**Logo

Description automatically generated**

**COMPUTER SCIENCE AND DATA ANALYTICS**

Course: Intro to Big Data Analytics

# TERM PAPER

Project title: “The role of Big Data and Machine Learning in Healthcare”

Student(s): **G32288185 Ilyas Karimov**

Instructor: **Dr. Abzatdin Adamov**

## Baku 2021

The role of Big Data and Machine Learning in Healthcare

Ilyas Karimov   
 School of IT and Engineering

ADA University

Baku, Azerbaijan

ilyaskarimov.f@gmail.com

# Dr. Abzetdin Z. Adamov Center for Data Analytics Research ADA University Baku, Azerbaijan aadamov@ada.edu.az

*Abstract*— Healthcare is a critical component of reality, which occasionally collects huge amounts of data. It is even arguable that its significance does not lag behind that of other sectors. It's fascinating to learn how to collect and use data using Big Data tools, and in this paper, the steps have been thoroughly researched. Precisely speaking, the purpose of this paper is to identify systems that support Vector Machine Support, Logistic and Linear Regression, Neural Network, Random Forest, Discriminant Analysis, and other Machine Learning and Deep Learning algorithms, as well as systems that will aid the Medicine industry, as well as to study and evaluate the role of Machine Learning in the healthcare industry. The research concludes that structures do exist, but they have been applied to specific cases and records, and their precision and accuracy have been extremely high. Regrettably, no centralized structure has been created for all hospitals due to the novelty of the field.

Keywords—Healthcare, Big Data, Machine Learning, Diagnosis, Classification, Detection

1. INTRODUCTION

Healthcare is a vital substance in the world of reality, where enormous amounts of data are occasionally collected. In a tempestuous major breakthrough, the health care industry is beginning to experience a lot. These are information that expect huge data characteristics and are appealing to analyze it and to create innate relationships between factors inside the health information system. Health information is rich in valuable information [1]. Health reform and fiscal restraint are the forefront and the focus of politics, the mainstream press and the world's business executives. With the global healthcare industry struggling with an unadjusted basis in costs, an expanding group of consumer patients and a growing chronic disease management demand for medical services is also on the increasing trend. The global costs of health, currently projected to reach over $12 trillion for a mere seven years, amounting to approximately $6 trillion to $7 trillion [2]. According to Organization for Economic Co-operation and Development (OECD) statistics [2], in North and South America, Europe and Asia/Pacific, medical spending equates to approximately of 9.50 percent of gross domestic product (GDP). The organization's annual rate of growth in healthcare spending is 4.9% (exceeding the U.S. rate). Health spending is growing at a rate of 16 percent in China, where most of the country is covered by insurance. Regardless the payer system, the worldwide phenomenon is that there is either a decrease in the pool of available capital to pay for healthcare or the numbers almost remain the same. However, this trend is not observed when it comes to costs, which is rising time-by-time. Despite these economic metrics, progress and digital change will have powerful, cost-consistent ways to improve the standard of treatment – but only if they are carried out correctly and it depends on the technology, the existing algorithms and their usage in healthcare [2].

1. BIG DATA IN GENERAL

Big data is central to modern engineering and science and its analysis. The data are collected and come from internet purchases, emails, photographs, audios, click streams, files, tweets, search questions, medical registries, social networking communications, scientific information, sensors and cell phones [5]. In large segments of the market, big data and analytics have caused revolutions. Big data is a modern, omnipresent concept in recent years. Big Data refers to extensive, complex databases which go beyond conventional data management systems for timely and cost-effective storage, management and processing. Big data techniques in petabytes and more can accommodate organized, semi-structured and unstructured data [6]. Structured data are classified as quantitative, and the data stored in Relational Database Management System

(RDBMS) are structured and structured records are examples of years, addresses, credit card details, sales orders, location data and etc. [7]. In semi-structured data, the data affiliated with a scheme are enclosed in the data sometimes referred to as "self-description." There’s no specific arrangement in some forms of semi-structured data, in others it only exists but places loose restrictions on data. Recently, semi-structured data became an important topic for a number of reasons such as the web, which we would like to treat as the database, but it doesn’t contain any scheme, and a high scalable format for the sharing of data between disparate databases [3]. The unstructured data are qualitative data, and they cannot be analyzed and interpreted using traditional techniques and processes and the example for the unstructured data are email, video, audio, telephone usage, social media and etc.

The size of the data collected so far is quite large, and the size of the data is increasing day by day. The Human Face of Big Data is a global initiative in 2012, which focuses on the collection, visualization and analysis of huge quantities of real - time information. Many figures are extracted from this media initiative. Facebook contains 955 million monthly active users with 70 languages, 140 billion uploaded images, 125 billion friendly contacts, 30 trillion of content per day and 2,7 billion likes and views. 48 hours of content is posted every minute and there are 4 billion views on YouTube every day, and Google provides resources for multiple programs, both monitoring 7,2 billion pages a day and processing 20 petabytes of daily data in 66 languages [4].

Health care is one of the lines of business which can benefit greatly from the rising data volume and availability [6]. According to [8], there are at least three fields rich of data that are in healthcare in order to help almost instantly with data science technologies: Hospital Claims Data, Patient’s Clinical Data and Trial Data for Research & Development.

1. Hospital Claims Data

Huge number and so many billion patients are treated every year in hospitals across the country. In-patient admissions, illnesses, casualties and efficient means of mitigating the visitation, population patterns and disease prevention dependent on predictive models, data analysis tools can help explain trends. Healthcare data comprehension will also help healthcare providers make informed decisions about how to improve the high-quality care.

1. Patient’s Clinical Data

Clinical data are laboratory outcomes, patient data, medical images, strongly linked and categorizable doctor notes (understanding the patients' health and wellbeing and indications of the disease) for better support for treatment practice. Gathering data from multiple sources makes it possible to see the broad picture for healthcare experts. As populations grow, clinical data are expected to just increase significantly and thus, as most of the time, almost implausible, to make decisions based on manual analysis.

1. Trial Data for Research & Development

There's an increasing number of trials around the entire planet to comprehend illnesses and methods for dealing with them. Substantial volumes of rich databases would be best used to implement analytical techniques in order to explore additional insight and connections in the data.

Big Data potentials were identified in five key themes: healthcare, public sector, retail, manufacturing, and personal data, according to  Mc Kinsey Global Institute [9]. Healthcare is used in support mechanisms for therapeutic decision-making, individual tests for patient profile, customized medicine, staff success prices, epidemic trends analyses, public health improvement

1. MACHINE LEARNING IN GENERAL

How will machines themselves be made to change with training? and what do we know about all learning processes, from basic computation and knowledge theory? Applied Machine Learning (ML) research is critical for tackling both basic mathematical and practical problems. Two decades ago, ML was a laboratory novelty, now it is a useful tool in industry [10]. Artificial intelligence (AI) arose as the preferred approach for creating functional machine vision tools, because of its ability to create methods for speech recognition, natural language processing, and robotics. Automated learning is concerned with the issue of how to create machines that can learn. Data science is currently one of the most rapidly increasing technological fields, connected to computer science and artificial intelligence, situated at the intersection of those three disciplines. Recent machine learning has been spurred by both the advancement of new machine learning methods and recent advances in affordable computational power. For evidence-based decision-making in many fields, including healthcare, manufacturing, education, and the financial sector, we're seeing the use of more data-intensive methods of ML [11]. Developers have learned that, with the use of an artificial intelligence device, it is easier to the use of machine learning has also affected various areas of computing, including software engineering, customer care, and the troubleshooting of complex systems. The ML techniques which have been developed to interpret high throughput data are affecting many different areas of scientific research, from biology to sociology

It can be difficult to provide a single explanation for the many interdisciplinary nature of ML; it can be attributed to statistics, data analysis andanalytics, and mathematics knowledge produced by ML is a specific type of AI because it comes from data training. In this phase, we're not instructing the systems where to look, but the system is sure to find several candidates busted into the following groups, according to the image in *Fig. 1*.

