People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Larbi Tébessi University - Tébessa
Faculty of Exact Sciences and Natural and Life Sciences
Department: Mathematics and Computer Science

End of study thesis
For obtaining the MASTER diploma
Field: Mathematics and Computer Science
Sector: IT
Option: Information systems

Theme

TOWARDS AN INTELLIGENT APPROACH FOR THE DETECTION AND CLASSIFICATION OF CANCER OF THE LYMPHATIC SYSTEM

Presented by :

Attia Djamilla

In front of the jury:

Dr.Djeddi Chawki	MCA	Larbi Tébessi University	President
Dr.Marzoug Soltane	MCB	Larbi Tébessi University	Examiner
Dr.Bendib Issam	MCB	Larbi Tébessi University	Framer

2020/2021

ACKNOWLEDGMENTS

First of all, I thank God who has guided us on the right path.

I thank my supervisor Mr. Bendib Issam for guiding me throughout my work, with a lot of effort, patience, experience and valuable advice.

Also, thank you all for his presence, his patience, his understanding, his human qualities, his interest in the subject of our work, and his confidence in us.

I also thank the members of the jury, the President Mr. Djeddi C. and the examiner Mr. Marzouk S. for agreeing to discuss our end of study project to benefit from their experience and advice.

Finally, I would like to thank all my teachers in the Department of Mathematics and Computer Science.

DEDICATION

To my parents,

To my brothers and sisters,

To my fiance.

Abstract

Determining cancer and its type is a very difficult task that requires high medical expertise and skills. With the development of image classification techniques, deep learning strategies have occupied the first positions in many medical image classification systems as part of computer aide decision (CAD).

The aim of this study is to accurately classify lymphoma subtypes using deep learning. A deep learning framework has been proposed to classify three types of lymphomas as follicular lymphoma (FL), chronic lymphocytic lymphoma (CLL) and Mantle Cell Lymphoma (MCL) by following pretrained CNN models (Transfer learning) such as Resnet and VGG and based on the available dataset from the National Institute on Aging (NIA). The data Patching was implemented for the first step of data processing, where the achieved results show that the proposed models were able to achieve better results compared to CNN built from scratch.

Keywords: Cancer, lymphoma, CLL, FL, MCL, CAD, Deep Learning, Transfer learning , CNN , VGG , Resnet ,Patching , NIA.

Résumé

Déterminer le cancer et son type est une tâche très difficile qui nécessite une expertise et des compétences médicales élevées. Avec le développement des techniques de classification d'images, les stratégies d'apprentissage en profondeur ont occupé les premières positions dans de nombreux systèmes de classification d'images médicales dans le cadre des décisions assistée par ordinateur (DAO)

Le but de cette étude est de classer avec précision les sous-types de lymphomes à l'aide de l'apprentissage en profondeur. Un cadre d'apprentissage en profondeur a été proposé pour classer trois types de lymphomes en tant que lymphome folliculaire (LF), lymphome lymphocytaire chronique (LLC) et Lymphome à cellules du manteau (LCM) en suivant des modèles CNN prétraité (apprentissage par transfère) tels que Resnet et VGG et sur la base des données public de l'Institut national sur le vieillissement (NIA). La fragmentation des données a été mise en œuvre pour la première étape du traitement des données, où les résultats obtenus montrent que les modèles proposés ont pu obtenir de meilleurs résultats par rapport à CNN construit à partir de zéro.

Les mots clés: Cancer , lymphome , LLC , LF , LCM, DAO , Apprentissage en profondeur , Apprentissage par transfer , CNN, VGG, Ressent , Fragmentation , NIA.

الملخص

ان تحديد مرض السرطان ونوعه مهمة صعبة للغاية تتطلب خبرة ومهارات طبية عالية .مع تطور تقنيات تصنيف الصور، احتلت إستراتيجيات التعلم العميق المراتب الاولى في العديد من انظمة تصنيف الصور الطبية ضمن أنظمة دعم القرار.

الهدف من هذه الدراسة هو التصنيف الدقيق للأنواع الفرعية من سرطان الغدد الليمفاوية باستخدام التعلم العميق. تم اقتراح إطار عمل تعليمي عميق لتصنيف ثلاثة أنواع من الأورام اللمفاوية مثل ورم الغدد اللمفاوية المسامي (FL ،)ورم الغدد الليمفاوية المزمنةCLL) (وورم الغدد الليمفاوية للخليةMCL)(وذلك باتباع نماذج الشبكة العصبية الالتفافية المدربة مسبقا مثل Resnetو VGGو بالاعتماد على مجموعة البيانات المتاحة للجمهور من المعهد الوطنيNIA). (تم تنفيذ تقسيم البيانات كخطوة اولى لمعالجة البيانات، حيث تظهر النتائج المحققة ان النماذج المقترحة تمكنت من تحقيق نتائج أفضل مقارنةً بالشبكات العصبية الالتفافية المبنية من الصفر.

الكلمات المفتاحية :مرض السرطان، أنظمة دعم القرار ، التعلم العميق، نماذج الشبكة العصبية الالتفافية المدربة مسبقا ، تقسيم البيانات ، VGG، Resnet ،NIA، LLC ، LF ، LCM .

Summary

Chapter 1: General study on the different categories of cancer of the lymphatic system

Chapter 2 : State of the art

Chapter 3: Detection and classification of cancer of the lymphatic system

List of tables

List of figures

General Introduction

1. Introduction

Cancer has spread all over the world and it has become more important than ever to understand its structure and try to control it in the body of the sufferers. Lymphoma is a type of cancer that affects the lymphatic system and has more than 38 subtypes. In this study, we were interested in classifying 3 types of lymphoma disease including chronic lymphocytic leukemia (CLL), follicular lymphoma (FL) and mantle cell lymphoma (MCL), which are descended from the category of non-Hodgkin's lymphoma (NHL). This category corresponds to 70% of cases of lymphomas whose classification and stratification remains a major challenge for pathologists. CLL represents a different type of lymphoma variant, so both are treated in the same way. FL is the second most common type of B-cell lymphoma. With regard to MCL, it is characterized by an aggressive clinical progression with few long-term survivors.

Visual assessment is a complex task due to the level of expertise, and diversity among pathologists. Digitization of specimens has therefore offered important advantages in the search for diagnoses using computational techniques, called computer-aided diagnosis (CAD). These systems have the potential to assist pathologists in their clinical decisions linking image information to disease features.

In order to make a research contribution, in this project we aim to build deep learning models with the aim of building a computer-aided diagnosis (CAD) system, which is a growing field of medical data analysis with the aim of helping clinicians make the correct diagnosis.

Our work targets the three main types of NHL and we rely on convolutional neural networks to build models.

2. Organisation

In order to describe the work done, and after presenting the general introduction, we divided the remainder of this research into 3 chapters:

- The first chapter is devoted to a general study of the various categories of lymphatic cancer.

- The second chapter is devoted to the study of the relevant works and the conclusion of our synthesis.
- The third chapter is devoted to our contribution and the steps to achieve it.
- Finally, we conclude the note with a general conclusion and some future works.

Chapter 1

General study on the different categories of cancer of the lymphatic system

1. Introduction

Although many new diseases threaten human life such as AIDS and viruses such as Ebola and bird flu, cancer remains one of the deadliest human diseases in the world. Statistics claim that the number of deaths is more than 200 deaths / day. According to the World Health Organization, cancer is the second leading cause of death in the world, with nearly one in every 6 deaths worldwide due to cancer.

The scientist has changed his way of life through processed foods, a sedentary lifestyle, stress and an unbalanced pace of life. There are many types of cancer and where they appear in the human body. In this chapter, we present a general study of the different categories of lymphatic cancer.

2. cancer definition

Cancer is a disease caused by transformed cells that become abnormal and proliferate excessively. These disturbed cells eventually form a lump called metastasis. Cancer cells tend to invade nearby tissues and move away from the tumor. Then they migrate through the blood vessels and lymph vessels to form another tumor (metastasis) [1].

3. Cancer causes

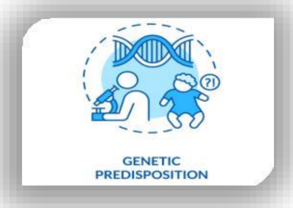
According to recent studies, very few cancers are caused by just one thing. Most seem to be caused by a complex set of many risk factors, but cancer can develop in people who do not have any risk factors. The latter is a substance or condition that increases the risk of Cancer. Cancer risk factors can play different roles in the onset of cancer and its development. These factors are:

Tobacco:

Tobacco has a strong link with the development of many cancers, especially lung cancer, as oncologists believe that tobacco is the number one cause of lung cancer. It is the cause of 90% of cases. Lung cancer is rare in non-smokers. Because less than 1% of non-smokers develop lung cancer [2].

Exposure to Rays:

There are several rays (UV, X…) which are all dangerous for health and which can cause skin cancer. The International Agency for Research on Cancer (IARC) has estimated that at least 85% of melanomas are caused by exposure to sunlight.



Genetic predisposition:

Some people are more predisposed to cancer than others due to mutations in genes which are passed down from generation to generation.

Obesity:

Obesity is a major challenge that appeared at the end of the 20th century, which appeared as a result of poor diet as well as a sedentary lifestyle that leads to overweight, thus increasing the chances of contracting cancer (stomach, colon, liver, pancreas, etc.).

Unhealthy Diet:

It is the main source of energy for the human body but unfortunately in modern times food contains a lot of industrial ingredients but also pesticide residues, according to some study 30% of cancer are caused by poor diet,

Sedentary lifestyle:

Physical inactivity has a very close link with cancer the more sedentary a person is, the more it increases the chance of getting cancer but also of other diseases such as diabetes and cardiovascular disease.

Alcohol:

Alcohol is a proven carcinogen. It is responsible for 11% of cancers in men and 4.5% of cancers in women (according to the World Health Organization).

Pollution:

Pollution is found in asbestos, gases emitted by car exhausts, as well as fine particles, plastic, etc.

according to research estimated that pollution is thought to be involved in the development of approximately 1 to 4% of all cancer [2] .

4. Cancer development

The body is made of cells brought together to form tissues and organs. Genes in the nucleus of each cell tell it when to grow, function, divide and die. Usually, our cells follow these guidelines. But when our DNA changes or is damaged, the gene can be transformed. Transgenes do not function properly [3] . Thus, cells that should rest can divide and grow in an unorganized manner, which may lead to cancer (Fig.1).

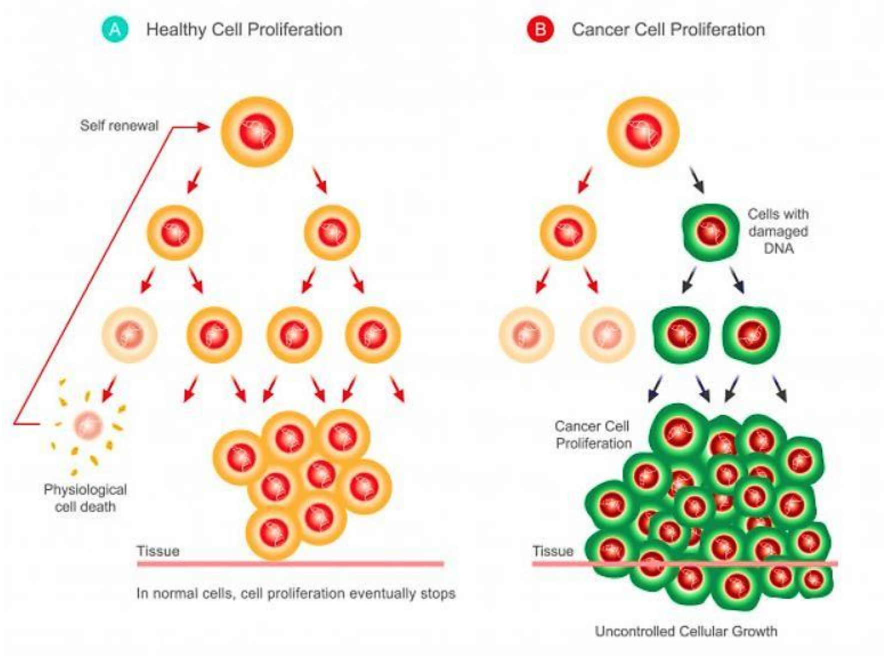

A
Healthy Cell Proliferation
B
Cancer Cell Proliferation
Self renewal
Cells with damaged DNA
Physiological cell death
Cancer Cell Proliferation
Tissue
Tissue
In normal cells, cell proliferation eventually stops
Uncontrolled Cellular Growth

Fig 1. 1:Cancer development [3]

When cells divide properly, they make exact copies of themselves. A cell divides into two identical cells then these two cells divide into 4 and so on. Usually, in adults, cells grow and divide to produce more cells only when the body needs them, such as replacing aging or damaged cells.

But cancer cells are different. Cancer cells contain genetic mutations that transform a normal cell into a cancer cell. These gene mutations can be hereditary, and develop over time just like the rest of us.

