Research Article

International Journal of Advanced Computer Research, Vol 10(46)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2019.940105

An approach for extracting chemical data from molecular representations

Amena Mahmoud[*], Taher Tawfiq Hamza and Magdi Zakaria Rashad
Professor, Department of Computer Science, Mansoura University, Egypt

Received: 03-September-2019; Revised: 11-November-2019; Accepted: 15-December-2019

Abstract
The ever-increasing quantity of chemical literature necessitates the creation of automated techniques for extracting relevant information. The digital conversion process of chemical molecule representations into their corresponding computerized illustration has been evolved using numerous applications. This work is a part of our contribution to computer-aided toxicity recognition research which predicts the possible toxic side effects of drugs on the process of drug design. The current proposed approach provides essential reviews for previous related researches in the field of automated chemical information extraction and mentions our current proposed technique that is considered as an intelligent module for converting chemical molecule structures to computerized structures. Mentioned mechanisms are used to detect bonds that are represented by letters through comparison with templates database, atoms, and lines for extracting data from graphs. A sample of 100 chemical compound structures was used to be converted into computer representation to get an overall result of 88.9% precision. Finally, a comparison between related approaches and the current proposed one at their precision rates for classification of substructure patterns was conducted.

Keywords
Chemical molecules, Data extraction, Bonds, Atoms, Molecule representation.

1.Introduction
The process of getting chemical data from different document formats is considered a complicated problem. Despite documents are found in some fundamental forms like 'image' or 'text', both types have various extensions (e.g. word document, rtf, html, txt, pdf, or gif file and jpg formats). Journal articles commonly exist in an integrated form of images embedded within text in addition to tables and figures. The capability of getting data form images and texts in all variations leads to a successful process of chemical structure recognition. Two instances of data sources that link chemical molecules structures with phenotypes of biomedical targets are PubChem [1]–a data warehouse containing over nineteen million chemical molecules structures and PubMed [2] – the medical literature corpus database and, both of them can have a link to bio-activity descriptions and comparable structures. Other resources for the chemical molecules' depictions in medical documents are commonly drawn manually using programs like ChemDraw, ChemWindow, ISIS/Draw and ChemSketch [3].

The structures of chemical compounds are integrated within document parts in a way that is not legible converted into a computerized format in scientific researches and books which are usually represented as image files (e.g. png, jpg or gif). Therefore, most depositories of biochemical knowledge cannot be simply connected to other chemical structures in biochemical researches, which are not adjustable for searching and analysis by related Bioinformatics applications [4]. As soon as a molecule is analyzed, the formulation representation can be converted to a computerized format, like the simplified molecular-input line-entry system (SMILES) [5], IUPAC [6] or MOLfile representations that characterize atoms, bond canonical orders, and linking modes of atoms in molecules according to pharmaceutical rules.

Several designed applications in the field of computer-aided drug design were proposed in the 1990s, which could get chemical formula diagrams in biochemical and medical documents and transform them into structured forms. Nowadays, through the continuing progress of the medical and cheminformatic techniques to process pharmaceutical and chemical data, further computer applications were developed with continued updates. Image-based

[*] Author for correspondence

chemical molecules definition [7], analysis of chemical components to recognize biological systems [8], atom identification using deep learning techniques of scanned chemical Images [9], structural states in atomic image recognition using computer vision [10] and chemical structures classification using neural network techniques [11] are recent instances for extracting chemical data from images.

An efficient ability for images searching might demand, converting the chemical molecules images into computer-readable formats like atoms linking tables or SMILES strings in common file representations. Newly synthesized molecules and novel drug compounds are typically indicated with chemical molecule structures in preference to molecule associated names. Additionally, a single molecule can have several alternatives that can be mentioned using exceptional names in distinct documents. The functionality of looking into patents or scientific research documents wherever the chemical compounds or similar diagrams were drawn might supplement present textual content-based search engines of medical and chemical data.

Most of the current introduced work in the field of extracting chemical information from images is commercially distributed, limited to a specific group or platform dependent. Our proposed approach will cover those limitations by being platform-independent and freely released on a designed web site which demonstrates an intelligent model depending on machine vision techniques for 2D chemical structure formulas recognition. We aim at describing and being able to identify canonical representations of chemical structure, as well as explicit and implicit information contained in the chemical molecule representations and also present a machine representation of chemical structure (connection tables, graphic visualizations, line notation). Converting chemical structures into a computer-readable format will be used at our proposed research of drugs' toxicity recognition which depends on the structure-activity characteristic.

