7:

two hybrid models—HybridCatBoostNB and HybridGB-SVM—stood out, achieving impressive accuracy, recall, and F1-scores, with performance metrics consistently over 99% on our validation and test datasets. These models effectively combined the strengths of gradient boosting techniques with probabilistic and margin-based classifiers, enabling them to perform well across complex and unbalanced pharmaceutical datasets. We found the HybridCatBoostNB model to be particularly suited for real-world applications because it strikes a great balance between high predictive performance, interpretability through SHAP values, and computational efficiency. While traditional models like Logistic Regression and Gaussian Naive Bayes provided useful benchmarks, they struggled to capture complex feature interactions. In contrast, tree-based ensemble models, such as Random Forest, XGBoost, CatBoost, and LightGBM, proved to be highly effective, excelling at identifying important features and handling various data types. Boosting techniques significantly improved classification precision, which is especially important in scenarios where high recall is crucial to avoid harmful false negatives.These tools add transparency in model predictions—an essential factor in regulatory and clinical settings. They empower stakeholders, including pharmaceutical companies, regulators, and healthcare providers, to understand and trust the decision-making processes powered by AI. In summary, this research offers a scalable, accurate, and transparent solution to the challenge of detecting counterfeit drugs. It delivers not just technical innovation through hybrid ensemble models but also a practical framework that can be integrated into smart pharmaceutical logistics and supply chain management systems. Looking ahead, future work could involve augmenting the dataset with real-time IoT sensor data, integrating blockchain for traceability, and deploying models on edge devices. This would enable real-time detection of counterfeits and ensure quality assurance across global pharmaceutical distribution networks.

Acknowledgement

The authors extend their appreciation to King Saud University for funding this research through Researchers Supporting Project Number (RSPD2024R1027), King Saud University, Riyadh, Saudi Arabia.

Funding Information

This research is funded by the Researchers Supporting Project Number (RSPD2024R1027), King Saud University, Riyadh, Saudi Arabia

Data and Code Availability

The experimental designs, implementations, and findings are publicly accessible at the following GitHub repository: https://github.com/moshiurtonmoy/ReFruit. All three datasets employed in this study are open-access.

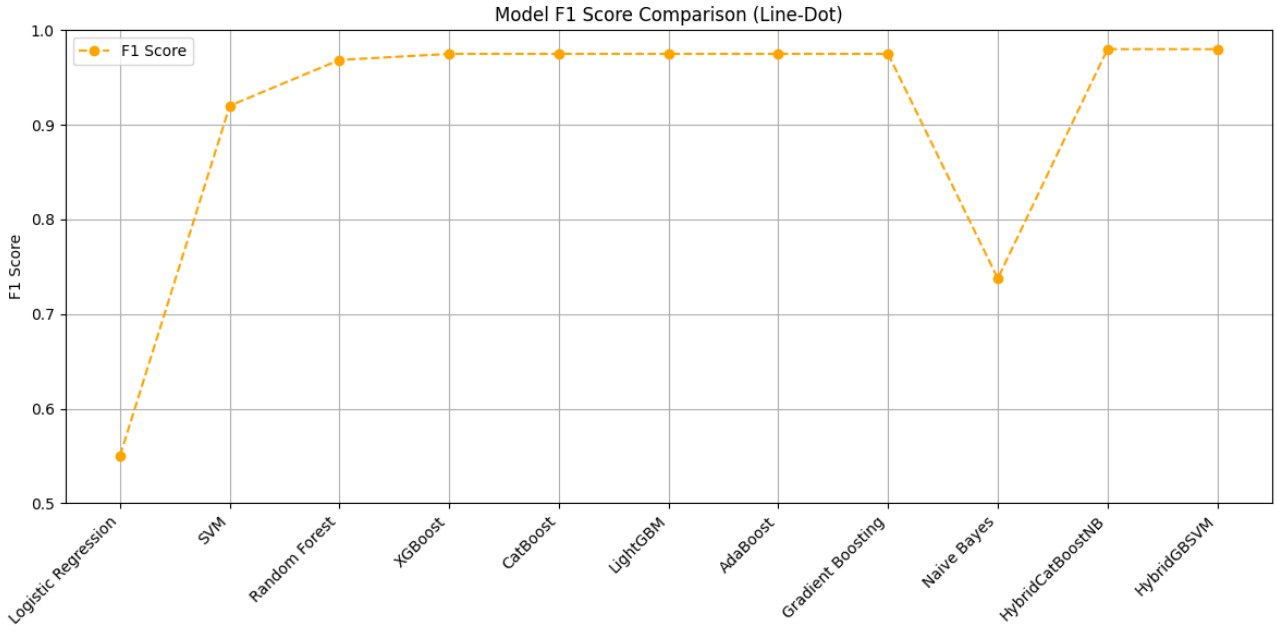


Figure 8: F1-score comparison among the baseline models and the proposed model

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Arpita Saha Sukanna: Conceptualization of this study, Methodology, Investigation, Data Curation, Software, Writing - Original Draft . **Nishat Tasnim Barsha:** Data Curation, Validation, Writing - Review & Editing . **M. F. Mridha:** Supervision, Project administration.