Cancer cells are affected by age and genetics differently from normal cells. It starts to grow and divide randomly instead of dying when it should.

5. Type of cancer

- **Liver cancer:** occurs when abnormal cells grow out of control in its tissue.
- **Pancreatic cancer: starts** in the cells of the pancreas. The cells of the pancreas sometimes undergo changes that make the way they grow or behave abnormally. These changes can lead to the formation of tumors [4].
- **Lung cancer:** starts in the cells of the lung. Globally, lung cancer is the most common cancer in terms of both incidence and mortality. Lung cancers are divided into non-small cell lung cancer and small cell lung cancer according to the type of cell they develop from.
- **breast cancer:** cells can stay in the breast or spread throughout the body through blood or lymph vessels. Most of the time, the progression of breast cancer takes several months and even a few years [5].
- **Skin cancer:** Skin cancer is characterized by malignant tumors that attack skin cells. One of the causes of skin cancer would be caused by too much exposure to UV, whether natural (the sun) or artificial (tanning booths).
- **Lymphoma cancer:** is a cancer of the lymphatic system, which is part of the body's germ-fighting network. The lymphatic system includes the lymph nodes (lymph glands), spleen, thymus gland and bone marrow. Lymphoma can affect all those areas as well as other organs throughout the body.

6. The lymphatic system

The lymph system is a series of lymph nodes and vessels that move lymph fluid through the body. Lymph fluids contain infection-fighting white blood cells. Lymph nodes act as filters, capturing and destroying bacteria and viruses to prevent infection from spreading. While the lymph system typically protects your body, lymph cells called lymphocytes can become cancerous. The names for cancers that occur in the lymph system are lymphomas. Doctors classify more than 70 cancer types as lymphomas. Lymphomas can affect any portion of the lymphatic system, including (Fig.2) [6] :

- ✓ Bone marrow
- ✓ Thymus
- ✓ Spleen
- ✓ Tonsils
- ✓ Lymph nodes

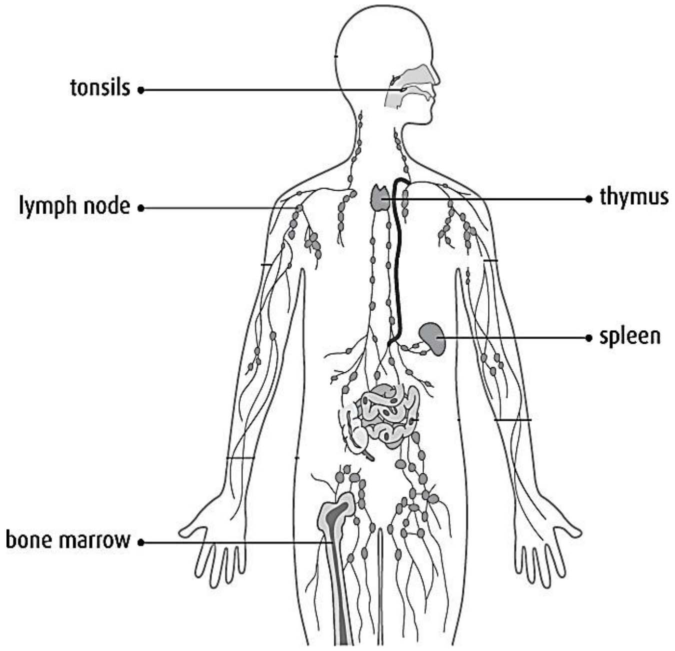


Fig 1. 2:The lymph system [7]

Medical science usually divides lymphomas into two categories:

- ❖ Hodgkin's lymphoma
- ❖ non-Hodgkin's lymphoma (NHL).

6.1. Causes of lymphoma

Cancer is the result of uncontrolled cell growth. The average lifespan of a cell is brief, and then the cell dies. In people with lymphoma, however, the cell thrives and spreads instead of dying.

It’s unclear what causes lymphoma, but a number of risk factors are connected with these cancers.

6.2. Symptoms of lymphoma

Common symptoms include:

- Swollen lymph nodes in the neck, armpits or groin
- Abdominal pain or swelling
- Chest pain, coughing or trouble breathing
- Fatigue
- Fever
- Night sweats
- Weight loss (usually unexplained weight loss and as much as 10% or more of their body weight)
- Loss of appetite

6.3. Types of lymphoma

The lymphoid neoplasms are classified according to multiple properties (Fig.3):

✓ Whether they are Hodgkin lymphoma or non-Hodgkin lymphoma

✓ Whether they originate from T or NK cells or from B cells

✓ Whether they are aggressive, indolent (not so aggressive) or in-between (“other”)

✓ Whether they originate from precursor cells (which don’t express CD20) or mature cells (which express CD20)

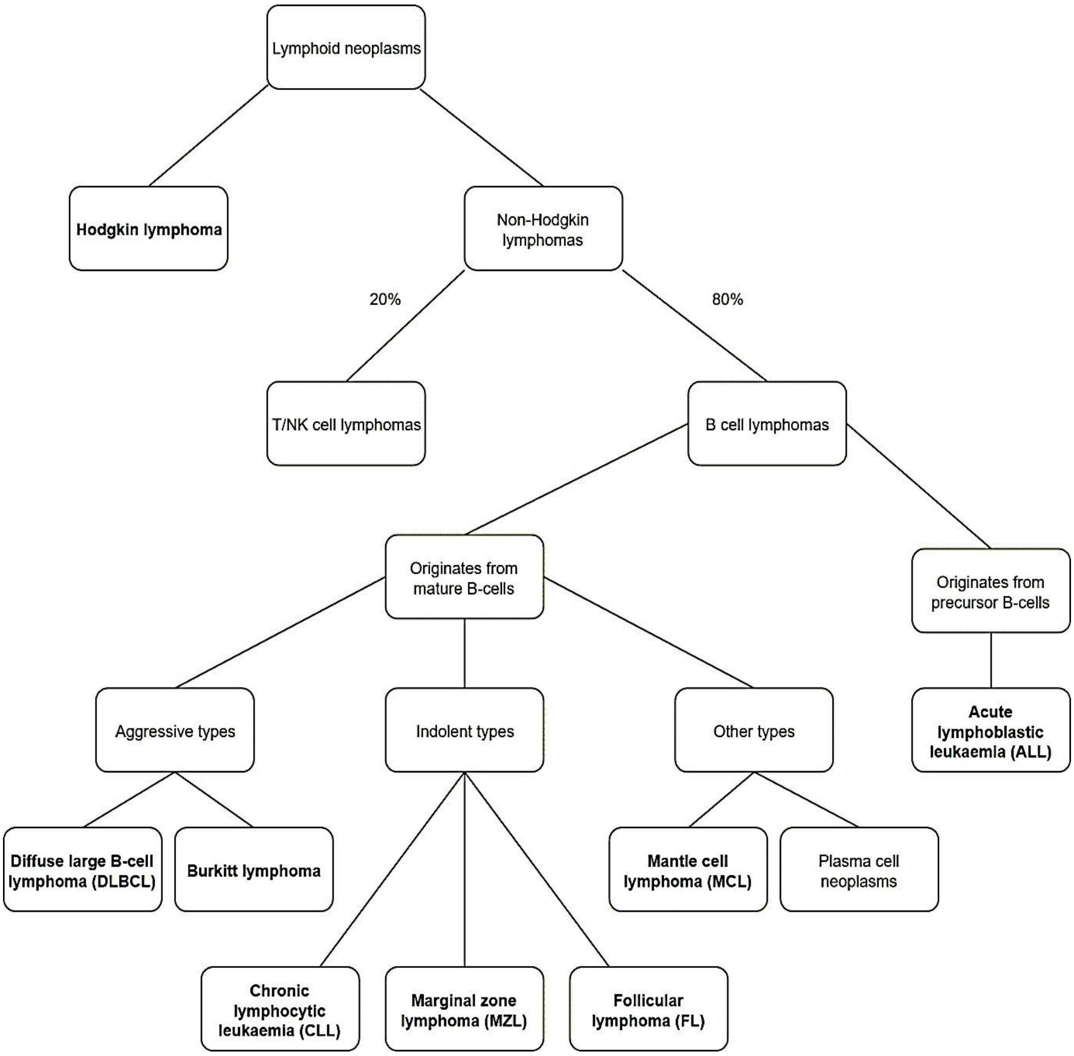

Lymphoid neoplasms
Hodgkin lymphoma
Non-Hodgkin lymphomas
20%
80%
T/NK cell lymphomas
B cell lymphomas
Originates from mature B-cells
Originates from precursor B-cells
Aggressive types
Indolent types
Other types
Acute lymphoblastic leukaemia (ALL)
Diffuse large B-cell lymphoma (DLBCL)
Burkitt lymphoma
Mantle cell lymphoma (MCL)
Plasma cell neoplasms
Chronic lymphocytic leukaemia (CLL)
Marginal zone lymphoma (MZL)
Follicular lymphoma (FL)

Fig 1. 3:Type and sub-types of lymphoma [9]

6.3.1. Hodgkin lymphomas:

usually begin in B cells or immune system cells known as Reed-Sternberg (RS) cells. While the main cause of Hodgkin lymphoma is unknown, some risk factors can increase the chances of developing this type of cancer. There are four subtypes based on lymph node disease, including the common subtypes of HL and mixed nodal sclerosis, and a rare variety. Clinical presentations of these types vary. Classically streptococcal disease appears as an early-stage disease in young adults, with it predominates in females, and generally has an excellent response to treatment. Mixed cellular and dominant forms of lymphocytes may present later in life, with a tendency to present with advanced or external disease. The differences between lymphoma subtypes help determine the most appropriate treatment and can predict the clinical behavior of a specific tumor and response to treatment [8]

6.3.2. Non-Hodgkin's lymphoma

According to the Leukemia & Lymphoma Society (LLS), NHL is three times more common than Hodgkin's lymphoma.Many lymphoma types fall under each category. Doctors call NHL types by the cells they affect, and if the cells are fast- or slow-growing. NHL forms in either the B-cells or T-cells of the immune system. According to LLS, most NHL types affect B-cells [8].

As you can see from the figure above are there many lymphoid neoplasms that are classified as "non-Hodgkin lymphoma". Most (80%) of these originate from B-cells.

Lymphomas may be nodal or extranodal, where "nodal" refers to a lymph node. A lymphoma is nodal if it originates in a lymph node, and primary extranodal if it starts in other lymphoid tissues, like the lymphoid tissues in the GI tract, the skin or in the CNS. A lymphoma can be "secondary" extranodal if it spreads from a lymph node to other tissues.

Aggressive lymphomas have worse prognosis than the indolent ones. They can be considered "high-grade". They proliferate rapidly and are more frequent in children. Aggressive lymphomas counterintuitively have a higher chance of being treatable.

Indolent lymphomas are much less harmful than the aggressive ones; in some cases they may not even require treatment. They can be considered "low-grade". These tumors are more frequent in older patients. The tumor cells don't proliferate much. In most cases are indolent lymphomas uncurable.

Lymphoid neoplasms that affect lymph nodes cause painless and firm lymphadenomegaly. Infiltration of tumor cells into the spleen and liver can cause spleno-and-hepatomegaly as well. Proliferating tumor cells in the bone marrow displace and destroy the normal cells of the bone marrow, which can cause bone marrow failure with anaemia, thrombocytopaenia and neutropenia.

The neutropenia can predispose to systemic infections. Autoimmune haemolytic anaemia may occur in patients with lymphoid neoplasms.

The term B symptoms refer to systemic symptoms of fever, night sweats and weight loss, which are common symptoms in patients with lymphoid neoplasms [9] .

1. Follicular lymphoma

Follicular lymphoma (FL) is an indolent, mature, non-Hodgkin B-cell lymphoma. It's the most common indolent non-Hodgkin lymphoma and the second most common non-Hodgkin lymphoma overall. It mostly affects adults older than 50 years.

It very frequently involves at (14;18) translocation where the gene BCL2 on the chromosome 18 is fused to the immunoglobulin heavy chain gene (IgH) on chromosome 14. This causes overexpression of the antiapoptotic protein Bcl-2, which contributes to tumor cell survival.

It usually manifests as slowly progressive, painless, generalized lymphadenopathy. The name "follicular" comes from the fact that the neoplastic cells form "neoplastic follicles" in lymph nodes. The histology of follicular lymphoma can be studied in the slide section.

The disease is not curable, and therapy is therefore usually not performed. 40% of cases of FL progress into diffuse large B-cell lymphoma. [9]

2. Chronic lymphocytic leukaemia

Chronic lymphocytic leukaemia (CLL) is an indolent, mature, non-Hodgkin B-cell lymphoma. It's the most common adult leukaemia and the most frequent lymphoma in Hungary. It also mostly affects adults older than 50 years. It's called "leukaemia" because tumor cells are found in the blood in 90-100% of cases. CLL/SLL cells surpresses normal B-cell functions, often resulting in hypogammaglobulinemia. Some patients also have autoantibodies from nonmalignant bystander B-cells against RBC or platelets, suggesting tumor cells somehow impair immune tolerance.

