Development of AI Algorithms to Detect Explosives in Waste Receptables Using Vision Transformers

Mohammad Monirul Hoque, Malinka Kandane Arachchige Dona,
Chanira Katugampala, Navneet Kaur,

Abstract—**The threat posed by improvised explosive devices (IEDs) hidden in public waste receptacles has led to the widespread removal of these bins, causing a host of environmental and health issues. Current solutions, such as blast-resistant or transparent bins, have significant drawbacks, including high costs and aesthetic concerns. This study explores the use of artificial intelligence, specifically Vision Transformers (ViT), to detect explosives within waste receptacles, offering a novel approach to mitigate these risks. Using the augmented TrashNet dataset with additional explosive images, our model was trained using a Vision Transformer architecture. The results are promising, with our model achieving a training accuracy of 87.28%, a validation accuracy of 94.95%, and a testing accuracy of 93.03% after just 10 epochs on a Colab T4 GPU, with a total training time of 64 minutes. This research demonstrates that Vision Transformers can be an effective and efficient tool for enhancing public safety by detecting hidden explosives in waste receptacles.**

Index Terms—**IED, Explosive, Trash, ViTIED, Explosive, Trash, ViT**

I. Introduction

THE detection of concealed explosives remains a critical area of focus for defense and security forces worldwide, given the ever-present threat posed by improvised explosive devices (IEDs) and homemade explosives (HMEs). Traditional methods of explosive detection, such as the use of canines with their enhanced olfactory capabilities, high mobility, and efficient standoff sampling, have been complemented by advanced technological approaches [1].

Simultaneously, the chemical analysis of explosive-related materials has gained importance, especially in understanding the volatile organic compounds (VOCs) associated with these explosives. Studies have focused on profiling explosive odors to enhance detection technologies. These profiles help in identifying key VOCs that are critical for the rapid field detection of explosives, thus aiding in the development of advanced chemical and biological sensors [2].

One notable study demonstrated the use of convolutional neural networks (CNNs) for detecting explosive hazards in conflict zones, achieving promising results even with limited training data by leveraging dimensionality reduction through multiple projected images.[3]. In another pioneering effort, researchers applied CNNs to the detection of conventional explosives in security X-ray scans of passenger baggage, introducing the Passenger Baggage Object Database (PBOD). This study achieved high reliability with an area under the curve (AUC) of 0.95 and utilized heatmaps to provide valuable visualizations of threat locations.[4]. Additionally, The development of portable inspection systems using electronic noses with MOX chemical sensor arrays has also been explored. This research evaluated various machine learning models on a dataset of 140 samples, including combinations of TNT or gunpowder with everyday items like soap and toothpaste. [5]. Furthermore, Addressing the critical security concern of detecting IEDs in public places, another study developed a deep learning model to classify X-ray images of baggage and objects. [6]. Moreover, Machine learning methods have also been applied to improve the selectivity of ion mobility spectrometry (IMS) for detecting nitrate-based explosives, a common challenge in security settings like airports and border crossings. By examining five classes of nitrate explosives and environmental nitrates, the study demonstrated that machine learning significantly enhances IMS selectivity.[7]. On the contrary, The limitations of transmission X-ray systems in airport security, particularly the lack of material-specific information leading to false alarms and operational delays, have been addressed by exploring X-ray Diffraction Tomography (XRDT) with a coded aperture. Training a 1D Convolutional Neural Network (CNN) on simulated data, the study achieved relative improvements in classification accuracy compared to previous correlation-based methods. However, the reliance on simulated data and the need for a comprehensive material library are notable limitations [8]. The challenge of enhancing public safety through the detection of explosive and related audio events has been explored by introducing a novel multimodal learning approach for automated detection of explosive sounds. This research significantly outperformed existing methods in classifying explosive and non-explosive sounds, although challenges remain in accurately detecting explosive sounds due to variations in audio environments [9].

The detection of IEDs in rural and urban environments poses significant challenges due to varied materials and concealment. A study leveraged a CNN to improve autonomous detection using a sensor array in hazardous terrains, achieving an accuracy of 98.7% in well-lit conditions. While the study demonstrated high accuracy and real-time detection capabilities, the model's dependency on well-lit conditions and the need for further testing in diverse environments are potential limitations [10].

The most recent work on IED detection in public dustbins using AI-based classifiers highlights the potential of such technologies to enhance public safety. By augmenting the TrashNet dataset with new explosive images and leveraging transfer learning with state-of-the-art CNNs, the study achieved a notable Top-1 accuracy of 80% with DenseNet-121. This approach underscores the importance of comprehensive datasets and the effectiveness of transfer learning in specialized detection tasks.[11].

In this research work, we have introduced ViT to Detect images, and shown how it performs. Previously, the author used DenseNet-121, MobileNet-V2, NASNetMobile, EfficientNetV2B0, EfficientNetV2L,ConvNeXtSmall, and ConvNeXtLarge.

II. Related Studies

The critical issue of detecting explosive hazards in conflict zones using advanced technologies like CNN architectures, demonstrating promising results even with limited training data, is addressed by the study [3]. By leveraging dimensionality reduction through multiple projected images, the research optimizes computational efficiency. Despite the preliminary nature of the results and challenges with data imbalance, the study shows that deep learning can potentially outperform skilled experts in identifying side attack explosive hazards. This highlights the potential of AI to enhance safety in civilian and military contexts, suggesting further validation and exploration of other explosive types as future research directions.

