

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Mohamed Khider University - Biskra
Faculty of Exact Sciences and Sciences of Nature and Life
Computer Science Department

Order Number: :.....



THESIS

In Candidacy for the Degree of
DOCTOR 3rd CYCLE IN COMPUTER SCIENCE
Option : Artificial Intelligence

TITLE

Big Data analytics using Artificial Intelligence techniques in medical PHM

Presented by **Abir Belaala**

Defended on:

In front of the jury composed of:

Mr. Okba Kazar	Professor at University of Biskra	President
Mr. Labib Sadek Terrissa	Professor at University of Biskra	Supervisor
Mr. Nouredine Zerhouni	Professor at ENSMM, Besançon, France	Co-supervisor
Mr. Soheyb Ayad	Associate Professor at University of Biskra	Examiner
Mr. Salim Chikhi	Professor at University of Constantine 3	Examiner
Mr. Adel Kermi	Associate Professor at ESI School, Algiers	Examiner
Mrs. Zeina Al Masry	Associate Professor at ENSMM, Besançon, France	Guest

Academic year : **2020 – 2021**

Abstract

Today, With the development of information technology, the concept of smart healthcare became a trending research area. Smart healthcare uses a new generation of information technologies such as big data, cloud computing, and artificial intelligence (AI). These new techniques helps to transform the traditional medical system to be more intelligent, efficient, convenient, and personalized.

Computer-aided diagnosis (CAD) has become one of the major research subjects in medical computing and clinical diagnosis. However, how to efficiently and effectively make accurate diagnosis remains a challenging problem in data-driven models.

In this thesis, we are interested in improving the performance of computer-aided diagnostic systems in the medical field by increasing the quality of medical data and the analytical techniques. To this end, several contributions have been proposed. First, we proposed an extension of Prognostic and Health Management (PHM) approaches in order to exploit its potential by adapting advanced industrial diagnostic models to medical diagnostics. Secondly, we focused on improving computer-assisted diagnosis, particularly in the dermatology field, using AI techniques as well as those of Big data. The proposed methods and the results obtained were validated by an extensive comparative analysis using benchmarks and private medical data.

Keywords: *Computer Aided Diagnosis (CAD), Medical PHM, Big data, Dermatology, Machine learning, Deep learning.*

Résumé

De nos jours et grâce à l'évolution rapide des technologies de l'information et de la communication, le concept de la santé intelligente devient de plus en plus un domaine de recherche très attractif. Ce concept se base principalement sur les techniques de l'intelligence artificielle et les technologies du Cloud Computing et du Big data et cela dans le but de transformer la médecine traditionnelle en médecine digitale à la fois intelligente et personnalisée.

L'un des principaux sujets de recherche en Informatique Médicale et en diagnostic clinique est le diagnostic assisté par ordinateur (CAD). Cependant, effectuer un diagnostic exact de manière efficace demeure une problématique importante notamment dans les modèles guidés par les données.

Dans cette thèse, nous nous sommes intéressés à l'amélioration des performances des systèmes de diagnostic assisté par ordinateur dans le domaine médical, en augmentant la qualité des données médicales et en améliorant les techniques analytiques.

A cette issue, plusieurs contributions ont été proposées. Dans un premier temps, nous avons proposé une extension des approches du Prognostic and Health Management (PHM) afin d'exploiter ses potentialités en adaptant les modèles de diagnostic industriels avancés au diagnostic médical. Dans un second temps, nous nous sommes focalisés sur l'amélioration du diagnostic assisté par ordinateur en particulier dans le domaine de la dermatologie en utilisant les techniques de l'IA ainsi que ceux du Big data. Les méthodes proposées et les résultats obtenus ont été validés par une analyse comparative approfondie en utilisant des benchmarks et des données médicales privées.

Mots clés: *Diagnostic assisté par ordinateur (CAD), PHM médical, Dermatology, Big data, Apprentissage automatique, Deep learning.*

LIST OF PUBLICATIONS

Journal Papers

1. Belaala, A., Terrissa, L. S., Zerhouni, N., & Devalland, C. Computer-Aided Diagnosis for Spitzoid Lesions Classification Using Artificial Intelligence Techniques. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 16(1), 16-37. (2021).
2. A Belaala, Y Bourezane, LS Terrissa, Z Al Masry, N Zerhouni. Skin cancer and deep learning for dermoscopic images classification: A pilot study. *Journal of Clinical Oncology*, 38 (15_suppl), e22018-e22018.(2020).
3. A Belaala, LS Terrissa, Z Al Masry, Y Bourezane,N Zerhouni.Towards improved skin lesions classification using Automatic Hyperparameters Selection and Transfer Learning. *Computer Vision and Image Understanding* (submitted)

Conference Papers

1. A Belaala, Al Masry, Z., Terrissa, L. S., & Zerhouni, N.Retargeting PHM tools: from industrial to medical field. In *PHM Society European Conference* . Italy. (Vol. 5, No. 1, pp. 7-7). (2020, July).
2. A Belaala, Terrissa, L. S., Zerhouni, N., & Devalland, C. Spitzoid Lesions Diagnosis Based on SMOTE-GA and Stacking Methods. In *International Conference on Advanced Intelligent Systems for Sustainable Development* (pp. 348-356). Morocco, Marrakech.(2019, July).
3. A Belaala, Labib Sadek Terrissa, Noureddine Zerhouni, “Predictive Big Data analysis in Healthcare” , in *Workshop ”Applications Medicales de l’Informatique Nouvelles Approches AMINA ”.*, Monastir, Tunisia.(2018).

Chapter book

1. A Belaala, Terrissa, L. S., Zerhouni, N., & Devalland, C. “Big Data Analytics in Healthcare: review. In Adaptive Health Management Information Systems. Jones & Bartlett Learning, (pp. 348-356), (2019).
2. A Belaala, Terrissa, L. S., Zerhouni, N., & Devalland, C. Spitzoid Lesions Diagnosis Based on SMOTE-GA and Stacking Methods. Advanced Intelligent Systems for Sustainable Development (AI2SD’2019): Volume 2-Advanced Intelligent Systems for Sustainable Development Applied to Agriculture and Health, Springer. 1103, 348.(2020).

Acknowledgement

First and foremost, praises and thanks to **Allah**.

I would like to express my deep and sincere gratitude to my research supervisor **Pr.Terrissa Sadek Labib** for giving me the opportunity to do research and providing invaluable guidance throughout this research. His dynamism, vision, sincerity and motivation have deeply inspired me.

I would also like to thank deeply **Pr.Noureddine Zerhouni**. It was a great privilege and honor to work under his guidance. I am extremely grateful for his listening and relevant advices.

I also thank **Dr.Zaina Almasry**, **Dr.Christine Devaland**, and **Dr.Yazid Bourezane** for their collaboration and support to complete this thesis successfully.

I also thank all the members of the jury **Pr.Kazar Okba**, **Dr.Soheyb Ayad**, **Pr.Chikhi Salim**, and **Dr.Kermi Adel** for the time they spend to review this work, I am grateful for the attention they paid to my work.

Finally I thank all those who helped me in some way for the realization of this work.

Abir Belaala

I dedicate my dissertation work to my parents.

Contents

Abstract	i
Résumé	ii
Publications of the author	iii
List of figures	xi
List of tables	xii
List of algorithms	xiii
List of abbreviations	xiv
I General introduction	1
1 Context	1
2 Problem statements	3
3 Contributions	4
4 Dissertation plan	5
II Preliminaries and Basic Concepts	7
1 Big data	7
1.1 Features of Big Data	8
1.1.1 Volume	8
1.1.2 Velocity	8
1.1.3 Variety	9
1.2 Big data process in healthcare	10
1.2.1 Big data generation	11
1.2.2 Big data storage	13
1.2.3 Big data analysis:	17
2 Machine learning and Deep learning	19
2.1 Machine learning categories	19
2.1.1 Supervised learning	19
2.1.2 Unsupervised learning	20

2.1.3	Reinforcement learning	21
2.2	Deep learning	22
2.3	Transfer learning	23
2.4	Machine learning process	25
2.4.1	Data preprocessing	25
2.4.2	Feature selection	27
2.4.3	Choosing a model	28
2.4.4	Model evaluation	30
3	Medical PHM	33
3.1	Engineering PHM VS medical PHM	33
3.2	M-PHM analytics	34
3.2.1	Computer Aided Detection (Descriptive analysis) . . .	35
3.2.2	Computer Aided Diagnosis (Diagnostic analysis) . . .	36
3.2.3	Computer Aided Prognostic (Predictive analysis) . . .	36
3.2.4	Computer Aided Decision making (prescriptive analysis)	37
4	Conclusion	37
III Retargeting PHM tools: from industrial to medical field		39
1	Motivation	39
2	Adapted PHM tool	41
2.1	Data description	41
2.2	Data preprocessing	41
2.2.1	Missing data	42
2.2.2	Scaling data	43
2.2.3	Feature selection	43
2.3	Diagnosis	43
3	Experimentation	45
4	Conclusion	48
IV Computer Aided Diagnosis for Spitzoid lesions classification using Artificial Intelligence techniques		49
1	Medical overview	49
2	Motivation	51
3	Proposed method	53
3.1	Data description	53

3.2	Pre-Processing Phase	56
3.3	The Feature Selection Phase	57
3.4	The Classification Phase	59
4	Performance metrics and experimentation	62
4.1	Performance Metrics	62
4.2	Data Sampling Results	63
4.3	Feature Selection Results	66
5	Discussion	70
6	Conclusion	72
V Toward efficient Automatic Hyperparameters selection using Big Data tools to improve Skin Lesions classification		73
1	Motivation	73
2	Related work	75
2.1	Architecture selection	75
2.2	Data preparation	76
2.3	Model's hyperparameter optimization	76
3	Proposed method	77
3.1	Data description	77
3.2	Data preparation	77
3.3	Data augmentation	81
3.4	Data analysis	82
3.5	Model's Hyper parameters optimization	85
3.5.1	Tools	85
3.5.2	Implementation	86
4	Experimental Results	89
4.1	Results with /without metadata	90
4.2	Results with Automatic hyperparameters Selection (CNN-AHPS)	92
5	Discussion	95
6	Conclusion	96
VI General conclusion		97
1	Summary	97
2	Perspectives	99

List of Figures

II.1	Main features of Big data	9
II.2	Ordinary data VS Big data	10
II.3	General process of Big data in healthcare	12
II.4	Sources of big data in healthcare	13
II.5	Big data storage tools	14
II.6	Supervised learning.	20
II.7	Unsupervised learning.	21
II.8	Reinforcement learning.	22
II.9	Example of CNN architecture.	23
II.10	Fine tuning strategies for pre-trained models.	24
II.11	Machine learning techniques.	30
II.12	Evaluation metrics.	31
II.13	General process of Medical PHM.	33
II.14	Four types of M-PHM analytics.	35
III.1	A general scheme of our CAD of heart disease.	41
III.2	Features selected by RF algorithm on UCI dataset and their score.	44
III.3	ROC curve of our proposed method.	47
IV.1	Example dermoscopic images of Spitz Nevus (1A), and Atypical Spitz Tumors (1B) Source: Rubegni et al. (2016)	50
IV.2	General schema of our proposed process	54
IV.3	The impact of number of generations on accuracy of classifier	60
IV.4	Change of Accuracy in terms of nearest neighbor's k value	61
IV.5	ROC curve of classifiers without oversampling.	65
IV.6	Impact of several different SMOTE's k values on the accuracy	65
IV.7	Distribution of our data with/ without SMOTE technique	66
IV.8	ROC curve of classifiers with SMOTE technique	66

IV.9 ROC curve of classifiers with GA based feature selection	68
IV.10 Most selected features with existing classifiers as fitness evaluation	68
IV.11 Number of selected features by GA and accuracy of classifiers	70
V.1 Overall scheme of our proposed method	79
V.2 Dermoscopic images' example in ISIC 2019	81
V.3 Distribution of dermoscopic images per class in ISIC2019 dataset	82
V.4 Self-training with Noisy Student illustration.	84
V.5 MapReduce illustration.	87
V.6 ROC curve of our CAD system with metadata and AHPS.	93
V.7 Performance comparison of three applied experiments.	94

List of Tables

II.1	Big data platforms with storage type.	16
II.2	Comparison of different feature selection techniques.	29
II.3	Industrial system vs. human body contrast in PHM view.	34
III.1	UCI heart disease dataset description.	42
III.2	Experimental performance metrics.	46
III.3	Performance comparison of our proposed method along with previous work on UCI heart disease Cleveland dataset.	47
III.4	Processing time comparison of our proposed method and existed work.	48
IV.1	Spitz nevus (SN) VS Atypical Spitz tumors (ASTs) vS Spitz melanoma(SM) Source: Adapted from World Health Organization (2018).	51
IV.2	Spitz nevus dataset details.	55
IV.3	Parameter settings of our genetic algorithm based feature selection.	58
IV.4	Experimental performance on our dataset without /with existing over-sampling methods.	64
IV.5	Experimental performance on our data without / with genetic algorithm based feature selection.	67
IV.6	Feature selection results obtained by different classifiers as fitness evaluation.	69
V.1	Comparison of previous work in skin lesions classification.	78
V.2	Description of public and private dataset.	80
V.3	Results obtained with and without metadata.	91
V.4	Hyperparameters selected manually VS automatically.	92
V.5	Results obtained by applying CNN-AHPS.	93
V.6	Results of our CAD system on Private dataset.	94
V.7	Comparative study with previous work.	95

List of Algorithms

1	CNN-AHPS algorithm	88
2	Map algorithm	89
3	Reduce algorithm	89

List of abbreviations

AK	Actinic keratosis
Adadelta	adaptive delta (Adaptive Learning Rate Method)
AdaMax	adaptation of the Adam optimiser
AdaSyn	Adaptive synthetic sampling
AI	Artificial Intelligence
AUC	Area Under Curve
AST	Atypical Spitz Tumor
BCC	Basal cell carcinoma
BK	Benign keratosis
CAD	Computer-Aided Diagnosis
CNN	Convolutional neural networks
CNN-AHPS	CNN- Automatic Hyperparameters Selection
CPOE	Computerized Provider Orders Entry
CT	Computed Tomography
CVD	CardioVascular Diseases
DF	Dermatofibroma
DL	Deep Learning
DM	Data mining
DT	Decision Tree
EHRs	Electronic Health Records
ELM	Extreme Learning Machine
FIFO	First In First Out
GA	Genetic Algorithm

GPU	Graphics Processing Unit
HDFS	Hadoop Distributed File System
HAM10000	Human Against Machine 10000 samples
ICT	Information Communication Technology
ID-RELM	Improved Dragging Regularized Extreme Learning Machine
IoT	Internet of Things
ISDIS	International Society for Digital Imaging of the Skin
ISIC	International Skin Imaging Collaboration
IT	Information Technology
K-NN	K-Nearest Neighbours
LR	Linear Regression
MN	Melanocytic Nevus
MEL	Melanoma
ML	Machine Learning
MLP	Multiple Layer Perceptron
M-PHM	Medical PHM
MPI	Message Passing Interface
MRI	Magnetic Resonance Imaging
mRMR	Minimum Redundancy Maximum Relevance
NB	Naïve Bayes
NoSQL	not only Structured Query Language
OpenMP	Open Multi-Processing
PC	Personal Computer
PCA	Principal Component Analysis
PHM	Prognostics and Health Management
RAM	Random Access Memory

ReLU	Rectified Linear Units
RF	Random Forest
RFE	Recursive Feature Elimination
RMSProp	Root Mean Square Propagation
ROC	Receiver Operating Characteristics
RUL	Remaining Useful Life
SFS	Sequential Forward Selection
SGD	Saccharomyces Genome Database
SMOTE	Synthetic Minority Oversampling TEchnique
SCC	Squamous Cell Carcinoma
SVM	Support Vector Machines
TCP	Transmission Control Protocol
TL	Transfer Learning
TPU	Tensor Processing Unit
UCI	University of California Irvine
UNK	Unknown Class
VASC	Vascular Lesion
WHO	World Health Organization
WSI	Whole Slide Imaging

Chapter I

General introduction

1 Context

Modern smart technologies such as cloud computing, Big Data, Internet of Things (IoT), and 5G technologies gave birth to the fourth revolution of industry. Industry 4.0 can autonomously exchange information, make actions, and control without human intervention [1]. In this context, Health 4.0 is a term that has emerged recently and has been growing as a vital strategic concept for the health domain [2]. Today, the healthcare sector has a big transformation from paper-based records to electronic health records (EHRs). This digitalization is coupled with a wide range of digital technologies that support Health 4.0 to deliver more effective and efficient healthcare services [3].

Prognostics and health management (PHM) discipline was initially developed in the engineering field. Many works piloted in PHM research focus on automatic detection, diagnostic, and prognostic for assessing components health state to support decision making [4, 5]. This success of PHM process in the engineering (industry 4.0) motivates researchers to think about implementing this process in the medical field (health 4.0). Medical PHM (M-PHM) uses diagnostic analytics to diagnose the disease by identifying the type, stage, and causes. Predictive analytics used to predict patients' survival, also predict whether a patient is at a high risk of having a disease based on the risk factor, or predict the disease's recurrence. Prescriptive analytics relate to finding the best course of actions by providing decision support for specific scenarios or

situation. The application of this process helps to improve the quality of care delivery, move towards personalized medicine, sharing of real-time decisions in diagnosis, and prediction of treatment outcomes at earlier stages [6].

Recently, diagnosis step in M-PHM has been an active area in computer science field. Various medical domains attract researches, more precisely oncology and chronic diseases. The main requirements to apply computer aided diagnosis (CAD) is the availability of digital dataset (data) and appropriate data analysis techniques (artificial intelligence).

For the data side in CAD development, medical data are generated massively in various types (numbers, images, text, videos). This large amount of data is emerging from various medical sources such as patient information, biomarkers (e.g., genomic, proteomic, metabolomic), diagnosis results (e.g., radiology, blood test), as well as pharmacies (e.g., prescriptions, medications), administrative (cost and claims data, population and public health data) and behaviour data (e.g., those from mobile apps, social media, sensors, wearable devices, and fitness monitors) [7]. With the fast growth, increased complexity, heterogeneity and size of these accumulated data, the big challenge now is how to collect, store, analyze and manage these Big Data in healthcare systems.

On the other side, artificial intelligence tools including machine learning and deep learning have been used to apply diagnosis analysis. They involve a set of tools and techniques such as classification, clustering, regression and association. Each technique serves a distinct purpose depending on the modelling objective. Often, choosing the right technique depends on the problem at hand and how the data is represented and stored. Deep learning (DL) is the newest iteration of machine-learning methodologies. DL is now performing at state-of-the-art levels in previously difficult tasks including image analysis, language processing, information retrieval, and forecasting. Deep learning is well suited for medical data as it can identify patterns in sparse, noisy data and requires little input-feature engineering [8]. Current successes of DL have shown performance that outperforms physicians and experts to diagnose patient's diseases. However, the performance still needs improvement to make an accurate diagnosis. This thesis deals with some problems encountered with the improvement of computer-aided diagnosis systems, more specifically those related to data quality. These problems are

addressed in the following section.

2 Problem statements

In this thesis, we tackled the problem of developing an accurate computer aided diagnosis to assist physicians during the diagnosis process. First of all, we want to reduce the gap between industry and medical field, by exchanging the applied techniques and benefit from the advancement in industry PHM. However, the complexity of biological system which is not predictable and the sensitivity of working on human lives makes us doubt the validity of this adaptation. Therefore, two research question addressed in this part are:

- *RQ1: what is the difference between industry PHM and medical PHM?*
- *RQ2: Models applied for machine's health diagnosis could be applicable for human's health diagnosis?*

In the second part of this thesis, we direct our research toward dermatology domain. Skin cancer is one of the most widespread types of cancer, and melanoma is the most severe form and causes most skin cancer deaths[9]. The mortality rate of this disease is expected to rise in the next decade, especially for cases diagnosed in later stages. Many computer-aided diagnosis (CAD) methods have been developed based on several research approaches, such as detection, segmentation, and classification using machine learning and deep learning. However this developed CADs still have many challenges that need improvements. Especially when we work with dermoscopic images.

The automatic classification of different skin lesions from dermoscopic images is challenging due to the high similarity in visual features among various lesion types in terms of size, shape, texture, and color. Other problems include artefacts in dermoscopic images, lack of data, and training of deep architectures requiring millions of parameters, which usually lead to overfitting and weak generalization. Thus, many research questions are addressed in this part:

- *RQ3: What is the impact of data quality on the performance of dermatology computer-aided diagnosis?*

- *RQ4: How can we deal with lack of dermatology data, mainly with rare lesions?*
- *RQ5: What about combining clinical features and dermoscopic images for improved performance?*
- *RQ6: What is the impact of hyperparameters selection on the CAD performance?*

3 Contributions

Dealing with the aforementioned problems, three contributions are proposed:

- **Retargeting PHM tools from industrial to medical field:** The first contribution is to apply an adaptation of a PHM model from fault diagnosis of an aircraft engine to diagnosis human heart disease. To this end, public UCI heart disease dataset is used[10]. The proposed scheme consists of (a) the pre-processing step to improve data quality (missing data imputation and scaling dataset); (b) the feature selection step to improve classification performance (based on embedded method); and (c) the diagnosis phase to identify the absence or the presence of heart disease (using Dragging Regularized ELM (ID-RLM)). The performance of the proposed scheme is compared with previous work.
- **Computer Aided Diagnosis to classify spitzoid lesions based on clinical features:** The second contribution focuses on classifying a challenging type of skin lesions called spitz nevus. This classification will be done using private data set contain clinical, histological, and immunohistochemical features to differentiate Atypical Spitz Tumors from regular Spitz nevus. The primary goals and contributions for this work include: (a) an effort to specify the exact type of a Spitz lesion, which is extremely difficult and challenging. Also, in the best of our knowledge, no one has used AI to classify spitzoid lesions before. (b) An attempt to extend past research results on the steps needed for the development of an automatic diagnostic system for Spitzoid lesion classification. (c) A move towards integrating clinical, histological, and immunohistochemical features to make an accurate diagnosis in distinguishing between SN vS AST. In addition, finding out the impact of these features on the classification.

A proposed three-phase approach is being implemented. In Phase I, collected data are preprocessed with an effective Synthetic Minority Oversampling TEchnique (SMOTE) which being implemented to treat the imbalanced data problem. Then, a feature selection mechanism using genetic algorithm (GA) is applied in Phase II. Finally, in Phase III, a ten-fold cross-validation method is used to compare the performance of seven machine-learning algorithms for classification. Experiments results will shed light on the impact of data quality on performance and the best features distinguishing between classical Spitz nevus and atypical Spitz tumors.

- **Computer Aided Diagnosis for skin lesions classification based on dermoscopic imaging:** The third contribution aims to develop a computer-aided diagnosis (CAD) that can accurately classify eight skin lesions using dermoscopic images from a public ISIC 2019 challenge dataset [11] and a private dataset. Three main tasks are proposed and implemented. Task 1 is data quality improvement by solving the imbalanced class problem, missing values, and dermoscopic multi-resolution. Task 2 is CAD development via a pretrained Noisy Student (EfficientNet-L2) architecture as a feature extractor using transfer learning. We incorporate additional metadata using a dense neural network concatenated with the CNN output. Then, the classifier follows with eight units representing skin lesion classes. Task 3 concerns the development of automatic hyperparameter selection (CNN-AHPS) using big data tools (MapReduce). Experiments results will shed light on the impact of hyperparameters selection, also the impact of combining metadata with dermoscopic images.

4 Dissertation plan

The rest of this thesis is organised as follows:

Chapter II presents a state of the art that briefly introduces three main domains related to this thesis by clarifying the basic concepts: Big data, Machine learning, and M-PHM.

Chapter III introduces the first contribution, which consists of applying an adaptation of a PHM model from fault diagnosis of an aircraft engine to diagnosis human heart disease in order to reduce the gap between them. We give first during

the presentation of our approach a quick overview of this adaptation. Then we detail each method steps: data description, pre-processing, feature selection, and classification. After that, experimental results to demonstrate the effectiveness of our system is provided. Finally, a conclusion and future works are presented.

Chapter IV presents our second contribution, which aims to test several artificial intelligence techniques to build a computer-aided diagnosis system to classify Spitzoid lesions. We present first an overview of spitzoid lesions, then a motivation for our contribution. The next Section is offering a detailed description of the proposed method used in three phases: preprocessing phase, feature selection phase, classification phase. next section highlights the key indicators, including performance measures, as well as the experimental findings. Finally, results obtained and discussion is presented.

Chapter V presents our main contribution, which aims to develop a computer-aided diagnosis (CAD) to classify accurately different skin lesions using dermoscopic images and metadata. In this chapter, We present first an overview of skin lesion classification and its challenges. Then, highlights previous work and presents our contribution. The next section presents the proposed method involves data preprocessing, data augmentation, data classification, and automatic hyperparameters selection (CNN-AHPS) technique for the training step. Then we show experimental results and a comparative analysis with state of the art. The last section discusses our observations, findings, and some limitations.

Finally, in **Chapter VI** we conclude the thesis with a summary of contributions and open perspectives.

Chapter II

Preliminaries and Basic Concepts

This chapter introduces briefly three main domains related to our work by clarifying some basic concepts. In Section 1, we introduce big data by defining its 5Vs characteristics. We emphasised on Big Health Data process from the collection step until the decision-making. In addition, we define tools and technologies which have been used to store and process big health data as the Hadoop ecosystem. In Section 2, we present a brief overview of Machine Learning and Deep Learning. Also in the same section we present the machine learning process from data processing until evaluation of model by focusing on supervised learning and more precisely classification task. In Section 3, we present the inspired Medical PHM (M-PHM) by comparing industrial PHM and Medical PHM, and finishing by defining the four health analysis types.

1 Big data

The term Big data founded for massive data sets having a large, complex and varied structures. Big data is generated from images, audios, emails, online transactions, clickstreams, posts, logs, search queries, social networking interactions, health records, science data, sensors and mobile phones and their applications [12]. These big datasets grow massively and become difficult to store, capture, form, manage, analyze, visualize, and share by traditional Information Technology (IT) and hardware/software tools within a sustainable time [13].

1.1 Features of Big Data

Big Data is primarily characterized by three Vs: volume, variety and velocity as presented in Figure II.1. Recent statistics declared that data is growing at a rate of 59% every year [6]. The growth of the data can be described regarding the following five Vs:

1.1.1 Volume

Big data volume refers to the size of data being created from all the sources including text, audio, video, social networking, research studies, medical data, space images, crime reports, weather forecasting and natural disasters [12]. Nowadays, a large amount of data is generated every day. In 2016 the whole amount of data is estimated to be 6.2 exabytes, and now in 2020, we exceeded 80000 exabytes of data [14].

Due to the fast generation of big data in massive sets, organisations and companies that want to join big data into their business strategies are starting to replace traditional methods and tools with business intelligence and analytics systems and software [13]. These advanced techniques allow them to effectively collect, store, process and visualise all of that data in real-time.

1.1.2 Velocity

Velocity defines the speed of generating, capturing, and sharing the dataset. As the flow of data nowadays is massive and continuous, the speed at which data can be accessed directly impacts the decision-making process. Most of the traditional approach face problems associated with data, which keeps adding up but can not be processed quickly. They generally take batch processing or manual processing that takes several hours or days for analysis [15]. The main objective is to collect, process and visualize data closer to real-time to extract information and insights that will lead to better business results.

