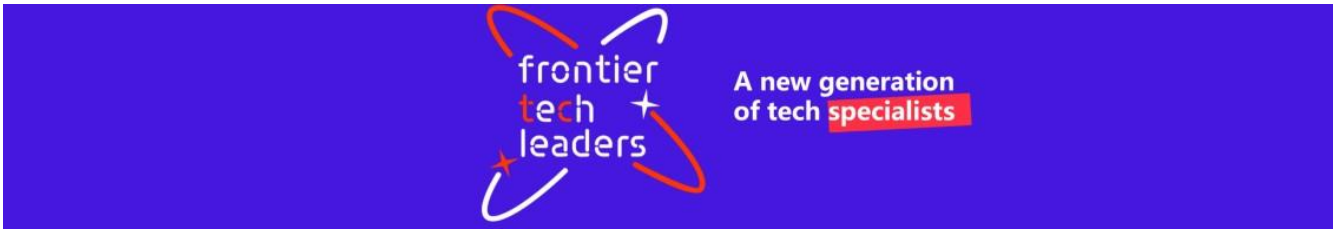

Capstone Project

Project : Machine Learning-Based Counterfeit Drug Detection Using Image Recognition and OCR

Literature Review, Data Research, Technology Review

Group Members's Name

1.Efrata Abebe

2.Hermela Hailegiogris

3.Mihret Tamene

April 10, 2025

Addis Ababa, Ethiopia

Machine Learning-Based Counterfeit Drug Detection Using Image Recognition and OCR

Literature Review

Introduction

The prevalence of counterfeit drugs poses a significant threat to public health globally, leading to ineffective treatments, adverse health outcomes, and increased mortality rates. As counterfeit medications become increasingly sophisticated, traditional detection methods often fall short, necessitating the exploration of innovative approaches. This research is crucial as it aims to address these challenges by investigating the use of advanced technologies, such as imaging and chemical composition analysis, to enhance the detection of counterfeit drugs.

A review of the existing literature is necessary to identify gaps in current knowledge and to understand the advancements made in counterfeit drug detection technologies. By analyzing studies that employ imaging techniques and chemical composition analysis, this review will provide insights into the effectiveness of these methods, the challenges faced in their implementation, and their potential impact on public health. Additionally, synthesizing existing research will inform future developments and strategies in combating counterfeit drugs, ensuring that efforts are grounded in established findings and best practices.

Chronological Organization of Literature Review on Fake Drug Detection

- “Fast Detection and Identification of Counterfeit Antimalarial Tablets by Raman Spectroscopy" M. De Veij (2007)
- "Detection of Counterfeit Drugs Using Electronic Nose Technology" by J. H. Lee (2018)
- "Image-Based Detection of Counterfeit Pharmaceuticals” by R. Kumar (2020)
- "Blockchain Technology for Detecting Falsified and Substandard Drugs in Distribution: Pharmaceutical Supply Chain Intervention"(2021)
- "A Novel Image Recognition Framework for Counterfeit Drug Detection" by M. J. Lee (2022)

Summary and Synthesis

1. **"Fast Detection and Identification of Counterfeit Antimalarial Tablets by Raman Spectroscopy" by M. De Veij (2007)**
    a) **Key Findings**: The study demonstrates that Raman spectroscopy can effectively identify counterfeit antimalarial tablets by analyzing their chemical composition.
    b) **Methodology**: Utilized Raman spectroscopy to analyze the spectral profiles of various tablets, comparing counterfeit products to authentic ones.
    c) **Contribution**: Established Raman spectroscopy as a rapid and reliable method for field detection of counterfeit medications, particularly in low- resource settings.
2. **"Detection of Counterfeit Drugs Using Electronic Nose Technology" by J. H. Lee (2018)**
    a. **Key Findings**: The electronic nose (e-nose) technology can successfully detect counterfeit drugs by analyzing the volatile organic compounds emitted from them.
    b. **Methodology**: Employed an array of gas sensors to capture and analyze the scent profiles of pharmaceuticals, differentiating between genuine and fake products.
    c. **Contribution**: Introduced an innovative approach to counterfeit detection using olfactory technology, expanding the toolkit available for drug verification.
3. **"Image-Based Detection of Counterfeit Pharmaceuticals" by R. Kumar (2020)**
    a. **Key Findings**: Image processing techniques can be effectively utilized to detect counterfeit drugs by analyzing visual features of packaging and tablets.
    b. **Methodology**: Utilized computer vision algorithms to extract features from images and classify them as genuine or counterfeit.
    c. **Contribution**: Highlighted the potential of image recognition technology in drug safety, paving the way for smartphone applications in counterfeit detection.
4. **"Blockchain Technology for Detecting Falsified and Substandard Drugs in Distribution: Pharmaceutical Supply Chain Intervention" (2021)**
    a. **Key Findings**: Blockchain technology can enhance traceability and authenticity in the pharmaceutical supply chain, reducing the risk of counterfeit drugs.
    b. **Methodology**: Proposed a framework for integrating blockchain into existing supply chain systems to improve transparency and security.
    c. **Contribution**: Advocated for technological innovation in logistics and supply chain management to combat counterfeit drugs at a systemic level.

