People's Democratic Republic of Algeria
Ministry of Higher Education

University of Dr. TAHAR MOULAY SAIDA

Faculty : TECHNOLOGIE

DEPARTEMENT : Computer Science

Master's Thesis

OPTION :MICR

Thesis

Big Data Analytics in Ehealth

Presented By: **Salhi Belgacem**

Supervised By : Pr.Amine Abdelmalek

Promotion : June 2018

To my family , Friends, and

My dear Supervisor Pr.Amine Abdelmalek

Abstract

The proliferation of data warehouses and the rise of multimedia, social media and the Internet of Things (IoT) generate an increasing volume of structured, semi-structured and unstructured data. Towards the investigation of these large volumes of data, big data and data analytics have become emerging research fields, attracting the attention of the academia, industry and governments. Researchers, entrepreneurs, decision makers and problem solvers view ‘big data’ as the tool to revolutionize various industries and sectors, such as business, healthcare, retail, research, education and public administration.

Index Terms—Big Data, Data Analytics , IoT, Healthcare

Summary

Table of Figures

References

Introduction

Big Data revolution is making our lives different . The last years have seen a massive generation of data that had a total impact on our way of life and health care . The healthcare industry welcomed these news just like any other important field in our life , The hospitals process and generate huge data for every patient but accessing, managing and interpreting the data are critical to create actionable insights for better care and efficiency. Earlier physicians used their judgements to make treatment decisions, but now things have changed. Specialists review the data and make an informed decision about a patient's treatment.

IOT adds a great value to the healthcare industry. Devices that generate data about a person's health and send it to the cloud will lead to a plethora of insights about an individual's heart rate, weight, blood pressure, lifestyle and much more. Big Data allows real time monitoring of patients, which leads to proactive care. Sensors and wearable devices will collect patient health data even from home. This data is monitored by healthcare institutions to provide remote health alerts and lifesaving insights to their patients.

We gladly present and put into details our contribution in this new revolutionary invention, building a full system of receiving EMG data, processing it, and with the help of big data techniques , we can have more insight and knowledge from the analytics

Going step by step, defining and explaining the new era of analytics with Spark streaming, and explaining the magic that micro-processors and sensors are offering to the health domain, and giving full details of the implementation for both hardware and software parts , at the very end we compare and discuss the results of our models and finishing with a conclusion

CHAPTER ONE : BIG DATA

I .Big Data

I.1 Introduction

Big data and analytics are hot topics in both the popular and business press. Today, many organizations are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as “big data” because of its volume, the velocity with which it arrives, and the variety of forms it takes. Big data is creating a new generation of decision support data management. Businesses are recognizing the potential value of this data and are putting the technologies, people, and processes in place to capitalize on the opportunities. A key to deriving value from big data is the use of analytics. Collecting and storing big data creates little value; it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value. Big data and analytics are intertwined, but analytics is not new. Many analytic techniques, such as regression analysis, simulation, and machine learning, have been available for many years. Even the value in analyzing unstructured data such as e-mail and documents has been well understood. What is new is the coming together of advances in computer technology and software, new sources of data (e.g., social media), and business opportunity. This confluence has created the current interest and opportunities in big data analytics. It is even spawning a new area of practice and study called “data science” that encompasses the techniques, tools, technologies, and processes for making sense out of big data.

I.2 . Definitions

Words and phrases are the fundamental components of construction of any language, much as genes and cells in biology life. Words are what we use to express ideas, so tracing their origin, development and spread is not merely an academic pursuit but a window into a society’s intellectual evolution.“BIG DATA” is one of the terms that made the buzz in the last few years. This term hadspread up to describe the explosion of data over the web. Meanwhile, there is no exact definition of the concept of Big Data. Some experts define it as more than can fit on a personal computer. Others go further more by defining it as not only the massive amounts of data but the tools that show the patterns within it. While others has chosen to be more metaphorical by defining BIG DATA as the process of helping the planet grow a nervous system in which humans are just another type of sensors. However, Rick Smolan, writer and editor of the book “The Human Face of Big

Data", had wrote an essay on that book entitled "A Planetary Nervous System" in which he had defined BIG DATA as: ". . . an extraordinary knowledge revolution that is sweeping, almost, invisibly through business, academia, government, health care, and everyday life. . "

As mentioned before, big data had been widely studied. Many definitions had been written and several experts had given their point of view. Some of the definitions were technical while others were abstractive. This subsection is reserved to the presentation of some tech- nical definitions.

Definition 1 :(Webopedia.com) According to Vangie Beal 1 , big data is a buzzword, or catch-phrase, meaning a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. However, she had gone further more by defining it as more than just large volumes of data but it is a technology that an organization often requires to handle the large amounts of data and storage facilities.

Definition 2 : (Techopedia.com) Cory Janssen – an editor at Techopedia.com – had defined big data as the process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines.

Definition 3 : (Forrester.com) Mike Gualtieri had posted what he called "a pragmatic definition of big data" in which he had defined it as the frontier of a firm's ability to store, process, and access all the data it needs to operate effectively, make decisions, reduce risks, and serve customers. In other words, the exponential growth of data containing nonobvious information that entities can discover to improve outcomes makes it continuously difficult to manage.

I.3 Big Data Challenges

Objectively, the main point of the V-based characterization of big data is to highlight its most serious challenges: the capture, cleaning, curation, integration, storage, processing, indexing, search, sharing, transfer, mining, analysis, and visualization of large volumes of fast-moving highly complex data

I.4 Big data Characteristics

In General, there are few aspects and characteristics that define Big Data. These are called the big V's. In early days there were 4 V's that defined Big Data. These 4 V's can be defined as Volume, Velocity, Variety & Veracity. Recently this was changed as 7 V's as Variability, Visualization and Value was added to the list of big V's.

Let's first discuss the main V's

Volume

This is the main characteristic that define Big Data as **"Big"**. we find that storage in Gigabytes is actually not enough. Due to this reason, in today's world, data warehouse storage is discussed in

Exabytes, Zettabytes and Yottabytes. With the introduction of new technology, and the number of new devices that come in to use, different types of data are generated and stored in an increasing manner each year. For example, the world population is somewhere around 7 Billion, and according to statistics, almost 6 billion people use mobile phones. Just imagine how much data is generated and uploaded to Internet every single second. So the volume of data that is used in data warehouses and in Big data field is so large. That is why "Volume" as a big "V" is most important.

Velocity

Velocity can be simply defined as the speed of change. In other words Velocity means how fast the data needs to be accessed and processed. In some situations, data is updated in batches during the night daily, and may be monthly or quarterly. As another example, just imagine, how fast Facebook, Twitter & YouTube data are getting updated? People around the world share some sort of thing every minute and may be every second. In that case, speed of change is really fast. Now a days, trends are on real-time data requirements so that daily night batches are not at all useful. In today's world required speed of change is almost real-time.

Variety

Variety can be simply defined as having different forms of data sources. Though it is explained simply, it is one of the biggest challenges of big data. When talking about different forms of data sources, data can come from structured data sources and unstructured data sources. However with unstructured data the process is very much different. Unstructured data does not have any predefined set of rules that say that a particular data element should be in this particular format etc. Unstructured data can have wanted and unwanted data filled in it without any set of rules that separates them from the important data. For example, just think of Video files, text messages, Facebook comments, Twitter feeds etc. These can have values in many different ways and sometimes not even in proper language. Those are ways people share their ideas and thoughts and those do not have any type of rule. Unstructured data can also be stored as images, audio files, web pages etc. To make it more complicated, these data change very fast as explained earlier. Though it is the most challenging aspect, it is a fundamental concept in big data and the best way to understand unstructured data is by comparing it with structured data. The goal of big data is to use technology and make unstructured data understandable.

Veracity

Veracity refers to the trustworthiness of the data that is being used. As we know, data warehouse is used for decision making and focuses on top management. Can a top level manager, or a board director of the company can rely on the data that is being given to them for their critical decision making?

Variability

Variability is different from Variety in the first place. Variability mostly focuses on properly understanding and interpreting the correct meanings of raw data that depends on its context. This is

mostly required when working on Natural Language Processing. As we all know, in natural languages such as English, there are some words that can multiple meanings and the exact meaning is depending on the context it is being used. For example the word "Great" gives an positive idea, however if it is said as "Greatly disappointed", that doesn't mean as positive. Due to this reason, the exact meaning needs to be properly interpreted for an organization to do proper analysis on their business .Of course doing developing such algorithms are quite complex and challenging, but it is not impossible with current technology that is available now.

Visualization

Visualization refers to how the data is presented to the management for their decision making. Data can be presented in many different ways such as long excel files with rows and columns of data, word docs, graphical charts etc. Whatever the format is, the data should be easily readable, understandable and accessible. This is why data visualization is important. Have a guess yourself, which one is easily readable, the excel sheet full of data or a nice graphical view or chart that represents the same set of data? Of Course, the graphical view is the best. Having a graphical view of critical data for a board of directors to take a critical decision will make their decisions much effective and accurate rather than allowing them to dig into large sheets of long data.

Value

Lastly the Value which is known as the end game. Value being the last one on the line, it is important to understand that the organization needs to get some sort of value after the immense efforts and resources spend on the above V's. Big Data can provide the business with immense value if it is done correctly and each step is properly processed. We all know that data on its own is actually worthless. data is worth when its integrated and analyzed in many different views and that's what generates value by giving the ability to make effective, efficient & accurate decision making on the opportunities and threats of the organization. Once the organization get the grip of what is done, the power of big data is limitless.

1.5.Challenges “big data” is posing

Data is growing and moving very quickly. Mobile devices, sensors, and social media all are contributing to this explosion. However, to gather, store, and derive value from this data is what that gives the organizations a competitive edge. Information can be thought of as the critical business asset of “big data.” With so much explosion in data, the need for a backup and disaster recovery process can’t be ruled out. Demand for data storage is expected to increase, which is the reason why companies like IBM (IBM), HP (HPQ), EMC (EMC), Cisco (CSCO), and DELL (DELL) are pouring funds into “big data” appliances.

