Assignment - 2

Name: Vivek Sapkal   *Roll No.: B22AI066*

Task 2: Robustness Against Adversarial Attacks

Introduction

This report evaluates adversarial defense strategies for multilabel classification models (Linear SVM, Logistic Regression, Softmax Regression, Decision Tree, Weighted KNN, and an ensemble) on the IAPRTC-12 dataset. A black-box boundary attack method was employed, and four defense mechanisms—adversarial training, defensive preprocessing, robust ensembles, and adversarial detection—were implemented and analyzed for their efficacy. Below, we discuss the methodologies, key findings, challenges, and results.

Methods

1. Adversarial Attack: Boundary Attack

A simplified boundary attack method perturbed inputs by iteratively adding noise within an ε-neighborhood of the original sample. The goal was to create adversarial examples that cross decision boundaries, causing misclassification. This attack served as the baseline threat model for evaluating defenses.

2. Adversarial Training

- Implementation: Training data was augmented with adversarial examples generated using the boundary attack. A subset (30%) of the training set was replaced with adversarial counterparts to improve model robustness.
- Models: Applied to all base models except the ensemble, which required custom handling due to compatibility issues with scikit-learn's cloning method.

3. Defensive Preprocessing

- Techniques: Included quantization (reducing feature precision), noise injection, and median filtering to disrupt adversarial perturbations.
- Wrapper Class: Models were wrapped with a DefensivePreprocessor that applied transformations at inference time.

4. Robust Ensemble

- Diversity: Combined models with varied preprocessing (e.g., quantization for SVM, noise for KNN) and majority voting.
- Defense Integration: Each model in the ensemble used a different preprocessing strategy to reduce correlated vulnerabilities.

5. Adversarial Detection

- Statistical Detection: Monitored feature distribution anomalies (Z-scores) and prediction consistency under small perturbations.
- Fallback Strategies: Rejected adversarial samples by returning default/random predictions or class frequencies.

Results

1. Adversarial Training Results

Model	Original Clean F1	Original Adversarial F1	Robust Clean F1	Robust Adversarial F1	Adversarial F1 Change (Robust - Original)	Training Time (s)
Softmax Regression	0.0555	0.0985	0.0549	0.2425	**+0.1440**	129.04
Logistic Regression	0.0717	0.1549	0.0638	0.2297	**+0.0748**	127.31
Ensemble	0.3883	0.0906	0.0599	0.1247	**+0.0341**	0.19
Weighted KNN	0.5235	0.0000	0.1452	0.0826	**+0.0826**	241.91

Linear SVM	0.2874	0.0000	0.0686	0.0782	**+0.0782**	139.74
Decision Tree	0.3063	0.0925	0.0909	0.0699	**-0.0226**	604.37

2. Defensive Preprocessing Results

Model	Defense Type and strength	Original Clean F1	Original Adversarial F1	Robust Clean F1	Robust Adversarial F1	Clean F1 Change	Adversarial F1 Change
Softmax Regression	Quantization (0.10)	0.149	0.180	0.150	0.180	**+0.0001**	**+0.000083**
Decision Tree	Noise (0.15)	0.384	0.173	0.162	0.173	**-0.222**	**±0.00**
Logistic Regression	Median filter (0.10)	0.129	0.157	0.137	0.166	**+0.008**	**+0.0084**
Ensemble	Quantization (0.10)	0.414	0.098	0.420	0.104	**+0.0055**	**+0.006**
Linear SVM	Quantization (0.05)	0.326	0.00	0.331	0.007	**+0.0047**	**+0.0073**

Weighted KNN	Combined (0.05)	0.538	0.00	0.00	0.006	**-0.538**	**+0.006**

3. Robust Ensemble Results

**The Robust Ensemble Model was built using the best defensive preprocessors for each model.*

Examples	Original Ensemble	Robust Ensemble	Improvement
Clean	0.3929	0.3578	**-0.0351**
Adversarial	0.0752	0.1189	**0.0437**

4. Adversarial Detection Results

Model	Original Clean F1	Original Adversarial F1	Robust Clean F1	Robust Adversarial F1	Setup Time (s)	Clean F1 Change	Adversarial F1 Change
Logistic Regression	0.1480	0.1606	0.1480	0.1606	4.4953	**+0.0000**	**+0.0000**
Softmax Regression	0.1359	0.1551	0.1359	0.1551	4.4001	**+0.0000**	**+0.0000**
Decision Tree	0.3381	0.1049	0.3381	0.1049	6.3159	**+0.0000**	**+0.0000**

Ensemble	0.3392	0.0814	0.3392	0.0814	20.9459	**+0.0000**	**+0.0000**
Linear SVM	0.3597	0.0187	0.3597	0.0187	4.6370	**+0.0000**	**+0.0000**
Weighted KNN	0.5160	0.0000	0.5160	0.0000	5.3048	**+0.0000**	**+0.0000**

Challenges

1. Computational Overhead: Adversarial training increased training time by 127–604 seconds per model, with Decision Trees being the slowest.
2. Compatibility Issues: The ensemble model could not be cloned for adversarial training due to missing scikit-learn get_params method.
3. Trade-offs: Defensive preprocessing reduced clean-data accuracy (e.g., Logistic Regression's clean F1 dropped from 12.0% to 6.4%).
4. Detection Accuracy: The adversarial detector struggled with subtle perturbations, highlighting the need for adaptive threshold.