5. **"A Novel Image Recognition Framework for Counterfeit Drug Detection" by M. J. Lee (2022)**
    a. **Key Findings**: The framework developed enhances the accuracy of counterfeit drug detection through advanced image recognition techniques.
    b. **Methodology**: Combined deep learning algorithms with image datasets of pharmaceuticals to classify products based on visual characteristics.
    c. **Contribution**: Provided a sophisticated model for real-time detection of counterfeit drugs, emphasizing the role of artificial intelligence in drug verification.

Commonalities

All papers contribute to the broader goal of improving drug safety and public health by addressing the critical issue of counterfeit drug detection and emphasizing the importance of reliable methodologies. Additionally, each study leverages advanced technologies to enhance detection capabilities.

Differences

- **Methodological Approaches**: The first two papers use chemical and olfactory detection, while later studies focus on image recognition and blockchain technology.
- **Scope of Application**: Some emphasize immediate detection in clinical settings, whereas the blockchain paper addresses systemic supply chain issues.
- **Technological Complexity**: Image recognition studies utilize machine learning, representing a more complex approach compared to chemical and olfactory methods.

Conclusion

In summary, the literature on counterfeit drug detection underscores the urgent need for reliable methodologies to combat this persistent public health issue. Key takeaways include the diverse technological approaches employed ranging from Raman spectroscopy and electronic noses to image recognition and blockchain technology. Each method offers unique advantages, enhancing detection capabilities and contributing to improved drug safety.

The importance of this research lies in its potential to protect public health by ensuring the integrity of pharmaceuticals. As counterfeit drugs pose significant risks to patients and healthcare systems, developing effective detection methods is crucial.

My project aims to build on this existing body of knowledge by integrating advanced image recognition techniques with real-time detection capabilities. By leveraging machine learning and deep learning algorithms, this research will provide a sophisticated framework for identifying counterfeit drugs, thus enhancing the accuracy and efficiency of detection methods. Ultimately, this project seeks to contribute valuable insights and tools that can further safeguard public health against the dangers of counterfeit pharmaceuticals.

Data Research

Introduction

Counterfeit medicine, also known as fake medicine, is a medication or phramaceutical product that is produced and sold with the intent to deceive [1]. They are often designed to mimic genuine products in appearance**,** including packaging and labeling [1]. Counterfeit medications are unauthorized replicas of genuine drugs that may contain incorrect, harmful, or inactive ingredients, posing significant health risks and undermining global healthcare systems [2].

Counterfeit medications represent a growing and critical threat to public health on a global scale [2]. According to the Internationl Federatio of Pharmaceutical Manufacturers & Associations (IFPMA), the global market for counterfeit pharmaceuticals was estimated to be worth $200 billion annually, making it one of the most lucrative sectors of the global trade in illegally copied goods The world Health Organization(WHO) reports that at least 1 in 10 medicines in low- and middle-income countries are substandard or falsified and countries spend an estimated US$ 30.5 billion per year on substandard and falsified medical products [3].

Addressing the global challenge of counterfeit medications requires international collaboration, stricter regulatory measures, and the adoption of advanced technologies for drug verification and tracking to safeguard public health [2]. The goal of our project is to develop a machine learning-powered application that can detect fake pharmaceutical products based on packaging images and text extracted via OCR (Optical Character Recognition)**.** We also plan to explore barcode verification and chemical composition checking as extended features.