The author [4] applies CNNs for detecting conventional explosives in security X-ray scans of passenger baggage, marking a pioneering effort in this domain. The researchers introduce the Passenger Baggage Object Database (PBOD) to aid the development of new threat detection algorithms. Utilizing advanced CNN models, the study achieves a high reliability with an AUC of 0.95. Additionally, the use of heatmaps provides valuable visualizations of threat locations. The strengths of this work include its introduction of a novel dataset and high detection accuracy. However, it may face challenges in generalizing across different X-ray machines and real-world variability.

The development of a fast, reliable, and portable inspection system for detecting hidden explosives using electronic noses with a MOX chemical sensor array is the focus of the paper [5]. The research evaluates five machine learning models on 140 samples, including combinations of TNT or gunpowder with everyday items like soap and toothpaste. The LSTM-based deep learning model derived from LeNet-5 achieves 100% classification accuracy using only 30 seconds of sensor data. The strengths of this work include its high accuracy, fast detection, and potential for easy implementation into embedded systems. However, the small sample size and limited variety of tested substances might limit the generalizability of the results.

The critical security concern of detecting IEDs in public places is addressed in this research [6] through the development of a deep learning model to classify X-ray images of baggage and objects. The study uses expert-generated sample images and various data augmentation techniques to mitigate the issue of limited training data. The proposed model achieves impressive accuracy, with the best rate being 0.985, surpassing related works. Additionally, the research highlights that a large training set might lead to overfitting and resource inefficiency. Strengths of this study include its high accuracy and effective use of data augmentation. However, reliance on expert-generated images may limit real-world applicability, and the findings on training set size suggest potential overfitting issues.

This study [7] explores improving the selectivity of ion mobility spectrometry (IMS) for detecting nitrate-based explosives using machine learning methods. IMS is widely used in security settings like airports and border crossings, but it struggles with differentiating between nitrate-based threats and ambient nitrates. The research examines five classes, including various nitrate explosives and environmental nitrates, using a small database. Preliminary results indicate that machine learning significantly enhances IMS selectivity. The strengths of this work include addressing a critical limitation of IMS and demonstrating the potential of machine learning to improve detection accuracy. However, the study's preliminary nature and small dataset limit the generalizability of its findings.

The author [8] addresses the limitations of transmission X-ray systems in airport security, particularly the lack of material-specific information which leads to false alarms and operational delays. It explores using X-ray Diffraction Tomography (XRDT) with a coded aperture to provide complementary chemical and molecular signatures for better material identification. The research highlights the challenges posed by noisy signals, variability in XRD form factors, and the absence of a comprehensive material library. By training a 1D Convolutional Neural Network (CNN) on simulated data, the study achieves relative improvements in classification accuracy compared to previous correlation-based methods. These improvements are validated with both simulated and real experimental data. Strengths include addressing a critical operational issue and demonstrating the potential of CNNs to enhance classification accuracy. However, reliance on simulated data and the need for a comprehensive material library are notable limitations.

The challenge of enhancing public safety through the detection of explosive and related audio events, thereby providing rapid aid and minimizing devastation, is addressed by the research [9]. It introduces a novel multimodal learning approach for automated detection of explosive sounds, such as gunfire and explosions, for audio surveillance. The proposed deep feature stacking method significantly outperforms existing methods in classifying explosive and non-explosive sounds. Strengths of this research include its innovative approach and high classification performance. However, challenges remain in accurately detecting explosive sounds due to variations in audio environments and the complexity of distinguishing between similar sounds.

The urgent need for automated demining methods in Ukraine due to the ongoing conflict is the focus of this paper [12]. It reviews various modern demining techniques, including the use of metal detectors and ground-penetrating radar (GPR), analyzing their advantages and disadvantages. The research develops an information system using a convolutional neural network and an autoencoder for the automated classification of explosive devices, achieving an accuracy of 97.83%. Strengths include its high accuracy and relevance to current demining efforts. However, the study may be limited by the specific conditions and types of explosive devices encountered in Ukraine, potentially affecting generalizability to other regions. The author [10] addresses the detection of IEDs in rural and urban environments, which pose significant challenges due to their varied materials and concealment. The paper leverages a CNN to improve autonomous detection using an autonomous sensor array in hazardous terrains. The proposed CNN, trained to distinguish IEDs from natural surroundings in real-time, achieved an accuracy of 98.7% in well-lit conditions. Strengths of the study include its high accuracy and real-time detection capabilities. However, the model's dependency on well-lit conditions and the need for further testing in diverse environments are potential limitations. Insights for your literature review could focus on the innovative use of CNNs for real-time detection and the challenges of deploying such systems in varied lighting and environmental conditions. The author [13] explores the application of deep CNNs with transfer learning for image classification and detection in X-ray baggage security imagery. The use of transfer learning addresses the challenge of limited data availability by optimizing pre-trained CNNs for specific security tasks. The study compares CNN features with traditional hand-crafted features, showing that fine-tuned CNN features significantly outperform conventional methods. Achieving an accuracy of 0.994 using AlexNet features with an SVM classifier, the research also evaluates various detection paradigms, such as YOLOv2, Faster-RCNN, and R-FCN. YOLOv2 demonstrates superior performance, achieving a mean average precision (mAP) of 0.885 for six-class detection and 0.974 mAP for two-class firearm detection, with quick processing times. Strengths include the high accuracy and efficiency of CNN-based methods in handling cluttered imagery. However, reliance on pre-trained models and the potential variability in detection performance across different object classes are limitations. Insights for your literature review could focus on the effectiveness of transfer learning in overcoming data scarcity and the comparative evaluation of CNN-based detection techniques in security imaging.