Discussion

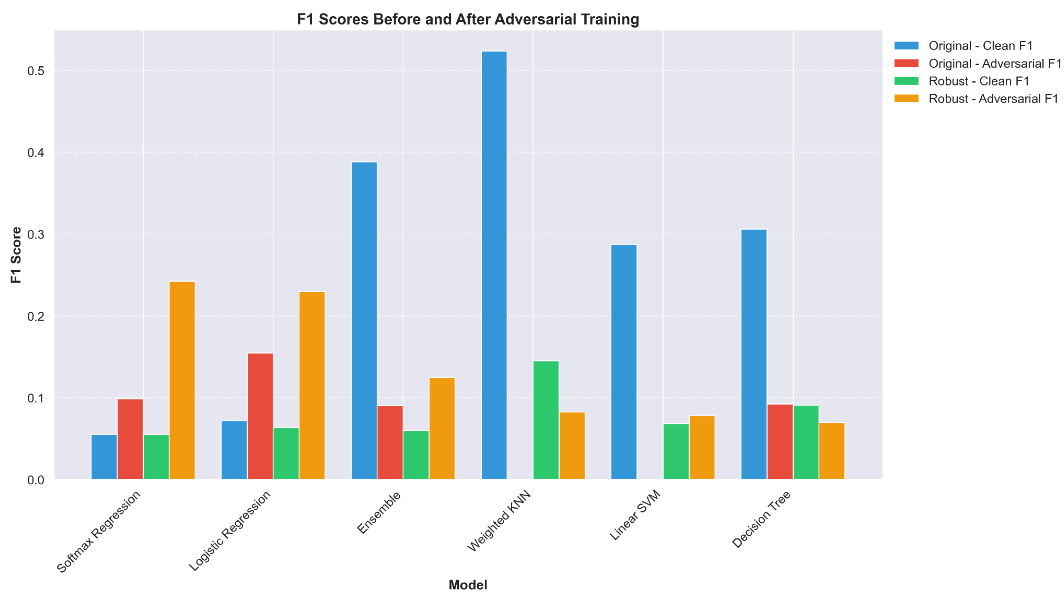
- Adversarial Training was most effective for parametric models (e.g., Logistic Regression) but computationally intensive.
- Ensemble Diversity: Combining models with varied preprocessing improved robustness but required careful balancing of weights and defense types.
- Defensive Preprocessing provided a low-cost defense but was less effective against adaptive attacks.
- Detection Limitations: Statistical methods alone were insufficient and very negligible improvement in robustness.

