N° d'ordre :
N° de série :

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research

ECHAHID HAMMA LAKHDAR UNIVERSITY - EL OUED
FACULTY OF EXACT SCIENCES
Computer Science department

End of Study Thesis
Presented for the Diploma of

ACADEMIC MASTER

Domain : **Mathematics and Computer Science**
spinneret: **Computer Science**
Speciality : **Artificial Intelligence and Distributed Systems**

Presented by :

- **Laibi Khadidja**
- **Djebali Manel**

Theme

Recognition of CAPTCHAs from a web page

Supported in: - 06 - 2021 In front of jury:

M.	OTHMANI Samir	MCA	President
M.	BELILA khaoula	MAA	Reporter
Dr.	LEDJDEL Brahim	MAA	Supervisor

University Year: 2020/2021

Dedicace

"

I dedicate this modest work to :
First of all to my parents; to my dear mother " Saliha " who gave birth to me and looked after my happiness and to for all her prayers and benedictions and help; to my father " Abdelkader " who gave everything and sacrificed. I express my deep gratitude to them and promise to always be by their side.
To my very dear brother " Soufaine " who was there for me throughout and gave me lots of support , for his constant encouragement and moral support, which always gave me hope and believed in my abilities.
My dear sisters(Messaouda and Abir and Asma), for their encouragement throughout my academic career and moral support.
To my affectionate little sister " Rayane " who gave me all the love and happiness.
To the children of my sister, whose presence gives me happiness (Soulaimane and naoufel).
To all of my colleagues and my friends, for the unforgettable memories. I hope that they fulfill all that they wish.
To all teachers who believed in my abilities and encouraged me in my education career. I am extremely grateful to them.

"

Djebali Manel

Dedicace

“

I dedicate my graduation project to the one who harvested thorns from my path, to pave the way for me to learn, my beloved father "Hamza".
To the one who nursed me love and tenderness To the symbol of love and healing To the white heart my beloved mother “Nadia”.
To pure and gentle hearts and innocent souls To the winds of my life my brothers and sisters "Dhia", "Ahmed", "soundes" , "Rhitedj".

”

Laibi Khadidja

Acknowledgement

“

First thanks to God Almighty, for giving us the strength and courage to complete this thesis.Then our parents for all of their efforts.
We convey our sincere thanks to our respected professor " Brahim Lejdel ".
We are also pleased to thank everyone who advised, guided, or contributed with us in preparing this research.

”

Abstract

In recent years, machine learning has developed on a large scale, as intelligent systems and computer programs are able to simulate the mental functions of the human brain due to the continuous development of machine learning algorithms, in addition to big data. This development has negative uses in terms of security and protection, including the ability to break CAPTCHA test.CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), a system used to verify whether a user is human or bot to protect web sites and web services from hacking by computer software (robots), the text-based CAPTCHA test is the most commonly used type.

Our goal is to study text-based CAPTCHA break mechanisms and and the great development of machine learning algorithms facing their security.In addition to studying an application to break text-based CAPTCHA by training a SVM model of data set containing numbers and letters automatically sent to target site by Python software.the results we get show that our model has achieved good results in breaking captcha text in several sites.

Keywords :Text-based CAPTCHA, CAPTCHA breaking, Machine Learning,Segmentation, SVM, python

---

Résumé

Ces dernières années, l'apprentissage automatique s'est développé à grande échelle, car les systèmes intelligents et les programmes informatiques sont capables de simuler les fonctions mentales du cerveau humain grâce au développement continu d'algorithmes d'apprentissage automatique, en plus des mégadonnées. Ce développement a des utilisations négatives en termes de sécurité et de protection, y compris la possibilité de casser le test captcha.CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), un système utilisé pour vérifier si un utilisateur est humain ou un bot pour protéger les sites Web et les services Web du piratage par un logiciel informatique (robots), le test CAPTCHA basé sur du texte est le type le plus couramment utilisé.

Notre objectif est d'étudier les mécanismes de rupture CAPTCHA textuels et le grand développement d'algorithmes d'apprentissage automatique face à leur sécurité. En plus d'étudier une application pour rupture le CAPTCHA textuel en formant un modèle SVM d'ensemble de données contenant des chiffres et des lettres automatiquement envoyés au site cible par le logiciel Python.les résultats que nous obtenons montrent que notre modèle a obtenu de bons résultats en rupture le texte captcha dans plusieurs sites.

Mots clés :CAPTCHA textuel, rupture de CAPTCHA, apprentissage automatique,SVM ,Segmentation, Python

---

ملخص

في السنوات الأخيرة ، تطور التعلم الآلي على نطاق واسع ، حيث أصبحت الأنظمة الذكية وبرامج الكمبيوتر قادرة على محاكاة الوظائف العقلية للدماغ البشري بسبب التطور المستمر لخوارزميات التعلم الآلي ، بالإضافة إلى البيانات الضخمة. هذا التطوير له استخدامات سلبية من حيث الأمن والحماية ، بما في ذلك القدرة على كسر اختبار كابتشا. كابتشا (اختبار تورينج العام المؤتمت بالكامل لإخبار الكمبيوترات والبشر بصرف النظر) ، وهو نظام يستخدم للتحقق مما إذا كان المستخدم بشرًا أم روبوتًا لحماية مواقع الويب وخدمات الويب من القرصنة من قبل برامج الكمبيوتر (الروبوتات) ، يعد اختبار كابتشا المستند إلى النص هو النوع الأكثر استخدامًا.

هدفنا هو دراسة آليات كسر كابتشا المستندة إلى النصوص و التطور الكبير لخوارزميات التعلم الآلي التي تواجه أمنها، بالإضافة إلى دراسة تطبيقية لكسر اختبار كابتشا المستند إلى النص من خلال تدريب نموذج SVM لمجموعة البيانات التي تحتوي على أرقام وأحرف يتم إرسالها تلقائيًا إلى الموقع المستهدف بواسطة برماجية بايثون ، تظهر النتائج التي تحصلنا عليها أن نموذجنا قد حقق نتائج جيدة في كسر نص كابتشا في عدة مواقع.

كلمات مفتاحية : الكابتشا النصية ، كسر الكابتشا، التعلم الآلي، SVM، التجزئة ، بايثون

Contents

List of Figures

List of Tables

List of abbreviations

CAPTCHA :*Completely Automated Public Turing Test to Tell Computers and Humans Apart)*

AI : *Artificial Intelligence*

ML : *Machine Learning*

DL :*deep Learning*

SVM :*Support Vector Machine*

RF : *Random Forest*

K-NN : *k-nearest neighbors algorithm*

DT :*Decision Trees algorithm*

Chapter 1

GENERAL INTRODUCTION

1.1 Context and Motivations

Since the inception of websites and their services, we have to verify that we are not machines in order to benefit from them by passing the so-called ”CAPTCHA test”. CAPTCHA (Completely Automated Public Turing Test for Computers and Humans Apart) This test designed to distinguish between human users and computer programs (bots) on websites in order to protect your services from harmful uses of these bots . This test has a lot of variations, including: CAPTCHA based on text, CAPTCHA based on images, and CAPTCHA based video. To test the CAPTCHA for the human, but the machine is not able to pass it. They relied on tasks that people perform to a machine like text understanding, character recognition, detection of objects, speech recognition, and video classification.
From the first time the Alta Vista website used the CAPTCHA test in 1997 to today, attempts are made to overcome it, either in computer programmes (bots) or by scientists to identify its weaknesses and, every time, to find new strengths in keeping with the rapid development of smart systems. The vast amount of data (big data) available and the improvement of computer capabilities have made learning a top level, resulting in good results breaking the CAPTCHA test. This makes the CAPTCHA test a challenge that the machine can not overcome.
Compared with other types, text-based CAPTCHA is the most commonly used, due to its ease of use and low design cost, which makes it vulnerable to attacks and destruction. On the other hand, it has a wide range of research fields in view of machine learning algorithms and deep learning. Development, it is more robust and unbreakable, or proves its inefficiency.

1.2 Goals

Our goal in this work is to strengthen the verification code technology. By proving the weakness of text-based CAPTCHA in front of machine learning methods. This is why we try to crack the text-based verification code using a model trained on DataSet. The SVM algorithm is used for text recognition, using the image data set of The CAPTCHA library generated in python. Each CAPTCHA image contains 4 to 7 characters out of 26 small and capital English letters and chiffer 0-9.

1.3 Organization

This Thesis includes four chapters:
- **Chapter 1:** This chapter introduces some preliminaries. It is divided into two sections: The first introduces the basic concepts of machine learning and artificial intelligence, As for the second section, we present a definition of CAPTCHA, its history, and its most important applications, with a mention of its classifications, strengths and weaknesses .
- **Chapter 2:** In this chapter, we search for previous works related to the CAPTCHA

from scientific articles by researchers in artificial intelligence.

- **Chapter 3:** This chapter overviews the CAPTCHA test and its types. It presents the text-based CAPTCHA generation process also the state of the art in text-based CAPTCHA breaking.And we put the idea of our design to break this test, mentioning the most important stages we went through.
- **Chapter 4:** This chapter contains our experiment to break a text-based CAPTCHA using DataSet training by svm algorithm used for handwriting and text recognition. With dataset generated by python in which every image contains between four to seven characters.

Chapter 2

DEFINITIONS AND CONCEPTS

2.1 Introduction

The Internet plays a very important role in our daily life, with many web services such as email and search engines available. However, these are often threatened by attacks from computer programs such as bots. To address this problem, CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) was developed to distinguish between human users and computer programs. And based upon Artificial Intelligence.Although this mechanism offers good protection and security for web services, some CAPTCHAs have several weaknesses which allow hackers to infiltrate the mechanism of the CAPTCHA and uses simple machine learning (ML) analysis to test them.

In this chapter, we first introduced a simple concept and history of artificial intelligence, in addition to reviewing machine learning and its most important techniques, applications and main algorithms. In the second section, we presented the concept of the CAPTCHA and its history, followed by the most important types of CAPTCHA, in addition to discussing the strengths and weaknesses of each type. We concluded the chapter with a conclusion.

2.2 Artificial Intelligence

In 1956, John McCarthy coined a term for Artificial Intelligence. ”The science and technology of making smart machines,” he defined. AI is the IT industry that studies and designs intelligent agents who perceive their environment and act in order to maximize their chances of success.The AI could be defined as: ”The ability to simultaneously keep two different ideas in mind and to continue to flow.” However, AI needs to include knowledge from past experiences, reasoning, inference strength, and rapid response. It must also be capable of taking priority decisions and tackling complexity and ambiguity. Artificial Intelligence is said to be available to machines designed to carry out tasks, when they are executed by humans. AI's scientific objective is to understand intelligence through the creation of computer programs that show intelligent behavior by using symbolic inference. Time-independent definition of AI is not. It takes into account the judgment of any system [35].

Artificial intelligence (AI) is a collection of algorithms and intelligence that attempts to emulate human intellect. One of them is Machine Learning, and Deep Learning is one of the techniques used in Machine Learning. ” Chen, Frank (Chen, 2016)”. This is shown in Figure 2.1:

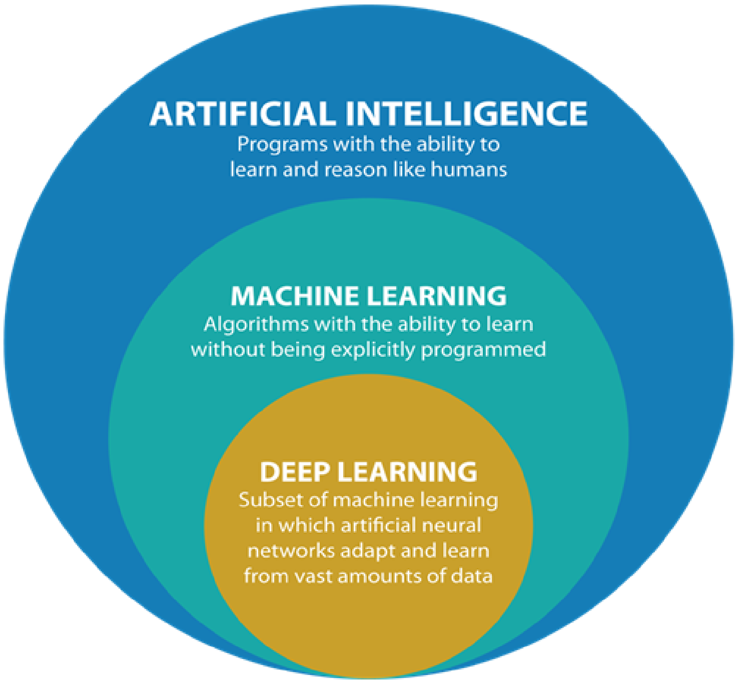


Figure 2.1: AI / ML / DL.

2.3 Machine Learning

Machine learning consists of extracting data from knowledge. Machine learning is a sub-category of artificial intelligence that allows machines to learn from past data or experiences without explicit programming. Machine learning enables a computer system to take predictions or decisions without explicitly programming them using historical data. Machinery learning uses a huge number of structured and half-structured data to allow a machine learning model to produce accurate results or predict based on these data.
Machine learning works by using historical data on an algorithm that learns by itself. It only works for specific domains, such as the creation of a machine model for dog images, which gives only results for photographs of dogs, but it does not respond to new data like a cat image. Machine learning is used in various locations such as for online advice systems, search algorithms for Google, spam filters, suggestions for Facebook automatic friend tagging etc.
It can be divided into three types:
1-Supervised learning.
2-Unsupervised learning.
3-Reinforcement learning.

2.4 Types of Machine Learning

According to the type of problem and input and output data, learning can be divided into different types, as shown in Figure 2.2:

2.4.1 Supervised Learning

Learning with supervision is one of the most essential types of machine learning. The machine algorithm is taught on marked data in this type. Even if the data must be accurately marked for this method to work, controlled education in the right circumstances is extremely powerful.

The ML algorithm provides a small training dataset for supervised learning. This dataset is a smaller part of the larger dataset and provides the algorithm with a basic idea of the problem, solution and data points.In terms of its characteristics, the training dataset is also very much like the final dataset and provides the labeled parameters needed for this problem. The algorithm then finds relationships between the given parameters,

mainly establishing a relation between the cause and effect of the data set variables. The algorithm gives an idea at the end of the training how the data works and how the input and the output are connected [36].

Regression:

The regression analysis is a statistical method for shaping the relationship of one (target) with one or more independent variables to independent (predictor) variables. The regression analysis allows us to understand how the dependent variable's value changes to an independent variable if other independent variables are retained. It forecasts real/continuous values like temperature, age, salary, price, etc [37].

Classification:

Classification is a process to find a feature that helps to divide the dataset into classes based on various parameters. The computer program in classification is formed on a training dataset and categorizes the data into various classes based on this training. It involves classifying the data.
Classified email as spam or not spam, for example. If the monitored learning algorithm marks input data into two separate classes the classification is called binary. Multiclass classification is called the selection between more than two classes. **SVM is the most popular algorithm** (Support Vector Machine).

2.4.2 Unsupervised Learning

Unchecked machine learning benefits from the ability to work with unlabeled information. This means that human work is not needed for the machine to read the data set and for the program to be able to work on much larger datasets [36]. Only parts of this class are presented here:

Clustering:

In unsupervised learning, clustering is an important concept. It mainly involves finding structures or patterns in a collection of unclassified data. The unsupervised learning, clustering algorithm will process your data and find the natural clusters (groups) that exist in the data. You can also modify the number of clusters that your algorithm should recognize. You can change the granularity of these groupings.

Estimation of Density:

Unattended learning problems involving the summary of data distribution.

2.4.3 Reinforcement Learning

Reinforcement learning is directly inspired by how people learn from data in their lives. It has an algorithm that enhances itself and uses a test and error method to learn from new situations. Favorable outputs will be promoted and unfavorable outputs will be discouraged or ”punished.” Q-learning is the most well-known algorithm.