The materials and methods section introduce the used mechanism of text extraction which is based on using characters recognizing algorithms (OCR) and lines extraction obtained from digital images using Hough transforms algorithm. Finally, we conduct a comparison between this approach and the previous related approaches which convert those images into its equivalent computerized format and discuss the current model limitations at the discussion and results in sections. The proposed mechanism aims to identify the current features like the chemical representation in the original images. The extracted chemical information of the original depiction is a significant part of the images' classification process.

2.Materials and methods

The (systematic) name and the 2D drawing of the chemical molecule are the most common ways to represent chemical molecules' structures. Both methods characterize the presented molecule, but are not being used for machine processing. Conventions illustrating how molecules could be drawn and named are required to prohibit opacities. line notations such as SMILES [5] and IUPAC [6] name recommendations are standardized instances for molecules labeling which are based on chemical graphs, that present bonds as edges and atoms as vertices, defining the current links within the molecule.

Graph-based representations are common as they present chemical molecules more simplistically. Molecules are those atoms who are connected using bonds. Specific atom functional groups give rise to some molecular properties and the molecule is considered a graph where atoms are vertices and bond edges. Using of the substructure searching and line notations have permitted creating molecular structures' databases with their properties.

Our approach aims at utilizing computer information technology to solve problems in the field of chemistry such as chemical information retrieval and extraction, compound database searching and drug discovery. Converting a compound structure in chemical information applicable for machine learning tasks requires multilayer computational processing from chemical graph retrieval, segmentation, information extraction and SMILES representation in which each layer builds upon the successful development of previous layers and often has a substantial impact on the quality of the chemical data for machine learning. *Figure 1* represents the block diagram of the proposed model.

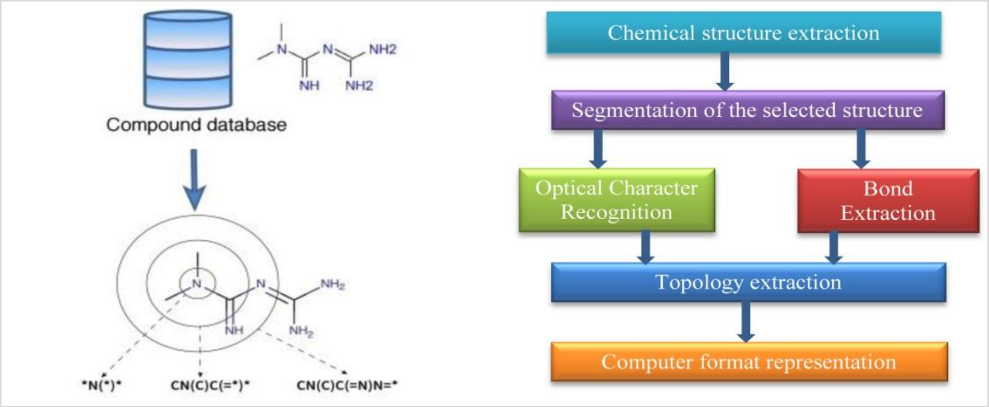

Compound database
N
N
NH2
NH
NH2
N
N
NH2
NH
NH2
N(*)
CN(C)C(=*)*
CN(C)C(=N)N=*
Chemical structure extraction
Segmentation of the selected structure
Optical Character Recognition
Bond Extraction
Topology extraction
Computer format representation

Figure 1 Proposed model block diagram

2.1Datasets

The proposed study presents an approach for molecule recognition, which accepts a molecule structure depiction from chemical structure databases or digitally (from a previously designed application that digitally generates chemical structures) and returns an equivalent SMILES representation for the compound. There is no dataset addressing molecular structure segmentation that has been published. To generate data from chemical structure images, the following steps were performed: i) remove structures from patent pages and journals, ii) overlay structures onto the pages, iii) Randomly crop images from the pages containing structures, and iv) Denoise the images using graphics software.

