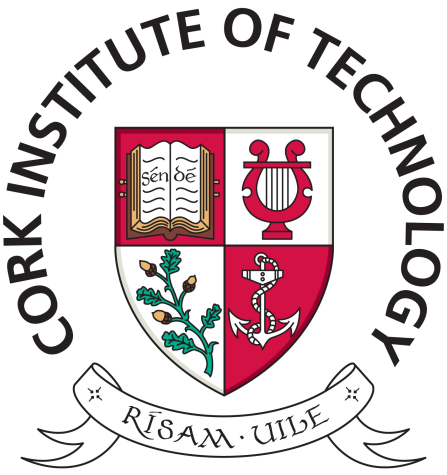


Comparative Cardiovascular Disease Prognosis Machine Learning models for Sub-Saharan African Population

by

Nirmal Pregassame

This thesis has been submitted in partial fulfillment for the degree of Master of Science in MSc in Artificial Intelligence

in the
Faculty of Engineering and Science
Department of Computer Science

May 2021

Declaration of Authorship

I, Nirmal Pregassame, declare that this thesis titled, 'Comparative Cardiovascular Disease Prognosis Machine Learning models for Sub-Saharan African Population' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for an masters degree at Cork Institute of Technology.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Cork Institiute of Technology or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- I understand that my project documentation may be stored in the library at CIT, and may be referenced by others in the future.

Signed: Nirmal Pregassame

Date: 17 May 2021

CORK INSTITUTE OF TECHNOLOGY

Abstract

Faculty of Engineering and Science
Department of Computer Science

Master of Science

by Nirmal Pregassame

Clinical diagnosis or prognosis is done mostly under doctor's expertise, and patients suffering from cardiovascular problems are advised to take several tests which may turn out to be long and expensive. Besides, in many cases, not all these tests contribute towards effective prognosis.

One of the fundamental notion of machine learning is data representation. This usually is done by detection of patterns or structure in the data. There is however a wide gap due to multidimensionality of data that feeds the model and performance of machine learning algorithms. My research study proposes a Comparative Cardiovascular Disease Prognosis Machine Learning models for Sub-Saharan African Population using latest dimensional reduction technique called autoencoder. An autoencoder is part of Artificial Neural Network (ANN), and this technique has been applied for dimensional reduction. Besides, a benchmark study to fully understand the efficiency of ANN techniques is done and it provides a comparative analysis of existing techniques of dimensional reduction (PCA, t-SNE). The objective is to determine the pertinence of autoencoders for cardiovascular incidents prediction. Lastly, the results will show how feature reduction leads to the desired accuracy for predicting the CVD disease.

Keywords—Cardiovascular disease, Sub-Saharan Africa, Artificial neural network (ANN), Autoencoders, PCA, t-SNE, Machine Learning, prognosis

Acknowledgements

When I started my master thesis, combining the clinical data to AI /Machine learning models was quite challenging due to its complexity in the format. Several aspects needed to be clarified and validated to conclude this work. This was made possible by the support of some people to whom I would really like to express my gratitude.

First and foremost, I would like to thank my thesis supervisor Dr. Farshad Ghassemi Toosi, Assistant Lecturer at Cork Institute of Technology who provided me with an unparalleled guidance, was always present to respond to my questions. He consistently allowed this paper to be my own work but steered me in the right direction whenever he thought I needed it.

I want also to express my profound gratitude to Dr. Farid Boumediene, Member of Executive Committee from University of Limoges (France) who proposed this Master thesis to me. This gave me the opportunity to work with him on this very passionating and crucial subject of cardio-vascular diseases and to build an AI prognosis model. His regular support has been vital during this project. My special thanks also to Dr. Julien Magne, expert in statistics and evaluation in University of Limoges and for providing me with valuable advice on dataset.

I would also like to thank Prof. Ted Scully for teaching me the standards and guidelines for carrying out scientist research on Machine learning and deep learning. This helped me in organizing my work in the standards defined by CIT.

I would like to express my special thanks to my friends Anuradha Khara and Abid Hussain for their encouragement and participation in proofreading of my work. A very special thought of gratitude for my brothers Saravana and Gopal, more especially to my partner, Vibhuti Khara, I really felt supported by them while I was pursuing my Master degree in AI, and this also provided me the courage to complete my thesis.

Last but not the least, I would like to express my deep and sincere gratitude to my parents for giving me the opportunities and experiences that have made me who I am. They selflessly encouraged me to explore new directions in life and seek my own destiny. This journey would not have been possible if not for them, and I dedicate this milestone to them.

The master journey has been a great experience and this accomplishment would not have been possible without you all by my side...

Nirmal Pregassame

Contents

List of Figures

List of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**euronal **N**etwork
CHU	**C**entre **H**ospitalier **U**niversitaire
CHUF	**C**entre **H**ospitalier **U**niversitaire de **F**rance
CNN	**C**onvolutional **N**euronal **N**etwork
CNHU	**C**entre **N**ational **H**ospitalier **U**niversitaire
CT	**C**omputed **T**omography
CVD	**C**ardio**V**ascular **D**isease
DL	**D**eep **L**earning
DR	**D**imensionality **R**eduction
ECG	**E**lectro**C**ardio**G**ram
ECDC	**E**uropean **C**enter **D**isease **C**ontrol
EIT	**E**uropean institute of **I**nnovation and **T**echnology
EMD	**E**mpirical **M**ode **D**ecomposition
GMM	**G**aussian **M**ixture **M**odel
IBM	**I**nternational **B**usiness **M**achine
ICAIH	**I**nternational **C**onference on **A**rtificial **I**ntelligence for **H**ealthcare
ICT	**I**nformation and **C**ommunication **T**echnology
IMF	**I**ntrinsic **M**ode **F**unction
KNN	**K**-**N**earest **N**eighbours
LR	**L**ogistic **R**egression
LDA	**L**inear **D**iscriminant **A**nalysis
ML	**M**achine **L**earning
NCD	**N**on-**C**ommunicable **D**isease
NN	**N**euronal **N**etwork

PCA	**P**rincipal **C**omponent **A**nalysis
RHF	**R**ight **H**eart **F**ailure
RTF	**R**andom **T**ree **F**orest
SMOTE	**S**ynthetic **M**inority **O**ver-**S**ampling
SSA	**S**ub **S**aharan **A**frica
TAHES	**Ta**nve **He**alth **S**tudy
t-SNE	**T**-distributed **S**tochastic **N**eighbour **E**mbedding
SVM	**S**upport **V**ector **M**achine
US	**U**nited **S**tate of America
VM	**V**irtual **M**achine
WHO	**W**orld **H**ealth **O**rganization

Dedicated to my dear parents, my dear brothers Gopal, Saravana and my dearest partner, Vibhuti...

Chapter 1

Introduction

1.1 Motivation

In today's era, many academic efforts and companies are getting involved in AI initiatives like for instance IBM has developed Watson for several health applications, and several start-ups are addressing all possible aspects of the health continuum. My research study, is in same lines and focuses on building predictive model using an autoencoder a new technique of artificial neural network. This work is aimed at exploring capabilities of classification techniques in extracting useful features for the representation of Cardiovascular disease (CVD) patients from Sub-Saharan Africa. The objective is to propose a robust Machine Learning model for prediction of disease. Autoencoder is basically used for dimensionality reduction, and my model is documented and compared with principal component analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (tSNE) techniques to validate its performance and efficiency.

The primary motivation behind this endeavour is that it would not only facilitate the identification of "at risk" patient but also enable a timely medical intervention so that patient can be supported. The patient can receive appropriate treatment and the doctor would be aware of the patient's medical situation much ahead of time. A greater lasting goal of this work is to assist medical experts in their practices for prognosis of cardiovascular disease, to encourage them to adopt new methods via persuasive Artificial Intelligence (AI) technologies and to empower doctors in creating significant change in their approach for CVD patients.

1.2 Executive Summary

Many studies have been carried out for detecting possible heart disease using various mathematical algorithms and recently through Machine Learning and Deep Learning approaches. However, those studies were only done for the patients of the developed countries. Unfortunately, no detailed studies have been performed for Sub Saharan African (SSA) population.

In this context, since 2015, the University of Limoges (France) has established a cohort study composed of SSA patients mostly based in Benin from whom clinical data have been regularly taken. Those data were taken every year for each subject. Out of them, some suffered from various cardiac disease. The primary intention of the University of Limoges (UOL) was to use classical mathematical model to make a prediction of heart disease for SSA population. The chosen classical method however had severe limitations in predictive analysis. UOL and it research department thereafter turned toward AI techniques as literature has shown that AI has proven its efficiency for such prediction.

My research worktherefore is precisely on the use modern AI techniques to build this SSA cohort prediction model for Cardio-Vascular Disease (CVD). The intent is to use a different approach of Machine Learning (ML) classification for making prediction. The project involved regular meetings with experts (doctors and statisticians) of UOL as well as my AI supervisor in CIT. Several sessions were planned to brainstorm and integrate ideas to define the specificities for this model.

Following initial stage of meetings, it appeared that in order to make a proper classification model, it was important to, firstly, address the challenges mostly related to the dataset itself. The challenges encountered were:

- The high number of parameters (high dimensionality)
- The low number of subjects: around 1964
- The cohort data sampling consistency: Data could not have been taken for all patients for each year since the track of some of them were lost or new patients were added in the cohort in the meantime.
- Missing data: Some values of clinical data are missing depending of subjects or the years
- Imbalanced data: The ratio between healthy patients and sick patients is disproportioned
- Noisy data: data patterns between sick people and healthy people is blurred

To address each of these challenges, the dataset needed to be modified and reworked in first place so as to ensure that built AI model would provide optimal results. Various techniques have been applied to improve the efficiency of the model. Typically, data transformation, data augmentation, feature reduction, cross validation techniques have been applied as main procedures to improve the predictive model classification.

The final model is the result of the different tests of those features techniques specified above. The accuracy reached at the end was 96% and the F1 scores is 96% as well.

As such, the performance of the model is very high, and this result is very encouraging for developing further model such a regression model to predict the cardiac accident occurrence.

In order to address each of these challenges, the dataset had been modified and reworked to give the best result in the built AI predictive model. Various techniques have been applied to improve the efficiency of the model. Typically, data transformation, data augmentation, feature reduction, cross validation techniques have been applied as main procedures to improve the predictive model classification.

The final model is the result of the different tests of those features techniques specified above. The accuracy reached at the end was 96% and the F1 scores is 96% as well.

As such, the performance of the model is very high, and this result is very encouraging for developing further model such a regression model to predict the cardiac accident occurrence.

1.3 Contribution

The medical team in CHU France have a big database with primary data collected on the SSA population. This part of the world has faced an increasing number of death due to heart problem. The doctors would like to build a tool using the AI technology so that they can predict the disease in the sensitive population before the symptoms arises in any individual and on studying his /her health history.

My research work and proposed model developed using the primary data will help French/African medical team to analyse data and predict heart disease in individuals. The contribution of my master's thesis has three main outputs as follows:

1. Using the raw primary data from SSA population, I have modelled this data using AI technology to propose a tool that will be able to predict the Cardiovascular disease in any fresh data that will be inserted into the system.

2. Testing of new AI technique, called autoencoder in Cardiovascular disease prediction, to evaluate its strength and potentiality in term of feature extraction, and comparing it to existing classical ML techniques. All this would help assess if the model can be further improved.

3. My work has recognised credibility and value in field of medicine as the entire work has been done on primary real data. This data set has been collected directly from population in the field, records of the parameter are real without any imaginary values. Besides, the experimental modelling of clinical data and techniques have been done wisely to get optimal results.

1.4 Structure of this Document

This research study focusses on building a ML model to predict cardiovascular accident using SSA population clinical data and evaluates a novel system for reducing dimensionality using autoencoder. This document is divided into 5 chapters.

The first chapter introduces the current study presenting the motivation for this works, its summary, its contribution and the structure of this document. The second chapters deal with the background of this study: The thematic area, the project scope as well as the thematic reviews and literature. The third chapter introduces the comparative study by explaining the problem definition, the objectives, the functional and non-functional requirements. The fourth chapter emphases the implementation of the work: The framework, the data processing methodology, the Feature Dimensionality Reduction Methodology, the classification methodology and the architecture are the main topics of this part. Finally, the fifth chapter gives a conclusion and describes the future work.

Chapter 2

Background

We are living in the age of digital, data and algorithms, in which Artificial Intelligence (AI) and more precisely machine learning (ML) systems are integrated into the system. Undeniably, AI technology is being embraced by various fields like marketing, commerce and education. Health care is another field to be revolutionized by AI techniques [1] [2] [3] [4].

With the rapid increase in data generated and methods to analyse data, there has been huge interest in using data trail to improve the wellbeing of population. Recently, data science has provided performance model in different domains—e.g., computer vision, text analytics, and speech processing, etc. and algorithms have begun to influence healthcare—a field that has traditionally been impervious to large-scale technological disruptions[5]. Several start-ups companies are addressing all possible aspects of the health continuum [6] based on massive digital data generated.

2.1 Thematic Area within Computer Science

2.1.1 Classification predictive models, Feature extration and Autoencoders

To construct a predictive or prognosis model with ML, there are a panel of techniques in of data science, and these are mainly as methods categorised under - 1. Supervised learning, 2. Unsupervised learning and 3. Reinforced learning. ML algorithms are capable of learning and training itself from input data without human intervention. These learning tasks can be of type – mapping of input data to output results, learning the hidden structure from the unlabelled data and learning from instances that are given.

The Figure 2.1 provides a schematic representation of the three cases -

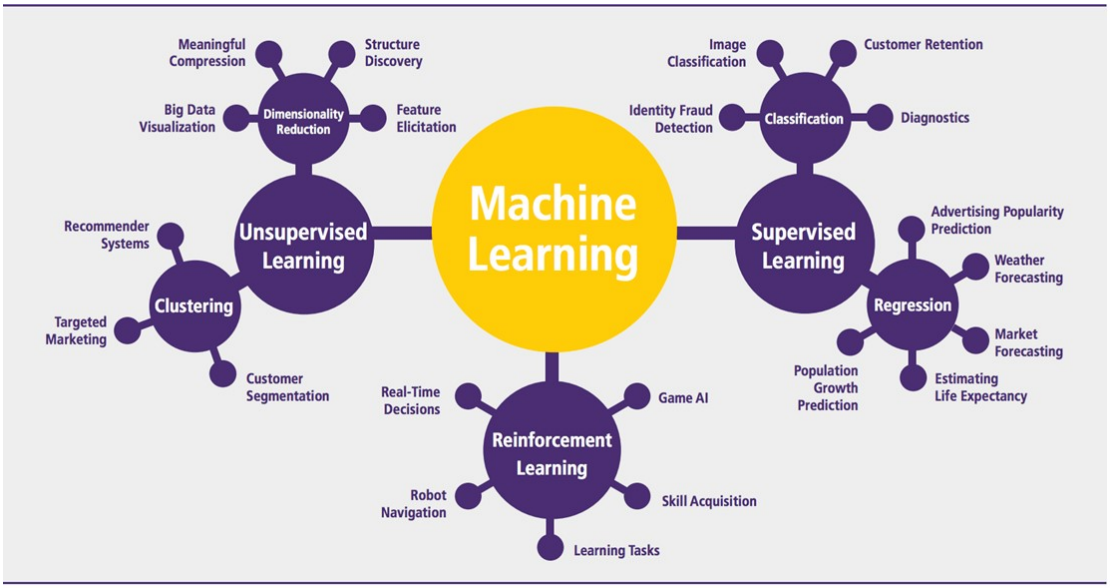


Image source from https://wordstream-files-prod.s3.amazonaws.com/s3fs-public

FIGURE 2.1: Machine Learning Algorithms Overview

The unsupervised learning has only an input variables (X) without any corresponding output variables. In this method the computer program uses unlabelled data to train the model for underlying data structures. The techniques in this category are association, clustering and dimensional reduction.

In the reinforcement learning, machine learning algorithm lets the model to decide on the best next action that it should consider while considering its current state. In fact, it works on the principle of rewards or suggestions to maximise the learning. The optimal learning often happens through trial and error like in the video games. It's learning by doing and understanding the environment to better perform.

In supervised learning approach, the algorithm uses labelled training data to learn the mapping function from the input variables (X) to the output variable (Y) and relation between X and Y can be represented as $Y = f(X)$. There are two types of techniques under this broad category: namely Classification and Regression.

Classification technique implies predicting the outcome of a given sample when the output is in the form of categories. Examples are like male and female, sick and healthy. In classification, the output data can be bi-class or multi-class. The model trains first itself from the input data, then later the same model once trained can categorize any new observations on its own. On the other hand, regression has its output prediction in form of real-valued labels like height of person, amount of rainfall in an area, amount of sugar in blood etc.