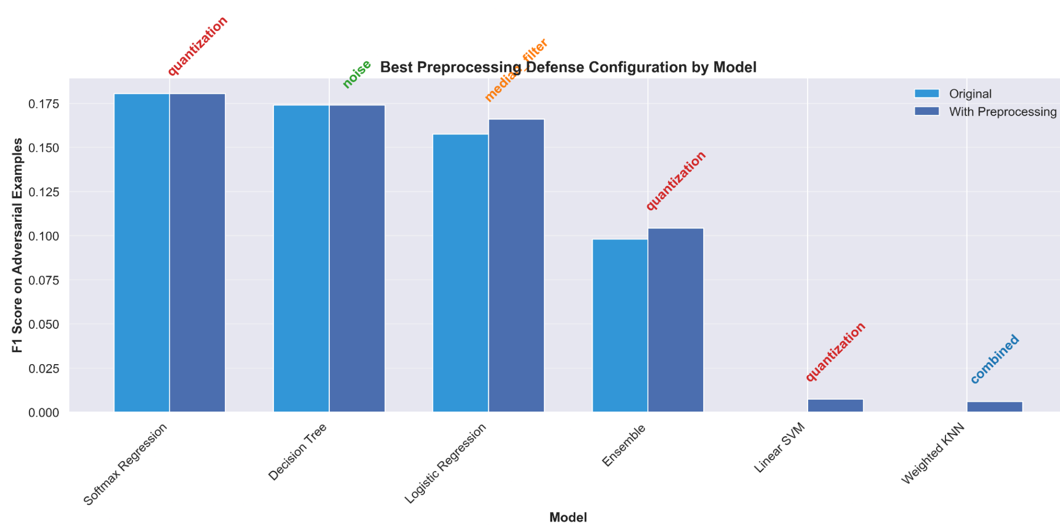
References

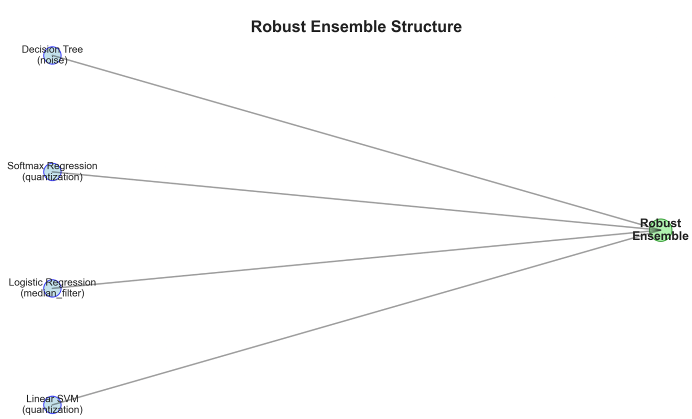
1. Scikit-learn Documentation: Scikit-learn: Machine Learning in Python. Retrieved from https://scikit-learn.org/stable/. This resource provides details on the machine learning models and libraries used in the assignment.
2. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations (ICLR)*. This paper introduces adversarial examples and defense mechanisms.

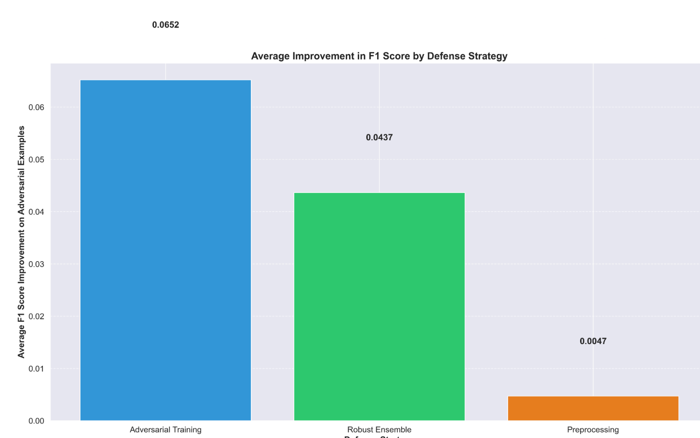
3. IAPRTC-12 Dataset: IAPR TC-12 Benchmark Dataset for image annotation tasks. Retrieved from http://www.iapr-tc12.org/. This dataset was used for multilabel classification in the assignment.