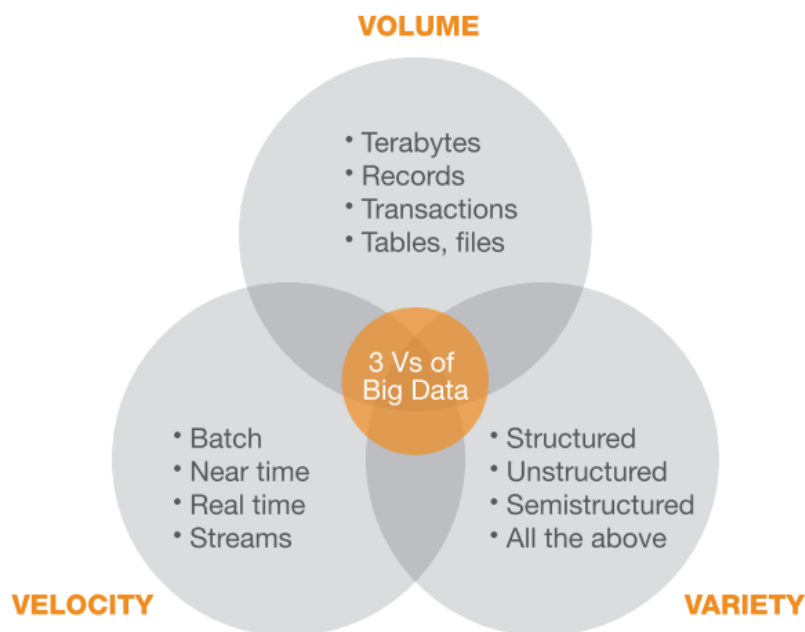


Figure II.1: Main features of Big data

1.1.3 Variety

The data is generated from different sources and forms such as structured, unstructured, and semi-structured.

- Structured data generally is well organized and it can be simply analyzed by humans and machines because it has a defined length and format.
- Semi-structured data is a mix between unstructured and structured data, therefore some components can be easily analyzed and organized, while other parts need a machine to organize it.
- Unstructured data is unorganized data that can be defined as chaotic data, and most of real data in nature is unstructured such as: videos, mobile data, texts, pictures [16].

In addition to these three main characteristics of big data, there are two additional features: Value and Veracity [17]. The veracity refers to the truthfulness of sources that influence accuracy, such as missing data, inconsistencies, ambiguities, duplication, spam, deception, fraud, and latency. Finally, the value represents cost-benefit to the

decision-making through the ability to take meaningful action based on insights derived from data [18]. Figure II.2 illustrates the difference between traditional data and big data according to five features.





	Traditional data	Big data
 Volume	<ul style="list-style-type: none"> • Kilobytes(10^3) • Megabytes(10^6) • Gigabytes (10^9) 	<ul style="list-style-type: none"> • Terabytes (10^{12}) • Petabytes (10^{15}) • Exabytes (10^{18}) • Zettabytes (10^{21})
 Variety	<ul style="list-style-type: none"> • Structured data 	<ul style="list-style-type: none"> • Structured data • Unstructured data • Semi Structured data • Various types of data
 Velocity	<ul style="list-style-type: none"> • Structured data 	<ul style="list-style-type: none"> • Structured data • Unstructured data • Semi Structured data • Various types of data
 value	<ul style="list-style-type: none"> • Analysis & reporting 	<ul style="list-style-type: none"> • Complex and advanced analysis • Predictive & insights analysis • Business intelligence

Figure II.2: Ordinary data VS Big data

1.2 Big data process in healthcare

Healthcare has a big transformation from a paper-based system to Electronic Health Records (EHR). This digitalization due to a massive amount of heterogeneous data, Which include patient medical information; biomarkers (genomic, proteomic, metabolomic); diagnosis results (radiology, blood test); pharmacy data(prescriptions, medications); administrative data (cost and claims data, population and public health data); Also behaviour data that comes from social media, sensors, wearable devices, and fitness monitors. [18].

Such big health data characterized by its complexity, heterogeneity, fast growth, and size, so the big challenge in healthcare systems is how to collect, store, analyze

and manage this data to improve the healthcare quality by understanding new diseases and therapies, predict outcomes at earlier stages, make real-time decisions, as well as personalized medicine. In this section, we will introduce the process of big data in healthcare from the collection of the raw data until the decision making. which can be generally divided into four phases: data generation, data acquisition, data storage, and data analysis (see Figure II.3).

1.2.1 Big data generation

Data generation is the first step of big data. Specifically, it is large-scale, highly diverse, and complex datasets generated through. In healthcare, data heterogeneity and variety of structured, semi-structured and unstructured data comes from diverse biomedical data sources .Healthcare Big Data includes data on physiological, behavioral, molecular, clinical, environmental exposure, medical imaging, disease management, medication prescription history, nutrition, or exercise parameters [18].

There is a several ways of classifying Sources of big data in the literature. According to [19] data sources divide on two classes:

1. Administrative (Government, National surveys (Medical Expenditure Panel Survey), commercial vendors (health plans, PBMs)).
2. Clinical (Hospital, Physician, Integrated delivery network, Clinical database).

On the other hand [20] categorized Big Health Data sources into:

1. Providers: medical data (EHRs).
2. Payers: claims and cost data.
3. Researchers: academic, independent.
4. Consumers and Marketers: patient behavior and sentiment data.
5. Government: population and public health data.
6. Developers: pharmacy and medical device.

We conclude that healthcare data comes from two types of sources, internal sources

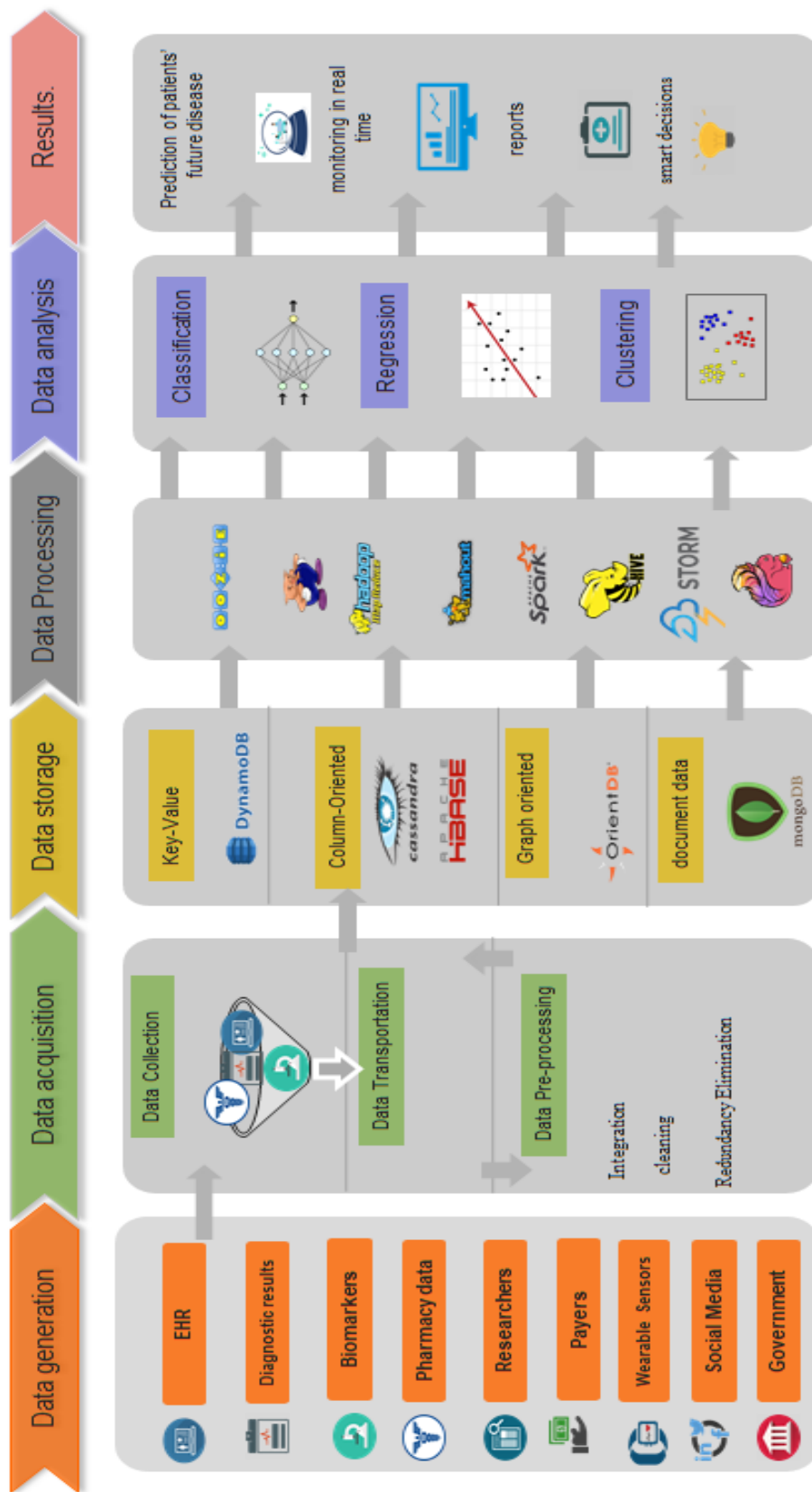


Figure II.3: General process of Big data in healthcare

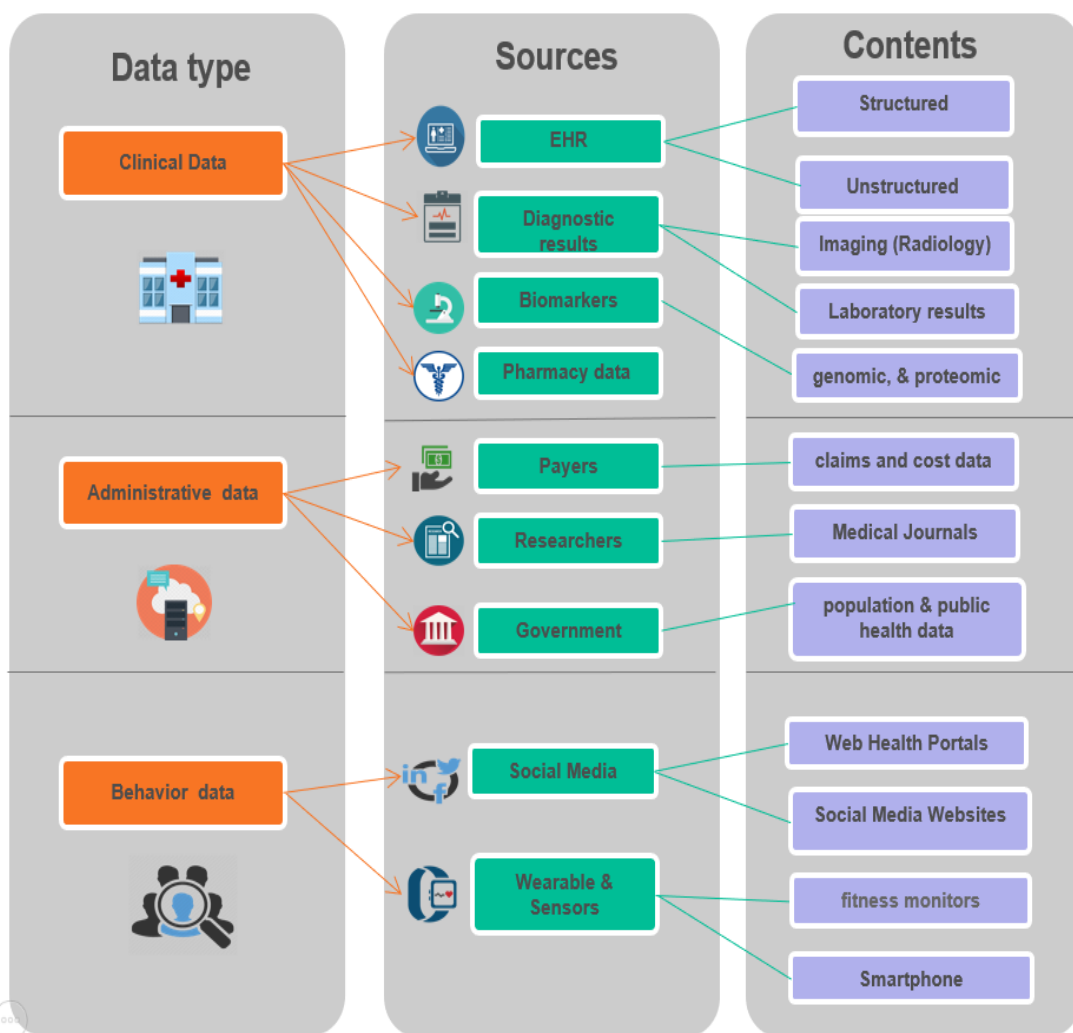


Figure II.4: Sources of big data in healthcare

such as EMRs, CPOE (computerized provider orders entry), imaging data, RD laboratories, pharmacy. In addition, external data sources such as government, insurance (claims billing), researches and social media. Based on previous classifications, we have inspired our health data classification, it shows type, source, and contents illustrated in Figure II.4.

1.2.2 Big data storage

Data storage refers to the management and the storage of large-scale datasets by achieving availability and reliability. A data storage system involves two parts: infrastructure and data storage mechanisms or methods. The hardware infrastructure

includes massive shared Information Communication Technology (ICT) resources used to feedback instant demands of tasks, and such ICT resources are organized flexibly [12]. The hardware infrastructure should provide elasticity and dynamic reconfiguration to adapt to diverse application environments. On the other hand, data storage methods are deployed on the top of hardware infrastructure to support large-scale datasets. Storage systems should be equipped with many interfaces, rapid query, and other programming models to analyze or interact with stored data. Figure II.5 illustrates storage tools for big data, and the following sections explain the techniques and technologies used to store and trait big data in details.

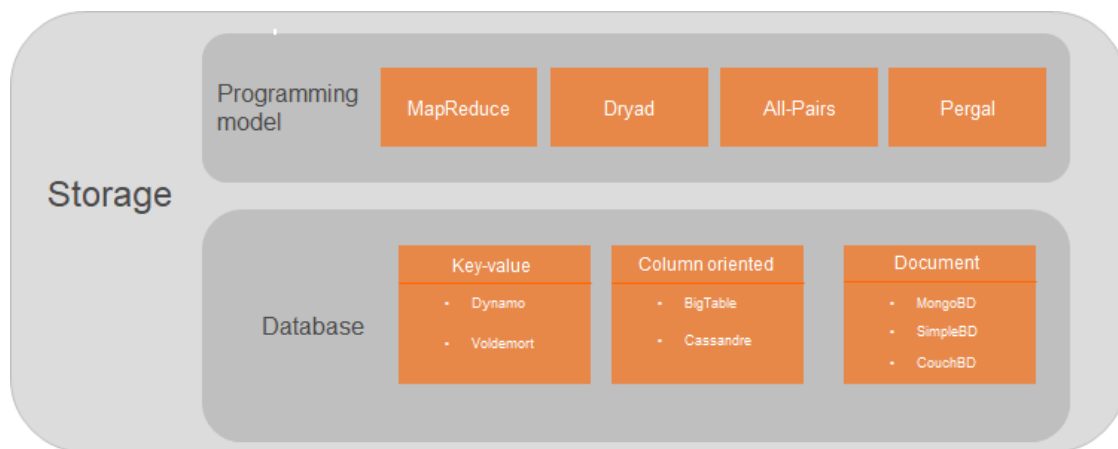


Figure II.5: Big data storage tools

Database Technology: Database technologies have been evolved these last years, and various database systems are developed to handle datasets at different scales and support various applications. This advancement allows us to avoid the limits of traditional relational databases that cannot meet the challenges on categories and scales brought by big data. NoSQL databases (nontraditional relational databases) are becoming more popular for big data storage. They provide a flexible model, simple API, support for simple and easy copy, eventual consistency, and extensive volume data support [12]. NoSQL databases are becoming the core technology for big data. We will examine the following three main NoSQL databases in the next sections: Key-value databases, column-oriented databases, and document-oriented databases.

Key-Value Databases: Key-value databases are developed by a simple data model, where data is stored corresponding to key-values. Every key is unique and users may input queried values according to the keys. Such databases provide a simple structure and the modern key-value databases are characterized with high expandability and smaller query response time higher than those of relational databases.

Column-Oriented Databases: The column-oriented databases store and process data according to columns other than rows. Columns and rows are segmented in multiple nodes to realize expandability. The column-oriented databases are mainly inspired by Google's BigTable which is a distributed, structured data storage system designed to process the large-scale data among thousands commercial servers [21].

Document Databases: Compared with key-value storage, document storage can support more complex data forms. Since documents do not follow strict modes, there is no need to conduct mode migration. Besides, key-value pairs can still be saved. There are three essential representatives of document storage systems, MongoDB, SimpleDB, and CouchDB (see Table II.1).

Database Programming Model: The massive datasets of big data are generally stored in hundreds and even thousands of commercial servers. The traditional parallel models such as Message Passing Interface (MPI) and Open Multi-Processing (OpenMP) may not be adequate to support such large-scale parallel programs. Some parallel programming modes have been proposed for specific fields. These models effectively improve the performance of NoSQL and reduce the performance gap between relational databases. Therefore, these models have become the cornerstone for the analysis of massive data.

MapReduce: MapReduce [30] model for large-scale computing using a large number of commercial PCs clusters to achieve automatic parallel processing and distribution. In MapReduce, the computational workload is caused by inputting key-value pair sets and generating key-value pair sets. The computing model only has two functions,

Big data platforms	Platform role	Store type
Cassandra [22]	Apache Cassandra is an open-source distributed NoSQL database management system designed to handle large amounts of data across many commodity servers, offers robust support for clusters spanning multiple datacenters [23]	Column oriented data stores
MongoDB [24]	MongoDB is an open source distributed document-oriented database. Classified as a NoSQL database program, .It uses JSON-like documents to store the data. Semi-structured data such as texts, time-stamped log, geo-info or even arrays and nested hash tables can be stored in MongoDB	document data stores
DynamoDB [25]	Dynamo is a highly available and expandable distributed key-value data storage system. It is used to manage store status of some core services in the Amazon. It's a set of techniques that when taken together can form a highly available key-value structured storage system [12]	Key-value stores
Hadoop [26] Distributed File System (HDFS)	HDFS is a popular type of cluster file system which is designed for reliably storing large amount of data across machines in a large scale cluster [27].	/
OrientDB [28]	OrientDB is an open source NoSQL database management supporting graph, document, key/value, and object models but the relationships are managed as in graph databases with direct connections between records.	Graph oriented data stores
HBase [29]	HBase is a column-oriented database management system that sits on top of HDFS. It uses a non-SQL approach [29]	Column oriented data stores

Table II.1: Big data platforms with storage type.

i.e., Map and Reduce, both of which are programmed by users. The Map function processes input and generates intermediate key-value pairs. Then, MapReduce will combine all the intermediate values related to the same key and transmit them to the Reduce function. Next, the Reduce function receives the intermediate key and its value set, merges them, and generates a smaller value set. MapReduce has the advantage that it avoids the complicated steps for developing parallel applications, e.g., data scheduling, fault-tolerance, and inter-node communications. The user only needs to program the two functions to develop a parallel application.

Dryad: Dryad [31] is a general-purpose distributed execution engine for processing parallel applications of coarse-grained data. The operational structure of Dryad is a directed acyclic graph, in which vertexes represent programs and edges represent data channels. Dryad executes operations on the vertexes in computer clusters and transmits data via data channels, including documents, TCP connections, and shared-memory FIFO. During operation, resources in a logic operation graph are automatically mapped to physical resources. The operation structure of Dryad is coordinated by a central program called job manager, which can be executed in clusters or workstations

of users. The user workstations can access clusters through the network. Besides, Dryad allows vertexes to use any amount of input and output data, while MapReduce supports limited computing, with only one input set and generating only one output set.

All-Pairs: All-Pairs [32] is a system specially designed for biometrics, bioinformatics, and data mining applications. It focuses on comparing element pairs in two datasets by a given function. The All-Pairs problem may be expressed as a three-tuples (Set A, Set B, and Function F), in which Function F is utilized to compare all elements in Set A and Set B. The comparison result is an output matrix M. It is also called the Cartesian product or cross join of Set A and Set B. All-Pairs is implemented in four phases: system modeling, input data distribution, batch job management, and result collection.

Pregel: The Pregel [33] system of Google facilitates the processing of large-sized graphs, e.g., analysis of network graphs and social networking services. A computational task is expressed by a directed graph constituted by vertexes and directed edges, in which every vertex is related to a modifiable and user-defined value. Directed edges are related to their source vertexes and every edge is constituted by a modifiable and user-defined value and an identifier of a target vertex. After the graph is built, the program conducts iterative calculations called supersteps among which global synchronization points are set until algorithm completion and output completion. In every superstep, vertex computations are parallel and every vertex executes the same user-defined function to express a given algorithm logic.

1.2.3 Big data analysis:

Data analysis is the final and the most important phase in the value chain of big data, with the purpose of extracting useful values, providing suggestions or decisions. Different levels of potential values can be generated through the analysis of datasets in different fields [34]. Big data analytics is often a complicated process of analyzing big data to reveal hidden patterns, correlations that can help organizations make the right decisions. Big data analytics is a form of advanced analytics, which involve complex applications with elements such as :

- Data mining: which sift through data sets in search of patterns and relationships;
- Predictive analytics: which build models to forecast customer behaviour and other future developments;
- Machine learning: which taps algorithms to analyze large data sets;
- Deep learning: a more advanced offshoot of machine learning.

2 Machine learning and Deep learning

Machine learning is part of artificial intelligence focused on constructing algorithms that make predictions based on data without programming it to perform the task. ML aims to identify a function $f: X \rightarrow Y$ that maps the input X into output Y [35]. Functions f are chosen from different function classes, dependent on the type of learning algorithm used. Machine learning algorithms can be classified mainly into three categories by the type of datasets used as experience.

2.1 Machine learning categories

Machine learning categories involve supervised learning, unsupervised learning and reinforcement learning. Other learning systems combine two categories, such as semi-supervised learning that use labelled and unlabeled data. More details in the followings sections.

2.1.1 Supervised learning

Supervised learning systems make use of labeled datasets, where x represents a data point and y the corresponding true prediction for x . This training set of input-output pairs is used to find a deterministic function that maps any input to an output, predicting future input-output observations while minimizing errors as much as possible [36]. Supervised learning problems can be further grouped into regression and classification problems:

- **Classification:** a classification problem is when the output variable is a category, such as “disease” and “no disease”. Classes can be called as targets/labels or categories (See Figure II.6).
- **Regression:** a regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some popular examples of supervised learning algorithms are:

- Linear regression

- Random forest
- Support vector machines
- Decision Tree
- Neural network (Multiple layer perceptron)
- K-Nearest Neighbours
- Naïve Bayes

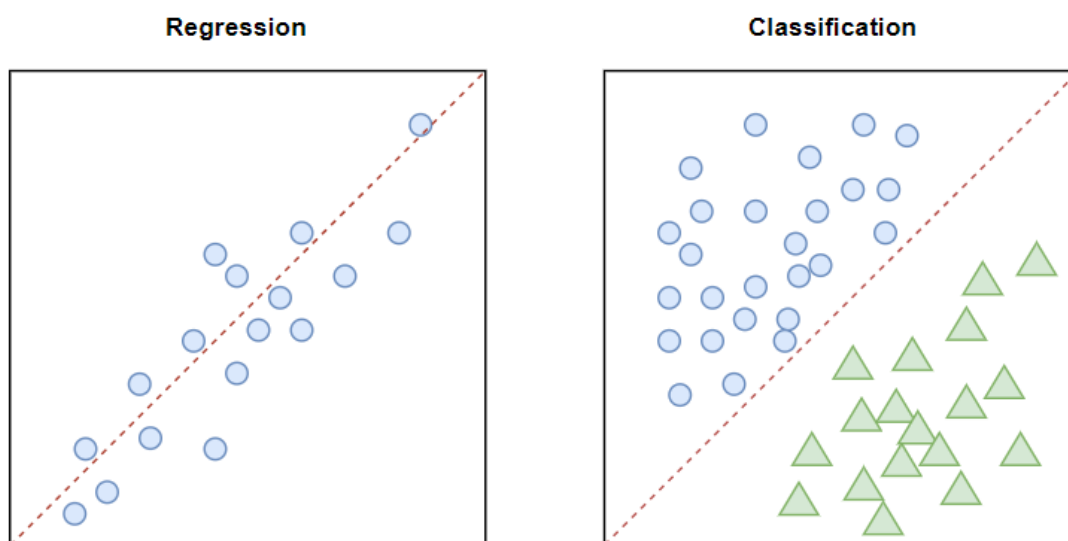


Figure II.6: Supervised learning.

2.1.2 Unsupervised learning

Unsupervised learning systems use unlabeled datasets to train the system. The objective of unsupervised learning is to derive structure from unlabeled data by investigating the similarity between pairs of objects, and is usually associated with density estimation or data clustering [16]. Unsupervised learning problems can be further grouped into clustering and association problems (See Figure II.7).

- Clustering: is a way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group. It does it by finding some

similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

- Association: checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable [37]. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

Some popular examples of unsupervised learning algorithms are:

- k-means for clustering problems.
- Apriori algorithm for association rule learning problems.

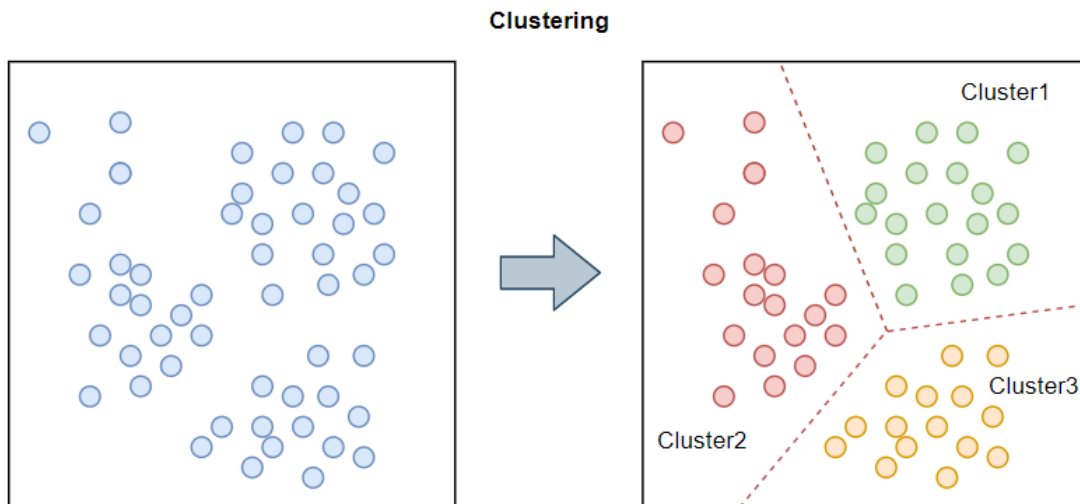


Figure II.7: Unsupervised learning.

2.1.3 Reinforcement learning

Reinforcement learning systems do not experience a fixed dataset, but a feedback loop between the system and its experiences [38]. As shown in Figure II.8, state-action-reward triples are observed as the data. The objective of reinforcement learning is mapping situations to actions with the goal of maximizing rewards .

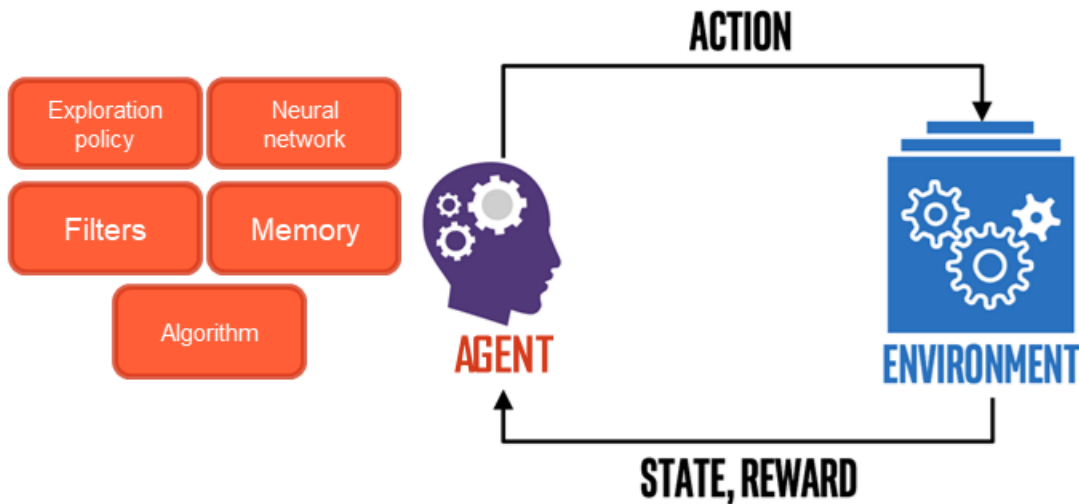


Figure II.8: Reinforcement learning.