The tumor cells of CLL contains high amounts of Bcl-2, however there is no rearrangement of the BCL2 gene involved. Some evidence suggests that the BCL2 gene is upregulated due to loss of regulatory microRNAs.

Symptoms of CLL include painless lymphadenomegaly and splenomegaly, symptomless leukocytosis and B symptoms. Due to the non-specific symptoms is CLL often an incidental finding, which is discovered during a routine blood test. The liver and bone marrow may also be affected – the bone marrow failure may cause anaemia and thrombocytopaenia. If the CLL is limited to a lymph node is the condition called *small lymphocytic lymphoma*.

Two types of CLL exist: the prefollicular type and the postfollicular type. The main difference is whether the immunoglobulin heavy chain gene (IgH) has been mutated or not. The differences are summed up here [9] :

Table 1. 1:comparison of types of CLL [9]

	Prefollicular type	**Postfollicular type**
IgH gene status	Not mutated	Mutated
Underwent somatic hypermutation	No	Yes
Prognosis	Much worse	Better
Treatment	More frequently required	Less frequently required

3. Mantle cell lymphoma

Mantle cell lymphoma (MCL) is a mature, non-Hodgkin B-cell lymphoma belonging to the “other” group. It originates from mantle zone B-cells and mostly affects adult males. It’s associated with lymphadenomegaly and bone marrow infiltration. It affects the peripheral blood in only 20% of cases and is therefore not considered a leukaemia.

It sometimes arises in the gastrointestinal tract, where it manifests as multifocal submucosal nodules that resemble polyps. These can be mistaken for polyps due to other diseases, like familiar adenomatous polyposis, hamartomatous polyps or ulcerative colitis.

MCL involves a t(11;14) translocation where the cyclin D1 gene is fused to the IgH locus. Cyclin D1 is involved in cell proliferation and is overexpressed as a result of this translocation.

It’s an aggressive cancer; the mean survival is 3-5 years [9].

4. Marginal zone lymphoma

Marginal zone lymphoma (MZL) is an indolent, mature, non-Hodgkin B cell lymphoma. The name comes from how it develops from marginal zone B-cells. Three subtypes exist:

- Extranodal marginal zone lymphoma – most common type
- Nodal marginal zone lymphoma
- Splenic marginal zone lymphoma – least common type

Extranodal marginal zone lymphoma, also called MALT lymphoma, arises most commonly from lymphoid tissue in organs such as the stomach, salivary glands, intestines, lungs, orbit or breast. They are associated with chronic infection, such as that seen in H. pylori gastritis. They are also associated with autoimmune disorders, like Sjögren's (MALT lymphoma in the salivary glands).

Nodal marginal zone lymphoma arises in lymph nodes, most commonly cervical lymph nodes. It causes lymphadenopathy.

Splenic marginal zone lymphoma arises in the spleen or bone marrow and causes splenomegaly [9].

6.4. Stage of lymphoma

Like any cancer, there are five stages of lymphatic system cancer that are summarized in the following table

Table 1. 2:Lymph cancer stages [10].

Stage	Definition (N = Non-Hodgkins, H = Hodgkins)
0	• Cancer is confined to site of origin, no spread
I	• Lymphoma in only 1 lymph node area or lymphoid organ • Or lymphoma found in only 1 area of a single organ outside the lymph system
II	• Lymphoma in 2+ groups of lymph nodes on the same side of diaphragm

	• Lymphoma extends from a single group of lymph node(s) into nearby organ
III	• The lymphoma is found in lymph node areas on both sides of (above and below) the diaphragm • The cancer may also have spread into an area or organ next to the lymph nodes, into the spleen, or both
IV	• The lymphoma has spread outside lymph system into organ not next to involved node • The lymphoma has spread to the bone marrow, liver, brain or spinal cord, or the pleura (thin lining of the lungs). Lymphoma Staging

6.5. Diagnosing of lymphoma

A biopsy typically is taken if a doctor suspects lymphoma. This involves removing cells from an enlarged lymph node. A doctor known as a hematopathologist will examine the cells to determine if lymphoma cells are present and what cell type they are. If the hematopathologist detects lymphoma cells, further testing can identify how far the cancer has spread. These tests can include a chest X-ray, blood testing, or testing nearby lymph nodes or tissues. Imaging scans, such as a computed tomography (CT) or magnetic resonance imaging (MRI) scans may also identify additional tumors or enlarged lymph nodes [11].

6.5.1. Physical exam

Your doctor checks for swollen lymph nodes, including in your neck, underarm and groin, as well as a swollen spleen or liver.

6.5.2. Removing a lymph node for testing

Your doctor may recommend a lymph node biopsy procedure to remove all or part of a lymph node for laboratory testing. Advanced tests can determine if lymphoma cells are present and what types of cells are involved.

6.5.3. Blood tests

Blood tests to count the number of cells in a sample of your blood can give your doctor clues about your diagnosis.

6.5.4. Removing a sample of bone marrow for testing

A bone marrow aspiration and biopsy procedure involve inserting a needle into your hipbone to remove a sample of bone marrow. The sample is analyzed to look for lymphoma cells.

6.5.5. Imaging tests

Your doctor may recommend imaging tests to look for signs of lymphoma in other areas of your body. Tests may include CT, MRI and positron emission tomography (PET).

7. The CAD

One of the biggest problems that classic image processing methods suffer from is the difficulty of analyzing images, the problem of false results, and also the requirement to provide more than one experienced specialist. In order to navigate this stage and overcome these difficulties, computer-aided decision-making systems have emerged, which act as a link between computer science and the field of medicine. Artificial intelligence has shown its role in this field by relying on machine learning and deep learning techniques, as it has been used to solve several problems in classifying and predicting many diseases, especially cancer [12].

Computer aided diagnosis (CAD) is the use of a computer-generated output as an assisting tool for a clinician to make a diagnosis. It is different from automated computer diagnosis, in which the end diagnosis is based on a computer algorithm only.

As an early form of artificial intelligence, computer aided diagnosis systems have been used extensively within radiology for many years. The most common applications are for detection of cancers on medical images and of pulmonary nodules on chest CT. These systems traditionally relied on manual feature engineering based on domain knowledge, but newer approaches are employing machine learning to discover latent features within imaging data.

The main goal of CAD systems is to identify abnormal signs at an earliest that a human professional fails to find. In mammography, identification of small lumps in dense tissue, finding architectural distortion and prediction of mass type as benign or malignant by its shape, size, etc.

8. Conclusion

In this chapter, we provide for the first time a comprehensive overview of cancer, including its definition, symptoms and types. In addition, we delve into more depth on the type of lymphatic system. We presented a general study of the various categories of lymphatic cancer, their symptoms, stages and how to diagnose them.

In the next chapter, we will present a related study on the set of techniques contributing to the diagnosis of lymphatic cancer.

Chapter 2

State of the art

1. Introduction

The "Deep Learning" is a solution for every subject of classification, regression or even exploration, and this is due to its satisfactory results for many researches, such as image processing and natural language processing.

In this chapter, we introduce the artificial intelligence groups namely machine learning and deep learning. We submit to the concepts of image processing as the second point of the chapter, and as a final point, this chapter presents a state of the art of some related work.

2. Intelligence artificielle

Artificial Intelligence (AI) aims to mimic how the human brain works, or at least its logic when it comes to making decisions. AI involves implementing a number of techniques to enable machines to mimic some form of real intelligence.

AI encompasses different subdomains such as business rules, Machine Learning (ML), Deep Learning (DL), etc. (See Fig 2.1).

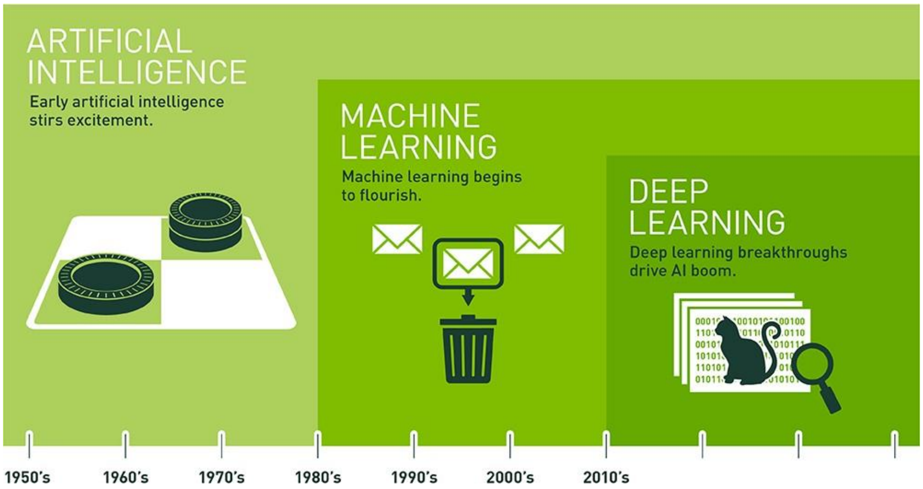

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Fig 2. 1:Relation between IA, LL, & DL [13].

3. Machine Learning

3.1. Definition

"*Machine Learning is the field of AI that enables a machine to learn. That is, to gradually improve performance on a specific task based on data without being explicitly scheduled.*" [14].

Machine learning takes place through neural networks designed to mimic human decision-making abilities. We must apply it to solve any problem that requires thought, be it human or artificial. Machine learning is generally divided into two main classes, which is "supervised learning" and "Unsupervised learning".

3.2. Supervised learning

The most common form of machine learning is supervised learning. Supervised learning is a method of transforming one dataset into another, the program is trained on a predefined set of training examples, which then facilitates its ability to come to a precise conclusion when new data is provided [15] [16]. ML's supervised classification algorithms are: Forest Random, Decision Trees, Logistic Regression, and the best known is SVM Support Vector Machines.

3.3. Unsupervised learning

Unsupervised learning, also known as learning from observations, shares a common property with supervised learning: it transforms one set of data into another. But the dataset it transforms into is not previously known or understood. Unlike supervised learning, for its part, it will be fed only by examples, and will itself create the classes that it deems the most judicious (clustering) or association rules (Apriori algorithms). The K-mean algorithm (Kmeans) makes it easy to understand the concept of unsupervised classification [17]

4. Deep learning

4.1. Definition

“Deep Learning is a class of machine learning techniques belonging to the field of Machine Learning in which multiple layers of iterative computing processing in hierarchical architectures supervised algorithms are exploited for unsupervised learning algorithms for analysis and classification tasks” [18]

Deep learning is essentially about calculate hierarchical characteristics of parameters of artificial neural networks for vector representations of observational or input data. The family of Deep learning methods are increasingly enriched, encompassing those of neural networks, hierarchical probabilistic models, as well as numerous supervised and unsupervised feature learning algorithms.

4.2. Deep learning VS machine learning

There are two main characteristics that distinguish deep learning from machine learning which are:

	feature extraction	**Performance**
ML	Most of the functionality of the application is required by an expert and then manually coded by domain and type of data.	the results prevail as the amount of data increases.
DL	whose algorithms try to know the high-level functionality of the data	When the data is small, the performance of deep learning algorithms gives poor results because it requires a large amount of data to fully understand it.

4.3. Deep Learning operation

The neural network consists of three important layers as shown in the figure:

- ✓ The input layer;
- ✓ The hidden layers;
- ✓ The output layers.

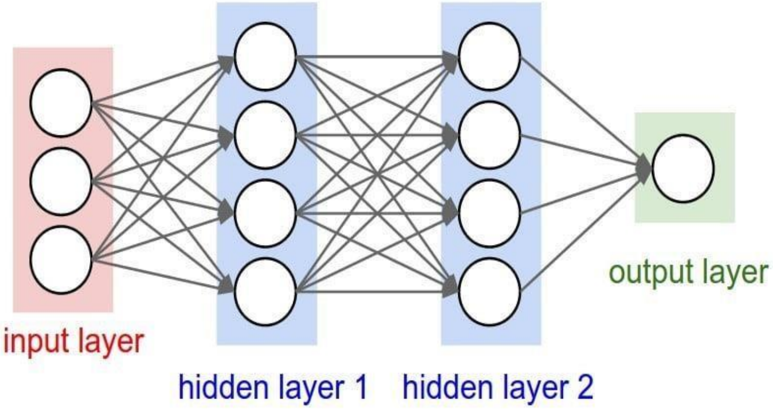


Fig 2. 2:Neural network topology [19]

The meaning of the word "deep" refers to the number of hidden layers in the neural network where traditional neural networks contain only a small number of hidden layers (2 or 3), while deep networks can have up to 150 hidden layers [19].

The neural network is made up of a set of nodes (neurons) connected via directed links (arrow), each arrow represents a link between the output of one neuron and the input of another, Each arrow carries a weight (W), reflecting its importance, each node being a processing unit that performs a static node function on its incoming signal to generate a single node output [20].

