VIFDD Data Card

VIFDD	VIFDD (Visual Intrusion and Fraud Detection Dataset) is a dataset designed for the task of detecting visual intrusions and fraudulent websites. The dataset consists of 240,000 images of webpage screenshots, evenly divided into two categories: SCAM and NORMAL.
DATASET LINK Dataset	DATA CARD AUTHOR(S) **Will be added after acceptance**

Dataset Owners

TEAM(S)	CONTACT DETAIL(S)	AUTHOR(S)
VIFDD Team	**Dataset Owner(s):** will be revealed on acceptance **Affiliation:** will be revealed on acceptance **Contact:** will be revealed on acceptance	Will be revealed on acceptance

Dataset Overview

DATA SUBJECT(S)

Data about systems or products and their behaviors

DATASET SNAPSHOT

Size of Dataset	5.65 GB
Number of Instances	240,000
Labeled Classes	2
Number of Labels	2
Average Labels Per Instance	1
Algorithmic Labels	None
Human Labels	Some[1]

Above: Summary of VIFDD dataset.

[1] Some websites needed to be cross-verified manually before being labelled as SCAM.

CONTENT DESCRIPTION

Each datapoint in the VIFDD dataset represents a screenshot of a webpage. The content of a datapoint includes:

- **Image Data**: A 224x224 pixel image in RGB format, stored as either a PNG or JPG file.
- **Label**: A categorical label indicating whether the webpage is a SCAM or NORMAL. This label is determined based on the source of the URL:
  - **SCAM**: The image is labelled as SCAM if the screenshot is from a known fraudulent website.
  - **NORMAL**: The image is labelled as NORMAL if the screenshot is from a legitimate website.

Sensitivity of Data

SENSITIVITY TYPE(S)

None

FIELD(S) WITH SENSITIVE DATA

Intentionally Collected Sensitive Data

The dataset does not intentionally collect sensitive data.

Unintentionally Collected Sensitive Data

Since the screenshots are taken from live webpages, there is a potential for unintentional capture of sensitive data. This could include:

- **Personal Information**: Names, email addresses, phone numbers, or other personal information visible on the webpage.
- **Pornographic Content**: Explicit content that might be displayed on some scam or adult websites.
- **Violent Content**: Images or text depicting violence that might be present on certain webpages.

Dataset Version and Maintenance

MAINTENANCE STATUS

VERSION DETAILS

MAINTENANCE PLAN

Limited Maintenance The data will not be updated, but any technical issues will be addressed.	**Current Version:** 1.0 **Last Updated:** 05/2024 **Release Date:** N/A	Since the dataset is static, there will be no additions or modifications to the dataset itself. Any discovered issues or updates will be documented and communicated through the dataset's documentation.

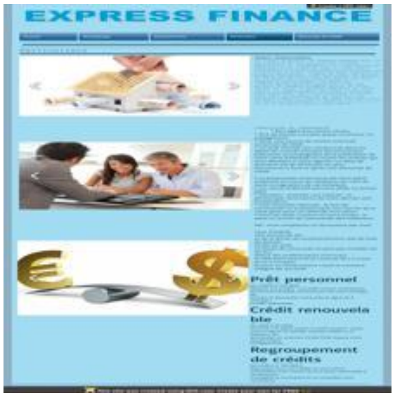
Example of Data Points

PRIMARY DATA MODALITY

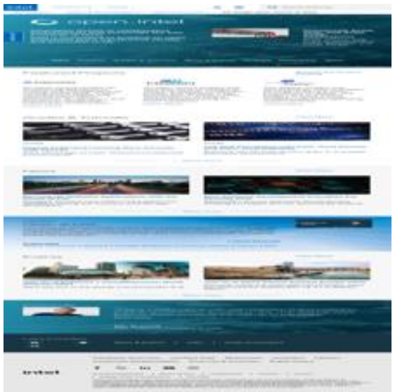
Image Data

SAMPLING OF DATA POINTS

Examples of data in VIFDD



label: SCAM



label: NORMAL

DATA FIELDS

Field Name	Type	Description
image_data	PNG/JPG	Pixel data of the screenshot.
label	Integer/String	Class of the screenshot.

Above: Summary of data fields in VIFDD.

TYPICAL DATA POINT

A typical data point.