To classify, the algorithms use a set of rules in calculations or problem-solving operations. Machine learning uses learning algorithms such as: 1. Linear Classifiers: Logistic Regression, Naive Bayes Classifier, 2. K-Nearest Neighbour, 3. Support Vector Machines, 4. Random Forest etc to name some. Each algorithm has its own specificity and function like for instance k-nearest-neighbours (KNN) algorithm in its working method takes the proximity as a proxy for 'sameness' [7]. KNN picks up labelled points and trains it-self on the labelling, and later to label a new point it will look for labelled points that are close to the new point.

The above stated supervised learning algorithms works well only if the data set with low dimension. When there is a dataset with high dimensionality [8], a severe challenge is seen in analysis of data. In this case, a second method of data science needs to be considered to carry out efficiently the analysis. This is called dimensionality reduction (DR) and it has two main techniques namely feature extraction [9] or feature reduction methods of in data science. dimensionality reduction [10] [9] is important area, where many approaches have been proposed.

DR is an important part pre-processing step of predictive model preparation [11]. This is because with the increase in number of features, the model starts becoming complex. In fact, a ML model that is trained on a large number of features starts to get increasingly dependent on the data it was trained on. This in makes the model overfitted and poor performer on real data. This beats the purpose. DR therefore can be very useful in removing irrelevant and redundant data, reduce variable while enhancing learning accuracy of the model, improve result comprehensibility.

In the category of unsupervised learning a recent technique of Artificial Neural Network (ANN) has received lot of acknowledgement, and this technique is known as autoencoder. In fact, ANN is a conceptual framework for executing AI algorithms [12], it mimics of the human brain—an interconnected network of neurons, in which there are weighted communication channels between neurons [13]. An autoencoder is mostly used in dimensionality reduction as it learns any representation of data by training the network to ignore the irrelevant data or to ignore signal "noise", they are very useful in DR. These have successfully been applied in several practical examples like image classification, pattern recognition, anomaly detection, and data generation. There are several kind of autoencoders. One of the interesting recent study done by Zhang and his colleagues [14] for detection of breast cancer, autoencoders are applied to learn concise features from gene expression profiles from the vast database. This model showed effective outcome for their prognosis model.

Both ML/DL techniques have shown their effectiveness in case of supervised and unsupervised tasks. These techniques have demonstrated being suitable to extract useful knowledge from medical big data [15].

2.1.2 ML in Healthcare

AI are an active research area, and several promising papers are published with ML/DL in healthcare and medical analysis [15]. Besides several recent ones have stressed on the use of autoencoder in healthcare. Clinical medicine has no doubt emerged as an exciting application for ML/DL models, and these models have already achieved efficient results and "human-level performance" in clinical pathology [16]. There are however several challenges as well that this field in terms of data set that can be very diverse and not limited to superficial representation, there can be many parameters to be taken into account and these can add a layer complexity. Nonetheless, early prediction and diagnosis of diseases from medical data are one of the exciting applications of ML.

Various studies have highlighted the potential of using predictive healthcare for the timely treatment of diseases. Currently, we are undergoing a pandemic crisis with Covid19 and literature shows that ML/DL can help accelerate solutions and minimize the impacts of the virus. Studying the heterogeneous sources of data can help the healthcare AI system to be more robust. In fact, the four major applications of healthcare that can benefit from ML/DL techniques are prognosis, diagnosis, treatment, and clinical workflow. Applications of ML in Healthcare service generates a large amount of data and information in the daily basis and using the newer techniques of AI can effectively analyze this data for actionable insights.

2.1.3 Data science and AI

While comparing traditional statistical methods to newer AI methods, the striking difference between the two is that statistics can proficiently help in understanding the relationship between a list of variables, and the AI contributed to identifying or remodelling the features from data to perform predictions. AI methods can benefit largely the statistical methods by providing tools and techniques to understand patterns from large, complex and heterogeneous data.

The higher-level field under which the above stated fields can be categorized is data science. In fact, data science is "concept to unify statistics" an amalgamation of data mining, AI and big data. A branch of data science, which is data mining, is a field combining computer science and statistics with an overall goal to extract information.

Artificial intelligence (AI) is categorized under data mining and are said to be intelligence of machines.

2.2 Project Scope

Project Tanve Health Study (TAHES) was initiated in the SSA. This project is a result of continued partnership between France and African Ministries of Health. TAHES is a pilot study by Limoges University in France under the guidance of medical doctors and university research scientists.

In fact, SSA has been facing a growing burden of non-communicable diseases (NCD) due to epidemiological transitions with increasing urbanization and changing lifestyle; and among those disease, a huge prevalence of cardiovascular diseases has been noted [17]. The Figure 2.2 shows an high preponderance of CVD in West Africa:

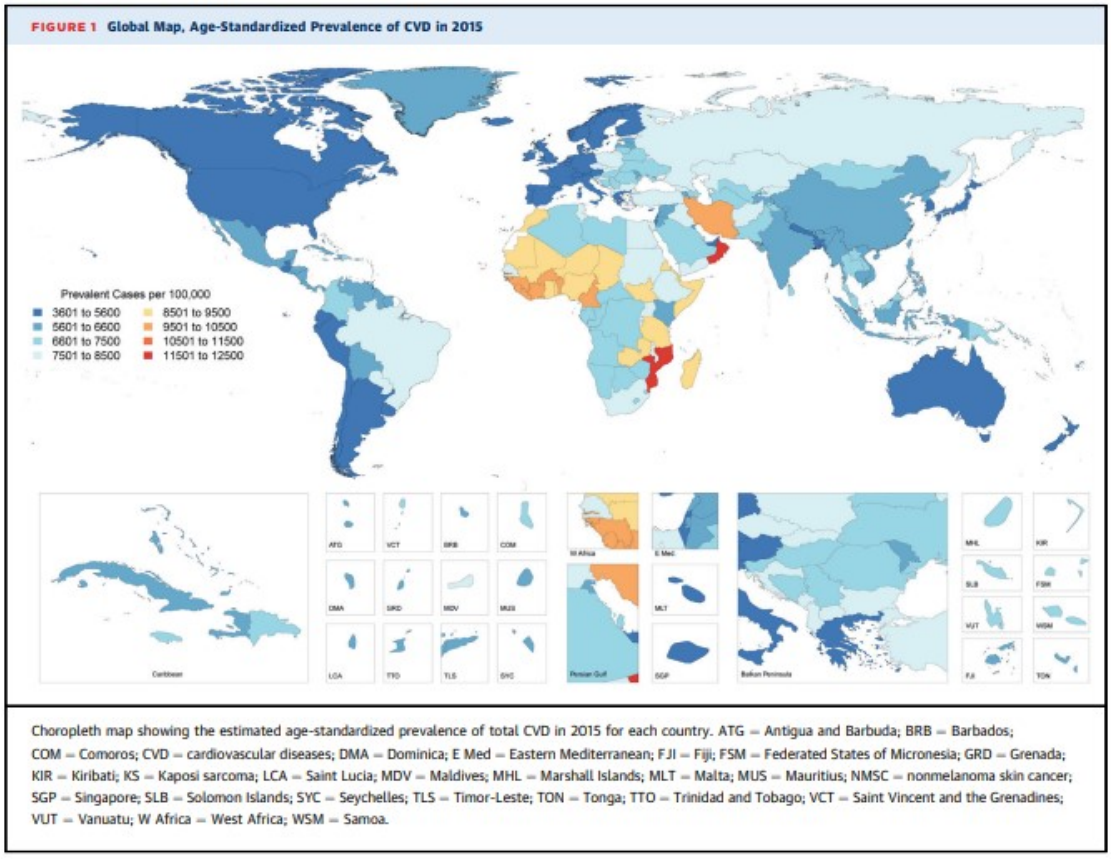


Image source from Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015, by Roth 2017 [18]

FIGURE 2.2: Global Map Age-Standardized Prevalence of CVD in 2015

The TAHES project started in February 2015, and thereafter a PhD student worked for initial data collection from people aged 25 years or above living at Tanv'e and

D´ekanmey, two villages of Agbangnizoun in the south-west of Benin. Amidou's research demonstrates that risk of CVD in Benin is high in the population and this doesn't limit to disability and death of people but has a bigger economic aftermath.

The individual who had CVD and survived are prone to high risk of disability, especially for stroke. This chain of medical health problem has repercussions both on the individual and on society. Besides, the constraints for those around them are both psychological and material. The society has to undertake the need to train new skills to fill the function that can no longer be exercised, it has to arrange special positions or schedules for people with disabilities. Looking at this balance sheet of consequences, control of CVD in SSA is therefore based on the global strategy for combating non-communicable diseases NCDs and other strategies by WHO for health promotion (WHO, 2005).

The aim of this TAHES cohort constitution therefore has taken a new turn in development of prognosis tool that can help in preparedness activities for cardiovascular incidence and its prevalence in the population. The innovative dimension of prognosis tool will help the medical unit be ready much before the incident may happen. In past, statistical tools and mathematical model were developed by the University of Limoges to analyze cohort's data, these are however limited for the predictive analysis.

The joint venture therefore calls on for development of the model using AI techniques. Unlike regular statistics, Machine-learning (ML) in this case would provide an interesting alternative approach to make prediction modelling and solve current limitations. The predictive model can be developed using the algorithms for learning all complex and non-linear interactions between parameters by minimizing errors between predicted and observed result [19]. The field of AI has a new paradigm for analyzing large biomedical datasets in various domains of health: New advanced analytics based on AI and more specifically on ML and DL are providing a great opportunity in recent years for improving the quality of healthcare [19] and cardiovascular risk. With help of this ML model, doctors would be able follow the CVD in ageing population in Benin, adapt the treatment and medication, and be aware of any immediate cardiovascular accident.

The ML techniques can identify latent variables inferred from other variables and cannot be observed [20]. As such, this powerful method can enhance the TAHES cohort data. Indeed, The TAHES Project has gathered values of multiple epidemiologic variables from the cohort constituting a big collection of biomedical datasets. Those data could make a fantastic support for training a ML for making prognosis on cardiovascular disease. To date, there has been no large-scale investigation applying machine-learning for prognostic assessment in the general population, using routine clinical data. The aim of this study was to evaluate whether ML can improve accuracy of cardiovascular risk

prediction within a large general primary care population. We also sought to determine which class of ML algorithm has highest predictive accuracy.

TAHES has primary data and there are many parameters that have been considered for constructing the data set. In fact, the large scale of parameters in the TAHES cohort had provided challenge to use directly a ML model for making a prediction since ML algorithms cannot handle large number of features sometimes. As such, dimensionality reduction [10] becomes a key step for pre-processing the cohort data.

DR plays a vital role in pre-processing of complex clinical data. It helps in ensuring that important information is conveyed on reducing the number of variables of a dataset. This technique formulates the data for predictive model preparation[11]. Among the dimensionality reduction techniques[9], two classes of methods are available:

- the feature selection
- the feature extraction

Feature selection is used for filtering irrelevant or redundant features from the dataset or alternatively we try to find best subset of input features. On the contrary, in feature extraction we create new features based on the transformation of original features. The Figure 2.3 highlights the difference between the Feature Selection and the Feature Extraction:

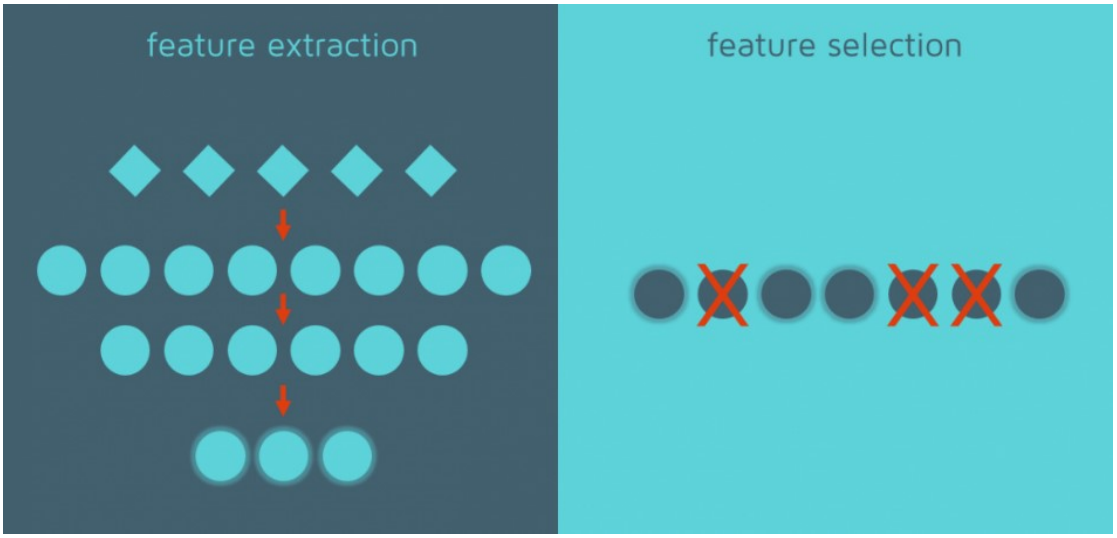


Image source from https://quantdare.com/what-is-the-difference-between-feature-extraction-and-feature-selection/

FIGURE 2.3: Feature Extraction and Selection Figure

In this study, the focus has been done on feature extraction technique rather than the future selection. Indeed, the project is dealing with clinical data which are in essence

extremely noisy: In the comparative study done by Khalid in 2014 [21] on medical data (ophthalmologists diseases), the two reduction techniques (feature extraction and feature selection) were compared. His work elaborated that Feature extraction seems to be more efficient than feature selection techniques in medical data as they are very noisy (i.e it has large amount of meaning less information and that needs to be cleaned) and Feature Selection does not propose a solution to reduce those noises. As such, dealing with the clinical data, the current study is implementing exclusively feature extraction solution for dimensionality reduction.

Among the feature extraction techniques, the study focuses on standard techniques like Principal Component Analysis (PCA) (Figure 2.4) and t-SNE (Figure 2.5) [8]. But in addition to them, the new technique as the ANN Autoencoders [15] (Figure 2.6) is also evaluated as well. The objective is effectively to benchmark this novel feature reduction technique in order to propose robust models of prediction in Healthcare.

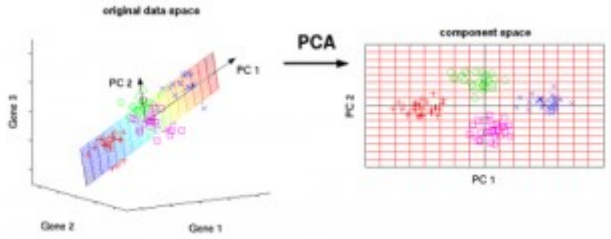


Image source from http://www.nlpca.org/pca_principal_component_analysis.html

FIGURE 2.4: PCA Representation

The concern to evaluate Autoencoder as a feature extractor comes indeed from the three following studies conducted in medical field which illustrated that Autoencoder is of high value:

- Zafar et al. in 2016 [9]: The proposed method used the autoencoder to predict risk factors for hypertension for a vulnerable demographic subgroup of patients. His model results showed that encoder representation learning outperform classic models.
- Miotto et al.in 2016 [22]: Autoencoders were applied for feature learning and representation of large scale electronic health records of individuals related to several diseases including schizophrenia, diabetes, and various cancers. Their model shoed high level of prediction, and thus showing the efficiency of Autoencoders.
- Zhang et al. in 2018 [14]: Autoencoder has been applied for detection of breast cancer, the model was able to learn concise features from gene expression profiles from the vast database systems and provide high efficiency predictive model.

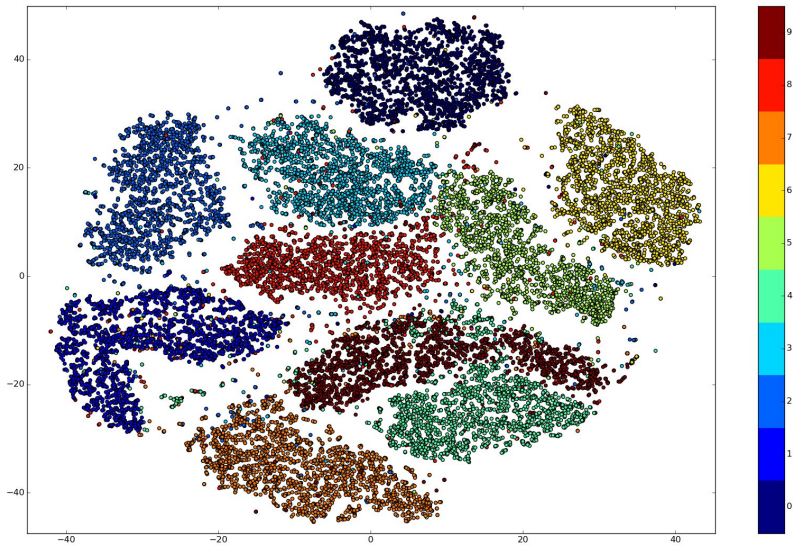


Image source from https://indico.io/blog/visualizing-with-t-sne/

FIGURE 2.5: t-SNE Visualization

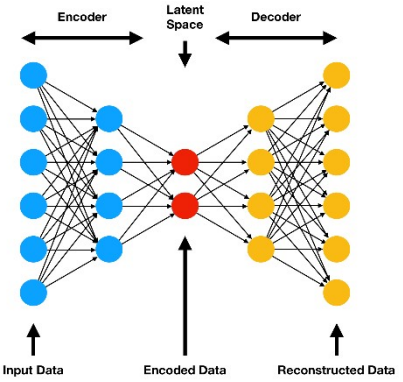


Image source from https://www.compthree.com/blog/autoencoder/

FIGURE 2.6: Autoencoder Architecture

These three studies inspired in my current research to apply the Autoencoder and test it along with other traditional solution for reducing the data dimensionality.