2.2 Deep learning

Deep learning is a sub-field of machine learning dealing with algorithms inspired by the structure and function of the brain called artificial neural networks. In other words, It mirrors the functioning of our brains and how nervous system structured where each neuron connected each other and passing information [39].

Convolutional neural networks is currently one of the most prominent algorithms for deep learning with image data. Whereas for traditional machine learning relevant features have to be extracted manually. deep learning uses raw images as input to learn certain features [40]. CNNs consist of an input and output layer, and several hidden layers between the input and output. Examples of in between layers are convolutional layers, max-pooling layers and fully connected layers explained in the following items and Figure II.9.

- Convolutional Layer (CONV): Convolutional filters are used to derive an activation map from the input data.
- Pooling Layer (POOL): Performs nonlinear down-sampling and cuts down the amount of parameters for a simpler output.
- Fully Connected Layer (FC): Computes the class probability scores by outputting of C dimensions, with C being the number of classes. All neurons are connected

to this layer.

- Activation function layer: the common used is ReLU (Rectified Linear Unit) which applies the non-saturating activation function $f(x) = \max(0, x)$ [55]. It effectively removes negative values from an activation map by setting them to zero. [66] It increases the nonlinear properties of the decision function and of the overall network without affecting the receptive fields of the convolution layer.

There are various CNNs architectures which have been developed in literature such as: VGGNet, GoogLeNet, ResNet, MobileNet, and the recent one EfficientNet.

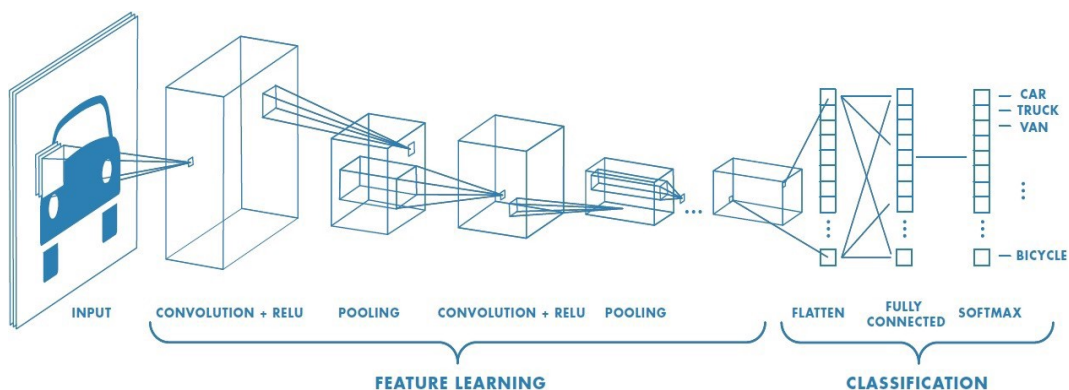


Figure II.9: Example of CNN architecture.

2.3 Transfer learning

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. In other words, Transfer learning is an approach in deep learning and machine learning where knowledge is transferred from one model to another [41]. In this method, pre-trained models are used as the starting point on computer vision instead of developing models from the very beginning. This allows us to handle the challenge of the large amount of computing and storage resources required to develop Deep Learning models. However, it should also be noted that transfer learning only works in deep learning if the model features learned from the first task are general [42].

Transfer learning is becoming the go-to way of working with deep learning models. The reasons are explained below:

- Lack of data : Deep learning models require a LOT of data for solving a task effectively. However, it is not often the case that so much data is available. In that case, a specific target task can be solved using a pre-trained model for a similar source task.
- Speed: Transfer learning cuts a large percentage of training time and allows for building various solutions instantly. In addition, it prevents from setting up a complex and costly Cloud GPU/TPU.

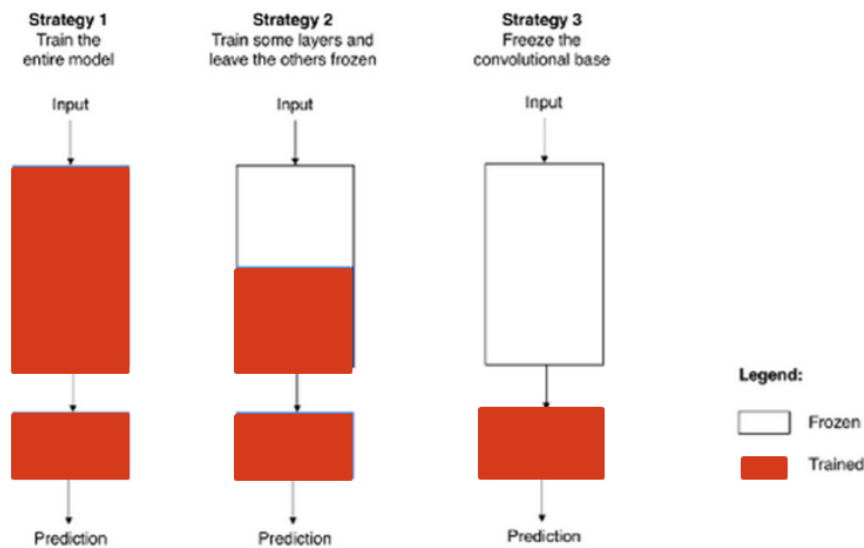


Figure II.10: Fine tuning strategies for pre-trained models.

As shown in Figure II.10 Transfer learning can be applied through several different strategies by fine-tuning the model according to one of three strategies:

1. Train the entire model: In this case, we use the architecture of the pre-trained model and train it according to our dataset. In other words, we are learning the model from scratch, so we'll need a large dataset and a lot of computational power.
2. Train some layers and leave the others frozen: lower layers refer to general features (problem independent), while higher layers refer to specific features (problem dependent). Here, we play with that dichotomy by choosing how much we want to adjust the weights of the network (a frozen layer does not change during training). Usually, if we have a small dataset and a large number of parameters,

we'll leave more layers frozen to avoid overfitting. By contrast, if the dataset is large and the number of parameters is small, we can improve the model by training more layers to the new task since overfitting is not an issue.

3. Freeze the convolutional base: This case corresponds to an extreme situation of the train/freeze trade-off. The main idea is to keep the convolutional base in its original form and then use its outputs to feed the classifier. we're using the pre-trained model as a fixed feature extraction mechanism, which can be useful if we're short on computational power, your dataset is small, and/or pre-trained model solves a problem very similar to the one we want to solve.

2.4 Machine learning process

Machine Learning is a data-driven process that starts by pre-processing the collected data until the model construction spits out predictions and insights. The method of performing machine learning usually requires many steps that are explained in the following sections:

2.4.1 Data preprocessing

Data preprocessing is the first and crucial step while creating a machine learning model. It is a process of preparing the raw data and making it suitable for a machine learning model. A typical healthcare data preprocessing procedure usually includes the following steps depending on the source and format of the data :

Data cleaning: Data cleaning is the method of detecting, correcting or removing wrong or inaccurate records from images, table, or database and identifying incorrect, incomplete, irrelevant or inaccurate parts of the data and then replacing, modifying, or deleting the row data [43].

Missing value interpolation: In health analytics, missing data may be unavoidable due to a variety of reasons, for example, faulty equipment, and/or imprecise or lost measurements; moreover, the errors of the caregivers, for instance, physicians or

nurses who forget and/or improperly record the information may also lead to missing information. Yet, the most serious problems of missing values are the resulting consequences, thereby effectively slowing down the analytic processing due to lower efficiencies, and/or the potential to compromise the information extracted from the data, there by leading to faulty conclusions.

Essentially, three strategies may be applied to deal with missing data. The first is missing data ignoring techniques that simply delete the cases that comprise the missing data [44]. In cases where the size of the data is small (as with the current study), deleting any information is not ideal. The second approach would be to deploy missing data modeling techniques. The strategy here is to define a model from the existing data and then generate inferences based on the distribution of the data [44]. The third strategy is to employ the missing data imputation techniques. These techniques complete the missing data in the dataset with a potential value [45]. Examples of such techniques include: Mean regression, K-NNs, and multiple imputations.

Data synchronization: Data synchronization ensures secure, accurate, compliant data. It assures harmony between each source of data and its different endpoints. As data comes in, it is cleaned, checked for errors, consistency and duplication before being used [46]. Data must always be consistent throughout the data record. If data is modified in any way, changes must upgrade through every system in real-time to avoid mistakes, prevent privacy breaches, and ensure that the most up-to-date data is the only information available. Data synchronization ensures that all records are consistent, all the time.

Data normalization: This step is usually needed to adapt to differences in the data recording process. For example, a daily heart rate may represent a daily average heart rate or a measurement during a specific time range. Moreover, a normalization step is usually performed to transform the original feature into a similar format by adopting and mapping standardized terminologies and code sets. A normalization process should sometimes be carried out to convert the original numerical values to nominal ones for a specific algorithm [47]. It is deserving of mentioning that the discretization

process may cause information loss and impact data quality.

Imbalanced data problem: Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally. The imbalance of medical data, as characterized by the non-uniformity of the class distribution among the classes, seriously affects the accuracy of medical diagnosis classification. Data imbalance exists widely in real-world datasets, especially those in the medical field. To resolve this challenge, a widely implemented technique for dealing with highly unbalanced datasets is resampling:

1. Under-sampling: is resampling consists of eliminating samples from the majority class which can cause wast of information.
2. Over-sampling is to duplicate random samples from the minority class, which can affect over-fitting [48].
3. Generate synthetic samples: is to sample the attributes from instances in the minority class randomly. We could sample them empirically within a dataset or use a method like Naive Bayes to sample each attribute independently when run in reverse. If data is different and no linear relationships between the attributes may not be preserved.

There are systematic algorithms that generate synthetic samples. The most popular of such algorithms is called SMOTE or the Synthetic Minority Over-sampling Technique [49]. As its name suggests, SMOTE is an oversampling method. It works by creating synthetic samples from the minor class instead of creating copies. The algorithm selects two or more similar instances (using a distance measure) and perturbing an instance one attribute at a time by a random amount within the difference to the neighbouring instances.

2.4.2 Feature selection

Appropriate feature identification has become an essential task to apply data mining algorithms effectively in real-world scenarios. Therefore, many feature selection

methods have been proposed to obtain the relevant features or feature subsets in the literature to achieve their classification and clustering objectives. There are three main approaches for feature selection:

Filter methods: The filter approach incorporates an independent measure for evaluating features subsets without involving a learning algorithm. This approach is efficient and fast to compute (computationally efficient). However, filter methods can miss features that are not useful by themselves but can be very useful when combined with others. Some existed techniques for filter methods are presented in Table II.2.

Wrapper methods: The filter and wrapper approaches can only be distinguished by the evaluation criteria. Different wrapper algorithms can be generated by varying the subset generation and subset evaluation measure (using dependent criterion). The wrapper approach selects an optimal subset that is best suited to a learning algorithm. Therefore, the performance of the wrapper approach is usually better (see Table II.2).

Embedded methods: This approach combines with the learning algorithm at a lower computational cost than the wrapper approach. It also captures feature dependencies. It considers relations between one input features and the output feature and searches locally for features that allow better local discrimination. It uses the independent criteria to decide the optimal subsets for a known cardinality. The learning algorithm is used to select the optimal subset among the optimal subsets across different cardinality.

2.4.3 Choosing a model

There are various existing models developed by data scientists which can be used for different purposes. These models are designed with different goals in mind. For instance, some models are more suited to dealing with texts, while another model may be better provided to handle images. We need to make the choice that meets our expected outcome. The options for machine learning models can be explored across three broad categories shown in Figure II.11 .

	Filter		Wrapper		Embedded
Time complexity	low		medium		high
Performance	Low		high		high
Suitable on big data	very high		high		high
Example of techniques	Univariate	Multivariate	Deterministic	Randomized	decision tree selection using the weight vector of SVM weighted Naive bayes
	X^2 Information gain Euclidian distance Gain ratio	correlation based feature selection	sequential forward selection (SFS) backward elimination (SBE), Beam search	simulated annealing genetic algorithm randomized hill climbing	

Table II.2: Comparison of different feature selection techniques.

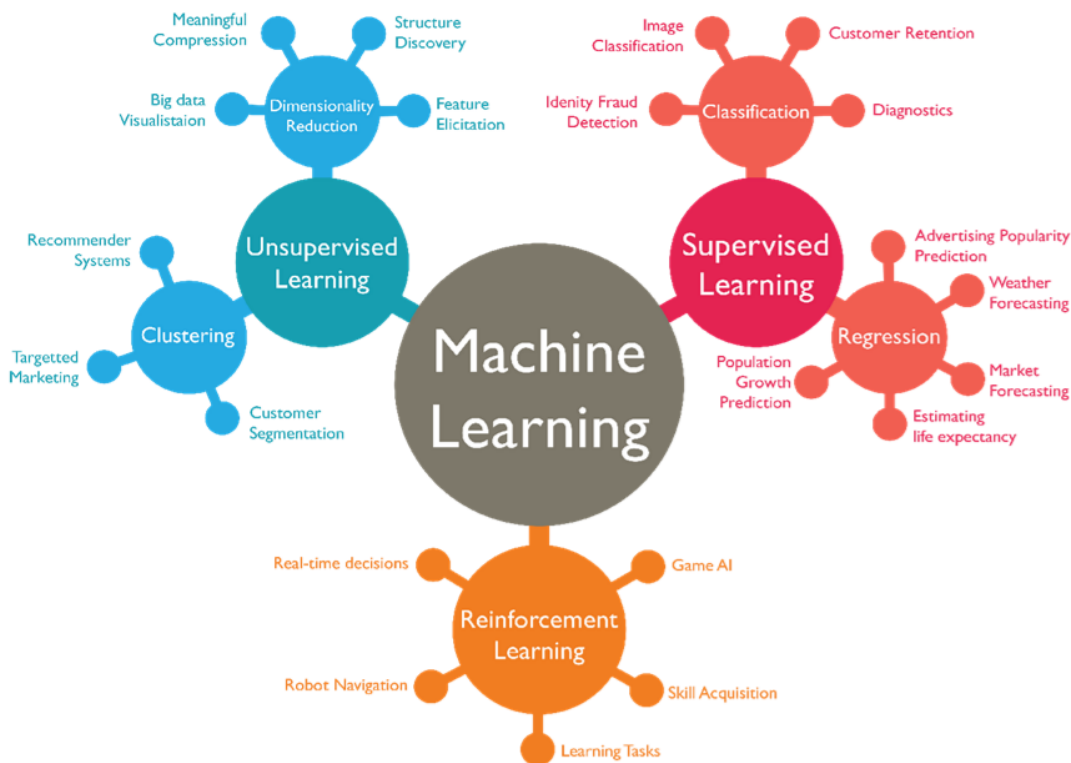


Figure II.11: Machine learning techniques.

2.4.4 Model evaluation

Evaluating a model is a crucial step throughout the development of the model. Evaluation metrics have a correlation with machine learning tasks. Figure II.12 illustrates various evaluation metrics based on the type of tasks(classification, regression, etc.) all have different metrics. In this section, we are going to shed light on the evaluation metrics used for classification.

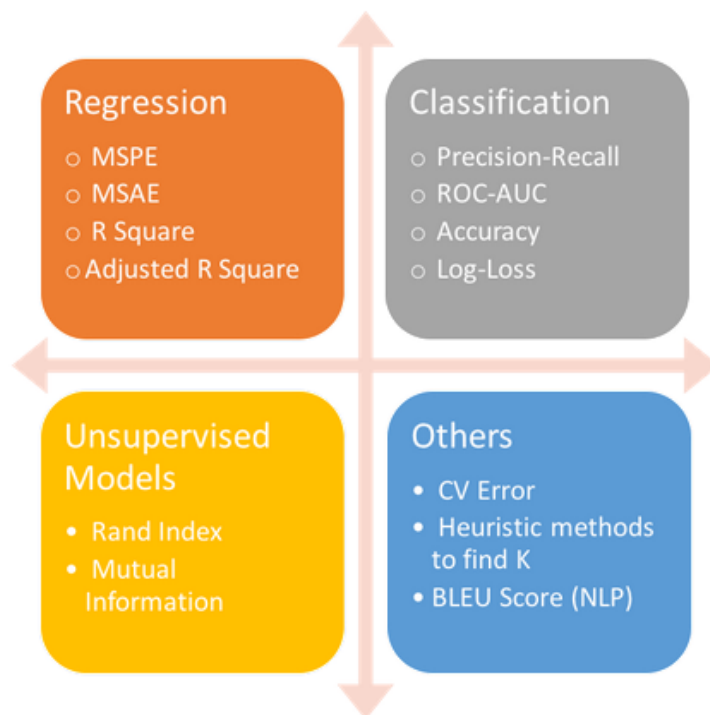


Figure II.12: Evaluation metrics.

- Accuracy refers to the whole number of instances that may be classified correctly. It is given by

$$\text{Accuracy} = \frac{TP + TN}{TN + TP + FP + FN} \quad (\text{II.1})$$

Where:

- TP= True positive;
- TN= True negative;
- FP= False positive;
- FN= False Negative.

- Sensitivity measures the quantity of TP instances, which are correctly identified by the classifier. It is given by

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (\text{II.2})$$

- Specificity measures the quantity of TN instances, which are correctly identified

by the classifier. It is given by

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (\text{II.3})$$

- Precision measures the amount of predicted TP that is truly related to the TP class. It is given by

$$\text{Precision} = \frac{TP}{TP + FP} \quad (\text{II.4})$$

- F1-measure is a combination of precision and sensitivity. Therefore, a high value of F-measure shows a high value of both precision and sensitivity [50]. It is given by

$$\text{F1-measure} = 2 \frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (\text{II.5})$$

- The Receiver Operating Characteristics (ROC) curve is a graphical plot used to compare the performance of a binary classifier. Area Under Curve (AUC)
- AUC is calculated for assessing performance of the classifier and provides an examination of the classifier stability and consistency.

3 Medical PHM

Prognostics and Health Management (PHM) approach, and theoretical models have had great success for industrial systems. Therefore, this accomplishment motivates us to think about potential extension of the PHM approach in such area as the medicine. Many researchers from various engineering fields have been focused on PHM tools, in order to decrease the maintenance cost of industrial resources and enhance system safety, availability, and reliability [51].

PHM has seven pillars: data generation, data acquisition, state detection, diagnosis, prognosis, decision-making, and human machine interface. The main duties of PHM expertise are to identify incipient system fault or component; implement failure diagnostics, failure prognostics, and health management [51]. All of those objectives are needed in the medical filed as well. In other words, we look for detect body or organ fault (disease), perform disease diagnosis, disease prognostics, and health management. the inspired process of Medical PHM (M-PHM) is shown in Figure II.13.

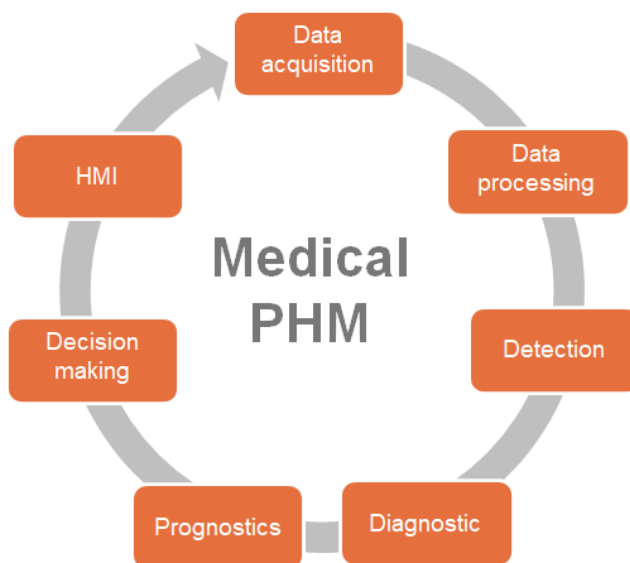


Figure II.13: General process of Medical PHM.

3.1 Engineering PHM VS medical PHM

Table II.3 presents the similarities and differences between machine and human body in PHM view. We may infer that an adaptation of PHM can be applied in the

medical filed with two key points being taken into account:

- Working on human body is more complicated, this is due to the complexity of organs and the interactions between them are sometimes unknown and less predictable. Furthermore, the sensibility of working on human lives, any mistake leads to critical consequences.
- The data generation and sharing in medical filed still hard to achieve, also we have to take in consideration the privacy of patients.

	Industrial system	Human body
Complexity	Interactions between components and failure modes may be well-defined	More complicated interactions. Biological failure modes of a human body or organs could be less predictable
Risk factors	Component aging, damage accumulation and fault progression	Concepts on natural history, clinical course, diseases progression, lifestyle, and environment
Data generation	Sensors data is the most type used in industry (Vibration, temperature, humidity, etc.)	Various types of data: wearable sensors (Blood pressure, heart rate, Fasting blood sugar, etc.), images (MRI,CT, ultrasounds , WSI, etc.), reports and prescriptions text, clinico-histological features, etc.
Diagnosis	Identify the system degradation behavior	Disease detection by identifying the type, and the cause based on clinico-histological data, etc.
Prognostics	Predict the component RUL	Predict getting the disease based on risk factors Predict the recurrence of disease
Decision-making	Determine optimal maintenance policies	Select optimal treatment, and prevention policies

Table II.3: Industrial system vs. human body contrast in PHM view.

3.2 M-PHM analytics

M-PHM analytics are steps that utilises various techniques including modelling, data mining, and statistics, as well as artificial intelligence (AI) such as machine learning and deep learning to evaluate historical and real-time data and make predictions about the future [52].

Figure II.14 illustrates the four health analysis models by showing the role and

difficulty of each one. Both descriptive analytics and diagnostic analytics look to the past to explain what happened and why it happened. Predictive analytics and prescriptive analytics use historical data to forecast what will happen in the future and what actions you can take to affect those outcomes. More details for each level are explained below:

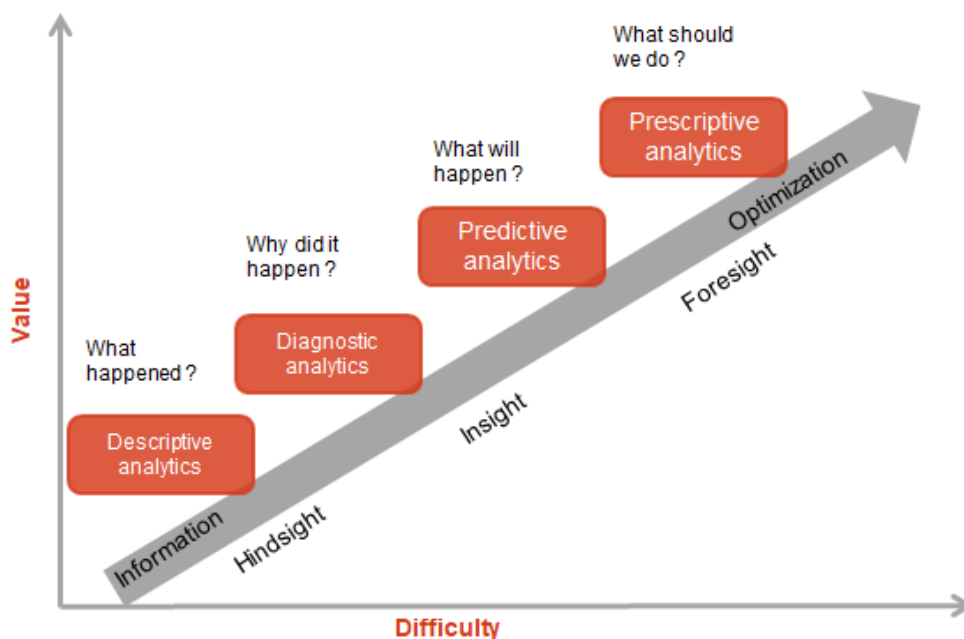


Figure II.14: Four types of M-PHM analytics.

3.2.1 Computer Aided Detection (Descriptive analysis)

Descriptive analytics is used to explain what was happening in a given situation [53]. This class of analytics can be used in healthcare to detect anomalies in a given dataset. for example:

- Screening mammography for the early detection of breast cancer.
- Detection of colorectal polyps in the colon in CT colonography.
- Identify subjects with Alzheimer's and mild cognitive impairment from normal elder controls

- Pathological brain detection (PBD)
- Automatic detection of significant coronary artery disease in coronary CT angiography

3.2.2 Computer Aided Diagnosis (Diagnostic analysis)

Diagnostic analytics takes descriptive data a step further and provides a more in-depth analysis to answer why this happened? This includes using processes such as data discovery, data mining, and drill down and drill through [54]. In the healthcare example mentioned earlier, diagnostic analytics would explore the data and make correlations for example :

- Diagnose the type and stage of disease
- Identify patterns of care and discover associations from massive healthcare records
- Interpretation of medical images such as X-ray, MRI, and ultrasound diagnostics precisely and in a short time.

3.2.3 Computer Aided Prognostic (Predictive analysis)

Predictive analytics can be described as a branch of advanced analytics utilized to make predictions about unknown future events or activities that lead to decisions [55]. For example:

- Predict whether a patient is at a high risk to have a disease based on risk factors.
- Predict the recurrence of breast cancer
- Predict the survival of patients
- Which patient is likely to be readmitted after surgery
- Whether a patient will stay longer than the average after surgery.

For this reason, predictive analytics in healthcare settings has received a significant amount of interest over the past few years. The knowledge gained through applying predictive analytics in health and medicine will change how medicine is practised while enhancing our ability to prevent and treat significant diseases and illnesses.

3.2.4 Computer Aided Decision making (prescriptive analysis)

Prescriptive analytics is an area of data analytics that focuses on prescribing possible actions and solutions for a problem. It uses modelling, data mining, and artificial intelligence to evaluate historical data and real-time data to make right decisions [56]. It gives the healthcare companies multiple ‘what if’ options to compare to find the best possible solution for the patient for example :

- Empowers healthcare providers with the capability to do something about it, helping them take the best action to mitigate or avoid a negative consequence.
- Determine the maximum dosage of the drug that is effective to maximize treatment outcome.
- Personalized medicine and evidence-based medicine are both supported by prescriptive analytics
- Allows health care providers to consider recommended actions for each of those predicted outcomes.
- Lower the cost of healthcare from patient bills to the cost of running hospital departments. In other words, it helps in making sound financial and operational decisions, providing short-term and long-term solutions to administrative and financial challenges.
- Provides enormous scope and depth as developers improve technologies in the future. It is making significant advances concerning patient care quality and timeliness and is reducing clinical and financial risks.

4 Conclusion

This chapter introduced the principal axes of our thesis: Big Data, Machine learning & deep learning, and M-PHM. Digitalization of the medical sector has led to a massive growth of data (Big Data) which come from various sources. The healthcare industry needs to work on detection, diagnosis, prediction, and prevention (M-PHM) to improve outcomes. To achieve this, we use artificial intelligence techniques (Machine Learning & Deep Learning) which examines such large data sets and uncovers hidden information

and patterns to discover knowledge from the data, as well as personalized medicine, all in real-time.

The following chapter aims to present the first contribution of this thesis, which aims to reduce the gap between industrial PHM and medical PHM. This work is a retargeting PHM model from fault diagnosis of an aircraft engine to diagnosis human heart disease.

Chapter III

Retargeting PHM tools: from industrial to medical field

This chapter presents the first contribution in this thesis which consists of applying an adaptation of a PHM model from fault diagnosis of aircraft engine to diagnosis human heart disease. For that adaptation, an algorithm for retargeting extreme learning machine (ID-RELM) is applied. We give first during the presentation of our approach a quick overview on this adaptation in section 1. Then we detail each of the method steps in the sections 2. In Section 3, experimental results to demonstrate the effectiveness of our system is provided. Finally, a conclusion and future works are presented in Section 4.

1 Motivation

Many researchers from various engineering fields have been focused on Prognostics and Health Management (PHM) tools, in order to decrease the maintenance cost of industrial resources and enhance system safety, availability, and reliability [51]. Recall that PHM has seven pillars: data generation, data acquisition, state detection, diagnosis, prognosis, decision-making and human machine interface. Works piloted in PHM research concentrate on developing accurate and robust models to evaluate the health state of systems by making diagnosis, prognostic, and support decision making.

Automatic diagnosis and prognostic in medical domain has been an active area

in computer science field in last decades. Various medical domains attract researches, more precisely oncology and chronic diseases. In this work, we are interesting in cardiology field.