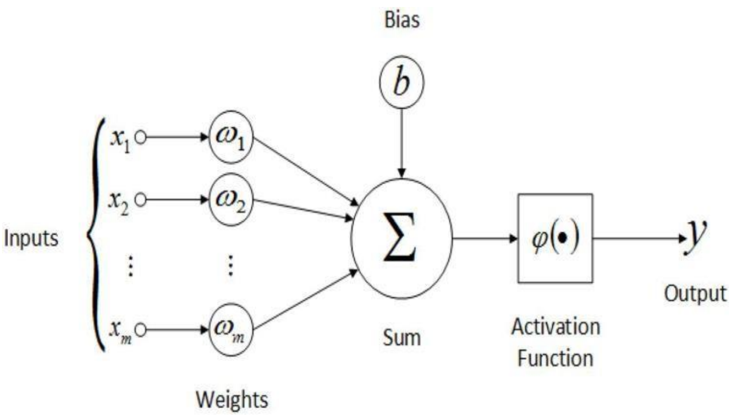


Fig 2. 3:Principle of operation of an artificial neuron [20]

Input values, or in other words, our underlying data, are passed through this "network" of hidden layers until they converge on the output layer. The output layer matches our prediction: it could be one node (we say a binary classification) or a few nodes if it is a multiclass classification problem.

The shape inside the neurons in the core layers represents an activating function which can be a Cube, Elu, Hardsigmoid, Hardtanh, Identity, Leakyrelu, Rationaltanh, Relu, RRelu, Sigmoid, Softmax, Softplus, Softsign, Tanh.

Back propagation

An important part of neural networks, including modern deep architectures, is the backward propagation of errors through a network. Each example of inputs propagates and back propagate to update the weights used by neurons closer to the input.

4.4. Deep Learning model

In the 1980s, most neural networks formed a single layer due to cost of computation and availability of data. Nowadays, we can take on more hidden layers in our neural networks, hence the nickname deep learning [21]. The different types of neural networks available for use have also proliferated, including pre-trained networks and trained from scratch networks.

4.4.1. Models from scratch

A model can be built from the beginning to solve a problem, as this depends first on choosing the appropriate type of model to solve the problem and then starting to design the model and choosing the appropriate number of layers and hyperparameters. For this there are three main families of neural network which are DNN, RNN and CNN.

4.4.1.1. DNN

Deep neural networks (DNNs) are forward feed networks (FFNNs) where data flows from the input layer to the output layer without going backward, and the connections between layers are one way that is forward and never touch any node repeatedly.

4.4.1.2. RNN

A repetitive neural network (RNN) is a class of artificial neural networks characterized by the formation of connections between nodes in the form of a directed graph along a sequence as feature links from one layer to previous layers, in this way information can flow back to the previous parts of the network and thus each model in Classes depend on past events, allowing information to continue.

4.4.1.3. CNN

According to the work of Yin W et al [22], convolutional neural networks "CNN" are specific types of artificial neural networks that use perceptrons and a machine learning unit algorithm for supervised learning, allowing to analyze data. CNNs apply much more to image processing, also to natural language processing "NLP" and other types of cognitive tasks. Like other types of artificial neural networks, a convolutional neural network has an input layer, an output layer, and one or more hidden layers (see Fig 7). Some of these layers are convolutional and use a mathematical and algebraic model to transmit the results to successive layers.

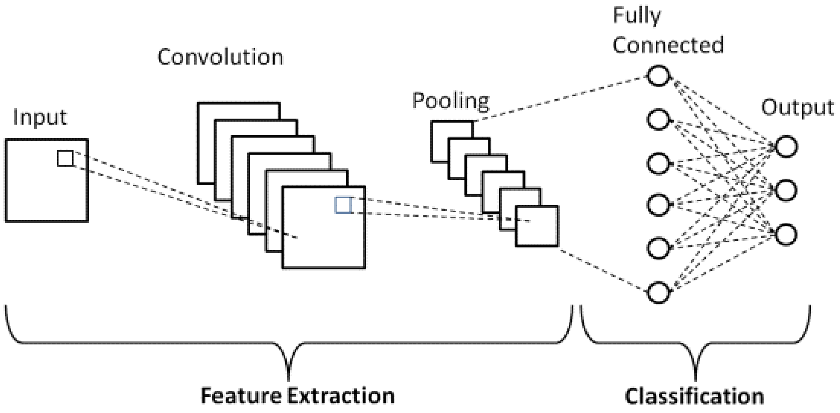


Fig 2. 4:CNN architecture [22]

Typical examples of the deep learning technique are "CNN" networks, where a more sophisticated model accelerates the evolution of artificial intelligence by providing systems that simulate different types of biological activities in the human brain.

Generally, “CNN” networks are made up of four main layers:

- **The convolutional layer "CONV":** which processes the input data or the results of the intermediate or hidden layers.
- **The Pooling layer "POOL":** which allows information to be compressed by reducing the size of the intermediate image.
- **The activation layer:** these are linear rectification functions often called by **"ReLU"** which allows the updating of the parameters of the current layer as well as the previous layers.
- **The fully connected layer "FL":** which is a perceptron-type layer.

4.4.2. Pretrained models

A pre-trained model is a pre-built model to solve similar problems. Instead of building a model from scratch to solve a similar problem, you can use the model trained on another problem as a starting point. This method overcomes the consequences of time, material potential, and the huge effort required, but it may not be 100% accurate. The most pretrained models used are VGG , ResNet and AlexNet.

4.4.2.1. VGG-16

VGG-16 (also called OxfordNet) is a convolutional neural network architecture that 16 layers deep. The model loads a set of weights pre-trained on ImageNet. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. The default input size for VGG-16 model is 224 x 224 pixels with 3 channels for RGB image. It has convolution layers of 3x3 filter with a stride 1 and maxpool layer of 2x2 filter of stride 2.[23]

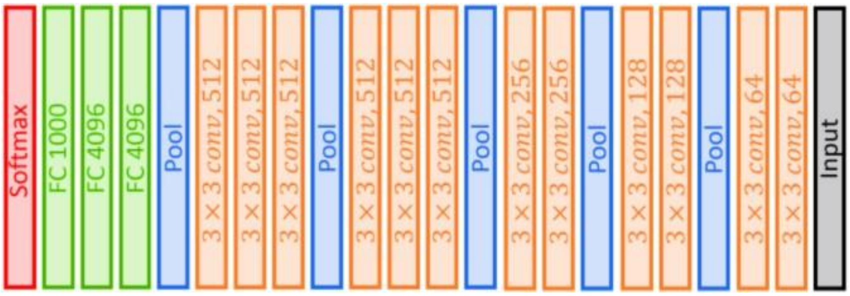


Fig 2. 5:VGG 16 Model [23].

4.4.2.2. VGG-19

Likewise, the VGG-19 performs the same way as the VGG-16, the VGG-19 differs over the VGG-16 only in that it deals with deeper layers for better, more accurate output.

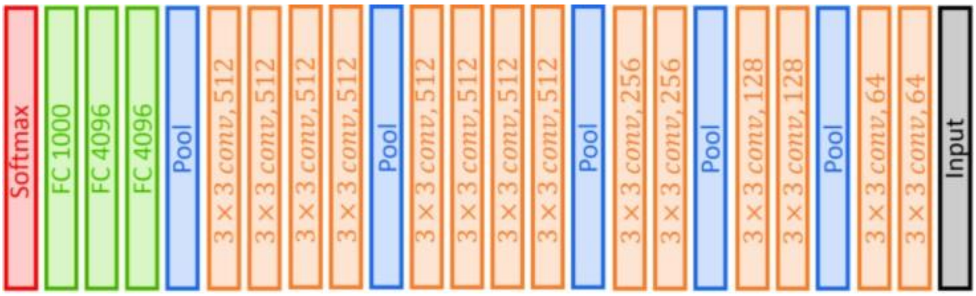


Fig 2. 6:VGG 19 Model [23].

4.4.2.3. ResNet

ResNet for “Residual neural network” was developed by Microsoft research, this network is able to provide and ease of training for networks that are subsequently deeper and hard to train. ResNet consists of 50 layers (for (ResNet-50), almost 3 times deeper than a VGG-19 network.

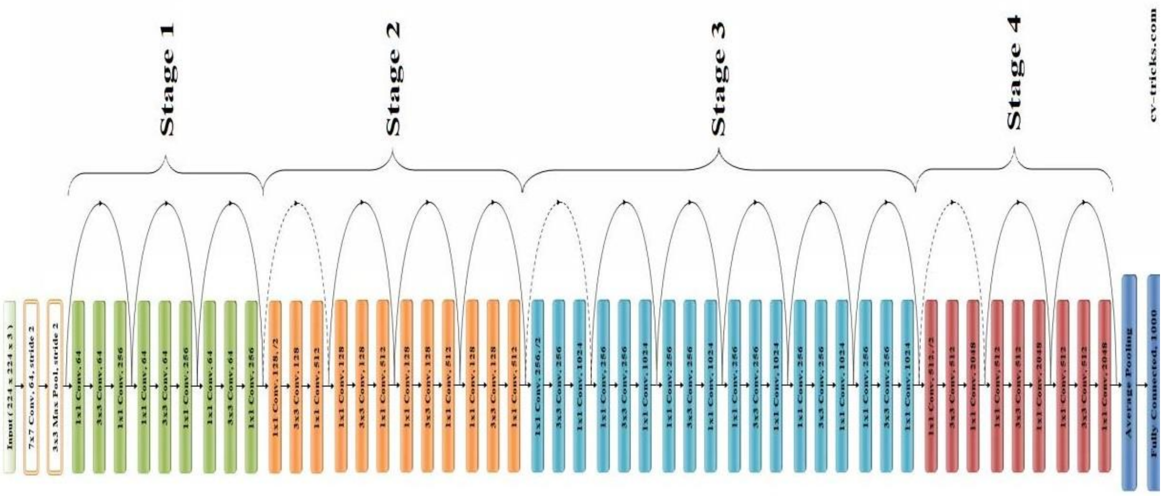


Fig 2. 7:ResNet Model [24]

4.4.2.4. AlexNet

Convolutional Neural Networks (CNNs) had always been the go-to model for object recognition, they’re strong models that are easy to control and even easier to train. The only problem: they’re hard to apply to high resolution images. At the ImageNet scale, there needed to be an innovation that would be optimized for GPUs and cut down on training times while improving performance.

Alexnet is the name of a convolutional neural network (CNN) architecture, designed by Alex Krizhevsky in order to resil the problem of cnn network . Her architecture consists of eight layers: five convolutional layers and three fully-connected layers. But this isn’t what makes AlexNet special; these are some of the features used that are new approaches to convolutional neural networks [25]:

- **ReLU Nonlinearity**: AlexNet uses Rectified Linear Units (ReLU) instead of the tanh function, which was standard at the time. ReLU's advantage is in training time; a CNN using ReLU was able to reach a 25% error on the CIFAR-10 dataset six times faster than a CNN using tanh.

- **Multiple GPUs:** Back in the day, GPUs were still rolling around with 3 gigabytes of memory (nowadays those kinds of memory would be rookie numbers). This was especially bad because the training set had 1.2 million images. AlexNet allows for multi-GPU training by putting half of the model's neurons on one GPU and the other half on another GPU. Not only does this mean that a bigger model can be trained, but it also cuts down on the training time.
- **Overlapping Pooling:** CNNs traditionally **"pool"** outputs of neighboring groups of neurons with no overlapping. However, when the authors introduced overlap, they saw a reduction in error by about 0.5% and found that models with overlapping pooling generally find it harder to overfit.

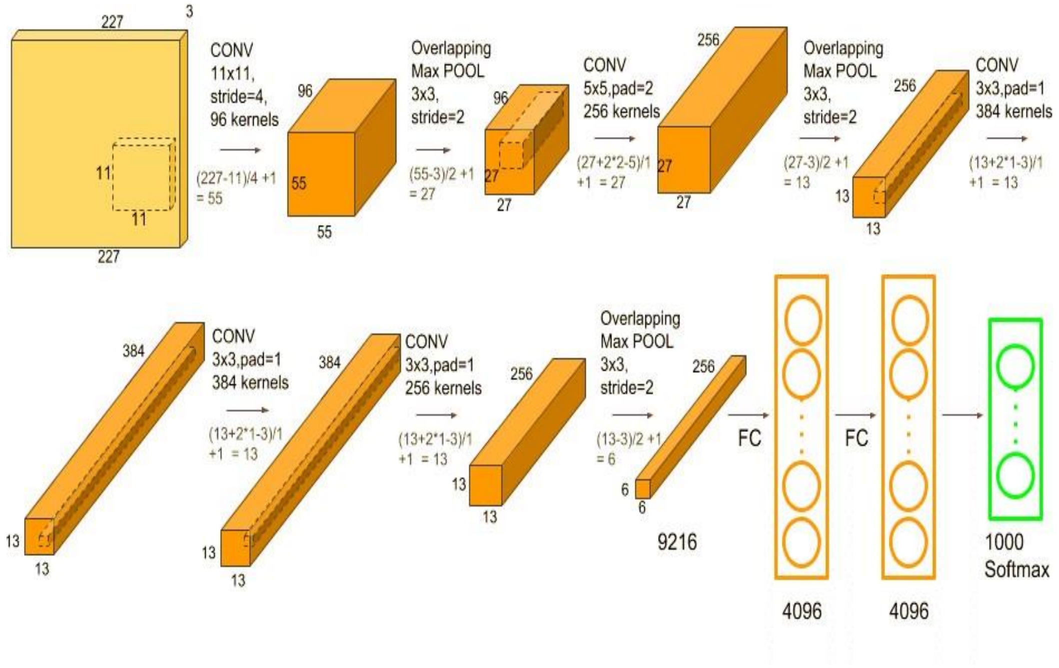


Fig 2. 8:AlexNet architecture [25].