1. Supervised Learning

During training, this learning devotes time to matching up inputs to outputs. It is typical for supervised learning tasks to use binary classifications, in which an input is either similar to or different to a desired output [12]. In contrast to unsupervised approaches, supervised approaches are approaches to model building. These approaches use a pre-processed sample training set to create a model in which each sample has a class label in the training dataset. The goal of supervised learning approaches is to define characteristic sets of rules that can be used to distinguish samples from various classes. For example, decision trees (DT) are an easy to understand and relatively easy to implement supervised learning approach but applying to complex nonlinear problems is challenging [13].

1. Semi-supervised Learning

Using the labeled information as well as the unlabeled information to construct the classifier is the most accurate. The actual and proper use of unlabeled information enables the system to achieve a very high classification rate of performance. To say that this program will be successful, one must assume a few things have to be true [14]. Clustering is the most frequently encountered unsupervised learning task: it involves identifying potentially valuable clusters of input examples. For instance, a taxi driver can gradually develop an understanding of "good traffic days" and "poor traffic days" without ever receiving labeled examples of each from a teacher [15].

1. Unsupervised Learning

The labels never receive any training. Unsupervised learning follows these three algorithms: K-means clustering, fuzzy clustering, and hierarchical clustering as defined by the K-means algorithm.  A cluster of trained data with unknown labels is created here. These algorithms are used with sample data for a development framework [16] .

Diagram

Description automatically generated

Fig. 1. Classification techniques for machine learning

1. Reinforcement Learning

The software system gives access to the changing environment to achieve a particular goal through this learning. The program receives feedback on rewards and penalties, as it navigates the disadvantage [17]. Reinforcement learning is the only viable method for training a program to work at a professional rate in a large variety of complex environments [15]. For instance, it is extremely difficult for a person to provide accurate and reliable assessments of a large number of positions, as would be required to train an evaluation function directly from examples, while playing a game. Other than that, the program can be informed when it has won or lost and use this knowledge to train an evaluation function that provides fairly reliable estimates of the likelihood of winning from any given position. Similarly, it is incredibly difficult to program an agent to fly a helicopter; but, if an agent is given sufficient negative reinforcement for crashing, wobbling, or deviating from a predetermined path, the agent will learn to fly independently.

One could argue that reinforcement learning encompasses all of AI: an agent is put in an environment and must learn to act effectively within it.

This research is aimed, first of all, at understanding and secondly, at showing people the role of technology in human health. Even when those are the principal goals, of course, other goals always exist, such as working with research which requires large amounts of data, collecting data and working on medical research. This research makes it possible not only to identify them, but also to discover and recognize very rare algorithms which, in spite of their rare development, are very successful in the field.

1. BIG DATA IN HEALTHCARE

This section will discuss how to collect and use large amounts of data, especially in the healthcare industry. The Big Data field is known to have a large number of data, so it is divided into collection and analysis steps, and this section examines the main functions of big data analysis in the development of health technologies and study [18]. This is used in order to capture, store, analysis and inspect vast quantities of organized, unstructured and semi-structured data generated by existing systems or organisations, likewise in healthcare, supported by computational methods and techniques. From data mining to information discovery, the different components of the major health data analysis are outlined as follows [18]:

1. Step 1: Collect, accumulate and store data

Many diverse data sets are acquired from a variety of sources during the first step (i.e. internal and external, respectively). Data may be recorded in different ways, depending on the application. Arbitrage is either conducted in the markets or placed in data warehouses. Finally, medical data processing is a huge problem in this case because of lack of data formats, requirements, scalability, and the possibility of breach of privacy In addition, it is difficult to identify where the required metadata are located. In an exploratory scheme, metadata collection of all patterns but potentially simplifies content discovery and identification of those patterns. Preprocessing and cleaning are critical issues in audio post-production. Information cannot be stored without use. Thus, any information that is stored as a result of the whole process being unproductive may be amplified.

1. Step 2: Clean up, extract and classify data

Step two in the medical analysis is to discover medical knowledge, and extract and combine it in a single database. Faulty health records are removed using data sanitization. Device and data collection, doctor's prescriptions, medical images, and social networking don't suffer from information inaccuracies. a major struggle during the life-cycle maintenance is removing and/adding values On the other hand, medical images are integral to the process of obtaining the data (e.g. MRI, CT, and ultrasound) and is time-consuming and challenging due to structure. Semi-structured, structured, and unstructured data should be compiled and processed for significant investigation

1. Step 3: Integrate, aggregate and visualize data

Source information can be provided from internal and external health databases for data aggregation processes. Medical Big Data comes from a variety of references and sometimes large data repository that has to be placed on a common platform for unified analysis. The aggregation challenge concerns a high volume and a wide range of data, from different data stores and real-time data. The solution can be to use high-speed technologies for file transfer [19]. To analyze the gathered medical data and use the results for aggregating results, the myriad of relevant findings may be useful. It's geared to specifics, including what any documents had been entered in relation to the patient, diagnosis and the patient's name, medical background, when records were last changed, and what date they were created, in order to assess a patient's level of vulnerability. According to these findings, the aggregated data is shared with hospitals, analysts, data scientists, and central, state, and even with regions and also the public health departments. The challenge is to connect complex and disparate medical knowledge with raw data in the real-world situations, which are challenging to coordinate. To get the most control possible, health care providers want the latest and most up-to-date medical history for the patient. In addition, in the final stages, critical details must not be changed to represent the frequency of accurate perceptions should be meaningful at the same magnitude as in relation to the noise, and after that, the prevalence of inaccurate ones must be limited in size. Selecting a more appropriate representation for the data could therefore increase the number of meaningful facts to select from.

1. Step 4: Model, analyze data and process queries

Convolatile or intractable data models are made use of to translate complicated medical data into text, diagrams, icons, and graphs. the main function of these analyzers is to detect abnormalities in health-related data and make sure everything is working correctly. Data models can be based on the following three concepts: physical, organizational, and conceptual. These models can be combined in various ways. Also, since it preserves semantics as well as other elements, we have the tendency to mark it as "quality". One thing a data miner might look at is any collection of healthcare data and the doctor may do is use different analytical methods and technology, such as data mining, to help the researcher and healthcare provider discover knowledge from a healthcare data sets. At the moment, the most common problems facing big data research is the lack of a proper database fit between various database systems. The next step is to combine the medical information with the value information to make an overall assessment. a. Broadening the queries to include users' circumstances, behavior, etc. to make them better understand their meanings is a method of aiding the end user with better query comprehension. the range of questions is all the way from the doctors and families to the patients There are a variety of different levels of tools and analysis that developers can use depending on the difficulty of the query that they face.

1. Step 5: Integrate data, deliver and get feedback

The last stage of data analysis is threefold: interpretation, presentation, presentation, and feedback. When all the preceding operations have been completed, the data must be evaluated, then it must be interpreted. In theory, the results of the results of the health study should be easy to grasp. If the data analyst/decision maker has not done their work, then the patient will not be able to interpret the findings for the doctor or other medical professionals. In addition to checking the assumptions, it is also critical to recheck and re- ment the data and information that was originally examined. A decision maker needs to carefully scrutinize the numerous assumptions throughout the process. This model is intended to give them the information they already have access to earlier on previous issues, so that they can avoid further complications. Feedback at the end of the project will be sought from patients and caregivers to better serve the needs of the patient population will be sought in all phases.

1. MACHINE LEARNING IN HEALTHCARE

In several areas, ML draws from the principles of informatics, statistics and optimisation. At its heart, almost all ML problems can be formulated in terms of optimizing a dataset. In such a context, the goal is to find a model that best explores the data in *Fig. 2.*

Most applications come under 3 groups, though there are several different ML types: supervised, unsupervised or reinforcement learning. One begins from the model as the system input with a conventional approach to data analysis. With an ML approach, a model may begin with the data and be applied to new data afterwards [20]

It is thought to help us to achieve an elusive pursuit of enhancing productivity and quality of health care with integrating Machine Learning into healthcare. There are, however, obstacles and opportunities. Machine learning offers a range of approaches and techniques, as described in the previous sections, that have cascaded into various instruments to help the diagnostic and forecasts challenges facing medical practitioners.