Heart disease is the major cause of death in the universal and the number of patients with heart disease is growing each year [57]. According to World Health Organization (WHO) reported data, around 17300000 persons died worldwide from cardiovascular diseases (CVDs). This statistics shows the need of having computer aided diagnosis (CAD) system that is able to give a preliminary assessment of a patient based on simple medical tests that are accessible to everyone [58]. CAD for heart disease was widely developed using data mining and machine learning [59], however the performance still needs improvement to make accurate classification.

The aim of this contribution is to apply an adaptation of a PHM model from fault diagnosis of aircraft engine [60] to diagnosis human heart disease. This PHM model is based on a new strategy by retargeting extreme learning machine (ELM) algorithm. ELM is a single feedforward neural network; its structure involves a single layer of hidden nodes, where the weights between inputs and hidden nodes assigned randomly, and remain constant during training and testing phases. On the other hand, the weights that connect hidden nodes to outputs can be trained very fast. The idea behind retargeting ELM is to avoid limits of the original one regarding the random hidden nodes generation by retargeting its label vectors. The proposed method need less hidden nodes to achieve the same classification performance, which means improving the processing real time.

In order to develop a CAD for heart disease, a new scheme is proposed based on PHM adaptation using UCI heart disease dataset [10]. The proposed scheme consists of: (a) the pre-processing step to improve data quality (missing data imputation and scaling dataset), (b) the feature selection step to improve classification performance (based on embedded method), (c) the diagnosis phase to identify the absence or the presence of heart disease (using Dragging Regularized ELM (ID-RLM)).

2 Adapted PHM tool

We here present the adaptation process of the proposed PHM tool for aircraft engine [60] to heart disease diagnosis. Figure III.1 shows the general schematic diagram of our proposed tool adapted from [60] to diagnose heart disease. The details of each processing stage are described in the subsequent sections.

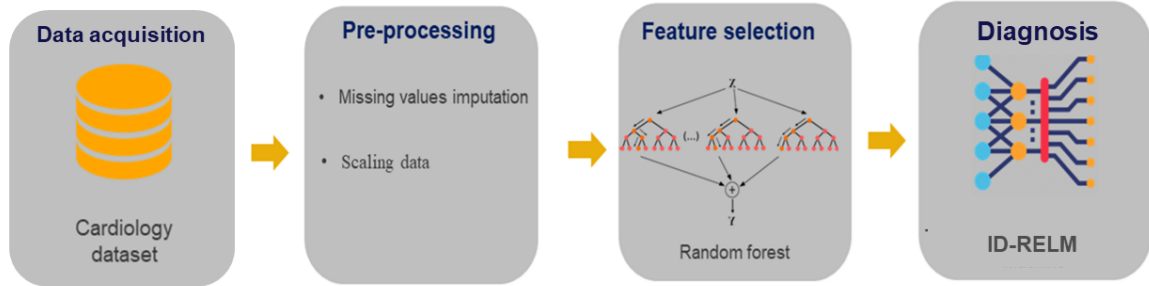


Figure III.1: A general scheme of our CAD of heart disease.

2.1 Data description

The Cleveland dataset [10] used in this study was created by the University of California Irvine (UCI) Machine Learning Repository heart disease dataset, including four independent databases funded by four independent medical institutions. The Cleveland dataset contains 303 cases of patient data, involving some missing values. Table III.1 shows the Cleveland dataset attributes with their definitions and type.

2.2 Data preprocessing

In order to achieve more accurate results, data pre-processing is an important step in changing raw heart disease dataset into a clean and understandable format for analysis. The following sub-sections discuss techniques applied to improve the quality of our dataset.

	Feature	Input type	Input range
age	Age in years	Numeric	[29, 77]
sex	Sex	Binary	0 = female 1 = male
cp	Chest pain type	Nominal	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic
trestbps	Resting blood pressure on admission to the hospital in mm Hg	Numeric	[94, 200]
chol	Serum cholestoral in mg/d	Numeric	[126, 564]
fbs	Fasting blood sugar is greater than 120 mg/dl or not	Binary	0 = false 1 = true
restecg	Resting electrocardiographic results	Nominal	0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
thalach	Maximum heart rate achieved	Numeric	[71, 202]
exang	Exercise induced angina	Binary	0= no 1=yes
oldpeak	ST depression induced by exercise relative to rest	Numeric	[0,6.2]
slope	The slope of the peak exercise ST segment	Nominal	1 = upsloping 2 = flat 3 = downsloping
ca	Number of major vessels (0-3) colored by flourosopy	Nominal	0-3
thal	The heart status	Nominal	3 = normal 6 = fixed defect 7 = reversable defect
num	Diagnosis of heart disease		0 = less then 50% diameter narrowing (normal) 1 = greater then 50% diameter narrowing (patient)

Table III.1: UCI heart disease dataset description.

2.2.1 Missing data

In this contribution, we are going to apply the KNN techniques, which is an algorithm that is useful for matching a point with its closest k neighbors in a multi-dimensional space [61]. It can be used for data that are ordinal, continuous, categori-

cal, and discrete which makes it particularly useful for dealing with all types of missing values.

2.2.2 Scaling data

In this step, data columns are re-scaled to a range of $[0 - 1]$ for two causes. The first is one is to simplify the complexity of digital computing. The second one is to get rid of attributes in the largest numeric range while controlling attributes in the smallest numeric range [62].

2.2.3 Feature selection

Feature selection process is very significant to find most relevant attributes to the classification and then to the diagnosis. It has many advantages: (1) To make the model simpler to interpret. (2) To decrease the variance of the model, and therefore over-fitting. (3) To decrease the computational time and cost. (4) And finally, the most important one, is to increase the performance of the model [63].

In the literature, there is three main types of feature selection: filter, wrapper, and embedded methods. In this contribution we will choose the third one, which combine the qualities of filter and wrapper methods as implemented by algorithms that have their own built-in feature selection methods. Random Forests (RF) are often used as embedded feature selection in data science. The reason is tree-based strategies used by RF can logically orders by how well they improve the clarity of the node [64]. at the start of the trees, we find nodes with the highest decrease in impurity, while nodes with the minimum decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, we can create a subset of the most important features. Figure III.2 shows the features selected by RF with their score, where x-axis is the feature indexes and y axis is the feature importance.

2.3 Diagnosis

Extreme learning machine (ELM), proposed and implemented by [65] as a single hidden layer feedforward network, which has expanded applications in many machine

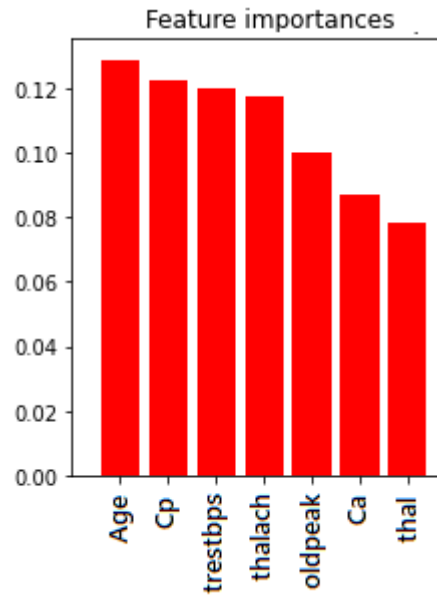


Figure III.2: Features selected by RF algorithm on UCI dataset and their score.

learning tasks, such as classification, regression, etc. The reason behind this success is efficiency and effectiveness of the model based on two main advantages: (1) ELM generates randomly hidden layer biases and the input weights, and fixes them without tuning by iterations the process of determining the output weights, Unlike the traditional neural network, as error backpropagation; (2) ELM searches to minimize both training errors and the norm of output weights, based on Bartlett's theory which benefits the generalization on the unseen data.

Despite the evidence of the drawback of ELM, it has a weak point presented in the generation of extra hidden nodes to reach the same generalization performance as the traditional neural networks. A large size network leads to more computational time in the testing phase, which is not suitable for testing time view. Therefore, a lot of proposed algorithms lean to compact the ELM architecture. The traditional methods overcome this disadvantage by optimizing the network structure. However, a data structure viewpoint is proposed by [60], which is different from the previous viewpoint of network structure.

They care about the margin instead of the reference points, so a flexible dragging strategy is developed. We here apply the improved version (ID-RELM) for diagnosis, which abandons the reference points and can improve the classification performance by

retargeting the ELM label vector. In this way, one can obtain a higher classification performance with a lesser network size and processing time. This seems to be beneficial for a real time application for heart disease diagnosis.

3 Experimentation

This experiment is compiled and run in google colab environment with python language using scikit-learn bibliography, and the default values of functions are used for all parameter values that are not explicitly stated. The used model is evaluated using the below metrics.

Let TP be the true positive means number of patient who don't have heart disease (Healthy) which are predicted correctly; TN the true negative means number of patient with heart disease (Not healthy) which are predicted correctly; FP the false positive means number of normal which are predicted as patient and FN the false negative means number of patient which are predicted as normal. The performance metrics used in this study are defined in chapter II section II.6.4.

Before experiments, we firstly prepared our dataset through the imputation of missing values using KNN technique. We have implemented KNN with $K = 4$. Secondly, a re-scaling method have been applied to re-scale continuous features into the range $[-1, 1]$. Then, we have splitted randomly UCI heart disease dataset into training set 70% and the testing set 30%. For the diagnosis step, we started by applying traditional ELM architecture to compare its performance with ID-RELM algorithm, We then applied RF for feature selection.

Table III.2 shows the performance of the three methods on UCI heart disease dataset without retargeted technique (ELM), with retargeted technique (ID-RELM), and RF based feature selection combined to ID-RELM (ID-RELM & RF). The application of the simple ELM classifier gives a modest results with 0.81 accuracy, 0.89 sensitivity of healthy patient and 0.74 sensitivity of not healthy patient. It also generates too much hidden nodes for this performance (153 hidden nodes), which are not as good as for the real time processing. We can observe that the classifier performance has been improved after using ID-RELM. We notice a higher accuracy 0.88, sensitivity

	ELM	ID-RELM	ID-RELM & RF
Accuracy	0.81	0.88	0.94
Sensitivity Healthy	0.89	0.98	0.98
Sensitivity Not healthy	0.74	0.79	0.93
Specificity Healthy	0.78	0.82	0.93
Specificity Not healthy	0.88	0.97	0.98
F1-measure Healthy	0.83	0.89	0.95
F1-measure Not healthy	0.80	0.87	0.95
Number of hidden nodes	153	42	40

Table III.2: Experimental performance metrics.

and specificity. This is obtained for only 42 hidden nodes (see Table III.2). ID-RELM method seems to be a good technique to classify UCI heart disease dataset with lower testing processing time.

Figure III.3 shows the same results using ROC curve. One can notice an improvement when ID-RELM technique is applied (from AUC = 0.81 to AUC = 0.89). By combining RF features selection with ID-RELM, the highest classification performances are achieved as we can remark in Table III.2 (the accuracy increases to 0.94 and AUC is about 0.95).

We now come to compare the obtained diagnosis results from the adapted PHM tool using ID-RELM & RF to several methods recently proposed to contribute in developing the decision support system for heart disease diagnosis (see Table III.3). It can be observed that ID-RELM & RF improves the performance of CAD for heart disease (accuracy = 0.94). The proposed method do not only improve the accuracy of the system, but it also reduces the processing time (see Table III.4). This indicator is a very important factor for real time applications.

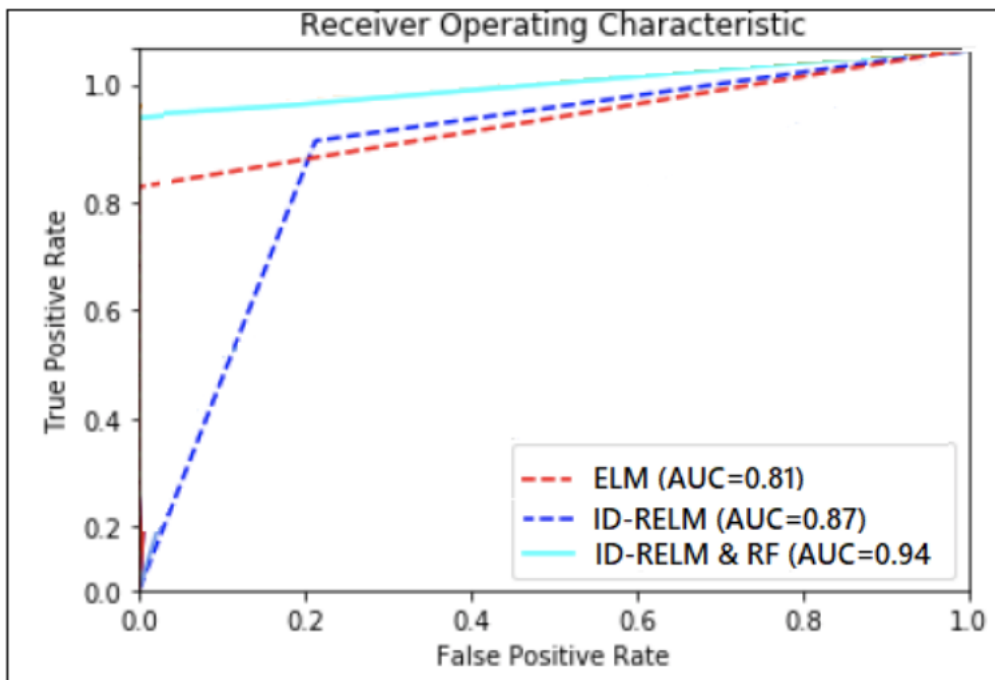


Figure III.3: ROC curve of our proposed method.

Previous work	Classifier	Feature selection	Accuracy
[66]	NB	SVM- RFE/10 F	0.84
	RF	SVM-RFE/8F	0.84
[67]	NB		0.83
	SVM		0.84
	SVM+MLP		0.84
	KNN (k=9)		0.83
	DT		0.77
	MLP		0.82
[68]	RF		0.91
[68]	Majority vote-based model (NB,DT,SVM)		0.82
[69]	FAMD+RF		0.93
[70]	Bagging with decision tree		0.81
[71]	NB		0.86
	ANN		0.85
	DT		0.89
[58]	Hard Voting Ensemble Method		0.90
[72]	LR	reflif	0.89
	NB	mRMR	0.84
	SVM	LASSO	0.88
Our approach	ID-RELM	RF	0.94

Table III.3: Performance comparison of our proposed method along with previous work on UCI heart disease Cleveland dataset.

Model used	Processing time (s)
Logistic regression	2.159
K-nearest neighbor	0.144
Artificial neural network	30.802
SVM (kernel=RBF)	60.589
SVM (kernel=linear)	0.179
Naive Bayes	1.596
Decision tree	1.831
Random Forest	2.220
ID-RELM & RF	0. 07

Table III.4: Processing time comparison of our proposed method and existed work.

4 Conclusion

The framework of this paper is to transfer the PHM approach from industrial to medical field. This work could be considered as a first step to reduce the gap between industry and medical filed, by exchanging the applied techniques, and proving that models applied for machine’s health diagnosis could be applicable for human’s health diagnosis. The suggested system accomplished higher classification accuracy rate, by improving the data quality, decreasing the number of attributes and obtained higher performance rate, with reduced processing time. The ID-RELM & RF model can be used as a medical decision support system for cardiologists to make accurate classification with lower time, cost, and effort.

In the following chapter, we will direct our work on dermatology domain. For this end, we have proposed a computer-aided diagnosis to classify a type of skin lesions in order to assist dermatologists in distinguishing between these challenging lesions called spitz nevus. This CAD system based on machine learning techniques and genetic algorithm-based feature selection.

Chapter IV

Computer Aided Diagnosis for Spitzoid lesions classification using Artificial Intelligence techniques

In this chapter, we present our second contribution which aims test several artificial intelligence techniques so as to build a computer aided diagnosis system. We present first an overview on spitzoid lesions in section 1 then a motivation of our contribution in section 2. Section 3 offering detailed description on the proposed method used in three phases: (a) the preprocessing phase; (b) the feature selection phase; and (c) the classification phase. Section 4 highlights the key indicators, including performance measure(s), accuracy, sensitivity, specificity, G-mean, F-measure, ROC curve, and area under the ROC curve (AUC) as well as overviews the experimental findings. Finally, Section 5 show results obtained and discussion.

1 Medical overview

Spitz nevus, a rare form of skin mole, tends to affect mostly young people and children with some 2016 statistics claiming that about 7 out of every 100,000 individuals may be inflicted [73]. Typically, patients diagnosed with Spitz nevus are under 21 years old [74]. Historically, such tumors had been treated as a melanoma, identifying

with the name, Benign juvenile melanoma; later on, Dr Sophie Spitz, a pathologist, characterized a new class of melanocytic tumor, which has now been popularized as Spitz nevus [75]. According to Harms, et al. [76], these Spitzoid melanocytic lesions may be clustered into three main types: (a) Spitz nevi; (b) Atypical Spitz Tumors; and (c) Spitzoid Melanomas (SM).

Figure IV.1 shows two dermoscopic images with Figure IV.1 A exhibiting Spitz nevus (SN), and Figure IV.1 B depicting Atypical Spitz tumor (AST). Although clinically indistinguishable, these lesions share some dermoscopic and histologic features (see Table IV.1). Arguably, the exact clinico-pathologic definition of AST is still incredibly challenging for dermatopathologists. However, the debate concerning AST prognosis is of highest priority, as their compartment cannot be easily predicted. SN displays a definite benign behavior, whereas SM is malignant and particularly aggressive [77]. Consequently, Spitzoid lesions, a subset of melanocytic skin lesions, are not only difficult to diagnose from a clinical viewpoint but from both histological and/or dermoscopic perspectives as well.

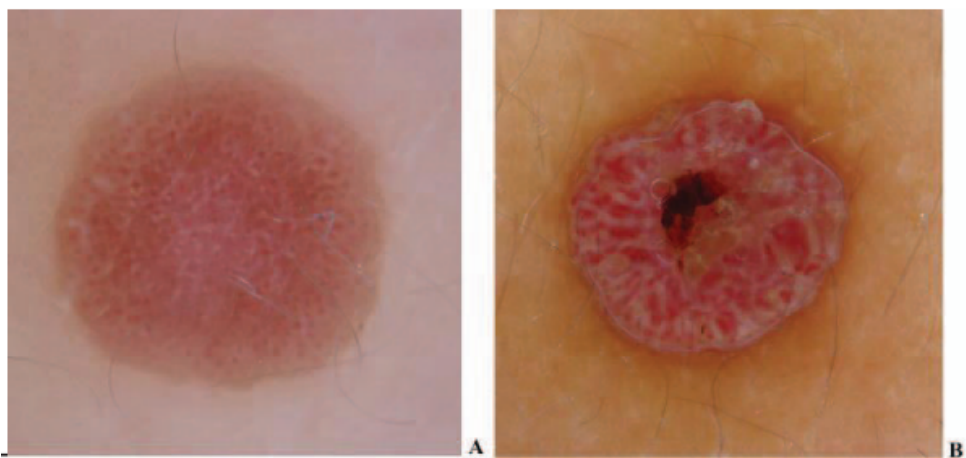


Figure IV.1: Example dermoscopic images of Spitz Nevus (1A), and Atypical Spitz Tumors (1B) Source: Rubegni et al. (2016)

Blum, et al. [78] argue that SN is diagnosed typically by dermatologists conducting visual inspections of a mole using clinical assessment tools such as ABCDE (Asymmetry, Border, Color, Diameter, and Evolution). Even so, a biopsy laboratory examination is often ordered to remove all or part of the mole to support the diagnosis. Indeed, a skilled and trained pathologist must be engaged to diagnose a sample, differentiating

it between SN v.a more severe melanoma.

Spitz nevus	Atypical Spitz tumor	Malignant Spitz tumor
Mean and median age 21 years (range 2-69 years) (< 40 years)	Can occur at any age, more common in younger patients	Can occur at any age (often > 40 years)e
Most commonly affects extremities	Occur in extremities, trunk	Occur in extremities, trunk, asymmetrical
Pink or reddish plaque, papule, or nodule	Plaque or nodule Color variegation Changing lesion	Enlarged Plaque or nodule Color variegation
< 5 to 6 mm	Often > 5 to 10 mm	> 5 mm, Often > 10 mm
Symmetrical	Symmetrical or Asymmetrical	Often Asymmetrical
Well circumscribed	Well or poorly circumscribed	Often poorly circumscribed
Epidermal hyperplasia	Ulceration possible	Ulceration
Vertically oriented nests with clefting	Irregular nesting	Irregular and confluent nesting
Central focal pagetoid spread	Increased cellularity	pagetoid spread may be extensive
Often wedge-shaped	Greater pagetoid spread than in SN	Ulceration
Maturation of dermal component	Deeper dermal than in SN Maturation may be partial or absent	Effacement of epidermis Lack of maturation
	26 dermal mitoses /mm ² Deep mitoses Possible necrosis	Often > 6 dermal mitoses /mm ² Deep / marginal or atypical mitoses Necrosis

Table IV.1: Spitz nevus (SN) VS Atypical Spitz tumors (ASTs) vS Spitz melanoma(SM) Source: Adapted from World Health Organization (2018).

2 Motivation

Considering the similarities of spitzoid lesions and the dependency on the dermatologist's skill level and/or pathologist to inform the diagnostic process, accurate diagnosis remains a problem. Data mining (DM) techniques have been successfully applied to

situations where such complexity exists, and the availability of advanced artificial intelligence (AI) techniques and data pre-processing techniques to build computer-aided diagnostic (CAD) system can be combined to provide effective solutions for the analysis of Spitzoid lesions.

In recent years, computer scientists have diverted attention to skin lesion analysis. A great majority of the proposed methodologies in the extant literature aim to develop a CAD to assist dermatopathologists in making an accurate diagnosis, thereby achieving a proper decision. Specifically, In Al-Masni et al. [79] suggest a segmentation method on dermoscopic images using full resolution convolutional networks (FrCN). They also argue that the proposed technique can generate full spatial resolution features for each pixel of the input dermoscopy images.

In contrast, a 3D skin lesion reconstruction technique using the estimated depth obtained from regular dermoscopic images, and the adaptive snake technique in the segmentation phase have been proposed by Satheesha, et al. [80]. Here, by fitting the depth map estimated to the underlying 2D surface, a 3D reconstruction can be achieved. This is then followed by a feature extraction (Color, texture and 2D shape) and feature selection to study decision-making features. Finally, AdaBoost and SVM classifiers can be applied in the classification phase.

In Jain, et al. [81], a CAD for the diagnosis of Melanoma Skin Cancer on dermoscopic Image Processing is presented. In Roffman, et al. [82], a multi-parameterized artificial neural network (ANN) using available personal health dermoscopic images for early detection of non-melanoma skin cancer with high sensitivity and specificity has been developed. Finally, in Xie, et al. [83], a novel method for the classification of melanocytic tumors as benign v. malignant using is proposed. Digital dermoscopy images have been advanced; precisely, in the feature extraction and reduction phase. The Principal Component Analysis (PCA) technique is used. In the classification phase, the ANN meta-ensemble model is applied by combining fuzzy NNs with Back Propagation NNs and evaluating the proposed method's performance using fuzzy NNs, RFs, Gentle Adaboost, k-NN, two SVM methods, and two systems using the Bag-of-Features (BoF) classification model.

Notably, most of the proposed methodologies in the CAD literature for differenti-

ating among skin lesions have been based chiefly on dermoscopic vision. It often fails to consider the clinical, genetic, molecular, and immunohistochemical information in making a holistic diagnosis. The primary goals and contributions for this work include:

- Develop an automatic diagnostic system for Spitzoid lesion classification to assist dermatologist during diagnosis process
- Specify the exact type of a Spitz lesion, which is extremely difficult and challenging, and to the best of our knowledge, no one has used AI to classify them before.
- Integrate clinical, histological, and immunohistochemical features to make an accurate diagnosis in distinguishing between SN v. AST and determine the impact of these features on the classification.

Broadly, this study evaluates various AI methods to classify Spitz lesions. Specific methods include Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Logistic Regression (LR), and Multi-Layer Perceptron (MLP), all of which have been commonly used in medical classification problems. Additionally, advanced pre-processing techniques and feature selection methods will be applied to improve the data quality and solve the imbalanced data problem, which will not only lead to a sizable improvement of the prediction time and classification accuracy but will also cleverly inform on the impact of histological and immunohistochemistry features on the classification.

3 Proposed method

Figure IV.2 shows the general schematic diagram of the proposed study technique. The details of each processing stage are now described in the subsequent sections.

3.1 Data description

A retrospective study of 54 Spitz lesions diagnosis from 2000 to 2018 has been conducted in the pathology department of Nord Franche Comte hospital (France). The cohort comprises 47 SN and 7 AST performed by five pathologists. The dataset contains 29 attributes computed from clinical, histological and immunohistochemical

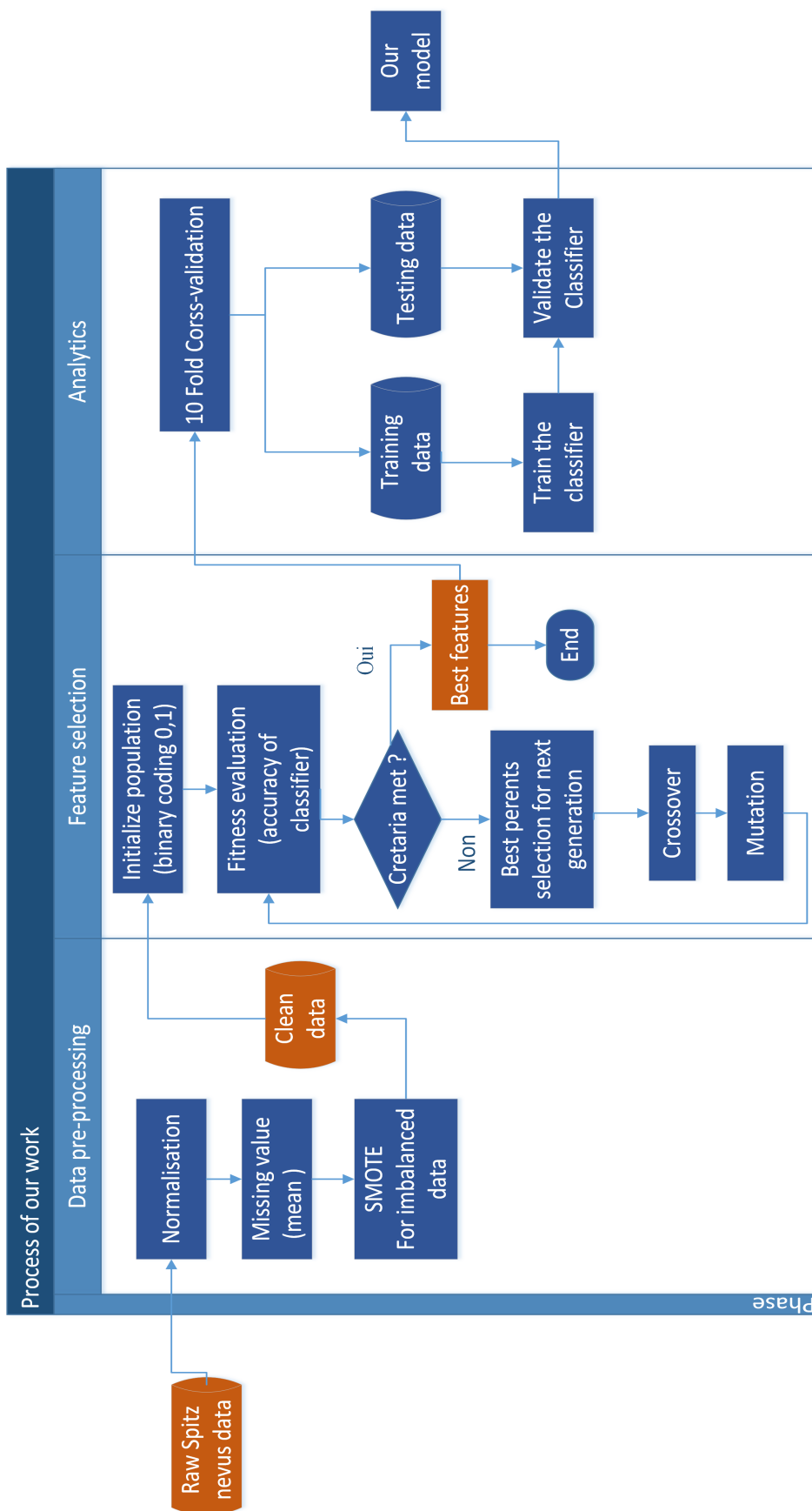


Figure IV.2: General schema of our proposed process

data, details of which are shown in Table IV.2.