5. State of the art

In order to become familiar with the various approaches to detection and classification of cancer of the lymph system, we divide the work in this section into two parts. First, we discuss the different classic techniques that researchers use to integrate semi-autonomy and autonomy in the classification of images of lymphoma. Next, we present the relevant work based on a deep learning algorithm for the task of lymphoma classification and diagnosis.

5.1. Classic Classification Methods

Various techniques and algorithms for computer vision like lymphoma diagnosis have been used to classify subtypes and semi-automatic detection of lymphoma lesion by many authors, a computer vision approach was used to quantitatively characterize image content. The two-stage approach used in this study does not use the segmentation method that allows for its generality contribution. The most popular techniques used, including:

- Use of Weighted Neighborhood Distance (WND) classifiers to compute similarities with training classes. [26]
- Use of Naive Bayes (NBN)
- Use the radial bias function (RBF) and arrangement to reduce the conventional dimensions [27].

❖ **Critics:**

Using these methods, it would not be possible to diagnose new cases on invisible slices without using standardization in sample preparation.

5.2. Dataset for lymph cancer

There are many datasets for the lymph cancer, some of theme are privates and some other are public and specified for the classification tasks, in this part we mention the top three datasets used in lymphoma classification tasks:

5.2.1. Hematoxylin-Eosin (H & E)

For a decade, pathologists have regarded histological classification as valuable and the most reliable way to determine the type of disease and its stage of spread to distinguish benign or malignant tissue by visually examining a sample of tissue glued to microscopic slides. by using a specific staining protocol which makes it possible to recognize the histological structures thanks to the impregnation of dyes which give them color.

The most widely used dye in pathological practice is hematoxylin-eosin (H & E). The drive to develop CAD systems has led to the scanning of samples to obtain digital tissue images that replace physical tissue slices for processing using computer algorithms.[28]

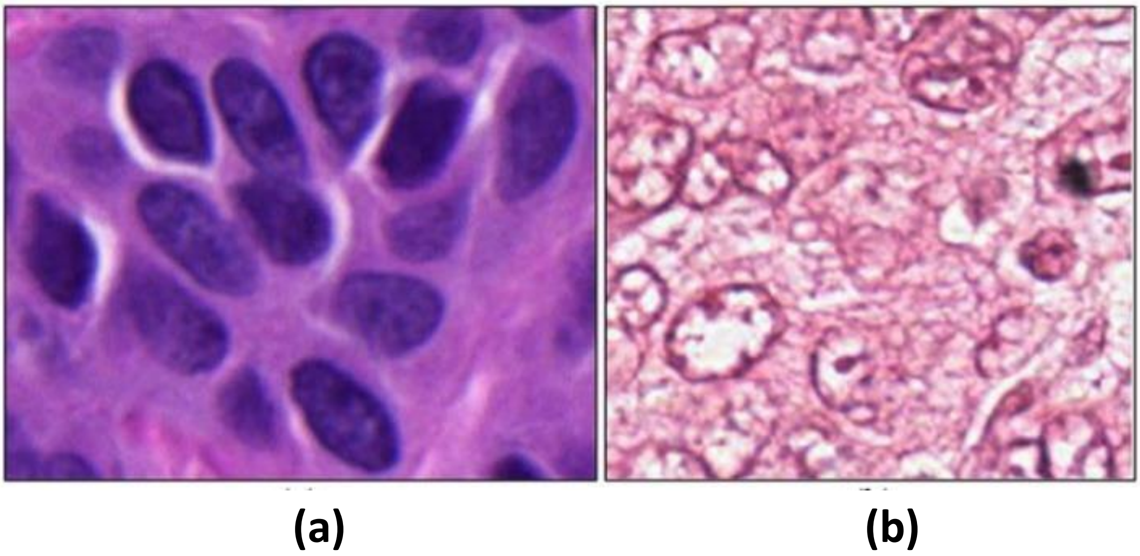


Fig 2. 9:Sample images of H&E dataset (a) image with nuclei that are well stained and have clear boundaries. (b) image with nuclei that have ill-defined nuclear contour due to uneven staining.

5.2.2. NIA

The National Institute of Aging provides a publicly available harmonized dataset for classifying lymphoma subtypes. The dataset contains samples prepared by different pathologists in different locations and is therefore diverse. The images in this dataset contain a wide variety of pigments in terms of color, as well as the variety of data used to train the neural network model. Experienced pathologists working with this type of lymphoma note that they can distinguish these subtypes based on hematoxylin and eosin staining, suggesting that digital pathology (PD) has the potential to work in these areas.

The dataset contains 374 images of 1,388 x 1,040 images saved as a color TIF file. The images and are classified into three classes: 113 for the class of chronic lymphocytic leukemia (CLL), 139 for the follicular lymphoma (FL) class, and 122 for the mantle cell lymphoma (MCL) class. Examples of these images.[29][30]

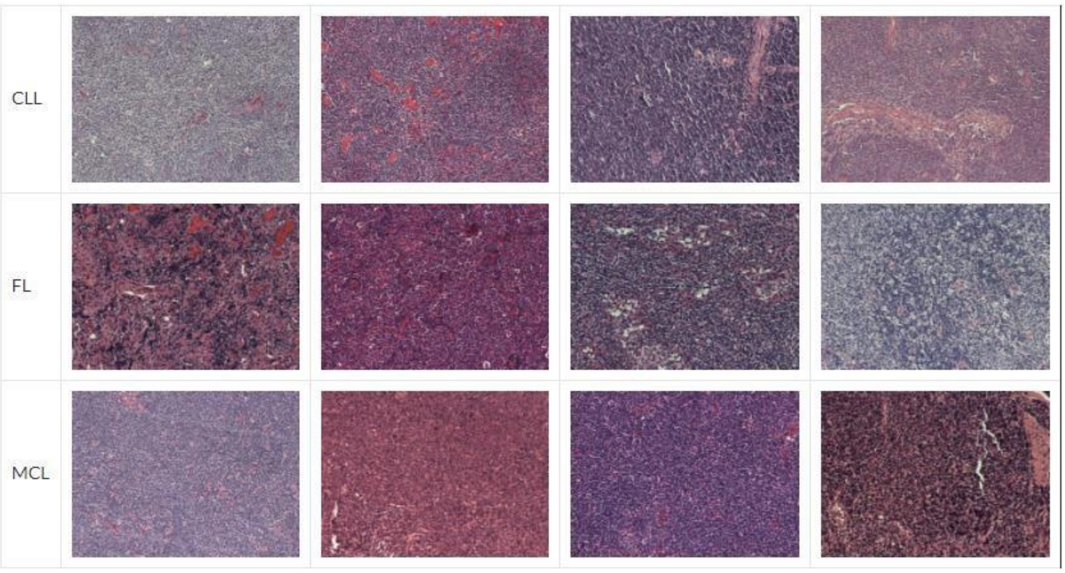


Fig 2. 10:Sample images for NIA dataset.

5.2.3. KIMIA PATH 960

The publicly available kimia path 960 is a histopathology treasure for a large amount of data contained over 400 new histopathology images whole slide images (WSIs) of muscle, epithelial and connective tissues with 960 non-magnified pathology images divided by 20 scans representing “visually” texture / pattern types (based only on visual cues).

48 regions of interest of the same size from each WSI were selected and then down sampled to 308 × 168 patches. Therefore, we obtained a dataset of 960 (= 20 × 48) images. Images are saved as TIF color files [31][32].

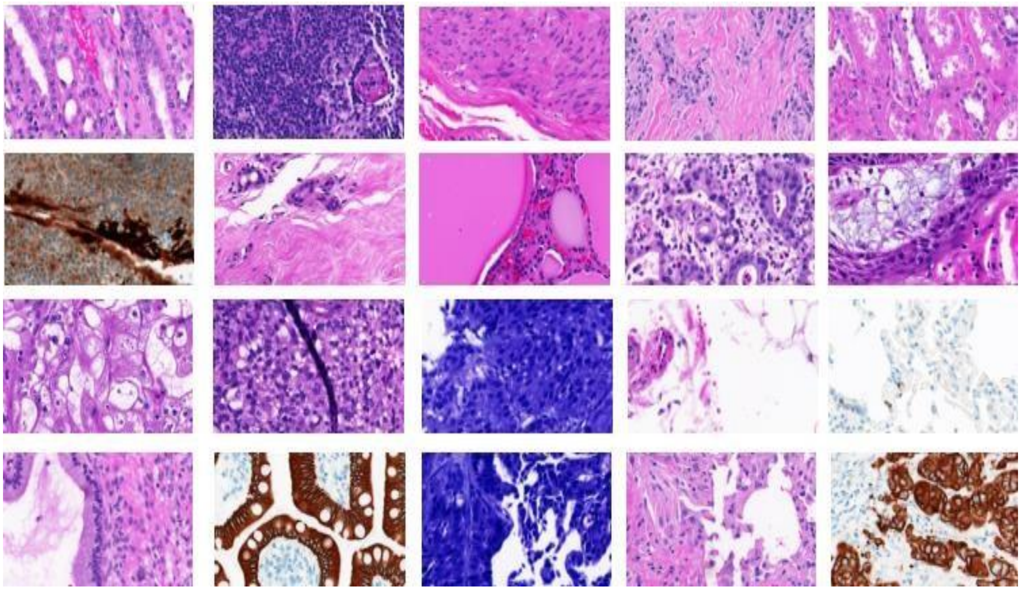


Fig 2. 11:Sample images for 20 classes in the KIMIA Path960 image dataset.

5.3. Deep learning in the medical field (Intelligent methods)

The use of deep learning techniques in the medical sector has proven its ability to develop a new tool that surpasses current technologies due to the accuracy and computational efficiency it brings. Image classification is one of the main areas in which neural networks have made an important contribution to intermediate image analysis. CNN is the most popular algorithm used to classify images, using a set of slices to extract the features of the images. The most effective method for training deep learning models on entire slide images is not to use the entire image as data, but to choose only small corrections from the images. The majority of approaches in the past decade have focused on looking at ways to derive features directly from images for later use for models. Thus, feature extraction and model design were two separate and independent stages that were performed sequentially. It has been shown that automating this method using Convolutional Neural Networks (CNNs) produces an advantage more appropriate to the nature of the problem. This is one of the factors behind the success of deep learning and more specifically, neural network-based learning in the medical field.

There are challenges in the form of choosing an appropriate magnification of the training image, and identifying the correct region to extract features from the training image in order to classify the lymphoma subtypes. The challenge of extracting suitable features from the training images to classify the lymphoma subtype is addressed using deep learning techniques, which is the technology that motivates our work.

Many works have been proposed depending on technology. We mention the following:

- Some authors use the AlexNet model which has applications in various fields such as scene classification and handwritten number classification.
- In another work the authors introduce the concept of image patching. Corrections are generated using the training and validation dataset, by dividing the dataset into 36 x 36 spots in a 32 step, to augment the training and validation dataset.
- Another work increased the size of the data set in order to avoid an over-processing problem. A similar type of corrections is generated on the test data set to create test samples of equal size to training and validation. The results are obtained by finding the category to which the largest number of corrections to the input test image belongs.

In order to conduct a comparative study between the various works related to our approach, we selected 4 recent studies to conduct an analytical study based on the following assessments: the purpose of the study, the data set used, the type of algorithm used, the advantages and the disadvantages of each work. Table 1 resume the differences between these studies.

- ❖ Fist work [32]: this work has classified the images in 12 different methods for the KIMIA Path 960 dataset. The pre-trained Resnet50 deep learning models with custom layers at the head of the deep learning model and data augmentation are most effective to classify the 20 tissue types. This model for the KIMIA Path 960 dataset with accuracy of 99.9% with 10 fold cross validation consistently outperforms the accuracies presented in the related works. To generalize whether this model is suitable for other datasets, we have tested the NIA curated dataset with the best performing model used in the KIMIA Path 960 dataset. The accuracy for this model is 98.13% with 5-fold cross validation is better than accuracies mentioned in the papers.