Text, application

Description automatically generated

Fig. 2. What is Machine Learning?

In this portion, we concentrate on the effects of machine learning to identify and analyze clinical parameters for the diagnosis and progression of the disease. ML is of major concern for the extraction of features that could contribute to the preparation and support of patient-specific treatment and ultimately reduce medical costs [21]. ML is often used to recommend patient clinical surveillance in real time. This includes real-time data processing, which adequately addresses data tracking on various sensors or devices and interpretation of continuous data for intensive care units (ICU) use [22]. Similarly, substantial efforts have been made in the last two decades to track and classify physical activity patterns from sensor-related data [23]. A variety of critical health applications inspired this initiative. For example, there is an increasing interest in the correlation between the physical activity levels and common health issues, such as diseas, cardiovascular illnesses and osteoporosis, as the trend towards more sedentary lifestyles. Since self-reported measurements have proven inaccurate in activity profiling, sensor data measurement in large-scale epidemiological studies in this field has begun to play an important role. Diagnosis assisted by computer (CAD) and its related methods is instrumental in the realization of machine learning potential. In the field of cancer research there are a wide variety of CAD tools [24]. The extensive resource that can be used for the creation of these resources is responsible for this. However, successful data and information integration from various data sources is required. The validation schemes for these instruments are also not successful. Radiology is the fastest field in which to implement CAD tools. Due to the lack of detailed datasets for information on a variety of ills, complications and injuries in many of these instruments during their inspection phases of production. Emergency medicine is another field which can be used in machine learning [24]. Although there are very few CAD instruments in clinical practice, existing technologies have demonstrated the ability to improve healthcare quality. Continued research in this field focuses on the creation of these instruments that address a broader range of trauma and disease scenarios. Machine learning has been unsuccessful in applying cardiovascular CAD tools because rigorous validation processes have not taken place [25]. Although many cardiovascular CAD instruments have high false positive rates, they are also useful for early detection of the disease. To minimize false positive rates, resources are therefore needed to integrate a wider range of details. As in the above-mentioned areas, orthodontics applies digital radiology [26]. They allow a dental complication to be diagnosed at an early stage. The CAD tools in this field are however relatively costly and a large adaptation bottle neck.

In the process of personal medical and clinical practice, the patient is more important in controlling his medical records in order to minimize medical costs. Cheaper technology for the detection, monitoring and understanding of diseases is growing [13]. After discussing the role of Machine Learning in healthcare, we may begin discussing algorithms, which is precisely what this chapter will do*. Fig. 4* brings together the algorithms used in Healthcare and shows how much or how little is used with the pie chart. One of the first algorithms is called Support Vector Machine (SVM).

The Vector Machine Support (SVM) is a data classification machine-learning method [15]. Typically, mathematical simulations are done to create an SVM model to identify variables that better predict medication adherence. The support vector machine (SVM) was created by Cortes and Vapnik [27] in the 1990s by the cooperation between the scientific community of statistics and the learning machine.  SVM attempts to categorize cases by finding the distinct hyperplane boundary.

SVM attempts to categorize cases by finding the distinct hyperplane boundary. The key benefit of the SVM is that the 'high dimension problems' can be relatively easily solved, i.e. the problem which occurs when there are numerous input variables in relation to the number of observations available [28]. The approach is an algorithmic implementation of statistical learning theory, using certain model features performance on a training package to create accurate data set estimators [27]. Supervised machine learning algorithms are among the most well-known classification techniques. A method known as Support Vector Machines (SVM) determines a model for the given training dataset, locating the best hyperplane which classifies the data while minimizing the distance between the two clusters [29]. SVMs for classification of patterns and are widely researched and implemented to a broad range of problems of classification and approximation of functions. The classification of patterns defines classification of certain objects into one of the recognized classes called [30]. There are generally two methods for classifier development: a) a parametrical approach based on data distribution prior data; and b) a non-parametrical approach in which data are not necessary to match a normal distribution [31]. A classifier's decision feature is established through a training process with a pre-determined data set of input-output. Then a test dataset can be classified into one of those classes [30]. SVMs use a separating hyperplane given marked data as discriminative classifiers. The hyperplane acts as a separator of both groups and SVMs are in nature binary classifiers [32].

SVMs use a separating hyperplane given marked data as discriminative classifiers. The hyperplane acts as a separator for both groups and SVMs are inherently binary classifiers [33].

And, according to the concept of the minimal risk of experience, SVM differs from the neural network method; the SVM seeks the best balance between model's complexity and learning capacity according to limited sample data, is ideal for small sample learning, and overcomes insufficient problem of typical negative types of data [34].

The input layers of neurons (or nodes, units), one or two (or even three) hidden neuron layers and the final layer of neuron outputs are formed into an artificial neural network (or simply neural network). *Fig. 4* illustrates a typical architecture, in which neuron lines are also shown. A numeric number called weight is connected to each link [35].

The neural network technology is an important tool to modify and adapt the data to classification, regardless of the primary model's specific functional or distributional requirements [36].  A multi-layer neural network (MLPNN) is also utilized to forecast heart disease using a back-dissemination algorithm (BP); and it was reported that neural networks can provide important diagnostic potential in healthcare [37].

Chart, sunburst chart

Description automatically generated

Fig. 3. ML in Health Industry

Diagram

Description automatically generated Fig. 4. Neural Network

Back propagation with an optimization technique, such as a gradient descent, can be used for the neural network training process. This approach would calculate the gradient of a loss function for all network weights [38]. However, multilayer perceptron (MLP) is a forward neural network feed and maps the inputs to a fitted output range. MLP consists of several layers of nodes within a chart, so each layer is fully connected with the next one, except the input nodes, by a nonlinear activation mechanism. For educational purposes and a non linear activation function [39], MLP uses a controlled learning technique called back propagation. In a further analysis, the authors claimed 100% accuracy with an increase in the number from 13 to 15 [40] by using neural networks on a data set made up of 573 records divided into the two sections of training and testing. In another paper [42], the support of a vector machine and neural backbone network have been used to identify 303 records of the online cardiovascular dataset. The seriousness of heart disease is diagnosed in five groups. To enforce this classification, the data set has been used. The findings showed a higher expected accuracy compared to neural networks in the SVM.

SVM is also used to detect electrocardiograms (ECG). The classification of arrhythmic beat is primarily used in ECG to detect heart problems. Preprocessing ECG signal and vector support arhythmic beat classes based on vector systems are done to classify into normal and abnormal topics. In view of the ECG signal analysis, a wide range of variations in cardiovascular beat-to-beat timing, called cardiac rate variability, have been provided over the years (HRV) [42].

An additional analysis compared SVM to a neural network, and the results were again in favor of SVM in the same paper. The topic was identical. In this article, you look at five types of ECG beats: (1) Normal beat (2) Right Bundle Branch Block, (3) Left Bundle Branch Block, (4) Premature Ventricular Contraction (5) Paced beat.  We compared the performance of the two separate classifiers in the Discrete Wavelet Transform (DWT) extraction technique. The technique of DWT uses the information coefficient (D1, D2, D3, D4, D5) and the approximation coefficient A5. The Vector Machine Support and the Artificial Neural Network classificators are provided separately with these feature sets. The results of the experiments showed that we were 98.80% more sensitive with the SVM [43].

Logistic Regression is a non-linear method used to represent binary dependent variables. There are only two classification variables, a success value or a failure-related value [13]. In recent years, the use in medical research of the logistic regression models has greatly increased [44] and Machine learning systems in different areas of science and the biomedical and healthcare sectors are increasingly being used [45]. For simple binary outcomes, the traditional logistic regression model applies. Models of illness (ill or healthy) and decision-making (true or false) are especially suitable. The most common approach for assessing logistic regression parameters is to use the highest probability estimator (MLE). We are aware that the existence of outliers in the data will seriously impact the estimation of MLE. In regression research, diagnostics play a significant role. Int is also found to have a great bearing on the covariate trend and thus its existence is misleading. In contrast to ordinary regression, logistic regressions in which all Ys are 0 or 0 must be replenished [44]. Professionals in the health sector with heart disease are limited by themselves and cannot predict the risk of high heart disease accuracy. One of the aims of Heart Disease improvement are to forecast accuracy through the Machine Learning Logistic Regression model taking into account the data set for health care which classifies patients, whether or not they are cardiac in their records.