1.6. Big data Analytics

1.61 What Is Analytics?

It is helpful to recognize that the term analytics is not used consistently; it is used in at least three different yet related ways [Watson, 2013a]. A starting point for understanding analytics is to explore its roots. Decision support systems (DSS) in the 1970s were the first systems to support decision making [Power, 2007]. DSS came to be used as a description for an application and an academic discipline. Over time, additional decision support applications such as executive information systems, online analytical processing (OLAP), and dashboards/scorecards became popular. Then in the 1990s, Howard Dresner, an analyst at Gartner, popularized the term business intelligence. A typical definition is that “BI is a broad category of applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help business users make better decisions” [Watson, 2009a, p. 491]. With this definition, BI can be viewed as an umbrella term for all applications that support decision making, and this is how it is interpreted in industry and, increasingly, in academia. BI evolved from DSS, and one could argue that analytics evolved from BI (at least in terms of terminology). Thus, analytics is an umbrella term for data analysis applications. BI can also be viewed as “getting data in” (to a data mart or warehouse) and “getting data out” (analyzing the data that is stored). A second interpretation of analytics is that it is the “getting data out” part of BI. The third interpretation is that analytics is the use of “rocket science” algorithms (e.g., machine learning, neural networks) to analyze data. These different takes on analytics do not normally cause much confusion,

1.6.2 Big Data Programming Models

Big Data programming models represent the style of programming and present the interfaces paradigm for developers to write big data applications and programs. Programming models normally the core feature of big data frameworks as they implicitly affects the execution model of big data processing engines and also drives the way for users to express and construct the big data applications and programs

A programming model is the fundamental style and interfaces for developers to write computing programs and applications. In big data programming, users focus on writing data-driven parallel programs which can be executed on large scale and distributed environments. There have been a variety of programming models being introduced for big data with different focus and advantages.

1.6.3 MapReduce

MapReduce the current defacto framework/paradigm for writing data-centric parallel applications in

both industry and academia. MapReduce is inspired by the commonly used functions - Map and Reduce

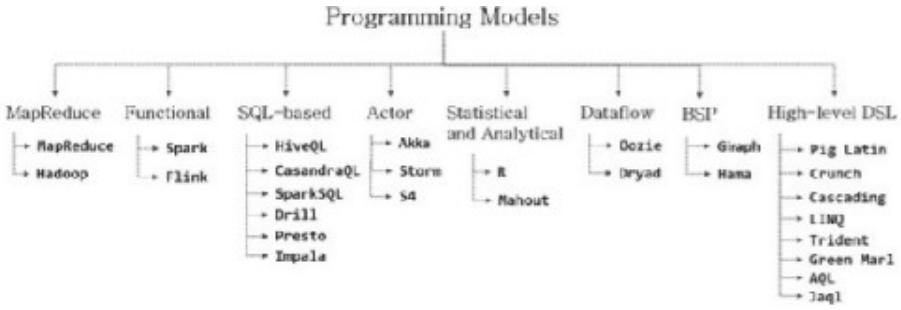

Figure 1.1:Taxonomy of programming models

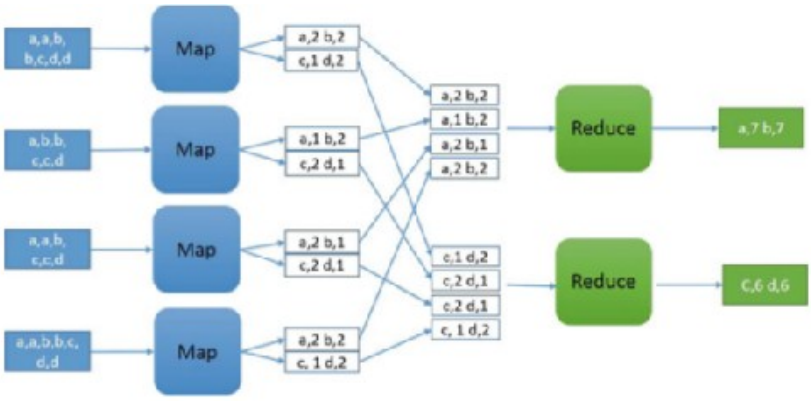

Figure 1.2: MapReduce Paradigm

• Map and Reduce functions. A MapReduce program contains a Map function doing the parallel transformation and a Reduce function doing the parallel aggregation and summary of the job. Between Map and Reduce an implied Shuffle step is responsible for grouping and sorting the Mapped results and then feeding it into the Reduce step.

• Simple paradigm. In MapReduce programming, users only need to write the logic of Mapper and Reducer while the logic of shuffling, partitioning and sorting is automatically done by the execution engine. Complex applications and algorithms can be implemented by connecting a sequence of MapReduce jobs. Due to this simple programming paradigm, it is much more convenient to write data-driven parallel applications, because users only need to consider the logic of processing data in each Mapper and Reducer without worrying about how to parallelize and coordinate the jobs.

• Key-Value based. In MapReduce, both input and output data are considered as Key-Value pairs with different types. This design is because of the requirements of parallelization and scalability. Key-value pairs can be easily partitioned and distributed to be processed on distributed clusters.

• Parallelable and Scalable. Both Map and Reduce functions are designed to facilitate parallelization, so MapReduce applications are generally linearly-scalable to thousands of nodes.

1.6.4 Hadoop

Hadoop is the open-source implementation of Google's MapReduce paradigm. The native programming primitives in Hadoop are Mapper and Reducer interfaces which can be implemented by programmers with their actual logic of processing map and reduce stage transformation and processing. To support more complicated applications, users may need to chain a sequence of MapReduce jobs each of which is responsible for a processing module with well defined functionality. Hadoop is mainly implemented in Java, therefore, the map and reduce functions are wrapped as two interfaces called Mapper and Reducer. The Mapper contains the logic of processing each key-value pair from the input. The Reducer contains the logic for processing a set of values for each key. Programmers build their MapReduce application by implementing those two interfaces and chaining them as an execution pipeline.

1.6.5 Spark

Spark provides programmers a functional programming paradigm with data-centric programming interfaces based on its built-in data model - resilient distributed dataset (RDD). Spark was developed in response to the limitations of the MapReduce paradigm, which forces distributed programs to be written in a linear and coarsely- defined dataflow as a chain of connected Mapper and Reducer tasks. In Spark, programs are represented as RDD transformation DAGs

Programmers are facilitated by using a rich set of high-level function primitives, actions and transformations to implement complicated algorithms in a much easier and compact way. In addition, Spark provides data centric operations such as sampling and caching to facilitate data-centric programming from different aspects. Spark is well known for its support of rich functional transformations and actions, Basically, programming primitives in Spark just look like general functional programming interfaces by hiding complex operations such as data partitioning, distribution and parallelization to programmers and leaving them to the cluster side.

1.7 Programming Platforms for Big Data Analysis

The necessity of increased computing speed and capacity offered by big data programming platforms has led to constantly evolving system architectures, novel development environments, and multiple third-party software libraries and application packages. Now, we are in an era where businesses, government sectors, small and big organizations have all realized the potential of big data analysis. The great demand for big data analysis systems is giving a thrust to the research and development in this area. Large amounts of data have to be handled in a parallel and distributed way wherein, and the computations have to be distributed across many machines in order to be finished in a reasonable amount of time. The issue of how the computation can be parallelized, how data is

distributed and how failures are handled in such a wide distribution are compelling, and call for special programming platforms for big data analysis.

1.8 . Big Data in the Era of Internet of Things (IoT)

Internet of things (IoT) is an emerging paradigm in the science of computers and technology in general. In the last years it has invaded our lives and is gaining ground as one of the most promising technologies. According to the European Commission, IoT involves “Things having identities and virtual personalities operating in smart spaces using intelligent interfaces to connect and communicate within social, environmental, and user contexts”. IoT has been in the centre of focus for quite some years now, but in the last years it has actually become reality allowing “people and things to be connected Anytime, Anyplace, with Anything and Anyone, ideally using Any path/network and Any service”

Internet of Things has a big impact in communities. Smart cities making use of sensors will be able to monitor air pollution or traffic, better manage waste and provide smart agriculture. In the case of health, sensors on patients will monitor their health condition and initiate an alarm in critical cases. IoT will also change retail, logistics and transportation whereas it can have a great impact on energy management. As IoT gradually becomes reality, novel appliances appear in the market, promising to make our homes smarter and our lives easier: e.g. a smart fork helping its owner monitor and track his eating habits or a smart lighting system adjusting automatically to the outside light intensity, the presence of people in the house and their preferences . But not only are homes made smart, but also our accessories have changed: small wearable devices monitor our everyday movements and are capable of calculating the steps we made and the calories we consumed, or can even store details about our sleeping habits .

Moreover, smart watches with e-SIM technology can substitute our mobile phones and jewellery has turned into powerful gadgets (aka smart jewellery) . This rapid evolution of non-mobile gadgets and mobile wearables has become the beginning of a new era! Imagine small, wireless devices on cars, in homes, even on clothes or food items . The tracking of merchandise will be automatically monitored, the conditions at which food supply is stored will also be recorded and our clothes will

be separated automatically—based on colour and textile—for the smart washing machines. But then, the production of IoT data will be enormous! So, the prevalence of Internet in our lives and especially of Internet of the Things will cause an explosion of data. According to a report from International Data Corporation (IDC) the overall created and copied data volume of the world was 1.8ZB (≈10 21 B) in 2011 which increased by 9 times within five years . On top of that, Internet companies handle each day huge volumes of data, e.g. Google processes hundreds of PB and generates log data of over 10 PB per month.

.8.2 (eHealth Big Data) Analytics in Internet of Things

On the one hand, in the health sector, we are faced with many problems, for which they come to give solutions the IoT and the new technologies. Some of these problems is the rapidly aging due to the low birth rate (demography problem), the chronic diseases due to the increasing aging of the population (such as hypertension, heart failure, diabetes mellitus, etc.), the rate diseases (e.g. Alzheimer's disease), the hereditary diseases, the lack of health personnel and the health infrastructure, the difficult treatment of emergency cases (e.g. the accidents, the emergency obstetric care and so on), the organizational problem, the patients with mild disease (avian, etc.) that they need no monitoring and binding site on the already congested hospitals' infrastructure, and also, the corruption of information over time . On the other hand, from our previous studies we know that the rapid development of IoT and CC has brought the rapid growth of data. So, we are faced with major challenges related with the management, the analysis, and the transfer of such data. These challenges of large-scale data mainly concern: their representation, reducing the redundancy that exists in them, the quality and the variety, the management of the life cycle, the confidentiality, their expendability, the energy management, the heterogeneity, the speed and the accuracy, the privacy and the security, the storing of them, the extracted knowledge from them, the creation or development of their analysis tools and algorithms or techniques as well as, other serious issues that need total improvement. The latest findings therefore show us that there are some "gaps" in the way in which such data are transmitted through the levels of management, analysis, and transportation, but also, some problems that arise from their use, and which we will try to optimize by proposing new techniques and new algorithmic solutions.