Feature	Input type	Input range	Details
Gender	Binary	0 or 1	Man : 0 Women: 1
Localization	quinary	1 or 2 or 3 or 4 or 5	1:Trunk, 2:Lower extremity, 3:Upper extremities, 4:abdo5:Face and neck
Age	Continuous	From 2 to 54	Majority of them are under 20 years old
Format	Ternary	1or2 or 3	1: junctional, 2: wholly dermal.3: compound
Size of spitz	Continuous	From 0.3 to 1.4	only 5 patients more than 1 cm , the rest under 1 cm
Thickness	Continuous	From 0.1 to 6	Majority \leq 2,5 mm
Mitotic index		From 0 to 2.2	Majority \leq 0,5 per mm square
Cytonuclear Atypia	Binary	0 or 1	0: no 1: yes
deep mitosis	Binary	0 or 1	0: no 1: yes
Atypical Mitosis	Binary	0 or 1	0: no 1: yes
Infiltration of the hypodermis	Binary	0 or 1	0: no 1: yes
Asymmetry	Binary	0 or 1	0: no 1: yes
Blurred boundaries	Binary	0 or 1	0: no 1: yes
Pagetoid spread	Binary	0 or 1	0: no 1: yes
Density of lymphocytic infiltrate	Binary	0 or 1	0: no 1: yes
Hypercellularity	Binary	0 or 1	0: no 1: yes
Ulceration	Binary	0 or 1	0: no 1: yes
Kamino's body	Binary	0 or 1	0: no 1: yes
desmoplastic cells	Binary	0 or 1	0: no 1: yes
epidermal alteration	Binary	0 or 1	0: no 1: yes
grenz zone infiltration	Binary	0 or 1	0: no 1: yes
irregular nests	Binary	0 or 1	0: no 1: yes
lack of maturation	Binary	0 or 1	0: no 1: yes
P16			100% no loss
KI 67	Continuous	From 0 to 18	most of them \leq 5
BRAF	Binary	0 or 1	0: mute, 1: not mute
ALK IH	Binary	0 or 1	0: negatif, 1: positif
ALK Fish	Nul	Nul	Nul
Melanin pigmentation	quaternary	0 or 1 or 2 or 3	

Table IV.2: Spitz nevus dataset details.

3.2 Pre-Processing Phase

In order to achieve more accurate results, data pre-processing entails a critical step in transforming raw SN data into a clean and understandable format for analysis. The following sub-sections discuss techniques applied to improve the quality of our dataset.

Categorical data: First, the majority of features in our dataset is categorical (Table IV.2). As machine-learning models are based on mathematical equations, we would only use numbers in the equations, which will then be converted into numerical values.

Missing values: In the current work, we apply the Mean imputation, one of the most commonly used methods, by replacing the missing value with the total sample mean. Accordingly, this strategy is simple and easy to implement.

Imbalance data: The imbalance of medical data, as characterized by the non-uniformity of the class distribution among the classes, seriously affects the accuracy of medical diagnosis classification. Data imbalance exists widely in real-world datasets, especially those in the medical field. The study dataset is found to be highly unbalanced, comprising 47 cases of classical SN v. only 7 cases of ASTs.

To resolve this challenge, a widely implemented technique for dealing with highly unbalanced datasets is resampling. Resampling consists of eliminating samples from the majority class (under-sampling) and/or adding more examples from the minority class (over-sampling). The simplest implementation of over-sampling is to duplicate random samples from the minority class, affecting over-fitting. In under-sampling [48], the simplest technique is to randomly remove samples from the majority class, which can cause wastage of information.

SMOTE (Synthetic Minority Oversampling TEchnique) consists of synthesizing elements for the minority class, based on those that have already existed [49]. It works by randomly picking “k,” a point from the minority class, and computing the k-NNs for this point. Synthetic points are then added between the chosen point and its neighbors. Other techniques discussed in the extant literature include SVM SMOTE

[84], or borderline-SMOTE [85], where only the minority examples near the borderline are over-sampled. Adaptive synthetic sampling (AdaSyn), as presented in [86], they include both minority and majority classes in processing and adds extra synthetic samples to the minority class.

Scaling data: In this work, the data columns are rescaled to a range of [0-1] for two reasons: (a) Simplify the numerical computational complexities; (b) Get rid of attributes in the bigger numeric range while controlling attributes in the lesser [87].

3.3 The Feature Selection Phase

Feature selection is a critical step in the SN diagnosis process. As the study dataset typically consists of several features, a critical goal is to identify the most relevant features to the problem at hand. Other advantages of feature selection include cost reduction, increasing classification accuracy, decreasing the complexity of the model, and reducing the learning time [63].

With far too many attributes specific to the current study dataset, this feature selection process is clearly non-trivial. Indeed, identifying those attributes that are the most relevant to the classification is complicated. To this end, our strategy is to apply a mix of three feature selection methods: filter, wrapper and embedded methods. The filter methods measure the significance of identifiable features by their association with the dependent variable. In contrast, the embedded methods combine the qualities of filter and wrapper methods as implemented by algorithms that have their own built-in feature selection methods. Finally, the wrapper methods measure the effectiveness of a subset of features by actually training a model on the two differing wrapper types: deterministic v. randomize. Herein, we apply the randomize wrapper method via the genetic algorithm, which is discussed next.

Genetic Algorithm(GA): To date, GAs have gained increasing popularity. Characterized by a heuristic and general adaptive optimization search methodology, these algorithms are inspired by Darwin's theory of evolution. Initially presented by Bledsoe [88], and mathematically formalized by Holland [89]. These GAs operate with diverse populations, with the dominant solution frequently achieved only after a sequence of

iterative steps. These GAs also develop sequential populations of periodic solutions presented by a chromosome until adequate results have been reached [87].

A predefined fitness function evaluates these chromosomes. Two major operators, which impact on the fitness value, are the crossover and mutation functions. For the next generation, chromosomes that obtained the higher fitness value will have the corresponding higher probability to be selected using either the roulette wheel or the tournament strategy [90]. In mutation step, genes may be changed randomly by pressing the probability. The parameter settings for the GA applied herein as feature selection are presented in Table IV.3.

Parameter	Value
Population size	100
Number of generation	50
Rate of crossover	0.8
Rate of mutation	0.1
Fitness evaluation	Accuracy of classifier
Size of chromosome	27
Coding	Binary 0: not selected 1: selected

Table IV.3: Parameter settings of our genetic algorithm based feature selection.

Different individual entities are assigned randomly in the initial population stage, with binary coding where 1 presents the selected feature and 0 not selected. All individual entities have a unique size (27 genes in each chromosome). The chromosomes characterizing the population represent a set of probable optimal features. At each generation, each potential solution's fitness value is derived from using a tenfold cross-validation method to calculate the accuracy of classifier and then intelligently applied to select the population for the next generation by roulette wheel selection method. To stop the solution set falling into a local optimal, crossover and mutation are used to generate populations that represented new sets of solutions. The basic process of the applied GA may be summarized as follows:

1. Initial population: The initial population size is 100 - several different numbers of generations in the experiment are tried and tested, before deciding to use 50 generations, which yields the highest accuracy as depicted in Figure 3.
2. Evaluation: Each population's fitness value determines if the population will

survive in future generations. Herein, the accuracy of the classifier serves as the fitness function.

3. Selection: The population with the better fitness value has a greater probability to be selected to the next generation; herein, a roulette wheel mechanism is deployed to choose the population sets for the next generation.
4. Crossover: Crossover is the process of generating a new individual entity from two parents by exchanging and reordering their parts. By crossing, the search power of the GA is dramatically increased. Crossover in the study is implemented using a single-point crossover operator is chosen with a rate of 0.8.
5. Mutation: Mutation is the process of changing some gene values of individual sequences to increase the population variety; herein, the mutation with a rate of 0.1 is applied.

3.4 The Classification Phase

The various methods applied for evaluation in classification phase are briefly highlighted at this point. These include:

Support Vector Machine (SVM): SVM is based on statistical learning theory and the structural risk minimization principle, and it has been used for classification and regression [91]. The main SVM concept applied here is to map the input data from the N-dimensional input space, through some non-linear mapping. Then, to classify our data, we should determine the optimal hyperplane that maximizes class boundaries' margins.

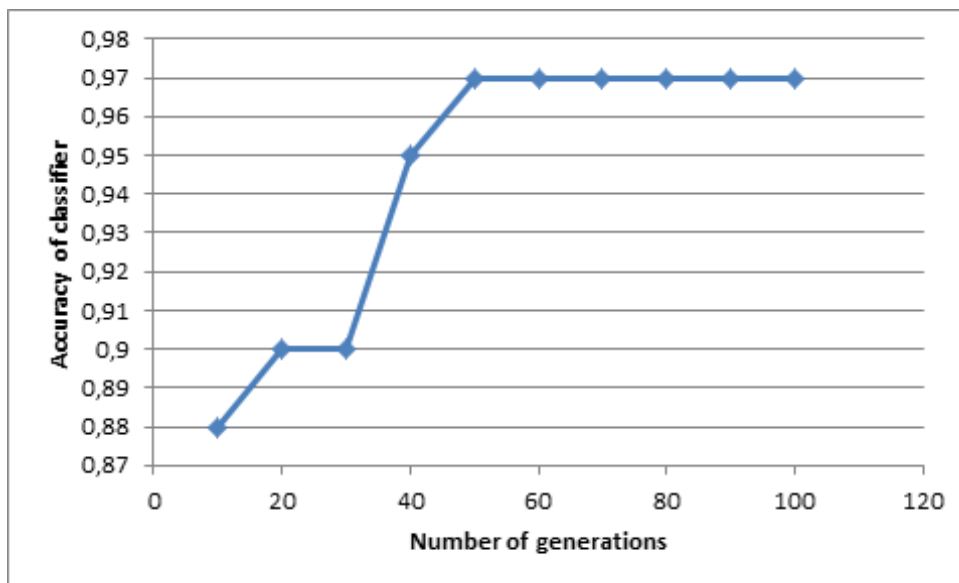


Figure IV.3: The impact of number of generations on accuracy of classifier

Decision Tree (DT): DT a popular and the most powerful supervised learning methods for classification where each internal node signifies a check on an attribute is applied herein. Each branch of the DT represents a result of the check, and each leaf node contains a class label. A DT algorithm is implemented by generating a DT with terminal nodes as the class label (Classical Spitz Nevus, Atypical Spitz Tumors). Additionally, sets of if-then conditions are employed to classify novel samples.

Logistic Regression (LR): The LR model originates as a result of modeling the posterior probability of K classes via linear functions in x while ensuring that the probabilities sum to one and remain in the range $[0, 1]$. The denominator selection is random in that the estimates are equally distributed under this choice [87]. When $K = 2$, as would be in our case (SN, AST), the model is straightforward as there is just a single linear function.

Naïve Bayes (NB): Bayesian Network describing sets of local conditional probabilities together with a set of conditional independence assumptions is applied herein to clarify the joint probability distribution for a set of variables. In the NB network, each node shows variable in the joint space; two types of information are detailed for

all variables. First, the variable is independent of its non-descendants, given its instant predecessors in the network. Second, a conditional probability table is given for each variable, indicating the probability distribution of this variable assuming the values of its immediate antecedents [92].

K-Nearest Neighbor (KNN): The basic concept of kNN is to compute the minimum distance between the stored feature vectors and the new feature vectors. Firstly, we compute the distances between all samples that have already been classified into clusters; then, we find the k samples with the smallest distance values; and finally, we approve the new data. A new sample will be classified into the largest cluster among the selected k samples [93]. We tried the values of k from 1 to 10 and found that k = 3 offers the best results with this classifier as illustrated in Figure IV.4.

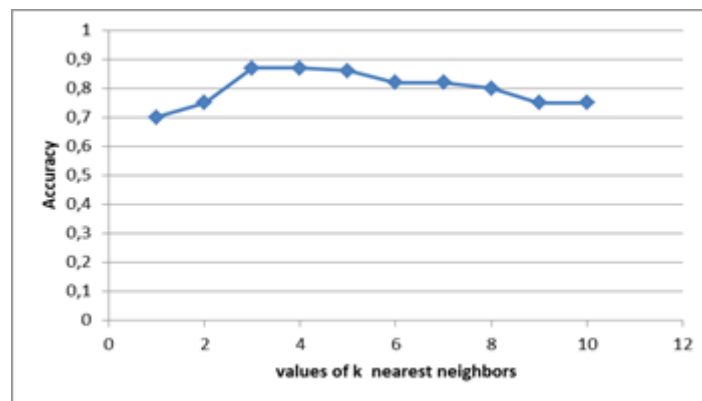


Figure IV.4: Change of Accuracy in terms of nearest neighbor's k value

Multilayer Perceptron (MLP): The MLP classifier applied herein has a three-layer structure. The input layer's size is equal to the number of the selected features ($1 < N < 27$). In contrast, the output layer contains one node for a possibility of only two classes to be classified (SN v. AST). Additionally, having selected and trained several potential combinations of the selected number of neurons in the hidden layer, we found the optimized number to be 50. We also added an activation function to make our MLP flexible vis-a-vis the non-linear decision boundaries' learning. Several kinds of activation function are discussed in the extant literature; herein, we used the Rectified Linear Units (ReLU), which applies the non-saturating activation function f

$(x) = \max(0, x)$ [55]. The function returns 0 if it receives any negative input, but for any positive value x , it returns that value back. Thus it gives an output that has a range from 0 to infinity.

Random Forest (RF): RF is defined as “combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest” [94]. Herein, when RF is used to perform the classification task, a class vote from each tree is generated, and the majority vote is then used to perform the classification task.

4 Performance metrics and experimentation

In this section, the various performance metrics applicable for evaluating and interpreting multiple experimental results are first highlighted prior to discussing the study findings and their interpretations. Notably, we conduct the experiments in the python language environment, and when no parameter values are given, the default values of these functions will apply.

4.1 Performance Metrics

A ten fold cross-validation scheme is performed to evaluate and compare the performance of all of the aforementioned classification methods being applied. One of the major issues when dealing with unbalanced datasets relates to the metrics used to evaluate the model’s performance, for example, using simple metrics like accuracy score alone can be relatively misleading. Accordingly, a range of different performance metrics is adopted for studying and comparing the various classification models differentiating SN v. AST samples. These metrics entail accuracy, sensitivity, specificity, precision, F-measure, G-mean, ROC and AUC with measures based on the correct and wrong prediction of the classifier. For the respective metrics, the below formulae are computed with:

- TP= True positive means number of SN which are predicted as SN;
- TN= True negative means number of AST which are predicted as AST;
- FP= False positive means number of SN which are predicted as AST;

- FN= False Negative means number of AST which are predicted as SN.

4.2 Data Sampling Results

In the first test, four different over-sampling methods have been applied with the unbalanced dataset so that their performance may be appropriately compared. Table IV.4 details the performance of the different machine learning (ML) classifiers on our dataset with and without oversampling methods.

All classifier's accuracy is high in the case of classifiers without oversampling methods, that is, between 0.72 - 0.87. Thus, each classifier's performance on other performance measures has to be investigated beyond just accuracy. Among the other measures, the sensitivity, specificity, and F-measure show a significant difference between SN (majority - very high) v. AST (minority - very low) classes. Especially with LR, KNN, MLP, and SVM the sensitivity and specificity of AST class is 0.00, which means these classifiers over-fits and the model predicts all cases as SN.

As shown in Figure IV.5, DT and RF get higher AUC scores than the other classifiers.

		Accuracy	Sensitivity		Specificity		F1-measure		G-mean
			AST	SN	AST	SN	AST	SN	
Without Over- sampling methods	DT	0.83	0.43	0.87	0.33	0.91	0.38	0.89	0.65
	RF	0.85	0.14	0.96	0.33	0.88	0.20	0.92	0.85
	SVM	0.87	0.00	1.00	0.00	0.87	0.00	0.93	0.50
	NB	0.72	0.43	0.77	0.21	0.90	0.29	0.83	0.59
	LR	0.87	0.00	1.00	0.00	0.87	0.00	0.93	0.50
	KNN	0.87	0.00	1.00	0.00	0.87	0.00	0.87	0.50
	MLP	0.85	0.00	0.98	0.00	0.87	0.00	0.92	0.48
SMOTE k=6	DT	0.95	0.98	0.94	0.94	0.98	0.96	0.96	0.95
	RF	0.97	1.00	0.96	0.96	1.00	0.98	0.98	0.97
	SVM	0.94	0.98	0.91	0.92	0.98	0.95	0.95	0.94
	NB	0.90	1.00	0.81	0.84	1.00	0.91	0.89	0.90
	LR	0.93	1.00	0.87	0.89	1.00	0.94	0.93	0.93
	KNN	0.94	1.00	0.89	0.90	1.00	0.95	0.94	0.94
	MLP	0.98	1.00	0.98	0.98	1.00	0.95	0.99	0.98
Borderline SMOTE	DT	0.95	0.98	0.94	0.94	0.98	0.96	0.96	0.95
	RF	0.97	1.00	0.96	0.96	1.00	0.98	0.98	0.97
	SVM	0.94	0.98	0.91	0.92	0.98	0.95	0.95	0.94
	NB	0.90	1.00	0.81	0.84	1.00	0.91	0.89	0.90
	LR	0.93	1.00	0.87	0.89	1.00	0.94	0.93	0.93
	KNN	0.94	1.00	0.89	0.90	1.00	0.95	0.94	0.94
	MLP	0.98	1.00	0.98	0.98	1.00	0.95	0.99	0.98
ADASYN	DT	0.95	0.98	0.94	0.94	0.98	0.96	0.96	0.95
	RF	0.97	1.00	0.96	0.96	1.00	0.98	0.98	0.97
	SVM	0.94	0.98	0.91	0.92	0.98	0.95	0.95	0.94
	NB	0.90	1.00	0.81	0.84	1.00	0.91	0.89	0.90
	LR	0.93	1.00	0.87	0.89	1.00	0.94	0.93	0.93
	KNN	0.94	1.00	0.89	0.90	1.00	0.95	0.94	0.94
	MLP	0.98	1.00	0.98	0.98	1.00	0.95	0.99	0.98
SVM- SMOTE	DT	0.95	0.98	0.94	0.94	0.98	0.96	0.96	0.95
	RF	0.97	1.00	0.96	0.96	1.00	0.98	0.98	0.97
	SVM	0.94	0.98	0.91	0.92	0.98	0.95	0.95	0.94
	NB	0.90	1.00	0.81	0.84	1.00	0.91	0.89	0.90
	LR	0.93	1.00	0.87	0.89	1.00	0.94	0.93	0.93
	KNN	0.94	1.00	0.89	0.90	1.00	0.95	0.94	0.94
	MLP	0.98	1.00	0.98	0.98	1.00	0.95	0.99	0.98

Table IV.4: Experimental performance on our dataset without /with existing over-sampling methods.

To verify the efficiency of the SMOTE method in handling the problem of the unbalanced dataset in the study, we have applied other existing methods as summarized in Table IV.4 to compare their performance. The input that needed to be determined

in SMOTE method is the number of nearest neighbors “k”. We tried several different k values in the experiment, finally deciding on using k = 6, which yields the best accuracy (see Figure IV.6).

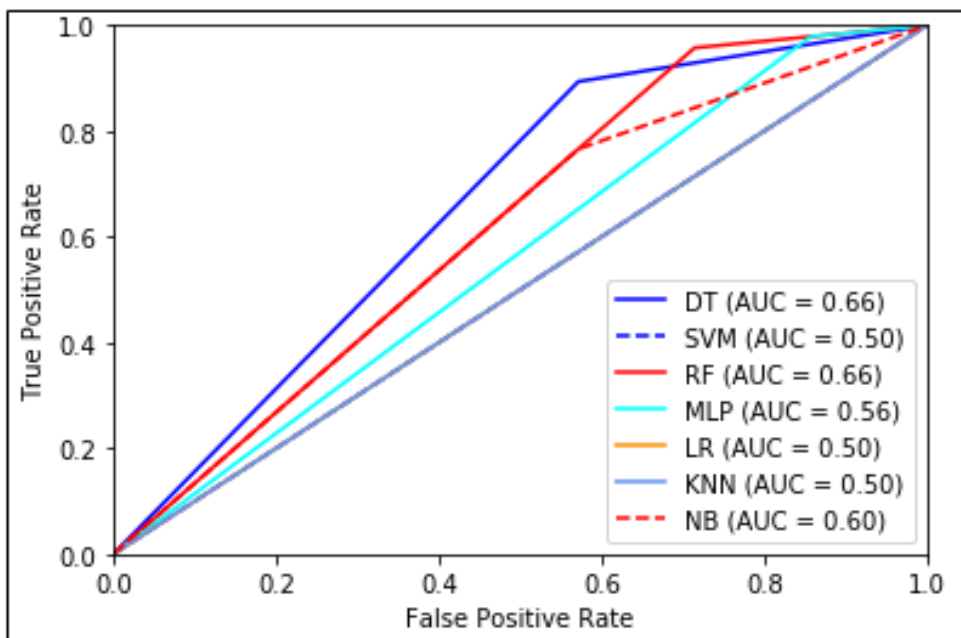


Figure IV.5: ROC curve of classifiers without oversampling.

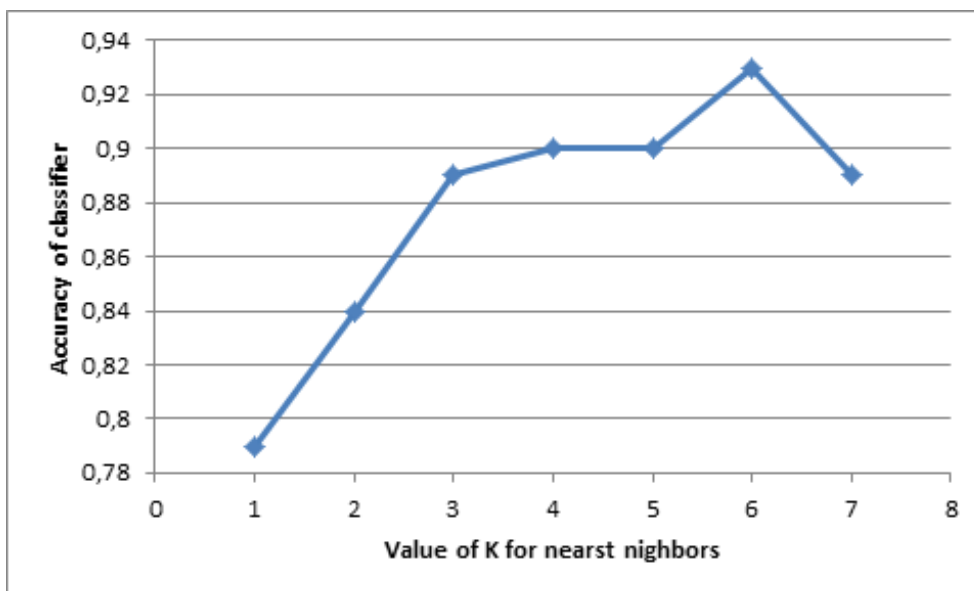


Figure IV.6: Impact of several different SMOTE’s k values on the accuracy

We now summarize performance of the four oversampling methods: SMOTE, Bor-

derlineSMOTE, ADASYN, SVMSMOTE. The results are relatively similar, where we see a balance in sensitivity, specificity, and F-measure of both SN and ASN class. Even so, SMOTE gives the highest accuracy 0.95 and G-mean 0.95 among all oversampling methods with random forest classifier. Figure IV.7 depicts the distribution of our data after applying SMOTE.

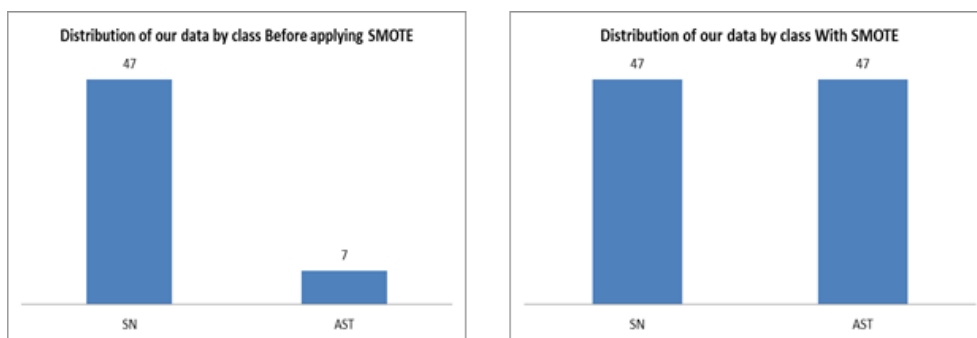


Figure IV.7: Distribution of our data with/ without SMOTE technique

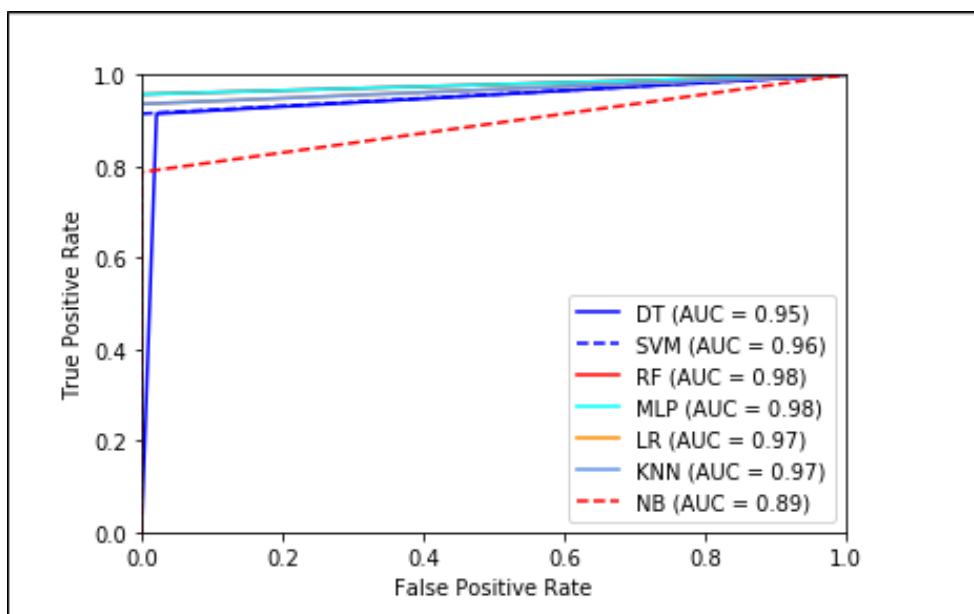


Figure IV.8: ROC curve of classifiers with SMOTE technique

4.3 Feature Selection Results

In the second test, we first used GA-based feature selection to select the best attributes; then, we used the same ML classifiers as in the first test. Table IV.5 shows the performance of the different ML classifiers on our dataset with and without

GA-based feature selection. Experimental results show that the highest classification performances are achieved when GA is used as feature selection with all classifiers. Notwithstanding, the MPL classifier does a good performance in predicting the AST instances correctly.

		Accuracy	Sensitivity		Specificity		F1-measure		G-mean
			AST	SN	AST	SN	AST	SN	
Smote Without GA	DT	0.94	0.94	0.85	0.82	0.95	0.88	0.90	0.89
	RF	0.95	0.94	0.94	0.98	0.98	0.96	0.96	0.95
	SVM	0.87	0.85	0.89	0.89	0.86	0.87	0.88	0.87
	NB	0.88	1.00	0.77	0.81	1.00	0.90	0.87	0.88
	LR	0.94	1.00	0.89	0.90	1.00	0.95	0.94	0.94
	KNN	0.70	0.87	0.81	0.65	0.53	0.75	0.64	0.70
	MPL	0.93	1.00	0.87	0.89	1.00	0.94	0.93	0.93
Smote With GA	DT	0.95	0.98	0.94	0.94	0.98	0.96	0.96	0.95
	RF	0.97	1.00	0.96	0.96	1.00	0.98	0.98	0.97
	SVM	0.94	0.98	0.91	0.92	0.98	0.95	0.95	0.94
	NB	0.90	1.00	0.81	0.84	1.00	0.91	0.89	0.90
	LR	0.93	1.00	0.87	0.89	1.00	0.94	0.93	0.93
	KNN	0.94	1.00	0.89	0.90	1.00	0.95	0.94	0.94
	MPL	0.98	1.00	0.98	0.98	1.00	0.95	0.99	0.98

Table IV.5: Experimental performance on our data without / with genetic algorithm based feature selection.