Field	Value
image_data	“\xFF\xD8\xFF\xE0\x00\x10JFIF\x00\x01\x01\x00\x00\x01\x00\x01\x...”
Label	[“SCAM”, “NORMAL”]

ATYPICAL DATA POINT

The dataset does not contain atypical data points as far as we know.

Motivations & Intentions

Motivations

PURPOSE(S)	DOMAIN(S) OF APPLICATION	MOTIVATING FACTOR(S)
Research	Cybersecurity, Fraud Detection, Machine Learning, Computer Vision, Web Security, Artificial Intelligence, Data Science, E-commerce Security, Internet Safety, Phishing Detection, Scam Prevention	1. **Enhancing Fraud Detection**. 2. **Improving Cybersecurity**. 3. **Supporting E-commerce Security**. 4. **Filling Data Gaps**. 5. **Encouraging Academic Research**. 6. **Promoting Internet Safety**. The VIFDD-2024 dataset aims to enhance fraud detection, improve cybersecurity, advance machine learning in visual content analysis, and support e-commerce security. It fills data gaps in scam detection research, encourages academic studies, and promotes internet safety by enabling the development of tools to identify and block scam websites.

Intended Use

DATASET USE(S)	SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
Safe for research use	• **Fraud Detection Research**: Developing and evaluating machine learning models for detecting scam websites based on visual content. • **Cybersecurity Studies**: Conducting research in web security to identify patterns and features indicative of fraudulent sites. • **Phishing and Scam Prevention**: Training and testing tools that detect and prevent phishing attacks and scam websites. • **Computer Vision Applications**: Exploring computer vision techniques to analyse and classify webpage screenshots.	• **Production Environments**: Using the dataset in live production environments for real-time fraud detection or security applications without additional validation and testing.
	RESEARCH AND PROBLEM SPACE(S)	CITATION GUIDELINES

	The VIFDD dataset addresses the problem space of visual intrusion and fraud detection on the internet. The dataset supports the advancement of computer vision techniques for webpage classification and the development of machine learning models that differentiate between legitimate and fraudulent sites.	**Guidelines & Steps:**
		1. Include the full citation in your references.
		2. Provide a direct link to the dataset wherever possible.
		BiBTeX:
		```
		 @dataset{vifdd,
		  title={VIFDD: Visual Intrusion and Fraud Detection Dataset},
		  author={on acceptance},
		  year={2024},
		  url={ https://www.kaggle.com/datasets/ae6753dd33076e09f4803961a000e8e5adbd6a1d6c16829195f422513720af3c }
		}
		```

Provenance

Collection

METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
Scraped or Crawled	**Collection Type**: Scraped or Crawled **Source**: • **Normal URLs**: Gathered from the Alexa Top 1 Million Websites list. • **Scam URLs**: Collected from various online sources including PhishTank, Scamwatch, Scamalytics, and Spamhaus. **Is this source considered sensitive or high-risk?** No **Dates of Collection**: Oct 2023 - Mar 2024 **Primary Modality of Collected Data**: • **Image Data**: Screenshots of webpages resized to 224x224 pixels, stored in PNG and JPG formats. **Update Frequency for Collected Data**: • Static	**Normal URLs:** • Normal URLs were sourced from the Alexa Top 1 Million Websites list, representing a diverse range of legitimate websites across various domains and industries. **Scam URLs:** • Scam URLs were collected from several reputable online sources, including PhishTank, Scamwatch, Scamalytics, and Spamhaus. These sources specialize in identifying and cataloging fraudulent websites, serving as valuable repositories for detecting online scams and threats. **Additional Notes:** The Alexa Top 1 Million Websites list is a widely recognized resource for gathering website popularity and traffic data. The selection from this list provides a comprehensive sample of normal webpages for the dataset. he inclusion of URLs from multiple sources enhances the diversity and representativeness of scam webpages in the dataset.

COLLECTION CADENCE	DATA INTEGRATION	DATA PROCESSING

Static

Data was collected once from single or multiple sources.