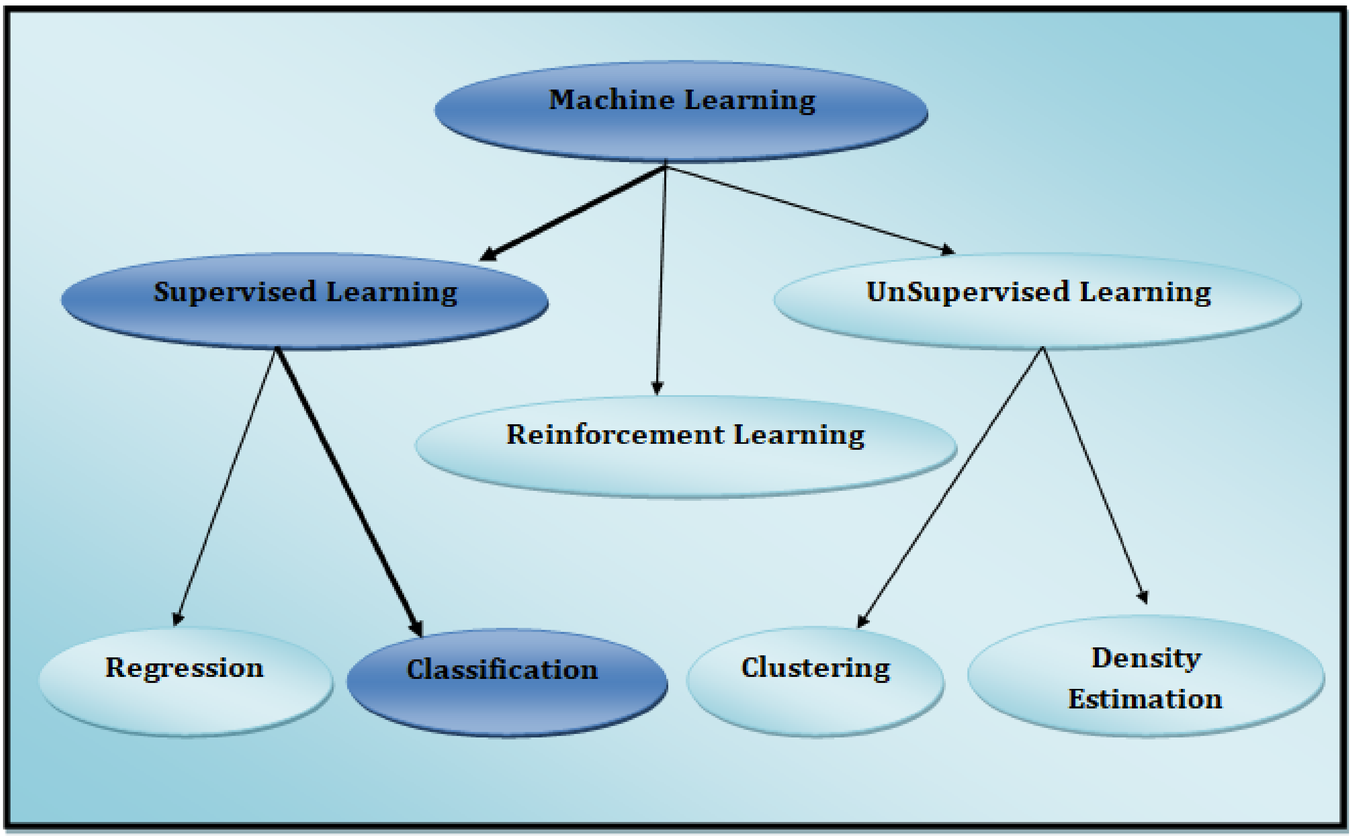


Figure 2.2: Types of Machine Learning.

Supervised Learning VS Unsupervised Learning

The following table 2.1 represents a comparison of supervised learning and supervised learning:

Parameter	Supervised Learning	Unsupervised Learning
DataSet	Labelled	Unlabelled
Method of Learning	Guided learning	The algorithm learns by itself using dataset
Complexity	Simpler method	Computationally complex
Accuracy	More Accurate	Less Accurate

Table 2.1: Supervised VS Unsupervised.

2.5 Advantages of Supervised Learning:

- With the help of supervised learning, the model can predict the output on the basis of prior experiences.
- In supervised learning, we can have an exact idea about the classes of objects.
- Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

2.6 Disadvantages of Supervised learning:

- Supervised learning models are not suitable for handling the complex tasks.
- Supervised learning cannot predict the correct output if the test data is different from the training dataset.
- Training required lots of computation time.
- In supervised learning, we need enough knowledge about the classes of object.

2.7 Algorithms for supervised learning

Algorithms for Supervised Learning There are several algorithms available for supervised learning. As shown below:

- **1-Decision trees**.
- **2-Rule-based algorithms**.

- **3-Naive Bayesian algorithm**.
- **4-Nearest neighbor algorithm**.
- **5-Neural networks**.
- **6-Linear Regression**.
- **7-Support vector machines**.
- **8-Random forest**.
- **9-Bagging**.

Brief description of popular algorithms

2.7.1 Support Vector Machines

SVM is an ML-based statistical model, on the other hand, that tries to separate nonlinear mapping input vectors to a higher dimension function space using the specific kernel function to linearly separate the problem in the input space **(i.e. To construct a linear decide area)**. (VAPNIK 1999). (SVM) In terms of complex function extraction and classification capabilities, CNNs are considered to be more powerful solutions than SVMs. However, some of the key benefits of SVMs are: (a) high capacity for generalization of sparse and noisy data (Meyer and Wien 2001), (b) solid capacity performance (Haussler et al. 2000). A contrast to CNNs where parameter tuning is a painful task).

2.7.2 Decision Trees

The algorithms used to illustrate the results of the decision are using a branching structure. Decision trees can be used to map a decision's possible result. A possible result is each node of a decision tree. Nodes based on the probability of the outcome are assigned the percentages.

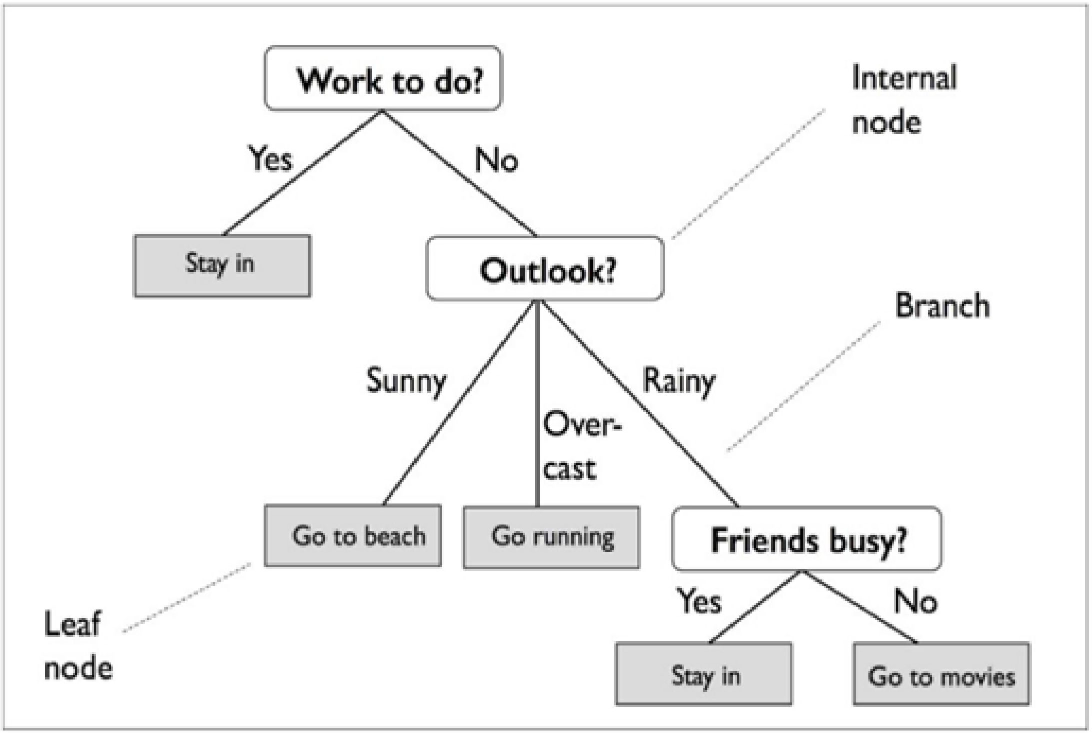


Figure 2.3: Exemple d'arbre de décision.

2.7.3 Random Forest

Random Forest is a popular algorithm for machine learning that is part of the supervised learning technique. It can be used in ML for both problems with classification and regression. It is based on the concept of ensemble learning that combines multiple classifiers in order to solve a complex problem and improve model performance.

"Random Forest is a classification system which contains a number of decision-making bodies on several sub-sets of the data set, using an average to improve the predictive accuracy of this data package". The random forest takes the forecasts from each tree instead of relying on a decision tree, and it forecasts the final output based on majority vote of predictions.
The greater number of trees in the forest, the more accurate it is and the issue of overfitting is avoided.
The below diagram explains the working of the Random Forest algorithm [38]:

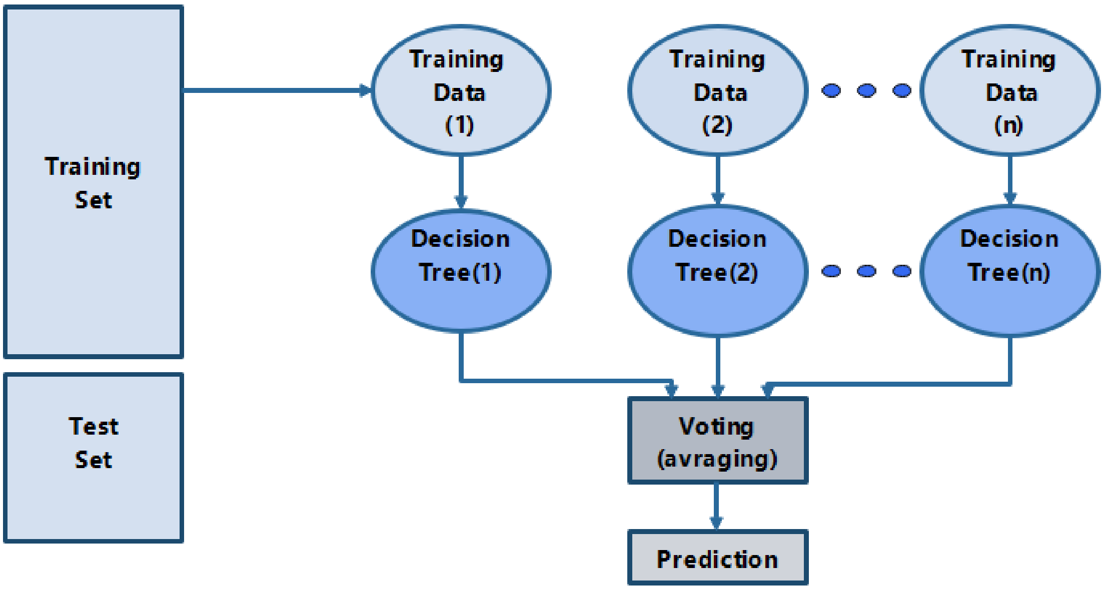


Figure 2.4: Diagram explains the working of the Random Forest algorithm.

2.7.4 Bagging

Bagging and boosting are the two primary types of ensemble machine learning. Bagging, also known as the bootstrap aggregating, is a method that may be used for regression and statistical categorization. Bagging is used with decision trees to increase model stability in terms of reducing variance and enhancing accuracy, which avoids the problem of overfitting.
In ensemble machine learning, bagging takes numerous weak models and aggregates the predictions to get the best one. The weak models specializes in certain areas of the feature space, allowing bagging leverage predictions to come from any model to get the highest level of accuracy [39].

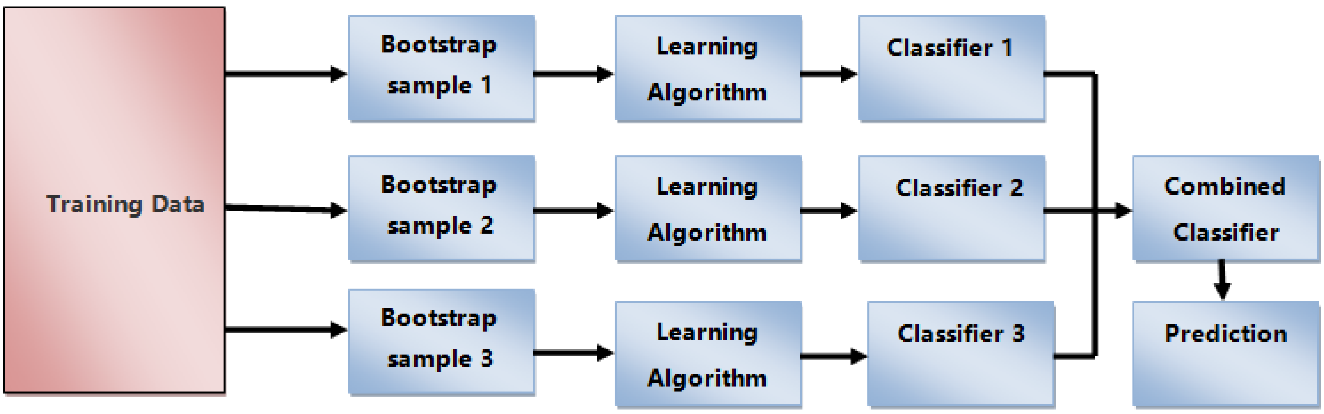


Figure 2.5: Bagging (Bootstrap Aggregation).

2.8 Machine Learning Application

Machine learning used in many domains, among which we mention:

- Image recognition (face detection, character recognition).
- NLP Natural Language Processing (word prediction, top modeling).
- Speech Recognition.
- Medical diagnosis (anomaly identification in medical images.)
- Information extraction.
- Big data visualization.
- Real-time decision and Robot navigation.

2.9 Deep Learning

Also called (Deep Neural Learning) or (Deep Neural Network), this is an artificial intelligence function that mimics the work of the human brain in processing data and creating models to use in the process of making decision making. Deep learning is a subset of machine learning in artificial intelligence (AI) that has networks capable of learning unsupervised from unstructured or disaggregated data.

Comparison of Machine Learning Vs Deep Learning

Below is a table 2.2 of Comparison between Machine Learning and Deep Learning [40]:

Machine Learning	Deep Learning
Machine learning uses algorithms to parse data,learn from that data,and make informed decisions based on what it has learned.	Deep learning structures in layers to create an ”artificial neural network” that can learn and make intelligent decisions on its own.
Can train on lesser training data.	Requires large datasets for training.
Takes less time to train.	Takes longer time to train.
Trains on CPU.	Trains on GPU for proper training.
The output is in numerical form for classification and scoring applications.	The output can be in any from including free from elements such as free text and sound.
limited tuning capability for hyperparameter tuning.	can be tuned in various ways.

Table 2.2: Machine Learning Vs Deep Learning.

2.10 Definition of CAPTCHA

CAPTCHA—or ”Automated Testing of Computers and Humans Apart”—is a security feature that can prevent unlawful use of websites. It is a sort of test on websites where the user must place or click certain parts of a picture in a number of letters and numbers in order to complete a task. This task can be to buy a product, open an account, post a review or any other task. This is used by web developers on sites to ensure that people with good intentions are indeed people using the site [41].

2.11 History of CAPTCHAs

1996	Moni Noar suggested using an automated Turing test to distinguish between human users and robots [27].
1997	Andre Brod et al. A mechanism was developed to distinguish between human users and computer programs. In the same year, the Altavista website used this method to prevent robot programs from entering by showing the user a distorted English word and asking the user to copy it [27].
2000	The concept CAPTCHA was coined by a team at Carnegie Mellon University led by Manuel Blum and Luis von Ahn [27].
2002	Broder announced that the CAPTCHA system has been implemented for more than a year, reducing the number of spam advertising URLs by more than 95% [27],.
2003	The Baffle text CAPTCHA was created by Californians Barid and Monica Chew [28].
2004	the Yahoo website utilized a simple version of the EZ-Gimpy method [27].

Table 2.3: History of CAPTCHAs.

2.12 Application of CAPTCHA

1- Protecting Website Registration:
CAPTCHAs protect multiple free email services such as Yahoo, Gmail and Hotmail from bot programs that register thousands of email accounts with an automated script every minute.

2- Protecting Email Addresses from Scrapers:
This can be accomplished by requiring a user to overcome a CAPTCHA before showing his or her email address, thereby shielding the user's email address from Web scrapers.

3- Online Polls:
CAPTCHAs are often used to prevent Web crawlers and bots from participating in online voting by requiring the voter to overcome a CAPTCHA before submitting their ballot. This method, however, cannot deter users from voting multiple times.

4- Preventing Dictionary Attacks:
After a series of failed logins, the user is asked to complete a CAPTCHA test to prevent computer programs from repeating through the entire password room. This mechanism is preferable to locking an account after a certain number of failed login attempts.

5- Search Engine Bots:
Administrators can use CAPTCHA to block search engines from indexing to prevent others from downloading or reading these sites, because sometimes they contain private information.

2.13 Classification Types

Verification code are used as Turing tests for users or humans, and they are classified based on distorted images, text and numbers. Some types of verification codes are as follows:

1. **CAPTCHA's based on Text.**
2. **CAPTCHA's based on Images.**
3. **CAPTCHA's based on Audio.**
4. **CAPTCHA's based on Video.**
5. **CAPTCHA's based on Puzzle.**

2.13.1 CAPTCHA's based on Text

Text-based CAPTCHAs became the first method of human verification. These CAPTCHAs may be made up of well-known words or phrases, as well as random digits and letters. Some text-based CAPTCHAs also include variations in capitalization. The CAPTCHA introduces these characters in an alienating and enigmatic manner that necessitates understanding. Scaling, rotation, and distorting characters are also examples of alienation. Overlapping characters with visual features such as colour, background noise, shapes, arcs, or dots may also be used. This alienation protects humans from bots with inadequate text recognition algorithms, but it can be difficult to understand for humans [42].

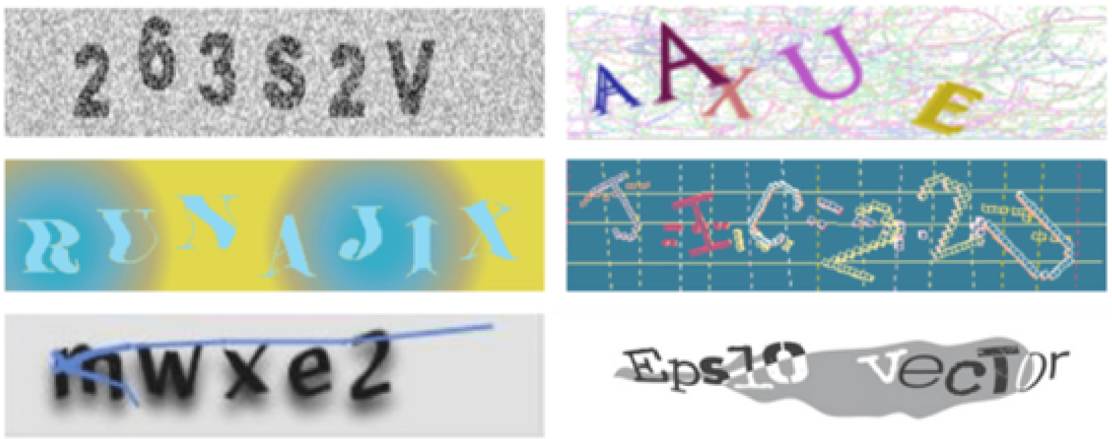


Figure 2.6: CAPTCHA based on Text.