The Autencoder techniques as well as the other feature extraction technique will further be associated with a Machine Learning classification methods in order to build a prognosis of cardiovascular disease event. ML classification tools will be used as control group for testing the feature extractions above mentioned models: 3 classifiers are used:

1. Logistic Regression
2. Random Tree Forest

3. K-Nearest Neighbors.

The end goal of this research work is to propose to the University of Limoges the high performance predictive model among the benchmarked methods and demonstrate the effectiveness of AI techniques for making a prognosis of cardiovascular disease.

2.3 A Review of the Thematic Area

This current project is exploring the power of Machine Learning algorithms for cardiovascular prediction purpose based on clinical data. The healthcare domain is a very trending topic. A study from Cbinsight [23] in 2016 reveals that the healthcare "is the hottest area of investment within AI". Indeed, more than 1.5 billion dollars are spent in this sector to support AI start-up in last 5 years and 80% of this funding goes into imaging and diagnosis. On the same note, the Global Health Estimates 2016 from WHO highlights [24] that the major causes of death in the world is due to the heart disease and stroke (Figure 2.7).

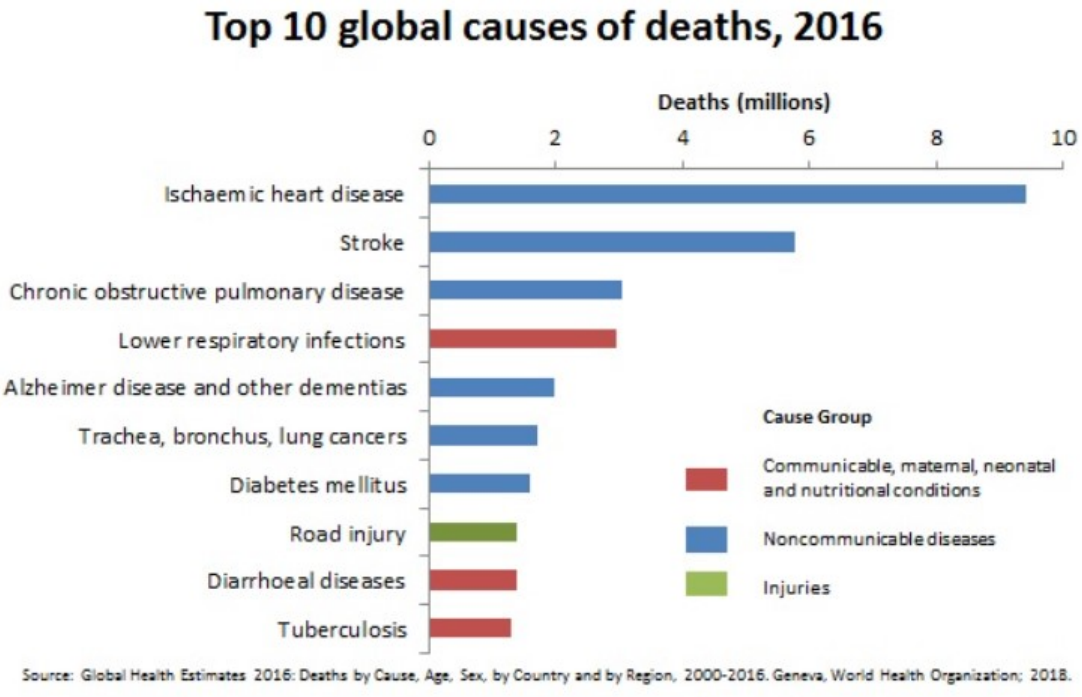


Image source from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

FIGURE 2.7: Top 10 Global Causes of Death in 2016

The huge interest in AI applied into healthcare and more specifically into cardiology emerges as various publications, exhibitions, conferences across the globe through numerous supports.

Major international conferences regrouping experts in both domain AI and healthcare/-cardiology are regularly organized. Among the countless ones:

- World Congress on Hypertension, Cardiology, Primary Health and Patient Care: In 2020, it will gather various specialists to discuss about new approach for hypertension in term of diagnosis, treatment and prevention.
- International Conference on Cardiology and Heart Failure: this session hosts mostly fellow researchers who presents thoughts, concepts and recent developments in Cardiology and Cardiovascular Medicine.
- Information Technology, Data Science and Digital Health Summit: Mostly dedicated to data scientists, this forum emphasis exchange of ideas in various AI application in particular in Healthcare.
- ICAIH (International Conference on Artificial Intelligence for Healthcare): its aims to bring together academic scientists to exchange on all aspects AI for Healthcare.
- International Conference on Artificial Intelligence, Machine Learning and Big Data: This deals with several fields particularly in industries like Healthcare; but its core field is related to AI.

Various channels are available to diffuse recent progress in AI and more specifically related to Healthcare. The mostly known are:

- The healthcareitnews, an authoritative source covering technology driving next-generation healthcare in the U.S. and the world: https://www.healthcareitnews.com/topics/artific intelligence?type=blog
- Facebook Artificial Intelligence gathers all paper, work, open source research and advancement in AI: https://ai.facebook.com/#notable-papers
- Healthcare.ai is a community with open source tools and education on machine learning in healthcare https://healthcare.ai/
- Intelligent Health AI is a youtube channel bringing global AI and health community: https://www.youtube.com/channel/UCe3uIGbukqwBYWG4XHBC-fw
- Partners Innovation is a coordinated group of medical inventors, thought-leaders, entrepreneurs and industry: https://innovationblog.partners.org/category/world-forum

Books publication on this topic are also accessible:

- Artificial Intelligence and Data Mining Methods for Cardiovascular Risk Prediction [25]. This book provides a comprehensive guide to the state-of-the-art in cardiovascular computing and highlights novel directions and challenges in this constantly evolving multidisciplinary field. The topics covered span a wide range of methods and clinical applications of cardiovascular computing, including advanced technologies for the acquisition and analysis of signals and images, cardiovascular informatics, and mathematical and computational modeling.

- Machine Learning in Cardiovascular Medicine 1st Edition [26]. Machine Learning in Cardiovascular Medicine addresses the ever-expanding applications of artificial intelligence (AI), specifically machine learning (ML), in healthcare and within cardiovascular medicine. The book focuses on emphasizing ML for biomedical applications and provides a comprehensive summary of the past and present of AI, basics of ML, and clinical applications of ML within cardiovascular medicine for predictive analytics and precision medicine. While the industrial applications of ML are nearly ubiquitous, its introduction into the medical field has been much more gradual. The landscape, however, is rapidly changing with the availability of computational power and the creation of large repositories of datasets.

- Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again 1st Edition [27] by Eric Topol. Eric Topol reveals how artificial intelligence can help. AI has the potential to transform everything doctors do, from notetaking and medical scans to diagnosis and treatment, greatly cutting down the cost of medicine and reducing human mortality. By freeing physicians from the tasks that interfere with human connection, AI will create space for the real healing that takes place between a doctor who can listen and a patient who needs to be heard.

- Machine Learning in Cardiovascular Medicine [28], Editors: Subhi Jamal Al'Aref Gurpreet Singh Lohendran Baskaran. This book focuses on emphasizing ML for biomedical applications and provides a comprehensive summary of the past and present of AI, basics of ML, and clinical applications of ML within cardiovascular medicine for predictive analytics and precision medicine. While the industrial applications of ML are nearly ubiquitous, its introduction into the medical field has been much more gradual. The landscape, however, is rapidly changing with the availability of computational power and the creation of large repositories of datasets.

- Deep Learning Approach to Cardiovascular Disease Classification Employing Modified ECG Signal from Empirical Mode Decomposition [29]. Author Nahian Ibn

Hasan and Arnab Bhattacharjee presents a method to classify multiple heart diseases using one dimensional deep convolutional neural network (CNN) where a modified ECG signal is given as an input signal to the network. Each ECG signal is first decomposed through Empirical Mode Decomposition (EMD) and higher order Intrinsic Mode Functions (IMFs) are combined to form a modified ECG signal. It is believed that the use of EMD would provide a broader range of information and can provide denoising performance. This processed signal is fed into the CNN architecture that classifies the record according to cardiovascular diseases using softmax regressor at the end of the network. The method is applied on three publicly available ECG databases and it is found to be superior to other approaches in terms of classification accuracy.

Concerning the actual project for building a prognosis tool predicting cardiovascular disease, its potentiality can be put in the perspective of being used in the markets. Indeed, different stakeholders in the industries or in public or international organization are working in the sector of healthcare to make diagnosis. My research work could be of great interest to companies that are in digital health or health related organisation were the direction is to modernise and direct toward eHealth supported application. Besides, universities and medical colleges, where the research is centred for building predictive model in health care. For instance,

1. IBM Watson Health in a joint venture with Broad Institute of MIT and Harvard is launching a research partnership to develop powerful predictive models that will enable clinicians to identify patients at serious risk for cardiovascular disease. This is a three-year project and will incorporate population- and hospital-based biobank data, genomic information, and electronic health records to build upon and expand the predictive power of polygenic scoring. My research could be a piece work to look at the framework and techniques that I have employed. They could benefit from a similar pre-cursor work as project.

2. EIT Health Is part of the European Institute of Innovation and Health and their ambition is to build a 'knowledge and innovation community'. EIT brings since 2015 their expertise in the sector of the Health. They're aim is to bring together the Education, the Business and the Research works together to create an optimal environment for Innovation. One of their key action is to finance the promising research to develop the most promising solutions into real-world commercially viable products.

3. CDC Centre for Disease Prevention and Control has mission is to strengthen Europe's defences against infectious diseases. CDC publishes numerous scientific and

technical reports covering various issues related to the prevention and control of Transmission (medicine) communicable diseases. Their literature shows their participation in using ICTs, innovative tools for Disease Prevention and Control. AI based application and in particular the predictive that I have developed in my research could potentially interest them applying the model for other communicable diseases.

4. Caption Health is based in US and has emulating expertise with AI. They are in field of healthcare and their AI software empowers healthcare providers with new capabilities to acquire and interpret ultrasound exams. My research could be in the same direction as their mission and can bring the predictive model developed from primary data rather different from their speciality which is in ultrasound data interpretation.

5. The university of Limoges and the partner university, who are the pioneer of this project are very interested to continue scaling-up my research project. They are aiming to introduce me and my research to cross-sectional teams and develop model for diagnosis. Besides, CNHU: Centre National Hospitalier de Cotonou and CHUF: Centre Hospitalier Universitaire France would directly be able to test the model in real and check on the benefits from it.

2.4 Current State of the Art

There have been developments in cardiology over the last century [30]. Several changes have happened since the time the initial practices and today like Electrocardiogram (ECG) in 1903, cardiac catheterization in 1929, heart and lung machine and first animal models in the 1950s, minimally invasive surgeries in 1958, diagnostic imaging by Magnetic resonance imaging (MRI) in 1980s, and then later another wave of practices came into force like cardiac image acquisition techniques, predictive in silico cardiac models [31], realistic image simulations [32], real-time patient monitoring [33], and large-scale cardiac databases [34]. With the digitalisation of the detection and diagnosis of the disease, there has been generation of huge amount of digital data, and this has paved way to more new and recent practices of using AI in CVD prediction, in the field of data science. In fact, the cardio-vascular prediction models are created done by using machine learning [35], and they can be promising in a follow-up of cardiac health of the patient by the doctors.

Recent studies on cardiovascular disease using machine learning have shown encouraging results like [36] have created heart disease risk prediction model. [35] developed a cardiovascular risk prediction system using fuzzy K-nearest neighbour (K-NN) classifiers.

These models provide a representation of patterns in data. In some domain reinforcement learning or unsupervised learning by the model is a critical step in improving the performance of machine learning algorithms. This is due to multidimensionality of data. Another critical gap listed is the size of the training sample obtained from patients. The sample may not be large enough to represent all the variation across patients and reflect the complexities of health problems.

Several other works have been carried out in finding efficient methods of medical diagnosis for diseases [6]. The idea is generally to refine models for making predictions on patient status using clinical data. A study was done on Neural Network and Decision Trees to build "Intelligent Heart Disease Prediction System" considered only 15 input parameters [20]. One suggestion of improvement out of this work was precisely to consider more parameters to refine the result. Besides, an extraction of relevant patterns is considered essential to get an acceptable model for the heart disease prediction.

With the recent progress in artificial neural networks, there are better solutions noticed to several pattern recognition problems and classification tasks. One of the inspirations for the present research study was a from the research done on autoencoder by team at Mount Sinai Hospital in New York City in 2015. The team used autoencoder in their work and without any expert instruction, they were able to discover hidden patterns in the hospital data that can predict the patient's future with respect to the development of certain diseases and with high accuracy. An autoencoder is an artificial neural network which is trying to reconstruct the input in the output while being trained and belongs to the class of unsupervised learning algorithms [38]. It finds patterns in a dataset by learning the internal structure and features of data.

Here below are some comparative analysis of four recent works done in the field of ML, healthcare and more precisely in the CVD with my own research work:

"Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 uk biobank participants," [39] and my research study

Similarity:

In a study conducted by Aala and her team, they focused to develop a predictive model for CVD using automated ML algorithmic tool. Similarly my research work is on development of a predictive model on the cohort THAES.In a study conducted by Alaa et

al.(2019) the focus was to develop a predictive model for CVD using automated ML algorithmic tool. Similarly, my research work is on development comparative Cardiovascular Disease Prognosis Machine Learning models for Sub-Saharan African Population.

Difference :

Alaa et al.(2019) study have employed algorithmic tool that automatically selects and assigns ML modelling in data imputation, feature processing, classification and calibration algorithms). Secondly the study compares the developed model with a well-established risk prediction Cox PH model. Cox proportional hazards (PH) model is based on familiar risk factors (i.e, age, gender, smoking status, systolic blood pressure, history of diabetes, reception of treatments for hypertension and body mass index). The CoX PH model has 473 available variables. Another difference is that the model has 23,604 participants without CVD at baseline and with 473 available variables.

In my research work, I have developed a comparative Cardiovascular Disease Prognosis Machine Learning models for Sub-Saharan African Population by using the latest classification and feature reduction technique like autoencoder of artificial neural network. I have done a comparative analysis with the classical classification methods of ML algorithms like t-SNE and PCA to get the most fitting and viable model for cohort TAHES. In the raw dataset, I had much smaller dataset with 37 different variables. As the number of records were relatively smaller therefore I used feature extraction techniques to get 26 most relevant variables to analyse the patterns within the data. My dataset is much smaller is size with 1900 participants and 37 variables.

"An effective approach for ct lung segmentation using mask region-based convolutional neural networks," [40] and my research study

Similarity:

Hu et al. (2020) and my research work, both have used AI and more specifically ANN in development of a AI model for the lethal disease treatment.

Difference:

Hu et al. (2020) work is more in the lung cancer diagnosis. In their study, the team have used computed tomography (CT) in the diagnosis of diseases for lung cancer. The method used before entailed is identification of lung regions in the CT images and then the regions were marked manually by a specialist. Later, the lung regions were segmented for clinical diagnoses. Hu et al. (2020) study proposes an automatic segmentation of the lungs in CT images, using the Convolutional Neural Network (CNN)

Mask R-CNN, to train a model for lung region mapping, combined with supervised and unsupervised machine learning methods (Bayes, Support Vectors Machine (SVM), K-means and Gaussian Mixture Models (GMMs)).

My research is more orientation toward CVD prediction model. In my research, I have used classification algorithm like logistic regression, random tree forest and KNN which falls under machine learning to train the model. Besides, I have done a comparative analysis to find which model would be optimal SSA dataset and with high accuracy of prediction. In my research I haven't use computed tomography technique, but it could be useful for extending the research from prognosis and to add the dimension of diagnosis.

"Effective heart disease prediction system using data mining techniques" [41] and my research study

Similarity:

The work from Singh et al. (2018) and my study are both on developing an effective heart disease prediction model using neural network of AI. These two studies have considered medium set of relevant variables for predicting the risk level of heart disease, and in same lines both studies have numerical and categorical data, therefore processing, cleaning and filtering are applied on records to remove irrelevant data from the database.