Research Questions

This project aims to address the following key research questions:

- What visual features (e.g., color, layout, printing quality) distinguish counterfeit drug packaging from authentic packaging?
- Can OCR-extracted text from drug packaging be reliably used to detect inconsistencies (e.g., spelling errors, wrong expiry date format) indicative of counterfeit drugs?
- Which machine learning models (e.g., YOLOv8, CNNs, OCR pipelines) are most effective in identifying fake drugs based on combined image and text data?

- What are the limitations of existing datasets in detecting fake drugs in low-resource regions like Ethiopia, and how can these limitations be addressed?

A comprehensive exploration of data is necessary to understand the intricate characteristics that differentiate genuine pharmaceutical products from their counterfeit counterparts. This research will encompass the identification of patterns and trends indicative of counterfeiting, the evaluation of various detection methodologies, and the selection of appropriate machine learning Furthermore, a deep understanding of the available data landscape including images of authentic and fake packaging, text extracted via OCR, and reference drug databases is key for recognizing the potential challenges and limitations in creating a resilient detection system.

Organization

The characteristics of genuine and fake drugs can be captured through various data types, each offering unique insights into a product's authenticity. The packaging details of pharmaceutical products offer a wealth of information that can be crucial in identifying counterfeits. These details include the drug's name, a list of active ingredients, the manufacturer, the expiry date, recommended storage conditions, and the name and address of the manufacturing company [4]. Security features such as barcodes, holograms, and tamper-evident seals are also important packaging elements that counterfeiters often attempt to replicate [4].

This data research is organized thematically into key components required for building a fake detection system. The section Include:

- Packaging image datasets
- OCR and text label data
- Barcode/QR code verification resources
- Drug chemical composition reference databases

Data Description

Packaging Image Data

1. **Roboflow Universe Image of Medicine Packaging**

   Format: Images (various), Annotations (multiple formats)

   Size: 2075 images

   Description: Object detection dataset of medicine packaging

   Relevance: relevant for training models to classify different types of medicine packaging.

   Link: https://universe.roboflow.com/fyp-hdzz0/medicine-packaging/dataset/8

2. **Kaggle Mobile-Captured Pharmaceutical Medication Packages**

   Format: Images (JPEG)

   Size: 3900 Images

   Description: Medication packages captured with mobile phones in real-world conditions

   Relevance: for training robust models that handle variations in image quality and real-world scenarios.

   Link: https://www.kaggle.com/datasets/aryashah2k/mobile-captured-pharmaceutical-medication-packages

3. **SIP.unige.ch (University of Geneva) Pharmaceutical packages**

   Format: Images (JPEG), SIFT/aKaZe descriptors, Fisher Vectors

   Size: Enrollment - 54,000 images. Recognition: 2,400 images

   Description: dataset of pharmaceutical package under controlled conditions with multiple views.

   Relevance: for training models to recognize genuine packaging from various perspectives

   Link: http://sip.unige.ch/projects/snf-200021-165672/pharmapack/

4. **Kaggle Medicine tablet pack image dataset**

   Format: Images(JPEG)

   Size: ~437 Images

   Description: Images of Indian medicine tablets and tablet boxes with details.

   Relevance: Useful for a multi-modal approach, analyzing both product and packaging.

   Link: https://www.kaggle.com/datasets/nitesh31mishra/medicine-tablet-pack-image-dataset

OCR and text label data

1. **HealthData.gov (FDA) Metadata for FDA Online Data Repository**

   Format: CSV in ZIP archive

   Description: Contains proprietary name, active ingredients, NDC, company name.

   Relevance: Structured data to verify OCR-extracted information.

   Link: https://catalog.data.gov/dataset/fda-drug-label-data

2. **FDA Full-text of drug product labeling**

   *Format***:** Web-based database (searchable text)

   *Size***:** >150,000 documents

   *Description***:** Searchable database of FDA-approved drug labeling.

   *Relevance***:** Allows direct comparison with OCR-extracted text.

   *Link***:** https://www.fda.gov/science-research/bioinformatics-tools/fdalabel-full-text-search-drug-product-labeling

3. **John Snow Labs NDC data and FDA labels databases**

   Description: Data package with NDC information and drug labels.

   Relevance: Could provide structured data for verification.