The most recent work on IED detection has been done in [11], where the author addresses the pressing issue of detecting IEDs in public dustbins using AI-based classifiers, enhancing public safety. By augmenting the TrashNet dataset with new explosive images [14] and leveraging transfer learning with state-of-the-art CNNs, it achieves a notable Top-1 accuracy of 80% with DenseNet-121. This approach highlights the potential of AI in public safety, the importance of comprehensive datasets, and the effectiveness of transfer learning in specialized detection tasks

III. Materials and Methods

A. Data used for model training

There is a notable scarcity of datasets specifically designed for detecting improvised explosive devices (IEDs) in waste receptacles. Currently, the primary dataset available for waste categorization is TrashNet [15]. To address the gap in IED detection, the authors [11] augmented the TrashNet dataset with images of IEDs, resulting in a new, more comprehensive dataset [14].

For the purposes of this research, we utilized this augmented dataset [14], which consists of two main categories: IED images and non-IED images. The non-IED images are further divided into six subcategories based on different types of waste: cardboard, glass, metal, paper, plastic, and general trash. Specifically, the dataset comprises a total of 3041 images, with 514 images of IEDs and 2527 images of non-IEDs. The non-IED images are distributed as follows: 403 images of cardboard, 501 images of glass, 410 images of metal, 594 images of paper, 482 images of plastic, and 137 images of general trash.

In this research, we combined all non-IED categories into a single class, effectively converting the problem into a binary classification task. This approach simplifies the model's objective to distinguishing between IEDs and non-IEDs.

The original images in both the TrashNet and IED datasets varied in size and were in color. To standardize the input for our models, we resized all images to 224x224 pixels, ensuring compatibility with the Vision Transformer (ViT) architecture [16] and the base models selected for deep transfer learning. This resizing aligns closely with the generally recommended dimensions of 256x256 pixels for deep learning models [17]. The transformation process is illustrated in Fig. 1.

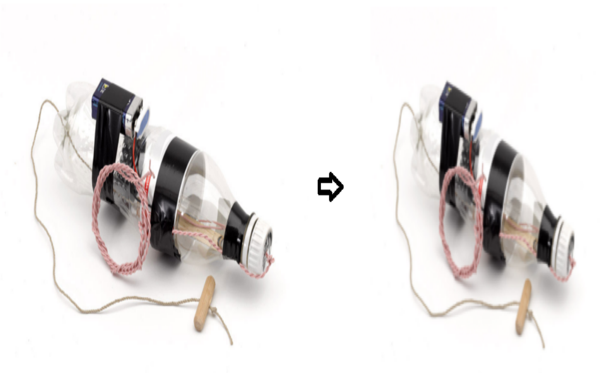

Fig. 1: Random Image Transformation to 224x224

To facilitate proper model training, evaluation, and testing, we divided the dataset into three subsets: 70% for training, 15% for validation, and 15% for testing. This stratification ensures that the model is exposed to a broad range of data during training, while still reserving sufficient data for unbiased validation and testing phases. By employing this structured approach, we aim to develop a robust model capable of accurately distinguishing between IEDs and non-IEDs in various waste receptacle contexts.

B. Method

For this research work, we have used Vision Transformer to classify IED and Non-IED images. Here we will discuss the concept of the algorithm that were implemented.

1) Vision Transformer: The concept of the ”Transformer” was initially introduced within the realm of natural language processing (NLP) [18]. Building on the success of transformers in NLP, Dosovitskiy et al. [16] adapted this technique for use in computer vision. This adaptation led to the development of the Vision Transformer (ViT), which processes input images by dividing them into patches that act as tokens. These tokens are then fed into a transformer encoder to determine their correlations. Following the introduction of ViT [19], [20], [21], several other models were developed to further explore this approach.

The integration of self-attention mechanisms, particularly transformers, from NLP into computer vision aims to enhance computational efficiency and address scalability challenges. In this context, a ViT replaces traditional convolutional layers with a transformer encoder that utilizes self-attention. This approach shifts away from the conventional convolutional methodology, offering a novel way to handle image data.

However, Vision Transformers typically require significantly more data compared to convolutional neural networks (CNNs) to achieve optimal performance. This data-intensive nature means that ViTs are often trained on large public datasets containing millions of labeled images. Consequently, using pre-trained ViT models is generally recommended when applying this technology to various datasets, as these pre-trained models have already learned from extensive datasets and can be fine-tuned for specific tasks with smaller amounts of data. This strategy helps to mitigate the issue of data scarcity and leverages the powerful capabilities of ViTs in new applications. The architecture of ViT is shown in:

ViT’s approach to image recognition starts with a novel way of processing the image itself. Unlike Convolutional Neural Networks (CNNs) that work by sliding filters across the image, ViT breaks the image down into smaller, manageable pieces called patches. These patches are essentially sub-regions of the image, like tiny squares focusing on specific areas.

But there’s a catch: simply flattening these patches into vectors loses valuable information about their original location within the image. To address this, ViT employs a technique called positional encoding. Imagine adding a special label to each patch that encodes its relative position in the larger image. This encoding could be a simple coordinate system or a more complex scheme that captures higher-order relationships between patches. By incorporating this positional information, ViT ensures that the model doesn’t treat all patches the same, even though they are flattened into vectors.

This combination of patch embeddings and positional encoding allows ViT to represent an image as a collection of informative vectors, each containing not only the visual content of a local area but also its relative position within the whole scene. This paves the way for the next stage of ViT’s processing pipeline - the Transformer encoder.

The Transformer encoder is where the magic of ViT happens. This is an architectural building block typically used for processing sequential data like text. In the context of ViT, however, the Transformer encoder takes the stage to analyze the relationships between the encoded patch vectors.

Think of the Transformer as a powerful tool that can identify how different patches relate to each other. It can learn that a patch containing a red object might be next to a patch with a blue object, or that a patch with a house might be positioned above a patch with a car. By analyzing these relationships across all patches, the Transformer encoder builds a comprehensive understanding of the global structure of the image.