Difference:

Singh et al. (2018) have a system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction, and have developed a system that predicts the likelihood of patients in getting heart disease. Their work highlights significant knowledge gain in terms of studying the relationships between medical factors related to heart disease and patterns. Singh et al. (2018) were able to study this by using a multilayer perceptron neural network with backpropagation as the training algorithm. My research has a bigger dimension with 37 variables, which were filtered to 26. Then pre-processing step of feature extraction have been applied. Thereafter dimensionality was reduced for PCA, 16 variables were obtained. In my study, I have employed ML algorithm to train Cardiovascular Disease Prognosis models for Sub-Saharan African Population.

"Unsupervised feature extraction with autoencoder: for the representation of parkinson" [42] and my research study

Similarity:

In a study conducted by Kazak, (2019) the author focuses on establishing a comparative study of feature extraction techniques with other method such as PCA. This study has a similarity with my research work as one of the goal of my work is to check the performance of autoencoder as a feature extraction to other traditional approach (like PCA).

Difference:

Kazak, (2019) research paper is a master thesis. The subject of interest covered by this work is applying the autoencoder algorithm for Parkinson's disease patient data and comparing feature extraction capabilities with other methods. Kazak, (2019) has examined the effectiveness of an autoencoder in feature extraction, and the algorithm has been examined in predicting scores according Parkinson's disease rating scale. Kazak, (2019) illustrated that by forcing an autoencoder to learn a compressed representation of patients it can be trained to find patterns in other patients' data. My research has a different subject and it is to develop a comparative Cardiovascular Disease Prognosis Machine Learning models for Sub-Saharan African Population. I have used 37 parameters as inputs variable. Another point is that the support of my study for making the predictive model is the clinical data which is not related to voice records but mostly non-categorial data. Finally, my model of autoencoder is not the same. I have employed a simpler model and then done a comparative analysis for get a robust and more accurate model for the SSA set of data.

"Deep learning approach to cardiovascular disease classification employing modified ecg signal from empirical mode decomposition," [29] and my research study

Similarity:

Hasan and Bhattacharjee, (2019) study and my research work are both on developing model on Cardiovascular Disease by using the classification techniques.

Difference:

Hasan and Bhattacharjee, (2019) have employed Modified Electrocardiogram ECG Signal from Empirical Mode Decomposition in their study. They illustrate that firstly a

modified ECG is formed by summing the Intrinsic Mode Functions (IMF) signals and then they have developed one dimensional CNN network to make the model learn about the features of the modified signal for classification purpose. They also have compared the method with other approaches to training the model with different combinations of IMF signals. Hasan and Bhattacharjee, (2019) assessed their method by evaluating it on three publicly available databases. This approach assessment makes their method robust and capable of classifying a broader range of cardiovascular diseases. In contrary, I have developed a prognosis model by using the machine learning algorithm on primary dataset. I have made 9 models and taken a comparative approach to provide an optimal model using the new technique of ANN auto-encoder.

"Can machinelearning improve cardiovascular risk prediction using routine clinical data?" [43] and my research study

Similarity:

Both the studies are on developing an effective heart disease prediction model using ML. The two studies have used clinical data to make their prediction and the number of features considered is approximatively at the same range (around 30). Some of the them classification on the benchmark are same just like the Random Tree Forest or the Logistic Regression.

Difference:

The work of Weng et al. is using a large database (more than 375,000 individuals) from UK population. Their main focus is to see the efficiency of AI prediction toward commonly used statistic. As such, ML and NN classifiers are studied. However, in their study, preprocessing data is limited. Typically, Dimensionality Reduction techniques are not implemented. This differs from my research study: The main focus is to estimate the performance the ANN, Autoencoder in term of Feature Extraction techniques relatively to other regular techniques and no comparison is done with the statistics. Moreover, my research is exploiting clinical data from SSA (and not from UK): The sample is not as proficient as in developed country and the base population is not the same.

"Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer," [14] and my research study

Similarity:

The work of Zhang and his colleagues is to propose a prognosis model which is like my research study. Moreover, the future extraction methods used in their work are PCA and Autoencoder and those are precisely part of my comparative study as well.

Difference:

Zhang et al, (2018) study has employed stacked feature extraction to train the model method by using first the PCA and then the Autoencoder for breast cancer prediction. Their method is well justified as they have large number of features (around 76 parameters). In my research, feature selection techniques have not been applied and just feature reduction technique has been used since the number of futures in my study is more limited and compared to their study I have relative less features (around 30). Their work focusses also on breast cancer prognosis using genes signatures as datasets to train their model. For my research, the data used are clinical samples just like blood pressure or heart pulse to make a prognosis of cardiovascular accident. Finally, Zhang et al, (2018) propose a more general way of learning features by integrating feature selection and feature extraction methods with several deep learning techniques like NN classifier to make its prediction after the Autoencoder. For my work, the classification method used was the ML techniques such as Logistic Regression, Random Tree Forest and KNN.

Chapter 3

Cardiovascular Disease Prognosis for SSA population

3.1 Problem Definition

Cardiovascular disease, mainly represented by strokes or Right Heart Failure (RHF), is the leading cause of death worldwide with 17.7 million deaths per year. Low- and middle-income countries pay the highest price for these diseases, with more than 3/4 of deaths most often occurring before the age of 70: According to the Global burden of Diseases in 2017, the highest prevalence is in West Africa. This situation is critical, since most of the studies for cardiovascular disease has been done in developed countries and very few data relative to this subject are available from Africa. Given this context and to respond to this issue, the TAHES project has been initiated in 2015 by the University of Limoges: Clinical data related to cardiovascular disease are collected in SSA, more particularly in Benin from a cohort study under construction. If the standard approach through statistics and mathematical models were adopted by the University of Limoges until now, the consideration of applying AI and Machine Learning techniques to refine the predictive model is examined. Therefore, for the continued progression of this work and to bring in data mining techniques for better predictive models, the current study was undertaken as my Master's thesis at CIT.

3.2 Objectives

The objective of my work is to propose a model to predict efficiently the risk of heart disease. The intent is also to extend the work of the TAHES project by applying fuzzy

learning models to evaluate the type of cardiovascular disease. This work will be achieved by reducing the number of attributes in the TAHES dataset using feature extraction methods. Among those methods, the Autoencoder, an ANN, seems promising. The concept of the study is to benchmark this feature extractor technique relatively to others traditional methods such as PCA or t-SNE. As such, two Research questions will be studied:

- *Research Question 1*: How efficient is Autoencoder in feature extraction when compared to standard techniques for given clinical data?
- *Research Question 2*: Which model, from this research comparative study, can be used to predict cardiovascular risk from clinical data?

By this study, the main concerned addressed here is to give more reliable diagnostics to African population. Until now most of the studies done for cardiovascular diseases have been conducted on western or developed countries and the cohort which have been studied are mostly from those countries only. The benefit of current work will be to develop predictive models based on the health data already collected on ground from the population in Benin.

3.3 Functional Requirements

3.3.1 Data Consideration

3.3.1.1 Data Overview

Data used for our study comes from the Project TAHES which has the ambition to create a cohort study in SSA, especially in Benin. Until now, a total population of 1962 subjects have been included in this cohort. Since 2015, TAHES project has made a regular follow-up of this population cohort every year except 2018. Because of technical issue, the data collection of year 2018 has been cancelled. As a result, the current dataset includes a collection of 4 years dataset (years 2015, 2016, 2017 and 2019) reported in Excel format. On top of those 4 yearly Excel Dataset, another collection has been created in order to report the status or more precisely, the cardiovascular event associated to each subject of the cohort and its occurrence date.

3.3.1.2 Datasets Handling

The current study has to deal with 5 datasets (4 yearly clinical data collection and the event data report). This means that a proper strategy had to be defined for the way to handle those data. The difficulty in those datasets is that the track of some patients has been lost during the data collection while some others have been included in the cohort study later: If, officially, 1962 subjects are part of the cohort study, in practice, all of them were not systematically and regularly tracked through the years. The datasets altogether show that:

- 444 individuals have been sampled for only 1 year
- 385 individuals have been sampled for 2 years
- 478 individuals have been sampled for 3 years
- 655 individuals have been sampled for 4 years

3.3.1.3 Missing Data

Through each year, the clinical parameters collected from the cohort population is not systematically identical. Some clinical features have been added through different years and doesn't make the data exploitation as straight forward as expected.

In addition, for a given yearly clinical feature, the collection of the data could be missing for some individuals in the cohort.

3.3.1.4 Data Type and Data scaling

The dataset includes various clinical parameters: Those parameters do not have the same type nor the same range of value. If most of the parameters are integer types, some of them are string type corresponding to categories label. Another consideration point is the high difference of values between parameters in the dataset. Neglecting this aspect could lead to a significant impact in the result of the prediction model as the high values tends to weight more than small values.

3.3.1.5 Imbalanced Data

Checking imbalanced data consists of checking the distribution of target result between each result categories. In TAHES dataset, there are 130 cases of cardiovascular events

out of 1964 individuals: The data therefore is clearly imbalanced, and some precaution has to be taken to prevent the data inequality. The risk with imbalanced data is that the Machine Model will be specifically trained for the majority class (data whose result is more represented). Therefore, we may have good accuracy only for the majority class and not the class underrepresented.

3.3.1.6 Data partition

With the aim of training the predictive model, an appropriate strategy has to be applied to split the dataset into training and testing sets.

3.3.2 Features Consideration

3.3.2.1 Features overview

The idea of the TAHES project is to collect clinical data from a cohort population based in SSA. Various clinical metrics are collected from the cohort population through the 4 years of data collections. The yearly clinical data collected, referred as features or parameters in our study, in the collection of datasets that are presented in following section.

It can be noted that, excluding the Id number (used to identify the individual) 37 parameters are yearly collected from the cohort population. Those 37 features represent mainly the input parameters which will be used for the model prediction. The list of all the parameters are presented in in the Appendix A.1.

In addition to these yearly samples, the event situation dataset used to report the cardiovascular event include 3 more parameters (excluding again the Id number) listed in the Appendix A.2.

This event situation dataset is the main reference for the output label for the model. From this dataset 2 types of output parameters could be used as a label output for building the model:

- *Event Type*: This parameter defines the type of cardiovascular event the individual has suffered. Those events are subdivided into 4 categories (Alive, Dead, Right Heart Failure – RHF–, and stroke). Using this parameter as output label will induce to build a classification type for the predictive model which aims to predict the type of event an individual is likely to meet.

- *Date*: This parameter defines the time when the event has occurred. The difference between this date of event and the *inclusion* date (date of individual sample collection) gives an integer which represents the *lifespan* of the individual before the occurrence of the event. Using this new *lifespan* parameter as output label will induce to build a regression type for the predictive model which aims to predict when an individual will suffer of an event.

3.3.2.2 Feature Dimensionality

As mentioned above, the yearly input data include a total of 37 parameters. However, considering this number of parameters in regard to the number of individuals in the cohort (1962), the model has only few observation points (individuals) to make its prediction. This situation is referred as high-dimensional data and infers various potential problems while building the predictive model. Among the possible issues which can be encountered in high-dimensional data situation are:

- The risk of massively overfitting the predictive model
- The hardship to define a cluster of the different predictive classes.

This situation is commonly known as "the curse of dimensionality" and lead the model to fail into converging to an acceptable prediction. Therefore, given the actual size and having looked at existing models, the main challenge while building the predictive model is finding an appropriate way to extract meaningful features from this dataset.

3.3.3 Predictive Model Consideration

3.3.3.1 Model overview

In the field of AI various techniques in Machine Learning (ML) or (Deep Learning) are available to build a prediction model. Here are the basic concepts of Machine Learning and Deep Learning:

- *Machine Learning*: This is a branch of artificial intelligence associated with creating algorithms that can change themselves the parameter weight without human intervention to train the model and get the optimal result. Common machine learning algorithms include decision trees, support vector machines or Linear Regression

- *Deep Learning*: This is a branch of machine learning where algorithms are created and function similarly to machine learning, but these algorithms are stacked at many levels, each providing a different view of the data. This network of algorithms is called artificial neural networks (ANN). Popular Deep learning algorithms include convolutional neural networks (CNNs).

Generally, DL needs a lot of computational resource and is dedicated to complex data solving (just like image or video) or to very large dataset in opposition to ML.

3.3.3.2 Model Type

The event situation database contains multiple elements which allows us to build different types of model. *Event Type* attribute from the database defines the type of cardiovascular event. There are 4 categories: Alive, Dead, Right Heart Failure – RHF—, and Stroke. This gives the possibility to build a classification model which aims to predict the type of cardiac event an individual is likely to get. On another side, *Date* attribute from the database defines the time when the event has occurred. The difference between this date of event and the *inclusion* date (date of individual sample collection) gives an integer which represents the *lifespan* of the individual before the occurrence of the event. Using this new *lifespan* parameter gives the possibility to build a regression model which aims to predict when an individual is likely to suffer of a cardiovascular event.

3.3.3.3 Model training

Model training requires a special attention for this dataset. The model has to work with biological clinical data. Those data are expected to be variational from one individual to another one. As a result, the global dataset of the cohort individuals could be extremely noisy.

3.3.3.4 Model Evaluation

Checking the performance of the model is an important aspect. The metrics has to be wisely selected in order to see how efficient the model is.

3.4 Non-Functional Requirements

The functional requirement presented in the above section has defined all the considerations that have to be addressed in the model. Technical solutions are available to address those concerns. on the following challenges:

- Data Processing
- Feature Reduction
- Models Exploration

3.4.1 Data Processing

3.4.1.1 Datasets Handling

Multiple ways could have been considered to work with the 4 yearly datasets collection and the event status report altogether. Possible methods could be:

- *Time series approach*: As the data collection is done yearly, the straight-forward approach is to consider the evolution of each parameter through time as a time series.
- *Dataset Flat Merging*: In this approach, all the datasets are merged without any consideration of feature evolution over time.
- *Dataset Evolution Rate Merging*: In this approach, the idea is to select individuals with at least 2 yearly samples and calculate the evolution rate with the following calculation:

  $$f(t_1) * (f(t_2) - f(t_1))/(t_2 - t_1)$$

  . This method has been suggested by the Limoges University experts.

3.4.1.2 Missing Data

Missing data implies either missing features (from a yearly dataset to another one) or missing values of the features. For missing features, the values cannot be exploited and implies that the features have to be removed. As for missing values, those have the possibility to be imputed. Two techniques can be used for this:

- *Mean or Median imputation*: The mean or the median of the related feature values is computed in order to replace the missing value.
- *KNN imputation*: The missing value is replaced with the estimated value based on the closest individuals' observation feature's values.

3.4.1.3 Data Type and Data scaling

The string values are all categorical. As such, a numerical encoding for each one of them could be done to avoid discrepancy of the data. At the same time, to avoid high disparities between the values, data needs to be scaled. For the scaling. Two methods are possible for the scaling:

- *Normalization*: this technique scales the value between 0 to 1
- *Standardization*: It transforms the data to have a mean of 0 and a standard deviation of 1

3.4.1.4 Imbalanced Data

For imbalanced data, Oversampling methods are available to face the issue. Among them:

- *Faker*: This is a tool available in Python libraries. The purpose of Faker is to create a new database from scratch by generating artificially "new individuals".
- *Synthetic Minority Over-sampling (SMOTE)* [44]: This algorithm works by creating synthetic observations based upon the existing minority observations. SMOTE uses the K nearest neighbours' observation of the minority to over sample them.

3.4.1.5 Data partition

Datasets has to be partitioned in order to use one partition as a training model and the other as a validation of the model. Two methods are commonly used:

- *holdout*: the datasets is split in 2 unequal proportion. The largest distribution will be used for training and the smallest for validating and testing the model

- *k-fold cross-validation*: datasets is partitioned into "k" subsets. k-1 subsets are used to train the model; the last one serves to test and validate the model. This process is then repeated k times, using each time a different k subset once as the validation set. The average of the results gives the global estimation of the model.

3.4.2 Feature Reduction

Dimensionality reduction is a critical concern for this dataset since the number of parameters is very high. For solving this problem, the Feature Selection and the Feature Extraction technique are available. For the Feature Selection, the principle is to select only few features (the most meaningful ones) among all of them to build our prediction model. The Feature selection approach will not be considered since the study aims to make a comparison of some Features Extraction technique. The concept of the Feature Extraction technique is to combine all the features and compress them into a smaller representation. the following are available Feature Extraction Technique:

- *The Principal Component Analysis (PCA)*: It uses linear projection. A simple schematic of PCA is visible in Figure 2.4
- *T-distributed stochastic Neighbour Embedding (t-SNE)*: It uses non-linear projection. A t-SNE projection is visible in Figure 2.5
- *Autoencoder*:It is an ANN which compresses the data representation. A simple representation of an Autoencoder is visible in Figure 2.6

3.4.3 Models Exploration

3.4.3.1 Model type

Given the size of the data is relatively small (less than 2000 subjects), and the type of data (single values and not images) using ML instead of DL would be more appropriate in this research. Moreover, due to time consideration, the regression model to predict the "lifespan" will not be considered during this study: The work will focus exclusively on classification models (classifiers) to predict the risk of cardiac event and more particularly on those ML techniques:

- *Logistic Regression*: This classification technique is very basic and is based on a specific function

- *Random Tree Forest*: This method uses ensemble and graph/flowchart to make a classification
- *KNN*: this approach utilizes distance clustering for classifying

This research will make a comparative study on these above specified ML classifiers.