In medical research, heart diseases are an important challenge. Machine learning can be a good option for predicting any human heart disease. The Neural Network, Decision Tree, KNN, etc. can be used to predict heart diseases [45]. Logistical regression is a regression analysis used in statistics to predict the outcomes of a dependent category variable of a predictor or independent variable set. The dependent variable is always binary in logistic regression. The logistic regression is primarily used to forecast and calculate success [46]. The analysis of logistic regression and the Artificial Neural Networks (ANNs) model to identify risk factors and forecast chronic conditions through case study on hypertension is presented in one of the research projects [47]. For a predefined p-value, the binary logistic regression system was applied to the experimental dataset gathered from Behavior Risk Factor Surveillance System (BRFSS). In addition, the Multi-Layer perception model (MLP) was developed and trained to predict hypertension with selected risk factors as inputs for ANNs with The Back Propagation algorithm (BP). Experimental findings demonstrated that their strategy was more than 72% accurate, in Table I, when hypertension was diagnosed and the Area Under the receiver-operator Curve (AUC) was more than 0.77. The findings show that logistical regression and artificial neural networks are a tool that offers an efficient way to pick risk factors and to predict hypertension, and also a general approach to the prediction of other chronic illnesses.

TABLE I. PREDICTION RESULTS FROM LOGISTIC REGRESSION

Table

Description automatically generated

The paper [48] focuses on the perspective of nursing. The purpose of this study is to provide a way for the voice of nursing emotion to be discriminated against. They first carried out the pronunciation experiment and derived effective speech parameters by means of the sound analysis. They also made several comparisons and discriminant analysis for clarifying the variations of these criteria between feelings and emotions.

In order to construct the speaking sample, they randomized 50 sounds to produce 8 word syllables, then we produced meaningless words and picked "ko-ne-yo-chi-yu-e-ho-te," the most easily pronounced words by two examinations. Also, when comparing the discriminant rate among groups we get some interesting results in Table II:

1. Sadness is a simple emotion to discriminate. This emotion is expressed in two different classes. (The rate of discrimination is 81.8%)
2. Surprise is quick to discriminate against emotions (59.1%). But in a community of nursing experiences, it is easy to misinterpret (The mistaken rate is 41.6 percent ). Groups who have not had a nursing experience are discriminated against better than those with a nursing experience (70 percent:50 percent of the discrimination rates). The next is Angry, which is equal to Surprise, quick to confuse for Sadness as a group of nurses (41.6%). Next is Angry, the discriminant rate is 40.9%.
3. Joy (the discriminating rate of nursing encounters is 9.1 percent) is the worst discriminated emotion; Sadness and Anger (the error rate is 33.3 percent) are easy to confuse in groups. Surprise is easy to mistake in a community with no nursing experience (60% of errors)

TABLE II. THE RESULT OF DISCRIMINANT ANALYSIS, ABOUT

EFFECTIVE PARAMETERS, DISTANCE, AND ERROR RATE

Table

Description automatically generated

To that end, the linear analysis methods of discrimination were rewritten to deal for non-crisp participants [47]. The test section of this paper stipulates various goals: 1) examines the performance gap, compares our suggestions with the fully controlled environment; 2) analyzes the possible advantages of refining class members through the methods proposed; and 3) tests the impact of other performance variables, such as classes number or bag size. However, more studies to study more complex data configurations should be promoted. The results of the experiment are encouraging.

Another research in connection with Discriminant Analysis concerns breast cancer. They identified the image analysis of brain tissue using hair wavelet texture features to distinguish Benign, DCIS and CA bridal organ lesion. A classifier development method consists of 2 steps: extraction and classification of features. In the extraction process, we have therefore extracted texture features from wavelet transformed images. In the classification process, they generated three classifiers using statistically discriminant analysis, neural networks (rear propagation algorithm) and SVM from each image of extracted features (support vector machines). In this study they conclude that it is the second-level wavelet transformed images used in discriminant functions that are the best classified histological parts of breast tissue in the texture.

Prediction of health status has become extremely important. The study of large data plays a crucial part in perfectly predicting this. Asthma is a serious symptomatic chronic condition. Asthma is a chronic death-driving condition. Researchers concentrated on this to help decide on the timely use of predictive analysis to forecast the disease [50]. A predictive analysis model for Random Forest asthma prediction is proposed in this report (PAM-RF). A random forest method has been developed to predict the classification of patients suffering from asthma. A multi class severity level for asthma disorder is forecast in this proposed Predictive Analytic Model for (PAM-RF). A multi class severity level for asthma disorder is forecast in this proposed Predictive Analytic Model for Four different groups, including Mild, Moderate, Severe and Weighted, were included in the seriousness of the disease, shown in Table III and . The PAM-RF approach proposed used the methodology for predicting disease through statistical classification.

TABLE III. QUANTITATIVE PERFORMANCE USING VARIOUS ACTIVATION FUNCTIONS FOR CLASSIFICATION OF IDC AND NO IDC TISSUES

Table

Description automatically generated

Chart, bar chart

Description automatically generated

Fig. 5. The Confusion Matrix

For the construction of a training model with information, the Random Forest machine learning algorithm is used. Random forests are known as decision-making trees [50]. The random forest algorithm uses two stochastic mechanisms to create separate decision trees. Firstly, before building each tree, the algorithm draws a sample bootstrap [51] from the data. Secondly, each decision tree is obliged to optimize its divisions over a randomly chosen predictor subgroup. Random forests are one of the most effective and popular supervised grading algorithms that can also perform regression and grading tasks. It's a key algorithm for random forest classification. These algorithms handle continuous values or values. This classification algorithm is supervised to create and mix multiple decision-making trees into a single forest and it works randomly [52]. As the name implies, the more trees in the forest, the more accurate the prediction and accuracy is typically generated by the Random Forest from many decision trees. This led to the prediction of each tree being called a vote in the proposed method to define an attribute-dependent new entity. The mark must be the class with the most votes. The random forest blends simplicity and adaptability of decision-making trees, which significantly improves precision. The model's results indicate that  PAM-RF will accurately identify and forecast asthma datasets for the future.

One of the most popular algorithms in the ML and health industry, where each algorithm is important, is the Naive Bayes callisification algorithm. One of the most interesting cases was the detection of Coronavirus disease using the naive bayes classifier.

One of the most popular algorithms in the ML and health industry, where each algorithm is important, is the Naive Bayes classification algorithm. Bayesian networks (BNs), alternatively referred to as belief networks or guided acyclic graphic models (DAGs), are graphical representations of the probabilistic dependencies between random variables and the approximate probabilistic inference obtained using statistical and computational methods within those variables [53]. Bayesian networks consider random variables, express them as a collection of nodes, and draw arcs to represent probabilistic causal relationships and conditional dependence between variables.

About 90% of heart attacks are preventable according to the paper [54]. Machine learning is a game changer in the health care industry when it comes to predicting heart disease. Heart disease is predicted in this research paper using Decision Tree, Naive Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbor, and logistic Regression algorithms. The algorithms' output was evaluated using Accuracy, Precision, AUC, and F1-score. The experimental results indicate that the Random Forest algorithm is more effective at predicting heart disease than other supervised machine learning algorithms, with an accuracy of 83.52 percent. Random forest classifiers have an F1-score of 84.21 percent, an AUC of 88.24 percent, and a precision score of 88.89 percent, respectively [54].

Biomedical signals' latest developments have been made possible by advances in Bayesian network analysis, which now incorporate biomedical signals and helps make decisions based on knowledge of unpredictable results More recent systems use probabilistic approaches, such as Bayesian Networks, to help model machine learning determine when things go wrong and diagnose machine learning issues [13]. Coronavirus was one of the most fascinating of the conditions discovered by a classifier that turned out to be virulent using the naïve bayes.