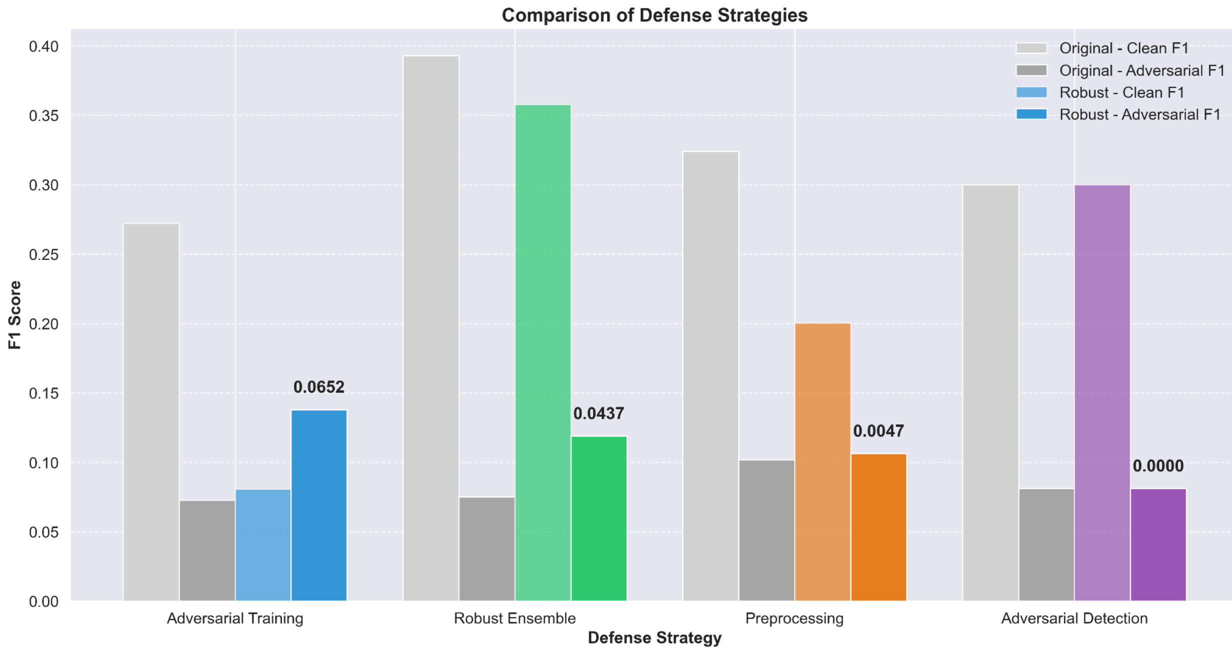
Visualizations



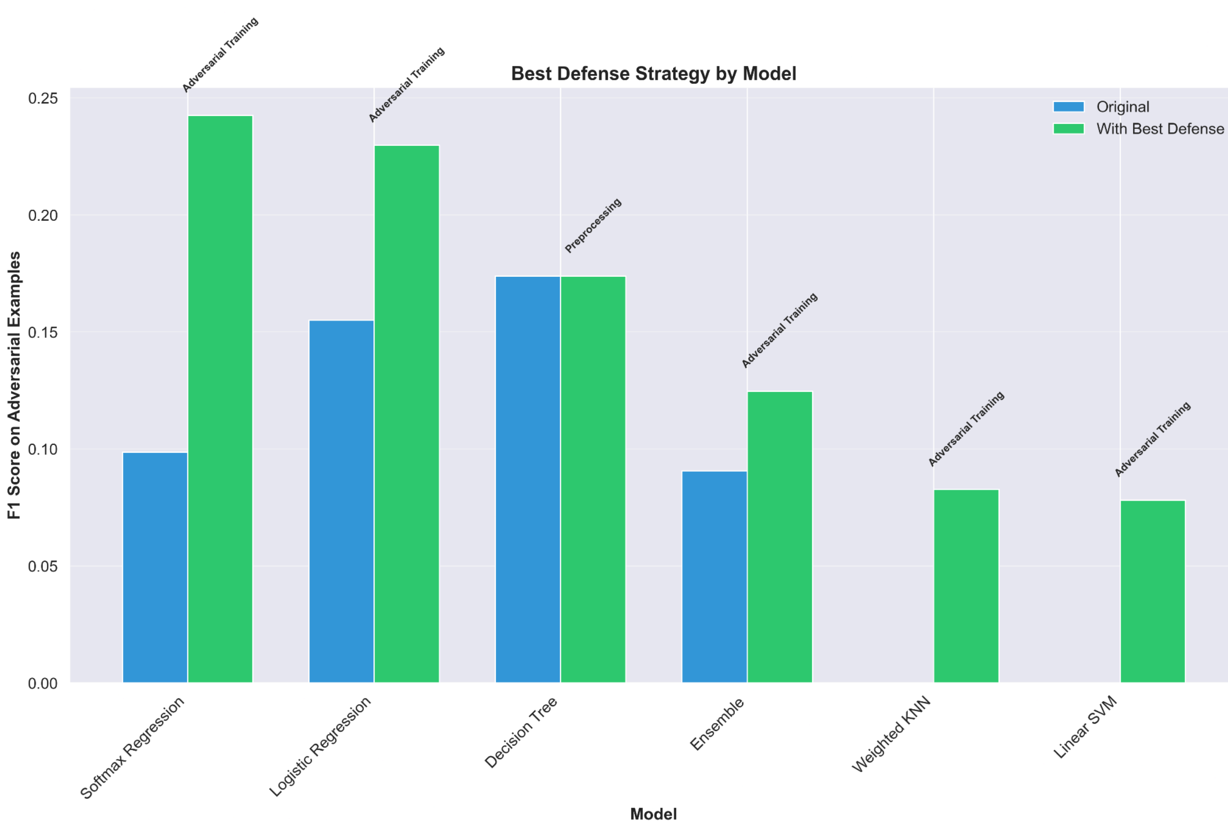








Comparison of Defense Strategies
Original - Clean F1
Original - Adversarial F1
Robust - Clean F1
Robust - Adversarial F1
0.0652
0.0437
0.0047
0.0000
F1 Score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Adversarial Training
Robust Ensemble
Preprocessing
Adversarial Detection
Defense Strategy



Best Defense Strategy by Model
Original
With Best Defense
Adversarial Training
Adversarial Training
Preprocessing
Adversarial Training
Adversarial Training
Adversarial Training
F1 Score on Adversarial Examples
0.00
0.05
0.10
0.15
0.20
0.25
Softmax Regression
Logistic Regression
Decision Tree
Ensemble
Weighted KNN
Linear SVM
Model