3.4.3.2 Model training

To train the model, data testing is required. However, only one dataset is available. The straightforward method will be to split the dataset into 2 parts one dedicated for the model training and another one for testing the trained model. But another method can be also applied: the cross validation. This method, also called the x-fold method, partitions the labelled dataset in x subsets of equal length, if possible. Then the first x-1 subsets are taken to train and the remaining subset is used for testing purposes.

3.4.3.3 Model Evaluation

The evaluation of the model is an important step in the study in order to make the comparison between each model. The metrics considered for this are:

- *Confusion matrix*: It represents a table containing two rows and two columns. The rows of the table are respectively the positive and negative, the predicted class and the columns define respectively the positive and negative real classes. The matrix reports at the end the number of true positives, false positives, false negatives, and true negatives. The table of the confusion matrix is in Figure 3.1 :

		Actual classes	
		positive	negative
Predicted classes	Positive	TP	FP
	Negative	FN	TN

TP: True Positive FP: False Positive
FN: False Negative TN: True Negative

FIGURE 3.1: Confusion Matrix

- *Accuracy*: It defines the number of correctly predicted data point among all the data points. The accuracy can be calculated with the following equation:

$$(TP + TN)/(TP + TN + FP + FN)$$

- *Precision*: The Precision also known as the positive predictive value, refers to the fraction of correctly predicted positive instance out of the total of all predicted positive instances. The precision can be found with this equation:

$$TP/(TP + FP)$$

- *Recall*: The Recall can be named as sensitivity and it refers to the fraction of correctly predicted positive instances over the total of actual positive instances. The recall can be set by the following expression:

$$TP/(TP + FN)$$

- *F1 score*: It refers to the weighted average of Precision and Recall using the harmonic mean between the two. The F1 score can be expressed as following:

$$F1 = 2 * precision * recall/(precision + recalll)$$

Chapter 4

Implementation Approach

4.1 Work Framework

4.1.1 Hardware overview

The Research study has been implemented in the cloud through Google Colab framework. This cloud service provides Virtual Machine (VM) running on Graphical Power Unit (GPU). The core specification of the VM used are presented below:

- 2 cores Intel(R) Xeon(R) CPU @ 2.20GHz
- 13 GB of RAM Memory
- 44 GB of HDD Memory

4.1.2 Software overview

For the implementation of the model, the code has been edited in Jupyter Notebook using Python language code. Various packages in Python libraries have been used to implement, build and test the models. The full list of the packages and libraries used are listed in the Table 4.1 :

4.2 Risk Assessment

The implementation of the research project could be disturbed or worst, disrupted due to potential hazard. Those have to be identified and managed as early as possible. Solving them or at least mitigating them is the key to carry out the project successfully.

	Libraries	Module	Usage
Data Processing	pandas	Excel_read	Read the excel file
	numpy	-	Various Operation with the Dataframe
	sklearn.impute	KNNImputer	Impute missing values in datasets
	sklearn.preprocessing	StandardScaler	Standardization of dataset values
	imblearn.over_sampling	SMOTE	Oversample the minority class
Model Training	sklearn.model_selection	train_test_split	Split the Dataset for training and testing
	sklearn.model_selection	KFold StratifiedKFold	Apply the cross validation to the Dataset
Feature reduction	sklearn.decomposition	PCA	PCA Features reduction
	sklearn.manifold	TSNE	TSNE Feature reduction
AutoEncoder	keras.models	Sequential	To build an ANN model for the autoencoder
	keras.layers	Conv1D MaxPooling1D UpSampling	Various Layers used to build the autoencoders
	keras	regularizers	Regularization to add sparsity in the autoencoder
Classification model	sklearn.linear_model	LogisticRegression	Logistic Regression classification
	sklearn.ensemble	RandomForestClassifier	Random Forest classification
Evaluation metrics	sklearn.metrics	classification_report confusion_matrix	Metrics (accuracy, precision, recall, F1) Confusion matrix reporting
Graphic	seaborn	-	Seaborn graphic
	matplotlib.pyplot	-	Graph plot

FIGURE 4.1: Python list of libraries and modules used

The possible risks which could arise during the project executing are listed below:

- *Data ethical concern*: The data correspond to clinical samples collected from Benin individuals as part of the TAHES project. Ethical issues regarding medical personal data collection must be considered. Otherwise, data could not be used as part of the study
- *Missing Libraries*: In part 4.1.2, the list of used libraries is presented. If some of them are missing, the usage of the module impossible.
- *Computational power shortage*: In part 4.1.1, the hardware specification used for the study is exposed. But for very big dataset or complex predictive model, this specification could be insufficient and leads to computational power shortage.
- *Size of the datasets*: If the size of the datasets is too large then loading them could be a problem. Reversely if it is too small, building a predictive model would be quite challenging
- *Unbalanced data*: Having disproportioned classes of data have a negative impact for building a prediction model

- *Noisy data*: This would impact the accuracy of the prediction model.
- *String categories*: ML cannot handle strings values in the datasets
- *Data scales*: the scale difference between features is impacting the efficiency of the predictive model.
- *Missing values*: ML cannot accept missing values in the datasets

The distribution of the mentioned risks could be reported in the below risk assessment diagram.

The distribution of the mentioned risks could be reported in the below risk assessment Table 4.1:

Frequency/ Consequence	**1-Rare**	**2-Remote**	**3-Occasional**	**4-Probable**	**5-Frequent**
4-Fatal				•Data ethical concern	
3-Critical	•Missing libraries			•Size of the datasets	
2-Major		•Computational power		•Unbalanced & •Noisy data	
1-Minor					•String Categories, •Data Scales, •Missing Values

TABLE 4.1: Project risk matrix

The hazards, registered in the risk assessment diagram, needs appropriate response to mitigate them. The possibilities are:

- *Data ethical concern*: To prevent this issue, data have been anonymized
- *Missing Libraries*: This can be solved with the pip installation
- *Computational power shortage*: Using special Virtual Machine could be considered to face the issue.
- *Size of the datasets*: when it is too small, datasets should be enriched. At extreme case, using Faker or possibly generative modelling can produce new data points.
- *Unbalanced data*: Possible solutions are listed in section 3.4.1.4
- *Noisy data*: Reworking the data or setting proper training model while partitioning the datasets could help to solve this problem.
- *String categories*: Technique for this issue are proposed in section 3.4.1.3
- *Data scales*: A proposition is mentioned in section 3.4.1.3.
- *Missing values*: some suggestions are presented in section 3.4.1.2

4.3 Methodology

Working on this research project grips multiple considerations to be reflected upon and proper methodology needs to be set to tackle those parts.

The datasets come from real samples on Benin population from TAHES cohort. Those datasets are, therefore, raw data with its flaws and a proper pre-processing of the data has to be executed. One of the biggest concerns about the datasets was, the way to handle them. Multiple discussion has been set with University of Limoges and CIT university to understand the best strategy to bring out the meaningful information for making a prediction. Various propositions were considered such as the time series, the flat merging or the evolution rate merging as suggested in section 3.4.1.1. The approach now is to analyse those different propositions through data visualisation technique.

The missing values management is also another concern to address. If the traditional approach for missing value is to fill it with the mean or the median values, then this has to be handled with great care since that the imputed value may not reflect a realistic value for the individual. This could be critical especially for noisy datasets as the one studied. Other propositions are mentioned in section 3.4.1.2.

Another reflexion point is the imbalance between the classes, an oversampling of the minority classes has to be considered. Possible methods are referred in section 3.4.1.4. Finally, the main subject of the study is to work on autoencoders which are relatively recent methods in ANN (Artificial Neuronal Network) for feature extraction. This technique and its architecture have to be assimilated through literature review before proposing an implementation.

4.4 Implementation Plan Schedule

As seen in the previous section 4.3, establishing this research work needs preparation to address various concerns. An implementation plan schedule is particularly useful in this situation to check the work progress. For this purpose, a Gantt chart has been elaborated on Figure 4.2:

4.5 Experiments and Architecture

This part will go through to the experiment done for implementing the research work and the adopted solution for each point addressed in Section 3.3.

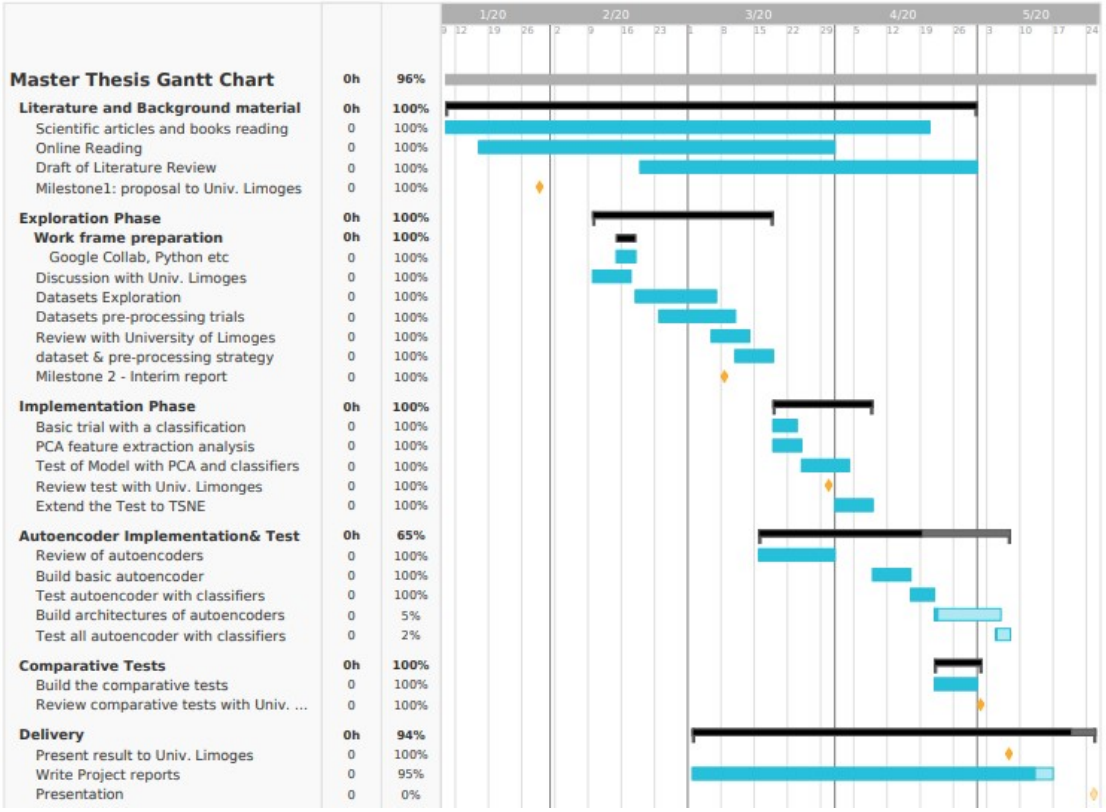


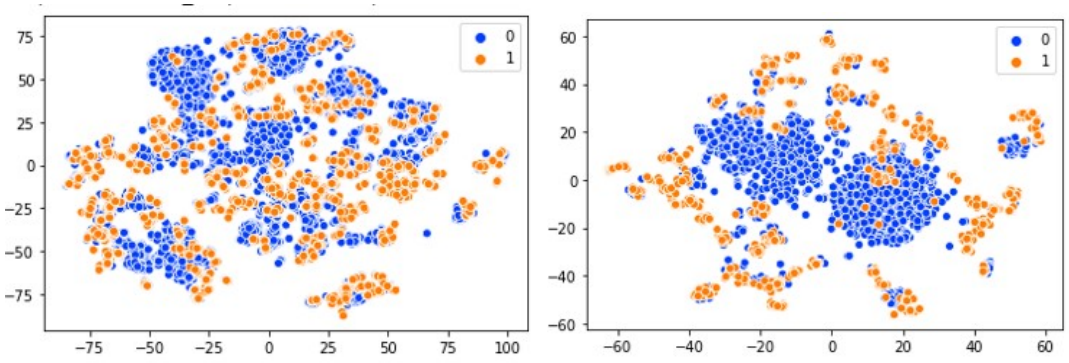
FIGURE 4.2: Project Plan Gantt Diagram

4.5.1 Data Processing Methodology

4.5.1.1 Datasets Handling

Datasets handling is the first step in the experiment and represents a critical process. Indeed, the performance of the predictive model depends greatly on the quality of the data fed into it for training. In section 3.4.1, three methods have been presented to handle the data: the time Serie approach, the flat merging and the evolution rate merging. If the Time Serie approach could have been the ideal method given the fact that each dataset represents samples of each year. Unfortunately, the small datasets (less than 2000 individuals) associated with fluctuant population over the year makes the exploitation of times series a bit difficult. This methodology indeed requires much more data observation. Therefore, the only option available to exploit the data is to merge them.

The flat merging is the fastest and immediate approach and offers the advantage to consider all the individuals (1964 in total). But the drawback is that the evolution of the feature over time is lost.



(a) Figure Flat Merge (b) Figure Evolution Rate Merge

FIGURE 4.3: t-SNE Representation of the merged Datasets

On the other hand, the evolution rate merging keeps the feature progression, but it requires at least observation with 2 samples. In other words, individuals with only one sample have to be discarded. In total 1518 individuals will be considered.

As criterion for sectioning the type of merging, a visual check on the t-SNE Representation of each merged data has been performed. Indeed, t-SNE is not just a feature reduction technique, it is a tool used to visualize high-dimensional data. As such, the t-SNE representation of the data can be plotted to check if a cluster between the 2 classes of population (the sick people and the healthy ones) is visible. The result of Flat Merge t-SNE Representation is visible on Figure 4.3 a) whereas the Evolution Rate Merge t-SNE is represented in the Figure 4.3 b).

A quick examination of the graph shows that the flat merging does not display any defined cluster between the 2 categories of population: the boundary between them is unclear, blurred and irrelevant. on the contrary, the evolution rate merging of the datasets shows a better result: the sick cases (orange points) are principally located in the outskirts of the figure. The choice of the type of merging is clearly in favour of the evolution rate merging datasets.

4.5.1.2 Missing Data

As mentioned in section 3.4.1.2, all the missing features from a yearly database to another have been removed. As such, from an initial database with 37 features, this has been downsized to 26 parameters. As for the missing values of a feature, the KNN imputation has been preferred over the mean or the median imputation. Indeed, the weakness of the mean or median imputation lies in the fact the imputed values does not reflect the reality of the individual condition. Conversely, with the KNN imputation, the imputed

value is close to the reel value since it is inferred from individuals close to the one with the missing value.

4.5.1.3 Data type and Data scaling

In the current study, data with categorical attribute are encoded numerically, and data are scaled with the standardization technique.

4.5.1.4 Imbalanced data

As explained in section 3.3.1.5, the datasets present a clear disproportion between the sick people and the healthy ones. To solve this problem, 2 solutions have been examined: Faker and SMOTE. Comparing them, SMOTE presents a great advantage since new individuals generated are done on the basis of existing ones. Generated points are close to the existing ones. On the other hand, Faker generates the individuals without any consideration of the initial database. Thus, new generated individuals may not be reliable since they may not be correlated with the original individuals. The SMOTE algorithm has been explained by Chawla, Nitesh V., et al.[44]. It generates fake samples from the minority class. A rudimentary explanation of the SMOTE algorithm is present below:

1. Among the minority class set C, for each $x \in C$, the Euclidean distance between x and other sample in set C in order to get the k-nearest neighbours of x.
2. Depending on the proportion of imbalanced data, the sampling rate N is fixed. For each $x \in C$, N instances $(x_1, x_2, \dots x_n)$ are randomly chosen among its k-nearest neighbours, and constitute the class set C_1.
3. For each instance $x_k \in C_1$ $(k = 1, 2, 3 \dots N)$, the following calculation is used to generate a new instance:

$$x' = x + rand(0,1) * |x - x_k|$$

$$with\ rand(0,1),\ the\ random\ number\ between\ 0\ and\ 1$$

Another consideration is to be highlighted in the current dataset: Among the classes of sick people, the data is also disproportioned. Out of 130 people with heart condition only 10 have a Right Heart Failure (RHF) and 16 had a stroke. Others are dead. Given the excessively low number of RHF and stroke in regard to the dead, all the categories

of people with heart condition have been gathered into one single class of "sick people". Then, the SMOTE technique has been applied.