- Second work [33]: They proposed a method for classifying the subtypes of lymphoma, lymphoma. Based on a dataset sponsored by NIA consisting of three titled subtypes of lymphoma, which are chronic lymphocytic leukemia, follicular lymphoma and mantle cell lymphoma, and the use of the CNN (V3 Network) algorithm. The method proposed produced an accuracy of 97.33% over the NIA harmonized dataset, shows that deep learning has the potential to aid in the automated diagnosis of lymphoma.

- Third work [34]: They proposed automatic classification of three common types of malignant lymphoma: chronic lymphocytic leukemia, follicular lymphoma and lymphoma in cap cells. The aim was to find patterns indicative of malignant lymphomas and allow classification of these malignancies by type. We used a computer vision approach to quantitatively characterize image content. A unique two-stage approach was used in this study. On the external level, the elementary pixels have been transformed by a series of transformations into spectral planes. The primary pixels and spectral levels were then directed to the second stage (the internal plane). Internally, the multipurpose global feature set was computed at each spectral level by the same feature bank. All of the computed properties were combined into a single trait vector. Samples were stained with hematoxylin (H) and eosin (E) stains.

- Forth work [35]: propose a deep learning model that uses transfer learning with fine-tuning to improve the identification of Follicular Lymphoma (FL) on images from new sites that are different from those used during training. Our results show that the proposed approach improves the prediction accuracy with 12% to 52% compared to the initial prediction of the model for images from a new site in the target environment.

Table 2. 1:Deep learning-based approaches to detect and classify cancer of the lymphatic system

Ref	Dataset	Objective	Method	Result	Advantage	Disadvantage
[32]	**KIMIA Path 960 dataset:** 960 images with 20 classes of tissues.	Ensure high classification accuracy by applying latest pre-trained models with custom head and employing cyclical learning rate method for faster learning	Resnet 50 with 10-fold cross validation	98.13%	The pre-trained Resnet50 deep learning models with custom layers at the head of the deep learning model and data augmentation are most effective to classify the tissue types.	use of the KIMIA960 dataset as well as the NIA curated dataset, the diversity of the images is limited by the number of labelled images. Further studies can be done on larger datasets.
	NIA curated dataset: 374 with 3 sub-types of lymphoma (CLL, FL, MCL)		Resnet 50 with 10-fold cross validation	98.13%		
[33]	**NIA dataset**	Demonstrate the potential of deep learning by proposing a method for classifying lymphoma subtypes to aid in the automated diagnosis of lymphoma.	CNN (V3 Network)	97,33%	The fully automated system does not require understanding of lymphoma in depth, its causes, symptoms or any other domain knowledge	Addition of more features will pave way to increase the quality of classifier
[34]	**H&E dataset:** Dataset composed of Hematoxylin and eosin (total of 400 microscopy images)	Find patterns indicative of lymphoma malignancies and allowing classifying these malignancies by type	WND Classifier	98%	The lassification accuracy is up to 98% without segmentation, without using multiple enlargements, or selecting training images containing diagnostic markers of lymphoma.	Failure to take into account a greater diversity of cases, as well as an expansion in the number of different lymphomas
			Naïve Bayes and RBFs	88%		
[35]	**NIA dataset and** **H&E dataset**	propose model to improve the identification of Follicular Lymphoma on images from new sites that are different from those used during training	The AlexNet pre-trained model is trained on the ImageNet dataset consisting 1000 classes of natural images	Best accuracy is 97.4%	Model help to deal with the problem of the needs of large amount of labelled data at the target sites.	Differences resulting from the sample preparation process may affect prediction accuracy
			CIFAR AlexNet			
			Fine-tuned AlexNet pre-trained model			

6. Synthesis

In this chapter, we have explored the field of artificial intelligence and we shed light on deep learning techniques and their role in the categorization of images. moreover, we have studied the approaches of diagnosis and the classification of CANCER OF THE LYMPHATIC SYSTEM namely the classical methods and the intelligent methods. We have chosen the deep learning approach because of the advantages provided by artificial intelligence over the rest of the technologies, as we have studied 4 related works that share the goal of diagnosing and classifying cancer of the lymphatic system, and differ in model and dataset. Finally, we compared this study and extracted the weaknesses. We concluded that reliable image-based detection and classification of cancer of the lymphatic system had certain consequences, including:

- ✓ Lack of large-scale training data: Many AI-based learning techniques rely on comprehensive training data, including medical imaging. However, the datasets available for AI are insufficient.
- ✓ Lack of diversity in deep learning algorithms to detect cancer of the lymphatic system and categorize it as other types of cancer.
- ✓ Transfer learning is used as a solution to solve the problem of insufficient training data, although the Alexnet algorithm is distinguished by the extraction of functionality without the need to split the images, its use in such problem has not gained popularity.

Based on these three points, we chose to move on to using the Alexnet algorithm and test it on the NIA dataset in order to build a classification model and compare the proposed work with the rest of the work. previously mentioned.

7. Conclusion

In this chapter, we present a related study on a set of deep learning algorithms and their role in the medical field and image classification. We analyze the different methods used to detect and classify cancer of the lymphatic system. Most of them used CNN's algorithm, and their goal is to provide better accuracy and create an appropriate system for patients and clinicians.

In order to overcome the problems, present in the current system, in the next chapter, we will provide a full explanation of our contribution as well as our sample algorithms used in our architecture.

Chapter 3

Project

1. Introduction

We are interested in our study mainly on the three main types of non-Hodgkins lymphoma. These types are mantle cell lymphoma (MCL), chronic lymphocytic leukemia (CLL) and follicular lymphoma (FL). The motive of this chapter is to develop a classification model using a set of learning transfer models using VGG 16, VGG 19, Resnet 50 and Resnet 101 in order to build a system capable of automatic discrimination between the mentioned types.

2. The purpose of diagnosing lymphoma

The most reliable method for diagnosing lymphoma is the method based on the analysis of histopathological images because of the morphological characteristics of the tumor that can be analyzed under a microscope.

Diagnosis is inherently difficult due to the influence of factors such as human skill requirements, complexities in slides, etc. The application of image processing and machine learning techniques in various cancer detection methods appears as the main tool. Researchers have used these techniques to identify CLL, FL, and MCL lymphoma subtypes, although current systems are not very effective due to the complex characteristics of these subtypes, and progress is still underway in improving lymphoma grading.

The main objective behind this work is to answer the following question:

What is the result of using the transfer learning approach in assessing the quality of lymphoma subtypes using limited potential?

3. Pre-trained

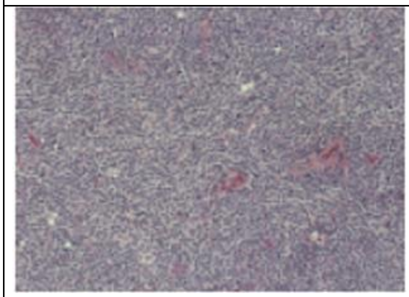
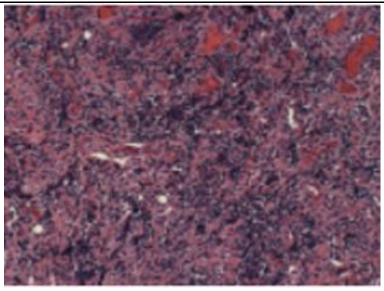
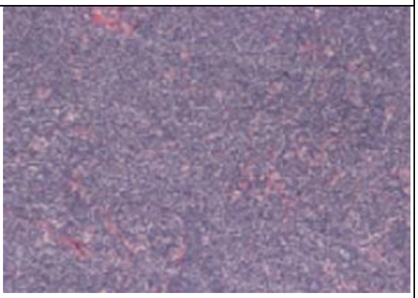
3.1. Lymphoma data

In our study, we chose the publicly available data set to classify lymphoma subtype. The National Institute on Aging (NIA) provides this data set.

As mentioned in the previous chapter, this data was downloaded from this link: https://ome.grc.nia.nih.gov/iicbu2008/lymphoma.tar.gz , with size 1.3GB.

Contains images of lymph nodes from biopsies sectioned and stained with Hematoxylin/Eosin (H+E). The size of these images is rather large with three channels RGB (**1388*1040*3**), it consists of three types of lymphomas:

Table 3. 1 : sample dataset

FL (1388*1040*3)	**CLL (1388*1040*3)**	**MCL (1388*1040*3)**
		
Total :140	**Total:112**	**Total:124**

3.2. Split the data set

The training set the largest corpus of our dataset that we reserve for training our model. After training, inference on these images will be taken with a grain of salt, since the model has already had a chance to look at and memorize the correct output.

We suggest allocating 80% of our dataset to the training set and divided it as follows: 90% train it represents 270 images, and 10% validation represents 30 images uses during training to gets a sense of how well the model and we set aside 20% as a test set represents 74 images to study how well the model performs.

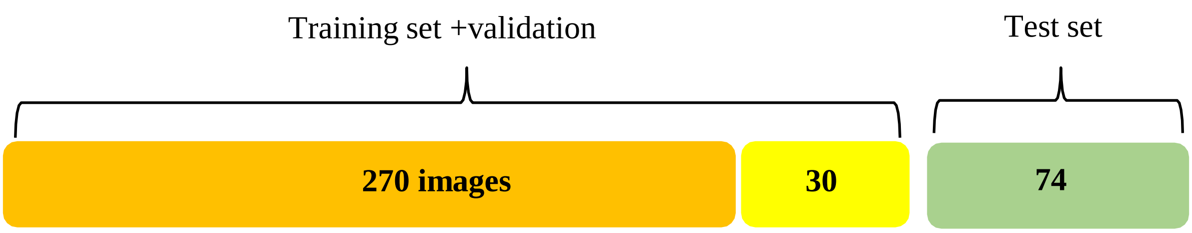


Table 3. 2: Protocol of split Dataset

	Train	Validation	Test
FL	104	13	21
MCL	88	5	28
CLL	78	12	24
Total	**270**	**30**	**74**

3.3. Patching

The result of models based on deep learning is primarily related to providing a large amount of data, but in our case, 374 images are not enough to train a model capable of extracting highly efficient properties.

Another problem the biopsy images of the entire slide amount to 4 330 560 pixels (1388 * 1040*3) Such huge images are time-consuming and expensive to comment on in detail, Also, they can't be put into the model transfer learning which requires the input images to be of size 224X224X3

Therefore, we introduced the nested image slicing strategy to increase the size of the data, where the image is sliced as follows:

Cutting lengthwise: $1388 \div (224 - 26) = 7\,.01$

Crosswise cut: $1040 \div (224\ - 20) = 5.09$

The overlap property maintains the continuity of data and properties within each image and adapts the input to the relevant model