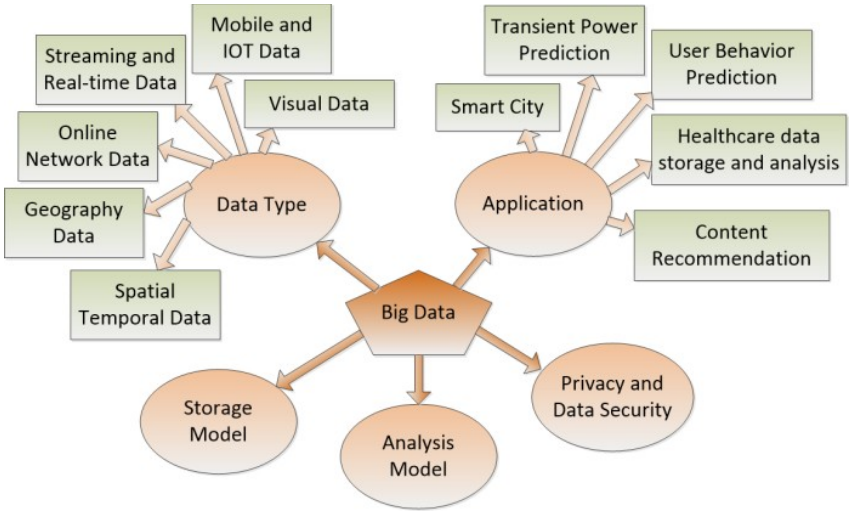

Figure 1.3 : Classification of BigData and IOT Literature

1.9. DATA TYPES

The era of big data has produced a variety of datasets from different sources in different domains. These datasets consist of multiple modalities, each of which has a different representation, distribution, scale, and density. How to unlock the power of knowledge from disparate datasets is of

paramount importance in big data research, essentially distinguishing big data from traditional data mining tasks.

1.9.1. Mobile and IoT Data

Another trend in network big data is the analysis of mobile and IoT data. With the development of 5G technology, converged mobile networks have resulted in significant improvements in machine-to-machine communications performance. Integrated mobile webs share unlicensed spectrum bands in cellulite networks, such as Long Term Evolution-Advanced, by using cognitive radio technology. This network generates large volumes of data, compared to former mobile networks . In addition to the increased volume of mobile data, the IoT also generates large amounts of data in this new context. Despite this large volume of data, the sensing elements of wireless body area networks (WBANs) to a certain degree restricts power use. The majority of researchers attach importance to energy efficiency in media access control (MAC) agreements in lengthening the lifetime of the sensors. One study addresses the recognition of classifications of power consumption attacks in MAC agreements in WBANs. It describes the straightforward operation of the attacks, resulting in power consumption in a variety of MAC agreements. This work is a good reference for research on the power efficiency of MAC agreements in WBANs of the future. Understanding the connection and interaction of mobile OSN data has been continuously broadened, although network big data in the IoT is a relatively new field. The data structure of the analysis is more likely to be Not Only Structured Query Language (NoSQL), which is adopted by many IoT systems. Some studies design the expected functions of a big loader and a convenient loading NoSQL system. The system allows the standard conceptual program to be loaded and lets the standard sources from which the data are supposed to be collected meet its requirements; finally, this study provides feasible strategies for the choice of NoSQL system where the conceptual program can be arranged well.

1.9.2 -Streaming and Real-time Data

Accompanied by the rise in online streaming services, network big data has evolved from spatial–temporal data to real-time spatial–temporal data. Network surveys in general require ongoing data analysis owing to constant renewal of reports and statistics over large-capacity data streams. In one study, researchers introduced DBStream, a system based on SQL that relies on surveys for continuous data analysis. They also discuss the respective properties of DBStream and the collateral data handling engine Spark. It is suggested that, on some occasions, a single DBStream node is capable of surpassing a set of 10 Spark nodes due to the renewed network survey capacity. The epoch of big data has begun, and much of the data are used to analyze the risks of a variety of industrial applications. There are technological trials in the collection of big data in a complex indoor industrial environment. Indoor wireless sensor network (WSN) technology is capable of overcoming such restrictions by gathering big data obtained from source nodes. The data are transferred to a data center, at present. In the study, representative housing, bureaus, and manufacturing environments were selected. Through analysis of tested data, it is possible to obtain

signal transferring features of an indoor WSN. On the basis of these features, a big data collection algorithm that relies on an indoor WSN was put forward for the analysis of industrial risk processing. City traffic also changes in real time. Traffic data are regarded as worthwhile resources in networks of vehicle.

Highlighting the significance of a survey of big data, an effective framework was put forward for current-time network data in vehicle networks . The system, in fact, reflects the newest trends in big-data paradigms. The framework put forward is composed of concentrated data memory principles for a series of processes, and dispersed data memory principles for stream processing in real time. The present big data streams from social networks and other associated sensor networks display the potential for relying on each other, thus enabling a special approach to the analysis of extended figures. Data from these figures are often gathered from data servers in various geographic locations, making it appropriate for dispersed handling in the cloud. While many measurements for large-scale immobile figure analysis have been brought forward, providing current-time analysis of the dynamics of social correlations requires novel methods that leverage increased stream handling and figure analysis in flexible cloud environments. The scope contains multiple fields, involving supervision, anti-terrorist applications, and public health supervision. Currently, space-borne sensors channel nearly constant streams of Earth-survey datasets. These tremendous multi-modal streams increase at a rapid rate, presently reaching several petabytes of satellite files. An extended platform for both geography and space was devised, developed, and assessed for online and current-time gains of worthwhile content from big Earth-survey data.

1.9.3- Challenges in Data

Each different data domain raises particular challenges that, properly addressed, may have an important impact on next-generation big data systems. In the first place, online network data is still waiting for better models, with increased support from sociologists. Mobile data and the IoT, which are generating large amounts of data, would benefit from the adoption of a big data infrastructure able to store and process information in current IoT infrastructures. As for geography data, a major trend seems to be to offer efficient integration among geographic data with records from the OSN domain, demanding efficient infrastructures able to meet the speed requirements typical of these domains. One challenge imposed by spatial data is the definition of proper mining algorithms that can be applied to special data; those algorithms could benefit from more efficient time-changing data. Besides, streaming and real-time data challenge current infrastructures, transforming current offline applications into online ecosystems, thus requiring the development of new algorithms that take into account offline and online data. Lastly, the ever-increasing amounts of image data challenge current learning algorithms to extract information, and also demand new algorithms to semantically classify and index images. In all these cases, all approaches would benefit from increased performance in big data infrastructures. By increasing efficiency in the different data domains, one can see how the amount of functionality improves; this is of special interest for the next generation of big data systems, which should integrate online and offline data efficiently.

1.10. APPLICATIONS

A. Transient Power Prediction

The prediction of transient power is valid in both distributed and streaming data. Machine learning was used in the study. In the classifier cultivation stage, researchers regard the tremendous amount of data from the past as a dispersed study target, and establish evaluation principles regularly. Zhiwei et al. designed a naive Bayes–category approach based on MapReduce handling, creating a map-and-decrease procedures method for calculating the chance rate of being tested in advance and the chance rate for conditions in dispersed means.

B. User Behavior Prediction

Many of the network big data predictions are based on data from OSNs. Big data is used for predictions based on ranked data, such as elections, car performance, and other areas in business and politics. One study discussed modeling and analysis approaches to democracy, as well as various cases of big data from elections; scenarios in established democracies such as the United States and Canada, and new democracies such as Tunisia, were studied. Another study gathered and explored user practices on Facebook. The model is capable of arranging entities with effectiveness and efficiency (for example, presidential candidates, specialized sport groups, and musical bands) according to their popularity .

C. Healthcare data storage and analysis

Big data in health and biology to tackle the challenges in new models is becoming significant. One study introduced two uses of mHealth, which gathers electronic medical records that are used for health services terminals. One is a blended system that enhances the user experience in high-pressured oxygen halls using VR glasses, which creates the feeling of being inside it. The other is a sound interaction game that is used by patients as a possible measurement for supplementary recovery tools. It is possible to analyze recordings of the sounds made by patients to assess long-term recovery results and further forecast the recovery process.

E. Smart City

A 3D Shenzhen city web platform based on a network virtual reality geographic information system (GIS) was put forward . A 3D worldwide browser is applied to load different kinds of required data from a city, such as 3D construction model data, inhabitants' messages, and traffic data from the past and present. These data are used to analyze and visualize city information on a 3D platform. A large number of messages are capable of being visualized on this platform, and a navigational project, taking the GIS as the premise, makes it possible to obtain a variety of data sources that are securable. The enhancive requirement for fluidity has resulted in great changes in fundamental facilities in transportation . Possessing certain features, such as a large scale, diversified foresee-ability, and timeliness, city traffic data represent the scope of big data Traffic visual analysis systems based on a virtual reality GIS represent the standard by which traffic data are controlled and developed. Aside from the fundamental GIS mutual functions, the system put forward also

contains smart functions for visual analysis and forecast accuracy. There is also a study that addresses the concept of smart and connected communities (SCCs), which are no longer solely defined as smart cities . Big data analytics in cyber-physical systems, which are engineered systems that are built from, and depend upon, the seamless integration of computational algorithms and physical components, will enable the move from the IoT to real-time control and towards the SCC. SCCs were conceived to represent earlier requirements (e.g., protection and redevelopment) in a cooperative way, and the requirements for current living (habitability) and planning for the future (sustainability). In consequence, the final objective of SCCs is to improve habitability, protection, redevelopment, and the attainability of a desirable society. This study uses mobile crowdsourcing and cyber-physical cloud computing for these two essential IoT technologies.

1.11.Conclusion

The term big data occurs more frequently now than ever before. A large number of fields and subjects, ranging from everyday life to traditional research fields (i.e., geography and transportation, biology and chemistry, medicine and rehabilitation) involve big data problems. The popularizing of various types of network has diversified types, issues, and solutions for big data more than ever before.

Big data is not always useful. The actual challenge of big data is not in collecting it, but in managing it as well as making sense of it. When we work on big data, it is crucial to determine whether the benefits outweigh the costs of storage and maintenance. Several tools are being designed to better understand the role of huge amounts of data in improving business. Researchers and practitioners are trying to look into the future of big data to extract more benefits.

CHAPTER TWO : EHEALTH

2.1.Introduction

The latest achievements in the information and communication technologies significantly change the relation between a user and a computer system. More and more ubiquitous computing is becoming the new reality, in which devices enhanced with intelligence and communication are all around us and these devices change the way we are interacting with our environment . eHealth solutions are just a part of this new “ecosystem” dedicated for medical prevention, supervision and treatment . In this context it is obvious that it is necessary to develop new technologies that allow integration and inter-communication between different devices or simply “things” through the Internet. Medical devices and mostly the portable ones should be part of this new approach. Medical supervision and healthcare services provisioning is about to change in the near future as result of these new technologies. Connecting intelligent devices on a network or even on the Internet is not a new subject. There are many examples of process control applications in which different automation devices (e.g. intelligent sensors, actuators, regulators, PLCs) are exchanging information through a network, in order to implement monitoring or direct control functionalities on large scale distributed systems. There are network protocols (known as industrial networks) specially developed for this purpose; these protocols satisfy the real-time, safety and security requirements of such applications. But these solutions are more or less particular for a given system and the access to devices is strictly restricted. Devices are not acting as individuals with an autonomous behavior as it is supposed to be in the IoT philosophy. For instance, an intelligent sensor (e.g. room temperature or movement sensor) may be part of a building automation system (e.g. heat regulation), a security system or an eHealth solution, at the same time. In such a scenario, a device should be accessed as an individual “thing” on the Internet, of course under a well defined access control policy.

