Capstone Project Concept Note and Implementation Plan

Project Title: Fake drug detection

Team Members

1. Efrata Abebe
2. Hermela Hailegeorgis
3. Mihret Tamene

Concept Note

1. Project Overview

Our project focuses on developing an advanced system for detecting counterfeit drugs through image recognition and chemical composition analysis using artificial intelligence. This project is particularly relevant to **SDG 3: Good Health and Well-being**. Counterfeit drugs pose a significant threat to public health, leading to ineffective treatment, increased morbidity, and even fatalities. By improving the detection of counterfeit medications, it contributes to ensuring Access to Safe Medicines and strengthening Health Systems.

Counterfeit drugs pose a serious global health threat, leading to ineffective treatments and increased mortality due to incorrect dosages or harmful ingredients [1]. This project aims to develop a detection system that combines image recognition and chemical composition analysis to effectively identify counterfeit drugs. We believe the solution we provide will improve patient safety, enhance drug quality assurance, strengthen public health systems.

2. Objectives

- **Specific Objectives of the Project**
- **To Collect Dataset**: Collect and curate a dataset of chemical compositions and images of both legitimate and counterfeit drugs to train the detection system.
- **To Implement Image Recognition Technology**: Create and train a machine learning model to analyze drug packaging and tablet images, identifying visual discrepancies indicative of counterfeits.
- **To Conduct Chemical Composition Analysis**: Utilize analytical techniques to analyze the chemical makeup of drug samples, ensuring accurate profiling against legitimate standards.
- **To Integrate Detection Methods:** Combine image recognition and chemical analysis into a cohesive detection system that provides a comprehensive assessment of drug authenticity.
- **To Validate and Test the System**: Conduct rigorous testing of the system with known counterfeit and legitimate samples to evaluate accuracy and reliability.

Overall, these objectives aim to create a robust, reliable detection system that enhances the safety and efficacy of pharmaceuticals. By accurately identifying counterfeit drugs.

3. Background

Counterfeit drugs pose a significant and growing threat to global public health, as they are manufactured to imitate legitimate products. According to the World Health Organization (WHO), approximately 10% of medicines in low- and middle-income countries are substandard or falsified, which results in treatment failures and increased mortality rates (WHO, 2021).

The health risks associated with counterfeit drugs are alarming. These medications may lack the active ingredients necessary for therapeutic efficacy. For instance, counterfeit antibiotics have been linked to the deaths of thousands of children suffering from pneumonia, as these drugs failed to provide the required treatment (Cohen et al., 2018). Additionally, some counterfeit medications have been found to contain dangerous substances such as mercury, arsenic, or even rat poison, resulting in severe health complications or death (Naylor et al., 2019).

The economic ramifications of counterfeit drugs are profound. They lead to increased healthcare costs due to additional treatments and hospitalizations, as well as significant financial losses for legitimate pharmaceutical companies. The U.S. economy alone suffers losses exceeding $200 billion annually due to counterfeit goods, including pharmaceuticals (International Chamber of Commerce, 2020).

Advancements in technology, particularly in artificial intelligence and machine learning, offer promising solutions for detecting counterfeit drugs. By employing image recognition and chemical composition analysis, it is possible to develop systems that can quickly and accurately identify counterfeit products, thereby enhancing drug safety and efficacy (Chong et al., 2021).

To combat counterfeit drugs, several solutions have emerged. **Advanced packaging technologies**, such as blockchain, enhance traceability by ensuring secure records of drug distribution. **Chemical and biological taggants** serve as invisible markers embedded in drugs, helping to verify authenticity through specialized equipment. **Serialization** involves assigning unique identifiers to each package, facilitating tracking. Additionally, AI and machine learning solutions, such as those developed by companies like Cypheme, utilize image analysis to detect counterfeits, making detection more accessible and efficient.