Learning is important for disease prediction. COVID – 19 has developed into a pandemic for humanity in the modern era. It is a communicable disease, and diagnosis results take between 12 and 24 hours to receive. It is not feasible to conduct the test on a large population in various remote and high-altitude regions, as well as due to the exponential growth of COVID – 19 in various parts of the world. The incubation period for COVID – 19 virus is between 2 and 14 days, but is usually about 5 days. The most common coronavirus symptoms [55] are fever, fatigue, cough, nasal congestion, runny nose, sore throat, diarrhea, and conjunctivitis. Decreased sense of smell and taste, Skin rashes, Ache and pain in the body Lips or face that are red, Speech Impairment, Respiratory Problems

While Nave Bayes appears to be a simple classification algorithm, it plays an important role in predictive modeling [56]. It is classified as a probabilistic classifier because it is entirely based on Bayer's theorem. Naive Bayes is, in essence, a model of conditional probability. It is possible to identify an instance of a given problem using this model, which is denoted by the vector X= (x1,..., xn), where n denotes the characteristics of independent variables. A collection of cases was considered, and data sets were programmed.

Probabilities were determined for each class and condition in Table IV. When test data is given, we obtain probabilities for different groups based on the provided symptom information. The data can be used to assign the patient to a probability-based class. A person's COVID – 19 status can be determined based on the likelihood value [57]. For example, looking at the figure, it can be said that the highest probabilities of coronavirus are Sore throat, cough or Runny nose, which is 0.666. The data mining techniques can be used in conjunction with the Naive Bayes classifier. Evidently, the collaboration of these methods is critical for diagnosing coronavirus infection, or COVID – 19, which has been declared a pandemic by the World Health Organization. The proposed mechanism demonstrated promising results, which could result in further refinements involving the use of data mining, machine learning, artificial intelligence, and information technology for coronavirus diagnosis [57].

Ultrasound (US) is a commonly used screening tool in a variety of fields of medicine. Regrettably, US photographs contain a plethora of objects. Lesions have been detected using Computer-Aided Diagnosis (CAD) techniques [58]. By analyzing the characteristics of lesions, CAD has been used to diagnose lesions using linear discriminant analysis (LDA) [59]. CAD has been particularly beneficial in the field of mammography [60], where it has been used to implement various techniques such as dynamic programming and radial gradient-based segmentation.

FCN, U-Nets, Dilated Residual Networks (DRN), and mask R-CNN are also promising deep learning strategies [61]. Joshi et al. used one-dimensional ultrasound in conjunction with random forest classifiers [62]. Additionally, linear discriminant analysis (LDA), support vector machines (SVM), K-nearest neighbor, and Naive Bayes classifiers are available [63]. Segmenting and detecting patterns in Ultrasound images is a difficult process due to the sensor's noise and exposure to the patient's position and physical features. Additionally, we must distinguish between internal structures and tumors. The aim of this paper is to demonstrate how we correctly identified and detected a cancerous mass in images using a variety of machine learning techniques and tumor properties. Among the approaches we tested, the Fast Shift algorithm produced the most promising results. This algorithm was more computationally complicated.

We want to integrate and compare additional segmentation and post-processing methods in the future to achieve more reliable and stable data.

TABLE IV. CORONAVIRUS SYMPTOMS AND PROBABILITIES

Table

Description automatically generated

1. LITERATURE REVIEW

The purpose of [64] was to detect breast cancer using Machine Learning algorithms, and the type of cancer they focus on is called Invasive Ductal Carcinoma (IDC), which is the most common subtype of all carcinomas.The genuine scanned images are split into 50x50 pixels to form samples of 277524 IDCs and non-IDCs. Pathology images were captured using a scanner to collect images of histopathology glass plate and also the size of this dataset is 5.8 Gigabytes. Their data are partitioned into nearly 2:1 ratios, non-IDCs and IDCs, and the training and test data ratios are 80:20. For this paper, two models of convolutional neural networks (CNN), Depthwise Separable Convolution (IDCDNet) and Standard Convolution (IDCNet), which is a class of multi-layer neural networks built for use on two-dimensional data such as photos and videos. The former algorithm involves 7-layers of CNN, 6-layers are for performing depthwise separable convolution and pooling, and the remaining layer is fully-connected. The latter algorithm also uses 6-layers standard convolution and pooling, and the rest is used for the same purpose. Here they used different activation functions, such as Rectified Linear Unit (ReLU), sigmoid and hyberbolic tangent functions. They were used for performance evaluation, and then each model was developed to minimize the binary cross-entropy loss function using the Adagrad optimizer.

Table I shows Precision (P1), Recall (R1), Sensitivity (S1) Specificity (S2), F-measure (F1) and Accuracy over different algorithms. P1 has been used to determine the proportion of IDC identified from the overall number of IDC patches. To approximate the proportion of IDC accurately predicted from the entire IDC automatically predicted, R1 and S1 are assessed. The proportion of Non-IDC patches successfully predicted from total actual Non-IDC patches is described as S2. The F1 and balanced precision are captured to display the trade-off between false positive and false negative minimization.To evaluate the accuracy of the IDC while using depthwise separable convolution and standard convolution, Gaussian noise of 0.1 and 0.2 is added to layers. As a result, they obtained 87.13% accuracy using Standard Convolution and 85.98% accuracy in Depthwise Separable Convolution, and concluded that performance had a positive effect on both models when the ReLU function was used, and that the best result was obtained when applying the ReLU function with standard convolution (see Table V).

TABLE V. QUANTITATIVE PERFORMANCE USING VARIOUS ACTIVATION FUNCTIONS FOR CLASSIFICATION OF IDC AND NO IDC TISSUES

Table

Description automatically generated

The paper [65] has focused on diabetes, one of the severe perils to human well-being, as that disease is correlated to the other disorders. The most significant manifestation of diabetic microangiopathy is diabetic retinopathy (DR), which is one of the most frequent complications of diabetes. The diagnosis of diabetic retinal complications is currently mostly determined by the diagnostic images. The fundus photographs are now the most important methods for diagnosing retinal disorders, but the procedure is difficult to detect and time-consuming. This paper uses digital data from diabetes medical records, diabetes glycosylation and diabetes biochemical test data, profound methods of learning and medical diabetes in combination with the use of the Convolution Neural Network Methodius (CNN) for a diagnosis model and uses the electronic medical records of 301 diabetically hospitalized patients for half a decade starting at 2009. The total data are more than 3500 pieces of data, and the distribution is divided into 1:1 ratio, DR and non-DR. As for the distribution of Training and Test data, this distribution was recorded as 75% and 25% respectively. This analysis makes two major contributions: 1) The CNN approach is used in this paper with one-dimensional unrelated data sets and solves the problem of how to convert uniformly useless data. 2) The composition of the CNN model and the BN layer stops gradient dispersal, speeds up training and increases model precision. An adaptive learning rate algorithm is also applied to refine the model of profound learning. The project overhauls the existing LeNet network structure, adds BN layers for the modern BNCNN model, essentially prevents the gradient diffusion, speeds training and improves model accuracy. In order to diagnose diabetic retinopathy, Multi-layer perceptron (MLP) and logistic methods in the same dataset were also used. The consistency of the preparation and testing of the three experimental approaches can be found in Table VI below. The research paper concludes that the model which is created has the best diagnostics precision of approximately 99%, increasing the overall machine learning algorithms by more than two percent. This research lays the groundwork for the early detection of diabetic problematic retinopathy and the automation of diagnostic procedures. It delivers excellent results by combining deep learning with electronic medical record content.