4.5.1.5 Data partition

In section 3.3.1.6, 2 methods are proposed to train and validate the model. In this research study, as the datasets contains biological data, the values tend to be very noisy. The drawback of the holdout method is precisely that the splitting of the training and the tests subsets are fixed and consequently, the model prediction is highly dependent on how the database is split between the training set and the test sets. In other words, with this method, the model is learning noises from the training sets: This will reduce the accuracy of the model. On the other hand, the k-fold cross validation avoids such problems. the main advantage of this method is that by splitting the database in multiple subsets and doing multiple training and testing with each time, different subsets, prevents the model to learn specifically noises from a single training set. Consequently, the model can make a better generalization of the underlying pattern which differentiate the different classes of the data. In the research study, the k-fold cross validation technique has been used and the k number of subsets has been set to 10.

4.5.2 Feature Reduction Methodology

For reducing the dimensionality of the datasets, a comparative study of three features extractors is made as part of the project study: PCA, T-SNE, and Autoencoder.

4.5.2.1 Feature Extraction technique Justification

In the chapter 3.3.2, the datasets present a large number of parameters (37) and processing them could be quite challenging. Feature Extraction techniques are precisely methods which from those initial parameters data build derived revealing non-redundant values; but the prerequisite for building such derived values are precisely to find some correlation between parameters data so that they could be compressed into significant reduced values. As such, the relationship between parameters has to be examined to determine the correlation between them. This can be done through correlation coefficients and with the help of a visualization tool such as the Pearson correlation matrix. The higher the correlation degree between original parameters, the fewer derived parameters will be needed to make a good representation of the data. The Figure 4.4 shows the Pearson correlation matrix of the merged dataset and the complete value table of this Pearson correlation is available in the Appendix B.1.

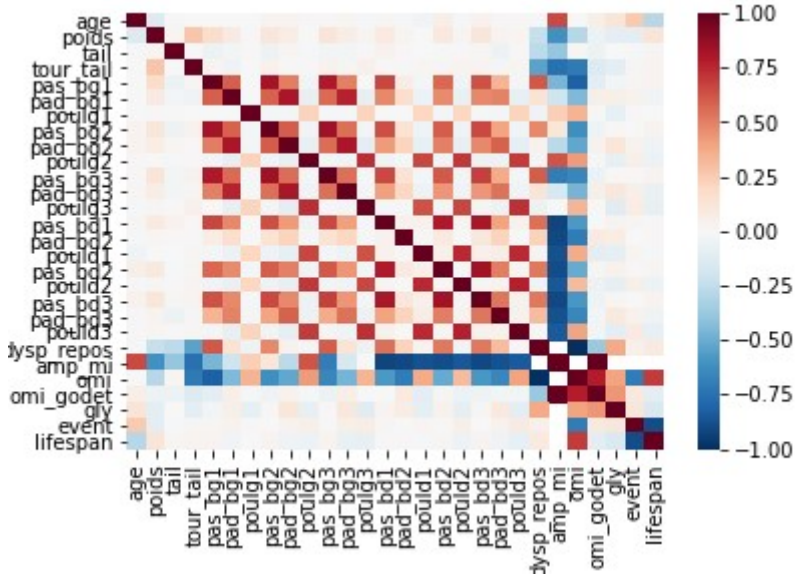


FIGURE 4.4: Pearson Parameters Correlation Matrix

4.5.2.2 PCA

PCA has the capacity to reduce the number of feature (as such the dimensionality of the database) by using linear orthogonal projection of these features. In the current merged database, after data processing, 26 features are in the database. The point is now to determine which will be the optimal number of reduced features to get the best representation of the data.

Various methods are available to set the optimal number of components. On this purpose, the current work has opted for a commonly used technique: the variance calculation for each Principal Component which represents the eigenvectors of the features' covariance matrix. The total cumulative variance of the Principal Component then gives an idea of the number of the Principal component required to get an optimal number of components. The graph 4.5 represents the variance of each Principal Component and their aggregated value: This figure shows us that with the first Principal Component explains around 25% of the total variance as for the second, it is accountable for about 15%. Altogether, they are responsible of 40% of the total variance. With the combination of the 16 first Principals Components, it is possible to reach to a total of more than 95% of explained variance. For the study, 16 components have been selected. As such, from the 26 original features, the implemented PCA feature extraction will reduce it into 16 Principals Components.

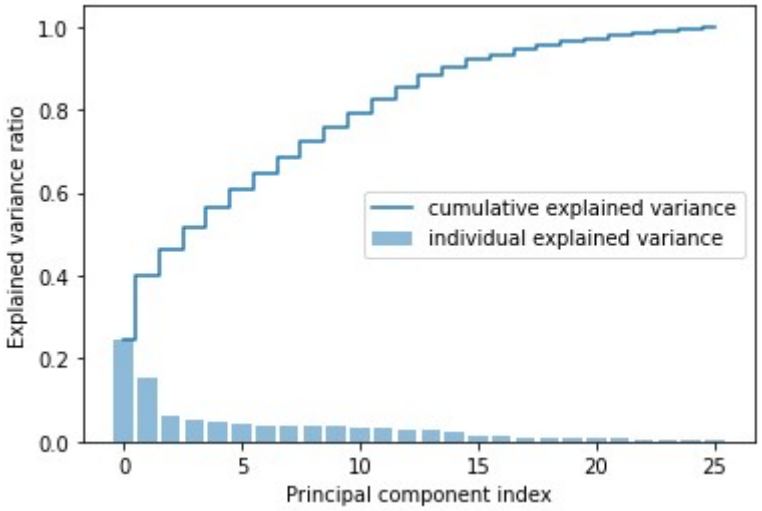


FIGURE 4.5: Principal Components explained Variance and cumulated explained Variance

4.5.2.3 T-SNE

As seen in section 3.4.2. The t-SNE reduces the features' number by using nonlinear projection. The basic concept of the algorithm is presented below:

1. The probability distribution in high-dimensional space which defines the relations between points and the probability of similarities of points are calculated. The similarity of points is defined as the conditional probability that a point chooses another one as neighbour:

$$p_{ij} = \frac{e^{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}}}{\sum_{k \neq I} e^{-\frac{\|x_k - x_I\|^2}{2\sigma_i^2}}}$$

   It is noted that this probability is proportionate to the probability density of a Gaussian (normal distribution) centred at x_i.

2. These conditional probabilities are then recreated in lower dimension and this reduction to the higher to the lower dimension is done by minimizing the difference between those 2 conditional probabilities through reducing the sum Kullback-Leibler divergence using gradient descent method:

$$\frac{\delta J}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_j - y_j\|^2)^{-1}$$

The t-SNE generally reduces the dimensionality to 2 or 3 parameters whatever the input data dimensions. In the current work, t-SNE is implemented to decrease the dimensionality to 2 parameters.

4.5.2.4 Autoencoder

As explained in section 3.4.2. The Autoencoder has the capacity to compress and simplify the representation of the data. For this, a hidden layer in the architecture of Autoencoder is used to encode the information in a smaller space since it has a lower number of nodes in comparison to the input. However, there is not a definite rule to set the appropriate number of nodes in the hidden layer. This number has been chosen empirically to half of the initial number of features and tested. As such, an Autoencoders with an input dimension of 26 nodes (as there is 26 parameters in the final merged dataset) and a hidden layer of 13 nodes has been implemented. As for its output, its size is identical to the input (26 nodes). The Autoencoder architecture is represented in Figure 4.6. After defining its architecture, the Autoencoder needs to be trained to learn

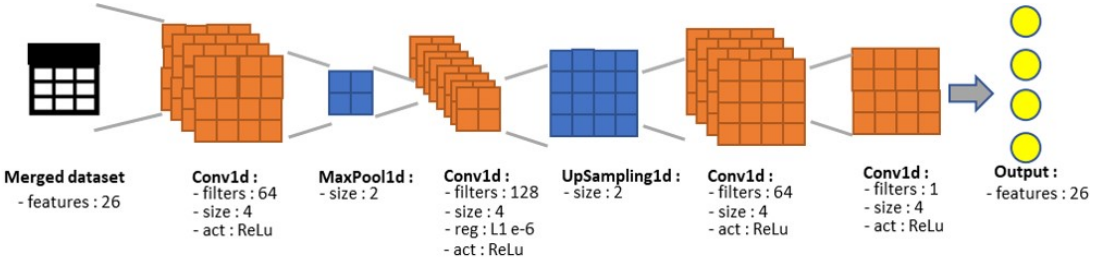


FIGURE 4.6: Autoencoder architecture implemented

how to encode the dataset. For this, training process hyperparameters have been set and refined after multiple trials. The final selection of the hyperparameters is mentioned in the Table 4.2: The result of the loss function is visible at the Graph 4.7: After 200

Parameters	*Settings*
Optimizer	Nadam
Loss function	Binary cross entropy
Learning Rate	0.001
Batch Size	200
Epochs	200

TABLE 4.2: Hyperparameters Selection for Training

epoch training, the graph shows a stabilization of the curve and the difference between the training curve and the test one is not reducing.

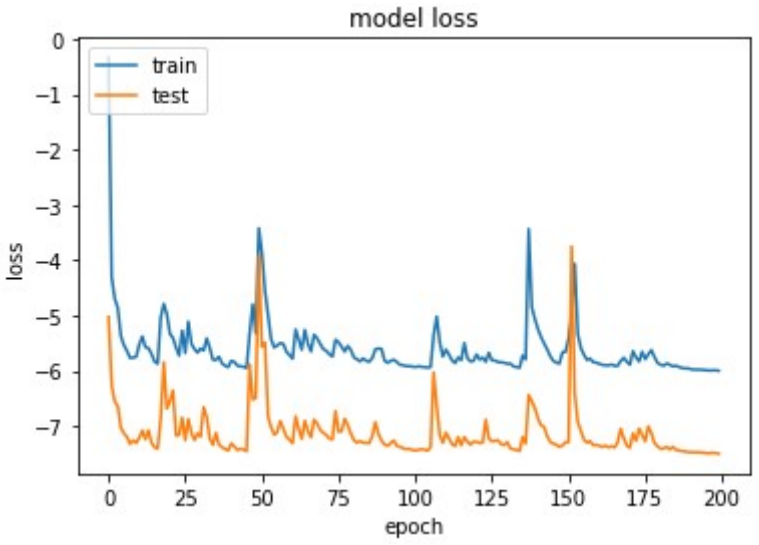


FIGURE 4.7: Model Loss Function Graph during the Training

4.5.3 Classification Methodology

As part of the study, different classification models have been tested.

4.5.3.1 Logistic Regression

Multiple logistic regression is commonly used by epidemiologists to estimate variables status such as alive or dead on the basis of multiple independent variables. Multiple logistic regression tries to define the optimal equation that best predicts the output value of Y variable on the basis of the multiple independent X variables. Equation of the logistic regression can be expressed as:

$$Y = \sum_{k=1}^{n} e^{\beta_k X_i}$$

$$\beta_k \ : \ regression\ coefficients$$

$$X_i \ : \ The\ independent\ variables$$

(which in our case represent the various features) The Y function depicts a sigmoid graph (Figure 4.8) which value is comprised between 0 and 1. In the study, the value of the function will classify the individual to either healthy (value close to 0) or with a cardiac condition (value close to 1): The process of training the logistic regression in the model means to optimize all the coefficients β_k in the equation to most accurately

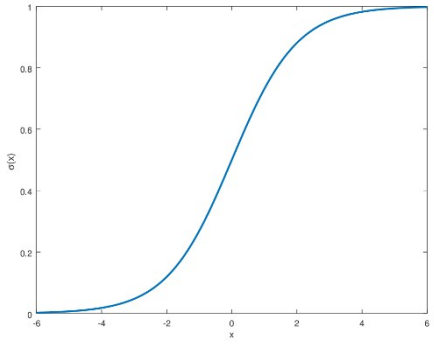


FIGURE 4.8: Sigmoid Function

fit in the solution Y. For this process, the mean squared error is minimized:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y - \widehat{Y})^2$$

Where is n – number of predictions, Y – real values of target variable, $\widehat{Y}$ *– predicted values of the ta*

4.5.3.2 Random Tree Forest

Random Forest is an ensemble (aggregation of multiple outputs made by a model) of Decision Trees. This one is an algorithm dividing the input dataset to smaller datasets on the basis of features until that the data falls under one label. The process for training the Random Tree Forest is described as below:

1. f features are randomly selected among the T total features ($k << T$).
2. From the f features, compute the node d using the optimal split point.
3. Split again optimally the node d into descendant nodes.
4. Repeat the above steps 1 to 3 until reaching the m number of nodes.
5. The forest is generated by repeating steps 1 to 4 for n number of times (n number of trees will be created).

For the prediction, the Random Tree Forest takes predicted outcomes and calculate the votes for each predicted outcome. The highest vote will be the final prediction. A graph of the Random Tree Forest is visible in Figure 4.9.

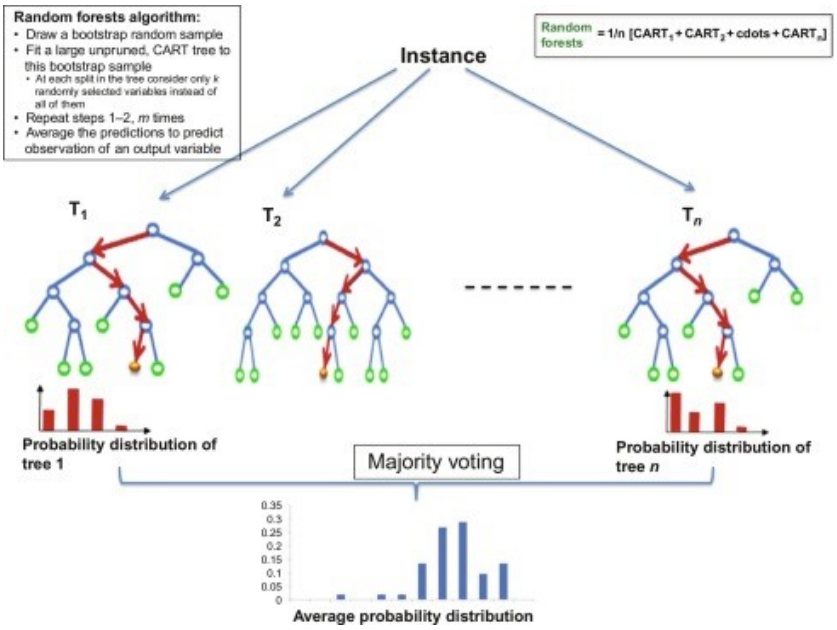


Image source from Anuj Mubayi, Disease Modelling and Public Health, Part A (2017) [45]

FIGURE 4.9: Random Tree Forest Graph

4.5.3.3 K-Nearest Neighbours

The KNN algorithm is mentioned below:

1. Define a positive integer k corresponding to the number of neighbours to consider.
2. For an observation point, compute its distance to other datasets points.
3. Determine the k closest points to the observation points (points to lowest distance to k): Those points are defined as the neighbours.
4. Assign the observation points to the majority neighbours' class.

The computed distance to the other points can be calculated with different metrics such as the Manhattan distance or the Minkowski Distance metrics. But the metric used in this study is the most commonly used i.e. distance metric: the Euclidean distance which expression is shown below:

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

with x_i and y_i the respective positions for each point

4.5.4 Research Methodology

Based on the experiment presented in the preceding sections, the current study can be framed in the phases illustrated on Figure 4.10.