Two sets of the images of chemical structure diagrams collected from different sources were used to test the precision of the proposed model. Fifty images in set I, are obtained by querying chemically and pharmaceutically significant molecules to Google Image Search while set II are collected from 120 scientific papers, as shown in *Table 1*.

Table 1 Proposed images sets

	Number of images	Image source
Set I	50	Google image search
Set II	50	Journals at the PubMed database

2.2Pre-processing

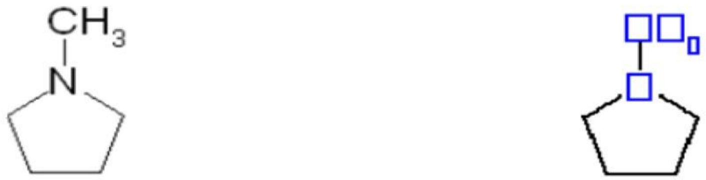
The first step includes re-sizing the chemical structure diagrams which were designed with different settings in the drawing applications, such as default bond lengths or character font sizes. Additionally, the image format and size are exposed to variations while moving to the final destination, for instance, scientific research or a web site publication. Therefore, it is essential to resize the input structure image to unify the diagram’s bond lengths and character sizes for all used chemical structures. Firstly, the length of any single bond is estimated and if the estimated bond length is larger or shorter than a certain threshold (presently 35 connected pixels), the image got resized to have the required length.

2.3Segmentation

After unifying the length of the components, the image is crossed from top to bottom, left to right searching for contained black pixel. When finding a pixel, the entire linked part including that pixel is labelled by the use of grass-fire algorithm [12] which looks for the unlabeled nearby black pixels recursively as shown in *Figure 2*. Labelling the entire component leads to the continuation of the systematic traversal.

The following procedure will be disassembling all connected components based on pixel connectivity using the 8-connectivity algorithm. Then, those connected components are categorized into graphics that represent bonds and characters as atoms. In the character detecting process, a character detection algorithm searches for objects with similar heights and areas. Generally, the most occupied height/area grouping will represent text components [13]. Using this technique, most text components can be segmented from the other chemical structure components.

To distinguish bonds that are represented by the small, isolated circles or lines from the characters' components, the relative location of each component is also checked. For instance, the character 'l' is wrongly recognized as a graphic component. Though, since it always found near to other characters, the character 'l' can be correctly recognized as a character and not a graphic component by estimating the relative location of each character. If text components cannot be recognized in this way, they will be corrected in the following steps.



(a) Original image **(b) After segmentation**

Figure 2 Segmentation of chemical structural formulas

2.4Optical character recognition

Different kinds of string-oriented chemical symbols are contained in molecule depictions, like super atoms, atoms and SMILES strings. To achieve that purpose, OCR is needed to appropriately recognize all founded character symbols. OCR is considered the most well-known system for computerized pattern recognizing. It interprets the images of written text that is usually analyzed by a scanner, into machine-editable text. Template based recognition approach can recognize those patterns.

The template-based recognition technique is structural-based, where the whole specified area is used as a feature; a comparison is conducted between each symbol which resulted from the early segmentation process and all alphabet representation as to its character templates. The similarity measurement between a character symbol and all templates is computed using suited distance function. Then there will be a comparison between the maximum similarity measured of the symbol with a certain threshold if it was found to be above that value so that the character is registered to the matched label. The characters' template contains various volumes for all character symbols that might be found at any chemical formula representation which can be scaled to suit template volumes that encounter the invariant problem. The results of the text extraction process are shown in *Figure 3*.

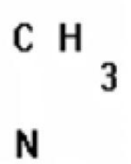


Figure 3 Text extraction

2.5Bond extraction

The most extremely used symbols in chemical depictions are bonds and atoms. Bonds usually contain the data about which atoms have to be linked and at which manner. Deducing a relatively accurate determination where the bond is situated in the picture is very important. Besides the linking data, bond collections can represent themselves as a set of atoms. The aromatic ring includes numerous carbon atoms and is symbolized using a group of linked bonds. It is necessary to process line drawings using a convenient technique for the reconstruction of molecule depictions, *Figure 4* shows types of bonds.