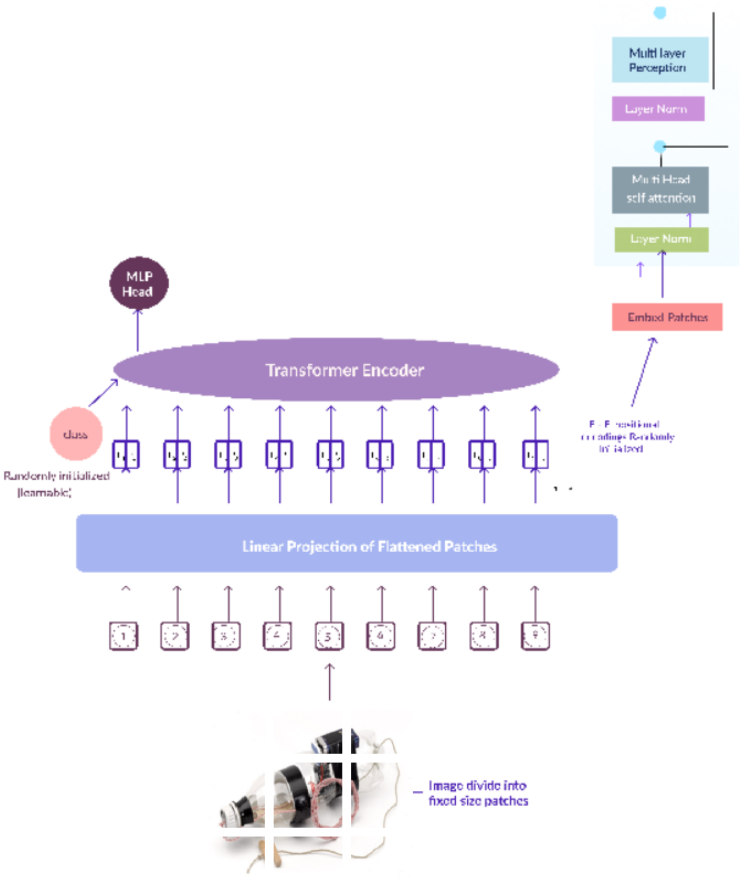
Multi layer Perception
Layer Norm
Multi Head self attention
Layer Norm
Embed Patches
MLP Head
Transformer Encoder
class
Randomly initialized (learnable)
Linear Projection of Flattened Patches
Image divide into fixed size patches

Fig. 2: ViT Architecture for IED Detection

This ability to capture long-range dependencies between image regions sets ViT apart from traditional CNNs. CNNs primarily rely on local filters, limiting their ability to grasp the overall composition of an image. ViT, with its Transformer encoder, transcends this limitation, allowing it to excel at tasks that require understanding the bigger picture within an image.

ViT’s true potential is unleashed through a two-stage process: pre-training and fine-tuning. Pre-training involves feeding ViT massive datasets of labeled images. During this phase, the model learns powerful image recognition capabilities by identifying patterns and relationships within the vast amount of data. Imagine ViT training on millions of images, progressively developing its ability to recognize objects, scenes, and their interactions within these images.

Once pre-trained, ViT is fine-tuned for a specific task. This involves taking the pre-trained model and adapting it to a smaller dataset relevant to the desired application. For instance, if you want ViT to excel at identifying dog breeds, you would fine-tune it on a dataset specifically containing images of dogs labeled with their breeds. This fine-tuning process leverages the pre-trained knowledge while specializing the model for the particular task at hand.

By combining pre-training with task-specific fine-tuning, ViT achieves remarkable performance in various computer vision tasks. It offers a compelling alternative to traditional CNNs, particularly for tasks that demand a strong grasp of the overall image structure and long-range dependencies between different image regions.

2) Approach: The process of predicting improvised explosive device (IED) images using Vision Transformers (ViT) begins with preparing the dataset. Initially, the dataset is divided into three subsets: 70% for training (T_r), 15% for validation (V_a), and 15% for testing (T_e). This ensures that the model is trained, validated, and tested on distinct data to prevent overfitting and to evaluate its performance accurately.

Next, a Vision Transformer model (V_a) is constructed. ViT, which originated from natural language processing, has been adapted for image classification tasks by splitting images into patches and using a transformer encoder to process these patches. This model leverages pre-trained weights from existing ViT models to benefit from previously learned features, making the training process more efficient.

The training phase involves fine-tuning the pre-trained ViT model on the (T_r) dataset. This step utilizes transfer learning, allowing the model to adapt its parameters to the specific task of IED detection. During training, the validation set (V_a) is used to monitor the model's performance and to adjust hyperparameters such as learning rate, batch size, and the number of epochs. Hyperparameter tuning is crucial to optimizing the model's accuracy and preventing overfitting.

Once the model is trained and fine-tuned, it is used to predict the presence of IEDs in the test set (T_e). The trained model processes the test images and generates predictions, identifying whether each image contains an IED.

Below we have shown the architecture we have followed:

Algorithm

```
Input: Image Dataset;
Output: Prediction of IED images $P$;
Fetch the dataset of train($T_r$), test($T_e$) and validation ($V_a$) set.
Build a model for images as V
Train model V with $T_r$ using pre-trained ViT algorithm.
Tune the hyperparameter of model V using $V_a$.
Get P by predicting $T_e$ based on model V
Evaluate the model
```

Below in Fig. 3 is the Framwork of our approach:

Finally, the model's performance is evaluated using the test set. This evaluation involves calculating metrics such as accuracy, precision, recall, and F1 score to determine the model's effectiveness in correctly identifying IEDs. The evaluation provides insights into the strengths and weaknesses of the model, indicating areas for potential improvement.