TABLE VI. INNOVATIVE OUTCOMES

Table

Description automatically generated

The article [66] concentrates on constructing the model that can predict near-term mortality in cirrhosis-hospitalized patients of the University of Virginia Health System with both a logistic regression and a long-term memory neural network. The primary goal of the study is that this model be integrated into an electronic medical process in order to provide precise predictions of decompensation and death in reality. The complete dataset employed in this study included 340.553 observational data from de-identified health record entries for 1783 patients with cirrhosis, each hospitalizing for some time for four years starting December 2012 at the University of Virginia Health System. Clinical colleagues indicated that a 12- to 24-hour timeframe before mortality would increase the effectiveness of treatments used to avert hospital death more effectively. This period was therefore used as the time frame for short-term mortality predictions. The actual data set encompassed binary index values in the resulting column: "1" refers to hospital mortality and "0" refers to non-mortality. Clinical colleagues indicated that a 12- to 24-hour timeframe before mortality would increase the effectiveness of treatments used to avert hospital death more effectively. This duration was therefore used as the time frame for short-term mortality predictions. The actual data set encompassed binary index values in the resulting column: "1" refers to hospital mortality and "0" refers to non-mortality. This research introduces a new method for forecasting near-term mortality in cirrhosis patients and contrasts a more conventional (logistic regression) methodology with the most advanced technology, namely RNN-LSM. During this study, the RNN-LSTM model established continues to demand refining. In future, a logistic regression model may be implemented into UVA's electronic medical record system or an updated version of RNN-LSTM model. This enables the model to automatically produce mortality forecasts as new observations are taken.  Furthermore, the method should either alert when the fast mortality is forecast or change the model so that each new calculation would give the system a revised mortality chance.  The use of machine learning, as indicated by the findings of this research, could help identify decompensation in cirrhosis patients sooner and more accurately. The use of machine learning, as indicated by the findings of this research, could help identify decompensation in cirrhosis sooner and more accurately. The professionals will be warned of the incidents in due time, helping them to respond when it is already possible to influence patient conditions and preventable deaths, with input from data based machine learning systems, as established in this research.

One of the primary causes of death has been declared by the World Health Organisation, pulmonary and high blood pressure heart attack. This disease especially becomes life-threatening if not controlled. The system tries to resolve this problem through the construction of a prototype of an interactive prediction system which provides a person with a risk factor calculated against heart attack. This model [67] is implemented using the data set from the University of California Irvine's (UCI) machine learning library, Rapid Miner for refining the data set, and Anaconda V2.7 for building the classifier. The findings are made accessible to the customer through a Web interface with a graphical user interface that is simple to use. Rapid Miner was used for establishing the best matching algorithm for a data set and comparing its processes in Rapid Miner with 4 algorithms, namely Naïve Bayes, Decision Trees, K-Nearest Neighbour, and Random Forest. The results of this operation demonstrate that Naïve Bayes gives the data set the greatest precision. A device prototype is introduced that classifies an individual as a potential cause. The data collection has been retrieved from the machine learning repository and pre-processed by UCI. The final data set consists of 13 variables (Age, Sex, Chest pain, Blood Pressure at Rest, Cholesterol, Fasting blood sugar, Electrocardiographic results at Rest, Peak heart rate archived, Exercise induced angina, ST depression induced by exercise relative to rest, Slope of the peak exercise ST segment, Number of major vessels, Thal) or functions of predictors and one number-name response variable (Num – response variable). When number is 0, it means that fewer than 50 percent of blood vessels decrease, i.e. the forecast is "low risk" and 1 indicates that the prediction is "high risk" with more than 50 percent of blood vessels reduced. Gaussian Naive Bayes algorithm has been used as a classification because the data follows a regular distribution. Naive Bayes showed the highest accuracy of 81.25% for all state-of-the-art classification algorithms on the datasets.

In latest years, the grading of breast cancer has been the subject of concern in information technology in the area of health care since it is the second leading cause for the deaths of women from cancer. Breast cancer can be determined by a biopsy in which the tissue is cut and microscopically analyzed. The diagnosis is dependent on the histopathologist's qualifications who searches for irregular cells. However, if the histopathologist is not well qualified, the interpretation and diagnosis may be incorrect. Recent progress in image processing and machine learning has led to an interest in trying to create trustworthy diagnostic system to improve diagnostic accuracy. In this article [68], two automated methods are compared to classify photographs of the histology of breast cancer into benign and malignant and into benign and malignant subclasses. The very first technique is focused on the extraction of a range of handcrafted features coded by two coding models (word sac and localization limited linear coding) and trained with vector supporting machines, while the second is a neural network architecture. The experiments on tiny data sets of 500 and 92 pictures were performed. The efficiency of the handmade characteristics approach, in which coding models were used to encode the descriptors to produce images, was too low compared to CNN. Convolutional layers neural networks have been used to fully replace classifiers with fully integrated layers to train manufactured features, which have contributed to an improvement in handcraft performance. Convolutional neural networks are state-of-the-art and show the ability to solve difficult classification tasks. In this paper, it is proposed that CNN topology has been the best in the past, where it have been achieved a performance between 96.15% and 98.33 for the binary classification task.

Diabetes Mellitus, or Diabetes, has also been depicted to be even worse than HIV and Cancer. It evolves just before insulin levels are high for a long time. It was referenced importantly as a health risk and a main cause for vision loss and renal failure for the development of Alzheimer. Almost one third of all diabetes patients are not aware of the disease, according to the Joslin Diabetes Center [69]. Diabetes is a lifelong metabolic condition that either develops if the pancreas does not produce sufficient insulin or if the organism cannot use insulin properly. Insulin is a blood sugar, type 1 diabetes, type 2 diabetes and gestational diabetes hormone. Three forms are available: Diabetes type 1, Diabetes type 2 and Gestational Diabetes, with the latter being the most treatable. Data were collected via questionnaires to perform this analysis. The conversation with the psychiatrist and the two diabetologists were the basis of the questionnaire. Through the questionnaires and negotiations, patients’ daily life choices and lifestyle have been asked. These include fast food, food alongside the street, the measure of sugar, potatoes and the everyday consumption of rice, the regular fitness frequency and duration to sleep, etc. and it turns out that as a result of the conversation with doctors they are the main cause for making people diabetic. Moreover, BMI (Body mass index), BP (Blood Pressure), genetic diabetes, era, waist circumference, chronic condition history and the diabetic type. Since the factors were different, different indications had to be determined. There were three Blood Pressure characteristics – the mild, average and extreme intakes of potatoes and rice, for instance. Sleep was split, for example, by the way you fell asleep sooner or later. Certainly, in one day each characteristic was answered and it would avoid overfitting to receive such answers in 50 days. The first mathematical test implemented on the visualized data was the Pearson’s Chi-Squared Test of Independence. With the use of the R Studio, the following variables were found to be mutually dependent upon the performance of the Chi-Squared Test (Table VII) for all variable combinations. Category - BP (a diabetics on non-diabetic depends whether they have or not a problem with their blood pressure); Category – junk; Category - rice; Category - Potatoes; Gender - Roadside,  Category - BP. These dependencies have created a way of more analysis, as this knowledge has been used to guide the growth of a tree during the modeling process, including branching and prunning. For a strongly categorical dataset, a CART prediction model with 75% accuracy has been implemented. *Fig. 6* states that a person with low or normal BP (0 to low, 1 to normal and 2 to high) who receives frequent or occasional food on the roadside without inherited diabetes is less likely than a person with a low or normal BP, who receives frequent or occasional food intake, has a lower likelihood of becoming diabolical and may also be diabetic if rice intake is intact Blood pressure, along with others such as roadside food, late sleeping, family history and legacy, intakes of rice and amount of physical activities an individual carries out every day, has been described as a major factor in the cause of diabetes. Blood pressure. Therefore, one should enjoy any part of one's life, but no harm is due to a little care in one's everyday routine

TABLE VII. CHI-SQUARED TEST RESULT

A picture containing text, cabinet, several

Description automatically generated

A picture containing diagram

Description automatically generated

Fig. 6. The CART plot

1. CONCLUSION

In fact, healthcare is one of the major areas of human life that play an important role and doing the right thing makes it grow in good fashion. Even if physicians have sufficient knowledge, systems need to be created to facilitate their work, and Big Data and Machine Learning can help them in this situation. Big Data is in general a data collection and can relatively increase its volume. Every large firm, for example Facebook and Youtube, users, stores megabytes of data per second and possesses petabyte machines. In general, data can be structured, unstructured or semi-structured. In order to create better systems, the data needs to be collected first. The health data can be collected for 3 purposes; Hospital Claims Data, Patient Clinical Data and Research & Development Trial Data. How are the information gathered? In 5 steps, this happens. To understand the state of the data, the data is collected and examined first. The first processes are, surely, preprocessing and cleaning. If the values are extracted for additional properties, i.e. columns, they are extracted. It is integrated and then consolidated. Finally, a database is created, queries are processed and feedback is provided. Once the data is ready, of course, the application of the algorithms remains, and this is done through Machine Learning.