Included Fields

Field Name	Description
URL	This field contains the URLs of normal webpages sourced from the Alexa Top 1 Million Websites list.
Screenshot	Screenshots of the webpages in PNG and JPG formats, resized to 224x224 pixels. Each screenshot provides a visual representation of the webpage content, capturing layout, design elements, and textual information. These screenshots are the primary data used for training and evaluating models in the dataset

Excluded Fields
None

Scraped or Crawled

Description: Data for the VIFDD dataset was collected through web scraping and crawling methods. Normal URLs were gathered from the Alexa Top 1 Million Websites list, while scam URLs were collected from various online sources including PhishTank, Scamwatch, Scamalytics, and Spamhaus.

Methods employed: Web scraping and crawling techniques were used to extract URLs from the selected sources. This involved programmatically navigating webpages, identifying relevant URLs, and extracting them for further processing.

Tools or libraries: Node.js-based libraries such as Puppeteer were utilized for web scraping and crawling tasks. These libraries provide functionality for parsing HTML content, navigating webpage structures, and extracting desired information.

Additional Notes: The collected URLs were then used as input to the custom script responsible for taking screenshots of the corresponding webpages. The screenshots were resized to a standardized format of 224x224 pixels and stored in PNG/JPG formats for further analysis.

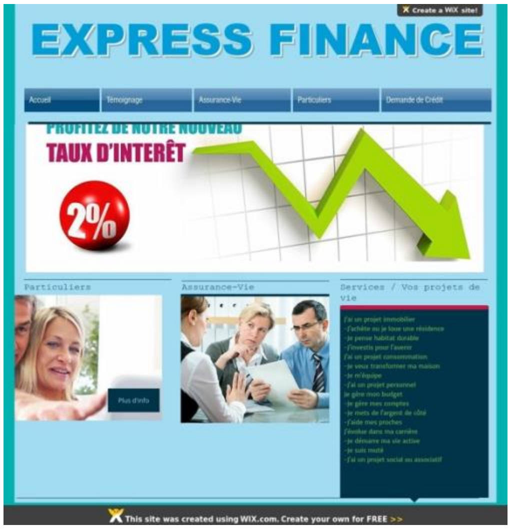
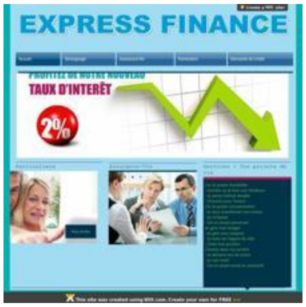
Transformations

Synopsis

TRANSFORMATION(S) APPLIED	FIELD(S) TRANSFORMED	LIBRARY(IES) AND METHOD(S) USED
Others (Resizing)	**Transformation Type:** Resizing • **Original Data:** Image files in PNG and JPG formats of all sizes. • **Transformed Data:** Image files resized to 224x224 pixels • **Description:** The image data, consisting of screenshots of webpages, was transformed by resizing each image to a standardized dimension of 224x224 pixels. This transformation ensures uniformity in the size of images across the dataset, facilitating consistent analysis and modelling. **Additional Notes:** • The transformation of image data to a standardized size simplifies the processing and analysis pipeline, as models can expect inputs of consistent dimensions. • Resizing the images to a common size also helps in reducing computational complexity during model training and inference.	**Transformation Type:** Resizing **Method:** The image resizing process was implemented using the OpenCV library in Python. OpenCV (Open Source Computer Vision Library) is a popular open-source computer vision and image processing library that provides a wide range of functions for manipulating images. The **cv2.resize()** function from the OpenCV library was utilized to resize each image to the desired dimensions of 224x224 pixels. **Platforms, tools, or libraries**: • OpenCV **Transformation Results:** The resizing process transformed the original images into a standardized size of 224x224 pixels, ensuring consistency in the dimensions of all images in the dataset.

Breakdown of Transformations

Others (Resizing)	METHOD(S) USED	COMPARATIVE SUMMARY

The raw image data, consisting of screenshots of webpages is of various dimensions depending on the web page content and length. The raw image data was transformed by resizing each image to a standardized dimension of 224x224 pixels.	**Platforms, tools, or libraries**: OpenCV	**Before transformation:**  Create a WIX site! EXPRESS FINANCE Accueil Témoignage Assurance-Vie Particuliers Demande de Crédit PROFITEZ DE NOTRE NOUVEAU TAUX D'INTERÊT 2% Particuliers Plus d'info Assurance-Vie Services / Vos projets de vie This site was created using WIX.com. Create your own for FREE >> **After transformation:**  **Above:** Before and after comparison of transformation.