C. Experimental Setup

We conducted an experiment to detect improvised explosive devices (IEDs) in waste receptacles using the Vision Transformer (ViT) model, specifically the ViT b-32 version. Our task was to classify images into two categories: IED and non-IED. The non-IED category included six types of waste: cardboard, glass, metal, paper, plastic, and general trash.

Hardware Configuration: Our experiments were run on the following hardware:

- GPU: NVIDIA Tesla T4
- Memory: 16 GB RAM
- Processor: Intel Xeon CPU
- Environment: Google Colab

Data Preprocessing: The dataset included images from the TrashNet dataset and an augmented set of IED images. These images varied in size and were in color. We resized all images to 224x224 pixels to match the requirements of the ViT model and to standardize them for processing. Then We divided the dataset into three parts: 70% for training, 15% for validation, and 15% for testing. This split ensures that we have enough data to train the model, validate its performance during training, and test its accuracy on unseen images. Fig. 4 shows all categories of Non-IED wastes.

Fig. 5 shows our target class IED image.

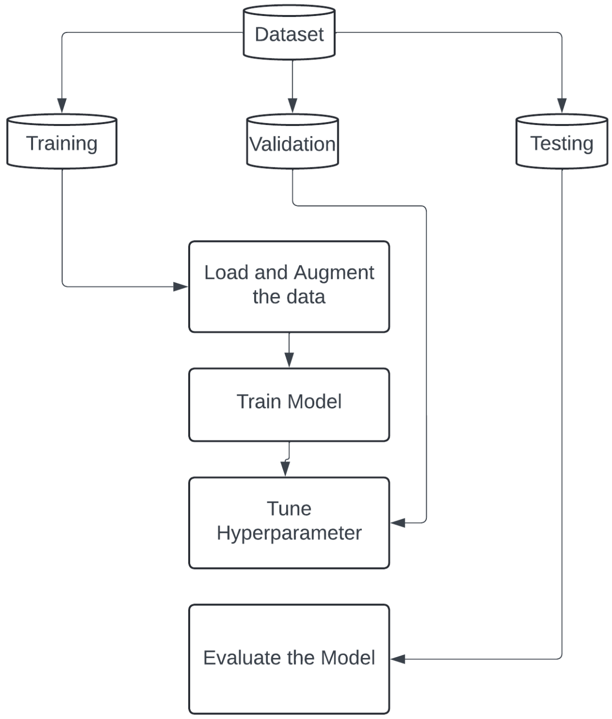

Fig. 3: Framework of our approach

D. Model Training and Experiments

To train the Vision Transformer (ViT) b-32 model for the task of detecting improvised explosive devices (IEDs) in waste receptacles, we followed a structured training process. We utilized transfer learning, leveraging a pre-trained ViT model to benefit from the knowledge gained from a large dataset. This approach allows the model to adapt to our specific task with less training data.

The training data consisted of 70% of our prepared dataset, which included both IED and non-IED images. The non-IED images comprised six categories: cardboard, glass, metal, paper, plastic, and general trash. These categories were merged into a single non-IED class for binary classification.

The model was trained for 10 epochs. During training, we monitored the model's performance using the validation set, which constituted 15% of the dataset. This monitoring involved adjusting hyperparameters, such as learning rate (set at 0.001) and batch size, to optimize the model's accuracy and prevent overfitting. We used the Adam optimizer for training, which is known for its efficiency and effectiveness in handling large datasets.

In the output layer, we employed the softmax function, typically used for multiclass classification but also applicable for binary classification. For the loss function, we used cross-entropy loss, which is well-suited for this type of classification task. Fig. 6 shows the hyperparameter tuning.

Fig. 5: IED image

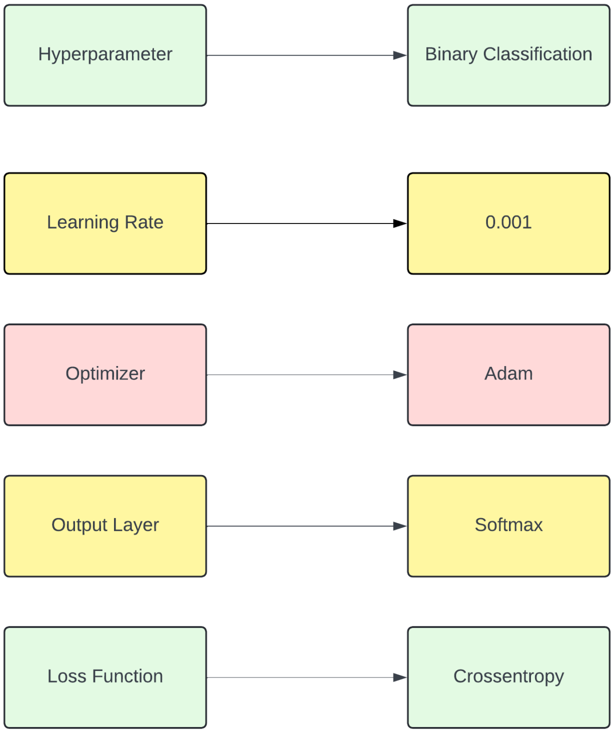

Fig. 6: Hyperparameters

Hyperparameter: Binary Classification
Hyperparameters are critical elements in machine learning that must be set before training the model. Unlike parameters that are learned from the data during training, hyperparameters are predetermined and significantly impact the model's performance. In this context, the hyperparameter is set for binary classification, which is essential because it defines the model's task of distinguishing between two possible classes: 'IED' (improvised explosive device) and 'Non-IED'. Setting the correct hyperparameter for binary classification ensures that the model is trained with algorithms and metrics suited for binary outcomes, leading to more accurate and reliable predictions. This setup is particularly important for applications where the accurate identification of a threat can have significant safety implications.