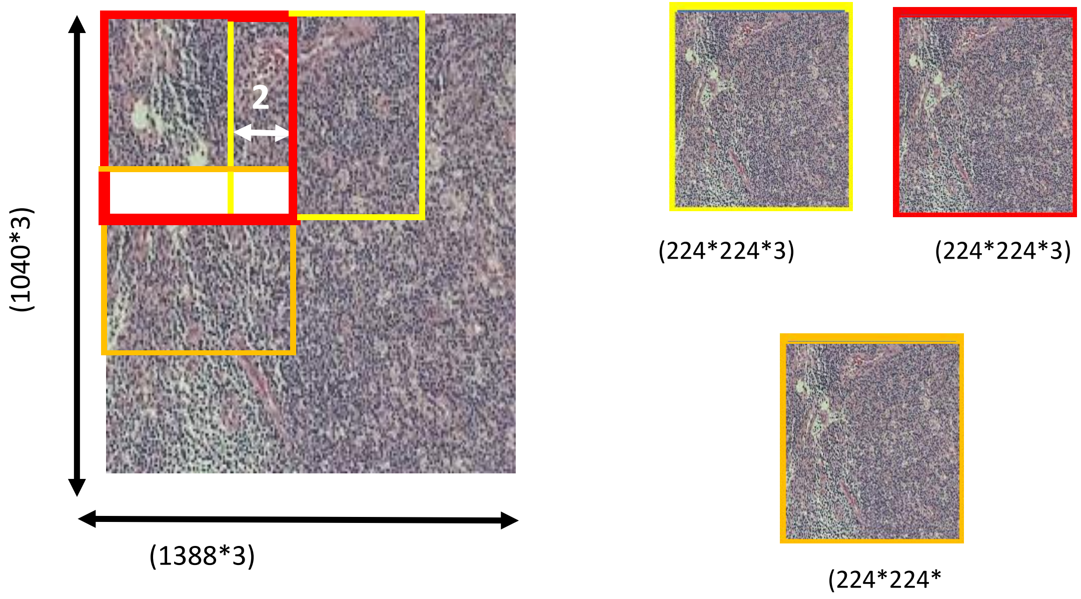


Fig 3. 1:patching with overlap

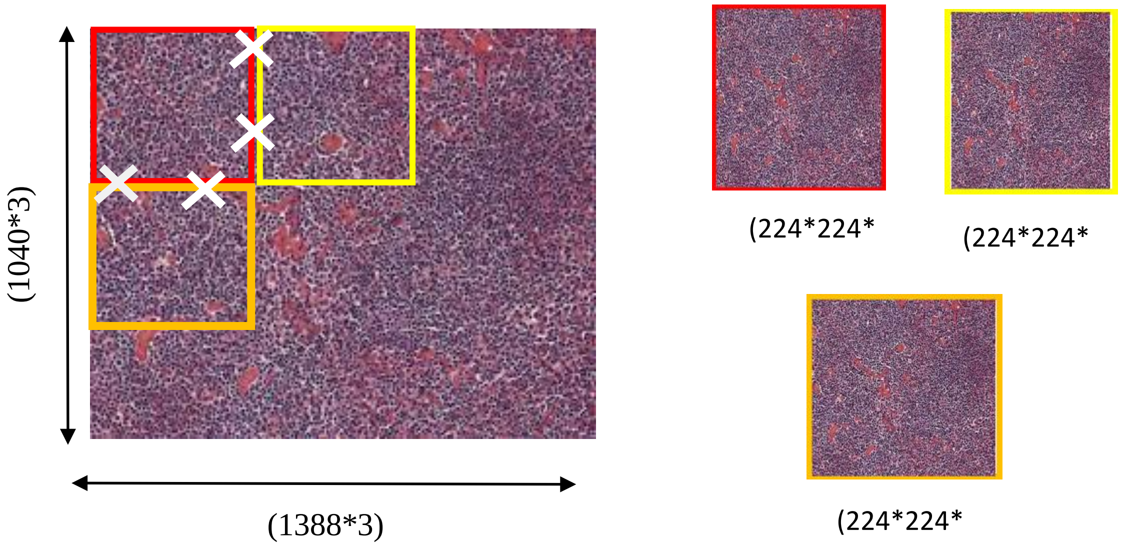


Fig 3. 2:without overlap

4. Conception

Our goal is to establish a system that can accurately classify lymphoma subtypes. So, we present in the following picture a design for the workflow.

We use a tissue image dataset that is organized into different folders based on categories. Where we first preprocess it by using patching technique to increase its size and then divide it into a test, train and val dataset.

Deep learning models are created based on pre-trained models (VGG and Resnet), and we finally evaluate the models using various evaluation metrics.

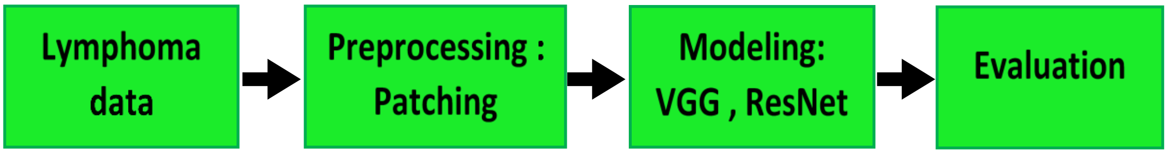


Fig 3. 3:Workflow design

5. Modeling

The convolutional neural network is mainly used to classify images. It is a neural network that allows. As discussed in previous chapter, there are five main layers of CNN , the convolutional layer is used for feature extraction. The Batch normalization is responsible for receiving the input from the previous layer and providing it to the next layer. Each neuron gives a linear output. When the output of one neuron is fed to another neuron, it again produces a linear output. To overcome this situation, there are many nonlinear activation functions available in CNN. These functions are Sigmoid, TanH, ReLU, and Leaky ReLU. The convolutional model produces a large number of features. As the number of features expands, the computational complexity also increases exponentially and the model becomes more sensitive. To overcome this situation, a sampling process is introduced. The sampling or aggregation process reduces overall dimensions and complexity. To specify the number of classes, each fully connected layer is connected to the previous one.

Transformed learning is usually used to solve the data shortage problem. We therefore chose to use transfer learning with the help of both **VGG16, VGG19, Resnet50 and Resnet101** networks.

Where we explain the architectures of each network in the following figures.

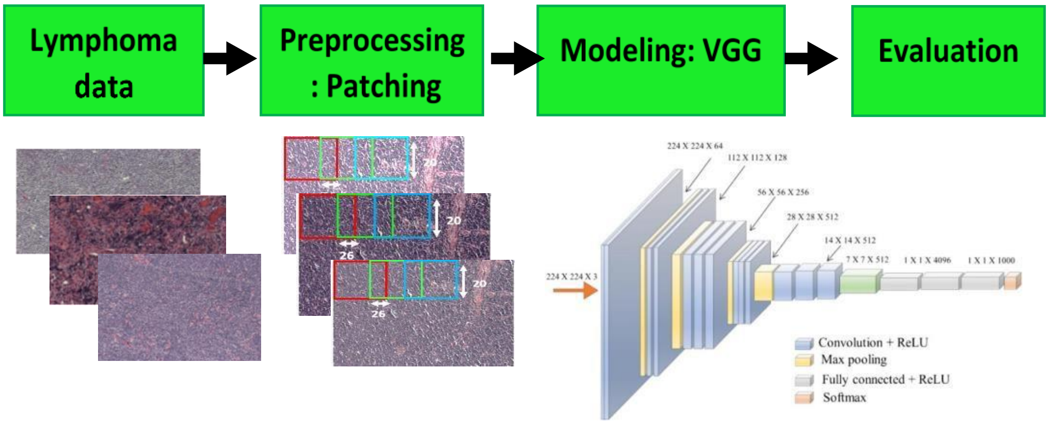


Fig 3. 4: VGG Architecture

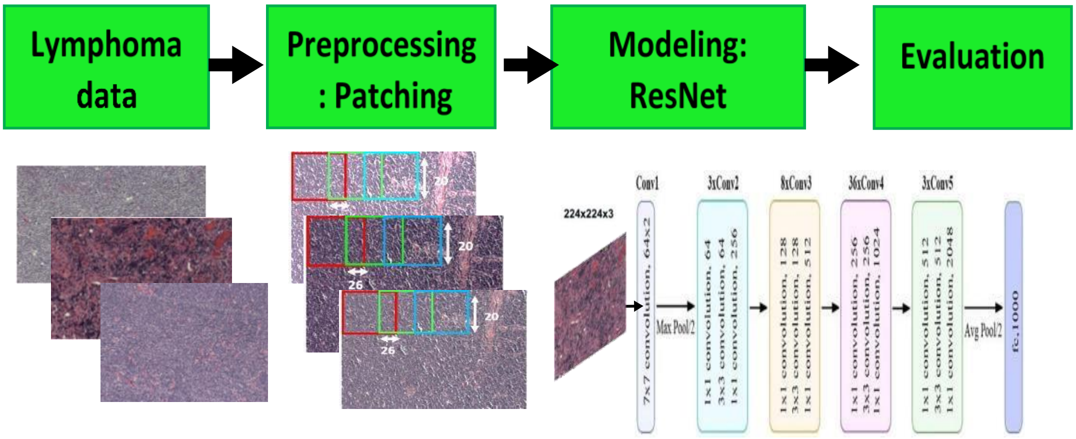


Fig 3. 5: Resnet architecture

6. Implementation

In this section we present the implications for the system and list all the tasks that have been undertaken to classify lymphoma subtypes. We first discuss the hardware requirements for the implementation, and then we detailed the tools used and the environment required for research.

Hardware Requirements

The hardware specifications of the system on which the research project is implemented are as follows.

Processor: Intel Core i5 8265U CPU @ 1.60GHz 1.80GHz

RAM: 8 G

Storage: 1TB HDD Operating System: 64-bit operating system, Windows 10.

6.1. Development Environment

We used the powerful programming language Python to build the proposed models. The reason we chose Python is because there are so many available libraries. Python has multiple libraries like Tensorflow and Keras that are dedicated to neural networks. Google Collaboration is used for the experiment. The data stored in Google Drive is accessed with the help of Python API. Other necessary libraries such as NumPy, Matplotlib.

TensorFlow: Is an open-source software library for high performance numerical computing. Created by the Google team in the to facilitate the creation of machine learning modules. Its flexible architecture allows easy deployment of compute on various platforms (CPU, GPU, TPU), from desktops to server clusters. Even though TensorFlow

Keras: Is considered a powerful, easy-to-use Python library for developing and evaluating deep learning models. It has a minimalist design that allows you to build a network layer by layer; train him and run it.

Matplotlib: is a plotting library for the Python programming language and its numerical math extension NumPy. It provides an object-oriented API for integrating plots into applications using general purpose GUI toolkits

NumPy: is a library for the Python programming language, adding support for large multidimensional arrays and matrices, as well as a large collection of high-level math functions to operate on these arrays.

Seaborn: is a matplotlib-based Python data visualization library. It provides a high-level interface for drawing attractive and informative statistical graphs.

Google colaboratory: or "Colab" for short, allows you to write and run Python in your browser, with. No configuration required; Free access to GPUs; Easy sharing.

Pandas: is a library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Sklearn: is a free Python library for machine learning. It offers in its framework many libraries of algorithms to be implemented, turnkey. These libraries are available in particular to data scientists.

It includes functions for estimating random forests, logistic regressions, classification algorithms, and support vector machines.

Fastai: is a deep learning library which provides practitioners with high-level components that can quickly and easily provide state-of-the-art results in standard deep learning domains, and provides researchers with low-level components that can be mixed and matched to build new approaches.

7. Evaluation

7.1. Case Study 1

Transfer learning with the help VGG 16 is applied on the dataset after the pre-processing. The model was trained for different set of epochs (10 ,50 and 100) and also was fitted for different batch size (16 ,32 ,64).

Table 3. 3:VGG 16 evaluation

		Batch size		
		16	32	64
Epochs	10	Test accuracy: 97.18%	Test accuracy: 96.94%	Test accuracy: 96.47%
	50	Test accuracy: 97.82%	Test accuracy: 97.19%	Test accuracy: 96.59%
	100	Test accuracy: 98.05%	Test accuracy: 97.80%	Test accuracy: 97.14%

The VGG 16 model is implemented as is but the size of the training data set is increased with the help of data augmentation using the partitioning method that we explained earlier, after performing the basic reinforcement methods. The model achieved an accuracy of 98.05 % in the original test after building the classifier for 100 epochs.

7.2. Case Study 2

Transfer learning with the help VGG 19 is applied on the dataset after the pre-processing. The model was trained for different set of epochs (10 ,50 and 100) and also was fitted for different batch size (16 ,32 ,64).

Table 3. 4:VGG 19 evaluation

		Batch size		
		16	32	64
Epochs	10	Test accuracy: 96.31%	Test accuracy: 97.54%	Test accuracy: 97.02%
	50	Test accuracy: 98.06%	Test accuracy: 97.98%	Test accuracy: 98.10%
	100	Test accuracy: 98.23%	Test accuracy: 98.60%	Test accuracy: 98.54%

The VGG 19 model is implemented as is but the size of the training data set is increased with the help of data augmentation using the partitioning method that we explained earlier, after performing the basic reinforcement methods. The model achieved an accuracy of 98.60 % in the original test after building the classifier for 100 epochs.

7.3. Case Study 3

Transfer learning with the help Resnet 50 is applied on the dataset after the pre-processing. The model was trained for different set of epochs (10 ,50 and 100) and also was fitted for different batch size (16 ,32 ,64).

Table 3. 5:Resnet 50 evaluation

		Batch size		
		16	32	64
Epochs	10	Test accuracy: 98.06%	Test accuracy: 97.99%	Test accuracy: 97.81%
	50	Test accuracy: 98.17%	Test accuracy: 98.46%	Test accuracy: 98.11%
	100	Test accuracy: 98.25%	Test accuracy: 98.19%	Test accuracy: 98.76%

The Resnet 50 model is implemented as is but the size of the training data set is increased with the help of data augmentation using the partitioning method that we explained earlier, after performing the basic reinforcement methods. The model achieved an accuracy of 98.76 % in the original test after building the classifier for 100 epochs.

7.4. Case Study 4

Transfer learning with the help Resnet 101 is applied on the dataset after the pre-processing. The model was trained for different set of epochs (10 ,50 and 100) and also was fitted for different batch size (16 ,32 ,64).

Table 3. 6:Resnet 101 evaluation

		Batch size		
		16	32	64
Epochs	10	Test accuracy: 97.93%	Test accuracy: 97.54%	Test accuracy: 97.66%
	50	Test accuracy: 98.01%	Test accuracy: 98.00%	Test accuracy: 98.14%
	100	Test accuracy: 98.56%	Test accuracy: 97.95%	Test accuracy: 98.49%

The Resnet 101 model is implemented as is but the size of the training data set is increased with the help of data augmentation using the partitioning method that we explained earlier, after performing the basic reinforcement methods. The model achieved an accuracy of 98.56 % in the original test after building the classifier for 100 epochs

8. Discussion

The objective of this study was to accurately classify sub types of lymphoma with deep learning. A significant drawback when deep learning is implemented is the shortage of training data that allows the model to learn features. Since the basic requirements for deep learning were not fulfilled, it was difficult to classify sub types of lymphoma. In this study 374 images were used to obtained the accuracy of 98,76% with Resnet 50. Image Normalization technique in pre-processing has shown a great effect on neural networks.