Techniques for creating Text-based verification codes include :
- **Gimpy-Choose** any number of words from a dictionary of 850 words and provide these words in a distorted manner.
- **EZ-Gimpy** — is a variant of Gimpy that uses only one word.
- **Gimpy-r-select random letters**, then deform and add background noise to the characters.
- **Simard's HIP-select random letters and numbers**, and then use arcs and colors to deform the characters.

2.13.2 CAPTCHA's based on Images

Text-based CAPTCHAs have been phased out in favor of image-based CAPTCHAs. These CAPTCHAs use graphical features that are easily recognized, such as animal portraits, patterns, or scenes. Image-based CAPTCHAs usually enable users to pick images that complement a pattern or mark images that don't. You can see an example of this type of CAPTCHA below. Note that it defines the theme using an image instead of text. Humans typically have an easier time deciphering image-based CAPTCHAs than text-based CAPTCHAs. These methods, on the other hand, have significant compatibility problems for visually disabled people. Picture-based CAPTCHAs are more difficult for bots to decipher than text-based CAPTCHAs since they require both image recognition and semantic classification [42].

Figure 2.7: CAPTCHA based on Image.

2.13.3 CAPTCHA's based on Audio

Audio CAPTCHAs were created as an option that allows visually disabled people to participate. These CAPTCHAs are often used in conjunction with CAPTCHAs that are built on text or images. Audio CAPTCHAs play a transcription of a sequence of letters or numbers, which the recipient would then type in.
The bots are unable to differentiate relevant characters from background noise in these CAPTCHAs. These technologies, including text-based CAPTCHAs, can be difficult for humans as well as bots to decipher [42].

Figure 2.8: CAPTCHA based on Audio.

2.13.4 CAPTCHA's based on Video

Instead of photographs or text, these CAPTCHAs use videos. Users are shown You Tube videos and asked to tag descriptive keywords in one such CAPTCHA. Users will be asked to watch a competition video and then annotate (or tag) it correctly. The challenges will be graded based on the user's answer being perfectly matched to a database of video ground reality tags. Fig.2.9 depicts an example of a video-based CAPTCHA method [26].

Figure 2.9: CAPTCHA based Video.

2.13.5 CAPTCHA's based on Puzzle

In a puzzle-based CAPTCHA, a picture (image) is divided into small fragments. A user's job is to put these segments together to shape the original image, as seen in Fig.2.10. This CAPTCHA is appealing to the user to solve. People with low intelligence and vision, on the other hand, can find it difficult to rearrange these segments in their proper positions.

Figure 2.10: CAPTCHA based Puzzle.

2.14 Comparison between CAPTCHAs types

2.14.1 Strength and weakness of text based CAPTCHA

Text based CAPTCHA is the simplest type of CAPTCHA where it is the first type which has been invented and implemented in Email services and search engine. Text based CAPTCHA consists of English letters and numbers.These characters are restricted, so robot programs and hackers can solve text verification codes by designing programs that scan text verification codes and type text in specific locations. This problem can be solved by making some modifications to the characters (such as adding some noise or rotating and distracting letters or displaying the characters as damaged and deformed letters or introducing the characters as 3D). These modifications will cause some problems for the user when the user recognizes the correct character, because some characters have a similar shape after the modification, additionally. Another problem with text-based verification codes are English, where some users cannot understand certain text-based verification codes, such as clickable verification codes and strangeness verification codes in sentences [29], [30].

2.14.2 Strength and weakness of CAPTCHA based on image

This type of image will introduce some similarity images to the user according to the questions under the CAPTCHA to select a suitable image. Although this type is simple, users may encounter some problems when trying to solve image-based CAPTCHAs:
1. Some users who have low vision or learning disability will meet some issues when they are attempting to solve this CAPTCHA
2. The likelihood of a bot machine breaking the CAPTCHA increases as the amount of options in the CAPTCHA decreases, so it is best to raise the number of options in the CAPTCHA to make it more robust, although this process would drain the database.
3. This CAPTCHA has been available just in English language, therefore English speakers and some others who have English vocabulary knowledge will solve the CAPTCHA only [29], [30].

2.14.3 Strength and weakness of CAPTCHA based on audio

This form of CAPTCHA is intended for people with visual impairments, in which the user is presented with registered words and must type the word that he or she has heard.
1. Although audio CAPTCHA is available for visually disabled users, there are some issues may face the users.
2. Adding noise to the reported words strengthens the CAPTCHA and prevents it from being broken by Bot systems, however, it can mislead the user, resulting in a wrong response.
3. Audio CAPTCHA is introduced by English language, therefor just users with English ability can solve this type.
4. In the English language, certain letters sound alike, such as J G, C K. This could be

perplexing to the user [29], [30].

2.14.4 Strength and weakness of CAPTCHA based on video

CAPTCHA based on video involves showing a short film of an individual doing an operation, and the user must choose the right explanation from a list. Since the video is too big in this process, users may have trouble retrieving it from the internet. This issue may cause users to abandon the website or email they were trying to use. Another thing that can concern certain users who do not speak English is that video CAPTCHA is only available in English [29], [30].

2.14.5 Strength and weakness of CAPTCHA based on puzzle

CAPTCHA is a puzzle that involves adding pieces of images and asking the user to combine them or describe a certain part of the image. Since this process takes longer to overcome, the user will become bored and leave the website. Additionally, people with poor vision would have difficulty identifying images [29], [30].

2.15 conclusion

We have devoted this chapter to basic concepts such as AI and its evolution, machine learning and its types and the most popular and most used algorithms and then comparing machine learning to deep learning. In the second section, we the presentation of concepts and history of CAPTCHAs, and discussed their applications. In addition, to an overview of CAPTCHA types (text, images, audio, video and puzzle). And, We conclude by discussing the strength and weakness of each category.

Chapter 3

Related works

3.1 Introduction

We have become familiar with the concept of CAPTCHA and its types and the study of ML and its algorithms.
In this chapter, we present a comprehensive review of recent developments in the field of CAPTCHA of all kinds and techniques, Many scientific articles supporting the processing, recognition and breakage of several types of text-based or image-based CAPTCHA have been summarized by ML and DL studied, their success rate in order to choose the path we will follow in the next chapter.

3.2 Related works

3.2.1 A Generative Adversarial Learning Framework for Breaking Text-Based CAPTCHA in the Dark Web:

Automated monitoring of dark web platforms (e.g., Dark Net Markets and carding shops) on a broad scale is required for cyber threat intelligence (CTI). While there are techniques for getting data from the surface web, anti-crawling tactics frequently obstruct large-scale dark web data acquisition. The text-based CAPTCHA is the most restrictive of these methods. The user must recognize a combination of difficult-to-read letters in a text-based CAPTCHA. To prevent automated CAPTCHA breaching, dark web CAPTCHA patterns are purposefully built with more background noise and varied character length. Existing CAPTCHA-breaking solutions can't solve these problems, thus they're useless on the dark web. Due to the especially loud backdrop and variable character length of CAPTCHA patterns, text-based CAPTCHA cracking has been a key bottleneck for large-scale dark web CTI monitoring. Leveraging GAN and CNN, In their research, they propose a novel framework to crack the text-based CAPTCHA in the dark web. The proposed framework utilizes Generative Adversarial Networks (GAN) to counteract dark web-specific background noise, and utilizes enhanced character segmentation algorithms. Compared with the state-of-the-art baseline methods, they checked the overall CAPTCHA cracking performance of their proposed framework. This experiment attempts to measure performance under specific conditions of the dark web that lacks labeled CAPTCHA images for training. Therefore, we extract 500 images from each dark web and benchmark CAPTCHA data source as the evaluation test platform. They compared the performance of the CAPTCHA cracking method against the benchmark CAPTCHA and dark web CAPTCHA modes. CAPATCHA cracking performance is measured by the success rate, which is often used for evaluation in previous CAPTCHA research. In particular, the success rate calculates the percentage of CAPTCHA patterns successfully resolved by each method. A successful attempt to crack a CAPTCHA is defined as the correct recognition of "all" characters displayed in the CAPTCHA pattern. In addition, three state-of-the-art methods are used as baselines.
- Image-level CNN only.
- Image-level CNN with pre-processing (grayscale conversion, normalization, and Gaussian smoothing).

- Character-level CNN with interval-based segmentation.

On both benchmark and dark web CAPTCHA testbeds, their suggested technique was examined. On all datasets, the suggested technique beats state-of-the-art baseline approaches, with a success rate of over 92.08% on dark web testbeds. their study enables the CTI community to create powerful dark web monitoring tools on a big scale [10].

3.2.2 Recognizing Objects in Adversarial Clutter:

The concept behind the CAPTCHA program stems from the real world problems faced by Internet companies such as Yahoo and AltaVista. Yahoo provides free email accounts. The target users are humans, but Yahoo found that various online pornographic companies and other companies are using ”bots” to register thousands of email accounts every minute, from which they can send spam. The solution is to require users to resolve the CAPTCHA test before receiving the account. The program selects a word from the dictionary and generates a distorted and noisy image of that word. The user sees the image and is asked to enter the words that appear in the image. Given the type of deformation used, most people succeeded in this test, while current procedures (including OCR procedures) failed the test. Achieved the purpose of screening ”robots”. The Manuel Blum team at CMU has designed many different CAPTCHAs. In the words of Ahn, Blum and Langford:Any software that passes the tests given by a CAPTCHA can be utilized to solve a challenging unsolved AI issue. Their hope is thus to provide challenge problems for AI researchers. If the problem cannot be solved by computer programs, it can be used as a CAPTCHA and provide practical help to companies like Yahoo, AltaVist, etc. If it can be solved, that marks scientific progress on a hard AI problem. It is strongly recommended that you visit http://www.CAPTCHA.net to see some of these examples. Gimpy is based on word recognition under cluttered conditions. The task is to recognize 3 out of 7 words in a cluttered image. The Yahoo test is a simpler version, the so-called EZ-Gimpy, which involves only one word. Pix is based on image recognition in generally more abstract categories. Chew and Baird also developed a test involving severely degraded fonts. CAPTCHA is also designed based on auditory recognition. For people with impaired vision, these may be the most appropriate tests. The basic principle behind the CAPTCHA design is to reduce to hard AI problems. As visual researchers, they decided to accept the challenge of defeating EZ-Gimpy and Gimpy.

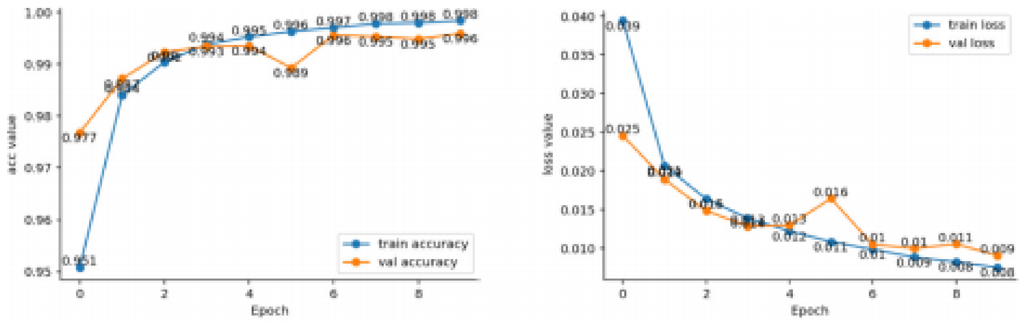
Their research describes an algorithm to achieve this goal. On EZ-Gimpy, their program correctly identified the word 92% of the time, which means that Yahoo can no longer use it to filter ”bots.” Their 33% success rate on Gimpy also makes it ineffective as such a screening tool. These two data sets provide more than just a color toy problem. In their paper, they explored two main aspects of object recognition: difficult clutter, and the trade-off between partial and whole object recognition. The CAPTCHAs we use are currently challenging chaos because they are designed to be difficult for computer programs to handle. The recognition of words can easily be understood as the recognition of parts (single letters or two-tuples) or entire objects (whole words). More importantly, these data sets are large. There are about 600 words to be recognized. In addition, since the source code for generating these CAPTCHAs is available (”P” stands for the public), they can access an almost unlimited set of test images to use. This is in sharp contrast to many object recognition data sets, where the number of objects is limited, and it is difficult to generate many reasonable test images. In terms of studying general object recognition, this data set must have limitations. Most notably, these are 2D objects and will not change due to 3D poses. In addition, there are no shadows and lighting effects in the text composite image. They believe that the ability to perform quantitative experiments with a large number of disorganized test images outweighs these shortcomings. They hope other researchers will try their techniques on these and other CAPTCHAs. Many computer vision researchers are dedicated to solving object recognition problems. Among these methods, the candidate most likely to succeed in this field is (LeCun et al) and (Amit et al.) (LeCun et al.) Use the convolutional neural network for handwritten character recognition. (Amit et al.) Use point features combined with graphics to match various deformable objects. These technologies are somewhat robust to clutter, but it is not obvious whether they can handle the ”adversarial” clutter found in Gimpy images. They have presented algorithms that

Figure 3.1: Results on Gimpy.The 3 words that are guessed using Algorithm B are shown below each image.

Are instances of a broader framework that may be applied to solve various recognition issues [14].

3.2.3 Neural CAPTCHA networks:

In order to prevent attacks from malicious computer programs, many websites use CAPTCHA (abbreviation for fully automated Public Turing Test to distinguish between computers and humans) technology for security protection. The distortion, rotation, and deformation of characters or puzzles in CAPTCHA increase the difficulty for the machine to automatically recognize them. The most advanced CAPTCHA recognition algorithms usually use Convolutional Neural Networks (CNN), regardless of the spatial order properties of character/image features. To solve this problem, in their paper, they proposed a verification code recognition model called Neural Verification Code Networks (NCNs). NCN uses a convoluted structure to extract image features, and uses a bidirectional loop module to learn spatial order information. Through NCN, we have processed text-based CAPTCHA, including arithmetic operations, character recognition and character matching CAPTCHA, and puzzle-based CAPTCHA. For arithmetic operations, character recognition, and puzzle-based CAPTCHA, they use NCN with CTC loss to predict the probability of letters, numbers, and slider coordinates, and output the correct results. For character matching verification codes, we built a CNN to automatically split the first part into four individual characters. Then, they input a pair of characters (one from the first part and the other from the second part) into the weight-sharing NCN. By comparing the loss, NCNs determines whether the two characters match, and finally outputs the matching orders. A large number of experiments have proved that NCN is superior to the most advanced CAPTCHA recognition method, because NCN not only learns the effective features of CAPTCHA images like previous CNN models, but also converts the spatial sequence information of character/image features in the CAPTCHA into accounts. They achieved 100% accuracy on the SOIEC CAPTCHA dataset. For character matching tasks, we achieved an accuracy of 99.3% on the SOIEC CAPTCHA dataset. For the puzzle-based CAPTCHA, we achieved an accuracy of 98.13%. These experimental results demonstrate the advantages of NCN over the related state-of-the-art CAPTCHA identification methods. Characters match the result of CAPTCHA [8].



(a) The accuracy curve (b) The loss curve

Figure 3.2: The accuracy and loss curves obtained by NCNs on the validation set of the character matching CAPTCHA task.

In the future, in order to improve the performance of NCNs, they plan to modify the convolution part of NCNs into stacked residual blocks or dense blocks. In addition, they integrated the attention mechanism into the two-way loop module of NCN to solve more complex and challenging CAPTCHA identification problems.

