

TABLE OF CONTENTS

<i>HALAMAN PENGESAHAN</i>	<i>i</i>
<i>PERNYATAAN BEBAS PLAGIASI</i>	<i>ii</i>
<i>FOREWORDS</i>	<i>iii</i>
<i>LIST OF FIGURES</i>	<i>vii</i>
<i>LIST OF TABLES</i>	<i>viii</i>
<i>ABSTRACT</i>	<i>ix</i>
<i>CHAPTER I</i>	<i>1</i>
1.1. Research Background	1
1.2. Research Problem.....	3
1.3. Research Scope	3
1.4. Research Objective	3
1.5. Research Advantage	4
<i>CHAPTER II</i>	5
<i>CHAPTER III</i>	8
3.1. Adversarial Attack	8
3.2. Whitebox Adversarial Attack.....	8
3.3. FaceForensics++ Dataset.....	10
3.4. DeepFakes	10
3.5. MesoNet	11
3.5.1. Meso-4.....	11
3.6. K-winner-takes-all activation function	12
3.7. FGSM Attack	14

3.8.	Iterative FGSM Attack	14
3.9.	C&W Attack	15
3.10.	Performance Evaluation	16
3.11.	Loss Function.....	17
CHAPTER IV.....		19
4.1.	Research Description.....	19
4.1.1.	Dataset Description.....	21
4.1.2.	Dataset Acquisition.....	21
4.1.3.	Dataset Preprocessing.....	22
4.2.	Why K-WTA?	23
4.2.1.	Average-takes-all activation function	24
4.2.2.	K-losers-takes-all (K-LTA) activation functions	25
4.2.3.	“Otherwise” Value.....	25
4.3.	Model Training.....	26
4.4.	Adversarial Example Generation.....	28
4.5.	Evaluation Plan.....	29
CHAPTER V.....		31
5.1.	Workflow.....	31
5.1.1	Data Collection.....	31
5.1.2.	Data Preprocessing	31
5.1.3.	Creating Model.....	34
5.1.4.	Crafting the Attack	36
5.1.5.	Model Training.....	42
5.1.6.	Evaluation Method	44
CHAPTER VI.....		47
6.1.	Results of Face Extraction and Preprocessing.....	47
6.2.	Results of Adversarial Attack.....	48

6.3.	Model Training	50
6.4.	Model Evaluation.....	53
<i>CHAPTER VII</i>	<i>.....</i>	<i>60</i>
7.1.	Conclusion	60
7.2.	Suggestion.....	61
<i>REFERENCES</i>	<i>.....</i>	<i>62</i>