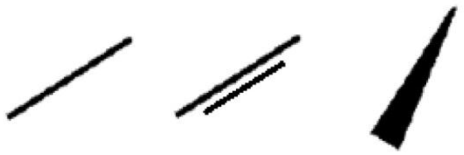


Figure 4 Bond types

Chemical structure representations, bonds are usually drawn as straightforward lines (single or double). Hence, a powerful algorithm for line-detection methodology is the fundamental software component for features' extracting of those bonds from chemical formulas. The Hough transform (HT) methodology [14] is a typical way that is used to extract features of the selected bonds in computerized image processing. It maps the depiction in the Cartesian space to the polar Hough space to recognize contained lines.

In the x-y space, collinear pixels intersect sinusoidal lines when a pixel corresponds to a sinusoidal curve in the Hough space. As a result, all potential lines passing via every possible pair of pixels in a chemical formula structure are recognized by checking the overlapping dots of curves in the HT space, as shown in *Figure 5*.

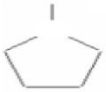


Figure 5 Lines detection

2.6 Data extraction and topology construction

For final data presenting, the chemical structure is defined using a graph and will be interpreting depending on the verified chemical symbols (atoms or bonds). All center points of the identified chemical symbols and other endpoints for the verified bonds are marked as a node. Then, between all those identified nodes, the nodes which exist in a specific space are integrated into a monocular node. A node-edge linking-table is created depending on this chemical graph structure, that will be transformed into a SMILE string [5].

Extracting the structure of a molecule equates to an enumeration of individual atoms taking place on the final proposed diagram, *Figure 6*. Besides that, there will be an addition of hydrogen atoms to all individual unsaturated carbon atoms in the chemical structure. The following algorithm generates the formula as:

```
Algorithm 1 (Generating chemical structure)
        In: diagram D (B,G)
        Out: chemical structure
        Methods:
    •   Set an initial counter for the structure
    •   For every b belongs to B
          a.  If b contains label C:
                i.   Adding C to the formula counter
                ii.  Computing valence v of b by
                     adding every g belongs to G
                iii. Adding 4-v H atoms to the
                     counter of the formula
          b.  Else, Adding v label to the counter
              of the formula
```

Applying this algorithm will result in the final molecule structure of C6H3N.

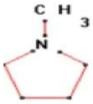


Figure 6 Final image reconstruction

2.7 SMILES representation

SMILES notation is a string representation of atoms, bonds, and their connectivity. In SMILES, all atoms are identified by their normal symbols except Hydrogen which can either be omitted or displayed and are characterized by their atomic symbols. Each non-hydrogen atom is stated independently by its atomic symbol surrounded by square brackets, [ ] (for instance, [Fe] or [Au]). Square brackets can be omitted for components at the "organic subset" (B, F, Cl, Br, P, S, C, N, O, and I) if the appropriate number of "implicit" hydrogen atoms is assumed. "Explicitly" attached hydrogens and formal charges are usually defined inside brackets. A formal charge is characterized by one of the symbols + or -. Single, double and aromatic bonds are characterized by the symbols, -, = and #. Aromatic and single bonds are omitted. Following are some instances for SMILES strings:

- C Methane (CH4)
- C=C Ethene (CH2CH2)
- CC Ethane (CH3CH3)
- CCO Ethanol (CH3CH2OH)
- CC=O Acetaldehyde (CH3-CH=O)
- C#C Ethyne (CHCH)
- COC Dimethyl ether (CH3OCH3)

Branches are itemized by enclosures in parentheses and can be nested or stacked, as shown in:

- CC(C)CO Isobutyl alcohol (CH3-CH(CH3)-CH2-OH)

Rings are characterized by breaking one single or aromatic bond in each ring, and labeling this ring-closure point with a digit directly following the atoms linked through the broken bond. Atoms in aromatic rings are itemized by lower case letters. Therefore, cyclohexane and benzene can be represented by the following SMILES:

- C1CCCCC1 Cyclohexane (C6H12)
- c1ccccc1 Benzene (C6H6)

Although the carbon-carbon bonds in these two SMILES are omitted, it is possible to deduce that the omitted bonds are single bonds (for cyclohexane) and aromatic bonds (for benzene). For instance, the following is a valid SMILES string for benzene.