3.2.4 Text-based CAPTCHA Strengths and Weaknesses:

They confirmed through in-depth research that many popular websites still rely on schemes that are vulnerable to automated attacks. For example, their automated DeCAPTCHA tool broke the Wikipedia plan about 25% of the time, as shown in the picture. Currently, 13 of the 15 most widely used schemes are also vulnerable to automated attacks by its tools. Therefore, it is clear that a comprehensive set of design and testing principles is needed to generate a more powerful verification code. Although previous work [24] showed that the security of the verification code depends on preventing segmentation, they found in their research that only relying on segmentation does not provide reliable automatic attack defense. For example, the fact that the length of the verification code is fixed can be used to guess where to split the verification code, even if its reverse split technology cannot be directly destroyed.
They discovered that this sort of attack may be used against a variety of CAPTCHA systems, including eBay and Baidu. Pre-processing, segmentation, post-segmentation, recognition, and post-processing are the five generic processes in the automated CAPTCHA-solving process, according to them. While segmentation, or the division of a sequence of characters into individual characters, and recognition, or the identification of those characters, are straightforward and well-understood, there are compelling reasons to include pre-processing and post-processing phases in a typical procedure. Pre-processing, for example, can remove background patterns or other picture enhancements that may interfere with segmentation, and post-segmentation procedures can "clean up" the segmentation output by normalizing the size of each picture or completing processes other than segmentation. Following recognition, post-processing can help enhance accuracy by adding spell checking to any CAPTCHA that is based on real words, for example (such as Slashdot). They experimented with numerous particular algorithms based on this basic CAPTCHA solving architecture and tested them on a variety of popular website CAPTCHAs. They discovered a collection of approaches that make CAPTCHAs more difficult to solve mechanically using this corpus. They were able to construct a bigger synthetic corpus by modifying these tactics, which allowed them to analyze the effects of each of these traits in more depth and develop their automated assault methods. Hey focused their attention on a spectrum of strategies that are within the grasp of human solvers, while they did explore hypothetical CAPTCHAs that are uncomfortably tough for certain individuals, based on their earlier analysis of how solvable CAPTCHAs are for humans.
They used genuine CAPTCHAs from Authorize, Baidu, Blizzard, Captcha.net, CNN, Digg, eBay, Google, Megaupload, NIH, ReCAPTCHA, Reddit, Skyrock, Slashdot, and Wikipedia to evaluate the effectiveness of our technology DeCAPTCHA. Prior to this study, none of these CAPTCHA schemes has been reported broken as far as we know. They had a 1%-10% success rate of two (Baidu, Skyrock), 10%-24% on two (CNN, Digg), 25%-49% percent on four (eBay, Reddit, Slashdot, Wikipedia), and 50% or more on five (eBay, Reddit, Slashdot, Wikipedia) (Authorize, Blizzard, CAPTCHA.net, Megaupload, NIH). They created the first effective assaults against CAPTCHA techniques that employ collapsed characters to attain such a high success rate (eBay, Megaupload, and Baidu). Only Google and ReCAPTCHA stood firm against our attacks, and they eventually figured out why we couldn't crack them. They were able to crack 7 of the 15 schemes (Authorize, Baidu, CNN, Megaupload, NIH, Reddit, and Wikipedia) without having to write

a new algorithm thanks to DeCAPTCHA genericity. We derived numerous criteria and suggestions from our examination of real-world and synthetic CAPTCHAs that we feel will be valuable to CAPTCHA designers and attackers. Randomizing CAPTCHA length and individual relative character size, for example, are crucial measures in preventing automated attacks, despite being quite pleasant for humans. Similarly, if all characters are the same size, partial segmentation can help the segmentation by providing a decent approximation of the number of characters. Creating a wave form and collapsing or overlaying lines, on the other hand, can be pretty successful. They also uncover that complicated character sets, which can be perplexing to humans, are ineffective, and we discuss the relative relevance of anti-recognition approaches, implementation faults, and developing alternative "backup" methods in the event that vulnerabilities are identified. This work's major contributions include [13]:

- DeCAPTCHA is a generalized evaluation tool for swiftly assessing CAPTCHA security.
- A current assessment of anti-recognition and anti-segmentation algorithms, as well as CAPTCHAs employed by prominent websites. A single program successfully attacked 13 out of 15 genuine CAPTCHA schemes from famous websites, as well as the first successful assaults on CAPTCHAs that employ collapsed characters (e.g. eBay and Baidu).
- A publicly accessible synthetic corpus meant to emulate the security aspects of real-world CAPTCHAs in human-acceptable ranges, so that designers may test new attack methods on them.
- A defensive taxonomy, as well as an assessment of the influence of anti-recognition strategies on the automated learnability of CAPTCHAs.

3.2.5 Cyber Security Using Arabic CAPTCHA Scheme :

Bots are programs that trawl through a website and register users automatically CAPTCHAs, which are written in Latin script, are extensively employed to prevent automated bots from misusing web services. Many of the existing English-based CAPTCHAs, on the other hand, contain flaws and cannot guarantee the security of these websites, (Bilal Khan et al.) offer the article "Cyber Security Using Arabic CAPTCHA Scheme" To produce a picture, the suggested system employs Arabic script. The image is corrupted by the addition of several forms of background sounds in the shape of dots, lines, and arcs. The background and foreground colors are chosen to make the CAPTCHA picture appealing to the user as a whole Because of the different amount of characters, font kinds, and font sizes, the OCR has a difficult time reading our CAPTCHA. The algorithm is effective, and the user has no difficulties when working with the system. Participants in the poll came from Arabic-speaking nations as well as non-Arabic South Asian countries that could read Arabic script. The overall reading rate of the photos was determined to be high in the study. As a result, this suggested CAPTCHA technique may be utilized in non-Arabic speaking nations where languages such as Urdu, Pashto, and Persian employ the Arabic script. The results were encouraging mentioned in the table 3.1. Furthermore, a comparison of their CAPTCHA and Persian CAPTCHA schemes reveals that our CAPTCHA scheme outperforms Persian CAPTCHA [12].

Figure 3.3: Arabic CAPTCHA Scheme Example.

Features	Arabic Scheme	Persian Scheme
Number of Font Types	50	1
Variation of Font Types	Randomly	n/a
Variation of Font Size	Randomly	n/a
Lines (as Background Noise)	Randomly	Randomly
Dots (as Background Noise)	Randomly Persian Scheme	n/a
Use of Dictionary Words	n/a	n/a
Number of Letters	5 to 9	3 to 9
Baseline Detection by OCR	Not Possible	Possible
Text Coordinates	Varies randomly	Static (center)
Background and Foreground Color	Different	Same

Table 3.1: Comparison of features of Arabic schemes (ours) and persian CAPTCHA scheme.

3.2.6 Empirical studies to investigate the usability of text- and image-based CAPTCHAs[1]:

This article provides a report on two empirical studies that explore the usability of image and text based CAPTCHA's to authenticate web users. First, four images-based CAPTCHAs, Asirra, ESP-PIX, SQ-PIX and IMAGINATION were examined, with twenty-participants interacting on the desktop with five of each type. The participant then completed the CSUQ, the NASA Task Load Index (NASA-TLX) and a preference-based final subjection test questionnaire for the CAPTCHAs. The results indicated statistically significant differences between CAPTCHAs for all dependent variables, with participants preferring Asirra and ESP-PIX, except for completion time. The second compared the usability of both of these with ReCAPTCHA and Google's CAPTCHA, the two most commonly employed text based CAPTCHA. After training on Asirra and the ESP-PIX, twenty-four participants completed 10 trials on the desktop computer, followed by a retrospective think-aloud session, followed by a CSUQ and NASA-TLX. The results of this second study showed that Google took the least time to complete CAPTCHA and ReCAPTCHA. Asirra and ReCAPTCHA had less workload than the workload of ESP-PIX and CAPTCHA from Google. For Asirra and ReCAPTCHA, perceived overall usability was often highest. Comments from the participants indicated that the CAPTCHA of Google was too visually distorting, and ESP-PIX had commonalities between the images, which were not immediately apparent. In all, because of its high perception of usability, low completion times and low workload, ReCAPTCHA is recommended for use on a desktop computer. Repeated one-way ANOVA measures with 95% confidence interval have been used to determine the existence of significant differences across various CAPTCHAs. Post-hoc LSD comparisons have been used to identify the locuses of significant differences. This study employed a within-subject experimental design with CAPTCHA type at 4 levels (Asirra, ESP-PIX, CAPTCHA, and ReCAPTCHA). All participants were exposed to two types of text-based CAPTCHAs and two types of image-based CAPTCHAs: 1) Microsoft Asirra. 2) ESP-PIX. 3) Google's CAPTCHA, a text-based CAPTCHA featuring a visually. distorted nonsense word that the user must correctly interpret and type into a response box as shown in Figure (3.4) ReCAPTCHA,another type of text-based CAPTCHA, in which the user must identify two visually distorted meaningful words by typing them into a response box. In addition to restricting access to web bots, ReCAPTCHA is also a book digitization project (von Ahn et al., 2008) as shown in Figure 3.5.

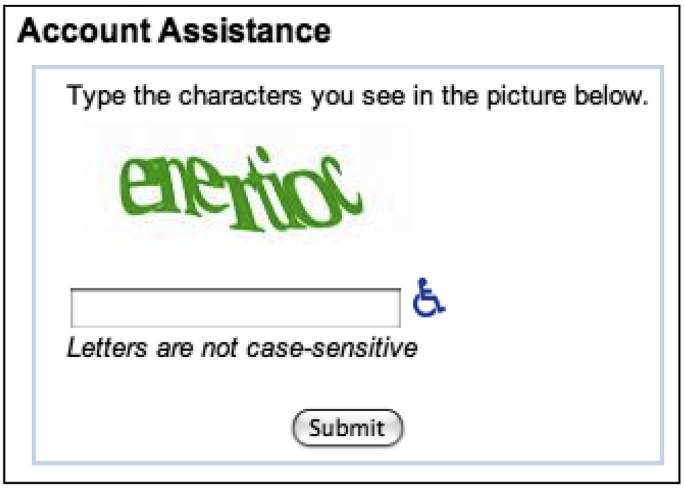


Figure 3.4: Google's CAPTCHA. A text-based CAPTCHA requiring the user to interpret a visually distorted nonsense word and type it into a response box.

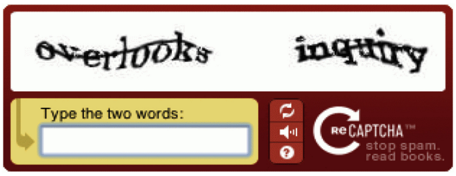


Figure 3.5: ReCAPTCHA. A text-based CAPTCHA and book digitization project that requires the user to interpret two visually distorted meaningful words and type them into a response box.

The Friedman test has been used to identify significant effects for this metric, due to the non-parametric nature of the ranked preference data. ReCAPTCHA text-based, high-perception usability, low completion time, low error rate, and low workload for desktop computers is recommended for use. Future work is needed to improve the usability of current CAPTCHAs and to develop CAPTCHAs that balance security and usability. In a recent paper, Golle (2008) claimed to have written a program that can break the Asirra CAPTCHA. This suggests the difficulty inherent in developing a CAPTCHA that is simultaneously easy to complete by humans and difficult to break with automation.

3.2.7 Usability study of text-based CAPTCHAs :

In this study, the effects of age groups and distortion types on the CAPTCHA job were analyzed in an experiment focused on text-based CAPTCHA. The experiments attracted 25 participants, of whom 12 were in the elderly (aged 50 to 56) and 12 in the young group (ages 23 to 24). The experiment was carried out by 12 participants. There was a single type without distortion and 5 types of common CAPTCHA distortion techniques. The results of non-parametric analyzes showed a substantial difference in response time, error rate and NASA-TLX score between the participants in two age groups. The type of distortion significantly influenced response time, error rate, critical flicker fusion (CFF), and NASA-TLX score.

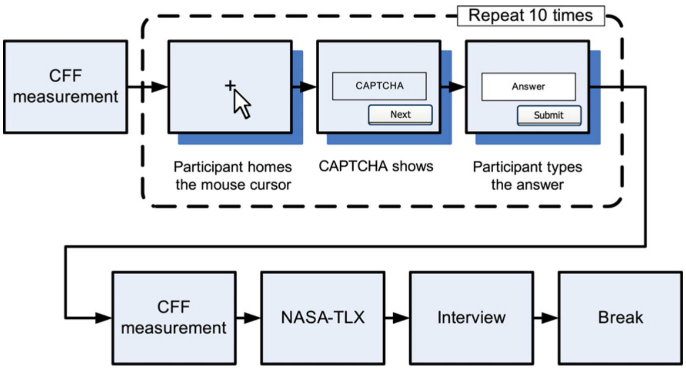


Figure 3.6: Experiment procedure of a session. Rectangles with shadows represent computer screens.

A posterior analysis showed the hardest CAPTCHas were the blot masks and the line masks while the noise was equal to Normal Type Thread, Global Warp, and Geometry Noise (no distortion). Also correlated were the dependent variables. This study has important implications for the design of CAPTCHA systems, given the inevitability of the security measure and the growing population of elderly Internet users. This study suggests that the type and age of CAPTCHA would affect the task, visual system and cognitive system performance. The main objective of this research was therefore to compare the effect of different techniques of CAPTCHA distortion on the visual and cognitive systems of human ages. The study used the experiment to gather objective and subjective data, including task performance, visual fatiguer, workload, and post-session interviews, from participants of two ages (young people and older people). A comparison of the usability issues of the CAPTCHA distortion techniques with and between age groups was made. The separation of the two CAPTCHA groups in this study has a useful influence on future designs of CAPTCHA. As discussed, the researchers found that when the occluded objects are visible, it is easier to recognize them rather than when they are not. This is because the occluders are interpreted as foreground objects that make fragments more visible in the background. Blot Mask had occluders that were not cohesive objects in this study, while Line Mask had occluders that were not visible. However, there were occlusers visible both in Thread Noise and in Geometry Noise. And the results of the current study have shown that the two previous types are harder than the other two. CAPTCHAs are suggested to use consistent, visible noise to challenge machines, while remaining friendly to the human cognition system. While this study focused on text-based CAPTCHAs, it is also worthwhile to explore other types of CAPTCHA like auditive or imagery. Also important issues are gender differences, curves of learning, and the interaction between the native speaker and CAPTCHA character set. In this direction further studies can be carried out [3].

3.2.8 On the security of text-based 3D CAPTCHAs :

CAPTCHAs are now a standard safety mechanism for the deterrence of automated abuse of human-led online services. To date, however, many existing CAPTCHA schemes have been breached successfully. Therefore, several CAPTCHA developers have explored other CAPTCHA design methods. 3D CAPTCHAs is a design solution proposed in order to overcome traditional CAPTCHAs limitations. These CAPTCHAs are designed to capitalize on the natural ability of the human visual system to picture 3D objects. A computer program finds it hard to identify the 3D content as an underlying assumption of security. Their paper examines the solidity of 3D CAPTCHAs based on text. In particular, three existing 3D CAPTCHA text-based systems currently being implemented on a variety of websites are examined. While it is impossible to correctly solve these 3D text-based challenges of using the Optical Character Recognition (OCR) software, They can use certain patterns in the 3D CAPTCHAs to identify important information with the CAPTCHA. This paper shows that automatic assaults can be used to resolve these 3D CAPTCHA with great success by extracting this information. This is a 3D CAPTCHA on text, now used on a number of websites (Cafe Milenium, 2009; Cafe Charlotte, 2006; EnterMir, 2013; SACHYonline. cz, 2010; HyperMedia, 2012). Whether this CAPTCHA scheme has an actual name or not is not known. They are therefore simply referring to '3dCAPTCHA' as this CAPTCHA scheme. 3D CAPTCHA Two examples:

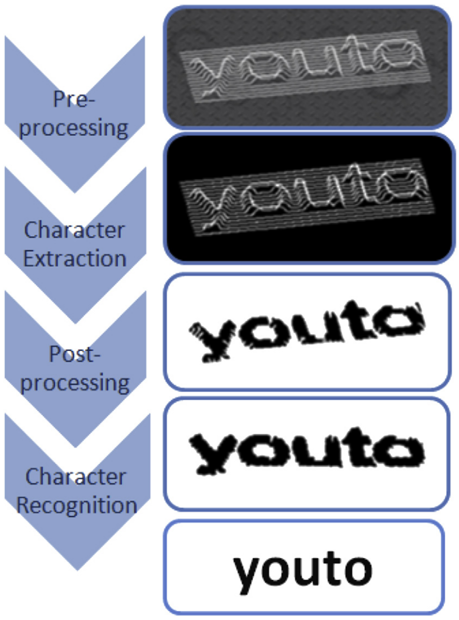
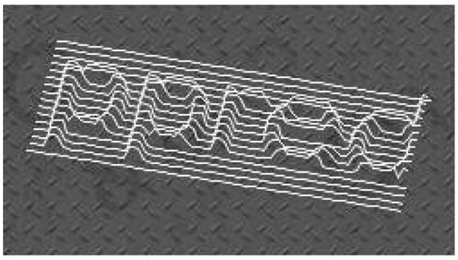
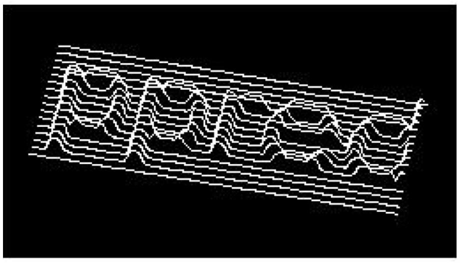


Figure 3.7: Overview of the stages.



(a) Original image



(b) After background removal and binarization

Figure 3.8: Pre-processing stage in attacking Super CAPTCHA.

They examine the robustness against automated attacks of text-based 3D CAPTCHAs. It is possible to obtain information on the CAPTCHA challenge by identifying key features in the construction of the 3D CAPTCHA system. This study has described a number of methods for doing so. Common methods include identifying background variations such as line direction changes, lines distance, pixel density, cell sizes, etc. It shows that while the concept of text-based 3D CAPTCHA schemes might seem safe, since OCR programs may not directly resolve the 3D CAPTCHA, it is possible to extract information from CAPTCHA first and then transfer it into the OCR programme. The study shows that, through automatic attacks, 3D CAPTCHAs can be resolved with a high level of success. Those results appear to be neglecting the assumption that 3D CAPTCHAs are actually safer than their traditional 2D counterparts, while CAPTCHA developers continue to seek new CAPTCHA alternatives [4].