While performing the experiment, different data augmentation methods have been tried on sample images to check the significance. It has been observed that patching is the label preserving transformations. The following graph in shows the comparative study of models in terms of accuracy.

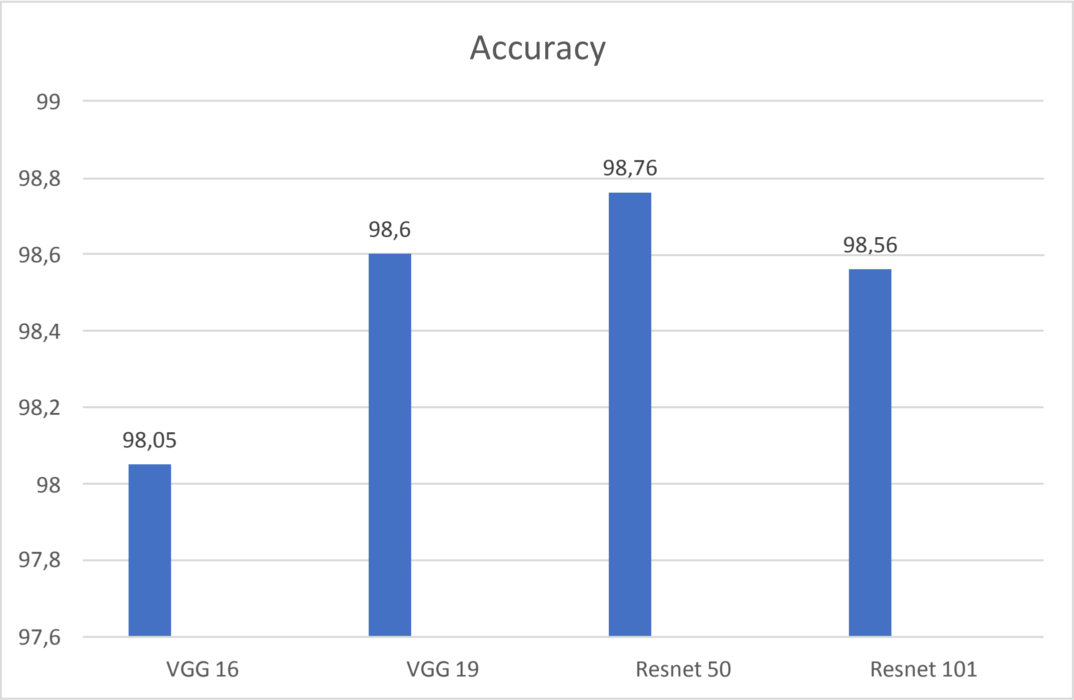


Fig 3. 6: Comparative study of models in terms of accuracy

Besides evaluating the models based on accuracy, a training time is also compared and understood how each one of these models perform against one another. As the models were built from different batch size and epochs, it took different time to train as compared to each other.

The results shows that Resnet models were better in terms of execution time and accuracy. The limitation of the study has been to obtain more set of histopathological images for lymphoma that could help train the model more. The field of research may be extended to many other types of disease.

The following table shows our result compared with the other works that detailed in the state of the art, as we see, our model achieves the best accuracy, it beats the previous best work by 0.63%.

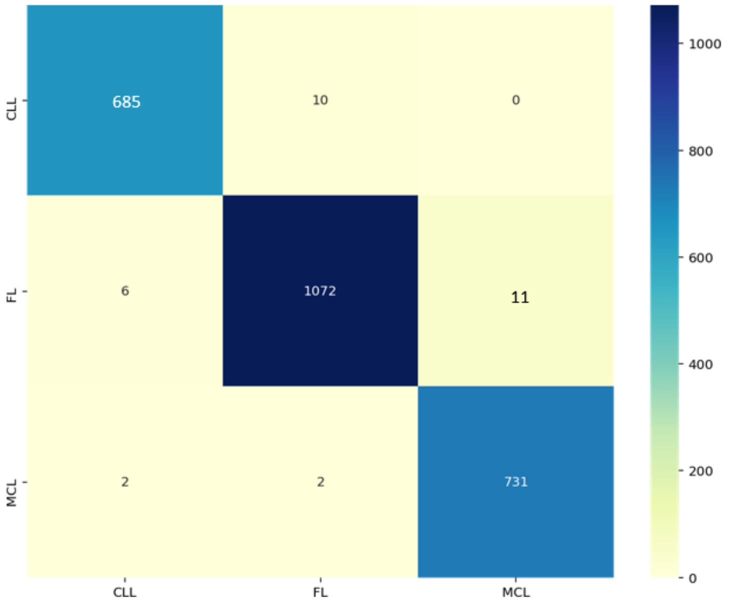


Fig 3. 7: Confusion matrix

Table 3. 7:comparative works in terms of accuracy

Ref	Dataset	Method	Result
Our work	**NIA**	Resnet 50	98.76%
[32]	**NIA**	Resnet 50 with 10-fold cross validation	98.13%
[35]	**NIA**	The AlexNet pre-trained model	97.4%
[33]	**NIA**	CNN (V3 Network)	97,33%

9. Conclusion

To classify lymphoma subtypes, we used deep learning, specifically and transfer learning with VGG and Resnet. The data set was tested with each model. Pre-processing was performed on the images to improve model performance. It is very important to apply data augmentation. Due to the limited computing power of the system, it was not possible to perform further image operations. Finally, it was noted that the Resnet 50 had an accuracy of 92.51%. When performing all the experiments, it has been observed that performance can be increased significantly by increasing the training data and epoch number. This research looked at only three types of lymphoma. In the future the target group can be increased.

General conclusion

In this work , we are interested in making a contribution based on artificial intelligence to help diagnose diseases. As a case study we tested the topic of lymphoma subtype classification.

Diagnosing this disease and determining its type is a difficult task that requires high resources, techniques, effort and expertise. So, in order to overcome these points and with the high popularity that machine learning algorithms have gained in classifying images, especially medical ones, we have followed this approach and we used Convolutional Neural Network and Transfer Learning using both VGG and Resnet networks. We chose the NIA open-source dataset.

The results obtained are interesting, and they can be considered as competitive results with the methods proposed in the literature.

In future work, we are primarily interested in building an integrated system that can receive images as they are and automatically apply the partitioning process and then diagnose and this is done behind the system so that only the final result appears to the reader, and in this way the system is complete and ready for use before the doctors.

References

Chapter 1:

[1] : https://www.e-cancer.fr/Dictionnaire/C/cancer (14/04/2021)

[2] :https://www.cancer.ca/fr-ca/cancer-information/cancer-type/lung/lungcancer/?region=on (15/04/2021)

[3] :https://www.researchgate.net/figure/Stages-of-tumor-development-and-mechanism-of metastasis_fig2_278644130 (15/04/2021)

[4] :https://www.cancer.ca/fr-ca/cancer-information/cancer-type/pancreatic/pancreatic cancer/?region=on (20/04/2021)

[5] : https://www.passeportsante.net/fr/Maux/Problemes/Fiche.aspx?doc=cancer_sein_pm (21/04/2021)

[6] : Cueni, L. N., & Detmar, M. (2008). The lymphatic system in health and disease. Lymphatic research and biology, 6(3-4), 109-122.

[7]: Loudon, A., Barnett, T., Williams, A. D., Visentin, D., Immink, M. A., & Piller, N. (2017). Guidelines for teaching yoga to women with breast cancer-related lymphoedema: an evidence-based approach. International journal of yoga therapy, 27(1), 95-112.

[8] : https://www.healthline.com/health/lymphoma#diagnosis (15/05/2021)

[9] :https://greek.doctor/third-year/pathology-2/theoretical-exam-topics/28-indolent-b-cell-lymphomas-fl-cll-mcl-mzl/] (16/05/2021)

[10] : Alevizos, L., Gomatos, I. P., Smparounis, S., Konstadoulakis, M. M., & Zografos, G. (2012). Review of the molecular profile and modern prognostic markers for gastric lymphoma: how do they affect clinical practice?. Canadian Journal of Surgery, 55(2), 117.

[11] : Cossman, J., Uppenkamp, M., Sundeen, J., Coupland, R., & Raffeld, M. (1988). Molecular genetics and the diagnosis of lymphoma. Archives of pathology & laboratory medicine, 112(2), 117-127.

[12] : Chan, H. P., Hadjiiski, L. M., & Samala, R. K. (2020). Computer- aided diagnosis in the era of deep learning. Medical physics, 47(5), e218-e227.

Chapter 2:

[13] : Ji, H., Alfarraj, O., & Tolba, A. (2020). Artificial intelligence-empowered edge of vehicles: architecture, enabling technologies, and applications. IEEE Access, 8, 61020-61034.

[14] : A. L. Samuel, "Some Studies in Machine Learning Using the Game of

Checkers," in IBM Journal of Research and Development, vol. 3, no. 3, pp. 210- 229, July 1959.]

[15]: Annina S., Mahima S, S. Venkatesan3, D.R. Ramesh Babu, An Overview of

Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE).

[16] : Andrew W. Trask, 2019. grokking Deep Learning.

[17] : Yann LeCun, Yoshua Bengio & Geoffrey Hinton., may 2015 REVIEW Deep learning

[18]: L. Deng, et D. Yu, Deep learning. Boston, p.217.2014

[19] : https://fr.mathworks.com/discovery/deep-learning.html (17/05/2021)

[20] : https://www.saagie.com/fr/blog/qu-est-ce-que-le-deep-learning/ (le 17/05/2021)

[21] : Chen, Q., Wang, W., Wu, F., De, S., Wang, R., Zhang, B., & Huang, X. (2019). A survey on an emerging area: Deep learning for smart city data. IEEE Transactions on Emerging Topics in Computational Intelligence, 3(5), 392-410

[22] : W. Yin, K. Kann,M. Yu, et H. Schütze. Comparative Study of CNN and RNN for Natural Language Processing, 2017.

[23] : Mojumder, U., Sarker, T. T., Monika, G. M., & Ratul, N. A. (2016). Vehicle model identification using neural network approaches (Doctoral dissertation, BRAC University).

[24] : Detailed Guide to Understand and Implement ResNets – CV-Tricks.com (27/052021)

[25] : Pedraza, A., Gallego, J., Lopez, S., Gonzalez, L., Laurinavicius, A., & Bueno, G. (2017, July). Glomerulus classification with convolutional neural networks. In Annual conference on medical image understanding and analysis (pp. 839-849). Springer, Cham.

[26] : R. Ferjaoui, M. A. Cherni, N. E. H. Kraiem and T. Kraiem, "Lymphoma Lesions Detection from Whole Body Diffusion-Weighted Magnetic Resonance Images," 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), Thessaloniki, 2018, pp. 364-369. doi: 10.1109/CoDIT.2018.8394840

[27] : N. V. Orlov et al., "Automatic Classification of Lymphoma Images With Transform-Based Global Features," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pp. 1003-1013, July 2010. doi: 10.1109/TITB.2010.2050695

[28] : Tosta, T. A. A., de Faria, P. R., Neves, L. A., & do Nascimento, M. Z. (2019). Computational normalization of H&E-stained histological images: Progress, challenges and future potential. Artificial intelligence in medicine, 95, 118-132.

[29] : https://www.britannica.com/science/lymphatic-system (21/05/201)

[30] : http://www.andrewjanowczyk.com/use-case-7-lymphoma-sub- type classification/#:~:text=Background,Mantle%20Cell%20Lymphoma%20(MCL) (21/05/2021)

[31] : https://kimialab.uwaterloo.ca/kimia/index.php/pathology-images-kimia-path960/ (21/05/201)

[32] : Ganguly, A., Das, R., & Setua, S. K. (2020, July). Histopathological Image and Lymphoma Image Classification using customized Deep Learning models and different optimization algorithms. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE

[33] : Tambe, R., Mahajan, S., Shah, U., Agrawal, M., & Garware, B. (2019, January). Towards Designing an Automated Classification of Lymphoma subtypes using Deep Neural Networks. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (pp. 143-149).]

[34] : Ferjaoui, R., Cherni, M. A., Kraiem, N. E. H., & Kraiem, T. (2018, April). Lymphoma Lesions Detection from Whole Body Diffusion-Weighted Magnetic Resonance Images. In 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT) (pp. 364-369). IEEE.

[35] : Somaratne, U. V., Wong, K. W., Parry, J., Sohel, F., Wang, X., & Laga, H. (2019, December). Improving follicular lymphoma identification using the class of interest for transfer learning. In 2019 Digital Image Computing: Techniques and Applications (DICTA) (pp. 1-7). IEEE.