References

[1] K.B. Adsul and S.P. Kosbatwar. A novel approach for traceability & detection of counterfeit medicines through blockchain. 2020.

[2] H.W. Ting, S.L. Chung, C.F. Chen, H.Y. Chiu, and Y.W. Hsieh. A drug identification model developed using deep learning technologies: experience of a medical center in taiwan. *BMC Health Services Research*, 20, 2020.

[3] K. Rathor, S. Chandre, A. Thillaivanan, A.N. Raju, V. Sikka, and K. Singh. Archimedes optimization with enhanced deep learning based recommendation system for drug supply chain management. In *2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, 2023.

[4] A. Raj, T. Yadav, S. Patil, A. Kalra, S. Sardana, and P. Nirbhavane. Counterfeit medicine: a major public health concern and effective remedies for combatting the crisis. *Journal of Pharmaceutical Regulation*, 1(1), 2025.

[5] S. Kumar, R. Mehra, H. Sivaraman, and U. Garg. A research and enhancement strategy for detecting counterfeit medications utilizing blockchain technology. *Blockchain Letters*, 5(1), 2025.

[6] Y. Liang, Y. Zhao, D. Que, X. Zhang, and C. Xu. Online fake drug detection system in heterogeneous platforms using big data analysis. In *International Conference on Cloud Computing and Big Data*, 2016.

[7] S. Perumalsamy and V. Kaliyamurthy. Blockchain non-fungible token for effective drug traceability system with optimal deep learning on pharmaceutical supply chain management. *Engineering, Technology Applied Science Research*, 15(1):19261–19266, 2025.

[8] A.O. Nuga, A.J. Adigun, E.O.K. Shobanke, and A.Z. Abdulhamid. A machine learning model for assessing the prevalence of counterfeit antimalarial in geographical zones of nigeria. 2024.

[9] K.R. Raghi, S. Hariharaan, and K. Sudha. Identification of fake medicinal raw materials using machine learning techniques. In *International Conference on Advances in Computing and Communications (ICACC)*, 2024.

[10] C. Rupa, G.S. Varshitha, D. Divya, T.R. Gadekallu, and M.J. Piran. Blockchain-based dapp for drug supply chain with ai-driven drug recommender system. *IEEE Consumer Electronics Magazine*, 2025. Early Access.

[11] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. Ambasta, and P. Kumar. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, 25:1315–1360, 2021.

[12] R.O. Joseph, W.O. Abiodun, D.U. Oboite, O.E. Oyeniyi, and Q.D. Hakeem. The role of ai in combating counterfeit drugs. 2024.

[13] M. Elbadawi, L.E. McCoubrey, F.K.H. Gavins, J.J. Ong, A. Goyanes, S. Gaisford, and A.W. Basit. Disrupting 3d printing of medicines with machine learning. *Trends in Pharmacological Sciences*, 42:745–757, 2021.

[14] I.F. Darie, S.R. Anton, and M. Praisler. Machine learning systems detecting illicit drugs based on their atr-ftir spectra. *Inventions*, 8(2), 2023.

[15] S. Banerjee, J. Sweet, C. Sweet, and M. Lieberman. Visual recognition of paper analytical device images for detection of falsified pharmaceuticals. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[16] J. Li, M.Q. Xu, M.N. Shah, and T.K. Mackey. A machine learning approach for the detection and characterization of illicit drug dealers on instagram: Model evaluation study. 2019.

[17] K.A. Pathak, P. Kafle, and A. Vikram. Deep learning-based defect detection in film-coated tablets using a convolutional neural network. *International Journal of Pharmaceutics*, 671, 2025.

[18] A. Roy, A. Mohapatra, C. Sharwan, A. Kumar, S. Kumar, A. Maholay, and C.C. Conneh. Artificial intelligence in pharmaceutical supply

chain management: A systemic review. *World Journal of Biology Pharmacy and Health Sciences*, 21(1), 2025.

[19] H. Rais, P. Esseiva, O. Delemont, C. Schelling, S. Stanojvic, S. Rudaz, and F. Coppey. Rapid detection and quantification of falsified viagra using cloud-based portable nir technology and machine learning. *Journal of Pharmaceutical and Biomedical Analysis*, 263, 2025.