3.2.9 Breaking text-based CAPTCHAs with variable word and character orientation :

A new approach is presented in this article in order to automatically separate and recognize CAPTCHAs with variable orientation and random collapse of covered characters. Furthermore, there is a discussion of the extension of the proposed break-out approach to 2012 reCAPTCHA. The original proposal includes enhancing the characters and words in the CAPTCHA, which then use a three-color bar code to segment them. The proposed original SVM-based learning classifier recognizes the enhanced characters and the entire word. The primary purpose of this study was to reduce CAPTCHA's vulnerability to spam and fraud, and to provide an approach to manuscript or manuscript recognition in OCR systems. The proposed approach to the breakdown of CAPTCHA tested the success of the average segmentation rate to 82% for 2011 reCAPTCHA and reaches 95.5% for 2012 reCAPTCHA by extended approach, with response time below 0.5 s per

2-word reCAPTCHA. The SVM classifier implemented shows a competitive accuracy of 94%. The very satisfying outcomes obtained confirm that the approach proposed can be used to develop new security mechanisms to protect users against Internet threats and cyber-criminal activities. Another demand for CAPTCHA automatic punching systems is the high speed recognition that is useful for applications in real time that are not always recorded in well-known approaches. These CAPTCHA beating schemes apply the following steps: pre-processing to remove background noise and confusion, segmentation of the picture in individual regions and character recognition. The most difficult task is segmentation, although it is also a challenge to develop a fast and robust classifier. In their study, They propose the division into the following phases of the CAPTCHA division process: image acquisition, pre-processing, segmentation and recognition. The step sequence is shown in Figure (3.9) Detailed description of the steps proposed is provided in the following sections.

On the stage, They use a morphological filter to detect unconnected image areas with a CAPTCHA word. They provide a sequence of steps for the breakdown of CAPTCHA for facilitating segmentation.

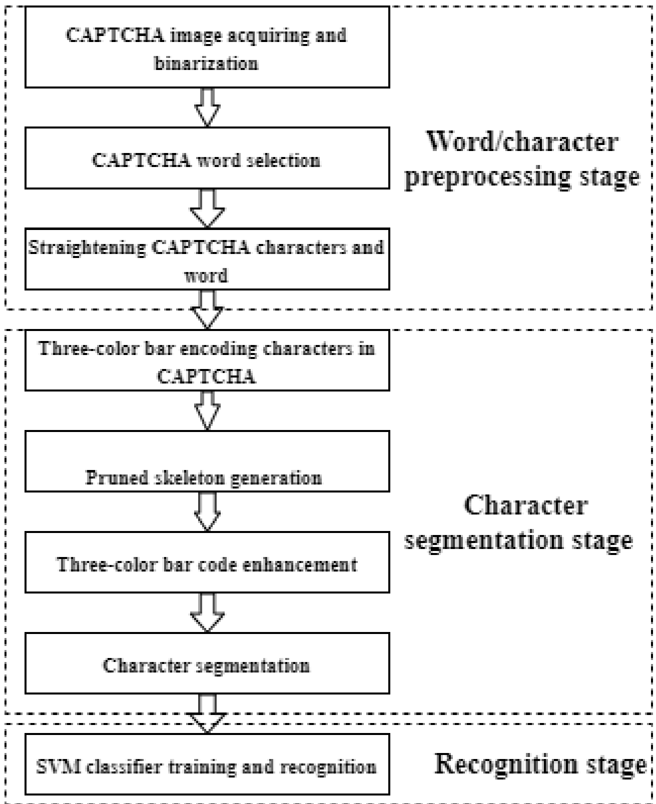


Figure 3.9: Sequence of steps for CAPTCHA breaking process.

For the reCAPTCHA 2011 approach, the segmented success rate is about 31% (for 60 single-character classes in training groups) and about 82%.(for 231 classes that includes also labeling double characters). Expanded reCAPTCHAs of 2012 are about 56.3%(for single characters alone in corpus) and 95.5% more precisely segmentation (for corpus with double characters). The time necessary to break reCAPTCHA by extensive approach is four times less than in the 2011 release (less than 0.5 s per two-word CAPTCHA). The SVM classifier implemented has practical accuracy of recognition of about 94% with an average break time for CAPTCHA. The obtained very satisfactory results confirm that the proposed approach may be considered as robust techniques for CAPTCHA improvement as well as a new way of developing systems able to protect users against cyber-criminal activities and Internet threats [2].

3.2.10 End-to-end attack on text-based CAPTCHAs based on cycle-consistent generative adversarial network :

To date, many researchers have conducted studies in various companies such as Microsoft, Amazon and Apple on approaches to attack text-based CAPTCHAS and have obtained specific results. Most attacks, however, have deficiencies such as poor portability of attack methods which involve a series of preprocessing steps and depend on large quantities of CAPTCHA labels. In this study, they propose an efficient method of end-to-end attack based on cycle coherent networks of generative opponents (Cycle-GANs). Its approach significantly reduces data labeling costs in comparison with previous studies. Furthermore, the high portability of this method. The attack can only be made easier by modifying several parameters of ordinary CAPTCHA text-based schemes. The first way to generate false samples is to train CAPTCHA synthesizers, based on a cycle going. Based on a convolutionary recurrent neural network, basic recognizers are trained with fake information. An active transfer learning method is then employed to optimize the fundamental recognizer by using small amounts of labeled CAPTCHA samples from the real world. Their approach effectively ruptured the 10 popular websites of the CAPTCHA schemes that show that our attack procedure may be universal. In addition, the latest anti-recognition mechanisms have been analyzed.The consequences show that the combination of more mechanisms against recognition can improve the safety of CAPTCHAs. The upgrade is limited, however. Instead, the more complex CAPTCHA generation can cost more resources and reduce the CAPTCHA usability. Multiple fonts, difficult background interference, rotation of characters, overlapping, distortion, two-layered structures, noisy arcs are the popular means of protection. These include These complicated security mechanisms force an attacker to invest significant work in pre-processing images such as the separation of character and denouncing images. They also develop methods which require a large volume of training data based on deep learning. Tens of 000 samples must generally be labeled in order to reach high success rates of attack. In addition, various measures to protect CAPTCHAs from being maliciously downloaded by attackers are being added increasingly to websites, thus making it more challenging to collect real-world CAPTCHA samples than before. In their study, a generic, effective and cost-effective method is proposed in order to break text-based CAPTCHA. It is divided into two parts: CAPTCHA and CAPTCHA. First, they used the cycle-compatible CAPTCHA synthesizer (Cycle-GAN) as the basis of their reliability in real-world data. The CAPTCHA Recognizers were subsequently trained based on a CRNN. Subsequence: Finally, they tested an attack on the most popular text CAPTCHA websites, such as Baidu, Wikipedia, Microsoft and eBay, which deployed the top 50 websites. The experimental results demonstrate that their method can achieve not only higher attack success rates, but also significantly reduce labor costs during the attack. In this study, the Commission proposed a text-based breaking method based on GANs, based on end-to-end text. The cycle-GAN-based synthesizers have been trained to generate a large number of synthetic CAPTCHA samples, resolving the inadequate training data issue. In order to improve the pre-trained profound learning models, they also proposed a model training method based on the active transfer learning, using small amounts of hard examples with more information. A large number of real world CAPTCHAs have been cracked, as rankings of Alexa.com, and used by the top 50 websites. The rate of success ranged from 33.8% to 97.6%. their approach requires less labeled data from 100 to 500 samples, and achieves the same or more successful attack rates, compared to previous studies. Also a better versatility of their method. Without expensive work and time, a text-based CAPTCHA system can be broken quickly by their method [7].

No	Sample	User Friendly	User Unfriendly	Success Rate
1	1F6XD5	80.0%	4.0%	97.4%
2	[illegible]	16.0%	46.0%	95.1%
3	[illegible]	16.0%	40.0%	96.1%
4	AGR3Q6	70.0%	4.0%	94.9%
5	[illegible]	28.0%	46.0%	97.5%
6	XSF5F	96.0%	2.0%	96.0%
7	KE14NW	96.0%	4.6%	96.6%
8	BME X6K	60.0%	4.0%	94.4%
9	[illegible]	14.0%	56.0%	80.0%
10	[illegible]	2.0%	86.0%	62.1%
11	[illegible]	0.0%	90.0%	47.4%
12	[illegible]	0.0%	96.0%	10.6%

Figure 3.10: User-friendliness test of 12 CAPTCHA security mechanisms.

Furthermore, 12 popular anti-recognition mechanisms were analyzed for safety. The test results show that the safety of the CAPTCHAs can be improved with more safety mechanisms. Nonetheless, these security mechanisms can no longer be useful if attackers receive a large number of training data using different methods. The current CAPTCHA text-based schemes have been shown to be uncertain and cannot withstand attacks on the machine and in-depth learning. Finally, they proposed proposals for future CAPTCHAs development on the basis of the weaknesses of the existing CAPTCHA schemes. They hope that this study will provide a reference for security experts to design CAPTCHA schemes featuring both security and usability.

3.2.11 Machine Learning Attacks Against the Asirra CAPTCHA :

The Asirra CAPTCHA is based on the challenge of differentiating photographs of cats and dogs (Asirra stands for ”Animal Species Image Recognition for Restricting Access”). It was first suggested at ACM CCS 2007. An Asirra challenge is made up to 12 photos, each of which depicts a cat or a dog. The user must choose all of the cat photos and none of the dog photos in order to solve the CAPTCHA. This is a job that humans excel at. According to Asirra, “people can do it 99.6% of the time in under 30 seconds.” When contrasted to CAPTCHAs based on identifying distorted sequences of letters and numbers, Asirra’s usability is a huge advantage. Asirra’s security is predicated on the difficulty of automatically identifying photos of cats and dogs. Cats and dogs are extremely difficult to discern apart algorithmically, according to data from the 2006 PASCAL Visual Object Classes Challenge. The accuracy of a classifier based on color features is just 56.9%. According to the authors of [25], “classification accuracy of greater than 60% will be challenging without a considerable increase in the state of the art,” based on an assessment of machine vision literature and vision specialists at Microsoft Research. The likelihood of completing a 12-image Asirra task with a 60% accurate classifier is just approximately 0.2%. They describe a classifier in their research that is 82.7% accurate in distinguishing between the photos of cats and dogs utilized in Asirra. They can use this classifier to answer a 12-image Asirra challenge with a probability of 10.3%. This success probability is far greater than the 0.2% estimate for machine vision assaults stated in [25]. While their success rate may look modest in absolute terms, it represents a severe danger to Asirra unless further measures are implemented to prevent machine adversaries from requesting and attempting to solve (at no cost) an excessive number of CAPTCHAs. Their classifier is made up of two support-vector machine (SVM) classifiers that have been trained on picture color and texture data. The classifier is completely automated, with the exception of the one-time labeling of training photos. Their classifier is 82.7% accurate using 15,760 color features and 5,000 texture characteristics per picture. The classifier was trained using 13,000 tagged photos of cats and dogs downloaded from the Asirra website on a standard PC. They also look at how their assaults can affect the partial credit and token bucket methods presented in [25]. Asirra is significantly weakened by the partial credit mechanism, and they advise against using it. The token bucket algorithm helps mitigate the impact of their attacks and allows Asirra to be deployed in a way that maintains an appealing balance between usability and security. Their study has helped to determine the selection of safety criteria in Asirra deployments. Beyond this immediate benefit, they believe that their research will help to popularize machine learning techniques in the security sector, for both offensive and defensive objectives. Machine learning and other artificial intelligence approaches haven’t been frequently employed in cryptography assaults until recently. However, current research reveals that these approaches are effective tools in the armory of cryptanalysts. SAT solvers have been used to identify collisions in hash functions and to thwart authentication methods, for example. The text-based Gimpy and EZ-Gimpy CAPTCHAs were successfully broken using object recognition techniques. Similarly, machine learning classifiers of the sort utilized in this study are likely to find wider uses in computer security [15].

3.2.12 The End is Nigh: Generic Solving of Text-based CAPTCHAs:

It has become well-established that a CAPTCHA's ability to withstand automated solving lies in the difficulty of segmenting the image into individual characters. The traditional method for solving CAPTCHAs automatically has been a sequential procedure in which a segmentation algorithm divides the picture into segments containing individual characters, followed by a machine learning-based character recognition phase. While this strategy has proved successful against specific CAPTCHA schemes, the segmentation stage, which is handcrafted to overcome the distortion at hand, limits its universality. The character collapsing anti-segmentation approach utilized by most popular real-world CAPTCHA methods has no known generic algorithm. Their study proposes a revolutionary way to solve CAPTCHAs in a single step that combines segmentation and recognition difficulties using machine learning. When both operations are done at the same time, their method can take use of information and context that would not be accessible if they were done separately. At the same time, it eliminates the requirement for any handcrafted components, allowing them to extend their methodology to new CAPTCHA schemes that the prior method could not. They were able to correctly solve all of the real-world CAPTCHA schemes we evaluated, with a recognition rate of 38.68% on Baidu 2011, 55.22% on Baidu 2013, 51.09% on CNN, 51.39% on eBay, 22.67% on ReCAPTCHA 2011, 22.34% on ReCAPTCHA 2013, 28.29% on Wikipedia, and 5.3% on Wikipedia. Their algorithm's ability to tackle a wide range of distortions demonstrates that it is a generic method for solving CAPTCHAs automatically. They also offer ideas on how reverse Turing tests should be improved in the future, based on their experience. The efficacy and universality of their results imply that integrating segmentation and recognition is the next step in CAPTCHA solution, and that it will eventually supplant the sequential technique utilized in previous studies. With these advancements, it appears that merely text-based CAPTCHAs may lose their use; major work may be required to rethink how reverse Turing tests are performed [11].

3.2.13 Segmentation-validation based handwritten Arabic CAPTCHA generation :

They describe a unique methodology for generating and certifying handwritten Arabic CAPTCHAs in their research. Arabic CAPTCHAs have lately piqued the interest of academics because of the increasing Internet penetration among Arabic speakers. CAPTCHAs based on handwritten text have been found to be more secure than CAPTCHAs based on printed text. As a result, the suggested approach generates CAPTCHAs using synthetic handwritten Arabic phrases. Users overcome CAPTCHAs by detecting segmentation points in cursive Arabic text during the user-validation phase of their approach.

The usage of synthetic words for handwritten Arabic CAPTCHAs is introduced in their work. The suggested approach is versatile enough to construct CAPTCHAs of varying difficulty levels. The following is how their CAPTCHA generating mechanism works. They begin by developing an Arabic handwriting synthesis system. His technique generates ground truth labeled synthetic Arabic Part of Arabic Words (PAWs). They can produce an endless number of Arabic PAWs in this method, which may be utilized to produce CAPTCHAs. To strengthen security, these synthetic PAWs are warped in the second phase utilizing multiple Arabic letter level operations and picture level noise. Finally, the user is invited to solve the CAPTCHA by using a digital pen or fingers to identify the segmentation points (also known as joining points of consecutive characters) in the PAW, or by clicking with a mouse. The suggested CAPTCHA method is illustrated with a simple example. As a result, their study makes a significant contribution to the field of CAPTCHA in two ways:

- The use of synthesized handwritten words for Arabic CAPTCHAs, which is reported for the first time in their work (to the best of their knowledge).
- For security and validation, Arabic script level characteristics are used.There are two parts to this. To begin, we use a variety of OCR processes to make the CAPTCHAs more secure.
- Second, user validation is carried out by the user detecting segmentation spots in the displayed CAPTCHAs. OCRs have a hard time breaking both of these elements, but a human user can readily answer the CAPTCHAs by leveraging context and script knowledge. To their knowledge, no previous research has employed segmentation-validation to solve CAPTCHAs. It's worth noting that, even if a user picks a segmentation point correctly (according to script rules), authenticating the user's pick is difficult due to the way Arabic letters join in ”valleys”. Although a word synthesis system will always unite two letters at the same position, the user may not always choose that place. Due to inconsistencies in touching the screen, the user nearly never selects the same connection points utilized during synthesis. As a result, allowing the user to choose any location in the connecting valley between two letters is a difficult operation, making segmentation-validation a non-trivial process. Experimental experiments were conducted to test the suggested CAPTCHA technique for both usability and security. The results reveal that, as compared to a machine method, the suggested system can be solved significantly more effectively by human users [6].

3.2.14 Breaking Visual CAPTCHAs with Naïve Pattern Recognition Algorithms :

Visual CAPTCHAs have been widely employed to protect against unwanted or dangerous bot programs on the Internet. They show how we defeated most of the visual schemes supplied by CAPTCHAservice.org, a publicly available online service for CAPTCHA production, in their study. Attacks employing a high-quality Optical Character Recognition software were effective against these techniques. They used fatal design flaws to create simple attacks that could break four visual CAPTCHA schemes (including one visual component of a multimodal scheme) provided by CAPTCHAservice.org with near 100% success–these schemes all used sophisticated distortions, were effective against OCR software attacks, and appeared to be secure. It's concerning that they discovered that many other visual CAPTCHAs used on the Internet contained identical flaws, making them vulnerable to easy assaults. The following are the main causes they believe can explain the degree of failure. Although a few pioneering initiatives have provided some insight on how to create visual CAPTCHAs, their cumulative grasp of the subject is still in its early stages. There were a few design standards and rules of thumb dispersed throughout the literature, but there are many more to be found. There is also no systematic way of determining if a CAPTCHA is truly resilient.Otherwise, the catastrophic design flaws described in this research may have been prevented with relative ease. As a result, their paper calls for more research into the design of practically secure and robust CAPTCHA schemes, a relatively new but important topic, and in particular into developing both a robustness evaluation method and a comprehensive set of design guidelines – the latter of which can include things like what should be included or avoided in CAPTCHA design, as well as what could be used. Many of today's CAPTCHAs are likely to create a false impression of security for the same reason. They believe that breaking representative schemes in a systematic way will produce persuasive evidence and, as proven in their study, significant insights that will aid in the design of the next generation of robust and useful CAPTCHAs [9].