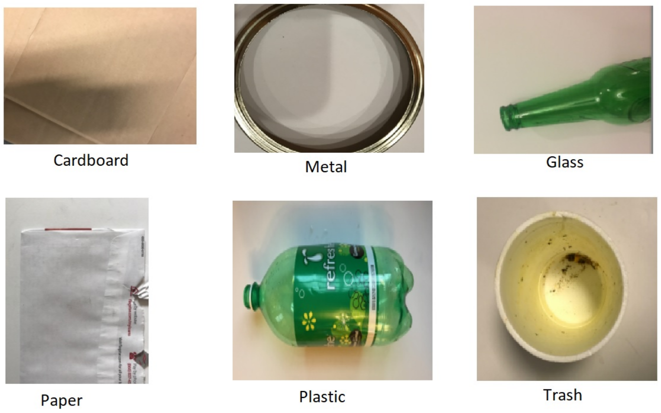

Fig. 4: All Categories of Non-IED Images

Learning Rate: 0.001
The learning rate is a pivotal hyperparameter in machine learning, controlling how much the model's weights are adjusted during training. It dictates the step size at each iteration as the model moves towards minimizing the loss function. In our setup, the learning rate is set to 0.001, a commonly used value that balances convergence speed and stability. A well-chosen learning rate can significantly affect training efficiency and effectiveness. If set too high, the

model might converge too quickly to a suboptimal solution or diverge entirely. Conversely, a very low learning rate can result in a slow training process, potentially getting stuck in a local minimum. Thus, setting the learning rate to 0.001 helps ensure that the model learns efficiently while maintaining stability throughout the training process.

Optimizer: Adam

An optimizer is an algorithm that adjusts the weights of the neural network to minimize the loss function, essentially guiding the model's learning process. The Adam optimizer, which stands for Adaptive Moment Estimation, is used in this context. Adam combines the benefits of two other extensions of stochastic gradient descent: AdaGrad and RMSProp, computing individual adaptive learning rates for different parameters. This optimizer is particularly favored for its efficiency and low memory requirements, making it suitable for large datasets and high-dimensional parameter spaces. By using the Adam optimizer, the model benefits from adaptive learning rates, which enhance both convergence speed and stability, thereby improving the overall training process and model performance.

Output Layer: Softmax

The output layer of a neural network is crucial as it produces the final predictions. The choice of activation function in this layer is essential for the type of problem being solved. In this case, we use the softmax function in the output layer. Although typically employed for multiclass classification problems, softmax can also be used for binary classification. It converts the raw output values (logits) into probabilities that sum up to one, facilitating easier interpretation of the outputs. By using the softmax function, the model can provide probabilities for each class, which is particularly useful for understanding the confidence of the predictions. This ensures that the output values are normalized and interpretable as probabilities, aiding in making more informed decisions based on the model's predictions.

Loss Function: Crossentropy

The loss function is a measure that quantifies how well the model's predictions match the actual labels in the training data. It plays a critical role in guiding the training process by providing feedback on the prediction accuracy. In this model, we use the crossentropy loss function, also known as log loss. Crossentropy loss measures the performance of a classification model whose output is a probability value between 0 and 1, making it particularly effective for both binary and multiclass classification tasks. This loss function is beneficial because it penalizes the model more for being confident and wrong, encouraging it to output probability distributions that closely match the true distributions of the data. By using crossentropy, the model is driven to improve its accuracy, ultimately leading to better performance in classification tasks.

IV. Results and Discussions

In this section, we present and analyze the results achieved with our proposed method. We begin by outlining the criteria and methodology used for evaluation. Following this, we discuss the performance of our approach based on various metrics.

1) Evaluation: We chose not to rely on accuracy as a performance metric due to the significant class imbalance in our dataset, which could lead to misleading results. Instead, we assessed the effectiveness of our approach using recall, precision, and F1-Score metrics.

1) Precision: The calculation involves taking the number of true positive predictions and dividing it by the total number of predictions made by the model. The formula of precision is shown in equation 1 below:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

2) Recall: Recall is determined by dividing the number of true positives by the total number of actual instances in that class. The formula of recall is shown in equation 2 below:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

3) F1: The F1-score is calculated as the harmonic mean of precision and recall. It is shown in equation 3 below:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

4) Accuracy: It is calculated by dividing the number of correctly classified instances by the total number of instances. It is shown in equation 8 below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

The performance of the Vision Transformer (ViT) b-32 model is evaluated using several key metrics: precision, recall, F1 score, and accuracy. Here is a detailed explanation of these metrics and what they indicate about the model's performance:

The precision, recall, F1 and Accuracy is shown in the below table I:

TABLE I: Result of ViT for IED Classification

Precision	Recall	F1	Accuracy
44%	42%	43%	93%

Precision: 44%: Precision measures the proportion of true positive predictions among all positive predictions made by the model. In this case, the precision of 44% indicates that out of all the instances the model identified as IEDs, only 44% were actually IEDs. This relatively low precision suggests that the model has a higher rate of false positives, where non-IEDs are incorrectly classified as IEDs.

Recall: 42%: Recall, also known as sensitivity, measures the proportion of true positive predictions among all actual positives. A recall of 42% means that out of all the actual IEDs in the dataset, the model correctly identified 42% of them. This low recall indicates that the model misses a significant number of actual IEDs, resulting in a higher rate of false negatives.

F1: 43%: The F1 score is the harmonic mean of precision and recall, providing a balance between the two. An F1 score of 43% reflects a trade-off between precision and recall, indicating that both metrics are relatively low. The F1 score is useful for assessing the overall effectiveness of the model, especially when dealing with imbalanced datasets where precision and recall are crucial.