   Link: https://www.johnsnowlabs.com/marketplace/national-drug-code-and-drug-labels-database-data-package/

Barcode and QR Code Verfication

1. **FDA Information about drug products (identified by NDC)**

   Format: Web-based directory, ASCII data files

   Description: List of drugs manufactured, prepared, compounded, or processed for sale

   Relevance: Can be used to verify NDCs extracted from barcodes.

   Link: https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory

2. **UPCItemDB API**

   Format: JSON (API)

   Description: contains Product name, manufacturer, barcode data

   Relevance: Enables validation of scanned product codes

   Link: https://www.upcitemdb.com/#google_vignette

Drug chemical composition reference databases

1. **Drug.com Information on prescription and OTC medicines**

   Format: Online database

   Size: > 24,000 entries

   Description: Detailed information for consumers and professionals.

   Relevance: Readily accessible information on drug ingredients and properties.

   Link: https://www.drugs.com/

Data Analysis and Insights

One of the main challenges we encountered is the lack of publicly available datasets specifically build for training AI/ML models to distinguish between real and counterfeit medicines. While the Pharmaceutical Security Institute Counterfeit Incident System (PSI CIS) database, tracks global incidents, it relies on a mix of open and nonpublic data sources, which limits its accessibility for open research [5]. In response to this scarcity, some researchers have adopted the strategy of creating their own datasets through web scraping collecting packaging images from online sources and using image editing tools to generate synthetic examples of fake packaging [6].

Given these limitations, our project adopts a **multi-source strategy**: combining real-world data and region-specific references where available. We also plan to generate or collect our own dataset by **scraping packaging images from recognized pharmacies and trusted pharmaceutical websites.**

Conclusion

This data research highlights the importance of a multi-faceted approach in tackling the challenge of counterfeit drug detection. Due to the limited availability of public datasets specifically designed for this task, we identified and evaluated a variety of alternative data sources, including packaging images, drug label text, barcode databases, and chemical composition references. By combining publicly accessible global datasets and web-scraped resources, our project aims to build a reliable fake drug detection app.

Citation

[1] B. G. Banik, "Detection of Counterfeit Medicines Using Machine Learning Techniques," *Medicon Engineering Themes,* vol. 7, no. 6, 2024.

[2] J. Berg, "The Impact of Counterfeit Medications on Global Health," TrueMed, 13 February 2024. [Online]. Available: https://truemedinc.com/blog/the-impact-of-counterfeit-medications-on-global-health/.

[3] World Health Organization, "Substandard and falsified medical products," WHO, 2024.

[4] digitallink Connexum, "Pharmaceutical packaging: a guide to global standards and local requirements," Digital-Link, LLC, 2025. [Online]. Available: https://digital-link.com/guides/pharmaceutical-packaging/.

[5] T. K. Mackey, B. A. Liang, P. York and T. Kubic, "Counterfeit Drug Penetration into Global Legitimate Medicine Supply Chains: A Global Assessment," *The American Journal of Tropical Medicine and Hygiene,* vol. 92, no. 6, pp. 59-67, 2025.

[6] K. MOTWANI, R. DSOUZA, R. DSOUZA and J. JOSE, "Counterfeit Medicine Detection using Deep Learning," *INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY,* vol. 9, no. 3, 2022.

Technology Review

Introduction

Counterfeit pharmaceuticals pose a serious threat to global public health, particularly in low- and middle-income countries where regulatory oversight is often weaker. According to the World Health Organization (WHO), 1 in 10 medical products in such regions is substandard or falsified [1]. These fake drugs not only lead to treatment failures and drug resistance but also undermine the trust in healthcare systems.

This technology review explores how machine learning (ML) technologies—specifically image recognition and Optical Character Recognition (OCR)—can be applied to detect counterfeit drugs efficiently and affordably. The relevance of this review lies in designing a hardware-free, accessible, and scalable counterfeit detection system that leverages mobile phones and ML models for real-time verification.

Technology Overview

Purpose:

The purpose of integrating ML in counterfeit detection is to automate and enhance the process of verifying drug authenticity using features like visual packaging characteristics and textual content extracted from drug labels and blister packs.

Key Features:

- ✓ Image Recognition (CNNs): For analyzing pill color, shape, size, logos, imprints, and packaging design.
- ✓ OCR (e.g., Tesseract, EasyOCR): To extract printed text like drug name, batch number, expiry date.
- ✓ QR/Barcode Scanning: Verifying encoded data with official records.
- ✓ Cross-check with Database APIs: Authenticating details with pharma databases.