3.2.15 Chinese Character CAPTCHA Recognition and performance estimation via deep neural network :

Many web apps are increasingly using the Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) to distinguish between machine and human. CAPTCHAs based on traditional English and digital characters are recognized with great accuracy. Because of the complexity of Chinese characters, automated recognition is extremely challenging. Because of intricate strokes with distortion, rotation, and background noise that are not easily identified by human developed feature detectors such as SIFT, color, and texture, Chinese Character CAPTCHAs are more challenging than English CAPTCHAs, as seen in the figure 3.11. Even Hanvon, the most well-known Chinese handwritten Character OCR software, is unable to solve the problem. As a result, they offer a CNN-based approach to handle the Chinese Character CAPTCHAs Recognition challenge in their work. There are two reasons to use CNN to recognize Chinese characters in CAPTCHAs. For starters, because there are so many various forms of distortion, rotation, and background noise, designing numerous universal human-designed feature detectors is not feasible. Fortunately, CNN is programmed to learn appropriate visual characteristics automatically, allowing it to deal with this issue. Second, several studies have demonstrated that CNN is effective in character recognition, and that a deep convolutional neural network is effective at capturing character structure. The following are the primary contributions of their article:

- They verify that the total training time of our approach is a constant for a given accuracy in the condition of enough and representative training samples.
- They demonstrate that the network structure of Chinese Character CAPTCHA Recognition is more complex and deeper than that of English and Digital CAPTCHA Recognition.
- They demonstrate that their network topology outperforms LENET-5 at this task.

They also develop a relationship between accuracy, the amount of training samples, and iterations, which they use to measure the performance of their method. To begin with, this method considerably enhances the detection accuracy of Chinese Character CAPTCHAs

that have distortion, rotation, or background noise.
The results of their research demonstrate that this method achieves over 95% accuracy for single Chinese characters and 84% accuracy for three varieties of Chinese Character CAPTCHAs with four Chinese characters. Finally, they confirm that their methodology recognizes Chinese Character CAPTCHAs better than the most well-known Chinese Optical Character Recognition (OCR) program, Hanvon [5].

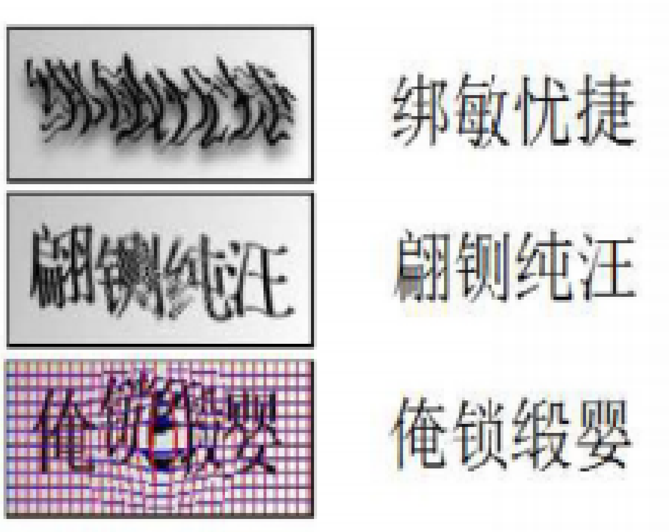


Figure 3.11: Three types of Chinese Character CAPTCHAs with answers (There are 3D shadow, Water ripple and Fish eye CAPTCHAs from top to bottom).

3.3 conclution

After reviewing the previous works and examining the success of each proposed idea and knowing the pros and cons of the various types of CAPTCHA, we will highlight the text-based CAPTCHA in order to identify and break, by exploiting the weaknesses that have been matched in the design of CAPTCHA some sites.

Chapter 4

CAPTCHAs PROCESSING

4.1 Introduction

The ability of hackers to sequence computer systems using computer software and robots led to the development of CAPTCHA Text-based CAPTCHA is captcha's most common scheme, an automated test that humans can pass, but computer software can't.
In this chapter, we present our planning for a simple, general and rapid attack on CAPTCHA based on text for several sites using machine learning techniques, by exploiting the flaws of website design to break this type of CAPTCHA by training a range of algorithms for testing.
This chapter is organized as follows: focus on the most commonly used CAPTCHA types,

in the following part how to break and attack by adopting pre-processing, hashing and recognition methods in order to make captcha's text-based image as simple as possible, in the last phase our model was developed for design.

4.2 CAPTCHA Generation

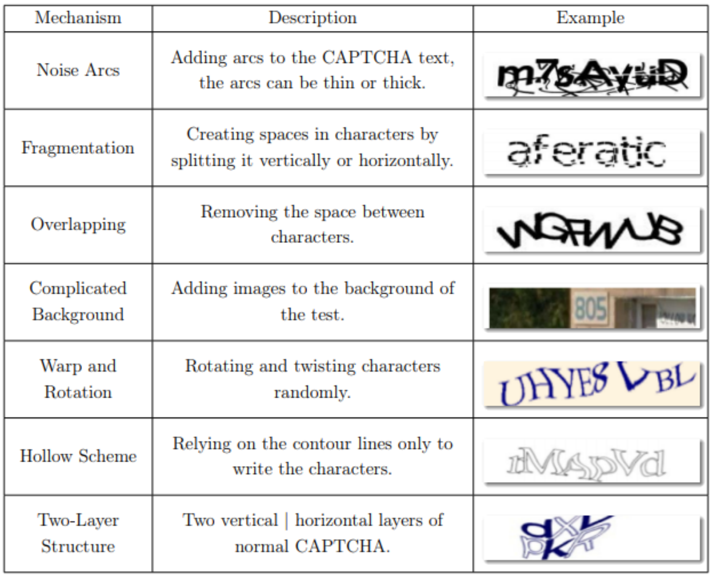
We concentrate on the most used and popular types of CAPTCHAs in this segment of our work due to the many types of CAPTCHAs: Text-based CAPTCHA

4.2.1 Text-based CAPTCHA Generation

While there is no universal algorithm or procedure for creating text-based CAPTCHAs, it is possible to identify a set of steps that have been followed the majority of the time and which can be tweaked marginal or supplemented with additional steps [15]. These are the steps:
1. Creating an image that contains the CAPTCHA text with the appropriate dimensions Is always the first move .
2. Make the image's backdrop. Depending on the resistance mechanism used, the image's context may be changed (distorted or simple background).
3. To generate CAPTCHA text, characters or words are picked at random from a previously chosen range based on those characteristics (often the difficulty to recognize by OCR softwares). The majority use English alphabets and Arabic numbers, but this does not preclude the presence of text in other languages like Arabic, Chinese, or Indian, for example.
4. Choose the font style, scale, and colour, as well as a range of distortions. Additionally, the better the CAPTCHA text would be the more fonts used each time. Then, on the previously generated image, write the CAPTCHA code.
5. Finally, the image is given resistance mechanisms to make it more stable and unbreakable.

4.2.2 Resistance Mechanisms

One of the most critical characteristics of the CAPTCHA is how difficult it is for computers to solve. It has been improved with the following techniques in order to do this:
Segmentation Resistance Text-based CAPTCHAs often use segmentation resistance techniques, in which the characters that make up the test are paired with a series of distortions to make them impossible to divide by OCR programs while still being simple to solve by humans [16], [17], [18]. Any of these mechanisms are described in the table below:



Mechanism	Description	Example
Noise Arcs	Adding arcs to the CAPTCHA text, the arcs can be thin or thick.	
Fragmentation	Creating spaces in characters by splitting it vertically or horizontally.	
Overlapping	Removing the space between characters.	
Complicated Background	Adding images to the background of the test.	
Warp and Rotation	Rotating and twisting characters randomly.	
Hollow Scheme	Relying on the contour lines only to write the characters.	
Two-Layer Structure	Two vertical | horizontal layers of normal CAPTCHA.	

Figure 4.1: Segmentation Resistances.

Resistance to Recognition Following the dilemma of identifying the letters (in the right order), there is the issue of understanding them. A collection of methods is used to make the recognition more complicated [17],[18].
- Use a different font: use a new font each time.
- Make the size, angle, width and position of the characters different.
- Make it difficult to differentiate between characters by measuring their pixels by ensuring that all characters have the same number of pixels.

Using random strings instead of words: by using words that aren't native to the language but instead generated at random
- Use a large database: Use a large number of characters that make up the test.

Other Mechanisms Many other methods are used to improve a CAPTCHA's resistance, such as restricting the test duration or number of attempts, exposing the IP addresses that have incorrect responses, using checks that involve human intervention, typically in the form of a game, using 3D images or constructs, and so on. All these resistances are added, whether for segmentation or recognition in the context of improving the robustness of the CAPTCHA, which makes it a challenge.

4.3 Text-based CAPTCHA Breaking

The text-based CAPTCHA is the most prevalent type of CAPTCHA that has been attacked or disabled, but most of the breakage techniques were programmed to target this type using conventional approaches before machine learning and deep learning techniques were created.

We must go through the steps seen in Figure 4.2 in order to crack any text-based CAPTCHA. The most crucial step is identified, while the rest of the steps can be called supporting steps [17], [19].

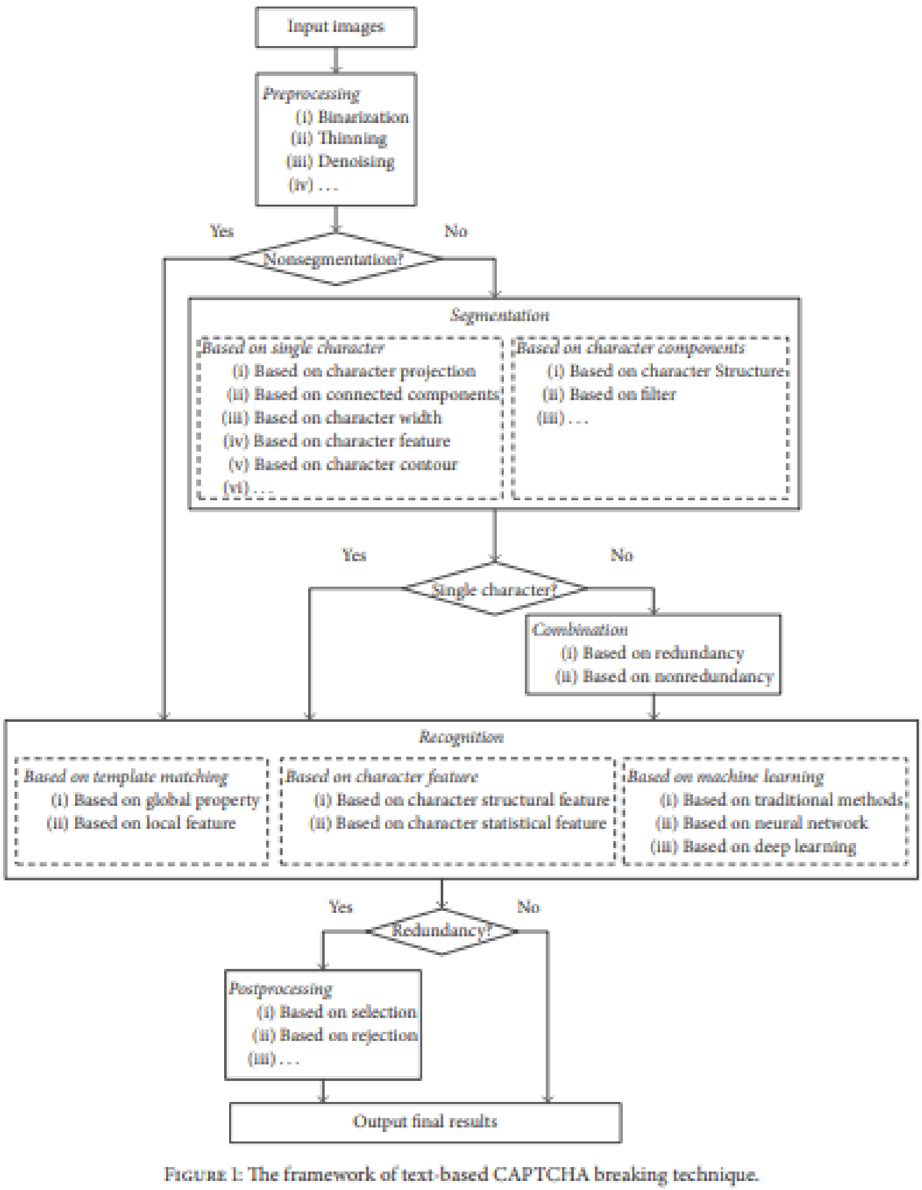


Figure 4.2: The framework of text-based CAPTCHA breaking technique.

4.3.1 Pre-processing

This move can be thought of as a warm-up for the next two measures (segmentation or recognition). The aim is to make the text-based CAPTCHA image as simple as possible using the procedures below:

• Image Binarization: By transforming the image to black and white using the threshold [19], Sauvola and Pietikainen's [20], or Otsu's [21], this technique is used to eliminate blurred backdrop and noise from the image.
• Image Thinning: It is considered to be one of the most prominent methods for analyzing any image pre-processing. It also plays an important role in image analysis and pattern recognition. Refinement involves reducing thicker objects in the image to thin skeletons [22]. The ZS algorithm [23] is used for this purpose.
• Image Denoising: A group of arcs and lines is often added to support the CAPTCHA test against breakage, so a variety of dance techniques are used to remove this noise. Filters, Gibbs and Hough transformations, and Wavelet transforms are all included [19].

4.3.2 Segmentation CAPTCHA

Attackers often use segmentation before attempting to break the recognition step; segmentation is one of the most significant supporting measures, particularly when CAPTCHA is assisted by segmentation resistance (overlapping, hollow, etc.). However, there are several methods for cracking the CAPTCHA that does not rely on segmentation [19]. As a result, there are two groups:

Text-Based CAPTCHA Breaking Methods Based on Segmentation

These methods use segmentation as the basic step to break CAPTCHA. A set of techniques and methods are used for this recommendation. Its purpose is to obtain each character, according to its segmentation resistance, from CAPTCHA to another character will be different. From these methods, we mentioned:

• Methods based on single (individual) characters: using a variety of techniques, fragment a CAPTCHA image into individual characters: – Character projection:A good way to segment non-overlapping characters. The character projection histogram confirms this (the projection can be vertical and horizontal).
– Character width: A limited method used only when the characters are equal in width and evenly distributed.
– Character feature: Segmentation of characters based on inside and outside characteristics such as (dot I j), circle (a, b, d, e, g, o, p, q), cross (t, f)).
– Character contour: Analyzing geometric characteristics of character contours determines the necessary segmentation lines.
• Methods based on character components: Instead of producing individual characters, these methods produce multiple character components based on:

– Character Structure: color filling is often used to discover shared character elements dependent on structural features of the characters.

– Filter: Use Log-Gabor filter to extract character components, which is suitable for various text-based CAPTCHAs.
Combination
Contrary to directly recognize individually segmented characters. Characterize components need to be combined for later recognition, depending on two technologies [19]:
– Combination methods based on redundancy: Since using the filter in the segmentation, this procedure is used. This is accomplished by integrating the filter's output to produce characters.
– Combination methods Based on non-redundancy: Create a complete character by using the boundaries of the overlapping area (shared character components).

Text-Based CAPTCHA Breaking Methods Based on Non-segmentation
With the increase in the use of segmentation resistors in text-based CAPTCHAs. What they rely on is to rely on the method of direct recognition, without the need to segment by using a set of technologies, which is what we will discuss in the next section.

4.4 Recognition Methods of Breaking Text-Based CAPTCHA

Nowadays, the recognition methods used in text-based CAPTCHA system include three categories: template matching, character feature, and machine learning.

4.4.1 Recognition Methods Based on Template Matching

Template matching is to compare the similarity of each pixel between the character and each template, and find the highest similarity. According to the matching range, there are matching recognition methods based on global attributes and matching recognition methods based on local features.

- **Matching Recognition Methods Based on Global Property:** The matching recognition method based on global attributes traverses scanning. In the search area, the best matching point with each pixel is found through regional correlation matching calculation. Because many templates that match each pixel are very slow, proposed a second template matching algorithm to improve efficiency. Only after the roughest match is successful, an exact match is required.

- **Matching Recognition Methods Based on Local Feature:** The shape context is a simple local feature shape descriptor. The basic idea is to transform the image matching problem into a feature point set matching problem. In 2003, Mori and Malik used shape context to break the verification codes of Gimpy and EZ-Gimpy. Provide good robustness for images.

4.4.2 Recognition Methods Based on Character Feature

We may describe various methods depending on the feature characters, which is mostly based a character, structural feature and character statistical feature, since the character of each CAPTCHA function differs in nature.

- **Recognition Methods Based on Character Feature:** The structural feature can describe the details and structural information about the characters, such as the number of loops, inflection point, convexo-concave degree, and cross points. For example to crack Yahoo CAPTCHA, uses character guidance (see Figure 4.3(a)) and closed loop detection (see Figure 4.3(b)).

Figure 4.3: the example of character structural features

- **Recognition Methods Based on Character Statistical Feature:** The recognition method based on character statistical feature uses commonly statistical features including pixel feature, projection feature, contour feature, and coarse mesh feature. This feature is robust to noise interference and is widely used in the CAPTCHA recognition field. Reference used the distinct pixel count for each of the letters A to Z (see Figure 4.4) to break captchaservice.org CAPTCHA with a near 100% success rate.

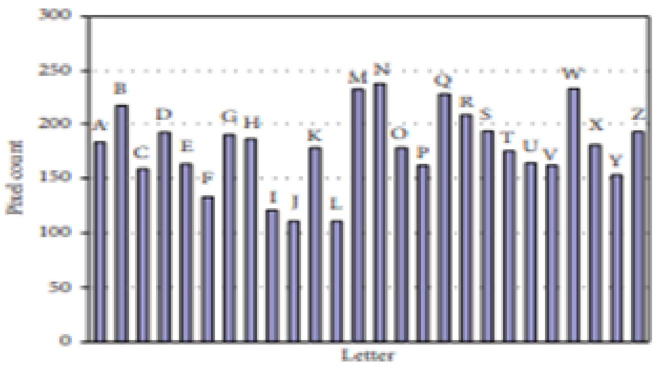


Figure 4.4: the pixel count for each of the letters A to Z

- **Recognition Methods Based on Machine Learning:**
CAPTCHA tests have often been backed up by recognition resistances that render them unbreakable, particularly when using the above-mentioned standard methods. As a result, with the complexity of machine learning algorithms, one of the

problems that machine learning can overcome is breaking CAPTCHA by correctly classifying CAPTCHA characters [19].As previously said, recognizing the characters of CAPTCHA using machine learning is a classification problem, and the most popular methods used for this are Support Vector Machine (SVM), k-nearest neighbor (KNN), and Artificial Neural Network (ANN) (ANN).

After the pre-possessing and segmentation phases, the authors of [31], break reCAPTCHA 2011 and 2012 versions (shown in Figure 4.5). Support Vector Machine (SVM), a supervised machine learning algorithm used in classification, was used to solve the recognition step as a classification problem.SVM uses kernel functions to separate classes by hyperplane. They trained 1000 subdivided reCAPTCHA images. The model achieved a success rate of 31% in the reCAPTCHA 2011 version and 56.3% in the reCAPTCHA 2012 version. Double characters (nn, vv, mm, etc.) and double characters formed by thin double characters (il, li, ii, ll, it, ti, 11, etc.) are the reasons for this result.Therefore, they created other classes for these combinations. The addition of these classes greatly improves the accuracy of the classification process. They achieved an 82% success rate in the reCAPTCHA 2011 version and a 95.5% success rate in the reCHAPTCHA 2012 version.

Figure 4.5: reCAPTCHA 2011 and 2012 versions.

In [32], the authors use two models of recognition, k-nearest neighbor (KNN) and convolutional neural network (CNN), to attack CAPTCHAs used by the top 10 most popular websites (Google, Microsoft, Yahoo!). KNN reported higher success rates.

K Means was used to solve the Recognition phase in [33]as a clustering challenge. Three separate data sets were used to test this model (shown in Figure 4.6). With 100 videos, the Finance Agency and the Umweltpramie have a 70%and 68% performance rate, respectively. Sparda Bank has 330 pictures and an 87% success rate.

(a) Umweltpramie (b) Finanzagentur (c) Sparda Bank

Figure 4.6: Attacking text-CAPTCHAs.

Bostik and Klecka [34],compared four different classification algorithms for recognizing CAPTCHA characters. ANN, SVM, KNN, and Decision Tree are all examples of artificial neural networks. A PHP script created the CAPTCHA dataset that they used. The highest performance rate was 100% for ANN, 98.80% for SVM, 98.99% for KNN, and 98.99% for Decision tree.

Through these works, we found that the ML method has better results than the standard method, but it is still limited in terms of extracting the features and segmentation on which the recognition model is based.

4.5 Global Architecture:

We can describe our architecture as the folowing:

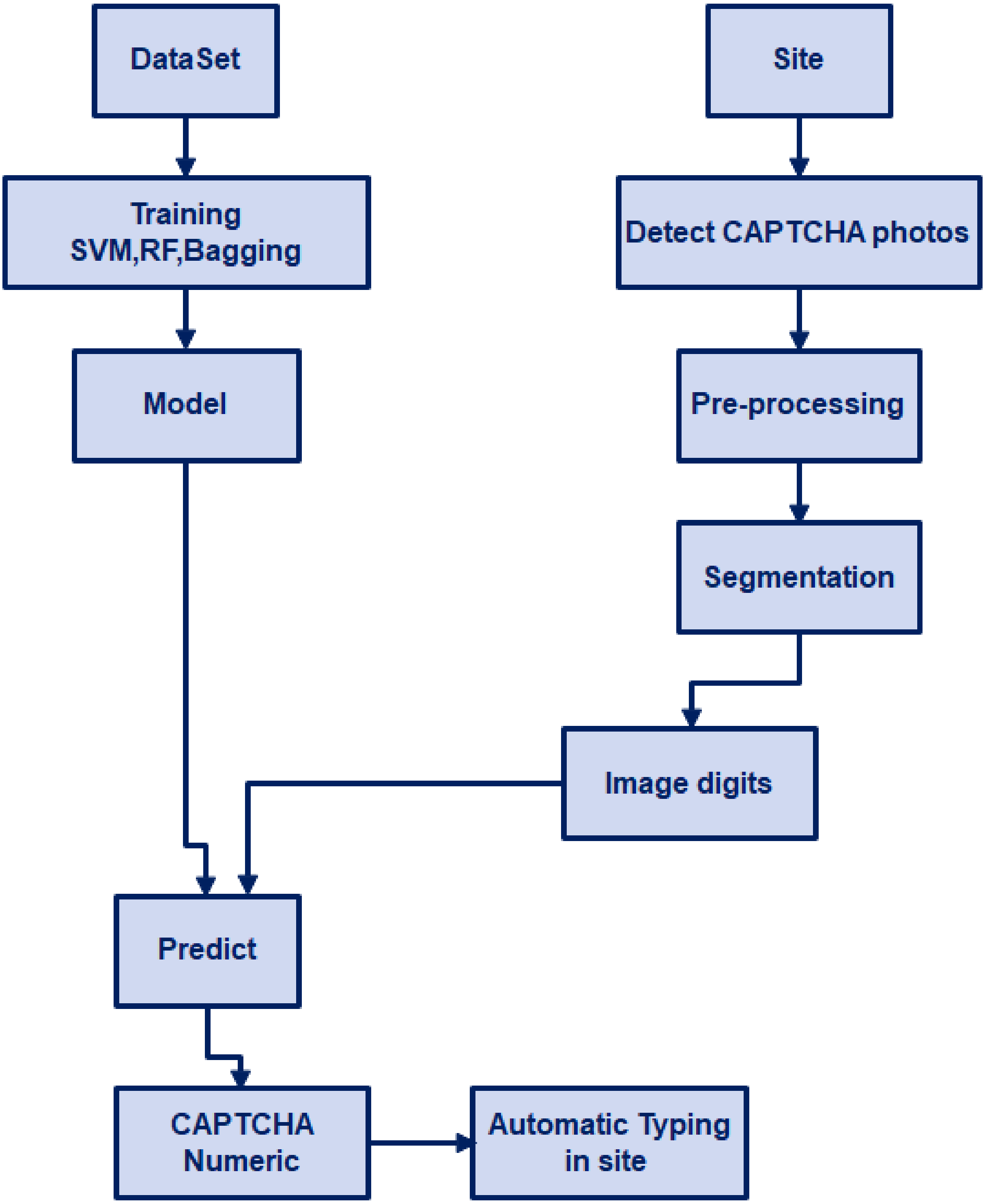


Figure 4.7: The global architecture of the CAPTCHA breaking process.

The most important operations of our Global Architecture:
Pre-processing :
CAPYCHA pre-processing is the first step in CAPTCHA image processing before segmentation and recognition The goal is to filter and remove noise and other interventions Pre-processing of current CAPTCHA-breaking methods mainly involves binary image encoding.
Image binarization :
Image encoding is the process of converting a digital image into a binary image. The binary image has only two possible values per pixel (black/white). It has been empirically proven that image recognition technologies work better and faster on binary features. So we encode a black and white CAPTCHA.
Segmentation :
Segmentation is very important. The most common method of segmentation analysis is to split a CAPTCHA image into multiple individual parts of one character. Segmentation is used to detect a large number of blurred CAPTCHA images containing lines and curves. Taking into account rotation, noise and even skewed characters. In the end, we get a clear picture free of blemishes

4.6 Detailed Architecture

In the first step: 7 Algorithms (SVM,RF,DT,Bagging, K-NN,Logitic Regression,Gaussian) are trained by a data set and all algorithms are tested and their results studied based on the accuracy of each model, the best of which are selected.

Figure 4.8: Choose the best trained algorithm.

In the second step: We choose the site whose CAPTCHA will be broken, and then we download the CAPTCHA image, but we have a problem, which is not being able to download this image, so we put a screenshot of the CAPTCHA.

Figure 4.9: Download the CAPTCHA image from the site to break its CAPTCHA.

In the third step: The image obtained from the website, we convert it into a binary image (black/white). This technique is used to remove the blurry background and noise from the image, so we get a processed image at the end of this stage.

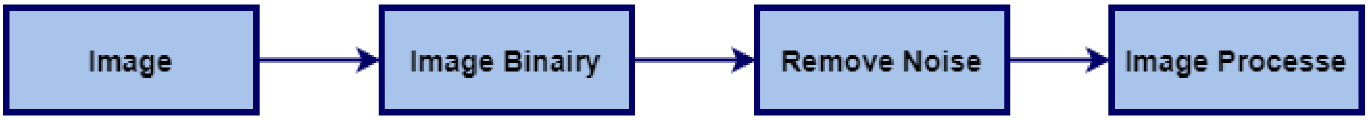


Figure 4.10: Processing image .

In the fourth step: The group of pixels connected to each other and that have the same color is selected in order to divide the basic image into separate images containing only one character by the process of vertical projection or contour selection.

Figure 4.11: segmentation the image into a separate images.

In the fifth step: Each image is checked in order to predict the content of each separate images of the letters(A-Z) and numbers(0-9) classes, and in the end it is sent to the site.

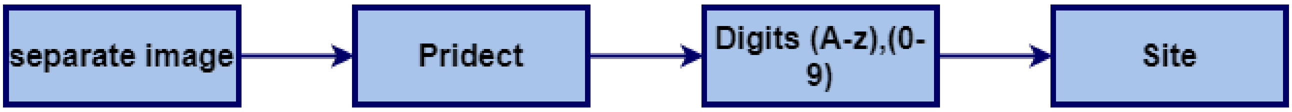


Figure 4.12: Predict the content of each separate image.

4.7 Conclusion

The search for a text-based CAPTCHA break is of great value in research and application, first of all, with the development of technology in artificial intelligence and the latest findings of machine learning techniques has been found technology that helps us with processing and recognition.
It is a theory to organize the work in the next chapter in order to get good results to work to increase CAPTCHA safety strength and prevent its breakage by the robot.

Chapter 5

IMPLEMENTATION AND EXPERIMENTS

5.1 Introduction

In this chapter, We will highlight our design and show how to achieve it, We conducted an experimental study to try to crack a text-based CAPTCHA. Test based on machine learning using several trained algorithms, Where each image contains between 4 to 7 characters out of the 26 English alphabet. After passing the CAPTCHA image hash, we try to find the best character recognition algorithm and break the text-based CAPTCHA test. We will work on the Python programming language for libraries for learning and classification to improve the performance of models.

5.2 DataSet Description

Dataset, To create our CAPTCHA image dataSet. you create an images containing character and numbers in many styles, different colors, a little bit of warping, distortion, rotation, sometimes overlapping, and noise such as a number of dots, lines, and arcs, which we can exploit to increase the diversity of our data and the difficulty of the recognition task and Breaking, resistance and attacks.
We have created a dictionary that contains 62 classes, as each class contains approximately 200 images for each type. We obtained a total number of 9156 images. As training dataset needs a large amount of data to construct the recognition model, We assigned 8000 of them as the training data set size and the test data set size was set to 1156.

Examples from our dataset:

Figure 5.1: Example of a capital characters dataset.

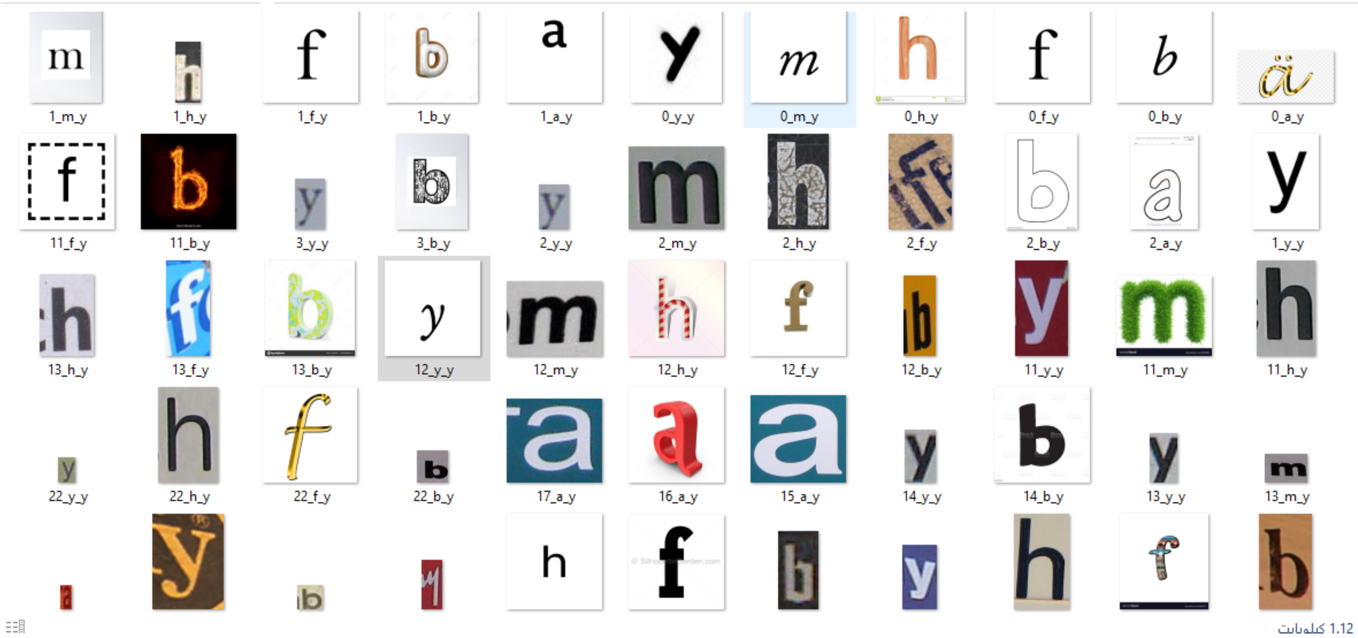


Figure 5.2: Example of a small characters dataset.

Figure 5.3: Example of dataset numbers.

5.3 Environment of the development

5.3.1 Hardware environment

The design and deployment of the web application serving as the user interface was carried out on a DELL laptop with the following characteristics:

- Processor : Intel(R) Core(TM) i5-4310U CPU @ 2.00GHz 2.60 GHz

- RAM : size 6,00 GB
- Operating system : Windows 10 64-bit
- Hard drive :size 500 GB

5.3.2 Language of programming and libraries

Python

Python is an excellent object-oriented, interpreted, and interactive programming language. It is often compared (favorably of course) to Lisp, Tcl, Perl, Ruby, C , Visual Basic, Visual Fox Pro, Scheme or Java ... etc. Python combines remarkable power with very clear syntax. It has modules, classes, exceptions, very high level dynamic data types and dynamic typing. There are interfaces to many system calls and libraries, as well as to various windowing systems.
The new built-in modules are easy to write in C or C ++ (or other languages, depending on the implementation chosen). Python can also be used as an extension language for applications written in other languages requiring easy-to-use scripting or automation interfaces [43].

Figure 5.4: The Python Logo

Jupyter Notebook

Jupyter Notebook is an open source web utility that allows you to create and share files containing dwell code, equations, visualizations, and narrative text content. Uses include data cleaning and conversion, numerical simulation, statistical modeling, data visualization, machine learning, and much more [44].

Figure 5.5: The Jupyter Notebook Logo

TensorFlow

TensorFlow is an open source software library for high performance digital computing. Its flexible architecture allows easy deployment of computing on various platforms (CPUs, GPUs, TPUs), and from desktops to server clusters, to mobile devices. Originally developed by researchers and engineers from the Google Brain team within Google's AI organization, it is based on machine learning and deep learning [45].

Figure 5.6: The TensorFlow Logo

Keras

Keras is a high-level neural networks API, written in Python and capable of running on TensorFlow, CNTK, or Theano. highly-productive interface for solving machine learning problems, with a focus on modern deep learning. It was developed with the aim of allowing rapid experimentation. Be able to go from idea to result as quickly as possible, is the key to doing research:

- Allows quick and easy prototyping (thanks to ease of use, modularity and scalability).
- convolutional networks and recurrent networks as well as combinations of the two.
- Works transparently on CPU and GPU [46].

Figure 5.7: The Keras Logo

Selenium

Selenium is a strong tool for automating web browsers and controlling them through programs. It is compatible with all major browsers, runs on all major operating systems, and has scripts written in a variety of languages, including Python, Java, C, and others; we will be using Python. Selenium Tutorial covers all topics such as – WebDriver, WebElement, Unit Testing with selenium.This Python Selenium Tutorial covers all there is to know about Selenium, from the fundamentals to sophisticated and professional applications.
Advantages Selenium Python:

- Open Source and Portable.
- Combination of tool and DSL: Selenium is a set of tools and a DSL (Domain Specific Language) that may be used to run a variety of tests.
- Easier to understand and implement.
- Less burden and stress for testers.
- Cost reduction for the Business Clients [47].

Figure 5.8: The Selenium Logo

5.4 Examples of source codes

In this section, we will present some source code examples.

```
In [1]: import numpy as np # linear algebra
        import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
        import os
        import pandas as pd
        from pandas.plotting import scatter_matrix
        #import matplotlib.pyplot as plt
        from sklearn import model_selection
        from sklearn.metrics import classification_report, mean_squared_error
        from sklearn.metrics import confusion_matrix
        from sklearn.metrics import accuracy_score
        from sklearn.linear_model import LogisticRegression , LogisticRegressionCV
        from sklearn.tree import DecisionTreeClassifier
        from sklearn.neighbors import KNeighborsClassifier
        from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
        from sklearn.naive_bayes import GaussianNB
        from sklearn.svm import SVC
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.svm import LinearSVC
        from sklearn import preprocessing
```

Figure 5.9: import libraries.

```
In [ ]: X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y,validation_size = 0.20)
        '

        model_svm=SVC(C=50,probability=True)
        model_svm.fit(X_train, Y_train)
        train_accuracy = model_svm.score(X_train,Y_train)
        test_accuracy = model_svm.score(X_validation, Y_validation)
```

Figure 5.10: SVM Training.

```
In [140]: def send_cap(M):
              if (M[0][0]==1):#la poste
                  M[0][1][1].send_keys("Yes, I am Robot")
                  M[0][1][2].send_keys("1234")
                  M[0][1][3].send_keys(M[0][1][0])
              if(M[0][0]==2): #mobilis
                  M[0][1][1].send_keys(M[0][1][0])
              if(M[0][0]==3): # jusitc
                  M[0][1][1].execute_script("document.getElementsByName('secu')[0].value='"+M[0][1][0]+"';")
              global my_list
              my_list = []
```

Figure 5.11: Select the site and textbox of the target site to send the CAPTCHA to.

5.5 Website interfaces

Figure 5.12 shows the website interfaces (E-CCP) that we will broke their CAPTCHA:

Figure 5.12: Example from site Algérie poste E-CCP.

Figure 5.13 shows the website interfaces (Mobilis) that we will broke their CAPTCHA:

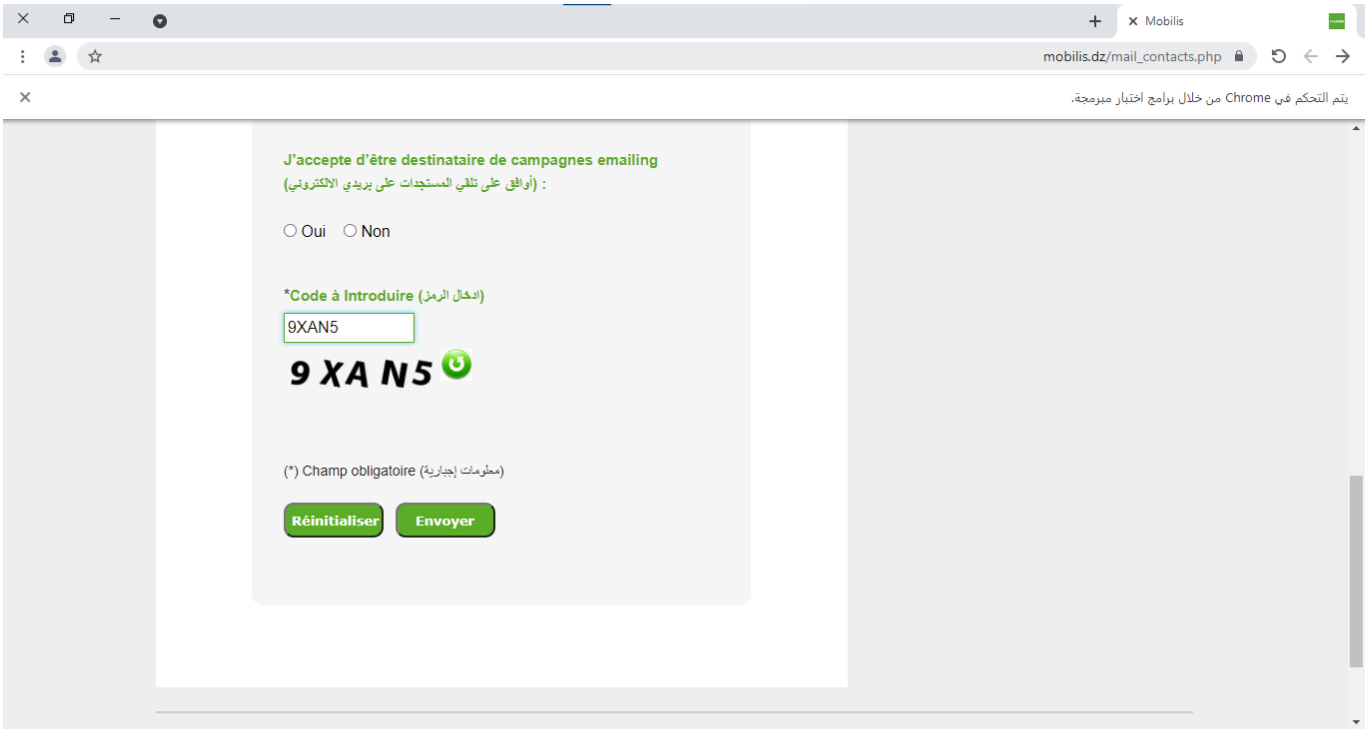


Figure 5.13: Example from site Mobilis.

Example segmentation prediction:

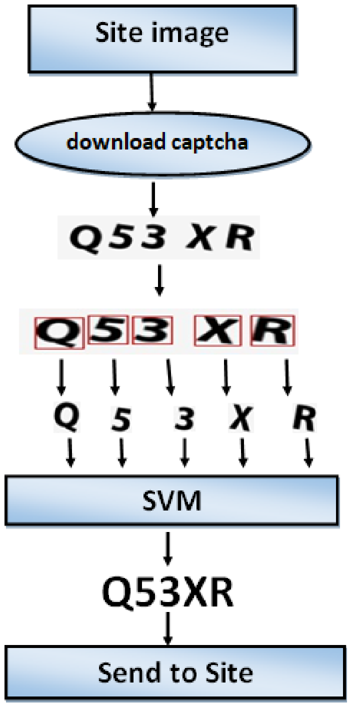


Figure 5.14: Example segmentation prediction.

5.6 Results and Discussion

In table5.1, we are using the supervised learning method, we discussed a group of algorithms used in this study, where we trained first, Then we conducted the experimentation phase for each model in order to find out which one can get the best results to be adopted in the study, In the training column,we trained SVM, BAGGING, RF and DT with high rates, as for GAUS-SIAN, Logitic Regression and k-NN with lower rates. When looking at the results of the experiment, We find that SVM achieved a higher percentage estimated at 81% , then followed by the BAGGING with a lower percentage 75% , followed by RF with a percentage of 70% and then lower results for other models.
So through the results obtained by our trained model, the SVM is generally the best classifier because it achieved the best results in the set of tests.

model	Train accuracy	Test accuracy
Support Vector Machines(SVM)	99%	81%
Bagging	99%	75%
Random Forest(RM)	99%	70%
Decision Trees(DT)	99%	50%
Gaussian	17%	17%
Logitic Regression	58%	35%
k-nearest neighbors algorithm (k-NN)	70%	50%

Table 5.1: Results Train accuracy and Test accuracy for each model .

5.7 conclusion

In this chapter, we conducted a text-based CAPTCHA break-up experiment using a range of algorithms(SVM,Bagging and DT...) where datasat were identified from the CAPTCHA image collection in the target locations. We found SVM is the best model for the best results compared to the rest of the algorithms, Where he achieved a success rate of 81% recognition and breakage ,this poses a risk to the security of the sites from automatic access.

GENRAL CONCLUSION

General Conclusion

The aim of this work is to break text-based CAPTCHA using machine learning techniques. We start with an overview of machine learning and artificial intelligence. Then, we introduced the generality and types of CAPTCHA tests. We introduce also the latest technology to break the text-based CAPTCHA, starting with matching algorithms, up to machine learning techniques. Subsequently, we conducted an experimental study using model-based DataSet training to crack text-based CAPTCHA through the SVM algorithm for handwriting and text recognition and the text-based CAPTCHA image dataset generated by python. Experimental results show that our model can break the 81% accuracy rate generated by text-based CAPTCHA, which can be improved by increasing the data set and training it for a longer period of time.

Despite achieving SVM good results, there are many problems and difficulties that we encountered in this design:

- Many classes, as we have more than 50 classes.
- Similar to uppercase and lowercase letters such as C, c, X and x.
- The difficulty of processing some CAPTCHA images, such as images of overlapping letters.
- The different types and forms of CAPTCHA from one site to another.

From the perspective of improving text-based CAPTCHA cracking, we recommend improving more CAPTCHA image datasets. In order to improve the verification code and make it unbreakable, we recommend using a CAPTCHA google.

However, due to advances in machine learning technology, cracking CAPTCHA is still an active field, and generating an unbreakable
CAPTCHA is still a challenge. "A step backward for CAPTCHA is still a step forward for AI - every defeat is also a victory" (Luis von Ahn 2006).

Bibliographie

[1] K. Madathil, J. Greenstein, K. Horan ”Empirical studies to investigate the usability of text- andimage-based CAPTCHAs” .Computer Science.International Journal of Industrial Ergonomics 2019

[2] Oleg Starostenko , Claudia Cruz-Perez ,Fernando Uceda-Ponga, Vicente Alarcon-Aquino, ”Breaking text-based CAPTCHAs with variable word and character orientation.” Department of Computing, Electronics, and Mechatronics, Cholula 72820, Mexico Volume 48, Issue 4, April 2015, Pages 1101-1112

[3] Ying-LienLee ,Chih-HsiangHsu.”Usability study of text-based CAPTCHAs.” Department of Industrial Engineering and Management, Chaoyang University of Technology, Taichung County 413, Taiwan, Volume 32, Issue 2, April 2011, Pages 81-86

[4] Vu Duc Nguyen, Yang-Wai Chow, Willy Susilo, ”On the security of text-based 3D CAPTCHAs,”School of Computer Science and Software Engineering, Volume 45, September 2014, Pages 84-99

[5] DazhenLina, Fan Linc, Yanping Lva, Feipeng Cai, Donglin Caoa,”Chinese Character CAPTCHA Recognition and performance estimation via deep neural network,” Cognitive Science Department, Xiamen University, China,Volume 288, 2 May 2018, Pages 11-19

[6] Mohammad Tanvir Parvez,Suliman A.Alsuhibany,”Segmentation- validation based handwritten Arabic CAPTCHA generation”,Department of Computer Engineering, College of Computer, Qassim University, Buraydah, Saudi Arabia.Volume 95, August 2020, 101829

[7] Chunhui Li, Xingshu Chen,Haizhou Wang,Peiming Wang, Yu Zhang , Wenxian Wang, ,”End-to-end attack on text-based CAPTCHAs based on cycle-consistent generative adversarial network,” College of Art, Sichuan University, Chengdu 610065, China, Volume 433, 14 April 2021, Pages 223-236

[8] Ying Ma , GuoqiangZhong, Wen Liu, Jinxuan Sun ,Kaizhu Huang, ”Neural CAPTCHA networks,”Department of Computer Science and Technology, Ocean University of China, Volume 97, Part A, December 2020, 106769

[9] Jeff Yan, Ahmad Salah El Ahmad,”Breaking Visual CAPTCHAs with Naïve Pattern Recognition Algorithms,” School of Computing Science, Newcastle University, UK, IEEE, 02 January 2008

[10] Ning Zhang; Mohammadreza Ebrahimi; Weifeng Li; Hsinchun Chen,”A Genera-

tive Adversarial Learning Framework for Breaking Text-Based CAPTCHA in the Dark Web,"International Conference on Intelligence and Security Informatics (ISI), pages1-6,IEEE,08 December 2020

[11] Elie Bursztein and Jonathan Aigrain and Angelika Moscicki and John C. Mitchell,"The End is Nigh: Generic Solving of Text-based CAPTCHAs", USENIX, Workshop on Offensive Technologies, San Diego, CA, 2014

[12] Bilal Khan, Khaled Alghathbar, Muhammad Khurram Khan, Abdullah Alkelabi, Abdulaziz Alajaji,"Cyber Security Using Arabic CAPTCHA Scheme,"Center of Excellence in Information Assurance,Saudi Arabia ,Vol. 10, No. 1, January 2013

[13] Elie Bursztein, Matthieu Martin, J. Mitchell ,"Text-based CAPTCHA Strengths and Weaknesses",Computer and communications security, CCS,Pages 125–138, 17 October 2011

[14] G. Mori and J. Malik, "Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA,"IEEE Computer Society Conference on Computer Vision and Pattern Recognition,pages(1-1), 2003

[15] Banday, M. Tariq, and Nisar A. Shah. "A study of CAPTCHAs for securing web services." arXiv preprint arXiv: 1112.5605, 2011

[16]Tang, M., Gao, H., Zhang, Y., Liu, Y., Zhang, P., Wang, P. "Research on deep learning techniques in breaking text-based captchas and designing image-based captcha". IEEE Transactions on Information Forensics and Security, 13(10), 2522-2537,2018

[17]Roshanbin, Narges, and James Miller. "A Survey and Analysis of Current CAPTCHA Approaches." J. Web Eng. 12.12: 1-40,2013

[18]Chellapilla, K., Larson, K., Simard, P. Y., Czerwinski, M. "Building segmentation based human-friendly human interaction proofs (HIPs)". In International Workshop on Human Interactive Proofs (pp. 1-26). Springer, Berlin, Heidelberg ,2005

[19]Chen, J., Luo, X., Guo, Y., Zhang, Y., Gong, D. (2017)."A survey on breaking technique of text-based CAPTCHA. Security and Communication Networks, 2017.

[20]Sauvola, Jaakko, and Matti Pietikäinen. "Adaptive document image binarization." Pattern recognition 33.2: 225-236,2000.

[21]Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." IEEE transactions on systems, man, and cybernetics 9.1: 62-66,1979.

[22]Boudaoud, Lynda Ben, Abderrahmane Sider, and Abdelkamel Tari. "A new thinning algorithm for binary images".In: 2015 3rd International Conference on Control, Engineering Information Technology (CEIT). IEEE, p. 1-6,2015

[23]Zhang, T. Y., and Ching Y. Suen. "A fast parallel algorithm for thinning digital patterns." Communications of the ACM 27.3 : 236-239,1984

[24]K. Chellapilla, K. Larson, P.Y. Simard, and M. Czerwinski. Computers beat humans at single character recognition in reading based human interaction proofs (hips). In CEAS, 2005.

[25]J. Elson, J. Douceur, J. Howell and J. Saul. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. In Proc. of ACM CCS , pp. 366–374,2007.

[26]Rao, Kameswara, Kavya Sri, and Gnana Sai. ”A novel video CAPTCHA technique to prevent BOT attacks.” Procedia Computer Science 85 : 236-240,2016.

[27]Hasan, Walid Khalifa Abdullah. ”A survey of current research on captcha.” Int. J. Comput. Sci. Eng. Surv.(IJCSES) 7.3 : 141-157,2016. .

[28]Cui, J.S., Zhang, W.Z.,Y .,Liang. Y., Xiao ,B., Mei., J.T., Zhang .,D. Peng , W. , ”A 3- layer Dynamic CAPTCHA Implementation”, Education Technology and Computer Science (ETCS), 2010 Second International Workshop on. Wuhan, vol. 1, no. pp. 23-26,2010.

[29]Singh Ved. Pal preet. (2014), ”Survey of different Types of CAPTCHA”, international Journal of computer science and information technologies, vol.5, no. 2 pp. 2242-2245,2014.

[30]Rizwan Rahman . ”Survey on CAPTCHA systems”, Journal of Global Research in Computer Science, 2012 vol.3, no. 5 pp,2012.

[31]Starostenko, O., Cruz-Perez, C., Uceda-Ponga, F., Alarcon-Aquino, V. ”Breaking text-based CAPTCHAs with variable word and character orientation”. Pattern Recognition, 48(4), 1101-1112,2015.

[32]Gao, H.,Yan, Jeff. ”A simple generic attack on text captchas.” 2016.

[33]Baecher, P., Fischlin, M. G. L., Langenberg, R., Lützow, M., Schröder, D. (2010). Captchas: The good, the bad, and the ugly. Sicherheit . Sicherheit, Schutz und Zuverlässigkeit,2010.

[34]Bostik, Ondrej, and Jan Klecka. ”Recognition of CAPTCHA characters by supervised machine learning algorithms.” IFAC-PapersOnLine 51.6: 208-213,2018. .

[35] Web Site :https://www.sciencedaily.com/terms/artificial$_{i}ntelligence.htm$

[36] Web Site : https://www.ibm.com/cloud/blog/supervised-vs- unsupervised- learning

[37] Web Site : https://www.javatpoint.com/regression-analysis-in-machine-learning

[38] Web Site : https://www.javatpoint.com/machine-learning-random-forest-algorithm

[39] Web Site : https://corporatefinanceinstitute.com/resources/knowledge/ other/bagging-bootstrap-aggregation/

[40] Web Site :https://www.mygreatlearning.com/blog/deep-learning-vs-machine-learning/

[41]Web Site :https://www.linkedin.com/pulse/top-15-mostly-used-captcha-examples-janki-thakkar-1d

[42]Web Site :https://www.imperva.com/learn/application-security/what-is-captcha/

[43] Web Site :https://www.python.org/

[44] Web Site :https://jupyter.org/

[45] Web Site :https://www.tensorflow.org/

[46] Web Site :https://keras.io/

[47] Web Site :https://www.geeksforgeeks.org/selenium-python-tutoria/