In the eHealth area there are a number of examples of connecting medical devices on the Internet in order to perform different remote medical services such as: remote patients’ monitoring, elderly persons’ supervision, on-line medical consultations, or even robotic arm control for surgical interventions . A real “boost” in this area was given by the latest spread of miniaturized portable medical devices and gadgets. These devices may be used for continuous measurement of medical parameters (e.g. ECG, blood pressure, temperature), for activity recognition and monitoring or for remotely-made medical evaluations . Again there are a number of particular solutions present on the market, but with very restricted accessibility. In most cases, the communication protocol is not open and a given device cannot be integrated in other (or multiple) applications. We consider that the Internet of Things is in a incipient stage and in order to connect a huge number of “things” on the

Internet, many things have to be changed and new, more adequate protocols and technologies must be developed. The eHealth domain, as an important beneficiary of these developments should set some requirements regarding data acquisition, access control, security and safety.

2.2.Iot In Healthcare

In this section, we briefly discuss some of the applications of IoT eHealth ecosystem and architecture.

A. mHealth

Thanks to the cloud platform, whether patients are traveling on the road or relaxing at home, the health information always stays with them, and they can access it using mHealth smartphone apps or web-based cloud dashboard. The care givers can also leverage the platform and mHealth smartphone apps with P2P video/audio capabilities to help and guide patients at anytime, and anywhere. Patients conveniently get diagnosis, treatment, as well as prescription refills from care givers whenever they require it. Considering the fact that health professionals have access to the holistic health database of patients on the cloud platform 24/7, patients receive the most accurate treatment possible. Therefore, the proposed holistic IoT eHealth ecosystem ensures that patients will always receive the best care available .

B. IoT in Ambient Assisted Living

The proposed IoT eHealth ecosystem allows the incapacitated and aging individuals to live longer and healthier. It is observed that the share of the aging population is significantly increasing and it is estimated that about 20% of the world population will be over 60 years old by 2050 .At the same time, aging brings fast growth of various chronic diseases (e.g., stroke, cancer, type II diabetes, and obesity). Thanks to the growing acceptance of technology, in Abient Assitet Living (AAL), IoT enables indoor positioning as well as location-aware real-time monitoring of living parameters (e.g., heart rate) and environmental conditions

C. IoT Medication

IoT can also enable us to detect adherence to medication and to prevent fatal Adverse Drugs Reaction (ADR) , combination of smart pill bottle technologies, wearable audio sensors, and classification techniques is capable of assessment of medication adherence with high accuracy. As stated in , ADR rate is about 6.5% in worldwide hospitals. IoT medication with help of NFC-enabled smart pill bottle, cloud-based Electronic Health Record (EHR), and a knowledge-based system can significantly prevent the adverse consequences of wrong drug usage.

D. IoT for Individuals with Disabilities and Special Needs

In 2011, WHO conducted its first survey on disability and reported that more than a billion people (equivalent to 15% population of world) live with disability. IoT eHealth can bring a great deal of comfort to this vulnerable population and enhance their life significantly through automated, timely, reliable resistive technologies. For example, a number of smart gloves are developed with low-cost inertia sensors enabling hearing loss to communicate with those who are not very familiar with the American Sign Language (ASL) . Smartwatches are being used for patients with speech disorders to train their speech functions in remote settings . IoT systems are built in smart cities for improving the access for wheelchair users who face mobility related challenges in everyday life. IoT also can help gather information from individuals about their special needs remotely in their comfortable environments Schools can leverage IoT platforms to make special needs education more efficient and accessible to the children with disabilities. The Wireless Nano Retina Eyeglasses is a kind of IoT device that is already out in the market, allowing to communicate with retina implants in blind individuals for real-time fine-tuning of the visuals . E. IoT for Smart Medical Implants Beyond wearable devices, IoT eHealth brings new promises to implantable medical devices that are highly sophisticated, miniaturized, reliable systems inserted inside the body to restore or enhance the human functions . Some of the examples of electronics implants are: 1) pacemakers that stimulate the heart muscle to help regulating its rhythm , deep brain stimulation (DBS) systems that are also known as brain pacemakers provide highly-controlled electrical impulses into deep brain regions to reduce movements symptoms in motor disorders such as Parkinson's disease and Essential Tremors , and 3) cochlea implants that stimulate electrodes placed inside the inner-ear to store hearing functions. Such electronics implants have miniaturized circuits including an analog front-end, a micro-controller, and a battery-based power management module. The paradigm of IoT has been explored in the area of medical implants to make them more contextual, power-efficient,and secure. For example, there are ongoing research efforts in optimizing deep brain simulations through limb-worn inertia sensors . IoT provides a basic framework for tele-management and programming of cochlea implants that are non-trivial tasks for patients who generally have to travel to implant centers for programming services .F. IoT-based Early Warning Score (EWS) Due to large volume of incoming medical data generated by a wide-range of bio-sensors and hundreds of thousands of patients, it is infeasible to monitor every patient directly. To assist the health professionals, an IoT-aware Early Warning Score System (EWS) can be utilized to effectively detect and forecast deterioration of patients' conditions early in time. The basic idea behind EWS is to process and analyze six cardinal vital signs including temperature, respiratory rate, systolic blood pressure, pulse rate, oxygen saturation, and level of consciousness. The measured vital signs are then mapped to a composite patient deterioration risk score. Indeed, each vital sign is processed and a score is assigned to it in such a way that the magnitude of the score represents the deviation of the parameter form its corresponding norm. Combining all the scores leads to a composite reflecting the overall deterioration risk of the patients. Note that EWS is an established approach which is widely used by many hospitals across the world. Several studies have shown that EWS can predict complications of patients about 24 hours in advance . In this context, the proposed IoT eHealth ecosystem enables

caregivers to continuously collect vital signs remotely and compute the deterioration risk of patients entirely in an automatic way remotely. This revolutionary approach potentially is capable of identifications of deteriorating patients early in time, and hence it can save lives and reduce mortality of patients. G. IoT-based Anomaly Detection EWS has two shortcomings. First, it does not consider full spectrum of bio-signals such as Cutaneous water/sweat. Moreover, it is developed based on supervised machine learning techniques, and thus it might not be able to effectively capture those abnormalities that are not within the knowledge of the supervisor . Indeed, an accurate anomaly detection system requires to learn continuously over time. To overcome these critical issues, we propose to use an accurate anomaly detection system that learns and evolves continuously over time. This system is based on Hierarchical Temporal Memory (HTM), a biologically inspired machine learning unsupervised intelligence technology . Thanks to this technique, health professionals can quickly identify temporal anomalies which might be a sign of severe problems such as heart attack, and stroke. Fig. 8 shows the key steps that the system uses to detect anomalies. Time series bio-signals captured by connected sensors and IoT eHealth devices are forwarded to fog nodes. Fog nodes filter, process, extract features, and compress the raw data. Next, the processed data and its corresponding time stamp are transferred to the cloud using a secure connection. Then, an encoder converts the received data to a Sparse Distributed Representation (SDR). The generated SDRs are fed into the HTM machine learning module. HTM indeed tries to mimic the behavior of neocortex of brain by learning the temporal patterns of SDRs continuously. Over time, HTM generates an adaptive sophisticated model to predict the next incoming sequence of SDRs. If the incoming sequence of SDRs does not match with what predicted, an anomaly alarm is triggered. In order to reduce the false positives mainly due
to variations in data and/or noise, HTM computes a time-varying average of the error and compares it with the distribution of errors. Based on this comparison, HTM can also estimate the accuracy of the predicted anomaly and warning. Finally, the outputs of the risk analysis and warnings can be visualized in the dashboard and also they can be directed to patients and those engaged in the patients care.

H. Population Health Management

As estimated by IBM, medical data is expected to double every 73 days by 2020 . Big data analytics can enable us to understand the medical data and extract deep insight in order to personalize the care plan, perform early interventions, improve the case and outcome, while reduce the healthcare cost. In this context, as reported by IBM , Medical Center of Columbia University applied big data analytics to examine medical data from patients who suffer from bleeding strokes in order to forecast the major complications 48 hours in advance compared to the traditional techniques . Rizzoli Orthopedic Institute utilized big data analytics to understand the clinical variability within families with hereditary bone diseases. Their technique results up to 30 percent decrease in annual patient hospitalizations. The Hospital for Sick Children studied different big data analytic methods to process several vital signs of patients in order to predict hospital infection up to 24 hours in advance than traditional methods. In addition, several machine learning techniques have been proposed automatically to detect psycho-physiological stress from bio-signals such as accelerometer

and skin conductance .

2.3.Architechture

In this section, we explain a holistic multi-layer IoT eHealth ecosystem/architecture that can fundamentally change the way organizations and caregivers deliver wellness and health., this system consists of three main layers: IoT eHealth Device Layer, IoT eHealth Fog Layer and IoT eHealth Cloud Layer. Note that IoT architecture for eHealth applications is already discussed in . However, the following provides a complete solution from data acquisition and processing, to cloud platform and big data analytics

A. IoT eHealth Device Layer

This rich set of smart IoT medical devices enables individuals to monitor their health data any time, from any computer or mobile device (all in real-time) and sync their data securely with the cloud eHealth platform. All they need to perform is to provide a connection using a suitable communication protocol to a gateway or a fog node. In this context, there is a vast variety of Personal Area Networks (PAN) and WSN protocols. Fig. 5 shows the IoT eHealth protocol stack. Note that selecting the best connectivity and the communication protocol highly depend on the application and the specific use-case. For example, to transfer a large amount of documents wirelessly, Wi-Fi is ideal. On the other hand, BLE best fits for short-range low power communications. The state-of-the-art IoT eHealth devices can be classified into two main groups:

• **Physical sensor:**

generally any medical device with a wired/wireless interface can be used in eHealth ecosystem to track patients' physical wellness, and digitally monitor their health . This includes ECG/EKG monitor , heart rate monitor , glucose monitor, blood pressure monitor , body temperature monitor, pulse oximeter , hemoglobin monitor,, activity monitor , smart shoes , smart garments or e-textiles , sleep monitor , knee sensor , skin conductance sensor

• **Virtual sensor:**

using software and mobile applications as well as eHealth services, virtual sensors capture patient's health data and contextual data from the environment . Virtual sensor includes many categories such as remote monitoring, remote consultation, diagnostic, patient health record, nutrition, and medical reference applications.