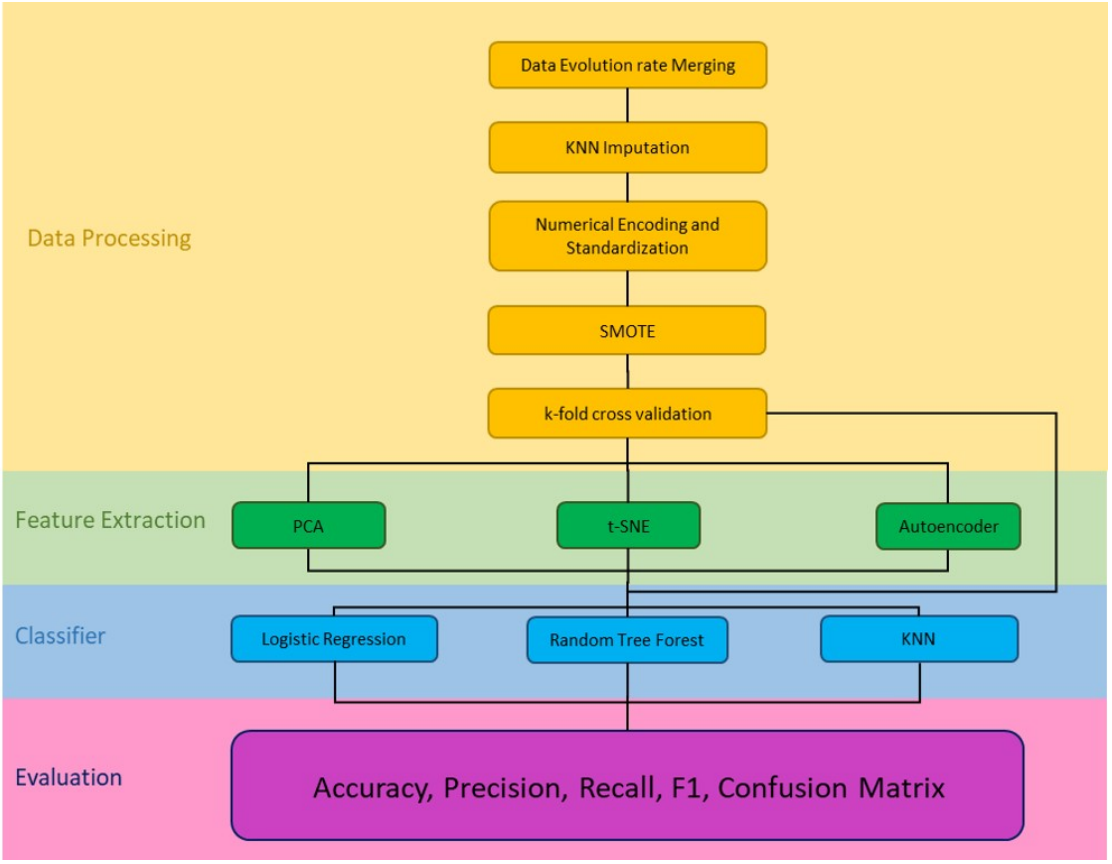


FIGURE 4.10: Research Methodology Graph

4.6 Evaluation

This part presents the resulting performance of feature extraction methods associated with the classifiers as presented in section 4.5.4. The Table 4.3 presents the results obtained after processing the proposed methodology.

The first aim of this research is to evaluate the performance of an autoencoder as a feature extractor with regard to other traditional feature extractor techniques for building a classification model. A comparative study was successfully carried out to compare the autoencoder with PCA and t-SNE as feature extraction techniques. For this, those feature extractors were associated with 3 different classifiers: the logistic Regression, the Random Tree Forest and the K-Nearest Neighbours. The PCA and t-SNE was used as

			Evaluated Model (Feature Reduction + Classifier)											
			PCA			*TSNE*			*Autoencoder*			—		
			LR	*RTF*	*KNN*	*LR*	*RTF*	*KNN*	*LR*	*RTF*	*KNN*	*LR*	*RTF*	*KNN*
Score	Metrics	*Accuracy*	0.75	**0.96**	0.94	0.55	0.95	0.94	0.74	0.94	0.88	0.60	0.91	0.89
		Precision	0.75	**0.94**	0.90	0.55	0.94	0.91	0.75	0.93	0.83	0.60	0.87	0.83
		Recall	0.75	0.98	**0.99**	0.52	0.97	0.98	0.72	0.96	0.96	0.57	0.98	0.98
		F1	0.75	**0.96**	0.95	0.54	0.95	0.94	0.73	0.94	0.89	0.58	0.92	0.90
	Confusion Matrix	*True Pos.*	1068	1401	**1419**	750	1390	1397	1024	1373	1371	814	1399	1405
		False Pos.	353	**90**	151	615	91	133	337	109	283	540	212	281
		False Neg.	363	30	**12**	681	41	34	407	58	60	617	32	26
		True Neg.	1078	**1341**	1280	816	1340	1298	1094	1322	1148	891	1219	1150

TABLE 4.3: Comparative Models Evaluation Table

a benchmark to evaluate the performance of the autoencoder and the result is available in Table 4.3 .

Furthermore, to complete the study and add more robustness, it would have been better to test various architecture of the autoencoder: This was originally the plan. However, pre-processing of the data took more time than expected and as such, the different tests of the autoencoders architecture have not been realized.

Another goal of the study was to propose an efficient predictive model of cardiovascular risk. As explained in the Table 4.3 , the PCA feature extraction associated with the Random Tree Forest could reached with a very high accuracy of 96% and an F1 score of 96% as well. If the scoring looks very high, this has to be tempered with the fact that this classification model is a binary one which predicts the cardiovascular risk; but not the type. Indeed, the initial proposal was to make a multi class classification model (healthy, death, stroke and Right Heart Failure); but due to the limited amount of stroke and Right Heart Failure cases, those classes have been merged with the death class to make one class of cardiovascular accident.

4.7 Prototype

Given the current state, the model cannot be exploited as it is. Multiple improvement has to be implemented to adjust the predictive tool. This tool is primarily designed to assist medical staff in making prognosis of cardiovascular risk more specifically for SSA population. For this purpose, the model has to improve its robustness by being trained with larger sample of data Fortunately, the TAHES cohort is still in progress and more and more data, and individuals are included in the study. Another development of this prognosis tool would be to identify the main risk factors among all the sampled features. Indeed, on the field, medical staff cannot practically measure all the 37 features. Many of them are probably redundant, uninformative or worst noisy. As such, instead of using feature extraction technique to build the model, it would be preferred to apply feature

selection approach and select only few meaningful features to train and build the model with probably some compromising in term of accuracy: The resulting model would be more convenient and accessible for medicals on the field as they would have to take only few sample from the patient to make the prognosis.

Chapter 5

Conclusions and Future Work

5.1 Result Discussion

5.1.1 Analysis

The aim of this study was to test the efficiency of the Autoencoder as feature reduction technique. Based on the results of this research, PCA tends to perform better than the other reduction methods especially in regard to the Autoencoder. The latter performs quite well with the Logistic Regression (LR) or the Random Tree Forest (RTF) classifiers (with an accuracy of 74% and 94%); but it tends give a poor result when it is associated with the K-Nearest Neighbours (KNN) Classifier (with an accuracy of 88%). This is because of the fact that Autoencoder is a feature reduction technique, it just simplifies the dataset representation thanks to its hidden encoding layer; but it does not reduce the dimensionality of its output. As such, the output presents as much parameters as its input features. Unfortunately, KNN tends to be less efficient with larger dimension since it has to evaluate distances for every parameter and a higher number of them induces sparsity in the overall distance. As a result, Autoencoder associated with KNN classifier delivers poor result. It is interesting to note that Autoencoder as a feature extractor technique associated with KNN gives almost the same accuracy as the KNN in standalone without any feature extractor with respectively an accuracy of 88% and 89%.

In term of the Classifiers model, without any surprise LR gives the lowest accuracy in comparison to the RTF or the KNN. This result is expected since the LR as a function-based technique provides a very simple classification method and cannot perform well with complex database including multiples features. On the contrary, RTF is an ensemble of Decision Tree (graph) technique which relies on voting and probability. This offers

a very powerful model to make its classification. Even without any feature extractor, the RTF reaches to an accuracy of 91%. Probably this is what makes the RTF too powerful to use as a benchmark test for feature extractor technique. It would have been wiser to use Decision Tree classification as a benchmark test, but it offers at the same time the highest score and thus the best model of classification associated with the PCA feature extractor with an accuracy of 96%.

5.1.2 Limitations

It is to be noted that the results and interpretations cannot be generalized due to several limitation upon which some solutions were implemented. Such as:

- The dataset size is particularly small. No satisfying solution has been provided for this. Ideally, more samples should have been collected to increase the size of the dataset.
- The dataset has a limited case of sick people. Out of 1964 individuals, only 130 suffered from cardiac disease; and among those 130 affected, just 10 of them have a Right Heart Failure (RHF) and 16 have a stroke. If initially, the intent was to classify the different type of disease, the few cases of stroke and RHF has led to merge all the sick individuals into one class "Heart disease" and then, the SMOTE technique was applied to oversample the "Heart disease" minority class. This process can have a significant impact in the prediction model. Ideally, the database should have been extended with other real cases of individuals with heart condition.
- The dataset is very noisy. This is in fact the direct consequence of the 2 above points. Since the dataset is small and there are only few observation points for the sick people, the weight of each sick people is extremely high and induces a lot of variance in the dataset. To overcome this issue, the dataset evolution rate (and not the flat) merging has been selected and the cross-validation technique has been applied. A bigger sample of the data over the time, could have helped to reduce the noises; not to mention that the time series approach instead of the merging the datasets could have been a more dedicated solution.
- For missing data, the KNN imputation technique has been applied. However, KNN is very sensitive to high dimension data which is the case in the current study since 26 parameters are present. It may be possible that KNN imputation is not the best approach to replace the missing values.

5.2 Conclusion

In the era of big data, when more is axiomatically better, it appears many noisy data inputs often leads to unsatisfactory model performance. Removal of uninformative or even worse, misinformative input might help to train a machine learning model with an overall better performance. The concept of applying Feature reduction technique responds perfectly to this challenge. The current study is aimed to see how performant the usage of the Autoencoder as a feature reduction technique could be. As such, 2 Research questions were stated:

- *Research Question 1*: How efficient is autoencoder in feature extraction when compared to standard techniques for given clinical data?
- *Research Question 2*: Which model, from this research comparative study, can be used to predict cardiovascular risk from clinical data?

To answer those questions a comparative study with existing Features extraction techniques such as PCA and T-SNE associated with 3 classifiers (Logistic Regression, Random Forest Tree, and K-Nearest Neighbours) was put in place.

If Autoencoder provides satisfying accurate result associated with the Logistic Regression or the Random Forest with respectively 74% and 94%, it performs very poorly (with only 88%) associated with KNN. Hence, it can be inferred that the Autoencoder would not be the optimal feature reduction technique when it is associated with classifiers sensitive to the high dimensionality (such as the KNN) as explained in section 5.1.

The comparative study between the feature extraction techniques reveals also that PCA tends to perform better than the Autoencoder. When PCA associated with LR and RTF can reach 75% and 96% of accuracy, the Autoencoder can reach only 74% and 94%. This result could be quite surprising at first since Autoencoder is known in the literature for being very efficient for denoising the dataset and the current dataset is very noisy. This result can be explained as Autoencoders is an ANN (Artificial Neuronal Network) and ANN performs very well with big dataset. Indeed, in order to efficiently encode the main pattern of the individual, the Autoencoder needs a decent amount of data to refine its encoding. But, in the current study, the dataset is relatively small and thus, the autoencoder cannot encode properly the main characteristics of individuals. With a bigger dataset, the Autoencoder might be more efficient.

In regard to using the optimal model to make prediction of the cardiac risk of event, it is noted that scoring is different from AI experts and Doctors. Indeed, while AI experts are aiming for the model with the highest metrics such as the accuracy, the

precision, the recall and the F1 score and for the lowest number of errors, medical experts are considering the "sensitivity" and "specificity" of the model. Medically, sensitivity is defined as the rate of people correctly tested positive for a disease; whereas, specificity is defined as the rate of people correctly tested negative. In another words, the "sensitivity" refers to the "True Positive" in the confusion matrix and the "specificity" refers to the "True Negative". As such, Medical experts give a special attention for the model confusion matrix. The aim here is to get the highest value of "sensitivity" ("True Positive") and by extension the lowest number of "False Negative". Indeed, a highly sensitive test will essentially detect most of individuals with actual cardiac disease even if False Positive comes up. Whereas, a highly specific test may discard people with real heart condition (False Negative).

In the study, it appears that the model with the best metrics (accuracy, precision, recall and F1 score) is not the one which has the best sensitivity rate (high number of True Positive and low value of False Negative). Typically, for AI experts, the model which presents the best predictive result is the model including PCA feature reduction and Random Forest Tree Classifier as it presents the best scoring metrics (accuracy of 96%, F1 score of 96%). Whereas, for medical experts the optimal model in the comparative study will be the one with the PCA features reduction associated with the KNN classifier (accuracy of 94%, F1 score of 95%). This model has the highest True Positive with the value of 1419 and the lowest value of False Negative with only 12.

5.3 Future Work

The present research has presented a preliminary model of Machine Learning predictive tool using a Feature Extraction technique associated with a classifier. If the PCA associated with the Random Forest Tree (or the K-Nearest Neighbours) offers interesting result with the current dataset, it should be noted that the model has been trained on a relatively small database.

The TAHES project is still ongoing and medical experts are continuing to integrate more individuals in the cohort and are collecting more samples. The objective of the TAHES project is to include at least 15000 individuals in observation which is 10 times more than the current database reference. It is expected that with a bigger dataset the predictive model built can be refined and more precise. We can think that with a bigger dataset the autoencoder will work better, a time series approach of the data (instead of the evolution rate merging) could be considered and neglected output classes such as the Right Heart Failure or the stroke could be integrated again among the predictive classes.

In term of feature reduction technique, the literature has introduced a new approach with very promising results: The Linear Discriminant Analysis (LDA). It could be interesting to integrate this technique in the comparative study to check its performance.

Taking samples on 26 parameters is not practically doable in this field; hence, it can be interesting to discuss with the epidemiologist in Limoges University to identify few main features on which the model has to make prediction. The work on feature selection techniques rather than feature extraction techniques is currently under discussion with the Limoges University. Another possibility is also to build a predictive model on the lifespan of the individual based on the same clinical data. As such, the model will no more be a classification approach; but a regression one.

Publication of a paper with the Limoges University team is also under consideration.

Bibliography

[1] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature medicine*, vol. 25, no. 1, pp. 30–36, 2019.

[2] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.

[3] K. J. Dreyer and J. R. Geis, "When machines think: radiology's next frontier," *Radiology*, vol. 285, no. 3, pp. 713–718, 2017.

[4] C. S. Kruse, R. Goswamy, Y. J. Raval, and S. Marawi, "Challenges and opportunities of big data in health care: a systematic review," *JMIR medical informatics*, vol. 4, no. 4, p. e38, 2016.

[5] M. Wesolowski and B. Suchacz, "Artificial neural networks: theoretical background and pharmaceutical applications: a review," *Journal of AOAC International*, vol. 95, no. 3, pp. 652–668, 2012.

[6] K. Itahashi, S. Kondo, T. Kubo, Y. Fujiwara, M. Kato, H. Ichikawa, T. Koyama, R. Tokumasu, J. Xu, C. S. Huettner *et al.*, "Evaluating clinical genome sequence analysis by watson for genomics," *Frontiers in medicine*, vol. 5, p. 305, 2018.

[7] T. Denoeux, "A k-nearest neighbor classification rule based on dempster-shafer theory," in *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 2008, pp. 737–760.

[8] R. Shah and S. Silwal, "Using dimensionality reduction to optimize t-sne," *arXiv preprint arXiv:1912.01098*, 2019.

[9] M. Z. Nezhad, D. Zhu, X. Li, K. Yang, and P. Levy, "Safs: A deep feature selection approach for precision medicine," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2016, pp. 501–506.

[10] H. Itoh, A. Imiya, and T. Sakai, "Dimension reduction and construction of feature space for image pattern recognition," *Journal of Mathematical Imaging and Vision*, vol. 56, no. 1, pp. 1–31, 2016.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[12] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[13] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[14] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer," *IEEE Access*, vol. 6, pp. 28 936–28 944, 2018.

[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[16] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[17] S. Amidou, Y. C. Houehanou, P.-M. Preux, D. S. Houinato, and P. Lacroix, "Feasibility of a population-based cardiovascular cohort in sub-saharan africa: experience of tahes study." 2018.

[18] G. A. Roth, C. Johnson, A. Abajobir, F. Abd-Allah, S. F. Abera, G. Abyu, M. Ahmed, B. Aksut, T. Alam, K. Alam *et al.*, "Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015," *Journal of the American College of Cardiology*, vol. 70, no. 1, pp. 1–25, 2017.

[19] A. Makhzani and B. Frey, "K-sparse autoencoders," *arXiv preprint arXiv:1312.5663*, 2013.

[20] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS international conference on computer systems and applications*. IEEE, 2008, pp. 108–115.

[21] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 Science and Information Conference*. IEEE, 2014, pp. 372–378.

[22] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[23] CbInsight, "Ai in healthcare heatmap: From diagnostics to drug discovery, deals heats up," 2018. [Online]. Available: https://www.cbinsights.com/research/artificial-intelligence-healthcare-investment-heatmap/

[24] W. H. Organization, W. H. Organization *et al.*, "The top 10 causes of death. 2014," *Fact sheet*, no. 310, 2018.

[25] S. Golemati and K. S. Nikita, *Cardiovascular Computing-Methodologies and Clinical Applications*. Springer, 2019.

[26] G. Singh, S. J. Al'Aref, M. Van Assen, T. S. Kim, A. van Rosendael, K. K. Kolli, A. Dwivedi, G. Maliakal, M. Pandey, J. Wang *et al.*, "Machine learning in cardiac ct: basic concepts and contemporary data," *Journal of cardiovascular computed tomography*, vol. 12, no. 3, pp. 192–201, 2018.

[27] E. Topol, *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.