As shown in Table IV.5, highest accuracy of 0.98, F-measure of 0.99, and G-mean of 0.98 with 14 selected features are attained with the MPL. Here, AUC = 98 as shown in Figure IV.9. Next is RF method with 16 selected features, an accuracy of 0.97, F-measure of 0.98, AUC of 0.98, and G-mean of 0.97. Then, DT comes with accuracy of 0.95, F-measure of 0.94, AUC = 0.95, and G-mean of 0.95. Lastly, LR and NB provide the lowest accuracy of 0.93 and 0.90, respectively.

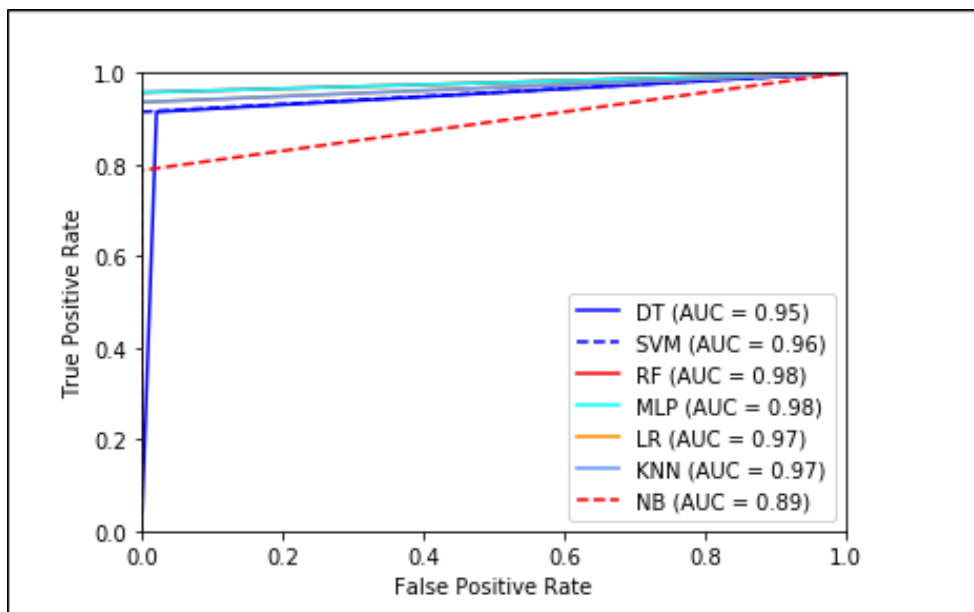


Figure IV.9: ROC curve of classifiers with GA based feature selection

Figure IV.10 and Table IV.6 highlights the most selected features with existing classifiers combined with GA as a feature selection method. Overall, for clinical features (colored in red), it is noted that localization is the most important as evidenced by its selection via five classifiers. Gender comes next, whereas only two classifiers have selected “age”. For histology features (colored in blue), it is clear that Cytonuclear Atypia is a most significant feature as selected by all classifiers. Finally, among immunohistochemistry features (colored in green), ki67 marker is the most significant.

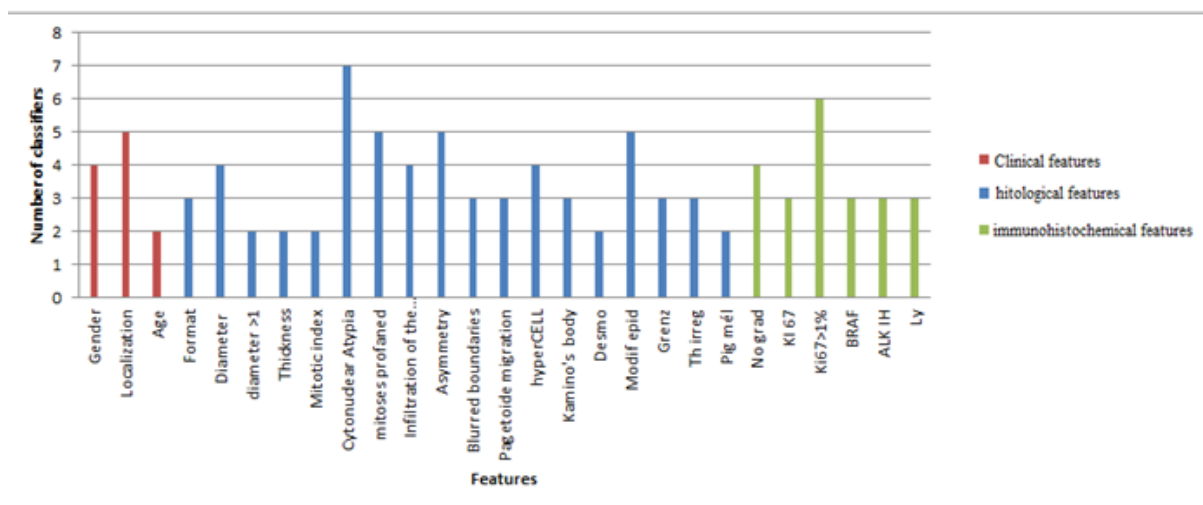


Figure IV.10: Most selected features with existing classifiers as fitness evaluation

	GA-DT	GA-SVM	GA-LR	GA-NB	GA-KNN	GA-MLP	GA-RF
Gender	×	✓	✓	✓	×	×	✓
Localization	✓	✓	×	✓	×	✓	✓
Age	×	×	×	✓	×	×	✓
Format	×	✓	✓	✓	×	×	×
Diameter	✓	×	✓	✓	×	×	✓
diameter > 1	×	×	✓	✓	×	×	×
Thickness	×	✓	×	×	✓	×	×
Mitotic index	×	×	×	✓	×	×	✓
Cytonuclear Atypia	✓	✓	✓	✓	✓	✓	✓
mitoses profaned	×	✓	✓	✓	✓	✓	×
Infiltration of the hypodermis	×	✓	×	×	✓	✓	✓
Asymmetry	×	✓	✓	×	✓	✓	✓
Blurred boundaries	✓	×	✓	×	✓	×	×
Pagetoide migration	×	×	✓	×	×	✓	✓
hyperCELL	✓	×	✓	✓	×	×	✓
Kamino's body	×	✓	✓	✓	×	×	×
Desmo	×	×	✓	×	×	✓	×
Modif epid	×	✓	✓	✓	×	✓	✓
Grenz	×	×	×	✓	✓	✓	×
Th irreg	×	×	✓	✓	×	✓	×
No grad	×	✓	×	×	✓	✓	✓
KI 67	✓	×	×	✓	×	×	✓
Ki67 > 1%	✓	✓	✓	✓	×	✓	✓
BRAF	×	✓	✓	×	×	✓	×
ALK IH	×	✓	✓	×	×	×	
Ly	✓	×	×	✓	×	✓	×
Pig mél	×	✓	×	×	×	×	✓

Table IV.6: Feature selection results obtained by different classifiers as fitness evaluation.

Figure IV.11 shows the performance of the different ML classifiers on our dataset with the number of selected features by GA. Experimental results show that when we used Naïve Bayes (NB) as the classifier with GA, it gave us the highest number of selected feature (17), and lowest accuracy (0.90), which means NB is the worst classifier applied. In contrast, DT offers the lowest number of selected features (8) with higher

accuracy (0.95). Notwithstanding, MLP provides the best classification accuracy (0.98) with 14 features; for this reason, we have chosen MLP as fitness evaluation and the classifier of choice in our model.

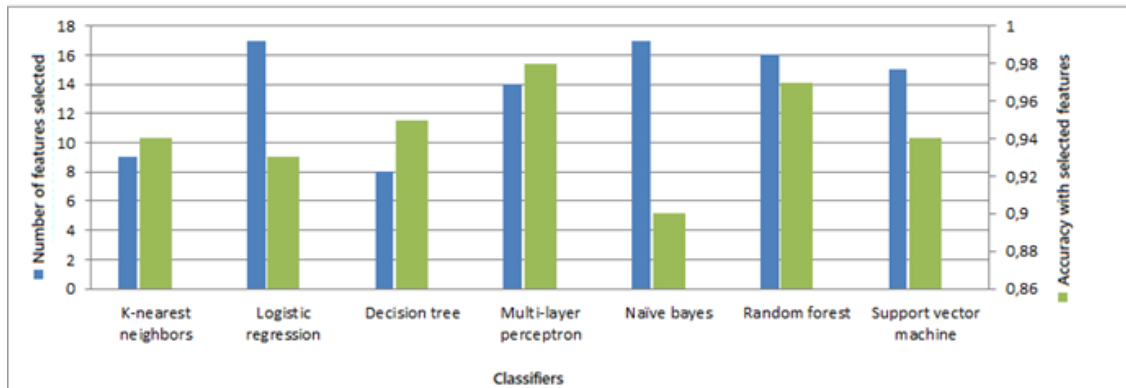


Figure IV.11: Number of selected features by GA and accuracy of classifiers

5 Discussion

The motivation for this research is to explore ways to improve the classification performance of different ML algorithms to identify Spitz nevus lesions accurately. The research presented here goes about the exploration by first removing class-imbalance in a real-world dataset and refitting then for analysis via the various methods of best features selection. Different class-imbalance techniques consisting of SMOTE, Borderline SMOTE, ADASYN, and SVM SMOTE were applied. All four class-imbalance techniques have been found to improve the classification results. This is consistent with previous research findings as provided in the cumulated literature [95, 85, 20]. However, in our case, the SMOTE method outperformed the other techniques. It uses k-nearest neighbors and generates excellent prediction results. Furthermore, the features, which are most appropriate for Spitzoid lesions classification, must be utilized as the inputs of the model. For this reason, it is found that GA has improved the performance for classifying the Spitzoid lesions data to achieve an accurate diagnosis.

Implications of the Findings As noted, we evaluated the scaled hospital dataset with seven (7) types of classifiers: K-NN, LR, DT, MLP, NB, RF and SVM. Where GA was applied, MLP achieves the highest accuracy at 0.98. Three key observations and implications may be drawn from the results: (1) GA can correctly rank

significant attributes since selected GA performs well in terms of classification performance; (2) SMOTE over-sampling method can correctly solve the problem of our imbalanced data and overcome the bias towards majority classes; and (3) MLP outperformed all other linear and nonlinear classification methods with respect to accuracy performance indicators. Therefore, the model proposed for similar type SN v. AST analysis will be a model where SMOTE is used for solving class imbalance problem, GA is used for feature selection phase and MLP to be applied for the classification phase. The proposed SMOTE-GA-MLP structure has previously been provided in Figure 2.

In summary, the suggested system accomplished higher classification accuracy rate, by improving the data quality, solving class-imbalance problem, decreasing the number of attributes and obtaining higher performance rate, while identifying the most critical features that can influence the classification. Results obtained in this study prove that the SMOTE-GA-MLP CAD system is valuable in aiding the dermatologists to identify ASTs, and to make the correct diagnosis. Accordingly, extending beyond this work may demonstrate a huge capacity in the area of medical decisions making in skin lesion analysis.

Study Limitations & Future Works This study has several limitations, mostly related to the data. First, the rarity of this disease and the lack of data in the hospitals about this type of lesions especially for the ASN case were an obstacle for us to collect enough data. We could use automatic data generation methods but we preferred to use only the real data to get realistic results. In our future work we aim to collect more real data and apply data generation methods to create new data from our real sampled data, and then we will compare results. Second, the nature of real-world skin lesion hospital datasets is not only imbalanced, but also heterogeneous and contains a lot of missing values and errors which can affect the analysis results.

In our future work we will concentrate on improving the data quality by applying various existing techniques in literature for data preprocessing and find out the one more suitable to our data. Also we aim to add the third class of Spitzoid lesions which is: Spitz melanomas. It is a malignant melanoma that is histologically similar to a benign skin lesion which makes the classification more challenging. Concerning the an-

alytical side, our future work will involve comparing and integrating the very promising approaches for classification, such as the ensemble techniques, by integrating multiple simple classifiers based on bagging, boosting, and stacking methods to improve the classification accuracy of Spitzoid lesions.

6 Conclusion

This chapter presented our first contribution that attempts to analyze Spitzoid lesions related to clinical, histological, and immunohistochemical features using AI techniques. Seven (7) classifiers: K-NN, LR, DT, MLP, NB, RF, and SVM have been applied for the analytic procedures. The hybrid technique of SMOTE-GA-MLP yields the highest performance overall.

The aided value of this research in the area of skin lesion classification is now summarized. First, it specifies the exact type of Spitz lesion, which is extremely difficult and challenging in real life. Second, it combines previous works on the steps needed to develop an automatic CAD system for Spitzoid lesion classification to assist dermatopathologists during the diagnosis process. Third, our work makes a classification based on various testes and types of data: clinical, histological, and immunohistochemical data. Contrary to previous literature work that only concentrates on the microscopic vision which cannot accurately classify them. Finally, the analysis for differentiating major classes of these lesions, namely SN (Spitz nevus) v. AST, is based on several features, including the immunohistochemical markers. Specifically, the findings indicate that localization of lesions, cytonuclear atypia, and Ki67 proliferative index are the most weighted features to differentiate AST from SN.

Based on the limitations of this chapter, we will propose in the next chapter a new CAD technique for skin lesion that eliminates the intervention of the dermatologist to extract the features from dermoscopic image. We will propose CAD technique that classifies skin lesions directly from dermoscopic images. The evaluation of the model will be done using public and private datasets.

Chapter V

Toward efficient Automatic Hyperparameters selection using Big Data tools to improve Skin Lesions classification

This chapter presents the last and the main contribution, which aims to develop a computer-aided diagnosis (CAD) that can classify accurately different skin lesions using dermoscopic images and metadata. In this research. We present first an overview on skin lesion classification and its challenges in section 1. Then, section 2 analyzes previous work and presents our contribution. The proposed method is presented in section3 involves data preprocessing, data augmentation, data classification, and CNN-AHPS technique for the training step. Section 4 shows experimental results and a comparative analysis with state of the art. Section 5 discusses our observations, findings, and some limitations.

1 Motivation

One of the most widespread types of cancer is skin cancer, with 5 million cases reported each year, and more than two persons die of skin cancer every hour in the

United States [96]. Melanoma is the most serious form and causes most of skin cancer deaths[9]. Since 2018, 178,560 new cases of melanoma are recorded in the US involving 87,290 cases of noninvasive and 91,270 invasive [97]. The mortality rate of this disease is expected to rise in the next decade, especially if diagnosed in later stages. However, if the skin cancer is diagnosed at primary stages, the survival rate is approximately 97% [98].

Traditional ways to diagnose skin cancer by dermatologists habitually follow three main steps: the first one is the observation of suspected lesion by the naked eye, then dermoscopy which is an imaging modality that shows more details. Finally, the biopsy step to extract histological characteristics [99]. The limits of this process is would consume time and the patient may advance to later stages. Furthermore, accurate diagnosis is depending on the expertise of the dermatologist, and the availability of skilled dermatologists is limited in public healthcare [100]. In order to solve some of these problems, there are many research solutions by developing computer-aided diagnosis(CAD) system based on several approaches such as: detection, segmentation, and classification using machine learning and image processing. These techniques could potentially help dermatologists and diagnose skin cancer accurately at the earliest stage, without the need for an invasive biopsy [101].

Deep learning (DL) algorithms have shown great performance on image classification and outperformed humans in many applications [9]. However, the application of DL techniques in medicine is still challenging and requires a large training dataset. Various Convolutional Neural Network (CNN) architectures are applied in skin cancer classification literature such as: DenseNet [102], ResNet [103], MobileNet [104], GoogleNet[105], VGG19 and AlexNet [101] etc. To solve the lack of dataset problem, the majority of cited work have used transfer learning (TL). The Principle of TL is to take a model trained on a certain source task and reuses it for a targeted task.

Since 2012 many CNN architectures are proposed for image classification in the ImageNet challenge dataset[106]. As models become more complex, the performance has increased. However, the majority are weak in terms of computing load. EfficientNet model[107], proved its effectiveness with 66 million parameters achieved 84.4% accuracy in the ImageNet classification problem. The idea behind EfficientNet is scaling width,

depth, and resolution while scaling down the model equivalently. On the other hand, EfficientNet and previous works on supervised learning need billions of labeled data to enhance ImageNet models. Noisy Student method [108] confirmed the potential to use unlabeled images to improve both robustness and accuracy of previous ImageNet models. They concluded that self-training is an effective and simple algorithm to benefit from unlabeled data on a large scale. Their experiments showed that self-training with EfficientNet and noisy Student achieved an accuracy of 88.4%, which is 2.9% greater than without Noisy Student. These results motivate us to apply the combination of these robust techniques to develop accurate CAD system to classify skin lesions.

2 Related work

Automatic classification of different skin lesions from dermoscopic images is an actual challenging task due to: (1) high similarity in visual features among various lesions types in terms of size, shape, texture, and color. (2) Existence of artifacts in dermoscopic images [109]. (3) Lack of data (4) class imbalance problem and much variety in dermoscopic image resolutions.

CNN architectures have recently been introduced to address these challenging dermoscopy image analysis problems. As shown in TableV.1, many contributions are proposed in this area using various architectures and datasets. In this section, we are going to analyze them by focusing on three main points that can improve CAD performance.

2.1 Architecture selection

Various CNN architectures are applied to truckle dermoscopic images classification, and each time researchers try to make proposed architectures deeper for better capability to classify these challenging lesions such as: Inception v3 [9, 103], DenseNet 201 [103], GoogleNet [105], MobileNet [104], EfficientNet [110], etc (see Table V.1). Train a whole CNN architecture from scratch takes time and needs a huge dataset. So, this problem can be fixed by using the power of transfer learning (TL) with fine-tuning pretrained models. The majority of cited works have applied pertained on ImageNet

challenge using transfer learning to solve the lack of skin data problem.

2.2 Data preparation

As shown in Table V.1, many public available datasets for skin lesion classification are used : ISIC 2016 [111], ISIC 2017 [111], HAM10000 [112], ISIC 2018 [113], ISIC 2019[11]. The largest one is ISIC 2019 contains more than 25,0000 images and clinical features for eight classes of skin lesions. ISIC2019 dataset contain many quality problems that should be addressed : 1. Missing data in metadata 2. Imbalance class problem 3. Multi resolutions for images: this is due to the variety of sources (HAM1000, BCN 20000 MSK dataset).

2.3 Model's hyperparameter optimization

In previous works cited in table V.1, hyperparameters selection for CNN architectures are based on trial-and-error approach or using sequential ways (grid search, random search, Bayesian optimization). It is typically used to search through a subset of a learning algorithm's hyperparameters. They are a simple tool for optimizing machine learning algorithms' efficiency. As DL architectures become more complex and datasets been larger, the training phase became more expensive and takes days or even weeks to train a model [114].

To evaluate each hyperparameter combination, we must train the model. We can imagine the difficulty in this context, and we can not wait for years to find a suitable configuration of hyperparameters. For this reason, the selection of sequential hyperparameters does not work in our case. We aim to design an automatic hyperparameter selection that can provide a parallel execution. This aim looks unimportant because random search and grid search could provide a parallel execution too. However, they often stack in the application, and they are limited in the context of hyperparameters choice, and the number of parallel resources. Therefore, we aim to develop our algorithm compatible with our needs and our CNN architecture via big data tools (MapReduce). Our contribution is clearer after this synthesis and we may summarise it as follows:

- We will fine-tune pretrained Noisy student(EfficientNet-L2) architecture that achieved top 1 accuracy in ImageNet challenge in our CAD as a feature extractor.
- We will improve data quality by solving its problems (missing data, class imbalance, image multi resolutions) and add metadata to our model to study its impact on our model
- We will develop the an Automatic Hyperparameters Selection (CNN-AHPS) implementation on Apache Hadoop. This latter is more aimed towards data locality, fault tolerance, commodity hardware, and simple programming with a strong link to Python.
- We will evaluate our work with results obtained in the last challenge for skin lesion classification ISIC2019.

3 Proposed method

Figure V.1 displays the overall scheme of our proposed system to classify skin lesions. The following sections describe the dataset, tools, and experimental configuration used for developing and testing our CAD for skin lesions.

3.1 Data description

This section describes the public ISIC2019 and private dataset used to train and test our proposed method. ISIC is an abbreviation of (International Skin Imaging Collaboration), sponsored by ISDIS (International Society for Digital Imaging of the Skin). Figure V.2 shows some examples of dermoscopic images in The last challenge ISIC2019 used in this work [11]. Another private dataset is used too for testing. This data collected from a dermatology office in Besançon, France. All the information about both datasets are listed in Table V.2.

3.2 Data preparation

Data preprocessing is a key step in transforming the raw skin lesion dataset into a clean and understandable format for analysis, And consequently enhance the efficiency of CAD system for skin lesions. par

Work	Architecture	Pretrained model	Dataset	Preprocessing step	imbalanced problem	Using metadata	Using AHPS
Esteve A, et al.2017 [9]	InceptionV3	✓	ISIC2017 +Private dataset	✓	✓	×	×
Rezvantlab, et al.2018 [103]	DenseNet 201 ResNet 152 InceptionV3 Inception ResNet v2	✓	HAM10000 + PH ²	✓	✓	×	×
Gessert N et al.2018 [102]	DenseNet SENet ResNeX	✓	ISIC 2018	✓	✓	×	×
Reddy et al. 2018 [115]	ResNet50	✓	ISIC 2017	✓	✓	×	×
Chaturvedi et al. 2019 [104],	MobileNet	✓	HAM10000	×	×	×	×
Sae-lim [116] et al. 2019	modified MobileNet	×	HAM10000	✓	✓	×	×
Gessert N et al.2019 [110]	Ensemble of EfficientNets	✓	ISIC 2019	✓	✓	✓	×
Li W.et al 2020 [101]	AlexNet VGG19 ResNet50 DenseNet16 SENet154 PNASNet-5	✓	ISIC 2018	✓	✓	✓	×
M.A.Kassem et al. 2020 [105]	GoogleNet	✓	ISIC 2019	×	✓	✓	×
Our proposed approach	Noisy-Student(EfficientNet-L2)	✓	ISIC 2019 + Private data	✓	✓	✓	✓

Table V.1: Comparison of previous work in skin lesions classification.

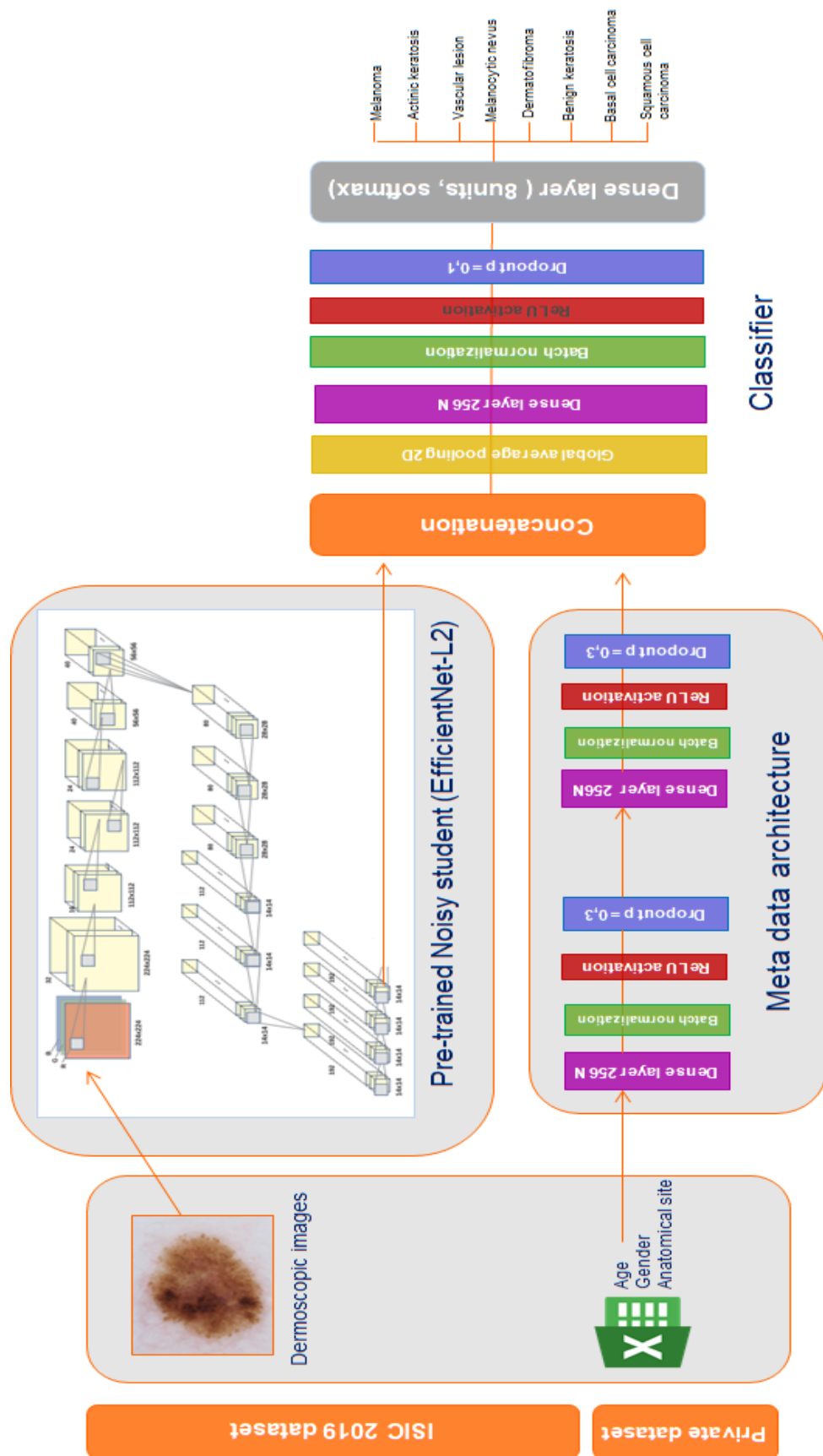


Figure V.1: Overall scheme of our proposed method

Concerning metadata, as machine learning models are based on mathematical equations, encoding categorical features is needed [117]. For the gender and anatomical site, we have applied label encoding. The anatomical site is represented by numbers from 1 to 8, the same applied to sex (1, 2), and missing values are encoded as 0.

Data contents	Public dataset	Private dataset
- Melanoma (MEL)	✓	✗
- Melanocytic nevus (NV)	✓	✗
- Vascular lesion (VASC)	✓	✗
- Actinic keratosis (AK)	✓	✓
- Squamous cell carcinoma (SCC)	✓	✓
- Dermatofibroma (DF)	✓	✗
- Basal cell carcinoma (BCC)	✓	✓
- Benign keratosis (BK) (solar lentigo/seborrheic keratosis/lichen planus-like keratosis)	✓	✗
- Unknown class (UNK)	✓	✗
- Dermoscopic images	✓	✓
- Meta data (age, gender, anatomical site)	✓	✓
- Macroscopic images	✗	✓
Training set	25,331 instances	/
Testing set	8,238 instances	45 instances

Table V.2: Description of public and private dataset.

Missing values: In health analytics, missing data can be evident. Many methods are available in the literature to deal with this problem. We are going to impute missing values in this work using the technique of K-nearest neighbors (KNN). The Principle of KNN is connecting a point with its closest k neighbors [118]. It may be used for continuous, ordinary, discrete, and categorical data. Therefore, it's useful for dealing with all types of missing values. In our work, we have implemented KNN algorithm with K= 6.