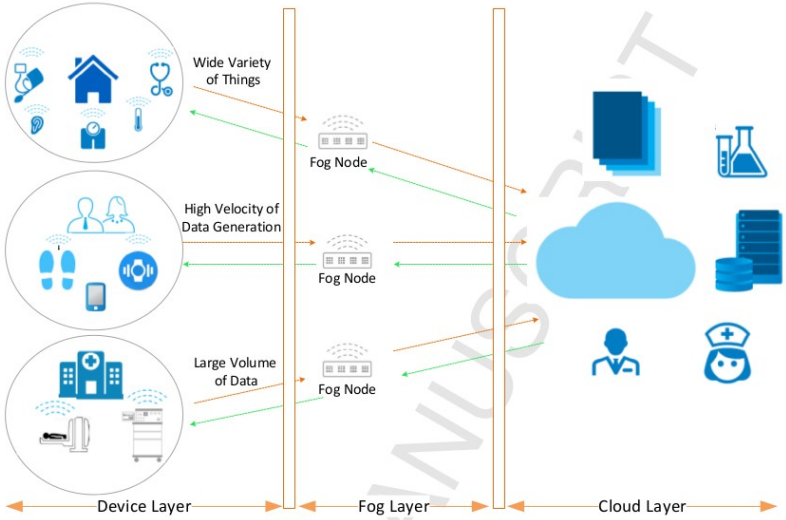

Figure 2.1 :IoT ehealth Architercute

B. IoT eHealth Fog Layer

eHealth is among those critical IoT applications which cannot afford latency. Indeed, it is a necessity to be able to analyze and act on the time-sensitive data and circumstances such as Myocardial Infarction (MI) in seconds. Therefore, it is not practical to rely on the traditional cloud model and architecture to collect and analyze the patients' sensitive medical data, vital signs, and bio-signals across a wide geographical area in the presence of various environmental conditions. The most efficient approach to tackle this issue is using fog or edge computing to take the cloud computing and services to the next level. A fog node is defined as a device with computing, storage, and network connectivity. In the proposed eHealth platform, we analyze the time-sensitive data, and make the extremely time-sensitive decision on the fog nodes. These nodes are placed closest to the medical devices that produce the data. On the other hand, we send the rest of the data to the cloud as the main storage and computing resource. It should be noted that another main issue with IoT eHealth is to conserve the network bandwidth. For example, an EMG device can generate several GBs of raw time-series data within a day. However, it is not feasible and not even necessary to send these huge amounts of data from thousands of patients to the cloud. As a result, fog nodes are reasonable to process, filter, and compress the data traveling between the medical devices and the cloud. As shown in the next figure, the main features of the fog node are as follow:

C. IoT eHealth Cloud Layer

The cloud benefits of a multi-layer architecture and it consists of the following layers

• Connectivity: the connectivity module is equipped with a vast variety of built-in capabilities to establish connectivity between eHealth devices, fog nodes and the cloud. This delivers an ultimate flexibility to select an appropriate communi- cation method that suits the requirements of the given

health application. As a result, the fog nodes and eHealth devices can be connected to the cloud using any hardware over any communication channel (wired networks, wireless networks, 3G/4G, or even satellite) based on a wide-range of protocols (MQTT, WebSocket, REST API, etc.).
• User, device and data management: the cloud integrates data from multiple sources. It captures the data from many fog nodes and stores the data safe and secure. So it is always there to be accessed by those engaged in the patients care. This platform is also integrated seamlessly with non-sensor sources such as data from EHRs, e-prescription platform, web sources, etc. Therefore, patient, physicians or anyone else in the patient's care team can access the data anytime and anywhere they need to. This significantly increases the collaboration across all disciplines increasing the efficiency of healthcare plan. Another advantage of the cloud platform is that it separates the data layer from the application layer while providing a unified schema in terms of capture and query transactions. This feature results in more flexibility to develop new applications. This module includes built-in capabilities for managing users, groups, devices and fog nodes, access permissions and roles.

2.4.IoT Ehealth Challenges

While IoT eHealth comes with the promises and visions of seamless connectivity across the physically distant locations where patients, clinics and hospitals could cooperate, coordinate and orchestrate the healthcare processes, there are several research challenges that IoT eHealth has to overcome before it could become a mainstream platform:
A. **Data Management**
Data management challenges for IoT eHealth is similar to those faced by IoT in other domains. However, the eHealth data come from medical sensors attached to humans. The human body is a dynamic system that changes its state continuously. Hence, as seen in IoT eHealth applications, there will be a constant flux of data coming from edge sensors via fog computing nodes. The cost of sensors and computing is declining and hence, it has become cheaper to collect the big data in a short time. In other worlds, IoT eHealth has to handle the complexity of the data in terms of their variety, volume and velocity . The challenge of data variety in IoT is quite newer than what existed 10 years ago. There are dozens of data formats depending on the healthcare end-user applications. For example, ECG data could be communicated in XML format, while detecting skin diseases using camera-based IoT device need to handle image formats. The data format support for edge computers is dependent on the manufacturers and their target customers. In addition to edge data format, the data model on the cloud also varies and therefore, demands standardization. The challenges of data volume and velocity are more associated with the capabilities of fog node hardware to receive, process, store and communicate the high-fidelity, high-resolution data coming from medical devices that could be with patients or in hospitals or clinics. Therefore, there will be a need of fog admins who could oversee the data flux between the fog and cloud computing.

B. **Scalability** To build a smaller scale of IoT, sensors on portable devices for data collection and secure central servers for processing users' requests are used to ensure all users can directly access medical services via portable devices such as smartphones. This facility can be scaled up to the entire hospital, so that patients in the hospital can use medical services, check updates and health

status monitoring by their smartphones. This eHealth model can be scaled up to the entire city, if there are sensors and antenna in the city to collect data, smart big data algorithms and APIs to process data and analyze users' requests, and intelligent interfaces to inform the status of users' requests in real-time. In an eHealth-aware smart city, all data can be collected, processed and analyzed by smartphones through mobile apps and feedback will be sent seamlessly to patients to allow them to know their health status and results of their medical checks. When patients use medical services in IoT, it can save their time to wait for appointments, wait for results and have direct access to certain level of medical resources. Benefits of scalability to a smart city level can include improvement in efficiency, saving quality time for waiting and building direct relationship and trust between medical staff and patients .

C. **Interoperability,** Standardization and Regulatory Affairs In general, IoT has raised concerns in the area of standardization. All manufacturers, service providers, and end users seek standards for operability both within and between the domains targeted by IoT applications. The standardization complexity lies in the fact that IoT aims to capture a wide range of disciplines that are, in general, regulated by different regulatory affairs. In the case of IoT eHealth, the complexity even increases due to the strict regulations mandated by medical standards. For example, in USA the standardization of wireless medical devices demands a multi-agency regulatory environment [128], involving three agencies: i) Food and Drug Administration (FDA), ii) Centers for Medicare and Medicaid Services (CMS), and iii) Federal Communications Commission (FCC). This implies that companies have to precisely consider the policies and rules mandated by all three agencies. Similarly, IoT eHealth will have to navigate through a complex multi-agency regulatory structure before we start to see the IoT eHealth products in the market. IoT eHealth will confront with similar trends in other parts of the world.

2.5.Conclusion

There is an increasing need from clinic-centric healthcare to patient-centric healthcare. IoT is expected to be a strong enabler by providing a seamless connection of devices and cloud storage as well as acting agents such as patient, hospital, analysis labs, and emergency services. A typical IoT eHealth system consists of four layers: 1) sensing layer, which integrates with all different types of hardware connect to the physical world and collect data, 2) networking layer, which offers networking support and data transfers in the wired and wireless networks, 3) service layer that creates and manages all types of services aiming to satisfy user requirements. 4) interface layer, which offers interaction methods to users and other applications. Due to the vast amount of different applications with different quality-of-service, storage and latency requirements, we observe *a tendency to split the networking layer into two sub-layers: fog and cloud layers. Fog layer handles the local buffering and different connectivity requirements to the device. Cloud layer handles the connectivity to fog, user/device/data management, and application services covering dashboard, rule engine, big data analytics, and integration framework within virtually any system, application or portal. We mostly observe that most of the work was done so far on the device sub-layer as part*

of sensing layer with examples in smart wearable devices, monitoring fitness devices, and medical-grade devices used in the hospitals but less work on fog and cloud sub-layers.

CHAPTER THREE : IMPLEMENTATION AND RESULTS

3.1.Introduction

For a long time, that is more than a decade , A real time Ehealth has been an agenda for many developed countries , there is a need of sharing the data with a group of specialists simultaneously ,Collaborative platform with parallel and synchronized data streaming are strongly needed ,In this chapter , the design and the implementation of a real time ehealth system is introduced by describing both the hardware and software levels and it's architecture and implementation

3.2 Real time Emg data classification

Electromyography (EMG) is an electrodiagnostic medicine technique for evaluating and recording the electrical activity produced by skeletal muscles and the central nervous system. EMG is performed using an instrument called an electromyograph to produce a record called an electromyogram. An electromyograph detects the electric potential generated by muscle cells when these cells are electrically or neurologically activated. The signals can be analyzed to detect medical abnormalities, activation level, or recruitment order, or to analyze the biomechanics of human or animal movement.

EMG testing has a variety of clinical and biomedical applications. EMG is used as a diagnostics tool for identifying neuromuscular diseases, or as a research tool for studying kinesiology, and disorders of motor control. EMG signals are sometimes used to guide botulinum toxin or phenol injections into muscles. EMG signals are also used as a control signal for prosthetic devices such as prosthetic hands, arms, and lower limbs.

Figure 3.1 : EmgSensors

3.3.EMG signal processing

Rectification is the translation of the raw EMG signal to a signal with a single polarity, usually positive. The purpose of rectifying the signal is to ensure the signal does not average to zero, due to the raw EMG signal having positive and negative components. Two types of rectification are used: full-wave and half-wave rectification.Full-wave rectification adds the EMG signal below the baseline to the signal above the baseline to make a conditioned signal that is all positive. If the baseline is zero, this is equivalent to taking the absolute value of the signal. This is the preferred method of rectification because it conserves all of the signal energy for analysis. Half-wave rectification discards the portion of the EMG signal that is below the baseline. In doing so, the average of the data is no longer zero therefore it can be used in statistical analyses.
EMG can be used to sense isometric muscular activity where no movement is produced. This enables definition of a class of subtle motionless gestures to control interfaces without being noticed and without disrupting the surrounding environment. These signals can be used to control a prosthesis or as a control signal for an electronic device such as a mobile phone or PDA[*citation needed*].

EMG signals have been targeted as control for flight systems. The Human Senses Group at the NASA Ames Research Center at Moffett Field, CA seeks to advance man-machine interfaces by directly connecting a person to a computer. In this project, an EMG signal is used to substitute for mechanical joysticks and keyboards. EMG has also been used in research towards a "wearable cockpit," which employs EMG-based gestures to manipulate switches and control sticks necessary for flight in conjunction with a goggle-based display.

Unvoiced speech recognition recognizes speech by observing the EMG activity of muscles associated with speech. It is targeted for use in noisy environments, and may be helpful for people without vocal cords and people with aphasia.

