REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : SIOD 4/M2/2018

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : **système d'information optimisation et de décision (SIOD)**

Svm mono classe pour la détection des documents administratifs numériques falsifiés

---

Par :

AGTI SANA

Soutenu le 26/06/2018, devant le jury composé de :

Kazar Okba	Professeur	Président
Abdelhamid Djeffal	M.C.A	Rapporteur
Meklid Abdessalem	M.C.B	Examinateur

Dédicace

Toutes les lettres ne sauraient trouver les mots qu'il faut... Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, la reconnaissance..Aussi, c'est tout simplement que. Je dédie ce ce modeste travail à :

À ma chère mère, tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi. Ta prière et ta bénédiction m'ont été d'un grand secours pour mener à bien mes études..

À mon chèr père, aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour vous. Ce travail est le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation.

À ma chère sœur Intissar, qui a toujours été de mon côté, je te souhaite un avenir plein de joie, de bonheur, de réussite et de sérénité..

À mes chère frères : Saad ,Ayman ,Islem et Wassim. Votre affection et votre soutien m'ont été d'un grand secours au long de ma vie professionnelle et personnelle.

À mes chers amis Malika, Aycha, khadidja, Zeyneb, Sara, Fatima, Meryem et Iman. En souvenir de notre sincère et profonde amitié et des moments agréables que nous avons passés ensemble .

Une spéciale dédicace à tous ceux qui m'ont aimé le bien et le succès et qui m'ont soutenu tout au long des années passé et qui était la raison de mon succès ..

À tous mes collègues de l'Université de biskra .Et à tous ce qui ont enseigné moi au long de ma vie scolaire

Remerciement

Tout d'abord, je remercie le Dieu, notre créateur de m'avoir donné la force, la volonté et le courage afin d'accomplir ce travail modeste.

Je veux adresser le grand remerciement les plus sincères à mon encadreur "Dr. Abdelhamid Djeffal" qui a proposé le thème de ce mémoire, pour ses conseils son encouragement, son patience, son aide précieuse et pour le temps qu'il m'a consacré. et ses dirigés du début à la fin de ce travail pour l'attention et la disponibilité dont il a su faire preuve au cours de la réalisation de ce mémoire.

Je tiens également à remercier messieurs les membres de jury pour l'honneur qu'ils m'ont fait en acceptant de siéger à notre soutenance.

Finalement, je tiens à exprimer mon profonde gratitude à ma famille qui ma soutenue et encouragée tout au long de mes études.Ainsi que l'ensemble des enseignants qui ont contribué à mon formation.

RÉSUMÉ

Dans la vie quotidienne, nous comptons sur les documents numériques dans presque toutes les transactions. Et avec la large utilisation de ces documents dans toutes les administrations, un nouveau crime est apparu. C'est la falsification des documents à l'aide d'un scanner, et des outils avancés de retouches et d'édition des images qui permettent à un simple utilisateur d'altérer un document numérique et de changer sont contenu.

Afin de prévenir un tel crime et de lutter contre les fraudeurs qui ont développé de nouvelles méthodes dans ce domaine, de nombreux chercheurs ont essayé de trouver des moyens automatiques de détection de la fraude en utilisant des techniques de traitement d'image et de data mining.

Ce travail vise à étudier la falsification des documents administratifs dans le cas où on ne dispose que de documents authentiques et on cherche à découvrir les documents qui représentent des anomalies par rapport à ces documents. Cette tâche peut être effectuée par l'utilisation de l'apprentissage automatique par la méthode SVM mono classe.

Mots clés : Document numérique, Falsification des documents, Data mining, Apprentissage automatique, Outils de retouche d'image, SVM mono classe.

ملخص

في الحياة اليومية ، نعتمد على المستندات الرقمية في كل المعاملات تقريبا ومع الاستخدام الواسع النطاق لهذه الوثائق في جميع الادارات، ظهرت جريمة جديدة و هي تزوير الوثائق باستخدام ماسح ضوئي وأدوات متقدمة لتغيير وتحرير الصور التي تسمح لاي مستخدم بتزوير مستند رقمي وتغيير محتوياته.

لمنع مثل هذه الجريمة ومحاربة المحتالين الذين طوروا أساليب جديدة في هذا المجال ، حاول العديد من الباحثين العثور على طرق تلقائية للكشف عن الغش باستخدام تقنيات معالجة الصور و التنقيب عن البيانات .

يهدف هذا العمل إلى دراسة تزييف المستندات الإدارية في الحالة التي تتوفر فيها الوثائق الأصلية فقط و يحاول كشف الوثائق التي تمثل حالات مجهولة بالنسبة للوثائق الاصلية. يمكن تنفيذ هذا العمل باستخدام التعلم الآلي .

كلمات مفتاحية: وثيقة رقمية، تزوير الوثائق، تعلم الآلة ، أدوات تحرير الصور

ABSTRACT

In everyday life, we rely on digital documents in almost every transaction. And with the widespread use of these documents in all jurisdictions, a new crime has emerged. This is the falsification of documents using a scanner, and advanced tools for editing and editing images that allow a single user to alter a digital document and change its content.

To prevent such a crime and to fight against fraudsters who have developed new methods in this area, many researchers have tried to find automatic ways to detect fraud using image processing techniques and datat mining.

This work aims at studying the falsification of the administrative documents in the case where one only has authentic documents and one seeks to discover the documents which represent anomalies with respect to these documents. This task can be performed by using automatic learning by the SVM mono class method.

Keywords: Digital document, Falsification of documents, Machine learning, Image editing tools, SVM one class.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLES

INTRODUCTION GÉNÉRALE

De nos jours, en raison de l'avancement de la technologie numérique, logiciel de traitement et outils d'édition, une image peut être facilement manipulée et modifiée. Il est très difficile pour les humains d'identifier visuellement si l'image a été modifiée ou non. Il y a une augmentation rapide des contrefaçons manipulées numériquement dans les médias grand public et sur Internet. Cette tendance indique de sérieuses vulnérabilités et diminue la crédibilité des images numériques.

Alors que ces documents numériques sont largement utilisés dans notre vie, en particulier dans l'environnement officiel et commercial, de nombreux types de falsification ont commencé à apparaître avec différentes techniques.

Ces techniques sont classées en trois catégories générales : copier/déplacer, la contrefaçon et retouche d'image. Ces types de falsification numériques sont principalement effectués sur des données, des documents, des chèques, des images et des vidéos.

Les documents fabriqués sont générés pour obtenir illégalement des avantages à court terme ou à long terme. Cela constitue une menace sérieuse pour le système et l'économie d'une nation.

Par conséquent, développer des techniques pour vérifier l'intégrité et l'authenticité des images numériques est très important, surtout si l'on considère

que les images sont présentées comme des preuves légales, des informations, des documents médicaux, des documents financiers, ou des documents administratif.
En ce sens, la détection de falsification d'image est l'un des principaux objectifs de la crédibilité de documents.

Plusieurs recherches tentent de trouver une solution au problème de fabrication de documents en construisant et en proposant des modèles, approches et techniques pour détecter la fabrication à partir de documents, en utilisant une série de caractéristiques extraites d'une image d'un document tel que les alignements de lignes, les caractères en lettres de langue anglaise et aussi avec quelques propriétés des pixels.

En outre, il n'y a pas d'ensemble de données publiques disponible pour l'expérimentation jusqu'à présent surtout dans le cas de documents administratifs. Ainsi à des fins d'expérimentation, une taille considérable d'échantillons de données pour les tests doit être recueillie.

Ces échantillons peuvent inclure des certificats de conférence, documents officiels, certificats de naissance, ou diplômes d'études. Parfois, il est difficile de recueillir des documents falsifiés pour les utiliser dans l'apprentissage et seuls les documents authentiques sont disponibles [30][11].

Dans ce travail, nous avons proposé et implémenté une méthode pour résoudre le problème de fabrication de documents administratifs numériques, par l'extraction des caractéristiques des images des documents numérisés par un scanner pour ensuite apprendre à détecter automatiquement les documents falsifiés en utilisant la méthode de classification (SVM mono classe).

Ce mémoire comporte quatre chapitres, il est organisé comme suit : Dans le chapitre 1, nous présentons quelques notions fondamentales liées aux documents numériques et leur falsification, les méthode et les outils de falsification. Il contient également des informations de base sur les images numériques afin de faciliter la compréhension des processus d'extraction des caractéristiques nécessaires à notre méthode proposée.

Le second chapitre comprend des travaux connexes et présente l'apprentissage automatique et plus particulièrement la méthode de SVM (mono classe) que nous utilisons dans notre projet où nous détaillons son principe, ses avantages et ses inconvénients.

Le troisième chapitre est réservé à la conception de notre système, son architecture globale et détaillée. Le dernier chapitre est consacré à l'expérimentation et à la discussion des résultats obtenus. Le mémoire est terminé par une conclusion générale contenant les perspectives envisagées.

CHAPITRE

1

FALSIFICATION DES DOCUMENTS NUMÉRIQUES

1.1 Introduction

Ce n'est plus un secret le phénomène de la falsification des documents numérique ou imprimé. Il ne fait aucun doute que le développement technologique a rendu la contrefaçon relativement facile ; maîtrisant le travail sur le programme (Photo shop) avec une imprimante laser couleur moderne et un scanner peut imiter n'importe quel document.

Dans ce chapitre, nous présentons les concepts de base qui tournent autour des documents numériques, la numérisation, et la falsification des documents. Nous mentionnons l'ensemble des différences entre le document numérique et imprimé, la numérisation, ses techniques et ses outils. Nous présentons également les objectifs souhaités de la documentation et l'objectif de la falsification des documents administratifs en particulier. Ensuite nous détaillons les différents types de fraudes la falsification des documents, et plus particulièrement les méthodes et les outils de falsification. De plus, le chapitre discutera les techniques de protection contre la falsification, et les signaux d'alerte.

1.2 Document numérique

1.2.1 Définition d'un document

Le document est un ensemble d'informations mises à la disposition d'utilisateurs. Il peut être vu comme un ensemble formé par un support(papier, analogique, numérique) et une information(textes, images,...) qui peut être lue par l'homme ou la machine [41].

1.2.2 Types de documents

Un document peut prendre plusieurs formes selon la technologie proposée. On peut citer les types suivants :

- **Papier :** Un document papier est lisible directement et ne nécessite aucun autre moyen pour le décoder. Le document papier peut servir à récolter les informations, les diffuser ou les conserver.
- **Multimédia :** Un document multimédia comporte des images, des vidéos, des sons numérique. Ces fichiers sont obtenus grâce à des outils d'acquisition numérique comme des appareils photo, caméras, etc..
- **Numérique :** est une forme de représentation de l'information consultable à l'écran d'un appareil électronique. Un document numérique est codée en binaire et n'est pas lisible directement. Il peut être conçu directement sous forme numérique ou issu d'un document papier par numérisation [10].

1.2.3 Définition d'un document numérique

Un document numérique est un document qui a pour caractéristique d'être sur un support électronique, d'être perceptible via la technologie numérique, il peut être une image, un fichier son, un ensemble de données organisées en fichier, un écrit électronique, ...etc. [29]. Il est considéré comme un fichier informatique (et donc représenté à la base par une suite de 0 et de 1) dont le contenu, structuré selon les spécifications d'un format de fichier, représente une information compréhensible par un humain et/ou par un ordinateur [42].

Définition de l'ISO Un document numérique est : "Ensemble formé par un support et une information, généralement enregistrée de façon permanente, et tel qu'il puisse être lu par l'homme ou la machine" [43].

Les documents numériques issus d'une numérisation servent a :

— l'impression,
— la documentation,
— la recherche et a la publication en ligne,
— le plus souvent pour la gestion des collections, la préparation de catalogues et la promotion d'expositions [42].

1.2.3.1 Formats de représentation des documents numériques

Pour les documents numériques, il existe plusieurs types, et pour chaque type, il y a des formats adaptés pour la représentation des documents. On cite à titre d'exemple les documents textes, les documents images, et les documents multimédias. Pour le premier, les formats Word, HTML, XML, PDF, SGM,... peuvent être des moyens pour stocker et représenter ce type de documents.

Pour les documents images, les formats utilisés sont :TIFF, JPEG, GIF. Pour les documents multimédias qui ne représentent pas une part très importante dans les bibliothèques numériques, il y a par exemple le format MPEG pour la vidéo et les formats WAV, MP3, MIDI,.. pour le son [43].

1.2.4 Différence entre un document imprimé et un document numérique

Avec l'utilisation généralisée d'Internet et l'existence de nombreuses transactions administratives qui reposent principalement sur des documents numériques, l'utilisation de documents papier est nettement réduite par rapport à l'utilisation de documents numériques. La table suivante montre la différence entre les deux [43].

document imprimé	document numérique
— Formes figées — Le couple support-contenu constitue une unité indissociable.	— Formes variés : multimédia, virtuelles, dynamiques — Chemins de lecture : bouquets de liens, réseaux hypertextes — Formes non figées — Document constitué de plusieurs fichiers informatiques disjoints — Support invisible — Programmes encodant la structure physique et intellectuelle du document aident à développer des études structurelles, des recherches fines, compilations d'index, manipulation et transformations du texte, etc.

TABLE 1.1: La différence entre un document imprimé et un document numérique.

1.2.5 Numérisation

On utilise le terme "*numérique*" du fait que par cette procédure, des données analogiques sont traduites en langage binaire (une suite de 1 et de 0) par un numériseur (scanner). La numérisation consiste à créer une copie en mode image d'un document physique existant [21].

1.2.5.1 Niveaux d'informations numériques

il existe trois niveaux d'informations numériques :

— **Les données :** ce sont des informations stockées sous forme numérique, considérées indépendamment de tout contexte de production et d'interprétation.

— **Les ressources :** sont des informations construites dans une logique de médiation et d'usage, évolutives. Leur fonction est d'être utiles et de rendre des services. Elles fournissent du renseignement .

— **Les documents** ils sont élaborés par des logiques de production qui leur donnent naissance, et des logiques de médiation qui les rendent accessibles. Ils doivent être authentifiés et stabilisés [36].

1.2.5.2 Numérisation des documents

La numérisation des documents est le processus de convertir des documents papiers en formats numériques. Un service de numérisation professionnel applique un processus en plusieurs étapes afin de digitaliser les documents physiques :

— Le triage et la préparation
— L'imagerie
— L'indexation
— L'entreposage [3].

1.2.5.3 Objectifs de la numérisation

Les objectifs de la numérisation de documents peuvent être de quatre ordres, soit :

— **La préservation des documents** La numérisation à des fins de préservation vise les documents dont le support est obsolète, qui présentent des altérations ou dont la manipulation peut causer une détérioration irréversible. Les documents originaux seront conservés, à moins qu'ils ne soient complètement irrécupérables. La copie numérisée constitue la copie de consultation privilégiée auprès des utilisateurs.
— **La diffusion des documents** La numérisation à des fins de diffusion vise les documents qui seront utilisés dans le cadre d'un projet de diffusion telle une exposition ou pour rendre accessibles des documents aux utilisateurs sur place ou à distance.Les documents originaux seront conservés, mais comme dans le cas précédent, la consultation se fera à partir de la copie numérisée.
— **La sauvegarde des documents** La numérisation à des fins de sauvegarde de documents vise essentiellement des documents d'une importance vitale pour les institutions (documents essentiels) et qui nécessitent la conservation d'un deuxième exemplaire, par mesure de précaution (copie de sécurité). Habituellement, cette copie de sécurité sera effectuée sur un support différent et, de préférence, conservée dans un autre lieu que les originaux.
— **La substitution des documents** La numérisation à des fins de substitution vise à rationaliser les coûts de conservation liés aux espaces et aux ressources matérielles nécessaires pour l'entreposage des documents. Elle vise également à faciliter l'accès et la consultation des documents [19].

1.2.5.4 Chaîne de Numérisation

La chaîne de numérisation représente l'infrastructure technique rendant possible l'imagerie numérique. Elle se compose de matériel, de logiciels et de réseaux. Une vision complète de l'infrastructure technique comprend également les protocoles et normes, politiques et procédures (concernant le déroulement du travail, la maintenance, la sécurité, les mises à jour, etc...)[14].

Elle suggère une série d'étapes logiquement organisées. Les principales étapes d'une chaîne de numérisation sont :

- **L'acquisition :** permettant la conversion du document papier sous la forme d'une image numérique. Cette étape est importante car elle se préoccupe de la préparation des documents à saisir, du choix et du paramétrage du matériel de saisie (scanner), ainsi que du format de stockage des images.
- **Le prétraitement :** dont le rôle est de préparer l'image du document au traitement. Les opérations de prétraitement sont relatives au redressement de l'image, à la suppression du bruit et de l'information redondante, et enfin à la sélection des zones de traitement utiles.
- **La reconnaissance du contenu :** qui conduit le plus souvent à la reconnaissance du texte et à l'extraction de la structure logique. Ces traitements s'accompagnent le plus souvent d'opérations préparatoires de segmentation en blocs et de classification des médias.
- **La correction des résultats :** de la reconnaissance en vue de valider l'opération de Numérisation. Cette opération peut se faire soit automatiquement par l'utilisation de dictionnaires et de méthodes de correction linguistiques, ou manuellement au travers d'interfaces dédiées [12].

1.2.5.5 Techniques de numérisation

Vu la volumétrie des documents papiers dans les différents organismes que ce soient dépôts d'archives, bibliothèques, centres de documentation, entreprises ou administrations. La numérisation s'avère le mode d'acquisition le plus utilisé.
L'acquisition des documents papier s'effectue principalement par numérisation à l'aide d'un scanner [44].

1.2.5.5.1 Définition d'un scanner Scanner ou numériseur de document. Est un périphérique informatique qui permet de transformer un do-

cument ou des informations (images, textes, etc.) figurant sur un support, généralement papier en une image numérique. Le document est soumis au balayage d'un rayon lumineux. Un capteur transforme la lumière reçue en un signal électrique qui est transféré à l'ordinateur, pour y être ensuite sauvegardé, traité ou analysé [61].

Les scanners sont déjà utilisés pour authentifier les documents présentés pour affirmer et prouver une identité. Régulièrement le partage des données, détectées par l'utilisation de scanners pourraient améliorer de manière significative la lutte contre la criminalité d'identité. Collation des données d'identité fausses informerait les activités d'application de la loi pour réduire la perte financière pour les secteurs privés et public. L'utilisation de scanners pourrait permettre une identification plus efficace des documents contrefaits au point de présentation, et le partage de ces données permettrait réduire encore la fraude dans le secteur public et privé [9].

1.2.5.5.2 Fonctionnement d'un scanner Il se basé sur trois principes : l'éclairage, la réflexion et la capture.

- Le scanner éclaire la page à numériser par une lampe
- Les rayons réfléchis sur la page sont conduits à l'aide d'un système de miroirs vers une barre de capteurs dite barre de CDD (Coupled Charged Devices)
- Les capteurs transforment la lumière reçue en signal électrique qui sera traité par la partie électronique du scanner [44].

1.2.5.5.3 Caractéristiques techniques des scanners Il existe de nombreuses caractéristiques techniques du scanner. Nous mentionnons les caractéristiques suivantes :

- **La résolution** C'est la précision de la capture de l'image, s'exprime en points par pouce (ppp), unité utilisée pour indiquer le degré de précision d'une image sur une surface donnée. La résolution dépend du type de l'image, la résolution de l'image d'origine, du matériel du sortie, de l'espace de mémoire disponible, etc.[44]
- **Types d'acquisition :**
  - **Mode monochrome :** utilisé pour les textes et illustration en noir et blanc, sans niveaux de gris.
  - **Mode niveau de gris :** utilisé pour les images contenant 256 niveaux de gris.

— **Mode couleur :** utilisé pour les images couleurs.[44]
— **Le format de papier :** c'est la grandeur maximale des feuilles que le numériseur peut accepter.
— **L'interface :** C'est le type de liaison utilisé pour relier le scanner à l'ordinateur.
— **La vitesse de numérisation :** C'est le nombre de pages numérisées par minute ; la vitesse de numérisation dépend non seulement de la puissance du numériseur mais aussi du format des documents à numériser et de la résolution choisie pour le travail [61].

1.2.6 Résultat de la numérisation

La numérisation de documents a deux objectifs différents. D'une part la reproduction numérique au sens strict. Et d'autre part la préparation d'une image pour une utilisation dans un contexte précis (impression, tirage, illustration de livre, montage pour une affiche,...) [25].

Numériser une image c'est lui donner une représentation électronique à partir de l'objet réel qui lui sert de support (papier, film, diapo, négatif, mais aussi objet 3D). Cette représentation sera la plupart du temps matricielle, c'est-à-dire une matrice (un tableau) où chaque point sera représenté par une couleur .
Donc Le résultat de la numérisation à l'aide d'un scanner c'est la production d'une image numérique à partir d'un document papier[4].

1.2.6.1 Image numérique

Une image numérique est une image dont le support est stocké sous forme binaire dans un fichier informatique. Celle-ci peut être obtenu soit à partir de capteurs optiques (appareil photo, caméra, scanner,...) ou créé à partir de logiciels (Paintbrush, libreoffice..) [13].

Définition : L'image numérique est une matrice de pixels dont chacun comporte une information de couleur et de luminance. Cette matrice simule une image quand la taille des pixels est suffisamment petite . Les images numériques sont formées d'une grille de petits carrés appelés pixels . Ce sont les plus petits éléments employés par les moniteurs et imprimantes d'ordinateur pour représenter des caractères, des graphiques ou des images [7].

1.2.6.2 Structure d'une image

Chaque image numérique est constituée d'un nombre donné de lignes. Chaque ligne comporte un nombre de point donnés. L'ensemble constitue une matrice. Chaque case de cette matrice contient des nombres caractéristiques à la couleur attribuée au pixel [13].

1.2.6.3 Formats d'images

Les formats de stockage des images varient selon les algorithmes mise en œuvre pour coder l'information. Cela se traduit par des caractéristiques propres à chaque format et donc des utilisations différentes.

De façon générale, on distingue deux types de formats d'images [4] :

— **Les formats matriciels (bitmap) :** L'image est considéré comme une matrice (un tableau) de points (ou pixels) ayant chacune une couleur. Elles sont utilisées pour stocker des images simples [4]. Une image matricielle est caractérisée notamment par :
  — sa dimension en pixels,
  — sa résolution,
  — et son mode colorimétrique [60]

— **Les formats vectoriels :** Le principe des images vectorielles est de représenter les données de l'image à l'aide de formules mathématiques. Cela permet alors d'agrandir l'image indéfiniment sans perte de qualité et d'obtenir un faible encombrement [13].

1.2.6.4 Caractéristiques d'une image numérique

L'image est un ensemble structuré d'information. Donc ces informations ont des caractéristiques définies par les paramètres suivantes :

1.2.6.4.1 Pixel : La plus petite unité constitutive d'une image. Chaque pixel affiche une seule couleur. Le pixel est également employé comme unité de mesure de la taille et de la résolution d'une image [7].

1.2.6.4.2 Dimension : C'est La taille de l'image. Elle présente le nombre de pixels de l'image, donné par le produit du nombre de lignes et le nombre de colonnes de la matrice associée à l'image [32].

1.2.6.4.3 Résolution : La résolution d'une image est définie par le nombre de points d'image ou "pixels" représentant l'image par unité de longueur de la structure à numériser (l'image initiale). On exprime cette résolution en points ou pixels par pouce ou (dpi) [32].

1.2.6.4.4 Luminance (Intensité) : C'est le degré de luminosité des points (Pixels) de l'image. Elle est définie aussi comme étant le quotient de l'intensité lumineuse d'une surface par l'aire apparente de cette surface [16].

1.2.6.4.5 Contraste : Le contraste est une propriété intrinsèque à une image qui permet de quantifier la capacité de distinguer deux régions distinctes. Il s'agit dans ce cas de distinguer deux régions suffisamment grandes d'après l'intensité des points présentés par des niveaux de gris [32].

1.2.6.4.6 Contours : Les contours représentent la frontière entre les objets de l'image, ou la limite entre deux pixels dont les niveaux de gris ou couleurs représentent une différence significative [32].

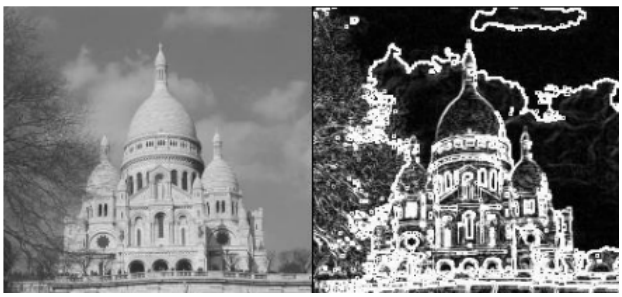


FIGURE 1.1: Exemple de Contour

1.2.6.4.7 Texture : Les textures décrivent la structure de l'image. L'extraction de contour consiste à identifier dans l'image les points qui séparent deux textures différentes [32].

FIGURE 1.2: Exemple de Texture

1.2.6.4.8 Couleur : La couleur est un des premiers descripteurs qui sont employés pour la recherche d'images. La couleur forme une partie significative de vision humaine. L'espaces de couleur le plus utilisée c'est le RGB. La forme la plus simple de descripteur de couleur est l'histogramme de couleur [32].

1.2.6.4.9 Forme : Au même titre que les caractéristiques de texture, les attributs de forme sont complémentaires de la description couleur. Les caractéristiques de forme sont extraites à partir des régions dans les images (contours), Nous distinguons deux catégories de descripteurs de formes :

— Les descripteurs basés régions.
— Les descripteurs basés frontières [32].

1.2.6.4.10 Bruit : C'est un signal qui lors de l'acquisition ou la transmission vient s'ajoutera à l'image. Il se matérialise par la présence dans une région homogène des valeurs plus ou moins éloignées de l'intensité de la région. Le bruit est le résultat de certains défauts électroniques du capteur et de la qualité de numérisation [32].

FIGURE 1.3: Exemple de Bruit

1.2.6.4.11 Histogramme : L'histogramme des niveaux de gris ou de couleurs d'une image est une fonction qui donne la fréquence d'apparition de chaque niveau de gris ou couleur dans l'image. Il permet de donner un grand nombre d'informations sur la distribution des niveaux de gris ou des couleurs [16].

1.2.6.5 Différents types d'images

L'image apparait comme une matrice où chaque case contient des nombres associés à une couleur. Il existe différentes catégories d'image selon le nombre de bits sur lequel est codée la valeur de chaque pixel.

— **Le noir et blanc :** Le mode le plus simple. Le contenu de chaque case de la matrice est soit un 0 (noir) soit 1 (blanc). Le nombre de

couleurs n'est que de 2 et le rendu de l'image le moins performant mais parfois suffisant dans le cadre par exemple de documents scripturaux.

FIGURE 1.4: Image en mode monochrome

— **Les niveaux de gris :** Le codage dit en niveaux de gris permet d'obtenir plus de nuances que le simple noir et blanc. Il offre des possibilités supplémentaires pour coder le niveau de l'intensité lumineuse [13]. La couleur de pixel prend des valeurs allant du noir au blanc en passent par un nombre fini de niveaux intermédiaires. Les valeurs peuvent être comprises entre 0 et 255. Les pixels sont alors codés non pas sur un bit mais sur un octet[32].

FIGURE 1.5: Image en niveau de gris.

— **La couleur :** La couleur d'un pixel est obtenue par le mélange de couleurs fondamentales : rouge, vert et bleu (RVB), on obtient toute une palette de nuances allant du noir au blanc. A chaque couleur est associé un octet (donc 256 niveaux de luminosité) de chacune des couleurs fondamentales. Un pixel 'couleur' est alors codé avec 3 octets et on a alors la possibilité d'obtenir 2^{24} possibilités de couleurs soit de l'ordre de 16 millions de couleurs différentes .

FIGURE 1.6: Image en couleur.

Ces types sont généralement à choisir lors d'une numérisation par scanner ou lors de la configuration d'un appareil photographique [13].

1.2.6.6 Amélioration et prétraitement d'images

Les images brutes permettent rarement de parvenir à une extraction directe des objets à analyser. Avant d'extraire les objets et d'analyser une image, il va donc être nécessaire d'améliorer l'image.

1.2.6.6.1 Amélioration d'images

- **En améliorant le rapport signal sur bruit :** Le transfert de l'image d'un objet jusqu'à l'ordinateur se produit avec un certain bruit. Le bruit est dû en particulier aux imperfections de la source qui génère l'image. On améliorera la qualité de l'image en modifiant le rapport signal sur bruit.
- **Par filtrage :** L'amélioration du rapport signal sur bruit, bien qu'intéressante, ne suffit pas toujours pour obtenir de bonnes images. L'amélioration de l'image est essentiellement obtenue par ce que l'on appelle une opération de filtrage. Un filtre est une transformation mathématique permettant, pour chaque pixel de la zone à laquelle il s'applique, de modifier sa valeur en fonction des valeurs des pixels avoisinants, affectées de coefficients.

1.2.6.6.2 Segmentation d'images Le segmentation d'image est une opération de traitement des images. C'est une des étapes critiques de l'analyse d'images qui conditionne la qualité de la mesure ultérieurement effectuée. Elle permet d'isoler dans l'image les objets sur lesquels doit porter l'analyse.

1.2.7 Raisons d'utilisation du scanner pour aider à détecter la Fraude :

Les utilisations du scanner sont nombreuses et variées et sont également utilisées dans la détection de la falsification de documents officiels et administratifs. Cela est dû à une variété de raisons :

- **Ils sont précis :** Les scanners utilisent un lecteur dédié pour prendre une image à haute résolution d'un document. L'image et les caractéristiques de sécurité intégrées, y compris la puce RFID sont analysées puis comparées aux données capturées par l'OCR. Dans certains cas, les bases de données sont utilisés pour confirmer si le document présenté est authentique ou non. Les scanners sont très efficaces par rapport à une inspection visuelle de base par une personne ayant une formation limitée dans l'examen des documents .
- **Ils sont rapides :** Avec des millions de documents numérisés chaque mois, la vitesse est essentielle. Le logiciel peut traiter un document et de livrer les résultats en quelques secondes, ce qui pourrait fournir des milliers de résultats par jour. Les meilleures solutions matérielles et logicielles peuvent lire des documents d'identité tels que les passeports et les permis de conduire, ainsi que les factures d'électricité et d'autres documents officiels.
- **Maintient la cohérence :** Les scanners permettront à un examen plus large et plus cohérent des documents que ce qui est actuellement possible [31].

1.3 Falsification des documents

1.3.1 Définition

C'est la modification d'un ou plusieurs éléments d'un document authentique. La falsification peut porter sur la date de validité, sur les mentions d'identité ou encore sur la photographie [37].
Les modifications plus importantes peuvent amener à une altération complète de l'identité portée sur le document original. L'altération peut être :

- **Physique :** un document peut être modifié physiquement (suppression d'éléments ou de références, ajout manuscrit d'informations altérant le document, par exemple) ;
- **Intellectuelle :** le contenu du document n'est plus conforme à la réalité (description inexacte des services rendus, contenu erroné d'un rapport, apposition de fausses signatures sur la liste de présence.)[23].

1.3.2 Un document administratif

Les documents administratifs sont des documents et formulaires normalisés. Ils sont établis pour constater un droit, une identité (carte d'identité, titre de séjour), une qualité (certificat de nationalité) ou accorder une autorisation (tous les documents qui délivrent des permis) certificat d'immatriculation d'un véhicule ou permis de conduire, certificat de mariage,.. etc [6].

1.3.3 Documents administratifs susceptibles d'être falsifiés [23] :

Tous les types de documents demandés par les différentes organisations et fournis par les bénéficiaires dans les différents secteurs pour l'obtention des différents services sont susceptibles d'être falsifiés :

- contrats de travail ;
- pièces d'identité ;
- garanties bancaires ;
- bilans ;
- factures (sous format papier ou électronique) ;
- rapports ;
- décompte horaires ;
- listes de présence ;
- Autres

1.3.4 Types de falsification

Il existe beaucoup de méthodes pour fabriquer des images ou des documents. Dans cette section, nous allons présenter les principales méthodes utilisées pour produire les documents falsifiés à partir des documents originaux. Il existe trois types de falsification :

1. **Falsification de la photo :** Cela signifie le remplacement d'une autre photographie à la place de la photo de la personne authentifiée.
2. **Falsification du texte :** l'ajout d'un paragraphe. Le fraudeur scanne le document puis modifie l'image en utilisant un outil de traitement pour ajouter une nouvelle ligne ou paragraphe au document et l'imprimer. Il peut remplacer aussi le contenu du document dans les régions variables, par coupure ou par la technique de copier-déplacer.

FIGURE 1.7: La forme d'un document.

3. **Falsification de cachet/signature :** Consiste à copier le cachet ou la signature existante dans un document authentique dans le document fabriqué [31].

1.3.5 Types d'opération de fraude documentaire

Les criminels utilisent souvent de manière frauduleuse des documents d'identité et de voyage pour mener à bien leurs activités illégales. Les documents faux et authentiques sont utilisés pour perpétrer diverses fraudes qui peuvent être classées comme suit :

La contrefaçon : Production intégrale par imitation d'un document d'identité. Elle tente de reproduire un document authentique [53].

La falsification : La modification illégale d'un document d'une manière ou d'une autre :
— Remplacer une page.
— Substituer une photo / image.
— Modification des données personnelles (les mentions d'identité).

Le document volé vierge : documents authentiques ayant été volés avant leur personnalisation et qui seront ensuite complétés par le voleur [37].

Le document fantaisiste : Le fraudeur va inventer un document qui n'existe pas, pariant sur la méconnaissance de la personne qui contrôlera, ils peuvent se présenter sous diverses formes et avoir l'aspect physique d'un passeport. Ils ne sont pas une déclaration acceptable de nationalité ou d'identité [37].

L'obtention indue : Document authentique délivré sur la base de faux documents (actes de naissance, justificatif de domicile, déclaration de perte, etc.) pouvant être contrefaits, falsifiés, usurpés ou obtenus indûment [37].

1.3.6 But de la falsification

Dans le monde de la contrefaçon, les fraudeurs essaient de répondre aux besoins et aux exigences des utilisateurs qui cherchent à falsifier certains documents administratifs à des fins différentes, entre autre :

- Pour cacher l'identité du détenteur.
- Pour endosser une autre identité.
- Pour cacher des antécédents jugés embarrassants en matière d'immigration ou déplacements .
- Pour bénéficier du même statut que le détenteur précédent .
- Pour éviter de devoir détenir un visa pour éviter tout contact avec une autorité de délivrance [2].

1.3.7 Méthodes de falsification

Les images numériques sont faciles à manipuler et à modifier en raison de la disponibilité de puissants logiciels de traitement et d'édition d'images.

De nos jours, il est possible d'ajouter ou de supprimer des éléments importants à partir d'une image sans laisser de traces évidentes de falsification.

L'image falsifiée est une grande menace, car des outils nouveaux et nouveaux sont disponibles avec un prix moins cher pour forger l'image numérique. Comme il existe de nombreux types de falsifications d'images parmi eux :

Copier et coller La falsification de la copie est populaire comme l'un des types de techniques d'altération d'images difficiles et les plus couramment utilisés. Dans cette technique, il faut couvrir une partie de l'image afin d'ajouter ou de supprimer des informations.
Dans l'image Copier-Déplacer, la technique de manipulation c'est que une partie de la même image est copiée et collée dans une autre partie de cette

image elle-même. Dans une attaque de copie, l'intention est de cacher quelque chose dans l'image originale avec une autre partie de la même image [51].

FIGURE 1.8: Exemple de falsification de copie-déplacement (a) image original (b) image falsifiée.

La fusion d'images (Image Splicing) : C'est une technologie de composition d'images en combinant des fragments d'images à partir des mêmes images ou des images différentes à l'aide d'outils numériques disponibles tels que Photoshop.

La falsification de l'image implique la composition ou la fusion de deux ou plusieurs images modifiant l'image originale de façon significative pour produire une image falsifiée. Dans le cas où des images avec un arrière-plan différent sont fusionnées, il devient très difficile de créer les bordures. La détection de l'épissage est un problème difficile par lequel les régions composites sont étudiées par une variété de méthodes. Des changements abrupts entre les différentes zones qui sont combinées et leurs arrière-plans fournissent des traces précieuses pour détecter l'épissage dans l'image considérée [28].

Retouche d'image Dans la retouche d'image numériques, les images sont moins modifiées. Cela améliore certaines fonctionnalités de l'image. L'image est réalisée pour réduire ou améliorer certaines caractéristiques de l'image. La retouche peut nécessiter la rotation, la mise à l'échelle ou l'étirement d'une image avant de la combiner avec une autre image. C'est un type de changement d'image très courant et fait très souvent des publicités. Le clonage de la partie de l'image est également très répandu dans la retouche d'image. La détection est très difficile car il n'y a pas de changement radical dans les différentes parties de l'image [28].

Imitation Le fraudeur ajoute ou remplace des informations en essayant de trouver les mêmes propriétés des polices d'un document. Par conséquent,

(a)

(b)

FIGURE 1.9: Exemple de retouche d'image (a) image original (b) image falsifiée.

le document numérique falsifié final contient des mots ayant un type de police,une taille de caractère, un dés-alignement ou un biais différent [48].

État de l'éclairage : Ce type de faux peut se faire facilement en épissant ensemble deux images différentes. Souvent, ces images épissées proviennent de différentes scènes et ont des conditions de foudre différentes et il est donc très difficile pour le forger d'image de faire correspondre l'état de foudre exact d'une image avec l'autre. Une telle variation dans les conditions d'éclairage peut être utilisée pour identifier la trempe dans l'image. Plusieurs fois, l'épissage de l'image se fait avec une telle précision qu'il est évidemment impossible d'identifier différentes conditions de foudre dans l'image combinée. Dans la

mesure où la direction de la source de lumière peut être estimée pour différents objets/personnes dans une image, des incohérences dans le sens de l'éclairage peuvent être utilisées comme preuve d'altération numérique [28] .

Recadrage : Le recadrage ou rognage est une technique permettant de couper les bordures ou de supprimer une partie périphérique d'une image. Généralement, ce type d'opération est utilisé pour supprimer les informations de bordure qui n'est pas très important pour l'affichage ou de l'adapter à un usage autre que celui pour lequel elle a été réalisée, ou de modifier son format [31].

1.3.8 Outils de falsification

Il est plus difficile de falsifier un document papier que de modifier un fichier informatique. Mais une fois imprimé, un faux document peut paraître authentique. Pour cela il y a plusieurs outils utilisés pour modifier ou falsifier un document numérique :

GIMP (GNU Image Manipulation Program) Est probablement le logiciel gratuit le plus puissant qui offre, lui aussi, une multitude d'outils favorisant la retouche photo et le montage photo. Il peut être utilisé comme un logiciel de retouche photo mais aussi comme un outil de peinture digital et de dessin, ou encore pour convertir des formats d'images. il a des outils utilisés pour l'édition d'image, le dessin à main levée, réajuster, rogner, photomontages, convertir entre différents formats d'image, et plus de tâches spécialisées [1].

Paint.NET C'est un outil d'édition et de retouche photo. Il dispose de toutes les options d'édition essentielles pour retoucher et optimiser les images telles que la correction des couleurs, du contraste, de la netteté et du flou, un remplacement de couleurs, la capture d'écran, la gestion en proposant notamment des pinceaux personnalisés, un remplacement de couleurs, la capture d'écran, la gestion du format Bitmap en 32 bits, des outils texte, ..etc [58].

Photoshop Outil de retouche d'images pour l'impression ou pour le Web. Logiciel phare de la société Adobe et mondialement il utilisé pour ses capacités hors du commun. Photo-shop est l'outil le plus utilisé actuellement pour la retouche d'images de qualité professionnelle. Ses nombreux outils lui permettent d'effectuer presque tout ce qui est possible de faire sur une image. Ses principales fonctionnalités sont :

— La conception de sélections dans différents modes qui permet de choisir avec une grande précision les partie de l'image devant subir des modifications.
— Les nombreux outils permettant de couper tout ou partie d'une image afin d'effectuer des montages.
— La création des masques divers permettant d'effectuer des montages entre différents visuels.
— De nombreux filtres permettant des effets spéciaux Photoshop est grandement utilisé pour le traitement de photographies numériques [5].

Photo scape PhotoScape est un programme gratuit qui permet de réaliser de nombreuses opérations liées à l'image au sein d'une seule et même interface. Photoscape sert à organiser ses photos, à convertir des images et à créer des Gifs animés. PhotoScape est non seulement capable d'effectuer les opérations de base avec de nombreuses possibilités d'édition (recadrage, retouche de luminosité, contraste, etc), mais ce logiciel propose également tout un lot de fonctions toutes aussi intéressantes les unes que les autres (Mise en page, Traitement par lots, Combinaison d'images, Impression, Fractionnement, Capture d'écran, Collecteur de couleur, Visionnage, etc.) [27].

Corel Photo-Paint Corel Photo-Paint est un logiciel de retouche, de traitement et de dessin assisté par ordinateur édité par Corel Inc. Il est essentiellement utilisé pour le traitement de photographies numériques [58]
.

Photo Filtre C'est un logiciel de retouche d'images très complet. Il permet d'effectuer des réglages simples ou avancés sur une image et de lui appliquer un large éventail de filtres. Outre les traitements classiques de l'image, il dispose d'une centaine de filtres pour améliorer et transformer des photos numériques [58].

1.3.9 Indicateurs de fraude (Signaux d'alerte)

Les indicateurs de fraude sont des signes très spécifiques ou "signaux d'alerte" indiquant l'existence d'une activité frauduleuse ou de corruption présumée, lorsqu'une réaction immédiate est nécessaire pour vérifier si d'autres actions s'imposent [38].

définition : Le signal d'alerte est constitué d'un élément ou d'un ensemble d'éléments s'écartant de la normalité ou qui par leur nature pré-

sentent un caractère inhabituel. C'est le signal d'une anomalie pour laquelle des recherches plus approfondies pourraient s'avérer nécessaires.

1. **Dans le format des documents** Il convient de s'interroger sur les documents dont la présentation s'écarte des normes établies et généralement admises. Par exemple, dans le cas des documents financiers on peut découvrir la falsification à partir : des factures ou des lettres n'affichant pas le logo de l'entreprise, des différences visibles dans le type, la taille, la netteté, la couleur de la police de caractères utilisée dans le document..etc.
2. **Dans le contenu des documents** C'est l'ensemble des modifications affectant le contenu du document. On se trouve dans le cas da falsification des documents financiers un ensemble d'indicateurs qui avertissent d'un changement suspect sur le contenu. Par exemple, caractère insolite des dates, montants, annotations, numéros de téléphone ou calculs, inscriptions manquantes (dans les vérifications séquentielles).etc[23].

1.3.10 Techniques de protection

Les moyens informatiques matériels, tels que les photocopieurs couleur, les scanners et les imprimantes, de même que les logiciels de retouches d'images, permettent aux fraudeurs de produire des falsifications de documents de plus en plus sophistiquées.

Cependant, ces mêmes moyens aux mains des responsables de la sécurité des entreprises et des organisations, peuvent également générer des protections efficaces, détecter les documents frauduleux et lutter efficacement contre les fraudeurs. Les techniques de protection couramment utilisées sont les suivantes [31] :

Images holographiques : Un hologramme est une technique d'impression avancée qui crée l'illusion de 3 dimensions sur une surface plane, l'image apparaîtra comme si certaines couleurs changent, certains éléments semblant être à premier-plan et autre en arrière-plan, Tandis que d'autres semblent être plus loin. La théorie générale derrière l'utilisation comme une caractéristique de sécurité est qu'ils Sont difficiles à copier, et qu'ils sont visibles à l'œil sans utiliser d'équipement.

Les hologrammes sont couramment utilisés sur les chèques de voyage, les cartes de crédit et les documents d'identité [56].

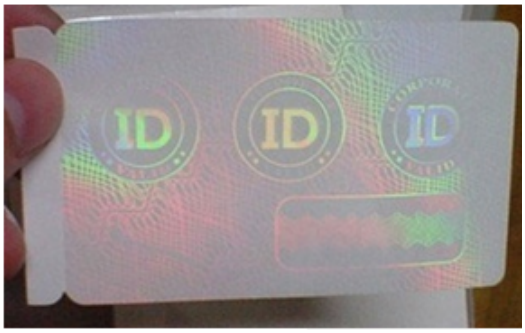


FIGURE 1.10: L'utilisation d'hologramme sur la cartes de crédit.

Le filigrane : Le filigrane est une information supplémentaire incrustée dans l'image et perceptivement indétectable. Cela se traduit d'un travail direct sur l'image, il est impossible de l'enlever sans endommager l'image. Son but principal est de prouver sans aucun doute la propriété de l'image. En tant que fonction de sécurité, ils ne sont pas très efficaces. Les contrefacteurs ont trouvé des contre-mesures, ils ont des méthodes pour reproduire l'effet d'un filigrane [31].

FIGURE 1.11: L'utilisation de filigrane sur le passeport.

Fibres de sécurité aléatoires : Les fibres de sécurité apparaissent dans des motifs aléatoires sur le papier. Ils peuvent être visibles à l'œil nu, ou fluorescents lorsqu'ils sont exposés à la lumière ultraviolette(UV) [20].

FIGURE 1.12: Fibres de sécurité sous la lumière UV.

Zone de lecture automatique : La plupart des passeports et cartes d'identité ont une zone normalisée lisible par la machine (ZLA). La présence de caractères qui ne correspondent pas exactement à cette zone pourrait indiquer que le document peut être falsifié [8].

FIGURE 1.13: L'utilisation de la Zone de lecture sur le passeport.

Fil de sécurité : Est un dispositif de sécurité présent dans de nombreux billets de banque afin de les protéger contre la contrefaçon, Le terme fil de sécurité désigne une fine bande (en plastique, en métal ou constituée d'un autre matériau) qui est incorporée au support au cours du processus de fabrication et fait ainsi office de sécurité supplémentaire [47][59] .

FIGURE 1.14: Fil de sécurité d'un billet de 100 francs suisse.

Impression taille-douce (Intaglio printing) : Est le maître de la défense ouverte pour la sécurité des documents imprimés. Est un processus d'impression qui se traduit par l'encre ayant une sensation élevée et rugueuse qui peut être ressentie par un doigt sur le papier, cette technique d'impression utilise des plaques d'impression rigoureusement sculptées et des presses d'impression extrêmement lourdes pour modifier physiquement la surface du papier imprimé.

Il se trouve sur la couverture intérieure de la plupart des passeports (mais pas de tous). Vous pouvez souvent trouver un motif caché, révélé lorsque la page est affichée dans un angle oblique [56].

FIGURE 1.15: Intaglio printing d'une partie d'un billet d'argent.

Chiffrement(cryptage) : Est un procédé de cryptographie grâce auquel on souhaite rendre la compréhension d'un document impossible à toute personne qui n'a pas la clé de déchiffrement. Ce principe est généralement lié au principe d'accès conditionnel.

Bien que le chiffrement puisse rendre secret le sens d'un document, d'autres techniques cryptographiques sont nécessaires pour communiquer de façon

sûre. Pour vérifier l'intégrité ou l'authenticité d'un document, on utilise respectivement un message d'authentification code (MAC) ou une signature numérique [57] .

1.4 Conclusion

Dans ce chapitre, nous nous sommes concentrés sur un ensemble de concepts concernant la falsification de documents, ceci est basé sur la connaissance de la base du processus de contrefaçon, à savoir le document numérique et comment le numériser afin de connaître son importance et de connaître les caractéristiques les plus importantes utilisées par les contrefacteurs dans le processus de contrefaçon.

Sur cette base, nous allons mentionner dans le chapitre suivant les moyens et les méthodes les plus importants utilisés pour la détection de fraude par la présentation de quelques travaux connexes. Ensuite, nous allons clarifier la méthode que nous avons utilisée dans notre système de détection des faux documents.

CHAPITRE

2

DÉTECTION DES FAUX DOCUMENTS ET APPRENTISSAGE AUTOMATIQUE

2.1 Introduction

Avec l'émergence de dispositifs électroniques bon marché et sophistiqués et la diffusion de l'utilisation par le public et avec l'utilisation fréquente de documents numériques par divers gouvernements et organisations et la tendance à l'administration électronique, il est devenu possible pour n'importe qui de manipuler les documents et les images par la falsification sans difficultés, seulement en utilisant ces appareils et logiciels de contrefaçon qui sont disponibles sur internet. Il est nécessaire donc de recourir à des dispositifs et des programmes capables de distinguer les données par une vérification automatique à la lumière de l'échec de l'utilisation de la vérification manuelle qui ne correspond pas au développement continu des techniques de fraude par les fraudeurs. Par conséquent, l'utilisation de techniques d'apprentissage automatisées et la vérification automatique est la solution la plus efficace pour surmonter ce problème et palier le phénomène de la falsification des documents.

Le but de ce chapitre est de présenter quelques travaux existants dans le domaine de la détection des documents numériques falsifiés et de se concentrer sur l'utilisation des techniques d'apprentissage, pour présenter ensuite une méthode puissante d'apprentissage automatique dite "machine à vecteurs supports (SVM)". Nous décrivons de manière générale le principe de fonctionnement de cette méthode et ses concepts clés tel que l'hyperplan, la marge et le noyau.

2.2 Types des méthodes de détection de falsification

Avec les systèmes qui font confiance aux documents numérisés, il est urgent de mettre en place des systèmes d'aide à la détection et à l'investigation de la fabrication dans les documents et dans la recherche sur ce sujet.
De nombreuses recherches tentent de mettre en œuvre des documents originaux au lieu de documents numérisés, tels que la vérification de signature, la détection de fausses signatures, la falsification d'écriture manuscrite, la falsification de données imprimées et l'authenticité de documents imprimés. De plus, cette tendance révèle que des tentatives de recherche ont été faites dans les cas suivants :

- Distinguer les chèques originaux des chèques contrefaits en utilisant certaines fonctions.
- Détecter la fabrication ou la manipulation de documents Par classification laser et des impressions jet d'encre pour comparer leurs empreintes digitales.
- Reconnaissance et vérification des billets de banque de différents pays en utilisant la société des réseaux de neurones avec l'adressage sur les monnaies bancaires falsifiées.
- Identification de l'écriture falsifiée en utilisant les rides comme une caractéristique tentée avec la comparaison de l'écriture authentique.
- Détecter la fabrication dans le document en mesurant les espaces entre les pixels du texte et l'alignement de la ligne de texte.
- Détecter la fabrication en trouvant la similarité entre les blocs dans l'image, et cette utilisation pour détecter la fabrication de mouvement de copie [30].

Beaucoup de travail a été fait pour identifier la source de l'image telle que l'appareil photo, l'imprimante ou le scanner.

2.3 Quelques travaux existants

De nombreuses recherches ont porté sur la falsification des images mais seules quelques-unes d'entre elles sont liées à la falsification de textes de documents et nous les classerons toutes en fonction des méthodes, des techniques et des caractéristiques qu'elles ont suivies.

2.3.1 Propriétés des imprimantes

- **Johann, Markus,Faisal et Andreas** ont produit un système pour détecter la différence du bord du caractère imprimé et pour distinguer les différents types de sortie d'imprimante. Ils ont créé un jeu de données avec 1200 images de documents par 7 imprimantes à jet d'encre et 13 imprimantes laser différentes. Ensuite, ils ont enregistré les caractéristiques de chaque imprimante pour clarifier les propriétés des bords des lettres et ensuite chercher des lettres différentes pour les distinguer dans le document pour montrer les lettres suspectes.

  L'ensemble du processus qu'ils ont utilisé a été divisé en deux étapes principales : la première étape est l'extraction de caractéristiques, et la deuxième c'est la détection d'anomalies. Au cours de l'extraction des caractéristiques, ils ont classé la technique d'impression en classant les caractéristiques et en examinant les documents. Dans la deuxième étape, ils ont analysé les documents et les documents identifiés, qui ne sont pas imprimés avec la même technique d'impression que la majorité des documents.

  En plus de l'extraction de caractéristiques, deux algorithmes de détection d'anomalies non supervisés différents - Grubbs et kNN - ont été implémentés et testés. Les deux montrent des résultats prometteurs dans différents cas de test. Les raisons possibles des résultats variés dans les différents cas de test ont été examinées et présentées. Ils ont créé l'ensemble de données contient des documents uniques pour chaque imprimante utilisée ,Il existe trois mises en page différentes, avec différentes difficultés pour l'extraction des caractéristiques et le processus de détection des anomalies. Pour chaque imprimante, un jeu de données unique a été créé afin de garantir un système d'extraction de contenu indépendant du contenu.

FIGURE 2.1: Comparaison de la rugosité des bords entre un jet d'encre (à gauche) et une imprimante laser (à droite).

Par conséquent, le but de leur recherche était de créer un système capable de faire la distinction entre différents types d'imprimantes. Il devrait fonctionner avec la détection d'anomalie non supervisée et les documents scannés à une résolution modérément faible [34] .

- Un système de classification a été proposé en 2010 par **Christoph H. Lampert et al** pour analyser la technique d'impression utilisée pour imprimer un document. Chaque lettre d'un document est classée à l'aide d'une machine à vecteurs de support qui a été entrainée pour distinguer les impressions laser des jet d'encre. Une visualisation codée par couleur aide l'utilisateur à interpréter les résultats de classification par lettre.

  L'approche proposée a été utilisée pour détecter la fraude en utilisant une classification par lettre de la technique d'impression. Elle repose sur le fait que chaque imprimante a ses propres caractéristiques visuelles au niveau des caractères imprimées, en particulier dans les zones de bordure des lettres. À partir de la distribution des niveaux de gris dans cette zone de l'image, il peut être décidé pour chaque lettre dans le document quel type d'imprimante l'a créé. La méthode ne nécessite pas de matériel spécifique mais peut fonctionner avec un périphérique d'imagerie grand public standard comme un scanner ou un appareil photo numérique.

  Les lettres imprimées par laser affichent généralement des contours plus nets que ceux imprimés par jet d'encre. La Figure (2.2) représente les zone de contour agrandie de quelques caractères imprimés en jet d'encre et laser.

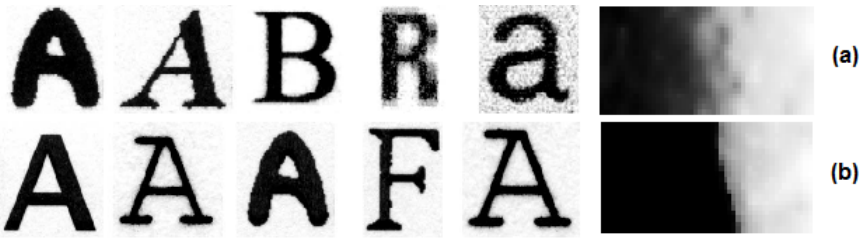


FIGURE 2.2: Exemple de différentes impressions : jet d'encre (a) et impressions laser (b).

La méthode proposée se compose de quatre étapes :
Prétraitement
Les auteurs ont supposé que le matériel est en format 8 niveaux de gris. Ils ont identifié des différents objets (caractères) présents sur la page par une analyse en composantes connexes. Parmi les composantes connexes, ils ont gardé celles ayant approximativement la bonne taille et la bonne forme que les caractères puis on utilisé ces composantes pour extraire les caractéristiques de classification.
Extraction de caractéristiques Ils ont utilisé les caractéristiques suivantes pour former un vecteur de caractéristique : la régularité du contour la de la ligne, l'homogénéité de la surface, le coefficient de corrélation et la texture.
Classification Pour la classification ils ont utilisé une machine à vecteurs supports avec noyau Gaussien (RBF) avec les paramètres $C = 20$ et $\sigma = 1$ (voir les sections suivantes pour plus de détails concernant ces paramètres).
Visualisation Le système n'est pas utilisé uniquement pour prendre des décisions, mais aussi pour guider l'utilisateur à prendre sa propre décision. Cela se fait en présentant à l'utilisateur les résultats de la classification sous forme d'une annotation couleur de l'image originale du document. Le vert représente les caractères imprimés en laser et le rouge ceux en jet d'encre.

- **Shize Shang et al. ont proposé en 2014 une méthode [49]** pour distinguer les documents produits par des imprimantes lasers, par des imprimantes jet d'encre et par des copieurs électrostatiques. L'approche permet de distinguer les documents produits par ces sources en fonction des caractères du document.
  L'utilisation de caractères séparés permet également de détecter et localiser les falsifications de documents créées à l'aide de différents types

d'outils.

Étapes proposées :
Les auteurs ont proposé de suivre les étapes suivantes :

1. Prétraitement : la méthode proposée est basée sur des caractéristiques dérivées de caractères individuels. Un seuil est utilisé pour diviser chaque image en trois parties : la région de texte, la région de bord et la région de fond.
2. Extraction de caractéristiques : les quatre caractéristiques suivantes a été extraites du texte et des régions de bord : L'énergie du bruit dans la région de texte, l'énergie du bruit dans la région de bord, la régularité du contour sur le caractère, le gradient moyen sur la région de bord du caractère.
3. Classification et décision : Après avoir extrait les quatre caractéristiques du caractère,la méthode SVM a été appliqué pour classer chacun des caractères.

La détection de faux documents Les caractéristiques extraites ont été utilisés pour la détection de falsification où un document falsifié contient des caractères provenant de différents types de dispositifs.
La figure (2.3) représente le résultat de détection de la zone falsifié par addition d'une impression à jet d'encre, seule la zone altérée est marquée par des carrés verts.

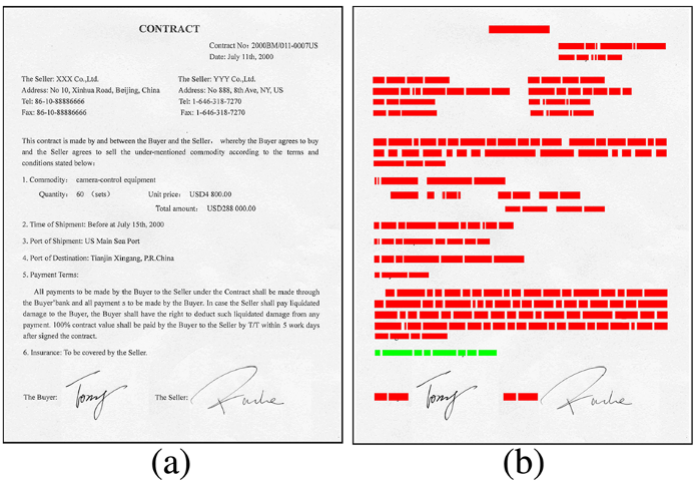


FIGURE 2.3: Résultats de détection d'un document falsifié (un contrat).

2.3.2 Propriétés du niveau de Caractère

- **Romain, Petra, et al [48]** ont présenté une méthode qui peut détecter automatiquement la fabrication en fonction des caractéristiques de certains documents au niveau du caractère. Cette méthode est basée sur la détection de caractères aberrants dans un espace de caractéristique discriminant et sur la détection de caractères strictement similaires.
  Ils ont calculé un ensemble de caractéristiques pour tous les caractères. Ensuite, ils ont classé le caractère si faux ou vrais basé sur une distance entre les caractères de la même classe.

  La méthode utilise des fonctions intrinsèques de document. Elle est basée sur deux techniques de falsification pour la détection d'un caractère frauduleux."Copier et déplacer (Copy-Move)" et "imitation". La méthode proposée passe par les étapes suivantes :

  Étape de Prétraitement : Dans cette étape,les caractères sont extraits et la structure de données hiérarchique est créée pour représenter la structure du document.

  Étapes de Détection de caractères similaires : La détection de caractères Copiés et déplacés ou imité est basée sur la comparaison des formes des caractères. Ils ont conclut que la probabilité de trouver deux caractères ayant exactement la même forme après un processus d'impression et de numérisation est très faible. Ainsi, dans le cas d'une détection de faux caractères sur un document directement envoyé par le fraudeur dans un format numérique, des caractères de forme identique seront considérés comme un indice pour une fraude. Leur objectif est de caractériser les caractères par un vecteur de caractéristique afin de détecter les similitudes de caractères, Les vecteurs de caractéristiques sont comparés à l'aide de la distance euclidienne.

  Étapes de Détection des anomalies : En cas de copier-coller ou d'imitation de faux, différent inexactitudes peuvent être trouvées au niveau du caractères, ils ont intéressé à la détection des erreurs de conception produites lors de la falsification.L'objectif est de détecter des inexactitudes dans la structure du document telles que le désalignement des caractères dans une ligne ou différentes tailles de caractères, positions ou orientations dans le même

mot. Ils ont définit un ensemble d'indices de falsification, regroupés dans un vecteur de caractéristique, calculés pour chaque caractère présent dans le document numérisé à savoir, la taille du caractère, l'axe d'inertie principal du caractère, l'alignement horizontal des caractères.

Ils ont utilisé la distance de Mahalanobis pour la comparaison des caractères. Cette distance est basée sur des corrélations de variables, elle est appropriée car elle intègre la variabilité de la distribution des données et peut être considérée comme une mesure de dissimilarité.
Ils ont mené trois expériences afin d'évaluer leur méthode :

1. Formes de similarités / dis-similarités par la détection des caractères frauduleux en utilisant la comparaison de forme, et ils distinguent trois cas :
   (1) Le fraudeur scanne un document, fraude en copiant et collant un ensemble de caractères et en l'envoyant par courrier électronique.
   (2) Le fraudeur scanne un document frauduleux en copiant et collant un ensemble de caractères et en ajoutant du bruit pour masquer ses manipulations et l'envoyer par e-mail.
   (3) Le fraudeur scanne un document, fraude en copiant et collant un ensemble de caractères appartenant à un autre document avec des propriétés de police différentes et les envoie par courrier électronique.
2. Détection des valeurs aberrantes - récupération des imperfections : la seconde expérience est liée à la détection de l'imperfection due à la manipulation de l'image par le fraudeur.
3. Détection de la fraude de documents : la dernière expérience consiste en une combinaison des deux précédents : les caractères copiés et collés qui sont également affectés par une ou plusieurs des trois imperfections communes.

2.3.3 Propriétés des lignes du texte [35]

- **Jost, Faysal et Thomas** ont décrit une approche pour examiner les caractéristiques intrinsèques des documents pour la sécurité des documents optiques. dont le but est de détecter automatiquement les lignes de texte manipulées ou insérées dans un document en

inspectant leur alignement (gauche, droite ou centre)par rapport aux autres lignes du texte du document. Cela constitue une caractéristique supplémentaire dans le but de développer une puissante boîte à outils pour l'inspection automatique des documents. Ils ont utilisé les lignes du texte et les marges d'alignement extraites. Les statistiques sur les distances entre les lignes de texte et les marges d'alignement servent à identifier les lignes qui auraient pu être falsifiés.

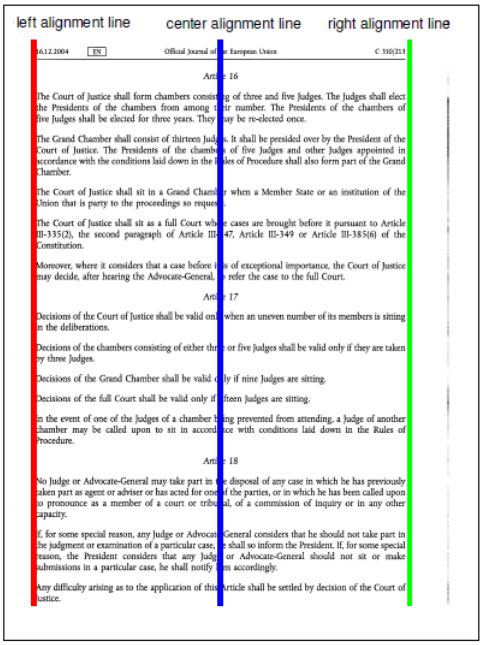


FIGURE 2.4: Visualisation des lignes d'alignement gauche, centrale et droite.

Leur approche proposée fonctionne comme suit : D'abord, les lignes de texte sont extraits. Ensuite, les lignes d'alignement sont calculées. Enfin, les distances entre les lignes de texte et les lignes d'alignement sont calculées et, en fonction de ces distances, il est décidé si une ligne de texte est normalement alignée ou non. Une visualisation de l'approche peut être trouvée sur la figure suivant.

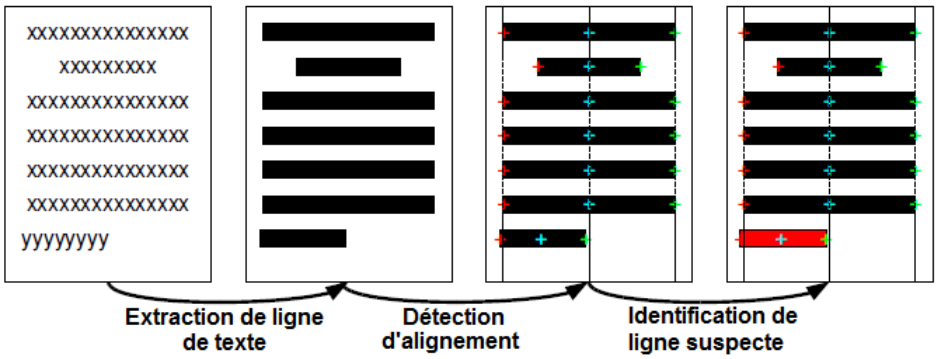


FIGURE 2.5: Visualisation de l'approche de détection des lignes de texte suspectes non alignées.

Il convient de noter que toutes les falsifications ne peuvent pas être détectées en utilisant cette méthode. Si le fraudeur est un expert ou si par hasard l'alignement est exactement le même que pour les autres lignes de texte, l'approche proposée ne sera pas capable de détecter les lignes de texte falsifiées.En fin de compte, de nombreuses fonctionnalités différentes devront être combinées, par exemple. la fonction d'alignement, la fonction d'orientation de la ligne de texte et l'espacement de la ligne de texte.

2.3.4 Propriétés des pixels

- **Tiago, Christian et. Al. [55]** ont suivi une approche différente en mesurant l'algorithme de constance des couleurs qui dépend des pixels spéculaires. Dans leur étude, ils ont détecté automatiquement les régions hautement spéculaires et les isolent. Ils ont proposé de segmenter l'image pour estimer localement la couleur de l'illuminant, recolorer chaque région d'image en fonction de son estimation d'illumination local donne une carte dite illumination comme le montre la figure (2.6) .

FIGURE 2.6: Visualisation de l'approche de détection des lignes de texte suspectes non alignées.

Ils ont utilisé cinq composants principaux pour leur méthode proposée :

— **Estimation locale des densité illuminant :** le but principal ici est de segmenter l'image d'entrée en régions homogènes. Une nouvelle image est créée où chaque région est colorée avec la couleur d'illumination extraite. Ce résultat est appelé carte illuminant.

— **Extraction de visage :** Cette étape peut nécessiter une interaction humaine. Un détecteur de visage automatisé peut être utilisé, puis l'opérateur définit une zone de délimitation dans l'image qui doit être étudiée, puis recadrer chaque zone de délimitation de chaque carte d'illumination, de sorte que la région restante sera la région des estimations d'illumination.

— **Calcul des caractéristiques de l'illumination :** Pour toutes les régions de visage des étapes ci-dessus, ils ont calculé sur les valeurs du carte illuminant. Chacun d'eux code certaines informations pour la classification.

— **Caractéristiques du visage apparié :** Pour une image avec des faces, ils ont construit des vecteurs de caractéristiques communs, constitués de toutes les paires de faces possibles.

— **Classification :** Enfin, ils ont utilisé une approche d'apprentissage automatique pour classer automatiquement les vecteurs de caractéristiques. Ils ont considéré une image comme une image falsifié si au moins une paire de faces de l'image est classée comme illuminée de façon incohérente.

- **Faidi H. Naser Hasan [30]** a proposé une nouvelle méthode pour détecter la fabrication du texte dans les documents numérisés. Cette méthode comporte deux étapes et cinq processus pour détecter la falsification sur un document numérisé. Les trois premiers processus sont utilisés pour pré-traiter les documents pour les deux prochaines étapes. Le quatrième processus est inclus dans la première étape, qui permet d'extraire la caractéristique d'intensité de fréquence maximale à partir des pixels des documents, et le cinquième processus extrait le dégradé de bord pour trouver la différence entre le texte original et le texte falsifié. Le processus final est la coloration et la localisation des pixels suspects.
  Étapes proposé : L'auteur a proposé de suivre les six étapes suivantes :

  — **Sélection de document numérisé :** Au début, une image est

importée avec une bonne résolution, pour obtenir des détails clairs sur les bords. Une image en couleur est obtenue puis convertie en niveau de gris.

— **Conversion en niveau de gris** La méthode des moyennes a été utilisée pour convertir le document couleur en niveau de gris. Elle calcule simplement la moyenne des valeurs.
— **Suppression du bruit :** Le deuxième processus est l'utilisation d'un filtre médian pour supprimer le bruit des documents numérisés.
— **Seuillage :** Le seuillage mappe l'image en niveau de gris à une image binaire. Après l'opération de seuillage, l'arrière-plan de l'image du document sera supprimé.
— **Extraction de l'intensité de fréquence maximale** L'intensité des mots fabriqués diffère de l'intensité du mot original dans le document numérisé, et il a souvent une intensité solide. Parce que les mots ou les lettres fabriqués sont fabriqués après l'impression du document original et ont un autre environnement tel que l'application de traitement de documents et le type de papier, spécialement, chaque type de papier a des spécifications internes. La figure suivant montre les différents niveaux d'intensité des niveaux de gris.

FIGURE 2.7: Exemple de niveau d'intensité de gris, où l'intensité de (A) est 0, (B) est 127, et (C) est 195.

Le nombre de pixels des mots fabriqués est souvent inférieur au nombre de pixels pour les mots originaux. Par conséquent, les pixels qui ont la fréquence d'intensité la plus élevée ont été isolé dans un document, ce qui a permis d'extraire les mots ou les lettres fabriqués dans un document fabriqué. Quatre étapes ont été utilisées dans cette phase pour extraire des mots qui ont une intensité de fréquence maximale.

— **Pré-traitement :** Certains processus ont été appliqués sur l'image du document, ces processus sont l'histogramme et le seuil.
— **Obtenir l'intensité de fréquence maximale :** Après tous

les processus précédents on obtient un tableau de niveau de gris pour l'image du document et le nombre d'itérations pour chaque niveau dans l'image du document.
L'auteur a extrait l'intensité de fréquence maximale de ce tableau, et construit un nouveau tableau qui a les positions d'intensité maximale dans le tableau de documents.

— **Supprimer les pixels d'intensité de fréquence maximale :** Il a enlevé tous les pixels qui ont une valeur d'intensité de fréquence maximale, en fonction de la position des pixels.
S'il y a des mots ou des lettres qui apparaissent ou disparaissent visuellement dans l'image, cela signifie que ce sont des mots suspects.
Si aucun mot n'apparaissait ou ne disparaissait visuellement, il trouve la prochaine valeur d'intensité de fréquence maximale à partir du tableau d'histogrammes et construit une nouvelle position d'intensité maximale dans le tableau, puis supprime les pixels de celui-ci à nouveau. Il répète ces étapes jusqu'à l'apparition de quelques mots ou lettres visuellement.

FIGURE 2.8: Résultat de la suppression des pixels d'intensité de fréquence maximale.

— **Extraire le dégradé de bord :** un filtrage maximal est utilisé pour trouver la différence entre les bords du texte fabriqué et non-fabriqué.

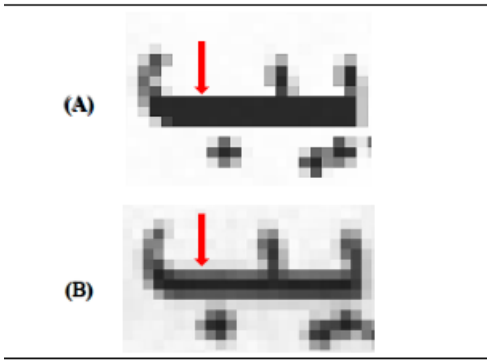


FIGURE 2.9: (A) Dégradé de bord pour le texte fabriqué. (B) Dégradé de bord pour le texte non fabriqué.

Dans cette étape, l'auteur applique un filtre Max sur l'image du document numérisé, pour faire en sorte que les pixels qui ont une intensité élevée soient répartis sur tous les mots en pixels. Tous les résultats de cette étape apparaissent visuellement.

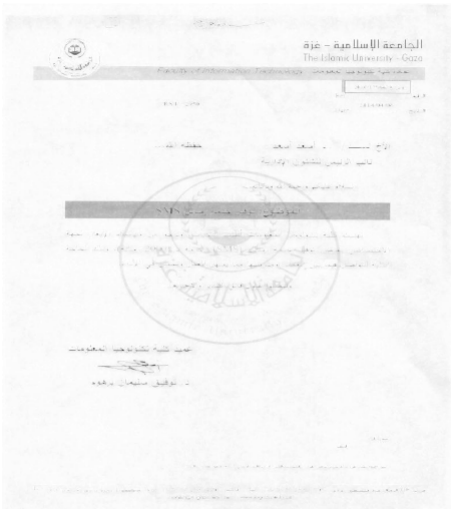


FIGURE 2.10: Résultat de l'application d'un filtre Max sur l'image du document numérisé.

Combiner les résultats : L'auteur a obtenu un tableau de positions de pixels suspectés. D'autre part, le résultat de la deuxième étape est

une image avec des mots suspects. Il a combiné les résultats précédents pour obtenir le résultat final de sa méthode, en changeant la couleur du pixel suspecté en une couleur rouge comme le montre la figure suivante.

FIGURE 2.11: Résultats finaux de la méthode proposée.

2.3.5 Propriétés de scanner

- **Ramzi M. Abed [11]** a proposé une nouvelle technique pour détecter les documents scannés altérés. Cette technique est basée sur l'identification du scanner utilisé en utilisant ses caractéristiques intrinsèques.

  Le système proposé commence par un document numérisé et extraire toute les lettres "e" dans le document, car c'est la lettre la plus fréquente dans la langue anglaise (leur système a été proposé pour les documents en anglais). Ensuite, le système extrait un ensemble de caractéristiques de chaque groupe de caractères "e", puis il forme un vecteur de caractéristique pour eux en divisant le document numérisé testé en blocs . Un ensemble différent de caractéristiques est extrait pour chacun de ces blocs. Chacun de ces vecteurs de caractéristiques a été ensuite testé et classé séparément à l'aide du classificateur SVM (Support Vector Machine), qui décide à la fin si l'image de document numérisée testée est authentique ou altérée.

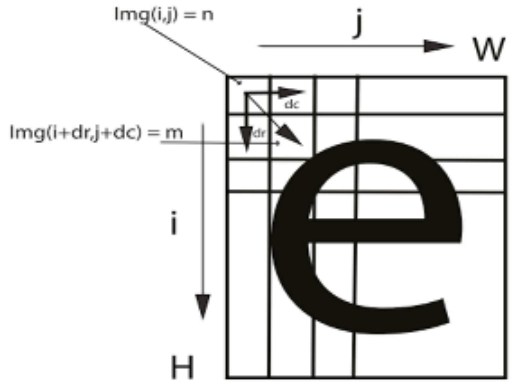


FIGURE 2.12: Image numérisée pour le caractère "e".

Le système proposé est développé sur la base de l'identification du scanner utilisé pour scanner le document testé, car cette technique dépend de l'identification de la signature du scanner. La qualité des bords des caractères dans les documents numérisés varie en fonction du scanner utilisé dans le processus de numérisation. Plus précisément, les scanners à haute résolution produisent des traits noirs plus nets avec des bords plus nets, tandis que les scanners à basse résolution produisent des caractères représentés par des lignes noires consistant en des variations du noir au gris et les bords de ces caractères sont plus graduels.

Ces différences entraîneront des changements dans les caractéristiques de texture. Donc la qualité des contours des caractères dans les documents numériques varie en fonction du scanner utilisé dans la numérisation.

Le sens du processus de numérisation peut produire des textures spécifiques car il peut provoquer une vacillation du niveau de gris dans les caractères numérisés. Ces textures extraites de la fluctuation du niveau de gris peuvent être représentées par la matrice de co-occurrence de niveaux de gris, et elles sont très robustes pour identifier des documents numériques et sont donc très robustes pour détecter des régions altérées sur les documents numériques.

2.3.6 Propriétés des images [31]

L'année dernière (2017), un mémoire sur la détection des documents numérique administratifs falsifiés utilisant les machines à vecteurs supports binaires a été présenté par l'étudiante H. Benhamza. L'objectif du travail était d'étudier la nature des images scannées et les méthodes utilisées pour leur falsification, ensuite de proposer des techniques permettant la détection des documents falsifiés en se basant sur des règles obtenus à travers l'apprentissage automatique.

Globalement, le système proposé suit les étapes suivantes pour arriver à la détection des faux documents :

- **Acquisition d'un document numérique :** Les documents numériques ont été importés sous forme d'images couleurs avec une bonne résolution ensuite transformées en niveau de gris.
- **Élimination du bruit :** Un filtrage médian a été appliqué pour l'élimination du bruit. La valeur de chaque pixel est remplacée par la valeur médiane de ses voisins.
- **Calcul des informations du fond :** Par :
  - Découpage de l'image en des blocs
  - Calcul de la couleur moyenne dans chaque bloc.
  - Calcul de la couleur fréquente dans chaque blocs
  - Calcul de la couleur moyenne du fond de l'image
  - Calcul de la couleur fréquente du fond de l'image
  - Calcul de l'écart-type de la couleur du fond
- **Calcul des informations de la zone du texte :** Par :
  - Délimitation de la zone du texte
  - Découpage du texte en des mots ou des caractères (blocs).
  - Calcul de la couleur de chaque mot comme étant la couleur fréquente après élimination du fond
  - Calcul de l'écart-type de la couleur du texte

FIGURE 2.13: Le découpage du texte.

- **Calcul des informations du cachet**par :
  - Délimitation de la zone du cachet par détection de la couleur bleue ou rouge dans l'image couleur
  - Élimination de la couleur du cachet correspondant aux pixels rouges et bleus dans l'image couleur

— Détermination de la couleur du fond du cachet comme étant la couleur fréquente des pixels restant

FIGURE 2.14: La détection du cachet.

— **Combinaisons des informations précédentes :** pour obtenir un vecteur de caractéristiques.
— **Construction d'une base de caractéristique :** en appliquant les étapes précédentes sur un ensemble d'exemples (documents) dont nous disposons de leurs états, authentique ou falsifié.
— **Construction d'un modèle de décision :** en utilisant la base des caractéristiques et le classifieur SVM binaire.
— **l'utilisation :** Utilisation du modèle construit pour la détection des documents falsifiés.

2.4 Limites

En résumé, tous les chercheurs précédents ont proposé des techniques de vérification et détection de la falsification en extrayant différentes caractéristiques du document en fonction de l'empreinte de l'imprimante, du papier, de scanner, des caractères ou des propriétés des pixels.

Bien que beaucoup de ces techniques soient très prometteuses et novatrices, elles ont toutes des limites, tel que la non prise en charge de la langue arabe, pour les chercheurs qui dépendent des caractéristiques du contexte et mesurent les espaces pour classifier les caractères originaux et falsifiés, et la mauvaise classification dans le cas d'une faible quantité de falsification comme changer un caractère ou une partie de caractère, aussi la restriction sur une langue particulière (l'anglais), ou condamnation d'un document à partir d'un seul caractère.

Nous concluons, donc, des travaux présentés dans le domaine de la découverte de la falsification dans les documents numériques que les recherches sont variée où chacun vise une propriété particulière utilisée en falsification.

Cependant, ils méritent d'être encore améliorés pour atteindre des solutions efficaces adaptées à tous les types de falsification.

Il est aussi important d'approfondir l'analyse des documents et aller au delà des textes pour analyser les logos, les signatures, les cachets, ...etc. Les résultats de telles analyses doivent être ensuite utilisés par des techniques d'apprentissage automatique appropriées pour automatiser la tâche de vérification et limiter l'intervention humaine.

2.5 Apprentissage automatique

L'apprentissage automatique est un des champs d'étude de l'intelligence artificielle. Il fait référence à la capacité d'un système à acquérir et intégrer de façon autonome des connaissances.
Cette notion englobe toute méthode permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle partiel ou moins général, soit en créant complètement le modèle.
L'apprentissage automatique fait référence au développement, l'analyse et l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer et de remplir des tâches associées à une intelligence artificielle grâce à un processus d'apprentissage. Cet apprentissage permet d'avoir un système qui s'optimise en fonction de l'environnement, les expériences et les résultats observés [45].

2.5.1 Définition

L'apprentissage est le processus de construire un modèle général à partir d'un ensemble de données (exemples). Le modèle sera créé à partir l'amélioration d'un modèle partiel ou général [54].

Il en existe de nombreuses approches. Elles sont généralement regroupées en trois familles qui se différencient par le type d'information dont dispose le système pour apprendre et le protocole avec lequel il interagit avec son environnement : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

2.5.2 Types d'apprentissage

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient :

— **L'apprentissage supervisé :** Si les classes sont prédéterminées et les exemples connus, le système apprend à classer selon un modèle de classement ; on parle alors d'apprentissage supervisé .
— **L'apprentissage non supervisé :**Quand le système ne dispose que d'exemples, mais sans étiquettes, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé. Dans ce cas le but d'apprentissage est de grouper les exemples selon leurs attributs en basant sur la notion de la similarité [45].

2.5.3 Domaines d'application de l'apprentissage automatique

Les principaux domaines d'applications de l'apprentissage automatique sont les fouilles de données et l'intelligence artificielle.

— La fouille de données (Data Mining, en anglais) est le processus d'extraction de la connaissance , il consiste à sélectionner les données à étudier à partir de bases de données (hétérogènes ou homogènes), à épurer ces données et enfin à les utiliser en apprentissage pour construire un modèle.
— L'intelligence artificielle, la vision par ordinateur, la robotique, l'analyse et la compréhension des images, la reconnaissance de formes, reconnaître des objets dans les vidéo et extraire des contenus sémantiques des images sont autant d'applications qui requièrent la construction de modèles par apprentissage automatique [54].

2.5.4 Classification

Les méthodes de classification ont pour but d'identifier les classes aux quelles appartiennent des objets à partir de certains paramètres descriptifs.

Elles s'appliquent à un grand nombre d'activités humaines et conviennent en particulier au problème de la prise de décision automatisée. La procédure de classification sera extraite automatiquement à partir d'un ensemble d'exemples. Un exemple consiste en la description d'un cas avec la classification correspondante. Un système d'apprentissage doit alors, à partir de cet ensemble d'exemples, extraire une procédure de classification, il s'agit en d'extraire une règle générale à partir des données observées. La procédure générée devra classifier correctement les exemples de l'échantillon et avoir un bon pouvoir prédictif pour classifier correctement de nouvelles descriptions. Les méthodes utilisées pour la classification sont nombreuses, citons : la les Séparateurs à Vastes Marges(SVM), les Réseaux de Neurones, etc [39].

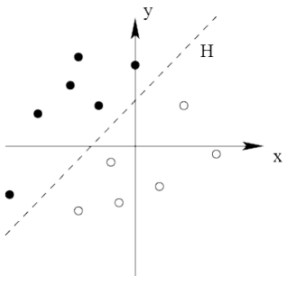
2.5.5 Machines à Vecteurs Support (SVM)

Les machines à vecteur support se situent sur l'axe de développement de la recherche humaine des techniques d'apprentissage. Les SVMs sont une classe de techniques d'apprentissage introduite par Vladimir Vapnik au début des années 90, elles reposent sur une théorie mathématique solide. Les SVMs sont dans leur origine utilisées pour la classification binaire et la régression. Aujourd'hui, elles sont utilisées dans différents domaines de recherche et d'ingénierie tel que le diagnostic médical, le marketing, la biologie, la reconnaissance de caractères manuscrits et de visages humains [24].

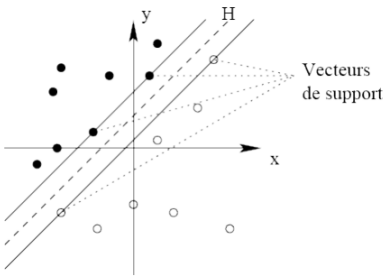
Principe de la technique SVM : L'idée principale des SVM consiste à projeter les données dans un espace de plus grande dimension appelé, espace de caractéristiques, afin que les données non linéairement séparables dans l'espace d'entrée deviennent linéairement séparables dans l'espace de caractéristiques. En appliquant dans cet espace la technique de construction d'un hyperplan optimal séparant les deux classes, on obtient une fonction de classification qui dépend d'un produit scalaire des images des données de l'espace d'entrée dans l'espace des caractéristiques.

Ce produit scalaire peut être exprimé, sous certaines conditions, par des fonctions définies dans l'espace d'entrée, qu'on appelle les noyaux.
Ce multiple choix de noyaux rend les SVM plus intéressantes et surtout plus riches puisqu'on peut toujours chercher de nouveaux noyaux qui peuvent être mieux adaptés à la tâche qu'on veut accomplir. Les trois noyaux les plus utilisés sont : le noyau linéaire, le noyau polynomial et le noyau gaussien .

Pour deux classes d'exemples donnés, le but de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan. Dans le schéma qui suit, on détermine un hyperplan qui sépare les deux ensembles de points.



Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support.



2.5.5.1 Hyperplan

On appelle hyperplan séparateur un hyperplan qui sépare deux classes, en particulier il sépare leurs points d'apprentissage. Il est évident qu'il existe une multitude d'hyperplan valide mais la propriété remarquable des SVM est que cet hyperplan doit être optimal. Nous allons donc en plus chercher parmi les hyperplans valides, celui qui passe « au milieu » des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan optimal.

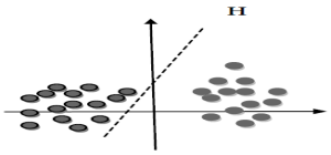


FIGURE 2.15: L'hyperplan H qui sépare les deux ensembles de points

Problème linéairement séparable : Un problème de discrimination est dit linéairement séparable lorsqu'il existe une fonction de décision linéaire (appelé aussi séparateur linéaire) est un hyperplan qui sépare les données s'appelle un (hyperplan de séparation).

La forme général d'un hyperplan est $\langle w, x\rangle + b = 0$, et la fonction décisionnelle pour un hyperplan, $f(x) \equiv \langle w, x\rangle + b$, peut être utilisé comme une règle de classification en assignant une observation à la classe positive $(y = 1)$ pour $f(x) \geq 0$ et à la classe négative $(y = -1)$ [15].

L'hyperplan qui sépare de manière optimale les données est celui qui minimise :

$$\phi(w) = \frac{1}{2}\|w\|^2$$

Hyperplan de séparation optimale généralisée : En général, les données d'entrainement ne sont pas linéairement séparables. Il existe deux approches pour généraliser le problème, qui dépend de la connaissance préalable du problème et une estimation du bruit sur les données. Dans le cas où il est prévu (ou peut-être même connu) qu'un hyperplan peut séparer correctement les données, une méthode d'introduction d'une fonction de coût supplémentaire associée à une mauvaise classification est appropriée [15]. Dans ce cas on doit définir l'hyperplan qui va minimiser l'erreur.

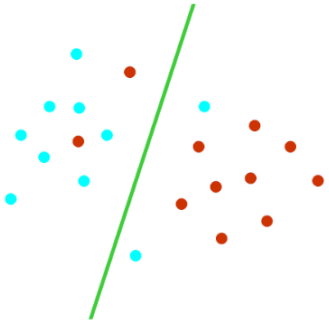


FIGURE 2.16: Hyperplan de séparation optimale généralisée.

2.5.5.2 Vecteurs supports

Pour une tâche de détermination de l'hyperplan séparable des SVM est d'utiliser seulement les points les plus proches (les points de la frontière entre les deux classes des données) parmi l'ensemble total d'apprentissage, ces points sont appelés vecteurs supports [45].

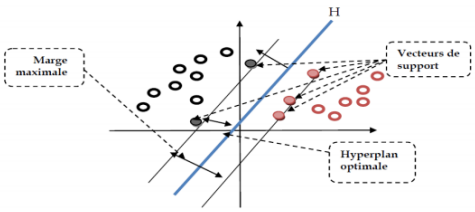


FIGURE 2.17: L'hyperplan H optimal, vecteurs supports et marge maximale

2.5.5.3 Marge

Il existe une infinité d'hyperplans capable de séparer parfaitement les deux classes d'exemples. Le principe des SVM est de choisir celui qui va maximiser la distance minimale entre l'hyperplan et les exemples d'apprentissage (la distance entre l'hyperplan et les vecteurs supports), cette distance est appelée la marge [45].

2.5.5.4 Maximiser la marge

Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple. De plus, si l'on trouve le classificateur qui se comporte le mieux vis-à-vis des données d'apprentissage, il est clair qu'il sera aussi celui qui permettra au mieux de classer les nouveaux exemples. Dans le schéma qui suit, la partie droite nous montre qu'avec un hyperplan optimal, un nouvel exemple reste bien classé alors qu'il tombe dans la marge. On constate sur la partie gauche qu'avec une plus petite marge, l'exemple se voit mal classé [45].

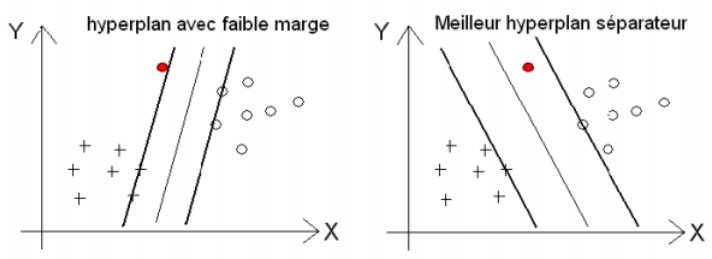


FIGURE 2.18: classification d'nouvel exemple dans deux cas

2.5.5.5 SVMs binaires

C'est Le cas le plus simple est celui où les données d'entrainement viennent uniquement de deux classes différentes (+1 ou -1), on parle alors de classification binaire. L'idée des SVMs est de rechercher d'un hyperplan (droite dans le cas de deux dimensions) qui sépare le mieux ces deux classes [24].

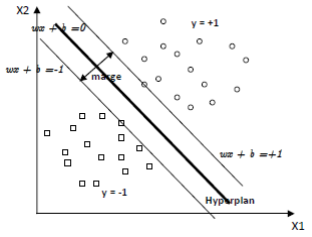


FIGURE 2.19: SVM binaire

2.5.5.6 SVMs multiclasse

Les méthodes des machines à vecteur support multiclasse réduisent le problème multiclasse à une composition de plusieurs hyperplans permettant de tracer les frontières de décision entre les différentes classes. Ces méthodes décomposent l'ensemble d'exemples en plusieurs sous ensembles représentant chacun un problème de classification binaire. Pour chaque problème, Un hyperplan de séparation est déterminé par la méthode SVM binaire. On construit lors de la classification une hiérarchie des hyperplans binaires qui est parcourue de la racine jusqu'à une feuille pour décider de la classe d'un nouvel exemple [24].

2.5.5.7 SVMs mono classe

Dans les machines à vecteur support binaires et multiclasse précédentes, nous avons toujours des exemples positifs et d'autres négatifs c-à-d des exemples et des contre-exemples. De telles informations ne sont pas disponibles dans tous les cas d'application. Parfois, il est très coûteux, voire impossible, de trouver des contre-exemples qui représentent réellement la classe négative.

Prenons l'exemple de reconnaissance d'une catégorie particulière de pièces par un robot dans une usine, il est facile d'avoir des exemples suffisants de cette pièce, mais il est difficile d'avoir des exemples de toutes les pièces différentes. Il est souhaitable, dans de tels cas, d'avoir un modèle de décision permettant de reconnaître autant d'exemples possibles de cette catégorie et de rejeter tous les autres. Ce problème est souvent appelé détection des nouveautés, puisque le modèle de décision connaît un ensemble d'exemples et détecte tous ce qui est nouveau(étranger).

Pour la classification SVM mono classe, il est supposé que seules les données de la classe cible sont disponibles. L'objectif est de trouver une frontière

qui sépare les exemples de la classe cible du reste de l'espace, autrement dit, une frontière autour de la classe cible qui accepte autant d'exemples cibles que possible. Cette frontière est représentée par une fonction de décision positive à l'intérieur de la classe et négative en dehors. La figure suivant représente, en deux dimensions, un cas de séparation d'une classe de toute autre classe [24].

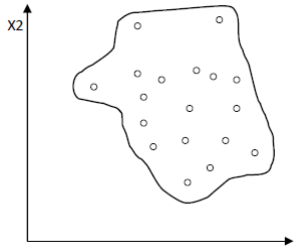


FIGURE 2.20: Séparation des exemples d'une classe du reste de l'espace

pour résoudre ce cas de problème ,la technique SVM mono-classe utilise le même modèle binaire avec une astuce en plus ; l'origine de l'espace est considérée comme étant la seule instance de la classe négative. le problème revient ,donc,à trouver un hyperplan qui sépare les exemples de la classe cible de l'origine ,et ce qui maximiser la marge entre les deux.

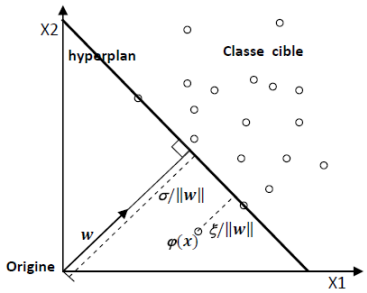


FIGURE 2.21: SVM mono classe à marge maximale .

Le problème est modélisé par le problème primal de programmation quadratique dont l'objectif est de maximiser la marge et minimiser les erreurs de classification.

2.5.6 Avantages et inconvénients des SVM

— **Avantages**

SVM est une méthode de classification intéressante car le champ de ses applications est large, parmi ses avantages nous avons :
— Un grand taux de classification et de généralisation par rapport aux méthodes classiques.
— L'algorithme est robuste face aux changements d'échelle .
— Résultat en général équivalents et souvent meilleurs.
— Décision rapide. la classification d'un nouvel exemple consiste à voire le signe de la fonction de décision.
— Les exemples sont comparés juste avec le supports vecteur et non pas avec tout les exemples d'apprentissage .

— **Inconvénients**
— Grand quantité d'exemples en entrées implique un calcul matriciel important.
— Temps de calcule élevé lors d'une régularisation des paramétrés de la fonction noyau
— Classification binaire d'où la nécessité d'utiliser l'approche un-contre-un.

2.6 Conclusion

SVM est une méthode de classification qui montre de bonnes performances dans la résolution de problèmes variés. Cette méthode a montré son efficacité dans de nombreux domaines d'applications tels que le traitement d'image, la catégorisation de textes ou le diagnostique médical et ce même sur des ensembles de données de très grandes dimensions. Dans ce chapitre, nous avons présenté certains travaux sur la détection automatiques des documents falsifiés. Nous avons présenté également la méthode des machines à vecteurs supports et ses détails pour l'apprentissage automatique.

CHAPITRE

3

CONCEPTION DU SYSTÈME

3.1 Introduction

Les documents numériques sont faciles à manipuler et à modifier en raison de la disponibilité de puissants logiciels de traitement et d'édition d'images. Le présent travail vise à étudier la nature des documents administratifs scannées et à analyser l'homogénéité de leurs pixels qui peut être perturbée en cas de sa modification. Ensuite de proposer des techniques permettant la détection des documents falsifiés en se basant sur des règles obtenus à travers l'apprentissage automatique.
Dans ce chapitre nous présentons la conception de notre système en commençant par sa conception générale puis sa conception détaillée en spécifiant les différents éléments le composant et précisant son fonctionnement.

3.2 Description et Objectif du système

Notre objectifs est de concevoir un système de classification des document administratifs numérique par la méthode SVM mono classe. Dans notre système, nous avons commencé par la détection des opération de type "Copier-

coller". Dont une partie de l'image elle-même est copiée et collée dans une autre partie de la même image. Ceci est généralement effectué avec l'intention de faire disparaître un objet de l'image en le recouvrant d'un segment copié d'une autre partie de l'image ou pour ajouter un cachet, une signature, ou n'importe quel objet ou même pour modifier un texte.

Nous avons trouvé que toute falsification de ce type introduit une non homogénéité entre les parties de l'image originale et la partie collée, ce qui provoque des changements sur le fond de l'image cible. Cette corrélation peut être utilisée comme une base pour une détection réussie de ce type de falsification.

Pour détecter cette opération, nous calculons les couleurs des fonds des différents blocs de l'image et considérer les blocs ayant un grand écart par rapport à la moyenne comme des blocs suspects.

Nous avons travaillé sur des images de documents administratifs contenant un en-tête, un texte, une signature et un cachet. Pour cela nous avons analysé et vérifié chacun d'entre eux séparément.

Pour la vérification des cas de changement ou l'ajout du texte dont la modification produit des caractères de couleurs différente avec un fond différent dans l'image cible, Nous avons utilisé des calculs pour vérifier l'homogénéité des caractères et détecter les mots étranges.
La confirmation de l'homogénéité du fond du cachet avec le fond global de l'image permet de détecter la modification dans l'image, par ce que les conditions d'acquisition des images source et cible sont absolument différentes.

3.3 Conception globale du système

Dans notre système de détection des faux documents, nous avons suivi une série d'étapes que nous mentionnerons ici globalement afin de comprendre la manière générale dont nous avons atteint notre objectif.

- L'acquisition d'un document numérique et le transformer en niveau de gris, puis faire le filtrage.

- Faire des calcules sur le fond pour l'extraction des informations :
  1. Diviser l'image en des blocs ;
  2. Calculer la couleur moyenne dans chaque bloc ;

3. Calculer la couleur moyenne du fond de l'image ;
4. Calculer la couleur fréquente pour chaque blocs ;
5. Calculer la couleur fréquente du fond de l'image
.

- Détecter la zone du cachet dans l'image et calcul de ses informations par :
  1. Déterminer la zone du cachet par la détection de la couleur bleue ou rouge dans l'image couleur.
  2. Transformer la zone de cachet sélectionnée en niveau de gris.
  3. Calculer la première et la deuxième couleur fréquente dans la zone du cachet pour déterminer la couleur du fond du cachet comme étant la deuxième couleur fréquente .
  4. calculer la couleur moyenne de la zone du cachet.
- Détecter la zone du texte et calculer ses informations par :
  1. Délimitation de la zone du texte par la détection des couleur du texte dans l'image
  2. Découper le texte en des blocs de texte (syllabes ou mots)
  3. Calculer la première couleur fréquent de chaque bloc et la deuxième comme étant la couleur du fond
  4. Calculer la couleur moyenne de chaque bloc.
- Combiner toutes les informations obtenues précédentes pour construire un vecteur de caractéristiques.
- Appliquer les étapes précédentes sur tous les document scannés pour construire une base de caractéristiques pour les documents authentique.
- Faire l'apprentissage par le classifieur SVM mono classe sur la base des caractéristiques construite.
- Construction d'un modèle de décision et son test.
- Utilisation du modèle construit pour la détection des documents falsifiés.

Le schéma général du système conçu est présenté dans la figure (3.1) et la figure (3.2), où notre méthode proposée est constituée de deux phase principales :

— Phase de construction de modèle.
— Phase d'utilisation.

Les figures suivantes schématisent ces phases.

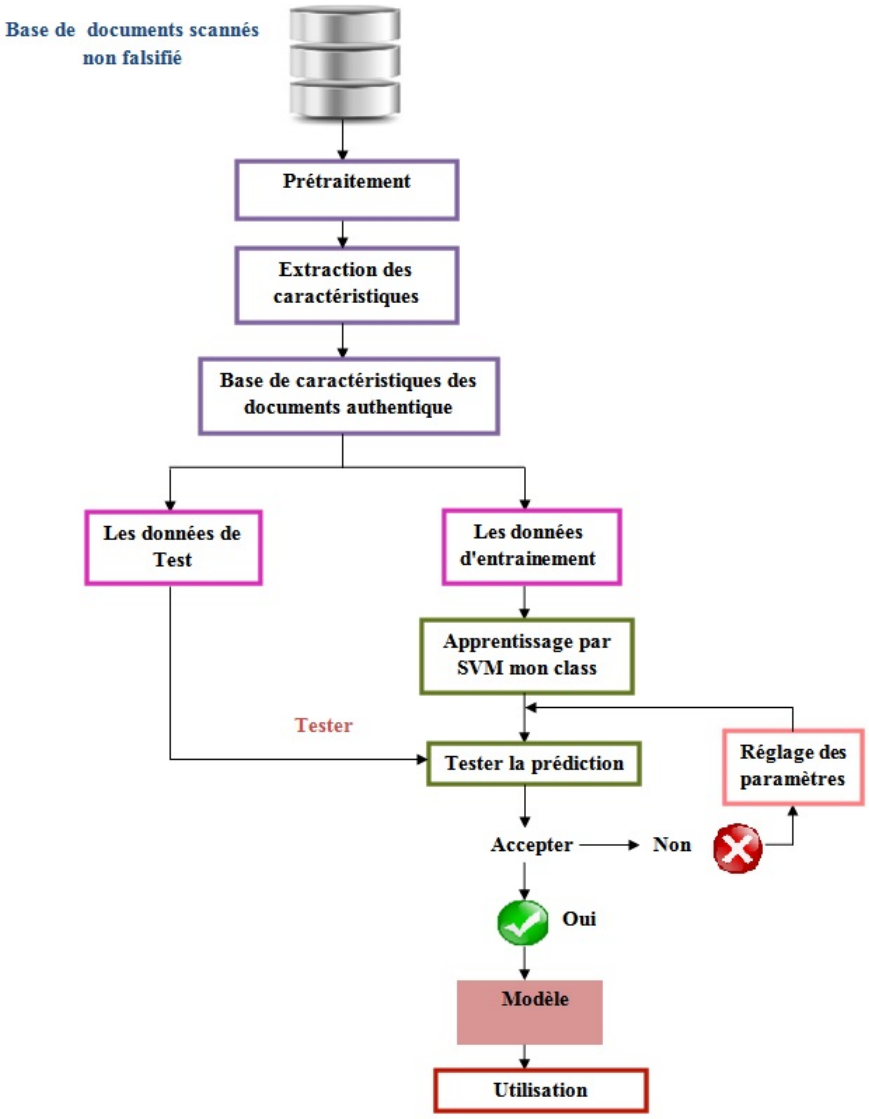


FIGURE 3.1: Conception de la phase de construction de modèle.

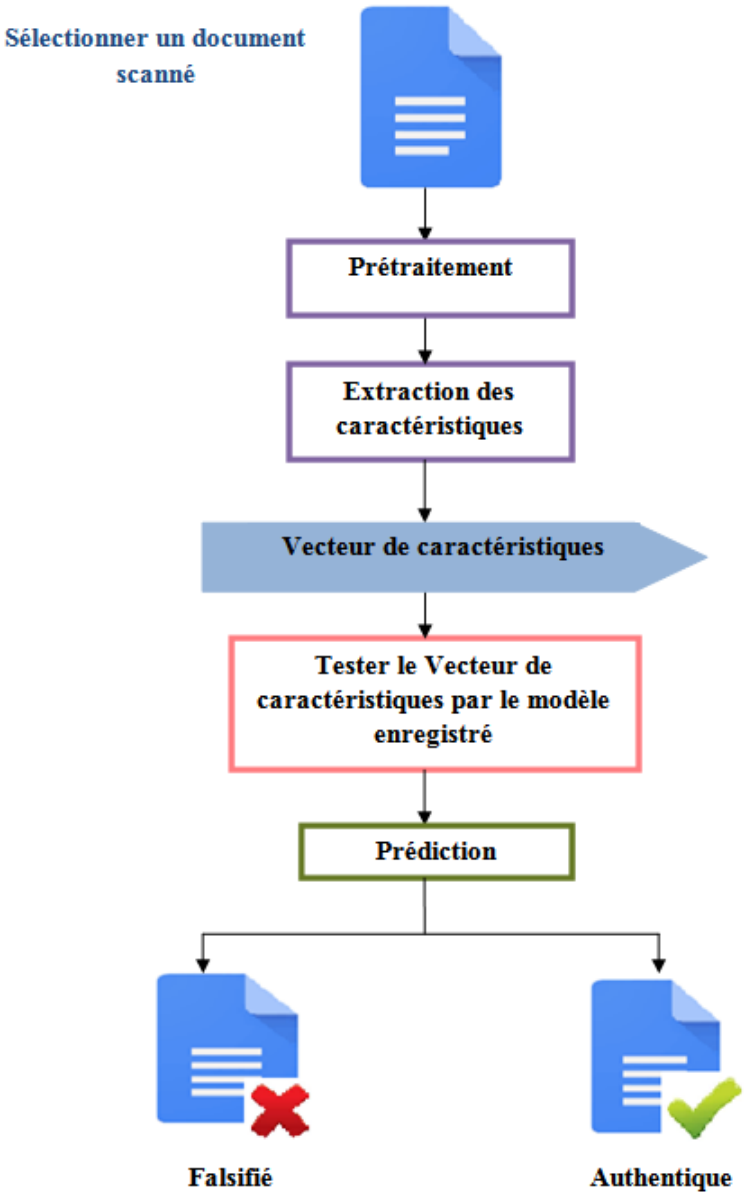


FIGURE 3.2: Conception de la phase d'utilisation.

3.4 Conception détaillée du système

3.4.1 Phase de construction de modèle

3.4.1.1 Sélection des documents numérisés

La première étape que nous avons faite a été d'importer des documents administratifs en les numérisant à l'aide d'un scanner. Nous avons utilisé des images numériques avec une bonne résolution pour obtenir des détails précis sur les bords et l'intensité des mots. Ces documents sur lesquels nous avons travaillés sont tous des documents authentiques et nous n'avons laissé qu'une partie pour les tests dont nous les avons falsifiés pour vérifier la validité des résultats plus tard.

3.4.1.2 Prétraitement

Après la sélection d'un document scanné, nous appliquons l'ensemble d'opérations suivant pour extraire les caractéristiques.

3.4.1.2.1 Conversion les images en niveau de gris et filtrage : Nous avons converti les images en niveau de gris pour garder uniquement les informations de luminance. Puis, nous avons utilisé le filtrage pour réduire le bruit dans les image résultant parfois des scanners pendant le processus de numérisation ou d'imprimantes de mauvaise qualité.

3.4.1.2.2 Analyse du fond : C'est la première étape dans l'analyse d'image et l'extraction des caractéristiques. Où nous divisons l'image en des blocs de même taille puis calculons la couleur fréquente de chaque bloc, sa couleur moyenne,la couleur moyenne générale du fond de l'image et sa couleur fréquente.

3.4.1.2.3 Analyse du cachet : Dans cette étape, nous délimitons la zone du cachet dans l'image couleur. Pour cela, nous cherchons la couleur bleue ou rouge dans l'image afin de détecter l'emplacement du cachet. Après la détection de cette zone, nous la convertissons en niveau de gris pour calculer sa première et deuxième couleur fréquentes. Nous calculons aussi la

couleur moyenne. Le but de tout ces valeurs est de comparer le fond du cachet avec le fond du document pour détecter tout changement survenu au niveau du cachet.

FIGURE 3.3: La détection du cachet.

3.4.1.2.4 Analyse du texte : Afin d'analyser le texte dans l'image en niveau de gris, nous parcourons toutes les lignes de texte et les découpons en morceaux basés sur le début et la fin de l'écriture. Sur la base de ce principe nous découpons le texte en des syllabes (mots et caractères) pour calculer la première et la deuxième couleur fréquente de chaque mot. Nous calculons aussi la couleur moyenne.

Cela permet de vérifier l'homogénéité de la couleur de l'écriture, car l'intensité de l'écriture représente un facteur important qui affecte la cohérence et l'homogénéité des mots dans le texte. Nous avons remarqué que l'intensité des mots ou des caractères originaux dans les documents est très différente de l'intensité des mots ajoutés par les outils de traitement d'image ou de texte. Cela est dû au fait que la couleur des caractères originaux est influencées par les caractéristiques de l'imprimante et du scanner utilisés pour leur production.

3.4.1.3 Extraction des caractéristiques

Après avoir analysé l'image des trois aspects, le fond, le cachet et le texte, nous allons maintenant extraire les informations dont nous avons besoin et qui représentent les caractéristiques requises par notre système pour détecter la fraude dans les documents numériques. Ces caractéristiques sont :

3.4.1.3.1 Caractéristiques extraites du fond

— La couleur fréquente de chaque bloc dans l'image.
— La couleur moyenne de chaque bloc dans l'image.
— La couleur moyenne/fréquente du fond de l'image.

3.4.1.3.2 Caractéristiques extraites du cachet

— La première couleur fréquente de la zone du cachet (la couleur du fond de cachet)
— La deuxième couleur fréquente de la zone du cachet (la couleur du cachet) .
— La couleur moyenne de la zone du cachet.

3.4.1.3.3 Caractéristiques extraites du texte

— La première couleur fréquente de chaque mots/caractère dans l'image.
— La deuxième couleur fréquente de chaque mots/caractère dans l'image.
— La couleur moyenne de chaque mots/caractère.

3.4.1.4 Construction de la base

Après l'extraction des caractéristiques dans l'étape précédente, nous construisons maintenant la base que nous utiliserons dans la phase d'apprentissage, en collectant les caractéristique extraites et nous les enregistrons dans un vecteur qu'on appelle vecteur de caractéristiques. Ensuite, le système combine les vecteurs obtenus dans un seul vecteur.

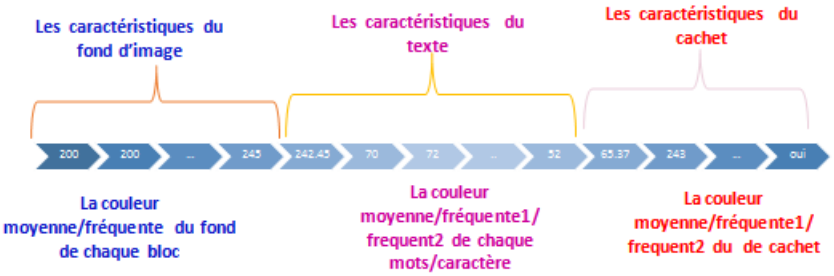


FIGURE 3.4: Le vecteur de caractéristiques d'un document.

Le même processus est fait avec chaque image jusqu'à ce que nous construisons la base qui est un ensemble de vecteurs. Après cela, nous consacrons une partie de la base pour les tests. Ce que nous appelons des données de test et la majeure partie pour l'apprentissage, ce que nous appelons les données d'entrainement.

Les données d'entrainement contiennent des caractéristiques spécifiques aux documents authentiques seulement parce que nous allons utiliser l'apprentissage par SVM mono classe, tandis que les données de test contiennent des caractéristiques de documents falsifiés et d'autres authentiques, afin de bien évaluer le modèle obtenu.

3.4.1.5 Phase d'entrainement

Dans la phase d'entrainement, le système apprend à partir les données d'entrainement dont d'appartenance est connue au préalable. Le modèle obtenu est testé ensuite sur les données de test pour vérifier sa validité. Le modèle qui détecte les documents falsifiés parmi les documents authentiques à un taux plus élevé va être utilisé par la suite dans la phase d'utilisation.

3.4.2 Phase d'utilisation

Dans la phase d'utilisation, nous devons passer par toutes les étapes précédentes sauf l'étape d'apprentissage où nous commençons à importer un document administratifs et extraire toutes ses caractéristiques pour construire un vecteur de caractéristiques. Enfin, on utilise le modèle enregistré pour décidé de l'authenticité du document.

3.5 Conclusion

Nous avons présenté dans ce chapitre notre méthode proposée. Ensuite nous avons présenté la conception de notre système, la conception globale des deux phases (apprentissage et utilisation). Enfin nous avons détaillé la conception de chaque phase.

Le chapitre suivant sera consacré à l'implémentation du système avec ses différents détails ainsi que les expérimentations menées et les résultats obtenus.

CHAPITRE

4

IMPLÉMENTATION

4.1 Introduction

Dans le chapitre précédent, nous avons présenté une conception du système. Le présent chapitre a pour objectif d'implémenter les étapes proposées pour réaliser le système conçu.
Nous allons présenter aussi l'environnement de travail, le langage de programmation, et des outils que nous avons utilisés pour construire le système. Par la suite, nous allons expliqué toutes les expérimentations que nous avons appliquées sur la méthode proposée et les résultats obtenus en utilisant un ensemble de documents que nous avons préparés.

4.2 Environnement et outils de programmation

Dans notre application, nous avons utilisé le langage c++ et l'environnement builder c++ afin de traiter les images des documents et faire l'extraction des caractéristiques.

Dans la phase d'apprentissage, nous avons choisi d'utiliser la plateforme Weka et plus particulièrement la bibliothèque LibSVM des machines à vecteurs supports.

4.2.1 Environnement de développement

4.2.1.1 Langage de programmation

Le C++ est un langage de programmation créé en 1983 par Bjarne Stroustrup des laboratoires Bell. Il se distingue de son prédécesseur (le langage C) en proposant une programmation orientée objet. Ce langage est particulièrement utilisé dans les applications demandant de hautes performances.

Le C++ est d'une grande efficacité car il a en plus des fonctionnalités puissantes, comme par exemple la notion de classe qui permet d'appliquer les techniques de la programmation-objet. Ils sont caractérisés par des avantages tel que la programmation générique, la structuration du code, l'encapsulation des données et la modularité accrue. D'un autre côté, il comporte un certain nombre d'inconvénients comme la lisibilité des programmes de bas niveaux et l'exécutable produit est plus lourd [40][52].

4.2.1.2 Environnement (builder c++)

Builder c++ est un environnement de développement basé sur C++ proposé par Borland. C'est un environnement de développement rapide d'applications (RAD) qui permet aux développeurs de développer du code avec une vitesse et une productivité. L'environnement est livré avec un certain nombre de composants qui rendent le codage logiciel plus simple et plus rapide et permet de réaliser de façon très simple l'interface des applications [33][17].

4.2.2 Outils utilisés

4.2.2.1 Logiciels de retouche d'images

Nous avons utilisé des programmes de retouche d'images pour falsifier les images des documents administratifs afin de pouvoir les utiliser plus tard dans la phase de test pendant le processus d'apprentissage.

PhoXo Est un outil d'édition d'image gratuit qui vient avec beaucoup de fonctionnalités. Il permet de recadrer, redimensionner, retoucher et éditer les images. En fait, il est dit être comme une mini version d'Adobe Photoshop.

Il possède tous les outils indispensables pour améliorer les photos. Cet outil dispose également d'une large gamme d'effets photo et texte. Il prend en charge les calques, le traitement d'images par lots, les masques d'image, les filtres et plus de 50 effets spéciaux à ajouter aux photos, ainsi que divers outils d'édition utiles pour ajuster les couleurs d'affinage, le traitement par lots,.. etc [46].

Nous avons utilisé ce programme pour modifier les images en termes de suppression des informations dans les documents et conserver la mise en forme et la couleur de l'arrière-plan.

Photo editor Une application d'édition d'image pour les photos numériques. Il est utilisé pour recadrer et retoucher des photos. Il ajuste les couleurs de l'image avec la possibilité d'ajouter des effets et des filtres dans chaque mise à jour. Photo editor a aussi l'avantage de l'écriture textuelle, du dessin et de l'écriture sur les images de manière cohérente. Il a également différents types de lignes de texte, y compris l'écriture en arabe [22].

Nous avons utilisé ce programme pour ajouter des mots aux champs que

nous avons précédemment supprimés d'une manière cohérente et non suspecte pour changer le contenu du document et le falsifier.

4.2.2.2 Weka

Weka (Waikato environment for knowledge analysis) est une suite populaire de logiciels d'apprentissage automatique écrits en Java. La suite weka contient une collection d'outils de visualisation et d'algorithmes pour l'analyse de données et la modélisation prédictive, ainsi que des interfaces utilisateur graphiques pour un accès facile à cette fonctionnalité. Il fournit de nombreux algorithmes différents pour l'exploration de données et l'apprentissage automatique pour les tâches data mining.

Weka contient des outils pour le pré-traitement des données : la classification, la régression, le clustering, les règles d'association et la visualisation [50].

4.2.2.3 Bibliothèque LibSVM

LIBSVM est une bibliothèque pour les machines à vecteurs de support. Son but est d'aider les utilisateurs à utiliser SVM facilement comme un outil.

LIBSVM est actuellement l'un des logiciels SVM les plus utilisés et pour cela, il a gagné une grande popularité dans l'apprentissage automatique et de nombreux autres domaines.

Cette bibliothèque est développée pour implémenter dans la résolution des problèmes liée a la classification avec l'utilisation des machine a vecteur support que ce soit la classification mono-class, binaire, multi-classes et même pour la régression.

LIBSVM prend en charge diverses formulations SVM pour la classification, la régression et l'estimation de la distribution [18].

Le fonctionnement de LIBSVM traduit bien ces deux phases, il consiste à définir des paramètres d'entrée, à utiliser des fichiers d'apprentissage pour générer un fichier modèle, puis exploiter ce dernier pour faire des prédictions sur le fichier de test.

4.2.2.3.1 Paramètres SVM pris en charge par la librairie [26] :

Afin d'entamer le processus d'apprentissage avec la LIBSVM, certains pa-

ramètres sont à renseigner selon qu'on souhaite faire une classification, une régression ou autre, le choix des bons paramètres est déterminant pour obtenir des résultats satisfaisants.

-s sv_type : C'est le type de l'algorithme SVM à utiliser, peut être l'une des fonctions : C_SVC, NU_SVC, ONE_CLASS, EPSILON_SVR, NU_SV.

-t kernel_type : C'est le type de la fonction noyau à utiliser, il défini par : LINEAR, POLY, RBF ou SIGMOID

$$Linaire = u' * v$$
$$polynomiale = (gamma * u' * v + coef0)^{degre}$$
$$Radiale = exp(-gamma * |u - v|^2)$$
$$sigmode = tanh(gamma * u' * v + coef0)$$

Tels que **u'** représente la transposé du vecteur contenant les valeurs des attributs de l'ensemble d'apprentissage, et **v** le vecteur des labels (étiquettes). **Le gamma**, **degré** et **coef0** sont des paramètres (rentrés pas l'utilisateur).

Paramètres des fonctions noyau

-d degree : Paramètre degré de la fonction noyau , par défaut 3

-g gamma : Paramètre gamma de la fonction noyau, par défaut 1

-r coef0 : Paramètre coef0 de la fonction noyau, par défaut 0.

Paramètres dépendants du type SVM choisi

-c cost : C'est le paramètre C (coût), qui représente la pénalité de l'erreur, à renseigner lors de l'utilisation du type SVM C-SVC, epsilon-SVR et nu-SVR, par défaut le coût est égal à 1

-wi weight : pour changer le paramètre C à (weight*C), s'il n'est pas renseigné weight est égale à 1 sa valeur par défaut, et par conséquent neutre.

-n nu : Paramètre nu du type nu-SVC, One-class-SVM et nu-SVR, par défaut 0.5

-p epsilon : Paramètre epsilon de la fonction de perte (Loss Function) pour le type epsilon-SVR, par défaut égal à 0.1

4.3 Application de détection des documents numérique administratifs falsifiés proposée

Nous avons développé une application qui implémente la méthode proposée décrite dans les chapitre précédents et qui se compose de deux phases :

— Construction du modèle (Par l'extraction des caractéristiques puis l'entrainement et le test).
— L'utilisation (Détection des documents scannés falsifiés).

4.3.1 Base de données utilisé

On sait que les documents administratifs ne sont pas largement disponibles sur Internet car ce sont des documents officiels d'individus ou d'institutions. Par conséquent, essayer de les collecter pour des expériences, des tests ou des analyses dans le domaine de la détection de documents frauduleux est difficile.

D'après cette perspective, et afin de réaliser notre approche proposée. Nous avions besoin d'éditer ces documents tout en conservant leur forme administrative, et de ne pas les lier à de vraies personnes, puis nous les avons numérisés pour construire la base de données dont nous avons besoin dans notre travail.

Nous avons utilisé différents type de documents administratifs, des certificats de scolarité, des attestations, des invitations, certificat d'authenticité, autorisation de licence, rapports administratifs,..etc.

Ces documents utilisés sont caractérisés par :

— Chaque document comporte un cachet.
— Chaque document comporte un signature.
— Chaque document comporte une en-tête et un texte .
— Chaque document a une bonne résolution.

Pour l'entrainement, nous avons utilisé une base qui contient 141 documents authentiques, tandis que pour le test du modèle, nous avons utilisé une base qui contient des document authentiques et d'autre falsifiés afin de vérifier la validité de modèle. Nous avons produit ces documents falsifiés par l'utilisation des logiciels de retouche d'image déjà mentionnés.

Après avoir construit la base des documents(images), nous passons à mettre en œuvre les étapes de notre méthode proposée. Nous commençons par la construction de la base de caractéristiques pour l'utiliser en apprentissage et obtenir le modèle. Pour cela, nous avons suivi les étapes suivantes :

— La détection des lignes du texte.
— Le découpage du texte en des mots et des caractères.
— Le découpage d'images en des blocs similaires.
— Détection de la zone du cachet.
— Calcul de la première et deuxième couleur fréquente et la couleur moyenne du fond de cachet.
— Calcul de la première et deuxième couleur fréquente et la couleur moyenne de chaque mots/caractère du texte. Puis,
— Calcul de la moyenne de toutes ces couleurs moyennes et calcul de la première et la deuxième couleurs fréquentes de toutes les couleur fréquentes.
— Calcul de la couleur fréquente et moyenne du fond de chaque bloc de l'image.
— Calcul de la couleur fréquente et moyenne du fond d'image.
— La combinaison des caractéristiques afin de construire un vecteur de caractéristique pour chaque image.

Le résultat de cette phase est un fichier de format "csv" contenant dans chaque ligne les caractéristiques numériques d'un document et la classe "one" pour marque tous les document d'authentique pour utiliser l'apprentissage mono-classe.

4.3.2 Apprentissage et test

Afin de faire l'apprentissage à partir de notre base de caractéristiques, nous avons utilisé Weka où nous avons gardé le format de fichier csv et nous l'avons utilisé directement.

Après l'entrainement, nous obtenons le résultat sous forme d'un modèle que nous enregistrons pour l'utiliser dans le test.

Puisque nous utilisons le SVM mono classe dans l'apprentissage, dans la phase de test, tout document qui n'a pas de caractéristiques ressemblant aux caractéristiques des documents authentiques sera considéré comme étrange et marqué par ' ?', et nous le considérons comme document falsifié.

Le résultat de la phase de test est obtenu sous forme de fichier texte. La

figure suivante illustre un résultat de test d'un ensemble de document.

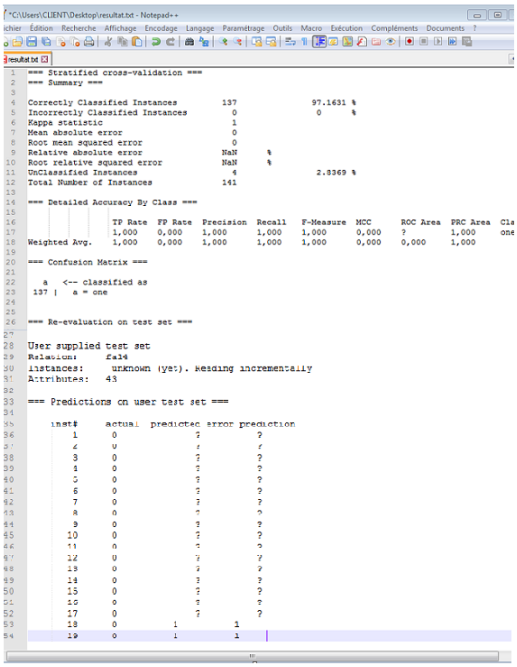


FIGURE 4.1: Résultats de phase de test d'un modèle construit

4.3.3 Utilisation

Pour classer un nouveau document et déterminer si il a été falsifié ou non, on le passe à toutes les étapes précédentes et les mêmes opérations d'extraction des caractéristiques précédentes afin de construire un vecteur de caractéristique. Ce vecteur est passé au modèle appris avec un '?' dans le champs classe.

4.3.4 Présentation des interfaces

Notre application est basée sur quatre interfaces, chaque interface offre une étape de notre système :

4.3.4.1 Extraction des caractéristiques

Cette interface permet de :
— Sélectionner un ou plusieurs documents scannés et les afficher.
— Extraire toutes les caractéristiques de ce document.
— Afficher le vecteur de caractéristiques du document ou l'ensemble de documents.
— Enregistrer les vecteurs construit dans un fichier csv.

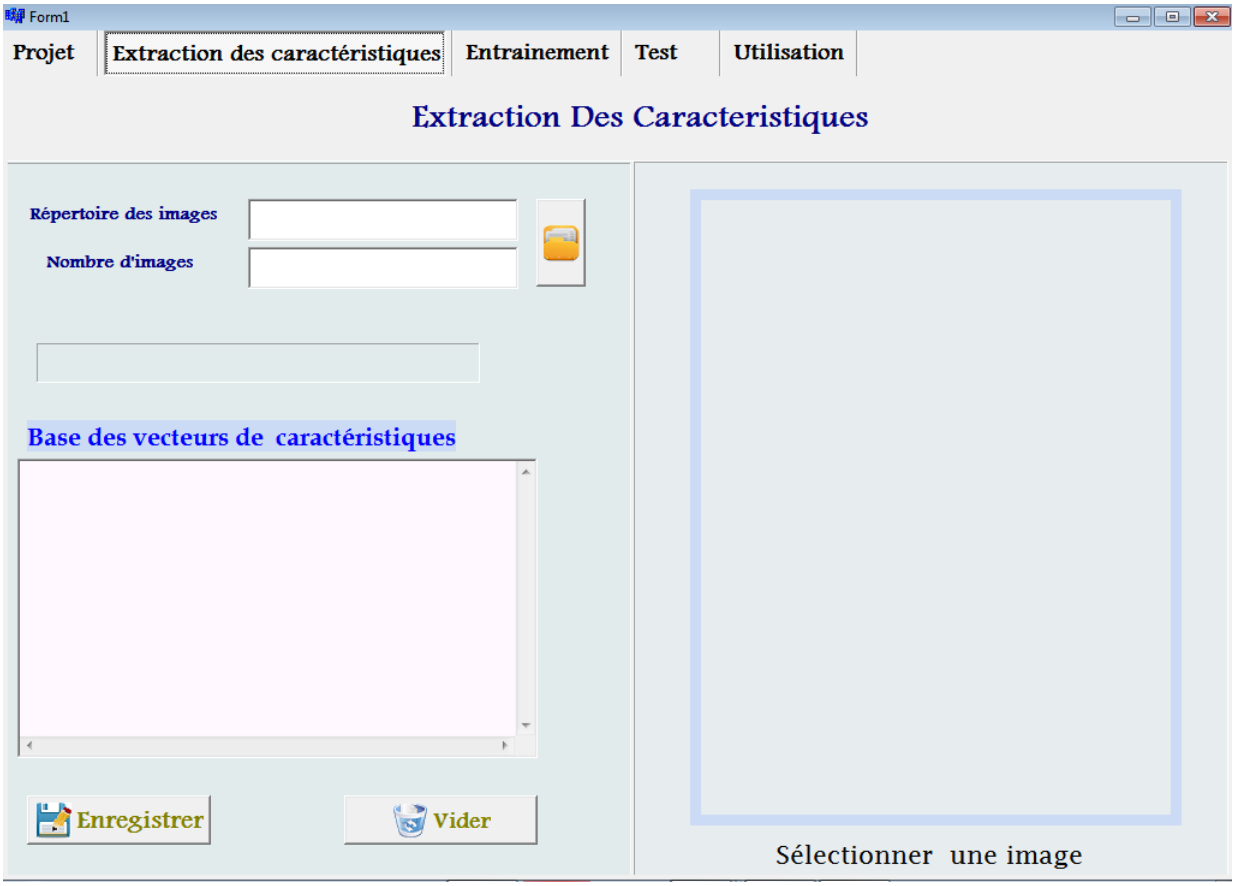


FIGURE 4.2: Interface d'extraction des caractéristiques.

4.3.4.2 Interface d'apprentissage

Cette interface permet de :
— Charger le fichiers d'entrainement.

— Modifier les paramètres de LibSVM.
— Faire l'apprentissage.
— Afficher le taux de reconnaissance.
— Enregistrer le modèle.

FIGURE 4.3: Interface d'apprentissage.

4.3.4.3 Interface de test

Cette interface permet de :
— Charger le fichier de test et le modèle.
— Prédire les classes des exemples du fichier de test.
— Afficher le résultat de prédiction.
— Enregistrer le modèle.

FIGURE 4.4: Interface d'apprentissage et test.

4.3.4.4 Interface d'utilisation

Cette interface permet de :
— Sélectionner un document scanné.
— Extraire les caractéristiques de ce document.
— Charger le modèle enregistré.
— Prédire la classe du document sélectionné.
— Afficher le résultat de prédiction.

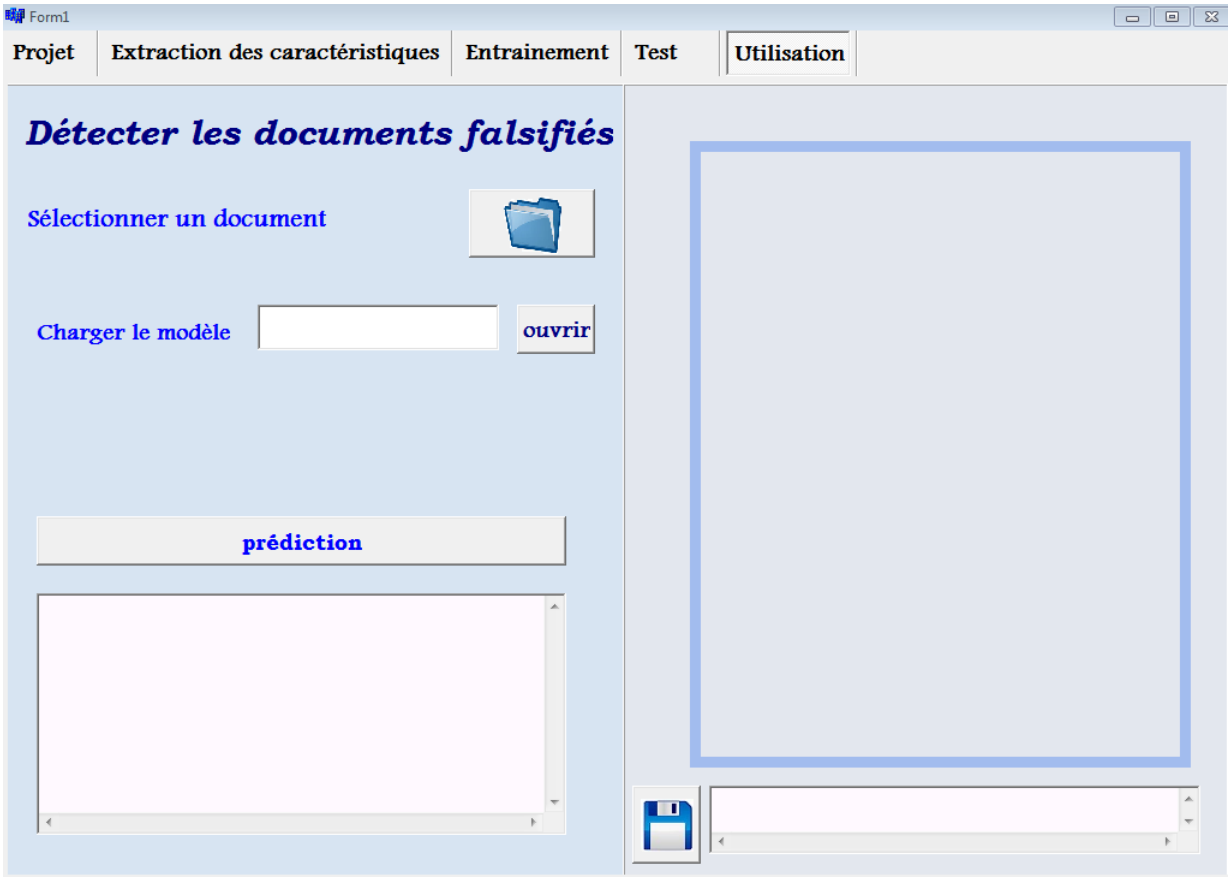


FIGURE 4.5: Interface d'utilisation.

4.4 Expérimentations et résultats

Pour tester notre méthode, nous avons utilisé les expérimentations suivantes :

4.4.1 Expérimentation

4.4.1.1 Première expérimentation

Dans la première expérimentation nous avons utilisé 141 instances, et le nombre d'attributs était 29 attributs (nous avons découpée l'image en 3 ligne

et 3 colonne).
Dans cette expérimentation, nous avons changé les paramètres utilisées avec la même base de 29 attribut comme suit :

- Nous avons utilisées dans la première fois "use training set" comme option de test et les paramètres suivants :
  — Type SVM : one-class svm classification
  — Type kernel : Radial basis function
  — gama : 0.04
  — nu : 0.02
  Le taux de reconnaissance obtenu sur les données de d'entrainement est 86 %

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.18 seconds

=== Summary ===

Correctly Classified Instances         122               86.5248 %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                NaN      %
Root relative squared error            NaN      %
UnClassified Instances                  19               13.4752 %
Total Number of Instances              141
```

FIGURE 4.6: Résultat de la première expérimentation avec le premier réglage des paramètres.

Le taux de reconnaissance obtenu sur les données de test est 94% de tels sorte les données de test sont 17 instances (12 documents falsifié et 5 authentique).
Le symbole ' ?' indique que le système ne connaît pas le document, que nous considérons comme des documents falsifiés.

```
=== Predictions on user test set ===

    inst#     actual  predicted error prediction
        1        0          ?          ?
        2        0          ?          ?
        3        0          ?          ?
        4        0          ?          ?
        5        0          1          1
        6        0          1          1
        7        0          1          1
        8        0          1          1
        9        0          ?          ?
       10        0          ?          ?
       11        0          ?          ?
       12        0          ?          ?
       13        0          ?          ?
       14        0          1          1
       15        0          1          1
       16        0          ?          ?
       17        0          1          1
```

FIGURE 4.7: Résultat de la première expérimentation avec le premier réglage sur les données de test .

- Nous avons fait un autre entrainement sur la même base et avec 'cross-validation' et les paramètres suivants :
  — Type SVM : one-class svm classification
  — Type kernel :polynomial
  — gama : 0.04
  — nu : 0.2
  Le taux de reconnaissance obtenu est 95% comme le montre la figure suivante :

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         135               95.7447 %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                NaN      %
Root relative squared error            NaN      %
UnClassified Instances                   6                4.2553 %
Total Number of Instances              141
```

FIGURE 4.8: Résultat de la première expérimentation avec le deuxième réglage sur les données d'entrainement .

Cette fois, nous avons testé le modèle sur la base d'entrainement. Le taux de reconnaissance obtenu est 98% .

```
     121       1:One       1:One       1
     122       1:One       1:One       1
     123       1:One           ?       ?
     124       1:One       1:One       1
     125       1:One       1:One       1
     126       1:One       1:One       1
     127       1:One       1:One       1
     128       1:One           ?       ?
     129       1:One       1:One       1
     130       1:One       1:One       1
     131       1:One       1:One       1
     132       1:One       1:One       1
     133       1:One       1:One       1
     134       1:One       1:One       1
     135       1:One       1:One       1
     136       1:One       1:One       1
     137       1:One       1:One       1
     138       1:One       1:One       1
     139       1:One       1:One       1
     140       1:One       1:One       1
     141       1:One       1:One       1

=== Summary ===

Correctly Classified Instances         139               98.5816 %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
UnClassified Instances                   2                1.4184 %
Total Number of Instances              141
```

FIGURE 4.9: Résultat de test de modèle obtenu sur la base d'entrainement.

Nous avons observé dans cette expérience que la valeur ν affecte considé-

rablement le taux de reconnaissance. Lorsqu'elle est réduite le taux atteint sa valeur la plus élevée comme indiqué la courbe de la figure suivante :

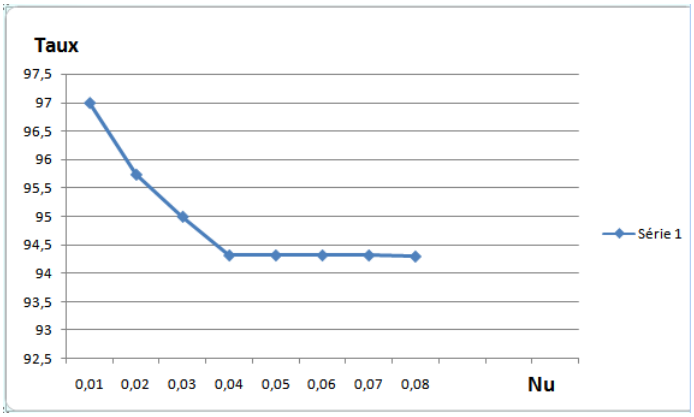


FIGURE 4.10: L'effet de la valeur de nu sur le taux de reconnaissance dans la première expérimentation.

Nous avons également changé le type de kernel avec les mêmes options de test pour voir son effet sur le taux de reconnaissance et trouvé les résultats dans le tableau suivant :

Réglage de paramètres	Type de kernel	Taux
Nu=0,02 Gamma=0,04	Linear	95.74
	polynomial	97.87
	Radial basis function	86.52

TABLE 4.1: L'effet du type de kernel utilisé sur le taux de reconnaissance

4.4.1.2 Deuxième expérimentation

Dans la deuxième expérimentation nous avons utilisé 141 instances et le nombre d'attributs était 60 attributs (nous avons découpé l'image en 5 lignes et 5 colonnes).

Nous avons utilisé les paramètres suivants :

— Type SVM : one-class svm classification

— Type kernel : Linéaire
— nu : 0.02
— gamma : 0.04
— validation croisée : 10

Le taux de reconnaissance obtenu sur les données d'entrainement est 96 %

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         136               96.4539 %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                NaN      %
Root relative squared error            NaN      %
UnClassified Instances                   5                3.5461 %
Total Number of Instances              141
```

FIGURE 4.11: Résultat de la deuxième expérimentation sur les données d'entrainement.

La table suivant montre l'effet du type du kernel sur le taux.

Réglage de paramètres	**Type de kernel**	**Taux**
Nu=0,01 Gamma=0,02	Linear	98.58
	polynomial	98.58
	Radial basis function	43.26

TABLE 4.2: Effet de type de kernel sur le taux dans la deuxième expérimentation

La même observation que nous avons trouvée dans l'expérimentation précédente sur concernant l'influence de la valeur de ν sur le taux de reconnaissance. Dans la courbe suivante nous avons fixer le γ à 0.04 avec l'option de test est en "cross validation".

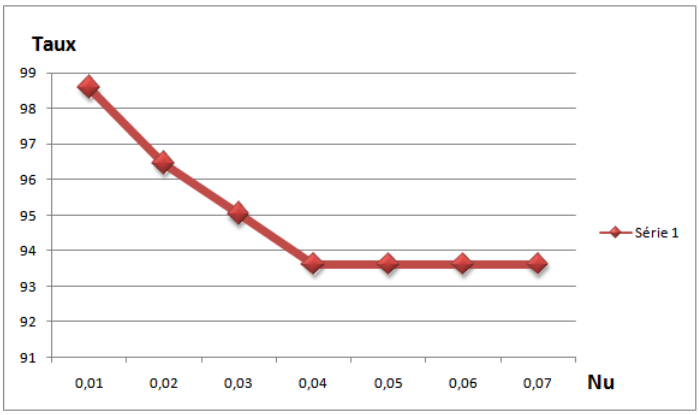


FIGURE 4.12: Effet de la valeur de ν sur le taux de reconnaissance dans la deuxième expérimentation.

4.4.1.3 Troisième expérimentation

Dans la troisième expérimentation nous avons utilisé 141 instances, et le nombre d'attributs était 42 attributs (nous avons découpée l'image en 4 lignes et 4 colonnes).

Nous avons utilisé les paramètres suivants :
— Type SVM : one-class svm classification
— Type kernel : polynomial
— nu : 0.01
— gamma : 0.04
— validation croisée : 10
Le taux de reconnaissance obtenu sur les données de d'entrainement est 98% .

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         139               98.5816 %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                NaN      %
Root relative squared error            NaN      %
UnClassified Instances                   2                1.4184 %
Total Number of Instances              141
```

FIGURE 4.13: Résultat de la troisième expérimentation sur les données d'entrainement .

Pour connaître l'effet du type de kernel utilisé sur le taux de reconnaissance, nous avons construit le tableau suivant où nous avons fixé γ à 0.04 et la valeur de ν à 0.02 :

Réglage de paramètres	**Type de kernel**	**Taux**
Nu=0,02 Gamma=0,04	Linear	97.87
	polynomial	98.58
	Radial basis function	86.52

TABLE 4.3: l'effet de type de kernel sur le taux dans la troisième expérimentation

Pour voir l'effet de la valeur de ν sur le taux, nous avons tracé la courbe de la figure suivante qui montre la variation du taux de reconnaissance en fonction de la valeur de ν.

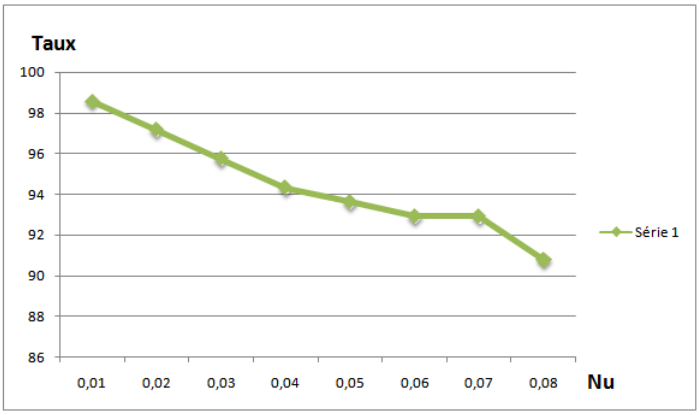


FIGURE 4.14: Effet de la valeur de nu sur le taux de reconnaissance dans la troisième expérimentation.

4.4.2 Discussion des résultats

A travers toutes nos expérimentations et résultats, nous pouvons dire que notre système proposé est très efficace pour détecter les documents frauduleux. Les résultats obtenus montrent la capacité de notre système à identifier correctement plus de 90% des documents dans le cadre de l'apprentissage automatique, ce qui permet d'accélérer le processus de vérification des documents et de manière automatique.

4.5 Conclusion

Dans ce dernier chapitre, nous avons représenté l'implémentation de notre système proposée : L'environnement, le langage de programmation et les outils de développement. Ensuite nous avons présenté quelques expérimentations effectuées et présenté tous les résultats et taux de reconnaissance.

CHAPITRE

5

CONCLUSION GÉNÉRALE

Avec l'utilisation croissante des documents numériques, la falsification de documents est devenue l'un des crimes les plus connus. Pour palier ce crime, nous avons étudié plusieurs travaux connexes et leurs méthodes et techniques pour la détection des documents numériques falsifiés.

Afin de lutter contre la falsification et faire face à la propagation de ce crime, nous avons proposé une méthode basée sur l'apprentissage automatique par la méthode SVM mono classe en utilisant certaines caractéristiques de l'image du document concernant le fond, le texte,et le cachet du document.

Pour réaliser notre travail, nous avons extrait ces caractéristiques de tous les documents administratifs authentiques disponibles.

Nous avons enregistrées les caractéristiques extraites dans une base et que nous avons utilisée pour construire un modèle de décision permettant de reconnaitre les document falsifiés des documents authentiques.

Pour valider notre méthode proposée, nous avons préparé une base d'entrainement de plus de 141 exemples représentant des documents administratifs authentiques. Les taux de reconnaissance obtenus sur les données d'entrainement dépassent les 90% ce qui est encourageant et démonte l'efficacité de la méthode proposée.

Pour les travaux futur, nous suggérons quelques idées qui peuvent améliorer notre système tels que :

— Essayez de traiter les timbres et les logos dans les documents avec le texte pour tester tous les composants des documents.

— Tenter d'extraire d'autres caractéristiques du document pour améliorer les performances du système.

— Travailler sur le traitement et l'analyse des propriétés dans d'autres documents administratifs tel que les passeports, les carte d'identité..etc, pour généraliser l'utilisation du système.

BIBLIOGRAPHIE

[1] Alternatives gratuites à photoshop 5 logiciels de retouche photo gratuits pour mac et windows. https://fr.softonic.com/articles/alternatives-gratuites-photoshop-logiciels-retouche-photo-mac-windows. Accessed : 2018-06-16.

[2] *Faq foire aux questions examen des documents.* (Prado) Public Register Of Authentic Identity And Travel Documents Online.

[3] La numérisation des documents. https://www.docudepot.com/la-ou-se-rejoignent-la-numerisation-et-la-gestion-des-documents. Accessed : 2018-06-16.

[4] Le traitement des images. http://www.unige.ch/cyberdocuments/didacticiel/unite2/module4.html. Accessed : 2018-06-12.

[5] Logiciel photoshop software. http://www.dicodunet.com/definitions/internet/logiciel-photoshop.htm. Accessed : 2018-06-16.

[6] Les faux commis dans un document administratif. *infpn*, 01/11/2010.

[7] *La numérisation des images et des textes (Cours numérisation)*, volume 52. D-Lib Magazine, 2010.

[8] *Guidance on examining identity documents.* National Document Fraud Unit, 2016.

[9] *Guidance on the use of document scanners.* 2016.

[10] Mkadmi Abderrazak. Documents électroniques. *Cours destiné aux étudiants de 2eme année Licence*, 2012.

[11] Ramzi M. Abed. Scanned documents forgery detection based on source scanner identification. *American Journal of Information Science and Computer Engineering*, 2015.

[12] H.Cecotti A.Belaïd. La numérisation de documents : principes et évaluation des performances. *Université Nancy 2 - LORIA*, 2005.

[13] Christophe Alleau. *Images numériques*. Académie de Poitiers.

[14] Richard Entlich Anne R. Kenney, Oya Y. Rieger. *Didacticiel d'Imagerie Numérique - Matières /de la théorie à la pratique didacticiel d'imagerie numérique*. Bibliothèque de l'Université Cornell/Département de Recherches, 2003.

[15] Jannsen Brad. Support vector machines for binary classification and its applications. 2008.

[16] Hammoudi Sihem Brahmi Sara. *Extraction des différents opérateurs de la morphologie mathématique*. PhD thesis, Université Abou Bakr Belkaid - Tlemcen, 2014.

[17] bruno garcian. *Introduction à C++ Builder*. ISIMA, 1988-1999.

[18] Chih-Chung Chang and Chih-Jen Lin. Libsvm : A library for support vector machines. *Department of Computer Science National Taiwan University, Taipei, Taiwan*, 2001.

[19] Sous comité des archivistes Sous l'égide du Comité des secrétaires généraux. Guide de gestion d'un projet de numérisation. Juillet 2014.

[20] cpni's web site. *A good practice guide on pre-employment screening document verification*. 2007.

[21] Service d'archives itinérant CDG 90. *La numérisation*. Maison des Communes - CDG 90, Septembre 2010.

[22] Ziff davis. Definition of : photo editor. `https://www.pcmag.com/encyclopedia/term/49191/photo-editor`. Accessed : 2018-06-17.

[23] Unité D.2 de l'OLAF. *Détection de faux documents dans le cadre des actions structurelles/ Guide pratique à l'intention des autorités de gestion*.

[24] Abdelhamid Djeffal. *Utilisation des méthodes Support Vector Machine (SVM) dans l'analyse des bases de données*. PhD thesis, Université Mohamed Khider-Biskra, 2012.

[25] E.Bacquet. *Préparer Des Images Numérique*. Eyrolles, 2009.

[26] Imane el hassani. Svr avec boosting pour la prévision à long terme. *école polytechnique de l'université de tours*, 2011-2012.

[27] Patrick Finot. Le couteau suisse du traitement de l'image avec photoscape. `http://www.informatique-enseignant.com/photoscape`. Accessed : 2018-06-16.

[28] R. D. Gaharwar G. K. S. Gaharwar, Prof.V.V. Nath. *Comprehensive study of different types image forgeries*. 2015.

[29] Michel Gorin. Le numérique : impact sur le cycle de vie du document. premier colloque ebsi-enssib du 13 au 15 octobre 2004, montréal (québec). *RESSI*, (1), 2005.

[30] Fadi H. Naser Hasan. *New Method to Detect Text Fabrication in Scanned Documents*. PhD thesis, Mémoire Master ,Islamic University in Gaza, 2015.

[31] Benhmza Hiba. *Détection automatique des documents numériques falsifiés*. PhD thesis, Mémoire Master. Université Mohamed Khider-Biskra, 2017.

[32] Halima Maamri Imene Trablelst. *Tatouage numérique fragile pour l'authentification d'images*. PhD thesis, Mémoire Master. Universite de kasdi merbah ouargla, 2016.

[33] Dale Janssen-Cory Janssen. C++ builder. `https://www.techopedia.com/definition/12728/c-builder`. Accessed : 2018-06-16.

[34] Faisal shafait Andreas dengel Johann gebhardt, Markus goldstein. *Document authentification using printing technique features and unsupervised anomaly detection*. German research center for artificial Intelligence, School of Computer Science and Software Engineering.

[35] Faisal Shafait Joost van Beusekom and Thomas M. Breuel. Document inspection using text-line alignment. *German Federal Ministry of Education and Research*, 2010.

[36] Sylvie Lainé-Cruzel. Documents, ressources, données : les avatars de l'information numérique. *Revue I3-information interaction intelligence*, (1), 2004.

[37] Aurélien langlade. *Éléments de connaissance sur la fraude aux documents et à l'identité en 2014*. La criminalité en France, Rapport annuel 2015 de l'ONDRP, 2015.

[38] les services de la Commission. *Note d'orientation relative à l'évaluation du risque de fraude et aux mesures antifraude efficaces et proportionnées*. Commission européenne direction générale politique régionale et urbaine.

[39] Hanifi Majdoulayne. *Extraction de caractéristiques de texture pour la classification d'images satellites.* PhD thesis, Université De Toulouse, 2009.

[40] O. Marguin. *C++ : LES BASES /Cours d'informatique.* 2003/2004.

[41] Maxicours. Cours de ressources humaines et communication terminale stmg-la notion de document. `https://www.maxicours.com/se/fiche/5/1/219515.html/tstt`. Accessed :2018-06-16.

[42] Jean-Michel mermet. Techniques de numérisation. *Cours licence BDAN*, 2010.

[43] Abderrazak Mkadmi. *Document numérique.* Institut supérieur de documentation université de manouba, 2008-2010.

[44] Abderrazak Mkadmi. *Gestion électronique de documents.* Institut supérieur de documentation, 2010.

[45] Marref nadia. *Apprentissage Incrémental and Machines à Vecteurs Supports.* PhD thesis, diplôme de Magister,Université hadj lakhdar– batna, 2013.

[46] Nirmal. Add effects to your photos with phoxo. `http://www.nirmaltv.com/2011/02/25/add-effects-to-your-photos-with-phoxo/`. Accessed : 2018-06-16.

[47] Christine Roger. *Des termes techniques relatifs aux éléments de sécurité et aux documents sécurisés en général.* Conseil de l'Union européenne, 2007.

[48] oriol ramos terradesy Romain bertrand, petra gomez-kramer. A system based on intrinsic features for fraudulent document detection. *Computer vision center, universitat autonoma de barcelona*, 2012.

[49] Nasir Memon Shize Shang and Xiangwei Kong. Detecting documents forged by printing and copying. *EURASIP Journal on Advances in Signal Processing*, 2014.

[50] Swasti Singhal and Monika Jena. A study on weka tool for data preprocessing, classification and clustering. *International Journal of Innovative technology and exploring engineering (IJItee)*, 2(6), 2013.

[51] Prof. Dr. Ajay A. Gurjar Snigdha K. Mankar. *Image Forgery Types and Their Detection.* April 2015.

[52] Nadir Soualem. Cours de c++ / introduction au c++. `https://math-linux.com/c-3/article/cours-de-c-1-introduction-au-c`. Accessed : 2018-06-16.

[53] Christophe Soullez Stéfan Lollivier. *La criminalité en France.* 2015.

[54] Mokhtar Taffar. Initiation à l'apprentissage automatique. *Support de Cours pour étudiants en Master en Intelligence Artificielle , Université de Jijel.*

[55] Christian Tiago. Exposing digital image forgeries by illumination color classification. *IEEE Transaction On Information Forrnsics And Security*, 2013.

[56] Sean Trundy. *Counterfeit Fraud Prevention – Tips, Tools and Techniques.*

[57] wikipedia. chiffrement. `https://fr.wikipedia.org/wiki/Chiffrement`. Accessed : 2018-06-16.

[58] wikipedia. Corel photo-paint. `https://fr.wikipedia.org/wiki/Corel_Photo-Paint`. Accessed : 2018-06-16.

[59] wikipedia. Fil de sécurité. `https://fr.wikipedia.org/wiki/Fil_de_sÃ'curitÃ'`. Accessed : 2018-06-16.

[60] wikipedia. *l'image numérique.* Accessed : 2018-06-12.

[61] wikipedia. Scanner (informatique). `http://www.brunomartin.be/cours/scanner.pdf`. Accessed : 2018-06-14.