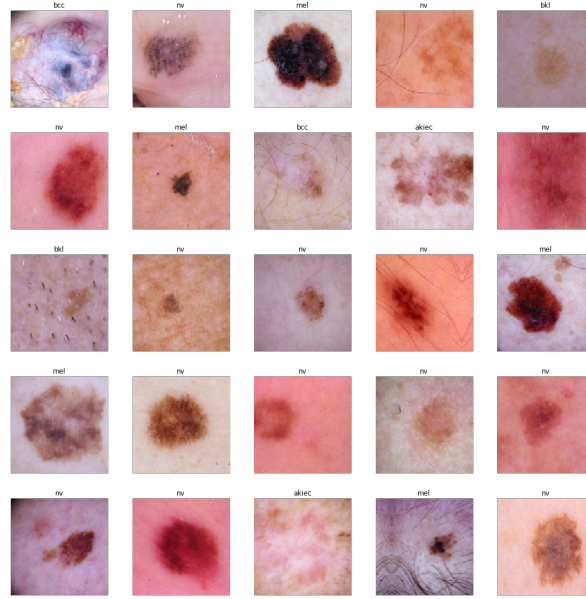


Figure V.2: Dermoscopic images' example in ISIC 2019

Image resolutions downscaled Dermoscopic images in the dataset have multi resolutions and sizes. This is due to multi-sources: a part of the dataset is from HAM10000 dataset, which contains images of size 600×450 . the second source is the BCN 20000 dataset, which involves images of size 1024×1024 , and MSK dataset contains images of various sizes. We have downsampled them to 255×255 pixel resolutions to uniform the images and make them compatible with our architecture NS-EfficientNet-L2.

3.3 Data augmentation

As shown in Figure V.3, the ISIC 2019 dataset has an unbalance distribution of images among the eight classes. To rebalance these classes, data augmentation is applied to increase minority classes: dermatofibroma, Melanoma, Basal Cell Carcinoma, Benign Keratosis, Actinic Keratosis, vascular. Data augmentation generated around 6000 images in each class to be the total of images in the training dataset is 38,569 images. parameters used for data augmentation of the images are:

- Horizontal flip = True
- Vertical flip = True
- Rotation range = 1000

- Width shift range = 0.1
- Fill mode = 0.9
- Height shift range = 0.1
- Zoom range = 0.1

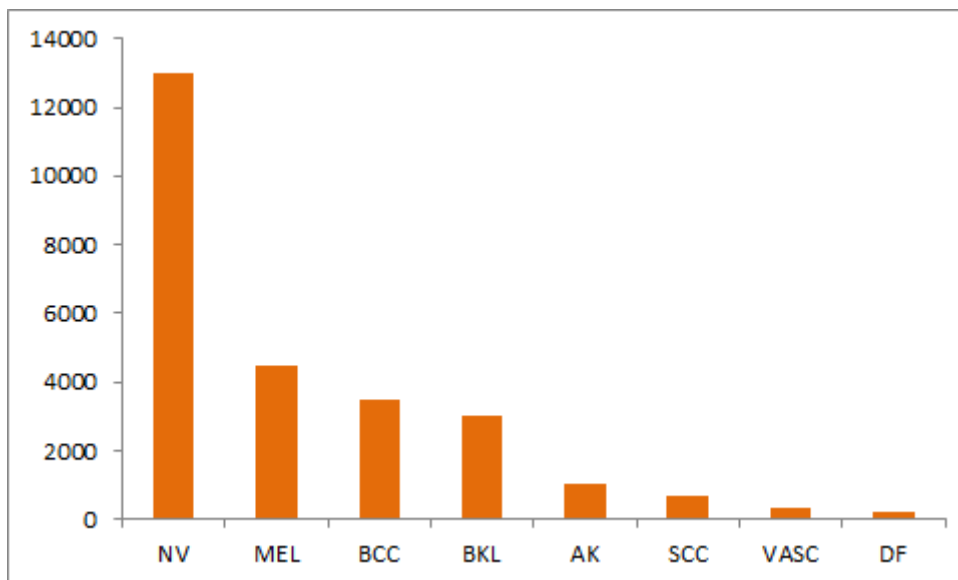


Figure V.3: Distribution of dermoscopic images per class in ISIC2019 dataset

3.4 Data analysis

In this section, we are going to present techniques used to construct our CAD system. This later contains three main components presented below:

Feature extractor For feature extraction, we are going to use pretrained EfficientNet-L2 trained with the Noisy Student method (NS-EfficientNet-L2) [107]. This technique achieved an accuracy of 88.4%, which is 2.9% higher than EfficientNet-L2 without Noisy Student, and outperformed the best methods proposed to classify the ImageNet challenge with 2.0%.

EfficientNet group involves eight models between B0 and B7 [107], and although the number of models grows, the number of calculated parameters does not increase considerably, while accuracy improves noticeably. Instead of the Rectifier Linear Unit

(ReLU) activation function used by most CNN architectures, EfficientNet uses the new Swish activation function [119]. EfficientNet-L2 is deeper and wider than EfficientNet-B7 but uses a lower resolution with more parameters to fit a huge quantity of unlabeled files[107]. This later is needed for the noisy student method, which has four key steps:

1. Use labeled images to train a teacher model

$$\frac{1}{n} \sum_{i=1}^n l(y_i, f^{noised}(x_i, \theta^t))$$

where:

x_i : Labeled images

y_i : Labels

θ^t : Teacher model

n : Number of labeled images

2. Generate pseudo labels on unlabeled images using the teacher model

$$\tilde{y}_i = f(\tilde{x}_i, \theta_*^t), \forall_i = 1, \dots, m$$

where:

\tilde{y}_i : Unlabeled images

\tilde{y}_i : Pseudo labels

m : Number of unlabeled images

3. Use of labeled images and pseudo labeled images to train a student model with noise added (dropout, data augmentation, stochastic depth) .

$$\frac{1}{n} \sum_{i=1}^n l(y_i, f^{noised}(x_i, \theta^s)) + \frac{1}{m} \sum_{i=1}^m l(\tilde{y}_i, f^{noised}(\tilde{x}_i, \theta^s))$$

where:

θ^s : Student model

4. Repeat this algorithm a few times by switching the student as a teacher train a new student and relabel the unlabeled data. This is schematically shown in Figure V.4.

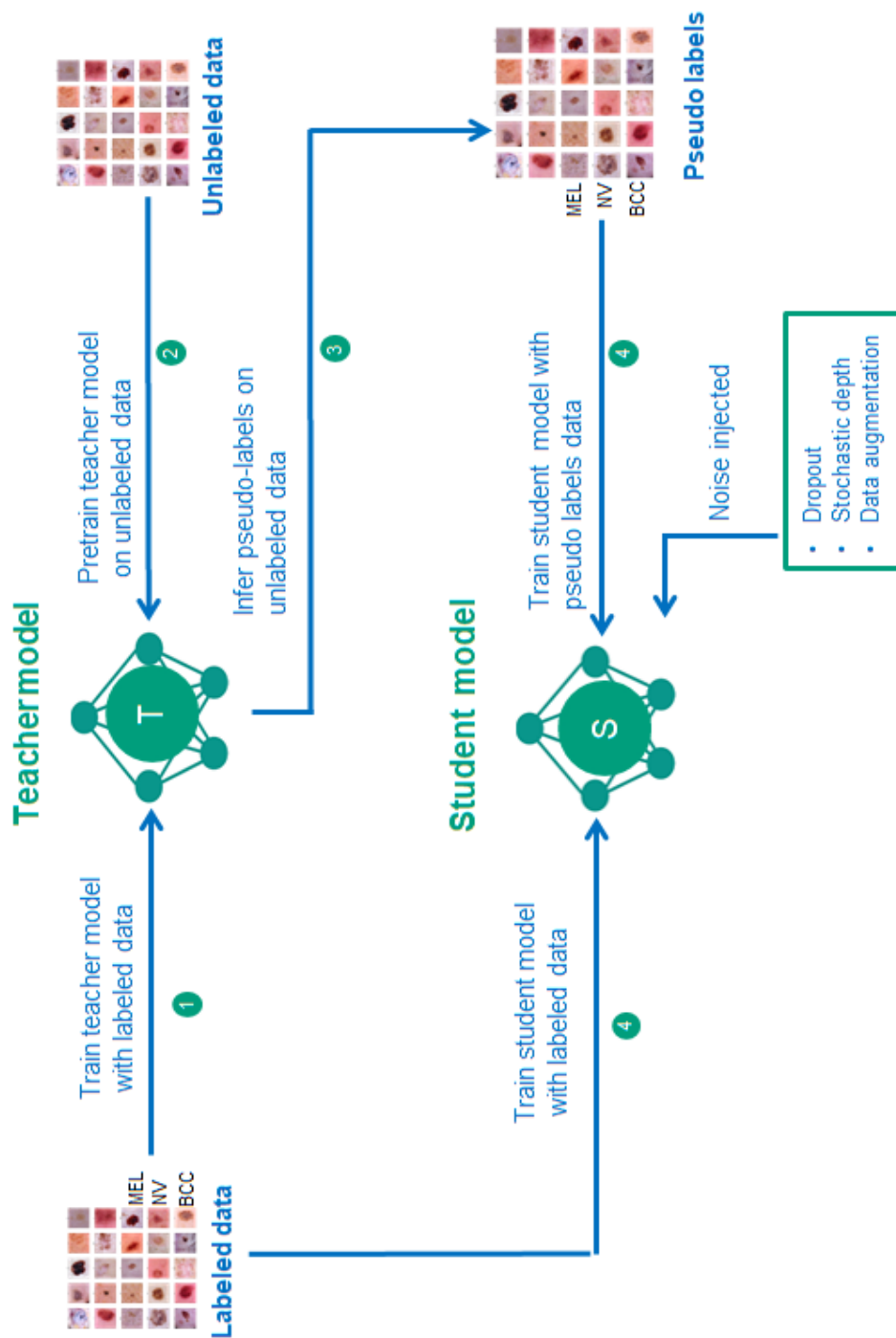


Figure V.4: Self-training with Noisy Student illustration.

Metadata architecture: For metadata architecture, two dense layers of Neural Network are constructed. Each layer contains 256 neurons, batch normalization, dropout = 0.3, and a ReLU activation. The output of the Metadata network is concatenated with the dermoscopic feature extractor output.

Classifier: In the classification level, a global average pooling2D layer has been applied to avoid overfitting by reducing the total number of parameters. Then we added another layer with batch normalization, dropout, and ReLU activation. Then it followed by one dense layer which contains eight output units for the classification, with Softmax activation function (see figure V.1).

3.5 Model's Hyper parameters optimization

For the training step, our combined architecture will be trained on two levels. We start training the feature extractor architecture with its hyperparameters. Then, we will freeze it to start training the second part(two metadata dense layers, the dense layer after concatenation, and dense layer for classification). In this section, we are going to explain our CNN-AHPS method by showing the tools and algorithms used for the implementation.

3.5.1 Tools

Hadoop [26] is a distributed computation and storage tool that supports the MapReduce programming model. Hadoop is more popular among other big data frameworks like Apache Storm [120] or Apache Spark [121], and it is used frequently in medium-sized data science research. This success due to its advantages: easy and quick to use and set up, moreover its compatibility on heterogeneous infrastructures [12].

Hadoop involves two main components. The first is a distributed data storage system called Hadoop Distributed File System (HDFS) [122], that manages the storage of extremely large files in a distributed, reliable, and fault-tolerant manner. The second component is the MapReduce model for distributed data processing[123]. We have used MapReduce in our work because it fits our requirements and simple to use[124]. It splits large jobs into two stages, called Map and Reduce. As shown in Figure V.5, we have

applied “Map” stage to separate a set of hyperparameters combinations into multiple parts, to be further treated in parallel, and each produces a final result, which is the value of the F1-score metric. The “Reduce” stage finds out the best configuration to our architecture as a final job result according to max F1-score value.

3.5.2 Implementation

The following list summaries all chronological phases for executing a distributed hyperparameter tuning via MapReduce presented in Algorithm 1.

1. An input file contains all the potential combinations of hyperparameters (Epochs, optimizer, Batch size, Learning rate) is created (one combination per line). This file generated automatically based on parameter ranges specified as following:
 - (a) Optimizer: SGD, Adam, and RMSprop are most used in image classification literature. in our work, we have added AdaMax, AMSGrad, Nadam, Adadelata Mini-Batch, and GD Momentum to test as much as possible optimizers
 - (b) Learning rate: min= 10^{-2} , max = 10^{-8} , step = 10^{-1} .
 - (c) Batch size: min= 30, max= 330, step = 30.
 - (d) Epochs: min= 20, max= 200, step= 40. After the file generation, it will upload to the HDFS, where it serves as the input file of the Hadoop job.
2. The Hadoop job begins by dividing the workload into N Map tasks, where N is the total number of lines as shown in Algorithm 2 in the file. Each task executes a setup function (only once per Map task) that contains the following steps (see Algorithm 2):
 - (a) Split the data into a training set and a testing set.
 - (b) Load the combination of hyperparameter for each task

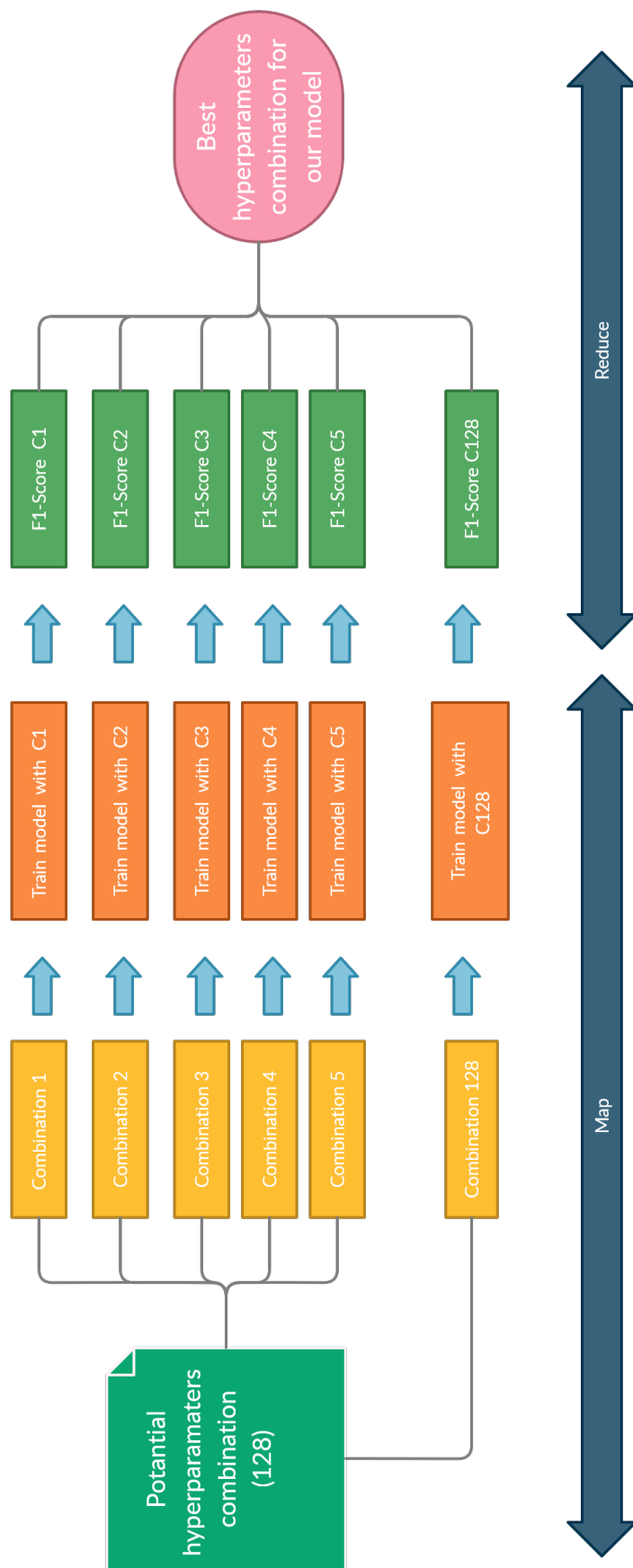


Figure V.5: MapReduce illustration.

- (c) Model training: Build the classifier using the current combination of hyper-parameters and the training set.
- (d) Model testing: classify each instance of the testing set using the previously built classifier model.
- (e) Get the value of F1-score of each combination

$$\text{Recall} = \frac{\sum_{i=1}^k \frac{TP_i}{TP_i + FN_i}}{K}$$
$$\text{Precision} = \frac{\sum_{i=1}^k \frac{TP_i}{TP_i + FP_i}}{k}$$
$$\text{F1-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- (f) Save constructed model with its hyperparameters and its F1score value
3. for Reduce task, we apply a search process to find out the highest F1-score achieved in all Map output, then return it and delete the other architectures (see Algorithm 3). $\max[L.F1 - Score]$

Algorithm 1: CNN-AHPS algorithm

Input:

C: a hyper-parameters Combination
N: number of combination in the file
M: untrained model
D: Data set

Result: Lbest

generateInput();

startJob();

for all task N **do**

- | LoadData();
- | $L < N, M', F1_{score} >.add(\text{Map}(C, N, M, D))$

endLbest = Reduce(L);

Algorithm 2: Map algorithm

Input:

C: a hyperparameters Combination
N: number of combination in the file
M: untrained model
D: Data set

Result: (N, M',F1-score)

D-train , D-test = splitData(D);

M'= Train (M, C , D-train); // Train model M on training data set
using hyperparameters combination C

F1-Score = Test(M',D-test);

// test trained model M' on testing data set and return

F1-score value

Algorithm 3: Reduce algorithm

Input:

L: list of vectors obtained by mapper tasks that contain :

N: number of combination in file

M': trained model on Cn combination

F1-score: Evaluation metric for each training task

Result: <LBest.N, LBest.M', Lbest.F1-score>**while** *L.hasNext()* **do** **if** *L.F1-score > max* **then**

max = L.F1-score;

LBest =L;

end**end**Delete (L);

4 Experimental Results

Experiments presented in this section are compiled and run on Hadoop framework with Python language using Scikit-learn bibliography. Characterises of hardware are: CPU : Intel (R) Core (TM) i7-7500U @ 4.00GHz, and 12GB RAM. In each experiment, Values represent the mean results (standard deviations) and evaluated using the standard metrics that introduced in chapter 2 section 2.6.4. In addition, Balanced Multi-class Accuracy is the arithmetic mean over the skin lesion classes. This metric

is used to evaluate ISIC2019 participants. Besides, the Area Under Curve (AUC) of the Receiver Operation Characteristic (ROC) curve has been plotted to compare the performance of each experiment[125].

- Balanced multiclass accuracy = $\frac{\sum_{i=1}^k TP_i}{\sum_{i=1}^k (TP_i + FN_i)}$

where :

K= number of classes

TP = True positive means the number of truly predicted class.

TN = True Negative means the number of non $class_i$ that are truly classified non $class_i$.

FP = the number of $non - class_i$ that are misclassified as $class_i$

FN=False negative means the number of the lost $class_i$ objects.

4.1 Results with /without metadata

Two experiments are presented in this section to evaluate the impact of metadata on our model. The first experiment is to classify skin lesions only with dermoscopic images. We have used NS-EfficientNet-L2 as a feature extractor adding two dense layers for classification. The second experiment is to add metadata to our model. At this point, we have concatenated our feature extractor with metadata architecture illustrated in figure V.1.

Architecture's hyperparameters are chosen manually by the Error-And-Trial method in both experiments. Table V.3 shows the performance of our CAD without and with metadata. Experimental results showed that the performance is slightly improved when metadata is used across all metrics except the sensitivity. Compared to Test 1, we can clearly observe the low sensitivity with the Unknown class and melanoma. This decrease in sensitivity in the Unknown class can be explained by the lack of metadata in this class. The results for the melanoma class remain without a logical explanation.

	Sensitivity	Specificity	AUC	F1-Score	Accuracy	Balanced Multiclass Accuracy	
Without metadata	SCC	54.3	96.9	92.7	69.5		
	VASC	63.2	98.8	93.3	77.0		
	DF	44.9	98.4	95.3	61.6		
	BKL	67.3	91.5	88.9	77.5		
	AK	39.7	98.1	91.5	56.5		
	BCC	66.3	93.9	93.2	77.7		
	NV	68.2	95.8	96.1	79.6		
	MEL	71.6	97.2	96.5	82.4		
	UNK	11.9	97.5	60.3	21.2		
	Mean	54.1	96.4	79.0	67.0	75.5	46.9
	With metadata	SCC	57.2	98.1	92.7	72.2	
VASC		62.4	98.7	93.9	76.4		
DF		49.9	95.3	98.3	65.5		
BKL		63.9	97.5	93.7	77.2		
AK		48.3	98.9	97.4	54.9		
BCC		63.5	91.7	93.2	75.0		
NV		68.2	95.8	94.0	79.6		
MEL		59.7	99.2	98.2	74.5		
UNK		09.9	99.4	55.4	18.0		
Mean		53.6	97.1	90.7	69.0	82.9	55.3

Table V.3: Results obtained with and without metadata.

4.2 Results with Automatic hyperparameters Selection (CNN-AHPS)

In the second experiment, we have applied our automatic hyperparameter selection algorithm. Table V.4 shows the hyperparameters that we have selected manually. The second columns represent hyperparameters selected by CNN-AHPS.

A significant performance improvement is achieved when CNN-AHPS is applied presented in Table V.5. we can observe a high sensitivity and F1-score compared with the first and second experiments. Figure V.6 shows the ROC curve and AUC of our model with CNN-AHPS. According to the results in the figure VASC class gets higher AUC scores than the other classes. We can also notice an improvement in MEL and UNK classes.

The bar chart in FigureV.7 shows the difference in the performance of all three experiments presented in this study across all evaluation metrics. Overall, we notice a significant performance improvement is achieved when CNN-AHPS is applied.

Table V.6 show results obtained when proposed CAD applied on private data set. we can notice that the system achieved results much better then on public data set, this is due to the absence of unknown class, also the size of data plays a role here with only 45 cases.

hyperparameters	Manually	Automatically using CNN-AHPS
Optimizer	Adam	RMSPROP
Learning rate	0.01	0.05
Batch size	30	35
Epochs	150	250

Table V.4: Hyperparameters selected manually VS automatically.

		Sensitivity	Specificity	AUC	F1-Score	Accuracy	Balanced Multiclass Accuracy
With CNN-AHPS	SCC	86.3	98.5	90.3	91.9		
	VASC	82.6	99.7	95.1	89.85		
	DF	89.9	95.1	94.6	92.5		
	BKL	78.9	97.5	83.9	87.3		
	AK	90.1	97.7	93.2	93.7		
	BCC	88.0	90.5	92.9	89.2		
	NV	91.7	97.2	91.0	94.3		
	MEL	86.7	99.7	94.8	92.7		
	UNK	24.5	89.5	60.9	39.2		
	Mean	79.8	97.5	85.5	87.7	91.2	69.4

Table V.5: Results obtained by applying CNN-AHPS.

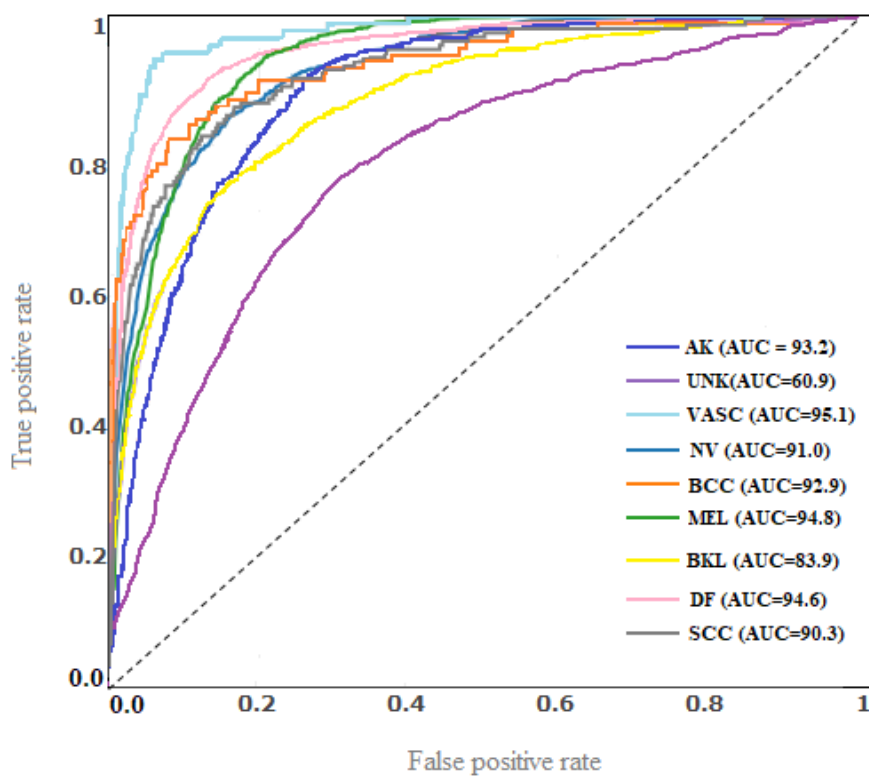


Figure V.6: ROC curve of our CAD system with metadata and AHPS.

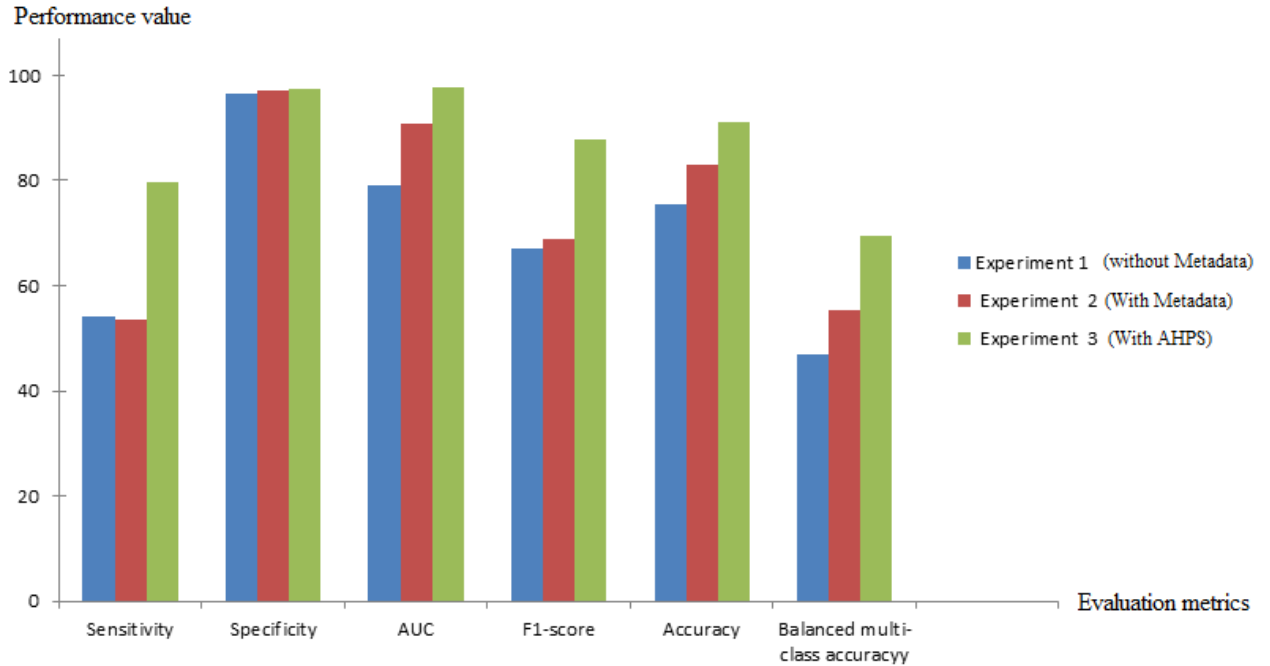


Figure V.7: Performance comparison of three applied experiments.

	Sensitivity	Specificity	AUC	F1-Score	Accuracy	Balanced Multiclass Accuracy
BCC	87.6	99.1	98.3	97.2		
AK	90.7	99.2	99.1	98.5		
SCC	97.9	94.2	98.6	98.5		
Mean	92.0	97.5	98.6	98.0	97.8	89.7

Table V.6: Results of our CAD system on Private dataset.