Machine Learning differs from the traditional method in that you compute and create a new model. Machine Learning is also divided into 4 categories; supervised, semi-supervised, unsupervised and reinforcement learning. A problem was then discussed, and it was observed that machine learning is very important in the healthcare industry. Indeed, the algorithmic tools have proved to be effective for many diagnostic purposes in the detection of many kinds of cancers, along with other methods, including stool tests for early stage detection, in categorizing disease that arises in the heart, as well as testing for ECG findings that suggest a form of cardiac infection, in assessing for diabetes, and checking blood test results for gestational hypertension. Many algorithms utilized in healthcare include the: the Support Vector Machine, which is widely used; the Neural Network, which is fairly standard; and Logistic Regression, which is less often used; these are some of the most commonly used algorithms in healthcare. Although originally an ambiguity and all in all of these methods, medical intelligence methods using Machine Learning has ended up being much more nuanced after this comparison. Finally, though machine learning's potential to aid medicine is astounding, no cohesive solution has arisen, and these solutions have been used mostly for academic purposes. This, I believe, is because Machine Learning is a relatively young area.

1. FUTURE WORK

Since Machine Learning is a fresh field of study, that has just begun, analysis and evaluation are underway to investigate how well the algorithms work. We aim to find information that is accurate and unbiased with regard to physician and medical information, thereby allowing us to reduce the number of incorrect conclusions or mistakes. I assume that the algorithms are yet to be explored, so novel varieties of outcomes can not be obtained, but it will be during the development of the field. Force predicting and classifying are an interesting feature of the future, both for researchers and doctors who need to understand what is happening in the healthcare field; however, many other researches should be done with existing algorithms or combination of all, to get better results. Using Naive Bayes network, Linear or Logistic Regression, Neural Network , classification and forecasting have been succesfully done because they're more practical. What is lacking is a mechanism established to help doctors with the treatment, in addition to which their authority is respected. Few authorities have made many attempts at the hospital level, but no one has tried to put together a system in the manner that will make the whole hospital work together with improved and advanced systems, to save time and patients.

REFERENCES

1. M. A. Lambay and S. Pakkir Mohideen, "Big Data Analytics for Healthcare Recommendation Systems," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2020, pp. 1-6, doi: 10.1109/ICSCAN49426.2020.9262304.
2. Bouwens, J., & Krueger, D. M. (n.d.). Embracing Change: The healthcare industry focuses on new growth drivers and leadership requirements. Retrieved April 13, 2021, from <https://www.russellreynolds.com/insights/thought-leadership/embracing-change-the-healthcare-industry-focuses-on-new-growth-drivers-and-leadership-requirements>
3. Buneman, P. (1997). Semistructured data. Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems - PODS ’97. doi:10.1145/263661.263675
4. C. Tankard, "Big Data Security", Network Security Newsletter, Elsevier, ISSN 1353-4858, July 2012
5. Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. 2013 International Conference on Collaboration Technologies and Systems (CTS). doi:10.1109/cts.2013.6567202
6. R. Bhardwaj, A. R. Nambiar and D. Dutta, "A Study of Machine Learning in Healthcare," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, Italy, 2017, pp. 236-241, doi: 10.1109/COMPSAC.2017.164.
7. Buyya, R., Calheiros, R. N., & Dastjerdi, V. A. (2016). Big Data: Principles and Paradigms (1st ed.). Morgan Kaufmann.
8. N. G. Maity and S. Das, "Machine learning for improved diagnosis and prognosis in healthcare," 2017 IEEE Aerospace Conference, Big Sky, MT, USA, 2017, pp. 1-9, doi: 10.1109/AERO.2017.7943950.
9. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011, http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI\_big\_data\_full\_report.ashx
10. T. Mitchell, “ Machine Learning”, McGraw Hill
11. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science. doi:10.1126/science.aaa8415
12. Allix, N. M. “Epistemology And Knowledge Management Concepts

And Practices”. Journal of Knowledge Management Practice.

1. Dua, S., Acharya, U. R., & Dua, P. (2016). Machine Learning in Healthcare Informatics. Berlin: Springer Berlin
2. K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 910-914, doi: 10.1109/ICECA.2018.8474918.
3. Sabharwal, A., & Selman, B. (2011). Book review. Artificial Intelligence, 175(5-6), 935–937. doi:10.1016/j.artint.2011.01.005
4. Anil Mathur, G. P. “Socialization influences on preparation for later

life”. Journal of Marketing Practice: Applied Marketing Science, vol.5, 163 – 176, 2007.

1. LP. Kaelbling and Aw. Moore, “Reinforcement Learning A Survey”,

Journal of Artificial Intelligence Research, vol-4, pp.237-285, 1996A

1. M. Ambigavathi and D. Sridharan, "Big Data Analytics in Healthcare," 2018 Tenth International Conference on Advanced Computing (ICoAC), Chennai, India, 2018, pp. 269-276, doi: 10.1109/ICoAC44903.2018.8939061.
2. G. Harerimana, B. Jang, J. W. Kim and H. K. Park, "Health Big Data Analytics: A Technology Survey," in IEEE Access, vol. 6, pp. 65661-65678, 2018, doi: 10.1109/ACCESS.2018.2878254.
3. Jenna Wiens, Erica S Shenoy, Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology, Clinical Infectious Diseases, Volume 66, Issue 1, 1 January 2018, Pages 149–153, <https://doi.org/10.1093/cid/cix731>
4. Mansi T, Mihalef V, Sharma P, Georgescu B, Zheng X, Rapaka S, Kamen A, Mereles D, Steen H, Meder B, Katus H, Comaniciu D (2012) Data-driven computational models of heartanatomy, mechanics and hemodynamics: an integrated framework. In: 9th IEEE international symposium on biomedical imaging (ISBI)
5. Mao Y, Chen W, Chen Y, Lu C, Kollef M, Bailey T (2012) An integrated data mining approach to real-time clinical monitoring and deterioration warning. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining

(KDD ‘12), New York, NY, USA

1. Calhoun B, Lach J, Stankovic J, Wentzloff D, Whitehouse K, Barth A, Brown J, Li Q, Oh S, Roberts N, Zhang Y (2012) Body sensor networks: a holistic approach from silicon to users. Proc IEEE 100(1):91–106
2. P. Naraei, A. Abhari and A. Sadeghian, "Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data," 2016 Future Technologies Conference (FTC), 2016, pp. 848-852, doi: 10.1109/FTC.2016.7821702.
3. A. Caceres C, Rikli A (2012) The digital computer as an aid in the diagnosis of cardiovascular disease. Transactions of the New York Academy of Science. 23(3 series II):240–245
4. Tracy K, Dykstra B, Gakenheimer D, Scheetz J, Lacina S, Scarfe W, Farman A (2011) Utility and effectiveness of computer-aided diagnosis of dental caries. Gen Dent 59(2):136–144
5. M. S. Kazemi, K. Kazemi, M. A.Yaghoobi, and H. Bazargan, “A hybrid method for estimating the process change point using support vector machine and fuzzy statistical clustering,” Applied Soft Computing, Vol. 40, pp. 507-516, 2016.
6. Verplancke T, Van Looy S, Benoit D, Vansteelandt S, Depuydt P, De Turck F, Decruyenaere J. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. BMC Med Inform Decis Mak 2008; 8: 56.
7. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20: 273-297.
8. S. Dreiseitl, and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review,” Journal of biomedical informatics, Vol.35(5), pp. 352-359, 2002.
9. S. Abe, “Support vector machines for pattern classification,” Springer, U.K., Vol.2, 2005.
10. K. Fukunaga, “Introduction to statistical pattern recognition,” Academic press, U.S.A., 2013.
11. N. G. Pavlidis, D. P. Hofmeyr, and S. K. Tasoulis, “Minimum Density Hyperplanes,” arXiv:1507.04201, 2016.
12. Tracy K, Dykstra B, Gakenheimer D, Scheetz J, Lacina S, Scarfe W, Farman A (2011) Utility and effectiveness of computer-aided diagnosis of dental caries. Gen Dent 59(2):136–144
13. Wang, S.-C. (2003). Artificial Neural Network. Interdisciplinary Computing in Java Programming, 81–100. doi:10.1007/978-1-4615-0377-4\_5
14. L. Liu, “From data privacy to location privacy: models and algorithms,” Proceedings of the 33rd international conference on very large data bases, VLDB Endowment, pp. 1429-1430, Austria, September 2007.
15. C. G. DeGroff, S. Bhatikar, J. Hertzberg, R. Shandas, L.Valdes-Cruz, and R. L. Mahajan, “Artificial neural network–based method of screening heart murmurs in children,” Circulation, Vol.103(22), pp. 2711-2716, 2001.
16. I. I. Ororbia, G. Alexander, C. L. Giles, and D. Kifer, “Unifying Adversarial Training Algorithms with Flexible Deep Data Gradient Regularization,” arXiv preprint arXiv:1601.07213, 2016.
17. B. G. Horne, “Progress in supervised neural networks,” IEEE Signal Processing Magazine, Vol. 10(1), pp. 8-39. 1993.
18. C. S. Dangare, and S. S. Apte, “A data mining approach for prediction of heart disease using neural networks,” International Journal of Computer Engineering and Technology (IJCET), Vol.3(3), pp. 30-40, 2012.
19. P. Naraei, A. Abhari and A. Sadeghian, "Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data," 2016 Future Technologies Conference (FTC), 2016, pp. 848-852, doi: 10.1109/FTC.2016.7821702.
20. C. Venkatesan, P. Karthigaikumar, A. Paul, S. Satheeskumaran and R. Kumar, "ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare Applications," in IEEE Access, vol. 6, pp. 9767-9773, 2018, doi: 10.1109/ACCESS.2018.2794346.
21. P. N. Ranaware and R. A. Deshpande, "Detection of arrhythmia based on discrete wavelet transform using artificial neural network and support vector machine," 2016 International Conference on Communication and Signal Processing (ICCSP), 2016, pp. 1767-1770, doi: 10.1109/ICCSP.2016.7754470.
22. S. Ahmad, N. M. Ramli and H. Midi, "Outlier detection in logistic regression and its application in medical data analysis," 2012 IEEE Colloquium on Humanities, Science and Engineering (CHUSER), 2012, pp. 503-507, doi: 10.1109/CHUSER.2012.6504365.
23. Y. Jiang et al., "SecureLR: Secure Logistic Regression Model via a Hybrid Cryptographic Protocol," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 1, pp. 113-123, 1 Jan.-Feb. 2019, doi: 10.1109/TCBB.2018.2833463.
24. M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal, "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104210.
25. A. Wang, N. An, Y. Xia, L. Li and G. Chen, "A Logistic Regression and Artificial Neural Network-Based Approach for Chronic Disease Prediction: A Case Study of Hypertension," 2014 IEEE International Conference on Internet of Things (iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom), 2014, pp. 45-52, doi: 10.1109/iThings.2014.16.
26. Yahan Gao, Ohno, Y., Feng Qian, ZhaoYan Hu, & ZhiXiong Wang. (2013). The discriminant analysis of the voice expression of emotion - Focus on the nursing experience -. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). doi:10.1109/embc.2013.6609737
27. M. Pérez-Ortiz, P. A. Gutiérrez, M. Carbonero-Ruz and C. Hervás-Martínez, "Adapting linear discriminant analysis to the paradigm of learning from label proportions," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 2016, pp. 1-7, doi: 10.1109/SSCI.2016.7850150.
28. Akbar, W., Wu, W.-P., Faheem, M., Saleem, S., Javed, A., & Saleem, M. A. (2020). Predictive Analytics Model Based on Multiclass Classification for Asthma Severity by Using Random Forest Algorithm. 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE). doi:10.1109/icecce49384.2020.9179467
29. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26
30. V. Shanmugarajeshwari and M. Ilayaraja, "Chronic Kidney Disease for Collaborative Healthcare Data Analytics using Random Forest Classification Algorithms," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-14, doi: 10.1109/ICCCI50826.2021.9402574.
31. Neapolitan RE (2004) Learning bayesian networks. Prentice Hall, New Jersey
32. P. Sujatha and K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-7, doi: 10.1109/INOCON50539.2020.9298354.
33. Zhang, Yong, et al. "Gastrointestinal tract symptoms in coronavirus disease 2019: Analysis of clinical symptoms in adult patients." 2020.
34. Zakikhani, K., et al. "A Failure Prediction Model for Corrosion in Gas Transmission Pipelines." Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 2020, p. 1748006X2097680, doi:10.1177/1748006x20976802.
35. S. Bhatia and J. Malhotra, "Naïve Bayes Classifier for Predicting the Novel Coronavirus," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 880-883, doi: 10.1109/ICICV50876.2021.9388410.
36. Karen Drukker, Maryellen L. Giger, Karla Horsch, Matthew A. Kupinski, Carl J. Vyborny, and Ellen B. Mendelson. Computerized lesion detection on breast ultrasound. Medical physics, 29 7:1438–46, 2002.
37. Karla Horsch, Maryellen L. Giger, Luz A. Venta, and Carl J. Vyborny. Computerized diagnosis of breast lesions on ultrasound. Medical physics, 29 2:157–64, 2002.
38. Sheila Timp and Nico Karssemeijer. A new 2d segmentation method based on dynamic programming applied to computer aided detection in mammography. Medical physics, 31 5:958–71, 2004.
39. ]Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40:834–848, 2016.
40. Neil Joshi, Seth Billings, Erika Schwartz, Susan Harvey, and Philippe Burlina. Machine learning methods for 1d ultrasound breast cancer screening. pages 711–715, 12 2017.
41. J. Jean Rossario Raj, S. M. K. Rahman, and Sneh Anand. Computational assessment of breast tumour differentiation using multimodal data. 2016.
42. H. Alghodhaifi, A. Alghodhaifi and M. Alghodhaifi, "Predicting Invasive Ductal Carcinoma in breast histology images using Convolutional Neural Network," *2019 IEEE National Aerospace and* Electronics Conference (NAECON), Dayton, OH, USA, 2019, pp. 374-378, doi: 10.1109/NAECON46414.2019.9057822.
43. Y. Sun, "The Neural Network of One-Dimensional Convolution-An Example of the Diagnosis of Diabetic Retinopathy," in IEEE Access, vol. 7, pp. 69657-69666, 2019, doi: 10.1109/ACCESS.2019.2916922.
44. E. Harrison, M. Chang, Y. Hao and A. Flower, "Using machine learning to predict near-term mortality in cirrhosis patients hospitalized at the University of Virginia health system," 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2018, pp. 112-117, doi: 10.1109/SIEDS.2018.8374719.
45. S. Manikandan, "Heart attack prediction system," *2017 International Conference on Energy*, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 817-820, doi: 10.1109/ICECDS.2017.8389552.
46. D. Bardou, K. Zhang and S. M. Ahmad, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks," in IEEE Access, vol. 6, pp. 24680-24693, 2018, doi: 10.1109/ACCESS.2018.2831280.
47. A. Anand and D. Shakti, "Prediction of diabetes based on personal lifestyle indicators," 2015 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 2015, pp. 673-676, doi: 10.1109/NGCT.2015.7375206.