Accuracy: 93%: Accuracy measures the proportion of all correct predictions (both true positives and true negatives) out of the total number of instances. An accuracy of 93% indicates that the model correctly classified 93% of the instances in the dataset. While this high accuracy suggests that the model performs well overall, it can be misleading in the context of imbalanced datasets. Given the low precision and recall, the high accuracy likely reflects the model's ability to correctly identify non-IEDs, which are more prevalent in the dataset.

In short, The high accuracy of 93% demonstrates that the model is effective in correctly classifying a large portion of the dataset. However, the lower precision and recall values indicate that the model struggles with accurately identifying IEDs, resulting in a higher rate of both false positives and false negatives. The F1 score of 43% highlights the need for further improvement in balancing precision and recall to enhance the model's reliability in detecting IEDs. These metrics collectively provide a comprehensive view of the model's performance, highlighting its strengths and areas for improvement.

During Training time, we have used 10 epoch to build IED classification using ViT.

TABLE II: Results for 10 Epoch

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	45.54%	82.68%	25.83%	93.03%
2	38.86%	85.71%	24.70%	94.23%
3	35.79%	86.89%	22.50%	94.47%
4	34.88%	86.86%	18.37%	94.23%
5	31.82%	87.28%	16.06%	94.95%
6	30.49%	88.05%	25.86%	88.22%
7	28.50%	88.35%	18.58%	94.23%
8	29.24%	88.26%	19.39%	94.95%
9	32.78%	85.07%	17.07%	94.71%
10	26.76%	88.81%	16.76%	93.75%

During the training process of the Vision Transformer (ViT) b-32 model, we meticulously tracked the performance metrics over 10 epochs. In the initial epoch, the model exhibited a training loss of 0.4554 and a training accuracy of 82.68%, while the validation loss was 0.2583 with a validation accuracy of 93.03%. This initial performance suggested that the pre-trained weights were already somewhat effective.

As training progressed to the second epoch, there was a noticeable improvement. The training loss decreased to 0.3886, and the training accuracy rose to 85.71%. Similarly, the validation loss dropped to 0.2470, with an enhanced validation accuracy of 94.23%. By the third epoch, these trends continued, with a further reduction in training loss to 0.3579 and an increase in training accuracy to 86.89%. The validation loss decreased to 0.2250, and the validation accuracy slightly improved to 94.47%.

The fourth epoch saw the training loss stabilize at 0.3488, with a training accuracy of 86.86%. The validation metrics showed a significant drop in loss to 0.1837 while maintaining a high accuracy of 94.23%. By the fifth epoch, the training loss further reduced to 0.3182, and training accuracy improved to 87.28%, with the validation loss reaching its lowest point at 0.1606 and the highest validation accuracy of 94.95%.

However, during the sixth epoch, while the training loss continued to decline to 0.3049, the validation loss unexpectedly increased to 0.2586, and validation accuracy dropped to 88.22%. This fluctuation indicated a potential overfitting scenario where the model was performing exceptionally well on the training data but not as consistently on the validation set.

In subsequent epochs, the training loss continued to decrease, reaching 0.2850 in the seventh epoch with a training accuracy of 88.35%. The validation metrics corrected themselves, with the validation loss dropping to 0.1858 and the accuracy rising back to 94.23%. This trend continued in the eighth epoch with a training loss of 0.2924 and a training accuracy of 88.26%, coupled with a validation loss of 0.1939 and a high validation accuracy of 94.95%.

By the ninth epoch, the training loss had increased slightly to 0.3278, with a drop in training accuracy to 85.07%. Despite this, the validation loss improved to 0.1707, and the validation accuracy remained robust at 94.71%. In the final epoch, the model achieved its lowest training loss of 0.2676 and the highest training accuracy of 88.81%, with a validation loss of 0.1676 and a validation accuracy of 93.75%.

These results indicate that the ViT b-32 model successfully learned and generalized well from the training data, maintaining high accuracy across both training and validation sets while effectively managing the loss metrics.

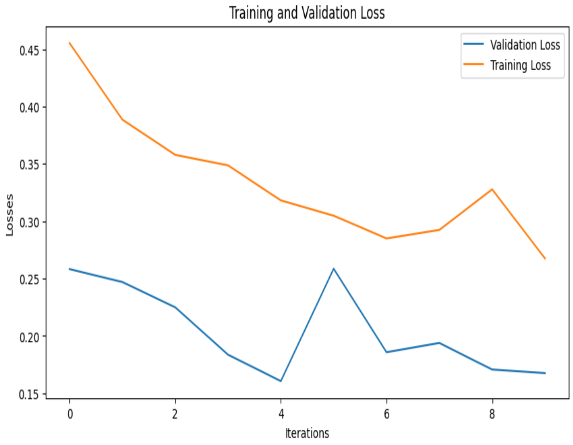

Fig. 7: Training and Validation Loss Curves

In summary, the loss curves of training and validation is shown in Fig. 7 and accuracy curves of training and validation

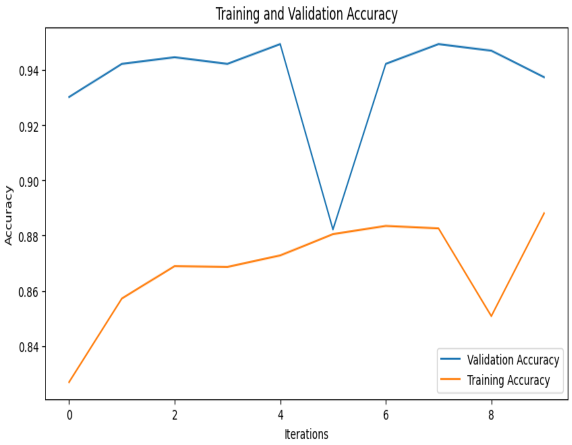

Fig. 8: Training and Validation Accuracy Curves

is shown in Fig. 8.