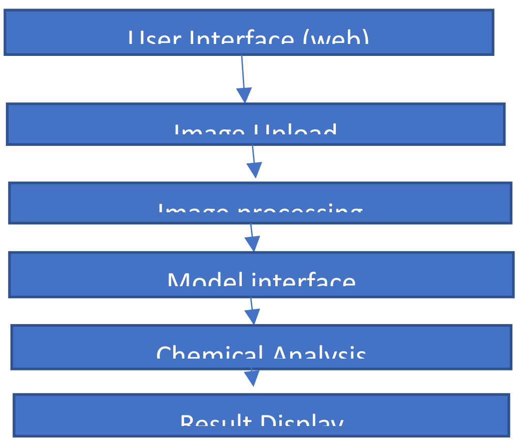
A machine learning approach offers significant advantages in this fight. It provides speed and efficiency by quickly analyzing large datasets and identifying counterfeits in seconds. The high accuracy of machine learning models allows them to recognize subtle differences that may elude human inspectors. Furthermore, these systems are scalable, easily updating with new data to combat evolving counterfeiting techniques. Finally, machine learning is cost-effective.

4. Methodology

To detect counterfeit drugs, we will utilize two primary methodologies: image recognition and chemical composition analysis. Each of these approaches leverages specific machine learning techniques to enhance accuracy and reliability.

- Image Recognition
  - ✓ **Algorithms and Model**: Convolutional Neural Networks (CNNs)
  - ✓ **Frameworks:** keras
- Chemical Composition Analysis
  - ✓ **Algorithm/Model**: Support Vector Machines (SVM)
  - ✓ **Frameworks:** Scikit-learn

5. Architecture Design Diagram



Component Descriptions

1. **User Interface (Web/Mobile App)**:
    a. **Role**: The front-end of the application where users interact with the system. It allows users to upload images of drug packaging for analysis.
    b. **Functionality**: Provides a user-friendly interface to facilitate image uploads and display results.
2. **Image Upload**:
    a. **Role**: Handles the intake of images from users.
    b. **Functionality**: Accepts images and sends them to the image processing component for further analysis.
3. **Image Processing (Preprocessing)**:
    a. **Role**: Prepares the uploaded images for analysis.
    b. **Functionality**: Applies normalization, resizing, and data augmentation techniques to enhance image quality and variability, making it suitable for model training.
4. **Feature Extraction**:
    a. **Role**: Extracts relevant features from the processed images.
    b. **Functionality**: Utilizes a Convolutional Neural Network (CNN), specifically VGG16, to identify and extract distinguishing features from drug packaging images.
5. **Model Inference (CNN Prediction)**:
    a. **Role**: Classifies the processed images based on the extracted features.
    b. **Functionality**: Uses the trained CNN model to predict whether the uploaded images are genuine or counterfeit.
6. **Chemical Analysis** (Optional):
    a. **Role**: Analyzes the chemical composition of drug samples.
    b. **Functionality**: If applicable, this component uses algorithms like Support Vector Machines (SVM) to assess the chemical composition data, providing an additional layer of verification.
7. **Result Display**:
    a. **Role**: Presents the classification results to the user.
    b. **Functionality**: Displays whether the drug is classified as genuine or counterfeit, along with any relevant information from the chemical analysis if performed.

6. Data Sources

For image detection, we will be using data sets from Kaggle and roboflow**.** The Roboflow Universe Image of Medicine Packaging dataset consists of 2,075 images with various annotation formats, focusing on medicine packaging and is relevant for training models to classify different types of packaging. In contrast, the Kaggle Mobile-Captured Pharmaceutical Medication Packages dataset includes 3,900 JPEG images of medication packages captured in real-world conditions using mobile phones, making it useful for training robust models that can handle variations in image quality. And for chemical composition analysis, we use Drug.com which is online database containing over 24,000 entries of prescription and OTC medicines, providing detailed information for both consumers and professionals. It offers readily accessible data on drug ingredients and properties.

7. Literature Review

The study "Image-Based Detection of Counterfeit Pharmaceuticals" by R. Kumar (2020) found that image processing techniques effectively detect counterfeit drugs by analyzing visual features of packaging and tablets, utilizing computer vision algorithms for classification. It highlighted the potential of image recognition technology in enhancing drug safety, paving the way for smartphone applications in counterfeit detection. Similarly, M. De Veij's study "Fast Detection and Identification of Counterfeit Antimalarial Tablets by Raman Spectroscopy" (2007) demonstrated that Raman spectroscopy can reliably identify counterfeit antimalarial tablets by analyzing their chemical composition, establishing it as a rapid detection method suitable for low-resource settings.

