# DECLARATION BY THE CANDIDATE

I hereby declare that the work, which is being presented in this project entitled **“SPECIOUS (Spectral Perturbation Engine for Contrastive Inference Over Universal Surrogates)”**, is an authentic record of my own work carried out by me under the supervision and guidance of **Ms. Himani Tyagi, Assistant Professor, University School of Automation of Robotics, GGSIPU.**

This project was undertaken as a part of the major project report for the partial fulfillment of B.Tech. (AI&ML).

I have not submitted the matter embodied here in this project for the award of any other Degree/Diploma.

**Dhruv Kumar**

**Enrollment No.: 00519011622**

**BTech (AI&ML)**

# CERTIFICATE

This is to certify that this MAJOR PROJECT REPORT **“SPECIOUS (Spectral Perturbation Engine for Contrastive Inference Over Universal Surrogates)”** is submitted by **Dhruv Kumar,** who carried out the project work under my supervision. I approve this project for submission of the Bachelor of Technology (B.Tech., AI&ML) in the University School of Automation and Robotics (USAR), Guru Gobind Singh Indraprastha University, Delhi.

## Date:

**Ms. Himani Tyagi**

**Assistant Professor**

**USAR, GGSIPU EDC**

**PLAGIARISM REPORT**

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to all those who have supported and guided me throughout the course of this project. First and foremost, I am deeply thankful to **Prof. A. K. Saini, Director**, East Campus, GGSIPU, for providing the necessary academic environment and resources that made this project possible. I would also like to extend my heartfelt gratitude to **Prof. Arvinder Kaur**, Dean USAR, for her constant encouragement and support during this academic journey.

I sincerely thank to **Prof. Dr. Abha Agarwal**, Program Coordinator, for his/her valuable guidance in structuring the project and ensuring smooth progress. I am especially grateful to **Assistant** **Prof. Dr. Ruchika Lalit**, Teacher-in-Charge, for the consistent academic supervision, helpful insights, and feedback that significantly enhanced the quality of my work.

A very special thanks to my project mentor, **Ms. Himani Tyagi,** for her unwavering support, expert guidance, and constant motivation throughout every phase of this project. Their dedication and mentoring were crucial to the completion of this work. Lastly, I would like to thank all my faculty members, friends, and family who contributed in any way toward the success of this project.

Thanking You

**Dhruv Kumar**

**Enrollment No.: 00519011622**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **TITLE** | **PAGE NO.** |
| 1 | Abstract | 1 |
| 2 | Introduction  2.1 Background and Problem Statement  2.2 Limitations of Existing Defenses  2.3 Our Approach | 2 |
| 3 | Literature Survey  3.1 Style-Cloaking and Dataset-Poisoning Attacks  3.2 Frequency-Domain Adversarial Perturbations  3.3 Perceptual Metrics in Adversarial Optimization  3.4 Y-Channel Specific Attacks  3.5 Universal and Multi-Model Attacks  3.6 Adversarial Patches and Physical-World Attacks  3.7 Object-Detection Patch Attacks  3.8 Certified Robustness via Randomized Smoothing  3.9 Extending Certification to Transformations  3.10 Feature De-Noising Networks | 5 |
| 4 | Materials and Proposed Methodology  4.1 Data Preparation  4.2 Model Architecture  4.3 Frequency Domain Perturbation Block  4.4 Specious Loss: Joint Perceptual-Feature Objective | 9 |
| 5 | Experiments & Results  5.1 Training Dynamics  5.2 CLIP Cosine Similarity Drop Distribution  5.3 ResNet-50 Classification Results  5.4 CLIP Zero shot Classification  5.5 Summary of Findings  5.6 Simulation and User-Friendly Website Screenshots | 17 |
| 6 | Conclusion | 26 |
| 7 | References | 27 |
| 8 | Annexure | 30 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **TITLE** | **PAGE NO.** |
| 4.1 | Working Flow Diagram of SPECIOUS | 9 |
| 5.1.1 | Plot of the smoothed LPIPS, feature loss, and total loss over ≈ 7,200 training steps (rolling window = 100) | 17 |
| 5.1.2 | Zoomed in LPIPS over training steps | 18 |
| 5.2 | The histogram and KDE of similarity drops | 19 |
| 5.3 | Scatter plot of the data points showcasing whether they got fooled or not, along with their LPIPS and Confidence Drop, and a bar plot showcasing the Fooling Rate when got tested on ResNet50 | 20 |
| 5.4 | Scatter plot of the data points showcasing whether they got fooled or not, along with their LPIPS and Confidence Drop, and a bar plot showcasing the Fooling Rate when tested on CLIP Zero Shot Prediction on the CIFAR-100 test dataset | 21 |
| 5.6.1 | Home Page | 22 |
| 5.6.2 | Website How It Works Section | 22 |
| 5.6.3 | Detailed Flow Diagram of Specious on the Website | 23 |
| 5.6.4 | Comparison of SPECIOUS with Glaze and Nightshade | 23 |
| 5.6.5.1 | Use it Tab on Website | 24 |
| 5.6.5.2 | SPECIOUS in Action on Website | 24 |
| 5.6.6 | Meet the team | 25 |

# ROLES AND RESPONSIBILITIES

|  |  |  |
| --- | --- | --- |
| **S.No** | **NAME** | **ROLES AND RESPONSIBILITIES** |
| 1 | Dhruv Kumar | Research, Planning, Methodology,  Execution, Drafting |
| 2 | Deepanshu Singh | Research, Methodology,  Execution, Drafting |
| 3 | Harshveer Singh | Data Curation and Processing,  Validation, Execution, Drafting |