TABLE OF CONTENTS

CHAPTER NO	DESCRIPTION	PAGE NO
1.	**INTRODUCTION**	1
	1.1.1What is phishing?	1
	1.1 Types of Phishing Attacks	2
	1.2 Anti-Phishing Solutions	3
	1.2.1 Phishing Prevention Solutions	3
	1.2.2 User Training Solutions	3
	1.2.3 Phishing Detection Solutions:	3
	1.3 A Machine Learning Approach for Phishing Detection	4
2.	**LITERATURE SURVEY**	5
	2.1 Phishing Detection Schemes	5
	2.1.1 Search Engine Based	5
	2.1.2 Heuristics and Machine Learning Based	6
	2.1.3 Phishing Blacklist and Whitelist Based	6
	2.1.4 Visual Similarity Based	6
	2.1.5 DNS Based	7
	2.1.6 Proactive Phishing URL Based	7
3.	**PROBLEM IDENTIFICATION**	8
4.	**OBJECTIVES**	9
5.	**REQUIREMENT ANALYSIS**	10
	5.1 Functional Requirements	10
	5.2 Non-Functional Requirements	10
	5.3 Software Requirements	11
	5.3.1 Kaggle	11
	5.3.2 Python 3	11
	5.3.3 MongoDB	11
	5.3.4 Flask	11
	5.3.5 Visual Studio Code	11
	5.4 Hardware Requirements	12
6.	**SYSTEM DESIGN**	13
	6.1 Data flow Diagram (DFD)	13
	6.2 Sequence Diagram	14

6.3 Use Case Diagram 15

7. METHODOLOGY 16

8. IMPLEMENTATION 21

8.1 Importing Python libraries 21

8.2 Creation of clean dataset 22

8.3 Creation of Blacklist Database 22

8.4 Data Visualization 24

8.5 Creation and Training of Random Forest Classifier 26

8.6 Obtaining the Feature Importance 26

8.7 Performance Evaluation of Trained Classifier 27

8.8 Comparison with other ML models 28

8.9 Extraction of URL Features 30

8.10 Combining Both Blacklist and ML Approach 35

8.11 Catch Phish - A Webpage for Phishing Website Detection 36

9. TESTING 40

9.1 Importance of Testing 40

9.2 Levels of Testing 40

CONCLUSION 43

FUTURE SCOPE 44

REFERENCE 45

List of Figures

Page No.

1. A classification of phished website detection schemes 5
2. Data Flow Diagram of Application 13
3. Use Case Diagram for Application 14
4. Sequence Diagram of Application 15
5. URL Features Dataset 16
6. URL Features Dataset (Continued) 17
7. Application Flowchart 20
8. Scatter plot of status vs URL features 25
9. Bar graph for visualization of URL features 26
10. Feature importance graph 27
11. Confusion matrix 28
12. Performance measurement 28
13. Comparison with SVM and Logistic Regression 29
14. Graph for comparing performance of different classifiers 30
15. Homepage of the web application 37
16. Catch a phish page 37
17. Invalid URL entry 38
18. Result for legitimate URL 38
19. Result for phishing URL 39