Common Use Cases:

- ❖ Mobile health apps for drug verification.
- ❖ Regulatory agency tools for spot checks.
- ❖ Pharmacy integration systems for supply chain inspection.

Relevance to our Project

This technology is central to a hardware-free counterfeit detection system, particularly:

- ✓ Accessible: Works on any smartphone with a camera.
- ✓ Affordable: Avoids spectrometers or lab-grade scanners.
- ✓ Efficient: Enables rapid verification at point-of-sale or consumption.
- ✓ Scalable: Easily deployed across regions via cloud and mobile applications.

By combining OCR and computer vision, the system can detect inconsistencies in text or visual features, two of the most common areas where counterfeiters fail to replicate authentic products.

Comparison and Evaluation

Technology	**Strengths**	**Weaknesses**	**Cost and Scalability**
Spectroscopy	Highly accurate chemical analysis	Requires costly lab equipment	Low scalability
RFID/Blockchain	Excellent traceability	Infrastructure-dependent, high setup cost	Medium
ML + Image + OCR	Low cost, camera-based, real-time detection	Needs large labeled datasets, can be spoofed	High Scalability

Given the project's constraints (no hardware), ML with computer vision and OCR is most suitable.

Use Cases and Examples

Mobile Apps for Drug Detection

A systematic review examined mobile apps like MedSnap and PharmaCheck that use cameras and image analysis to verify pills and packaging [2]. These tools have shown practical use in low-resource settings.

DLI-IT: Deep Learning for Drug Label Identification

A study proposed a multi-modal model combining image and text data using deep learning (CTPN + CNN) for pill label verification, achieving 88% accuracy [3].

Blockchain + AI Integration in Mongolia

The Mongolian government piloted a solution integrating blockchain with AI to improve transparency in drug tracking and verification [4].

Paper Analytical Devices + CNNs

Researchers used visual recognition with CNNs to classify color reactions from chemical paper devices for counterfeit detection, achieving over 94% accuracy [5].

AI-Based Fake Product Detection Systems

An AI model was used to detect fake consumer products (similar in principles) by analyzing visual discrepancies in branding and text [6].

Gaps and Research Opportunities

Despite strong progress, several gaps remain:

- ✓ Lack of Public Datasets: Few open-source labeled datasets of counterfeit vs genuine drug packaging.
- ✓ Evolving Counterfeiting Tactics: Models need frequent retraining with updated samples.
- ✓ Localization: Need to adapt models for local languages and pharmaceutical regulations.
- ✓ Integration with National Databases: For batch number and QR code validation.

Opportunities include:

- ✓ Creating open datasets from partnerships with local pharmacies.
- ✓ Exploring multimodal learning (image + OCR + metadata).
- ✓ Deploying adaptive models with continual learning.

Conclusion

This review concludes that ML-powered image recognition and OCR are promising, cost-effective tools for counterfeit drug detection. The technology:

- ✓ Enables real-time, mobile-based verification.

- ✓ Avoids the need for expensive hardware.
- ✓ Scales well across diverse environments.

For the proposed project, it offers the best balance of accuracy, cost, accessibility, and scalability, with high potential to make a real-world impact in public health.

Citations

1) World Health Organization (WHO). (2017). "1 in 10 medical products in developing countries is substandard or falsified." https://www.who.int/news/item/28-11-2017
2) Källander et al. (2021). Mobile apps for detecting falsified and substandard drugs: Asystematicreview.PLOSONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0246061
3) Wu et al. (2020). DLI-IT: A deep learning approach to drug label identification through image and text embedding. BMC Medical Informatics and Decision Making. https://d-nb.info/1212766504/34
4) Government of Mongolia. (2021). Counterfeit medicine detection using blockchain and AI. OECD OPSI. https://oecd-opsi.org/innovations/counterfeit-medicine-detection-using-blockchain-and-ai/
5) Bender et al. (2017). Visual Recognition of Paper Analytical Device Images for Detection of Falsified Pharmaceuticals. arXiv preprint. https://arxiv.org/pdf/1704.04251.pdf
6) Shaikh et al. (2022). An AI-Based Fake Products Identification System. ResearchGate. https://www.researchgate.net/publication/366445003