EMG has also been used as a control signal for computers and other devices. An interface device based on EMG could be used to control moving objects, such as mobile robots or an electric wheelchair. This may be helpful for individuals that cannot operate a joystick-controlled wheelchair. Surface EMG recordings may also be a suitable control signal for some interactive video games.
For signal processing and feature extracting there are a set of parameters that can ensure the optimal performance of the classifier
we mention next three of the most used feature extraction methods in signal processing

A. Mean Absolute Value (MAV) : Describes the energy of signals. The MAV of sEMG or ACC signal of non-dominant hand is used to distinguish one-handed or two-handed subwords

$$MAV = \frac{1}{N}\sum_{n=1}^{N}|x(n)|$$

B.Autoregressive (AR) Coefficient : AR model describes the current signal x(n) as the linear combination of previous k samples x(n − k) plus white noise w(n). AR coefficients (ak) have been proven to be effective in SLR.

$$x(n) = w(n) - \sum_{k=1}^{p} a_k x(n-k)$$

C.Linear Predication Coefficient (LPC): LPC model describes each sample of signals as a linear combination of previous *k* samples *x*(n − k)

$$x(n) = -\sum_{k=1}^{l} l_k x(n-k)$$

3.4.Hardware Part

To make this all happen , few instruments are needed,e-Health Sensor Shield . Arduino and/Or Raspberry Pi

3.4.1.Arduino

an open source computer hardware and software company, project, and user community that designs and manufactures single-board microcontrollers and microcontroller kits for building digital devices and interactive objects that can sense and control objects in the physical and digital world. The project's products are distributed as open-source hardware and software, which are licensed under the GNU Lesser General Public License (LGPL) or the GNU General Public License (GPL), [1] permitting the manufacture of Arduino boards and software distribution by anyone. Arduino boards are available commercially in preassembled form, or as do-it-yourself (DIY) kits.

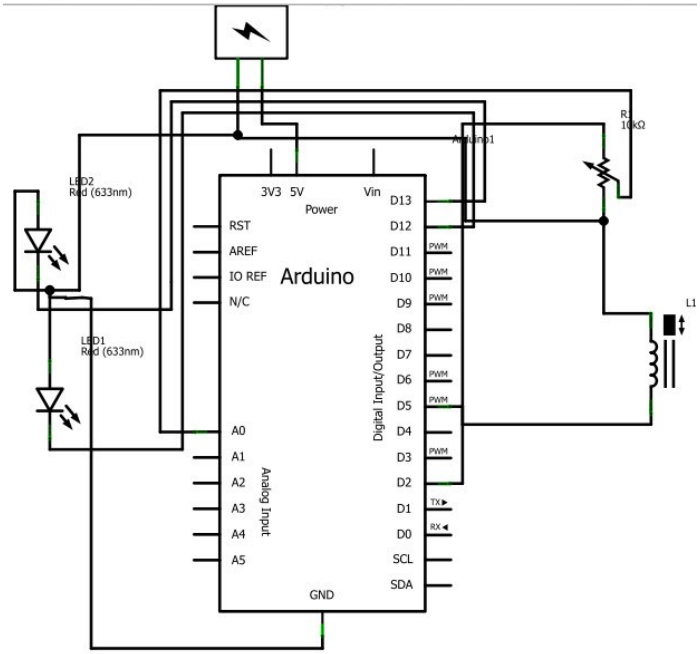

Figure 3.3: Arduino Uno with digital input/output

Arduino board designs use a variety of microprocessors and controllers. The boards are equipped with sets of digital and analog input/output (I/O) pins that may be interfaced to various expansion boards or Breadboards (*shields*) and other circuits. The boards feature serial communications interfaces, including Universal Serial Bus (USB) on some models, which are also used for loading programs from personal computers. The microcontrollers are typically programmed using a dialect of features from the programming languages C and C++. In addition to using traditional compiler toolchains, the Arduino project provides an integrated development environment (IDE) based on the Processing language project.

3.4.2.Ehealth kit

The e-Health Sensor Shield V2.0 allows Arduino and Raspberry Pi users to perform biometric and medical applications where body monitoring is needed by using 10 different sensors: pulse, oxygen in blood (SPO2), airflow (breathing), body temperature, electrocardiogram (ECG), glucometer, galvanic skin response (GSR - sweating), blood pressure (sphygmomanometer), patient position (accelerometer) and muscle/eletromyography sensor (EMG).

This information can be used to monitor in real time the state of a patient or to get sensitive data in order to be subsequently analysed for medical diagnosis. Biometric information gathered can be wirelessly sent using any of the 6 connectivity options available: Wi-Fi, 3G, GPRS, Bluetooth, 802.15.4 and ZigBee depending on the application.

This e-Health Sensor Platform Complete Kit allows to get a complete First Aid Kit for Makers. It includes:

Figure 3.4. : Ehealth Shield
Figure 3.5: Sensors Of the Shield

Before moving on to the software part, the last thing to do with the hardwares is preparing the instruments to receive and send data, as mentioned, Ehealth Kit can be used with Raspberry or Arduino, we've chosen to work with an arduino uno. Placing and connecting the two is enough to end the hardware part,

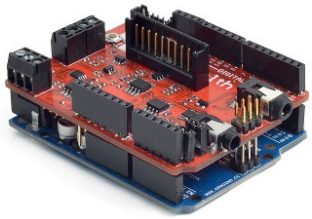

Figure 3.6 : Connecting the Board with the Shield

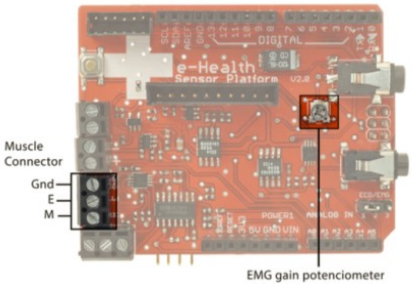

Figure 3.7 : Emg connectors

Figure 3.8: Connecting the sensor

3.5.Software part

the software part has been busy as well as the hardware part, such implementations provoked the need of new and recent techniques and libraries using different programming language to satisfy the need of each part of the project, Hardware programming Like Arduino and also big data techniques such as spark and Pyspark.Mllib , connecting all of this together couldn't be done without the good use of python's most powerful libraries such as Pandas and Sickit learn , we will go through details and explain the ideas step by step with the different techniques and algorithms used

3.5.1.Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Python features a dynamic type system and automatic memory management. It supports multi-programming paradigmes including oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

3.5.2.Pandas

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free softwarereleased under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals
DataFrame object for data manipulation with integrated indexing.

- Tools for reading and writing data between in-memory data structures and different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of data sets.
- Label-based slicing, fancy indexing, and subsetting of large data sets.
- Data structure column insertion and deletion.
- Group by engine allowing split-apply-combine operations on data sets.
- Data set merging and joining.
- Hierarchical axis indexing to work with high-dimensional data in a lower-dimensional data structure.
- Time series-functionality: Date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging.

3.5.3.Signal processing and data preparation

Root Mean Square (RMS) Value: Several factors can influence the capture of the EMG signal. These factors can be divided into physiological--type of muscle fiber, nerve fiber conduction, body temperature; anatomical (as a diameter of the muscle fiber), position (depth) of the muscle in relation to the electrode and thickness of the skin; and technical, related to instrumentation during EMG analysis, involving aspects related to the capture and processing of data

In statistics and its applications, the root mean square (abbreviated RMS or rms) is defined as the square root of the mean square (the arithmetic mean of the squares of a set of numbers). The RMS is also known as the quadratic mean and is a particular case of the generalized mean with exponent 2. RMS can also be defined for a continuously varying function in terms of an integral of the squares of the instantaneous values during a cycle.

$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}$$

The next Figure is an example of an EMG data with and without the application of the RMS function and we see the result with the RED plot .

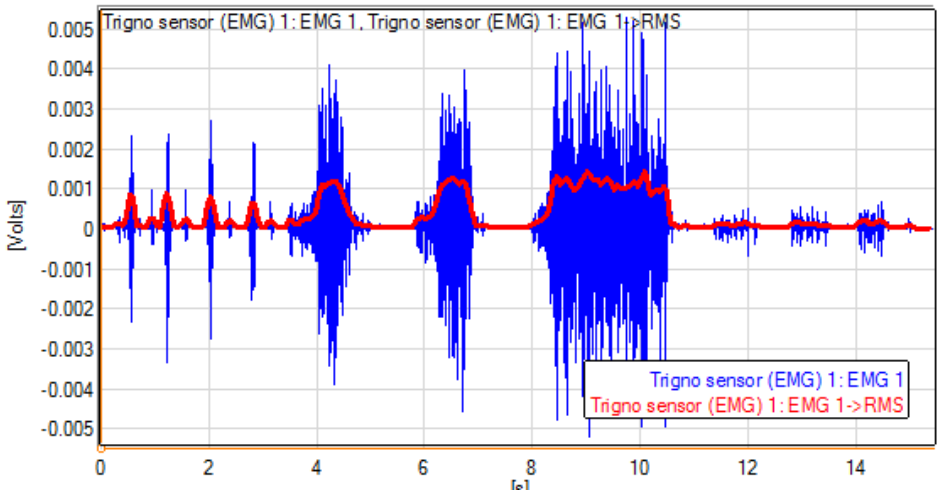

Figure 3.9: Emg data plotted in Red with RMS function

3.5.4.Arduino programming

A program for Arduino hardware may be written in any programming language with compilers that produce binary machine code for the target processor. Atmel provides a development environment for their 8-bit AVR and 32-bit ARM Cortex-M based microcontrollers: AVR Studio (older) and Atmel Studio (newer).

A. IDE

The Arduino integrated development environment (IDE) is a cross-platform application (for Windows, macOS, Linux) that is written in the programming language Java. It originated from the IDE for the languages *Processing* and *Wiring*. It includes a code editor with features such as text cutting and pasting, searching and replacing text, automatic indenting, brace matching, and syntax highlighting, and provides simple *one-click* mechanisms to compile and upload programs to an Arduino board. It also contains a message area, a text console, a toolbar with buttons for common functions and a hierarchy of operation menus. The source code for the IDE is released under the GNU General Public License, version 2.

The Arduino IDE supports the languages C and C++ using special rules of code structuring. The Arduino IDE supplies a software library from the Wiring project, which provides many common input and output procedures. User-written code only requires two basic functions, for starting the sketch and the main program loop, that are compiled and linked with a program stub *main()* into an executable cyclic executive program with the GNU toolchain, also included with the IDE distribution. The Arduino IDE employs the program *avrdude* to convert the executable code into a text file in hexadecimal encoding that is loaded into the Arduino board by a loader program in the board's firmware.

B.Sketch

A program written with the Arduino IDE is called a *sketch*.Sketches are saved on the development computer as text files with the file extension *.ino*. Arduino Software (IDE) pre-1.0 saved sketches with the extension *.pde*.