- C1=CC=CC=C1 Benzene (C6H6)

To represent different types of bonds, SMILES puts nothing between atoms connected via single bonds, a = between atoms connected via the double bond and a # to represent a triple bond. Also, stereotype bonds (3D bonds) are presented by a @@. Additionally, SMILES use numbers to indicate ring closures and parentheses to imply branches. SMILES notation is not necessarily unique. For example, ([NC5]) C([H3]) is considered a valid representation of our molecule in *Figure 1*, as the indication of the bonds between carbon atoms and entries in square brackets, are molecular substructures.

3.Testing and results

Despite only a few approaches are conducting with the chemical molecule recognition problem, the image processing for patterns and features recognition is not considered a modern research field. Examples of such approaches are; OSRA[15], CliDE[16] and chemOCR[17]. Two sets of chemical structure images collected from different sources were used to test the current approach precision and compare it to OSRA V4.01, CLiDE V3.1, and chemOCR V3.0 (*Table 2*). Set I contain 50 images obtained by querying chemical and pharmaceutically important molecules to Google image search (http://images.google.com/) so the images have various sizes, drawing styles, resolutions, and font. Fifty ligand images are contained in Set II which are selected from the PubChem database (about 120 scientific papers). The performances of chemical structure recognition are examined in two characteristics: the fraction of correct results and the ability to identify chemically significant substructure patterns. Precision measurement is the recognition of the extracted patterns that are correct. Though an error exists in the resulted molecule, it wouldn't be observed as a useless one if chemically important features-of-interests are well-classified. For instance, the misassignment of bondstereo or atom charge may not be so serious for finding molecules similar to the classified structure in a database. Therefore, we compute the precision to evaluate the approach's ability to extract chemical substructure patterns.

Table 2 Comparison of chemical-structure recognition systems precision

	Current		**OSRA**		**CLiDE**		**chemOCR**	
	Set I	**Set II**	**Set I**	**Set II**	**Set I**	**Set II**	**Set I**	**Set II**
Atom (Character recognition)	0.95	0.83	0.95	0.92	0.95	0.59	0.88	0.57
Bond (Lines)	0.94	0.86	0.89	0.73	0.83	0.60	0.89	0.68
Ring (Connected lines)	0.84	0.75	0.92	0.68	0.90	0.52	0.88	0.70
Simple patterns	0.89	0.89	0.93	0.71	0.80	0.68	0.82	0.70

4.Discussion

To produce an automated pattern recognition model, there are still more challenges that remain to be addressed. For the automated extraction of chemical structures and related information from scientific journal researches, it would be significant to rapidly differentiate between a diagram of a non-chemical structure and a chemical structure diagram between the extracted images. Such functionality still has to be merged into our proposed approach. Finally, since the transformation of the chemical structure of a digital image to a computerized chemical format as seen in the test is highly error-prone, output structures should be thoroughly examined before usage. Also, manual curation resulting in the high cost of system operation and filtering technique which can detect wrong outputs and filtered them out at the pre-processing stages may be accurate to expand the performance of automated systems for chemical structures recognition. In this way, accuracy can be increased at the expense of productivity. The limitations of the current approach are generally its need for manual feeding of images and its significant error rates, especially at ring pattern recognition. As an alternative, we have a plan to design a modern chemical structure designer software that will be easier at usage and contains the needed atoms and bond types to perform an accurate structure with the equivalent computer-readable format.

5.Conclusion

Computational drug development is an efficient strategy for accelerating and economizing the drug discovery process. Because of the dramatic increase in the availability of biological macromolecule and small molecule information. The applicability of computational drug discovery has been extended and broadly applied to nearly every stage in the drug discovery and development workflow, including target identification and validation, lead discovery and optimization and preclinical tests. Therefore, our proposed technique is considered an information extraction model that supports the work of scientists with a special interest in biochemical entities recognizing and recent enhancements for the sake of advanced extraction of chemical formulas representations. The precision rates for classification of substructure patterns (character recognition, lines, rings, and simple atoms) for the proposed approach that were mentioned in the previous section are accepted but not the optimum. More modifications are still to be done, as future suggestions, and hopefully, it will give better results.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. Pubchem substance and compound databases. Nucleic Acids Research. 2015; 44(D1): D1202-13.

[2] Chen H, Sharp BM. Content-rich biological network constructed by mining pubmed abstracts. BMC Bioinformatics. 2004.