[28] S. J. Al'Aref, G. Maliakal, G. Singh, A. R. van Rosendael, X. Ma, Z. Xu, O. A. H. Alawamlh, B. Lee, M. Pandey, S. Achenbach *et al.*, "Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the confirm registry," *European heart journal*, vol. 41, no. 3, pp. 359–367, 2020.

[29] N. I. Hasan and A. Bhattacharjee, "Deep learning approach to cardiovascular disease classification employing modified ecg signal from empirical mode decomposition," *Biomedical Signal Processing and Control*, vol. 52, pp. 128–140, 2019.

[30] E. Braunwald, "The ten advances that have defined modern cardiology," *Trends in cardiovascular medicine*, vol. 24, no. 5, pp. 179–183, 2014.

[31] P. Lamata, R. Casero, V. Carapella, S. A. Niederer, M. J. Bishop, J. E. Schneider, P. Kohl, and V. Grau, "Images as drivers of progress in cardiac computational modelling," *Progress in biophysics and molecular biology*, vol. 115, no. 2-3, pp. 198–212, 2014.

[32] M. Alessandrini, M. De Craene, O. Bernard, S. Giffard-Roisin, P. Allain, I. Waechter-Stehle, J. Weese, E. Saloux, H. Delingette, M. Sermesant *et al.*, "A pipeline for the generation of realistic 3d synthetic echocardiographic sequences:

Methodology and open-access database," *IEEE transactions on medical imaging*, vol. 34, no. 7, pp. 1436–1451, 2015.

[33] H. Xia, I. Asif, and X. Zhao, "Cloud-ecg for real time ecg monitoring and analysis," *Computer methods and programs in biomedicine*, vol. 110, no. 3, pp. 253–259, 2013.

[34] A. Suinesiaputra, D. A. Bluemke, B. R. Cowan, M. G. Friedrich, C. M. Kramer, R. Kwong, S. Plein, J. Schulz-Menger, J. J. Westenberg, A. A. Young *et al.*, "Quantification of lv function and mass by cardiovascular magnetic resonance: multicenter variability and consensus contours," *Journal of cardiovascular magnetic resonance*, vol. 17, no. 1, p. 63, 2015.

[35] K. Krishnan, L. Ibanez, W. Turner, and R. Avila, "Algorithms, architecture, validation of an open source toolkit for segmenting ct lung lesions," in *Proc. MICCAI Workshop on Pulmonary Image Analysis (Sept. 2009)*, vol. 365, 2009, p. 375.

[36] V. Khatibi and G. A. Montazer, "A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8536–8542, 2010.

[37] J. Patterson and A. Gibson, *Deep learning: A practitioner's approach.* " O'Reilly Media, Inc.", 2017.

[38] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp. 37–49.

[39] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. Rudd, and M. van Der Schaar, "Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 uk biobank participants," *PloS one*, vol. 14, no. 5, 2019.

[40] Q. Hu, L. F. d. F. Souza, G. B. Holanda, S. S. Alves, F. H. d. S. Silva, T. Han, and P. P. Rebouças Filho, "An effective approach for ct lung segmentation using mask region-based convolutional neural networks," *Artificial Intelligence in Medicine*, p. 101792, 2020.

[41] P. Singh, S. Singh, and G. S. Pandi-Jain, "Effective heart disease prediction system using data mining techniques," *International journal of nanomedicine*, vol. 13, no. T-NANO 2014 Abstracts, p. 121, 2018.

[42] V. Kazak, "Unsupervised feature extraction with autoencoder: for the representation of parkinson´ s disease patients," Ph.D. dissertation, 2019.

[43] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PloS one*, vol. 12, no. 4, 2017.

[44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[45] A. S. S. Rao, S. Pyne, and C. R. Rao, *Disease modelling and public health*. Elsevier, 2017.

Appendix A

Parameters in the datasets

Value Type	***param code***	***Parameter description***
integer	*id*	ID number
date	inclusion	date of sample collecton
integer	age	age
str cat	sex	gender
date	ddn	birthday
integer	poids	weight (kg)
integer	tail	height (cm)
integer	tour_tail	waist size (cm)
integer	tour_hanch	hip circonference (cm)
integer	pasbg1	systolic blood pressure left arm at rest (mm Hg)
integer	pad_bg1	diastolic blood pressure left arm at rest (mm Hg)
integer	pouls_g1	left pulse at rest
integer	pasbg2	systolic blood pressure left arm moderate activity (mm Hg)
integer	pad_bg2	diastolic blood pressure left arm moderate activity (mm Hg)
integer	pouls_g2	left pulse moderate activity
integer	pasbg3	systolic blood pressure left arm high activity (mm Hg)
integer	pad_bg3	diastolic blood pressure left arm high activity (mm Hg)
integer	pouls_g3	left pulse high activity
integer	pas_bd1	systolic blood pressure right arm at rest (mm Hg)
integer	pad_bd1	diastolic blood pressure right arm at rest (mm Hg)
integer	pouls_d1	right pulse at rest
integer	pas_bd2	systolic blood pressure right arm moderate activity (mm Hg)
integer	pad_bd2	diastolic blood pressure right arm moderate activity (mm Hg)
integer	pouls_d2	right pulse moderate activity
integer	pas_bd3	systolic blood pressure right arm high activity (mm Hg)

Value Type	***param code***	***Parameter description***
integer	pad_bd3	diastolic blood pressure right arm high activity (mm Hg)
integer	pouls_d3	right pulse high activity
str cat	dysp_repos	dyspnea at rest
str cat	amp_mi	limb amputation
str cat	amp_mi_sg	amputation location
str cat	omi	edemas at lower limbs
str cat	omi_godet	OMI, godet
integer	gly	glycemy
Integer	Creat	creatine
integer	nb_gross	Number of pregnancies for women

TABLE A.1: Parameters Dataset Description

Value Type	***param code***	***Parameter description***
date	Date	Date of event occurence
integer	Id	Individual Id
str cat	Event Type	Type of cardiovascular accident
integer	Age	Age of the individual when event occurs

TABLE A.2: Parameters Event Report Description

Appendix B

Pearson Features Correlation Values

	age	poids	tail	tour_tail	pas_bg1	pad_bg1	poulg1	pas_bg2	pad_bg2	poulg2	pas_bg3	pad_bg3	poulg3	pas_bd1	pad_bd2	pould1	pas_bd2	pould2	pas_bd3	pad_bd3	pould3	dysp_repos	amp_mi	omi	omi_godet	gly	event	lifespan
age	1	-0.139465	-0.008505	-0.004588	-0.002615	-0.005356	-0.001662	0.023489	0.027262	0.004464	0.006329	0.017591	-0.009072	0.01289	-0.013305	-0.03545	0.055457	-0.013739	0.039534	0.002793	0.000763	-0.012034	0.661223	-0.000147	0.054734	0.130781	0.261757	-0.279298
poids	-0.139465	1	0.020432	0.281045	0.174252	0.082873	0.000087	0.112538	0.070453	-0.000919	0.131659	0.064836	0.015399	0.090177	0.02021	-0.003167	0.09739	0.015348	0.110813	0.026413	0.021106	-0.232938	-0.603804	-0.274249	-0.047533	-0.097632	-0.097516	0.132713
tail	-0.008505	0.020432	1	-0.004284	-0.052998	-0.02638	0.001044	-0.050009	0.016141	-0.00946	-0.024889	-0.030289	0.003277	0.046871	0.006141	-0.006337	-0.008661	0.002869	0.015145	-0.050975	0.002032	-0.260829	-0.383205	0.019112	-0.052296	-0.006044	-0.00825	0.009679
tour_tail	-0.004588	0.281045	-0.004284	1	0.017689	-0.044777	-0.002318	0.042146	0.001383	0.007245	0.033993	0.011293	0.048452	0.000321	0.002004	0.010641	0.008768	0.023811	0.026809	-0.02327	0.019794	-0.534272	-0.744458	-0.705318	-0.126043	-0.094496	0.011123	0.024491
pas_bg1	-0.002615	0.174252	-0.052998	0.017689	1	0.600872	-0.00146	0.828176	0.5106	-0.031225	0.802119	0.506566	-0.005545	0.668469	0.06731	-0.022109	0.572376	-0.033006	0.638348	0.336616	-0.021911	0.61896	-0.453691	-0.818376	-0.100917	-0.027639	0.010005	0.028915
pad_bg1	-0.005356	0.082873	-0.02638	-0.044777	0.600872	1	-0.000866	0.592081	0.813242	-0.074434	0.565845	0.769431	-0.073946	0.479994	0.165372	-0.062466	0.470431	-0.068687	0.479067	0.497929	-0.076	0.135839	-0.200106	-0.44149	0.056856	0.057082	0.031712	-0.036576
poulg1	-0.001662	0.000087	0.001044	-0.002318	-0.00146	-0.000866	1	-0.017842	-0.020988	0.234269	-0.009809	-0.009688	0.225918	-0.006329	-0.003275	0.210685	-0.018236	0.209223	-0.012655	-0.017935	0.219634	-0.04621	0.235242	0.35007	-0.014647	-0.081714	0.005019	-0.006717
pas_bg2	0.023489	0.112538	-0.050009	0.042146	0.828176	0.592081	-0.017842	1	0.60556	-0.038778	0.847279	0.551603	-0.026486	0.642198	0.13955	-0.042595	0.605723	-0.045667	0.656713	0.392524	-0.031753	0.47776	0.120877	-0.610881	-0.040286	0.002614	-0.00086	0.03069
pad_bg2	0.027262	0.070453	0.016141	0.001383	0.5106	0.813242	-0.020988	0.60556	1	-0.077581	0.556113	0.819717	-0.094199	0.428446	0.22405	-0.085418	0.50282	-0.076563	0.483342	0.59183	-0.091993	0.041127	-0.267458	-0.517077	0.045572	0.123897	0.035487	-0.042933
poulg2	0.004464	-0.000919	-0.00946	0.007245	-0.031225	-0.074434	0.234269	-0.038778	-0.077581	1	-0.021885	-0.037992	0.737456	-0.034772	-0.010738	0.668182	-0.033025	0.705428	-0.047596	-0.052732	0.697755	-0.023171	0.631631	0.428352	-0.01214	-0.076319	0.056309	-0.057807
pas_bg3	0.006329	0.131659	-0.024889	0.033993	0.802119	0.565845	-0.009809	0.847279	0.556113	-0.021885	1	0.551913	-0.014408	0.663608	0.140216	-0.020761	0.616974	-0.030841	0.68079	0.371068	-0.017128	0.606106	-0.69241	-0.650339	-0.049675	0.00328	0.003662	0.026225
pad_bg3	0.017591	0.064836	-0.030289	0.011293	0.506566	0.769431	-0.009688	0.551603	0.819717	-0.037992	0.551913	1	-0.018413	0.41027	0.205857	-0.042197	0.441575	-0.035152	0.440333	0.542786	-0.036796	0.128169	-0.162062	-0.463468	0.061958	0.108504	0.056233	-0.070087
poulg3	-0.009072	0.015399	0.003277	0.048452	-0.005545	-0.073946	0.225918	-0.026486	-0.094199	0.737456	-0.014408	-0.018413	1	-0.012871	-0.009753	0.637605	-0.019838	0.664063	-0.02264	-0.051803	0.729646	-0.050593	-0.005996	0.338067	-0.006049	-0.10972	0.07166	-0.072464
pas_bd1	0.01289	0.090177	0.046871	0.000321	0.668469	0.479994	-0.006329	0.642198	0.428446	-0.034772	0.663608	0.41027	-0.012871	1	0.045277	0.046379	0.800309	0.014426	0.818983	0.404633	0.035823	0.552338	-0.916951	-0.5777	-0.005929	0.026824	0.013327	0.003409
pad_bd2	-0.013305	0.02021	0.006141	0.002004	0.06731	0.165372	-0.003275	0.13955	0.22405	-0.010738	0.140216	0.205857	-0.009753	0.045277	1	0.002061	0.092952	-0.000471	0.103129	0.226298	-0.000958	0.233081	-0.902276	-0.69961	0.088996	0.071517	-0.003622	0.00362
pould1	-0.03545	-0.003167	-0.006337	0.010641	-0.022109	-0.062466	0.210685	-0.042595	-0.085418	0.668182	-0.020761	-0.042197	0.637605	0.046379	0.002061	1	0.052399	0.750635	0.047263	0.064004	0.747617	0.023455	-0.872101	0.368969	-0.095063	-0.089828	0.017562	-0.013668
pas_bd2	0.055457	0.09739	-0.008661	0.008768	0.572376	0.470431	-0.018236	0.605723	0.50282	-0.033025	0.616974	0.441575	-0.019838	0.800309	0.092952	0.052399	1	0.004952	0.839088	0.508529	0.022418	0.530419	-0.883613	-0.504707	0.012851	0.042375	0.006793	0.008025
pould2	-0.013739	0.015348	0.002869	0.023811	-0.033006	-0.068687	0.209223	-0.045667	-0.076563	0.705428	-0.030841	-0.035152	0.664063	0.014426	-0.000471	0.750635	0.004952	1	0.008826	0.027676	0.776617	-0.034744	-0.851032	0.318508	-0.060653	-0.088326	0.058051	-0.05555
pas_bd3	0.039534	0.110813	0.015145	0.026809	0.638348	0.479067	-0.012655	0.656713	0.483342	-0.047596	0.68079	0.440333	-0.02264	0.818983	0.103129	0.047263	0.839088	0.008826	1	0.536845	0.037743	0.524652	-0.9002	-0.582262	-0.045604	0.013078	-0.000631	0.028573
pad_bd3	0.002793	0.026413	-0.050975	-0.02327	0.336616	0.497929	-0.017935	0.392524	0.59183	-0.052732	0.371068	0.542786	-0.051803	0.404633	0.226298	0.064004	0.508529	0.027676	0.536845	1	0.032962	0.309785	-0.874912	-0.650766	-0.025289	0.093951	0.042515	-0.04479
pould3	0.000763	0.021106	0.002032	0.019794	-0.021911	-0.076	0.219634	-0.031753	-0.091993	0.697755	-0.017128	-0.036796	0.729646	0.035823	-0.000958	0.747617	0.022418	0.776617	0.037743	0.032962	1	0.001588	-0.838687	0.403451	-0.031124	-0.099239	0.079071	-0.083713
dysp_repos	-0.012034	-0.232938	-0.260829	-0.534272	0.61896	0.135839	-0.04621	0.47776	0.041127	-0.023171	0.606106	0.128169	-0.050593	0.552338	0.233081	0.023455	0.530419	-0.034744	0.524652	0.309785	0.001588	1	NaN	-1	-0.370791	0.379386	0.049074	0.092002
amp_mi	0.661223	-0.603804	-0.383205	-0.744458	-0.453691	-0.200106	0.235242	0.120877	-0.267458	0.631631	-0.69241	-0.162062	-0.005996	-0.916951	-0.902276	-0.872101	-0.883613	-0.851032	-0.9002	-0.874912	-0.838687	NaN	1	NaN	1	NaN	NaN	NaN
omi	-0.000147	-0.274249	0.019112	-0.705318	-0.818376	-0.44149	0.35007	-0.610881	-0.517077	0.428352	-0.650339	-0.463468	0.338067	-0.5777	-0.69961	0.368969	-0.504707	0.318508	-0.582262	-0.650766	0.403451	-1	NaN	1	0.785817	0.399547	-0.695346	0.695346
omi_godet	0.054734	-0.047533	-0.052296	-0.126043	-0.100917	0.056856	-0.014647	-0.040286	0.045572	-0.01214	-0.049675	0.061958	-0.006049	-0.005929	0.088996	-0.095063	0.012851	-0.060653	-0.045604	-0.025289	-0.031124	-0.370791	1	0.785817	1	0.439349	0.099236	-0.099576
gly	0.130781	-0.097632	-0.006044	-0.094496	-0.027639	0.057082	-0.081714	0.002614	0.123897	-0.076319	0.00328	0.108504	-0.10972	0.026824	0.071517	-0.089828	0.042375	-0.088326	0.013078	0.093951	-0.099239	0.379386	NaN	0.399547	0.439349	1	0.11072	-0.155055
event	0.261757	-0.097516	-0.00825	0.011123	0.010005	0.031712	0.005019	-0.00086	0.035487	0.056309	0.003662	0.056233	0.07166	0.013327	-0.003622	0.017562	0.006793	0.058051	-0.000631	0.042515	0.079071	0.049074	NaN	-0.695346	0.099236	0.11072	1	-0.89027
lifespan	-0.279298	0.132713	0.009679	0.024491	0.028915	-0.036576	-0.006717	0.03069	-0.042933	-0.057807	0.026225	-0.070087	-0.072464	0.003409	0.00362	-0.013668	0.008025	-0.05555	0.028573	-0.04479	-0.083713	0.092002	NaN	0.695346	-0.099576	-0.155055	-0.89027	1

TABLE B.1: Pearson Features Correlation Values