We now come to compare the obtained diagnosis results from our proposed model with two main published works that have used the same data for the same objective. We can notice that our model outperformed Nils Gessert et al.[96] who achieved Top1 Balanced Multi-Class Accuracy in the ISIC2019 challenge. Our work outperformed M. A. Kassem et al. [105] which have recently contributed to developing a CAD system for skin lesions (see Table V.7).

	Accuracy	Sensitivity	Specificity	AUC	F1-Score	Balanced Multiclass Accuracy
Nils Gessert et al.2019 [96]	92.6	50.7	97.7	85.1	51.5	63.6
M. A. Kassem et al.2020 [105]	81	74	84	/	/	/
Our approach	91.2	79.8	97.5	85.5	87.7	69.4

Table V.7: Comparative study with previous work.

5 Discussion

The primary goal of this study was to develop a computer-aided diagnosis to classify skin lesions and evaluate the model using public and private datasets. These datasets contain dermoscopic images of skin lesions and clinical information (age, gender, anatomical site). We have applied three experiments to evaluate the impact of adding metadata and our developed AHPS. The best results are achieved when metadata and AHPS are applied on our CAD with Accuracy = 94.3%, F1-score= 91.8%, AUC= 91.8%, and Balanced Multiclass accuracy = 96.7% in ISIC 2019 dataset and outperformed previous work. besides, Accuracy = 94.3%, F1-score= 91.8%, AUC= 91.8% and Balanced Multiclass accuracy = 96.7% on private data set. This model is based on a pretrained NS-EfficientNet-L2 as a feature extractor, concatenated with metadata architecture. Then, a classifier follows with eight neurons and Softmax function.

Three major observations can be concluded from the experimental results. First of all, the great performance achieved and outperform previous work that confirms the effectiveness of NS-EfficientNet-L2 to classify skin lesions. Second, speedup achieved 128 times less by applying AHPS using MapReduce, and improvement that validate the important role of hyperparameters on performance. In the last observation concerning the unknown class in the testing dataset, the performance of the unknown class is noticeably lower than the other classes. This could explain the large difference between tests on the private dataset and the results on the official test set.

There are three major limitations in this study that could be addressed in future research. First, the study focused only on automatic hyperparameters selection for

training, and not for model construction like the number of layers, nodes, dropout percentage, activation functions, etc. Second, Since deep learning methods are data-driven, their main limitations also come from the data itself. For metadata existed in the skin lesions' pubic data set contains only sex, gender, and anatomical site. Whereas many clinical features can be added to improve the diagnosis like: the history of sunburns, personal or family history of skin cancer, Some common medications, and drugs(antibiotics) can make the skin more sensitive to sunlight. Besides dermoscopic images, adding macroscopic images can also help for a more accurate diagnosis. The last limitation concerning the validation of the model in the real application, we have developed a desktop application that provides simple user interfaces for a dermatologist to apply real-time diagnosis with our developed CAD, and we will validate results obtained by histological examination.

Future work will consist of extending CNN-AHPS to model construction hyperparameters and fine-tuning pretrained models. Moreover, studying the impact of adding macroscopic images and other metadata on the classification of skin lesions. We believe that our model can improve the classification of other medical images as well. To this end, we aim to test this model in other medical datasets.

6 Conclusion

In this chapter, we have presented our main contribution that consist of developing a CAD system to assist dermatologists to classify skin lesions. This work is based on three main points: (1) Data quality: we have preprocessed the data in this phase by imputing missing values, solving images' multi resolutions problem, solving imbalanced classes problem, and adding metadata to the model. (2) Appropriate architecture and learning strategy: this is done using a pretrained NS-EfficientNet-L2 architecture that achieved Top1 accuracy in image classification literature. (3) Appropriate hyperparameters for CNN architecture: to this end, we have developed CNN-AHPS based on big data tools (MapReduce). This method gave us the best configuration of hyperparameters to our architecture in reduced time. Results from ISIC2019 confirmed that our model improved the classification for all of the specificity, sensitivity, AUC, and F1-score.

Chapter VI

General conclusion

This chapter summarizes the contributions of this thesis and presents several research directions that require further investigations in the future.

1 Summary

This thesis integrates three trending concepts: Big Data, PHM and Artificial Intelligence in the medical field. The research work presented in this thesis emphasizes computer-aided diagnosis systems to assist physicians in making an accurate diagnosis generally, and data-driven models in dermatology domain precisely.

The problems treated in this thesis gave birth to three main contributions that answer the questions presented in the introduction. These contributions can be summarised as follows:

- We proposed an adaptation of a PHM model from fault diagnosis of an aircraft engine to diagnosis human heart disease. This PHM model is based on a new strategy by retargeting extreme learning machine (ELM) algorithm. The main objective of this study is to transfer the PHM approach from the industrial to the medical field. This work could be considered a first step to reduce the gap between industry and medical field, by exchanging the applied techniques, and proving that models applied for a machine's health diagnosis could be applicable for human health diagnosis. The suggested system accomplished a higher classification accuracy rate, by improving the data quality, decreasing the number of

attributes and obtained a higher performance rate, with reduced processing time. The ID-RELM & RF model can be used as a medical decision support system for cardiologists to make accurate classification with lower time, cost, and effort.

- Second contribution attempts to analyze Spitzoid lesions related to clinical, histological, and immunohistochemical features using AI techniques. Seven (7) classifiers: K-NN, LR, DT, MLP, NB, RF, and SVM have been applied for the analytic procedures. The hybrid technique of SMOTE-GA-MLP yields the highest performance overall.

The added value of this research in the area of skin lesion classification are now summarized. First, it specifies the exact type of Spitz lesion, which is extremely difficult and challenging in real life. Second, it combines previous works on the steps needed to develop an automatic CAD system for Spitzoid lesion classification to assist dermatopathologists during the diagnosis process. Third, our work makes a classification based on various testes and types of data: clinical, histological, and immunohistochemical data. Contrary to previous literature work that only concentrates on the microscopic vision, which cannot accurately classify them. Finally, the analysis for differentiating major classes of these lesions, namely SN (Spitz nevus) v. AST, is based on several features, including the immunohistochemical markers. Specifically, the findings indicate that localization of lesions, cytonuclear atypia, and Ki67 proliferative index are the most weighted features to differentiate AST from SN.

- Based on the limitations of the second contribution, we have proposed a new computer-aided diagnosis for skin lesion that eliminates the dermatologist's intervention to extract features by developing a CAD system that classifies skin lesions directly from dermoscopic images. This contribution is based on three main points: (1) Data quality: we have preprocessed the data in this phase by imputing missing values, solve images' multi resolutions problem, solve imbalanced classes problem, and add metadata to the model. (2) Appropriate architecture and learning strategy: this is done using a pretrained NS-EfficientNet-L2 architecture that achieved Top1 accuracy in image classification literature.

(3) Appropriate hyperparameters for CNN architecture: to this end, we have developed CNN-AHPS based on big data tools (MapReduce). This method gave us the best configuration of hyperparameters to our architecture in reduced time. Results from ISIC2019 confirmed that our model improved the classification for all of the specificity, sensitivity, AUC, and F1-score.

2 Perspectives

Several improvements could be made to the work done in this thesis, and research directions that require further investigations in the future. We listed some of them in the items below:

- The developed CAD systems in this thesis are data-driven, their main limitations also come from the data itself. In future work we intend to study the impact of adding macroscopic images and more metadata on skin lesions classification. Used public data set in the current work contains only age, sex, and anatomical site. Many clinical features can be added to improve the diagnosis, such as a history of sunburns, personal or family history of skin cancer, and some common medications and drugs (antibiotics) that can make the skin more sensitive to sunlight. In addition to dermoscopic images, the support of macroscopic images could also help achieve a more accurate diagnosis.
- We intend to test the model in real applications. To this end, we will collaborate with experts in dermatology for better understanding their needs and capitalize their expertise by developing a desktop application that provides simple user interfaces for a dermatologist. This application will facilitate applying real-time diagnosis with our developed CAD. Finally, results obtained will be more validated by histological examination.
- We intend to extend our work on the whole M-PHM process by using results obtained from the diagnosis phase to make predictions and support decision making. For example, predict skin lesion growth, predict the spread and tumour stage if the lesion diagnosed as cancerous. Furthermore, based on these predictions, we will support the decision making concerning the treatment should be taking.
- We believe that our model can be relevant to the classification of other medical

images as well. To this end, we aim to test this model in other medical datasets such as brain tumors on MRI images, breast cancer tumors on mammography or whole slide imaging.

- We will focus on the healthcare situation in Algeria. To this end, a partnership with the Algerian medical community is needed. This collaboration can help to understand their needs, identify their challenges. Thus we can design and propose suitable solutions based on our developed CAD systems.

Bibliography

- [1] Li Da Xu, Eric L Xu, and Ling Li. Industry 4.0: state of the art and future trends. *International Journal of Production Research*, 56(8):2941–2962, 2018.
- [2] Vania V Estrela, Ana Carolina Borges Monteiro, Reinaldo Padilha França, Yuzo Iano, Abdeldjalil Khelassi, and Navid Razmjooy. Health 4.0: applications, management, technologies and review. *Medical Technologies Journal*, 2(4):262–276, 2018.
- [3] Rishi Duggal, Ingrid Brindle, and Jessamy Bagenal. Digital healthcare: regulating the revolution, 2018.
- [4] Rafael Gouriveau, Kamal Medjaher, and Noureddine Zerhouni. *From prognostics and health systems management to predictive maintenance 1: Monitoring and prognostics*. John Wiley & Sons, 2016.
- [5] Brigitte Chebel-Morello, Jean-Marc Nicod, and Christophe Varnier. *From Prognostics and Health Systems Management to Predictive Maintenance 2: Knowledge, Reliability and Decision*. John Wiley & Sons, 2017.
- [6] Javier Andreu-Perez, Carmen CY Poon, Robert D Merrifield, Stephen TC Wong, and Guang-Zhong Yang. Big data for health. *IEEE journal of biomedical and health informatics*, 19(4):1193–1208, 2015.
- [7] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and SS Iyengar. Computational health informatics in the big data age: a survey. *ACM Computing Surveys (CSUR)*, 49(1):1–36, 2016.
- [8] Fei Wang, Lawrence Peter Casalino, and Dhruv Khullar. Deep learning in medicine—promise, progress, and challenges. *JAMA internal medicine*, 179(3):293–294, 2019.
- [9] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.
- [10] Cleveland Heart Disease Dataset, 1990.
- [11] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.
- [12] Min Chen, Shiwen Mao, Yin Zhang, Victor CM Leung, et al. *Big data: related technologies, challenges and future prospects*, volume 96. Springer, 2014.
- [13] Ioannis Anagnostopoulos, Sherali Zeadally, and Ernesto Exposito. Handling big data: research challenges and future directions. *The Journal of Supercomputing*, 72(4):1494–1516, 2016.

- [14] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Big data in healthcare: Challenges and opportunities. In *2015 International Conference on Cloud Technologies and Applications (CloudTech)*, pages 1–7. IEEE, 2015.
- [15] Philip Russom et al. Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34, 2011.
- [16] Reza Mehmood and Arvind Selwal. Fingerprint biometric template security schemes: Attacks and countermeasures. In *Proceedings of ICRIC 2019*, pages 455–467. Springer, 2020.
- [17] Borko Furht and Flavio Villanustre. Introduction to big data. In *Big data technologies and applications*, pages 3–11. Springer, 2016.
- [18] Nishita Mehta and Anil Pandit. Concurrence of big data analytics and healthcare: A systematic review. *International journal of medical informatics*, 114:57–65, 2018.
- [19] Laura B Stokes, Joseph W Rogers, John B Hertig, and Robert J Weber. Big data: implications for health system pharmacy. *Hospital pharmacy*, 51(7):599–603, 2016.
- [20] Prabha Susy Mathew and Anitha S Pillai. Big data solutions in healthcare: Problems and perspectives. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–6. IEEE, 2015.
- [21] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wal-lach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):1–26, 2008.
- [22] Avinash Lakshman and Prashant Malik. Cassandra: structured storage system on a p2p network. In *Proceedings of the 28th ACM symposium on Principles of distributed computing*, pages 5–5, 2009.
- [23] ABM Moniruzzaman and Syed Akhter Hossain. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*, 2013.
- [24] Kristina Chodorow. *MongoDB: the definitive guide: powerful and scalable data storage.* ” O’Reilly Media, Inc.”, 2013.
- [25] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakula-pati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: amazon’s highly available key-value store. *ACM SIGOPS operating systems review*, 41(6):205–220, 2007.
- [26] Tom White. *Hadoop: The definitive guide.* ” O’Reilly Media, Inc.”, 2012.
- [27] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, pages 1–10. Ieee, 2010.
- [28] Claudio Tesoriero. *Getting started with orientDB.* Packt Publishing Ltd, 2013.
- [29] Lars George. *HBase: the definitive guide: random access to your planet-size data.* ” O’Reilly Media, Inc.”, 2011.
- [30] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
- [31] Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, and Dennis Fetterly.

- Dryad: distributed data-parallel programs from sequential building blocks. In *Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*, pages 59–72, 2007.
- [32] Christopher Moretti, Jared Bulosan, Douglas Thain, and Patrick J Flynn. Allpairs: An abstraction for data-intensive cloud computing. In *2008 IEEE international symposium on parallel and distributed processing*, pages 1–11. IEEE, 2008.
- [33] Grzegorz Malewicz, Matthew H Austern, Aart JC Bik, James C Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 135–146, 2010.
- [34] Arvind Sathi. *Big data analytics: Disruptive technologies for changing the game*. Mc Press, 2012.
- [35] Tom M Mitchell et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.
- [36] Laura Maruster. *A machine learning approach to understand business processes*. Citeseer, 2003.
- [37] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.
- [38] Khadija El Bouchefry and Rafael S de Souza. Learning in big data: Introduction to machine learning. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, pages 225–249. Elsevier, 2020.
- [39] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghahfoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [40] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 42(11):226, 2018.
- [41] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.
- [42] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [43] Shaomin Wu. A review on coarse warranty data and analysis. *Reliability Engineering & System Safety*, 114:1–11, 2013.
- [44] Rima Houari, Ahcène Bounceur, Tahar Kechadi, A-Kamel Tari, and Reinhardt Euler. Missing data analysis using multiple imputation in relation to parkinson’s disease. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, pages 1–6, 2016.
- [45] Ton J Cleophas, Ton J Cleophas, Aeilko H Zwinderman, and Aeilko H Zwinderman. *Clinical Data Analysis on a Pocket Calculator: Understanding the Scientific Methods of Statistical Reasoning and Hypothesis Testing*. Springer, 2016.
- [46] Kazuo Nakatani, Ta-Tao Chuang, and Duanning Zhou. Data synchronization technology: standards, business values and implications. *Communications of the*

- Association for Information Systems*, 17(1):44, 2006.
- [47] Jorge Sola and Joaquin Sevilla. Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on nuclear science*, 44(3):1464–1468, 1997.
- [48] Malgorzata Bach, Aleksandra Werner, J Żywiec, and Wojciech Pluskiewicz. The study of under-and over-sampling methods’ utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, 384:174–190, 2017.
- [49] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [50] Abdul Majid, Safdar Ali, Mubashar Iqbal, and Nabeela Kausar. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. *Computer methods and programs in biomedicine*, 113(3):792–808, 2014.
- [51] Vepa Atamuradov, Kamal Medjaher, Pierre Dersin, Benjamin Lamoureux, and Nouredine Zerhouni. Prognostics and health management for maintenance practitioners-review, implementation and tools evaluation. *International Journal of Prognostics and Health Management*, 8(060):1–31, 2017.
- [52] Wullianallur Raghupathi and Viju Raghupathi. An overview of health analytics. *J Health Med Informat*, 4(132):2, 2013.
- [53] Harry T Lawless and Hildegarde Heymann. Descriptive analysis. In *Sensory evaluation of food*, pages 227–257. Springer, 2010.
- [54] Yaroslav Faybishenko and Boris Faybishenko. Health diagnostic systems and methods, September 15 2015. US Patent 9,131,893.
- [55] T Eswari, P Sampath, S Lavanya, et al. Predictive methodology for diabetic data analysis in big data. *Procedia Computer Science*, 50:203–208, 2015.
- [56] Ronnie Ben-Zion, Nava Pliskin, and Lior Fink. Critical success factors for adoption of electronic health record systems: literature review and prescriptive analysis. *Information Systems Management*, 31(4):296–312, 2014.
- [57] VV Ramalingam, Ayantan Dandapath, and M Karthik Raja. Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, 7(2.8):684–687, 2018.
- [58] Rahma Atallah and Amjed Al-Mousa. Heart disease detection using machine learning majority voting ensemble method. In *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, pages 1–6. IEEE, 2019.
- [59] Burak Kolukisa, Hilal Hacilar, Mustafa Kuş, Burcu Bakır-Güngör, Atilla Aral, and Vehbi Çağrı Güngör. Diagnosis of coronary heart disease via classification algorithms and a new feature selection methodology. *International Journal of Data Mining Science*, 1(1):8–15, 2019.
- [60] Yong-Ping Zhao, Fang-Quan Song, Ying-Ting Pan, and Bing Li. Retargeting extreme learning machines for classification and their applications to fault diagnosis of aircraft engine. *Aerospace Science and Technology*, 71:603–618, 2017.
- [61] R Malarvizhi and Antony Selvadoss Thanamani. K-nearest neighbor in missing data imputation. *International Journal of Engineering Research and Development*, 5(1):5–7, 2012.
- [62] Bibhuprasad Sahu, Sachi Mohanty, and Saroj Rout. A hybrid approach for

- breast cancer classification and diagnosis. *EAI Endorsed Transactions on Scalable Information Systems*, 6(20), 2019.
- [63] Siyabend Turgut, Mustafa Dağtekin, and Tolga Ensari. Microarray breast cancer data classification using machine learning methods. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–3. IEEE, 2018.
- [64] Xiao-Yong Pan and Hong-Bin Shen. Robust prediction of b-factor profile from sequence using two-stage svr based on random forest feature selection. *Protein and peptide letters*, 16(12):1447–1454, 2009.
- [65] Nan-Ying Liang, Guang-Bin Huang, Paramasivan Saratchandran, and Narasimhan Sundararajan. A fast and accurate online sequential learning algorithm for feedforward networks. *IEEE Transactions on neural networks*, 17(6):1411–1423, 2006.
- [66] Kanika Pahwa and Ravinder Kumar. Prediction of heart disease using hybrid technique for selecting features. In *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, pages 500–504. IEEE, 2017.
- [67] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, and Juan Gutierrez. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 204–207. IEEE, 2017.
- [68] Shan Xu, Zhen Zhang, Daoxian Wang, Junfeng Hu, Xiaohui Duan, and Tiangang Zhu. Cardiovascular risk prediction method based on cfs subset evaluation and random forest classification framework. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*(, pages 228–232. IEEE, 2017.
- [69] Ankur Gupta, Rahul Kumar, Harkirat Singh Arora, and Balasubramanian Raman. Mifh: A machine intelligence framework for heart disease diagnosis. *IEEE Access*, 8:14659–14674, 2019.
- [70] My Chau Tu, Dongil Shin, and Dongkyoo Shin. Effective diagnosis of heart disease through bagging approach. In *2009 2nd International Conference on Biomedical Engineering and Informatics*, pages 1–4. IEEE, 2009.
- [71] Nidhi Bhatla and Kiran Jyoti. An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, 1(8):1–4, 2012.
- [72] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 2018.
- [73] Ana F Pedrosa, Jose M Lopes, Filomena Azevedo, and Alberto Mota. Spitz/reed nevi: a review of clinical-dermatoscopic and histological correlation. *Dermatology practical & conceptual*, 6(2):37, 2016.
- [74] Daryl J Sulit, Robert A Guardiano, and Stephen Krivda. Classic and atypical spitz nevi: Review of the literature-response, 2007.
- [75] Sophie Spitz. Melanomas of childhood. *The American journal of pathology*, 24(3):591, 1948.
- [76] Kelly L Harms, Lori Lowe, Douglas R Fullen, and Paul W Harms. Atypical spitz

- tumors: a diagnostic challenge. *Archives of pathology & laboratory medicine*, 139(10):1263–1270, 2015.
- [77] Elvira Moscarella, Aimilios Lallas, Athanassios Kyrgidis, Gerardo Ferrara, Caterina Longo, Massimiliano Scalvenzi, Stefania Staibano, Cristina Carrera, M Alba Díaz, Paolo Broganelli, et al. Clinical and dermoscopic features of atypical spitz tumors: a multicenter, retrospective, case-control study. *Journal of the American Academy of Dermatology*, 73(5):777–784, 2015.
- [78] Andreas Blum, Gernot Rassner, and Claus Garbe. Modified abc-point list of dermoscopy: a simplified and highly accurate dermoscopic algorithm for the diagnosis of cutaneous melanocytic lesions. *Journal of the American Academy of Dermatology*, 48(5):672–678, 2003.
- [79] Mohammed A Al-Masni, Mugahed A Al-Antari, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Computer methods and programs in biomedicine*, 162:221–231, 2018.
- [80] TY Satheesha, D Satyanarayana, MN Giri Prasad, and Kashyap D Dhruve. Melanoma is skin deep: a 3d reconstruction technique for computerized dermoscopic skin lesion classification. *IEEE journal of translational engineering in health and medicine*, 5:1–17, 2017.
- [81] Shivangi Jain, Nitin Pise, et al. Computer aided melanoma skin cancer detection using image processing. *Procedia Computer Science*, 48:735–740, 2015.
- [82] David Roffman, Gregory Hart, Michael Girardi, Christine J Ko, and Jun Deng. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. *Scientific reports*, 8(1):1–7, 2018.
- [83] Fengying Xie, Haidi Fan, Yang Li, Zhiguo Jiang, Rusong Meng, and Alan Bovik. Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE transactions on medical imaging*, 36(3):849–858, 2016.
- [84] Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21, 2011.
- [85] Weihong Han, Zizhong Huang, Shudong Li, and Yan Jia. Distribution-sensitive unbalanced data oversampling method for medical diagnosis. *Journal of medical Systems*, 43(2):39, 2019.
- [86] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.
- [87] Emina Aličković and Abdulhamit Subasi. Breast cancer diagnosis using ga feature selection and rotation forest. *Neural Computing and Applications*, 28(4):753–763, 2017.
- [88] Woodrow W Bledsoe. The use of biological concepts in the analytical study of systems. In *ORSA-TIMS national meeting*, 1961.
- [89] John Henry Holland et al. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.
- [90] Thendral Puyalnithi and Madhuviswanatham Vankadara. A unified feature selection model for high dimensional clinical data using mutated binary particle

- swarm optimization and genetic algorithm. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 13(4):1–14, 2018.
- [91] V Vapnik. Statistical learning theory new york. NY: Wiley, 1998.
- [92] Shalini Gambhir, Sanjay Kumar Malik, and Yugal Kumar. The diagnosis of dengue disease: An evaluation of three machine learning approaches. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 13(3):1–19, 2018.
- [93] Elias Ebrahimzadeh, Alireza Foroutan, Mohammad Shams, Raheleh Baradaran, Lila Rajabion, Mohammadamin Joulani, and Farahnaz Fayaz. An optimal strategy for prediction of sudden cardiac death through a pioneering feature-selection approach from hrv signal. *Computer methods and programs in biomedicine*, 169:19–36, 2019.
- [94] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [95] Adnan Firoze and Rashedur M Rahman. Mining icddr, b hospital surveillance data and exhibiting strategies for balancing large unbalanced datasets. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 10(1):39–66, 2015.
- [96] Nils Gessert, Thilo Sentker, Frederic Madesta, Rüdiger Schmitz, Helge Kniep, Ivo Baltruschat, René Werner, and Alexander Schlaefer. Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. *IEEE Transactions on Biomedical Engineering*, 67(2):495–503, 2019.
- [97] Muhammad Attique Khan, Muhammad Younus Javed, Muhammad Sharif, Tanzila Saba, and Amjad Rehman. Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification. In *2019 international conference on computer and information sciences (ICCIS)*, pages 1–7. IEEE, 2019.
- [98] Abir Belaala, Yazid Bourezane, Labib Sadek Terrissa, Zeina Al Masry, and Nouredine Zerhouni. Skin cancer and deep learning for dermoscopic images classification: A pilot study., 2020.
- [99] Mohammad Ali Kadampur and Sulaiman Al Riyaae. Skin cancer detection: applying a deep learning based model driven architecture in the cloud for classifying dermal cell images. *Informatics in Medicine Unlocked*, 18:100282, 2020.
- [100] Pegah Kharazmi, Jiannan Zheng, Harvey Lui, Z Jane Wang, and Tim K Lee. A computer-aided decision support system for detection and localization of cutaneous vasculature in dermoscopy images via deep feature learning. *Journal of medical systems*, 42(2):33, 2018.
- [101] Weipeng Li, Jiaxin Zhuang, Ruixuan Wang, Jianguo Zhang, and Wei-Shi Zheng. Fusing metadata and dermoscopy images for skin disease diagnosis. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1996–2000. IEEE, 2020.
- [102] Nils Gessert, Thilo Sentker, Frederic Madesta, Rüdiger Schmitz, Helge Kniep, Ivo Baltruschat, René Werner, and Alexander Schlaefer. Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting. *arXiv preprint arXiv:1808.01694*, 2018.
- [103] Amirreza Rezvantalab, Habib Safigholi, and Somayeh Karimijeshni. Dermatologist level dermoscopy skin cancer classification using different deep learning con-

- volutional neural networks algorithms. *arXiv preprint arXiv:1810.10348*, 2018.
- [104] Saket S Chaturvedi, Kajol Gupta, and Prakash S Prasad. Skin lesion analyser: An efficient seven-way multi-class skin cancer classification using mobilenet. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 165–176. Springer, 2020.
- [105] Mohamed A Kassem, Khalid M Hosny, and Mohamed M Fouad. Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning. *IEEE Access*, 8:114822–114832, 2020.
- [106] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [107] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.
- [108] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.
- [109] Mohammed A Al-Masni, Dong-Hyun Kim, and Tae-Seong Kim. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Computer Methods and Programs in Biomedicine*, 190:105351, 2020.
- [110] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *MethodsX*, page 100864, 2020.
- [111] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.
- [112] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.
- [113] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.
- [114] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. Massively parallel hyperparameter tuning. *arXiv preprint arXiv:1810.05934*, 2018.
- [115] Nithin D Reddy. Classification of dermoscopy images using deep learning. *arXiv preprint arXiv:1808.01607*, 2018.
- [116] Wannipa Sae-Lim, Wiphada Wettayaprasit, and Pattara Aiyarak. Convolutional

- neural networks using mobilenet for skin lesion classification. In *2019 16th International Joint Conference on Computer Science and Software Engineering (JC-SSE)*, pages 242–247. IEEE, 2019.
- [117] Abir Belaala, Labib Sadek Terrissa, Nouredine Zerhouni, and Christine Devaland. Spitzoid lesions diagnosis based on smote-ga and stacking methods. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, pages 348–356. Springer, 2019.
- [118] Abir Belaala, Zeina Al Masry, Labib Sadek Terrissa, and Nouredine Zerhouni. Retargeting phm tools: from industrial to medical field. In *PHM Society European Conference*, volume 5, pages 7–7, 2020.
- [119] Linh T Duong, Phuong T Nguyen, Claudio Di Sipio, and Davide Di Ruscio. Automated fruit recognition using efficientnet and mixnet. *Computers and Electronics in Agriculture*, 171:105326, 2020.
- [120] Apache storm. <https://storm.apache.org/>. Accessed: 2020-12-02.
- [121] Apache spark™ - unified analytics engine for big data. <http://spark.apache.org/index.html>. Accessed: 2020-12-02.
- [122] Dhruba Borthakur et al. Hdfs architecture guide. *Hadoop Apache Project*, 53(1-13):2, 2008.
- [123] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. 2004.
- [124] Sergio Ramírez-Gallego, Alberto Fernández, Salvador García, Min Chen, and Francisco Herrera. Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with mapreduce. *Information Fusion*, 42:51–61, 2018.
- [125] Ikram Remadna, Sadek Labib Terrissa, Ryad Zemouri, Soheyb Ayad, and Nouredine Zerhouni. Leveraging the power of the combination of cnn and bi-directional lstm networks for aircraft engine rul estimation. In *2020 Prognostics and Health Management Conference (PHM-Besançon)*, pages 116–121. IEEE, 2020.