Implementation Plan

1. Technology Stack

Programming Languages: Python

Libraries

- **OpenCV**: For image processing and feature extraction.
- **NumPy**: For numerical operations and handling multi-dimensional data.
- **Pandas**: For data manipulation and analysis of datasets.
- **Matplotlib**: For data visualization and plotting results.
- **SciPy**: For scientific computing and advanced mathematical functions.

Frameworks

- **Keras**: High-level API for TensorFlow, simplifying the creation and training of neural networks.
- **Scikit-learn**: For implementing traditional machine learning algorithms, such as SVMs for chemical composition analysis.
- **Roboflow**: For managing, preprocessing, and augmenting image datasets.

Software Components

- **Jupyter Notebook**: For interactive development and analysis.

Hardware Components

- Cameras and Imaging Systems
- Spectroscopy

2. Timeline

Stage 1: Data Collection and Preprocessing (Weeks 1-3)

- **Data Collection: Gather datasets**
- **Data Annotation**
- **Data Preprocessing**

Stage 2: Model Development (Weeks 4-6)

- **Image Recognition Model**:
  - Select appropriate architecture (e.g., CNN).
  - Implement data augmentation techniques to improve model robustness.
- **Chemical Composition Analysis Model**:
  - Choose suitable machine learning algorithms (e.g., SVM, Random Forest) for analyzing spectroscopic data.
- **Integration**:
  - Develop a combined model that incorporates both image recognition and chemical analysis for comprehensive detection.

Stage 3: Training and Evaluation (Weeks 7-9)

- **Model Training**:
  - Train the image recognition model on the annotated dataset.
  - Train the chemical analysis model on the processed spectroscopic data.
- **Evaluation**:
  - Validate models using a separate test dataset.
  - Assess performance metrics (accuracy, precision, recall, F1 score) for both models.

Stage 4: Deployment (Weeks 10-12)

- **System Integration**:
  - Combine the trained models into a single application.
- **User Interface Development**:

        - Create a user-friendly interface for end users to upload images and spectra for analysis.
    - **Testing**:
        - Conduct end-to-end testing to ensure the system performs well in real-world conditions.
    - **Deployment**:
        - Deploy the application on a local server or cloud platform.
        - Monitor performance and gather user feedback for future improvements.

Task Distribution

- **Team Member 1**: Focuses on data sourcing, image model development, and deployment.
- **Team Member 2**: Concentrates on spectroscopic data, chemical model development, and user interface.
- **Team Member 3**: Handles organization, integration tasks, and overall quality checks.

3. Milestones

- **Completion of Data Collection, Data Annotation, and Preprocessing (**End of Week 3): All images and spectroscopic data are gathered and organized and labeled for accuracy.
- **Image Recognition Model Developed:** The architecture for the image recognition model is selected and implemented.

4. Challenges and Mitigations

Challenges

- Incomplete or inaccurate data can lead to poor model performance.
- Models may underperform due to overfitting or underfitting.
- Limited computational resources can hinder model training and evaluation.
- Limited access to high-quality spectroscopic equipment.

Mitigation Strategies

- Collect data from multiple sources to ensure variability and comprehensiveness.
- Experiment with various hyperparameters to optimize model performance.
- Ensure that different components of the system (e.g., data processing, model training) are designed to be modular for easier integration.
- Continuously monitor system performance and optimize code to address bottlenecks.
- Collaboration with Research Institutions

5. Ethical Considerations

- **Data Privacy**: Protect sensitive data through anonymization, informed consent, and robust security measures.

- **Bias in Machine Learning**: Use diverse datasets to train models, conduct bias audits, and maintain transparency about limitations.
- **Impact on Community**: Engage with community members, provide education on the technology, and ensure equitable access to its benefits.

6. References

- Chong, J., et al. (2021). Utilizing Machine Learning for Detection of Counterfeit Pharmaceuticals. *Artificial Intelligence in Medicine*.
- Cohen, J., et al. (2018). The Impact of Counterfeit Antibiotics on Child Mortality. *Pediatrics*.
- International Chamber of Commerce. (2020). The Economic Impact of Counterfeit Goods. *ICC Report*.
- Naylor, D., et al. (2019). The Dangers of Counterfeit Medications: A Global Perspective. *Global Health Action*.
- World Health Organization. (2019). Antimicrobial Resistance: Global Report on Surveillance. *WHO*.
- World Health Organization. (2021). Substandard and Falsified Medicines. *WHO Fact Sheet*.