A minimal Arduino C/C++ program consist of only two functions:

•*setup()*: This function is called once when a sketch starts after power-up or reset. It is used to initialize variables, input and output pin modes, and other libraries needed in the sketch.
•*loop()*: After *setup()* has been called, function *loop()* is executed repeatedly in the main program. It controls the board until the board is powered off or is reset.

3.5.6.Spark

Most of the time, if it's a big data application. Spark will be there! Spark streaming offers a great deal everytime you think of a real time classification, I have deployed spark on the run execution for the final step in our implementation that concerns classification of EMG data
In chapter : big data, more details have been written about Spark

3.5.7.Data Collection

With the help of the A.M Hospital and the department of ELN/ELT , The data has been collected from different subjects over many sessions. Each session consisted of the subject maintaining the 6 chosen movement states for 10 seconds each, thus providing a large data points per class. We use many sessions to make sure we prevent overfitting a given action may be slightly different each

time it is performed. Data acquisition system included PowerLab, DualBioAmp, recording signal amplitude 2 mV, Data are outputted in csv format which are readable in Python for data processing.

EMG signals of biceps, deltoid, triceps, tibialis anterior, and quadriceps muscles are recorded in three states of isometric contraction (ISO), maximum voluntary contraction (MVC), and dynamic contractions . A preprocessing filtering process is then applied to recorded signals.

3.5.8.Emg Data Plotting and visualization (For distance monitoring)

Visualizing the data can be very helpful in most cases, But with medical data, it is helpful in every case, Python's Matplotlib provides a scientific plotting of an EMG raw data .

The next figure shows the plotting of 4 raw emg data :

E1 represents the first 30 data from EMG1

E2 represents the first 30 data from EMG2

E3 represents the first 30 data from EMG3

E4 represents the first 30 data from EMG4

X → Time in seconds

Y→ values in the interval of [0.1 : 0.35]

The data has been visualized after the signal processing , A numerical type of data , and before any Dimensionality reduction or anomaly detection

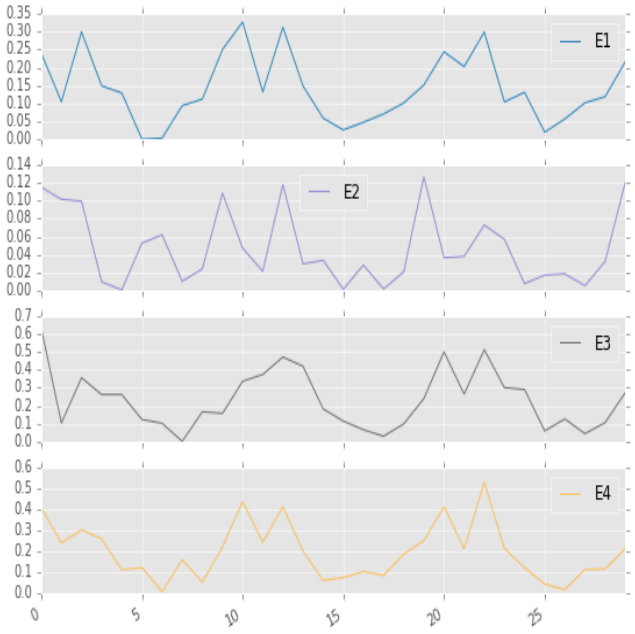

Figure 3,10: EMG graphs

3.5.9.Data preparation for streaming processing

Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams. Data can be ingested from many sources like Kafka, Flume, Kinesis, or TCP sockets, and can be processed using complex algorithms expressed with high-level functions like `map`, `reduce`, `join` and `window`. Finally, processed data can be pushed out to filesystems, databases, and live dashboards.

Figure 3.11 : Spark Streaming input/output

Internally, it works as follows. Spark Streaming receives live input data streams and divides the data into batches, which are then processed by the Spark engine to generate the final stream of results in batches.

Figure 3.12: Spark processing the divided batcheS

Spark Streaming provides a high-level abstraction called *discretized stream* or *DStream*, which represents a continuous stream of data. DStreams can be created either from input data streams from sources such as Kafka, Flume, and Kinesis, or by applying high-level operations on other DStreams. Internally, a DStream is represented as a sequence of RDDs.

3.5.10.Advanced concepts

A.Discretized Streams (Dstreams) is the basic abstraction provided by Spark Streaming. It represents a continuous stream of data, either the input data stream received from source, or the processed data stream generated by transforming the input stream. Internally, a DStream is represented by a continuous series of RDDs,

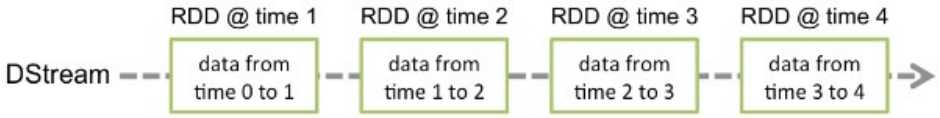

Figure 3.13: Spark continues RDD

B.Input DStreams and Receivers

Input DStreams are DStreams representing the stream of input data received from streaming sources. Every input DStream is associated with a **Receiver** object which receives the data from a source and stores it in Spark's memory for processing.

3.5.11 Applying Machine learning on our Data :

as we previously mentioned, in Chapter 2 the architecture of any IoT system mostly devided into 3 part , Data receiving and collecting , processing and analyzed , for our system we've shown the collection of data and the receiving with the EMG sensors using an Ehealth Kit, processing with Python and now we come to choosing the machine learning technique for real-time prediction/classification

A. Machine learning: is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.]

Machine learning tasks are typically classified into categories,

•Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. As special cases, the input signal can be only partially available
•Semi-supervised learning: the computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing.
•Reinforcement learning: training data (in form of rewards and punishments) is given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.
•Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

But in our case, the algorithms of machine learning should be adapted to streaming , and there it comes Spark, with it's reach libraries

3.5.12 Algorithms for the streaming

The machine learning algorithms provided by Mllib will be so helpful for the prediction step. First of all, there are streaming machine learning algorithms (e.g. Streaming Linear Regression, Streaming KMeans, etc.) which can simultaneously learn from the streaming data as well as apply the model on the streaming data. Beyond these, for a much larger class of machine learning algorithms, you can learn a learning model offline (i.e. using historical data) and then apply the model online on streaming data. And this last class of machine learning algorithms has proved it's usefulness for our case, we have chosen after many experiences to deploy Random forest classifier and Deep learning (Auto-encoders) for our classification and anomaly detection

Streaming applications impose unique constraints and challenges for machine learning models. These applications involve analyzing a continuous sequence of data occurring in real-time. In contrast to batch processing, the full dataset is not available. The system observes each data record in sequential order as they arrive and any processing or learning must be done in an online fashion.

Regarding the three characteristics of our data ,(Numerical, massive and time-series) and the fact that our machine learning should be a supervised learning , we have choose to perform the Data analysis phase using 4 different techniques :

A. Random Forests

Random forests are ensembles of decision trees. Random forests combine many decision trees in order to reduce the risk of overfitting. The `spark.ml` implementation supports random forests for binary and multiclass classification and for regression,

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it.

using both continuous and categorical features. Since that our Dataset consists of 6 classes and 4 attributes , The model has proven it's efficiency

we will see in the next figure a code for the initialization of the random forest algorithm on spark, and training the model on the given Training Data

```
ModelOne = RandomForest.trainClassifier(
    trainingData, numClasses=2, categoricalFeaturesInfo={},
                                    numTrees=3, featureSubsetStrategy="auto",
                                    impurity='gini', maxDepth=4, maxBins=32)
```

Figure 3.14 The Random Forest algorithm training on pySpark

In this next figure, we see the code to test, evaluate and compute the test errors of our model

```
predictions = ModelOne.predict(testData.map(lambda x: x.features))
labelsAndPredictions = testData.map(lambda lp: lp.label).zip(predictions)
testErr = labelsAndPredictions.filter(
    lambda lp: lp[0] != lp[1]).count() / float(testData.count())
print('Test Error = ' + str(testErr))
print(ModelOne.toDebugString())
```

Figure 3.15 The Random forest algorithm evaluation

B. GaussianNB

Naive Bayes classifier is a straightforward and powerful algorithm for the classification task. Even if we are working on a data set with millions of records with some attributes, it is suggested to try Naive Bayes approach. Naive Bayes classifier gives great results when we use it for textual data analysis. Such as Natural Language Processing. To understand the naive Bayes classifier we need to understand the Bayes theorem What is Bayes Theorem? Bayes theorem named after Rev. Thomas Bayes. It works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge. Below is the formula for calculating the conditional probability.

$$P(H\ |E) = \frac{P(E\ |H) * P(H)}{P(E)}$$

The next figure shows the Spark code for training a Naive Bayes model on our data

```
model = NaiveBayes.train(training, 1.0)

predictionAndLabel = test.map(lambda p: (model.predict(p.features), p.label))
accuracy = 1.0 * predictionAndLabel.filter(lambda pl: pl[0] == pl[1]).count() / test.count()
print('model accuracy {}'.format(accuracy))
```

Figure 3.16 The naive Bayes model on PySpark

C.SVMs

supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Computing the SVM classifier :

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(w \cdot x_i - b)\right)\right] + \lambda\|w\|^2$$

The next figures show the code for loading, building and training our SVM model

```
def parsePoint(line):
    values = [float(x) for x in line.split(' ')]
    return LabeledPoint(values[0], values[1:])
```

Figure 3.17 Labeled points on PySPark for SVM model

And in this next figure we show the code for fitting our model on our data . And running a test for our model

```
model = SVMWithSGD.train(parsedData, iterations=100)

# Evaluating the model on training data
labelsAndPreds = parsedData.map(lambda p: (p.label, model.predict(p.features)))
trainErr = labelsAndPreds.filter(lambda lp: lp[0] != lp[1]).count() / float(parsedData.count())
print("Training Error = " + str(trainErr))
```

Figure 3.18 SVM Model training on PySpark

D.K-nearest neighbor algorithm: In pattern recognition, *k*-NN is a non-parametric method used for classification and regression.[1]In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

3.5.13. Deep Learning for Real-time Anomaly Detection

A. anomaly detection: In data mining, anomaly detection (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset.

B. Autoencoder is an algorithm that uses neural networks to compress data into a simpler form, then decompresses it into a reconstructed version of the original data. In more mathematical terms, it's an approximation of the identity function: ? -> ?' .

This is a useful feature for anomaly detection because we're training the autoencoder on normal data. Thus, it has a very good approximation of the identity function for normal data. In other words, normal data as input leads to an output with a small reconstruction error.