[3] Li Z, Wan H, Shi Y, Ouyang P. Personal experience with four kinds of chemical structure drawing software: review on Chemdraw, Chemwindow, ISIS/draw, and Chemsketch. Journal of Chemical Information and Computer Sciences. 2004; 44(5):1886-90.

[4] Gkoutos GV, Rzepa H, Clark RM, Adjei O, Johal H. Chemical machine vision: automated extraction of chemical metadata from raster images. Journal of Chemical Information and Computer Sciences. 2003; 43(5):1342-55.

[5] Weininger D, Weininger A, Weininger JL. SMILES. 2. algorithm for generation of unique SMILES notation. Journal of Chemical Information and Computer Sciences. 1989; 29(2):97-101.

[6] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. In ChI, the IUPAC international chemical identifier. Journal of Cheminformatics. 2015; 7(23):1-34.

[7] Ofner J, Brenner F, Wieland K, Eitenberger E, Kirschner J, Eisenmenger-Sittner C, et al. Image-based chemical structure determination. Scientific Reports. 2017.

[8] Da Cunha MM, Trepout S, Messaoudi C, Wu TD, Ortega R, Guerquin-Kern JL, et al. Overview of chemical imaging methods to address biological questions. Micron. 2016; 84:23-36.

[9] Ziatdinov M, Dyck O, Maksov A, Li X, Sang X, Xiao K, et al. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. ACS Nano. 2017; 11(12):12742-52.

[10] Laanait N, Ziatdinov M, He Q, Borisevich A. Identifying local structural states in atomic imaging by computer vision. Advanced Structural and Chemical Imaging. 2016; 2:1-11.

[11] Mallea MD, Meltzer P, Bentley PJ. Capsule neural networks for graph classification using explicit tensorial graph representations. arXiv preprint arXiv:1902.08399. 2019.

[12] Pitas I. Digital image processing algorithms and applications. John Wiley & Sons; 2000.

[13] Fletcher LA, Kasturi R. A robust algorithm for text string separation from mixed text/graphics images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1988; 10(6):910-8.

[14] Illingworth J, Kittler J. A survey of the hough transform. Computer Vision, Graphics, and Image Processing. 1988; 44(1):87-116.

[15] Filippov IV, Nicklaus MC, Kinney J. Improvements in optical structure recognition application. In IAPR international workshop on document analysis systems, Boston, MA 2010.

[16] Ibison P, Jacquot M, Kam F, Neville AG, Simpson RW, Tonnelier C, et al. Chemical literature data extraction: the CLiDE Project. Journal of Chemical Information and Computer Sciences. 1993; 33(3):338-44.

[17] Algorri ME, Zimmermann M, Friedrich CM, Akle S, Hofmann-Apitius M. Reconstruction of chemical molecules from images. In annual international conference of the IEEE engineering in medicine and biology society 2007 (pp. 4609-12). IEEE.

Amena Mahmoud is an Assistant Lecturer in the Computer Science Department at Kafr El Sheikh University, Egypt. She is a Ph.D. Researcher in the Computer Science department at the Faculty of Computers and Information Mansoura University, Egypt. Her research areas are BioInformatics and Artificial Intelligence topics like Pattern Recognition, Machine Learning, and Image Processing.
Email: amena_mahmoud@fci.kfs.edu.eg

Prof. Taher Tawfiq Hamza is a Professor of Computer Science at Mansoura University, Egypt. Professor Taher holds a Ph.D. in Computer Science from the Faculty of Science Mansoura University in Egypt. He has served as head of the Computer Science Department and a Vice-Dean of faculty of Computers and Information Systems Mansoura University.
Email: taher_hamza@yahoo.com

Prof. Magdi Zakaria Rashad is a Professor of Computer Science at Mansoura University, Egypt. Professor Magdi holds a Ph.D. in Computer Science from the Faculty of Engineering at Cairo University, Egypt. He is the author of more than 160 papers published in refereed international journals. He has served as head of the Computer Science Department and a Vice-Dean of faculty of Computers and Information Systems Mansoura University. He has also served as a reviewer for various international journals, such as IJCSIT, MJCSIS and he is interested in the following fields: Artificial Intelligence, Pattern Recognition, Machine Learning, Image Processing, and Cloud Computing.
Email: magdi_z2011@yahoo.com