V. Conclusion

In this study, we successfully implemented the Vision Transformer (ViT) b-32 model to address the critical task of detecting improvised explosive devices (IEDs) in waste receptacles. Through rigorous training and validation over 10 epochs, we achieved a high overall accuracy of 93.03% on the test set, demonstrating the model's strong generalization capabilities. Despite this high accuracy, the precision and recall values were lower, indicating challenges in correctly identifying IEDs, which are critical in real-world scenarios. The performance metrics suggest that while the model is effective at overall classification, further improvements are needed to enhance its reliability in detecting IEDs specifically.

The adoption of the Adam optimizer, with its adaptive learning rate mechanism, played a crucial role in efficiently training the model. The use of the softmax function in the output layer and the crossentropy loss function provided a robust framework for binary classification, even though traditionally these are more associated with multiclass tasks. This adaptability highlights the flexibility of modern machine learning techniques in addressing diverse classification problems.

Future work could focus on expanding the dataset to include more varied and balanced samples, employing advanced data augmentation techniques, and exploring different model architectures or ensemble methods to further improve the detection accuracy of IEDs. Additionally, fine-tuning hyperparameters and exploring alternative optimizers could also contribute to enhanced model performance. Overall, this research underscores the potential of Vision Transformers in enhancing public safety through the accurate detection of explosive threats in waste management systems.

Acknowledgment

The authors would like to thank Dr. Amoakoh Gyasi-Agyei for his continous support throughout the project work.

References

[1] J. Kennedy, A. Sayedelahl, J. O. Castro, M. Circelli, P. Ghasemigoudarzi, D. Green, M. Henschel, Y. Ma, and P. McGuire, "The detection of concealed explosives using the midsix system," *IEEE Transactions on Radar Systems*, 2023.

[2] S. F. Gallegos, E. O. Aviles-Rosa, M. T. DeChant, N. J. Hall, and P. A. Prada-Tiedemann, "Explosive odor signature profiling: A review of recent advances in technical analysis and detection," *Forensic science international*, vol. 347, p. 111652, 2023.

[3] B. Brockner, C. Veal, J. Dowdy, D. T. Anderson, K. Williams, R. Luke, and D. Sheen, "Convolutional neural network based side attack explosive hazard detection in three dimensional voxel radar," in *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIII*, vol. 10628. SPIE, 2018, pp. 494–506.

[4] T. Morris, T. Chien, and E. Goodman, "Convolutional neural networks for automatic threat detection in security x-ray images," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 285–292.

[5] J. Torres-Tello, A. V. Guaman, and S.-B. Ko, "Improving the detection of explosives in a mox chemical sensors array with lstm networks," *IEEE Sensors Journal*, vol. 20, no. 23, pp. 14 302–14 309, 2020.

[6] C. Chamnanphan, S. Vorapatratorn, T. Boongoen, N. Iam-On, and K. Kirimasthong, "Improvised explosive device detection using cnn with x-ray images," *Journal of Advances in Information Technology*, vol. 14, no. 4, pp. 674–684, 2023.

[7] D. Fisher, S. R. Lukow, G. Berezutskiy, I. Gil, T. Levy, and Y. Zeiri, "Machine learning improves trace explosive selectivity: application to nitrate-based explosives," *The Journal of Physical Chemistry A*, vol. 124, no. 46, pp. 9656–9664, 2020.

[8] K. Brumbaugh, C. Royse, C. Gregory, K. Roe, J. Greenberg, and S. Diallo, "Material classification using convolution neural network (cnn) for x-ray based coded aperture diffraction system," in *Anomaly Detection and Imaging with X-rays (ADIX) IV*, vol. 10999. SPIE, 2019, pp. 63–71.

[9] V. Shukla and M. Singour, "Multimodal learning for early detection of explosive sounds using relative spectral distribution," in *2020 Sensor Signal Processing for Defence Conference (SSPD)*. IEEE, 2020, pp. 1–5.

[10] S. Colreavy-Donnelly, F. Caraffini, S. Kuhn, M. Gongora, J. Florez-Lozano, and C. Parra, "Shallow buried improvised explosive device detection via convolutional neural networks," *Integrated Computer-Aided Engineering*, vol. 27, no. 4, pp. 403–416, 2020.

[11] A. Gyasi-Agyei, "Detection of explosives in dustbins using deep transfer learning based multiclass classifiers," *Applied Intelligence*, pp. 1–34, 2024.

[12] L. Mochurad, V. Savchyn, and O. Kravchenko, "Recognition of explosive devices based on the detectors signal using machine learning methods." in *IntelITSIS*, 2023, pp. 249–260.

[13] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery," *IEEE transactions on information forensics and security*, vol. 13, no. 9, pp. 2203–2215, 2018.

[14] A. Gyasi-Agyei. [Online]. Available: https://github.com/jessieAmoakoh/I2Net

[15] T. G. Yang M, "Trashnet trash dataset," 2022, accessed 27 Nov, 2022. [Online]. Available: https://github.com/garythung/trashnet

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] O. Rukundo, "Effects of image size on deep learning," *Electronics*, vol. 12, no. 4, p. 985, 2023.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[21] Z. Zhang, H. Zhang, L. Zhao, T. Chen, and T. Pfister, “Aggregating nested transformers,” *arXiv preprint arXiv:2105.12723*, vol. 2, no. 3, p. 5, 2021.