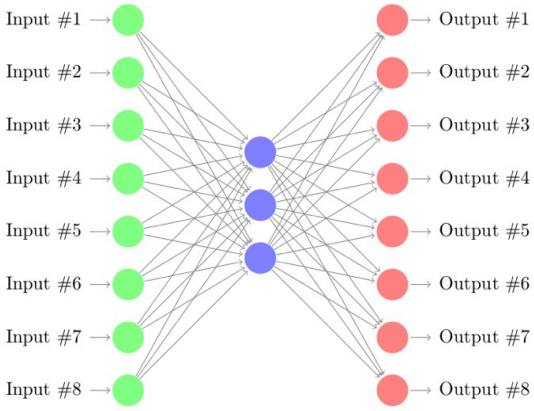

Figure 3.19 : Auto-encoder NN

When we use unusual input data, the output likely will be all wrong, as the model is only good at reconstructing normal data. This leads to a larger – hopefully detectable – output error.

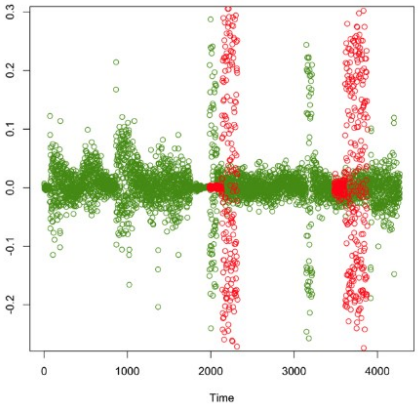

Figure 3.20 : Anomaly detected in Red

Figure 4.15 shows a graph representing an anomaly that is plotted int the red color

With what we have seen, we can simply say that our model can be deployed in the following flow

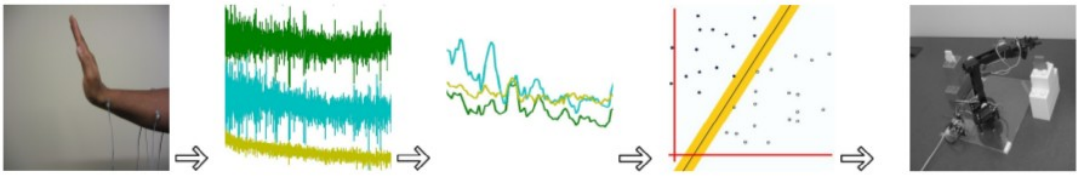

Figure 3.21: The Model Processes of Robotic control

3.5.14.Performance Tuning

Getting the best performance out of a Spark Streaming application requires a bit of tuning

1. Reducing the processing time of each batch of data by efficiently using the resources.

2. Setting the right batch size such that the batches of data can be processed as fast as they are received

3.5.15.Level of Parallelism in Data Receiving

Receiving data over the network (like Kafka, Flume, socket, etc.) requires the data to be deserialized and stored in Spark

Note that each input DStream creates a single receiver (running on a worker machine) that receives a single stream of data. Receiving multiple data streams can therefore be achieved by creating multiple input DStreams

One of the main problems faced during the implementation of different ehealth data is the synchronization and parallelizing of the same number of data for each sensor to be processed in the same interval , for an example :

we have Sensor1 sending 10 data in a known interval , while Sensor2 sends 5 data in the same interval, And for this, we propose to apply the RMS optimization function on a previously -determined batch of data, so the output of the processed data of both sensors will guarantee for us to have same amount of data in a known time interval

3.5.16.Discussing results

In our last experiments , we used our collected data to train four different techniques: GaussianNB, Random Forest , KNN, and a Multi-Class SVM classifier, First without the RMS function and without the anomaly detection , and then we applied both RMS and Autoencoders ,

We choose the precision(accuracy) metric for comparison

The Classification Models	Before applying the RMS and Auto-encoders	After Applying the RMS and Auto-encoders
GaussianNB	58,18%	85,20%
Random Forest	62,65%	**91,60%**
NuSvc(Multiclass SVM)	49,49%	70,55%
KNN	68,01%	90,90%

Table 3,1 : The results of the models

These results prove that after using the RMS and Auto-encoders on our collected data, the accuracy of the models had augmented , The Random Forest algorithm has shown to be the most effective, The RMS function represents the square root of the average power of the EMG signal for a given period of time and that helps for both optimizing and cleaning, We took the interval of 500 data to as an entry for the RMS equation, that guarantees two things, synchronization of streaming to receive the same amount of data as an output, and dimensionality reduction.

The anomaly detection using Auto-encoders could detect outliers before it can be processed by our classification model, Sensors are very sensitive , and the health care data are very critical ,so the system should make the anomaly detection as early as possible,

The random forest algorithm gave the best precision result and that's due to it's efficiency in the Over-fitting control and preventing with a controlled over-fitting improvement (Cross-Validation Technique) , we split our data into three groups -- a training set, a test set, and a validation set. Perhaps having noise data in any "signal form" input was a reason to reduce the over-fitting and carefully prepare our data to get the most augmented result there is .
Systematically applying the techniques on the IoT architecture did not show any shortcomings, But still, needs an enhance and improvement for better results, with the precision of 91%, The health section can't deal with any mistake

3.6.Conclusion

In this chapter we have explained how we can use both Health kit and Arduino board and Spark and big data techniques to deploy an efficient model of Emg data classification , both for health and robot controlling, starting from basics with small details until we reached the model's metrics, where we presented a comparison of different models and how efficiently the RMS and AutoEncoders augmented the precision metric for our system,

Conclusion

In the recent history, Internet of things is considered to be one of the major things to happen with the technology, hand in hand with big data analytics, millions of objects are connected, of different domains and fields, connecting the world making it more smarter

In this project, we tried and succeed to deliver a full EMG data classification in a real time streaming application, and in a second hand, making it more efficient throughout the explained steps during the different chapters

Our next work will be specifically related to the IoT industry and health technology , making our model more precised and having a scalable architecture , as we know , There is a big need from clinic-centric healthcare to patient-centric healthcare. IoT is expected to be the main change in human's health life

the ongoing process is extending our results to other types of movements, and to scale our model for online learning, rather than an offline supervised learning, and that can increase the possibility for more insight

References

[1] Jo, Minho. Taras, Maksymyuk. A survey of converging solutions for heterogeneous mobile networks." IEEE Wireless Communications 21, no. 6 2014

[2] Bär, Arian, Alessandro Finamore, Pedro Casas, Lukasz Golab, and Marco Mellia. "Large-scale network traffic monitoring with DBStream, a system for rolling Big data analysis." In Big Data (Big Data), 2014 IEEE International Conference on, pp. 165-170. IEEE, 2014.

[3] Ding, Xuejun, Yong Tian, and Yan Yu. "A real-time Big data gathering algorithm based on indoor wireless sensor networks for risk analysis of in dustrial operations." IEEE Transactions on Industrial Informatics 12, no. 3. 2016

[4] Simmonds, R. M., Paul Watson, Jonathan Halliday, and Paolo Missier. "A Platform for Analysing Stream and Historic Data with Efficient and Scalable Design Patterns." In 2014 IEEE World Congress on Services, pp. 174-181. IEEE, 2014

[5] Dimitrakopoulos, George, and Panagiotis Demestichas. "Intelligent transportation systems." IEEE Vehicular Technology Magazine 5, no. 1.2010

[6] Y. Sun, H. Song, A. J. Jara, & R. Bie . Internet of Things and Big Data Analytics for Smart and Connected Communities. IEEE Access, 4, 766-773.2016

[7] H. Song, S. Ravi, T. Sookoor, S. Jeschke, Smart Cities:Foundations, Principles and Applications, ISBN: 978-1119226390, Wiley, Hoboken, NJ, USA, 2017.

[8] Marx, V. Biology: The big challenges of Big data. Nature, 498(7453), 255-260.2013.

[9] Lv, Zhihan, Javier Chirivella, and Pablo Gagliardo. "Bigdata Oriented Multimedia Mobile Health Applications." Journal of medical systems 40, no. 5.2016.

[10] Pablo Basanta-Val,Anthony Steed , Zhihan Lv .” Next-Generation Big Data Analytics “, IEEE Transactions on Industrial Informatics . Volume: 13, Issue: 4, Aug. 2017

[11] Constandinos X. Mavromoustakis , George Mastorakis.- Advances in Mobile Cloud Computing and Big Data in the 5G Era , Springer; 1st ed. 2017 edition.2016.

[12] Koitzsch K. - Pro Hadoop Data Analytics. Designing and Building Big Data Systems using the Hadoop Ecosystem – 2017

[13] Andreas P. Plageras, , Christos Stergiou.” Efficient Large-scale Medical Data Analytics

in Internet of Things",2017 IEEE 19th Conference on Business Informatics (CBI) 24- 27 July 2017

[14] Kuan-Ching Li (Editor), Hai Jiang (Editor), Laurence T. Yang (Editor), Alfredo Cuzzocrea (Editor) "CRC.Big.Data.Algorithms.Analytics.and.Applications".Chapman and Hall/CRC; 1 edition . 2015

[15] Omar Said, Amr Tolba ."SEAIoT: Scalable E-Health Architecture based on Internet of Things" International Journal of Computer Applications (0975 – 8887) Volume 59– No.13,2012

[16] Gheorghe ,Sebestyen. Anca, Hangan. "eHealth Solutions in the Context of Internet of Things", 2014 IEEE International Conference on Automation, Quality and Testing, Robotics.2014

[17] Temitope O. Takpor and Aderemi A. Atayero, Members, IAENG. "Integrating Internet of Things and Ehealth Solutions for Students' Healthcare".Proceedings of the World Congress on Engineering 2015 Vol I WCE 2015, July 1 - 3, 2015,

[18] George Suciu & Victor Suciu . "Big data, Internet of Things and Cloud Convergence – An Architecture for Secure E-Health Applications" Journal of Medical Systems 39:141

[19] Yeliz Yengi, Kerem Küçük ."Context aware internet of things for large scale data analytics ",2017 International Conference on Computer Science and Engineering (UBMK),5-8 Oct. 2017

[20] Farahani, Bahar. Firouzi, Farshad. Chang, Victor "Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare" .Future Generation Computer Systems .2017.

[21] https://spark.apache.org/docs/latest/

[22] https://medium.com/swlh/apache-spark-streaming-simplified-3107f1580b30

[23] https://spark.apache.org/docs/latest/ml-classification- regression.html#classification

[24] https://spark.apache.org/docs/latest/api/python/index.html

[25] https://spark.apache.org/docs/2.2.0/streaming-programming-guide.html

[26] https://www.quora.com/What-is-the-impact-of-big-data-on-eHealth

[27] https://en.wikipedia.org/wiki/Machine_learning

[28] https://en.wikipedia.org/wiki/Support_vector_machine

[29] http://dataaspirant.com/2017/02/20/gaussian-naive-bayes-classifier-implementation-python/

[30] https://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/

[31] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[32] Ruiliang Su, Xiang Chen, Shuai Cao, and Xu Zhang .”Random Forest-Based Recognition of Isolated Sign Language Subwords Using Data from Accelerometers and Surface Electromyographic Sensors“.Sensors (Basel). 2016 Jan; 16(1): 100.

[33] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd