# **Measuring the Limits of Machine Learning Algorithms in Predicting Patient Prognosis**

**Table of contents**

1. ABSTRACT[…………………………………………………………………………………3](#_bookmark1)

CHAPTER 1 ….. [4](#_bookmark2)

1. Introduction[….….………………………………………………………………………………………4](#_bookmark2)
   1. Background of the Study [….………………………………………………………………………7](#_bookmark3)
   2. Objectives […....……………………………………………………………………………………..7](#_bookmark2)
   3. Research Questions […..……………………………………………………………………………8](#_bookmark3)
   4. Ethical Considerations [….…………………………………………………………………………8](#_bookmark4)
   5. Project Timeline […...……………………………………………………………………………….8](#_bookmark5)

CHAPTER 2 .. [………………………………………………………………………………………...10](#_bookmark2)

1. Literature Review [.………………………………………………………………………………………10](#_bookmark2)

CHAPTER 3  [27](#_bookmark2)

1. Project Design and Methodology [...……………………………………………………………………27](#_bookmark2)

CHAPTER 4  [31](#_bookmark2)

1. [Results and Discussion………….…………………………………………………………………..…31](#_bookmark30)

CHAPTER 5  [45](#_bookmark2)

1. Conclusion […………...………….…………………………………………………………………..…45](#_bookmark30)

REFERENCES  [46](#_bookmark2)

# **Abstract**

As globally, the population is increasing day by day and living standards are going to compromise due to the unavailability of basic needs. Nowadays major causes of death across the world are the unavailability of resources in the early stages. Medical experts can submit mistakes while recognizing ailment changes. The data visualizations and interpreters in the initial part of this access provide an assessment between the causes and risk factors of demise across states and age groups. Enormous information examination assumes an urgent part to foresee future status of wellbeing and offers superior well-being results to individuals. To mechanize the cycle and foresee infections all the more precisely Artificial intelligence strategies are acquiring notoriety. AI expands its roots in medical diagnostics. Various acute and chronic diseases can be identified accurately at the initial level by using AI methods to prevent the undergoing development of complications.

The advent of technology has made it easier for humans to automate the process of patient prognosis and deduction of disease/issue from symptoms and report analysis. However, Machine Learning is only as capable as the data and the resources it has. It cannot expand and learn at will. But the capacity of digesting huge amounts of data and statistically analyzing to provide accurate predictions in the blink of an eye is what draws the researchers to using ML for this purpose.

This research aims to render out the limits of Machine Learning in achieving prognosis for patient health using Machine Learning. There are many possible techniques and paradigms for use in this research. Classification algorithms like Multi-layer Perceptron, Decision trees, SVM, or Regression algorithms like Logistic Regression. The output of this research will be the experimental proof and analysis of how much a Machine Learning algorithm can perform given a medical dataset to predict patient health.

**Keywords:** Machine Learning, Prognosis, Classification, Regression, Multilayer Perceptron, Decision trees, SVM, Logistic Regression

**CHAPTER 1**

1. **Introduction**

In recent years, there has been a magnified focus on the use of artificial intelligence in several domains to resolve complex issues, such as urban infrastructure, resident living environment, transport management, medical treatment, shopping, security assurance, and so forth. Further, the endorsement of artificial intelligence in health care is growing whereas evolving the face of health care delivery.

The initial information and communication technology endorsement in the health care region's primary contribution cost reduction and efficiency. The overall mobile devices arrival as well as placement abilities set up mobile health, which added consolidates to the health care system, like global accessibility, quickness, and observation. Another growing concept is smart health that is based on both smart city and internet of things use [Dirks et al., (2009)]. IoT allows connecting identifiers, sensors, devices, and computers via wired and wireless networks [Albrecht et al., (2013)].

The emerging smart health is isolated intelligent customized health services, typically with sensor data gathering and cloud processing. Emerging smart health examples include person-to-person audio or video home patient’s direction through smartphones [Luxton et al., (2012)]. IoT encrypted systems [Anzanpour et al., (2015)] cloud-based digital health sensor variables synchronization and categorization concepts. Emerging smart health IoT apps may be classified into telemedicine, emergency, medication, pharmaceutical packs, and others.

The emerging smart health model is a resolution that gives a provisional position for medical care observation that requires both varied ages population as well as increasing online health services. Such functions could be supported by gathering and refining patient-related data from portable sensors. Raw data wouldn't sufficiently provide health care system functions, which may lead to risky miscommunications. The emerging smart health model is reasonably insufficient in terms of health care coverage and ICT architecture [Jara et al., (2010)].

The intelligent medical diagnosis can recommend affordable medical examination proposals according to the patient's situation [Gil et al., (2016)]. Intelligent services that assist clients to create smarter decisions are pervasive because the Quality of Experience (QoE) is intensely important in smart health applications [He et al., (2018)].

However, the request for intelligent expertise involves an extremely well-set data processing  
perfectibility, which inspires smart homes authorized with Artificial Intelligence (AI) [Rasch  
(2014)]. More particularly, artificial intelligence uses various learning techniques to investigate  
self-regulating support activity and sensible decision-making. Therefore, the smart health  
empowered with artificial intelligence is predicted to be intelligent.

Machine Learning is a subfield of artificial intelligence where software programs become able  
to classify and predict results accurately without programming them explicitly [Das et al., 2015)].  
The learning process of machine learning software modules involves providing some data for those models, allowing those models to look for patterns into data, and make better decisions in the future based on the data provided. The main aim of machine learning is to allow software programs to be learned directly from provided data and adjust their results according to this data without the aid or interference of humans. Machine learning algorithms are often categorized into four categories: Supervised Algorithms (SA), Unsupervised Algorithms (UA), Semi-Supervised Algorithms (SSA), and Reinforcement Algorithms (RA) [Khan et al., (2013)]. In supervised algorithms data that is used to training modules (training set) is provided with labels. However, in unsupervised algorithms training data has no labels. In semi-supervised algorithms both labeled and unlabeled data are used for training modules. In reinforcement algorithms modules are trained from the environment through a trial and error process.

Deep extreme machine learning is a subfield of machine learning. Deep extreme  
machine-learning generally uses sequences of several layers to accomplish the feature extraction  
and classification tasks. Layers used in deep extreme machine learning are connected in a cascade  
manner so that the output of each layer is connected to the input of the following layer [Nguyen et  
al., (2019)]. With deep extreme machine learning, software modules can be learned and trained to  
accomplish classification and prediction tasks from images, sounds, videos, or text data. The  
performance and accuracy of deep extreme machine learning models can be very excellent and  
exceed human beings' performance. Deep extreme machine learning models are trained to  
accomplish classification or regression tasks by using a large number of datasets (data with labels)  
and powerful neural network structures.

An artificial neural network is another subfield of machine learning that is briefly  
stimulated by the human neural network and it employs different neurons to perform amassed  
tasks. This technique has high-level accuracy because it leans towards the updates itself without  
human interaction. The forward technique of neural network process different instructions, and if  
it finds errors, it back propagates automatically, and improves back neurons to process data with  
high accuracy. The study reveals that it was up to mark the level of accuracy in the diagnosing  
system of medical fields [Sibi et al., (2013), Soltani et al., (2016), Zebardast et al., (2013)].

The fuzzy logic controller is widely utilized in a different field, especially medical  
sciences to design diagnosis systems. It is a successful tool that is currently in use, different input  
parameters based on the disease are converted into fuzzy input sets, and these are processed by the  
fuzzy controller with expert rules. The input data is converted into crisp outputs and presented in  
the form of human-readable or understandable outputs. The fuzzy logic controller normally uses  
more than one rule to provide accurate results. It is a complete approach that process inputs in the  
inference engine along with the composition [Ahmad et al., (2019)].

Allahverdi *et. al.* (2007) stated that computer technologies have played a vital and decisive role in the development, investigation, and research of the medical field. The use of computer technology has been increased over time for diagnostic purposes in medical fields but there exist complexity and uncertainty in the diagnosis process due to the application of traditional measures in the shape of personal expertise and experiences as well, to overcome such shortcomings and flaws in diagnosis, intelligent and efficient system like fuzzy logic, machine learning is developed.

Palaniappan and Awang (2008) found that the medical field is entirely dependent on the application of investigation based on lab examination. For this reason, in the history of medical science, right from the developmental stages, time to time different decision support systems were developed. Due to the variations in the symptoms of the disease, these decision support systems were found inefficient and inaccurate as per the generation of medical reports, the reason behind was that, they were based on just past limited instructions which were now due to change in symptoms of heart disease failed to perform well. Heart disease diagnosis is very difficult for medical practitioners due to many uncertain risk factors. It is very difficult to decide whether a patient is suffering from heart disease or not.

Obenshain (2004) pointed out that many factors play an important role in the sickness of cardiovascular disease. These factors have a great relationship with the physiological outlook and lifestyle of the patients, so the detection of heart disease needed to be considered all these factors about the lifestyle and eating habits. These factors have certain variant elements about different cultural and social groups even within the country. In this situation a comprehensive decision support system that has the strength of variability and capacity of adjustment of new factors in the diagnosis of heart disease.

Avci (2009) observed that many simple computer technologies are in use for the diagnosis of heart disease which is operating under their defined platforms. It is always felt a need for such an exclusive hybrid computer technique that can express the capabilities of performing efficient diagnosis with the help of devised input classification technique based on unique features.

Wu *et. al.* (2002) declared that Fuzzy logic is a qualitative approach that has the potential to create precise things out of the imprecise nature of the world environment. Decision support system (DDS) based on a fuzzy logic system that relates to the fuzzy rule-based system. In the medical industry, DDS is used frequently for diagnosis purposes. Many computer-based systems consist of decision support systems. Decision Support System is using by some hospitals for simple data management purposes instead of advance medical utilities concern with the disease diagnosis and its dimensions. Medical decisions based on a doctor’s experience and expertise are established by just applying the comparative skills of different medical reports which reflects the current understating of the doctors regarding the nature of the disease. But the logic and application of the medical reports made at a single platform are altogether absent in the decision-making of the doctors. Fuzzy Logic has the capability of fulfilling such a critical deficiency invalid and reliable decision making.

Since the Centuries humanity has found different demonstrated medical services frameworks. To robotize the interaction and anticipate sicknesses all the more precisely AI techniques are acquiring fame in research local area. AI strategies work with improvement in the knowledge to a machine for more efficient performance later on utilizing insight from learning (Santosh A. Shinde and P. Raja Rajeswari, 2018).

Nowadays, wellbeing expectation in current life turns out to be a lot of fundamental. Huge information examination assumes a vital part to anticipate the future status of well-being and offers superior well-being results to individuals. A great deal of examination is going on prescient investigation utilizing AI methods to uncover better dynamics. Large information investigation cultivates extraordinary freedoms to anticipate future well-being status from wellbeing boundaries and give the best results (Venkatesh, Balasubramanian, and Kaliappan, 2019).

This research project aims to utilize Machine Learning for Patient Prognosis and Diagnosis using Data Analysis. The main aim behind this is to measure the limits i.e. how much Machine Learning can achieve in terms of Patient Prognosis when provided with the data resources.

**1.3 Background of the study**

Technology development has taken healthcare to zeniths never seen before. The stage now is automation. The necessity comes with the involvement of Machine Learning in the medical domain. The huge amount of data required to process for predicting patterns and identifying disease symptoms can only be targeted using Machine Learning. This data is normally used for patient prognosis.

Therefore, researchers around the globe are working on the automation of patient prognosis. The analysis of patient data and reports to predict whether the patient has a disease or not and provide a report on it is what has got the attention of the researchers in the medical domain. This research is focused on testing the mettle of Machine Learning algorithms to the limit given a dataset for prognosis.

**1.4 Objectives**

This research project aims to see how much Machine Learning can achieve in terms of predicting patient prognosis. To achieve this, the following steps are traced:

* Literature Review of papers already written in the same domain
* Dataset Finalization, Comprehension, and Analysis
* Data Pre-processing and Machine Learning Algorithm Design
* Performance Metrics for the Machine Learning Ensemble

**1.5 Research Questions**

The question that provides the drive to this research is as follows:

* What are the Limitations of using Machine Learning for the automation of patient prognosis?

**1.6 Ethical Considerations**

Ethics is a complicated subject that has only become more prominent during the advent of Big Data. The UK Data Service department also provides guidelines for ethical research with specific relation to Big Data. These guidelines will form the basis for this report's ethical approach. Some of the concerns that will be addressed are:

* Maintaining confidentiality in line with Birmingham City University (BCU) and DC guidelines,
* Anonymizing information that violates group privacy,
* Ensuring transparency in reasons for data collection,
* Ensuring data is only used for the direct purpose it has been requested,
* Referencing sources for all information used within the research project,
* Ensuring all data is stored in the correct location. DC information must remain on DC servers.

**1.7 Project Timeline**

Research projects are random and time-bound and the ability to meet a deadline is key to success. A project timeline lays out key project deliverables and the scope of their completion. This research project identifies time as its key resource. By efficiently allocating time to various tasks resource overload is minimized. Preventing resource overload minimizes the risk of quality decreasing.

The Gantt chart is identified as a strong tool for time management. The Gantt chart designed for this project is laid out below in Figure 1.1. Tasks are laid out in chronological order on the left-hand side. The timeframe for their completion is found along the X-Axis. By sticking to this schedule, the project will be delivered promptly to a high standard.

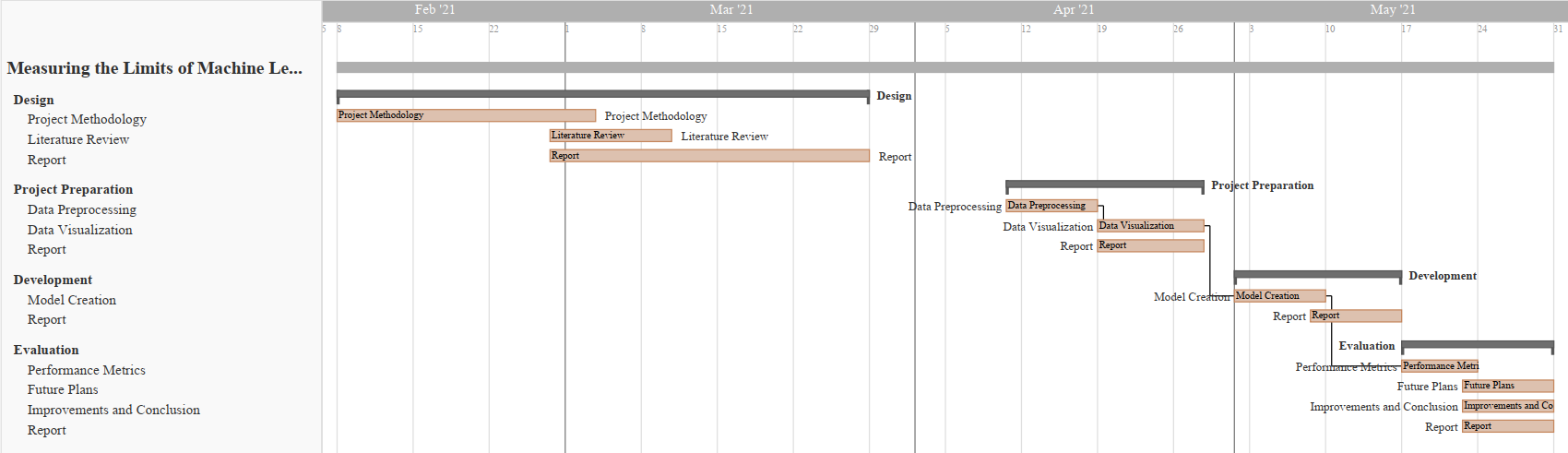


Figure 1.1: Gantt chart

**CHAPTER 2**

1. **Literature Review**

The process of reviewing, studying, and understanding research was done by peer researchers in the same domain through their research papers is known as Literature Review (LR). A strong LR provides validity to the integrity of the research using proven, published facts. The papers selected for LR in this research are summarized as follows.

Missing information is the most outstanding issue experienced by AI topic specialists while investigating certified information. In different applications going from quality verbalization in computational science to inspect reactions in humanistic frameworks, missing information is open to different degrees (Bertsimas, Pawlowski, and Daisy Zhuo, 2018). As different quantifiable models and AI assessments depend upon complete illuminating records, it is significant to dealing with the missing information suitably. To a great extent, major methods of reasoning may take care of business to oversee missing information. For instance, an absolute case appraisal utilizes essentially the information that is known and obstructs all experiences with missing qualities to incite irrefutable evaluation (Bertsimas, Pawlowski, and Daisy Zhuo, 2018).

These limits excellently a few observations contain missing qualities, and when the information is missing totally at irregular, complete case appraisal doesn't quick lopsided results (Bertsimas, Pawlowski and Daisy Zhuo, 2018). Then, several AI assessments normally address missing information, and there is no need for pre-dealing with it. For example, CART and K-surmises have been adjusted to issues with missing information. In different conditions, missing attributes should be credited to going before running quantifiable appraisals on irrefutably the instructive document (Bertsimas, Pawlowski, and Daisy Zhuo, 2018).

Examiners have been working on this issue for quite a while previously and have composed innovative, new, and novel approaches to manage to deal with Data Imputation. The Literature Review for this investigation fixates on approaches to manage Data Imputation after 2014. The philosophies for Data Imputation are generally quantifiable, yet more imaginative and novel techniques are appearing in the Literature i.e., significant learning, generative opposing learning, fuzzy logic, etc.

Missing information is universal in massive information clinical starter. Through different evaluations don't unequivocally report how they handle missing information, some obvious techniques are utilized in quantifiable programming. As such, various packs may shockingly oversee missing information (or the default frameworks are phenomenal) and results may not be copied definitively by utilizing arranged quantifiable programming gatherings (Zhang, 2016). Now and then, this may not incite on an extremely fundamental level various outcomes, yet the steady adequacy of the assessment is attacked. The best practice is to unequivocally state how missing attributes are managed. For ease, different specialists kill inadequate cases (Listwise prohibition), which is comparatively the default strategy in different lose the faith packs (Zhang, 2016).

This technique gets reliable outcomes precisely when the measure of missing qualities isn't enormous and the missing model is missing eccentrically (MCAR) or missing MAR. Another affront to finish case appraisal is data affliction. This can be a critical issue when there is an enormous number of (segments) (Zhang, 2016). A liberal number of cases can be erased because annulment depends on missing qualities on at any rate one component. Also, an absolute case assessment can prompt erratic propensity (Zhang, 2016). The reaction to this issue is attribution. Missing qualities are uprooted by ascribed values. Since credit is a space of dynamic evaluation, there are various methodologies and gatherings made for attribution (Zhang, 2016).

The missing characteristics are by and large evaluated using central tendency assessments like mean, center and mode in various kinds of investigation (Zhang, 2016). The mean and standard deviation are lopsided. Attributions with mode and focus work thusly and they are left to customers for planning (Zhang, 2016). However, unforgiving attribution gives fast and central systems to missing qualities, it puts down change, manages the relationship among factors, and propensities synopsis encounters. In this manner barbarous attributions ought to be utilized when a subtle bunch of qualities is missing, they are not for general use (Zhang, 2016).

Figure 2.1 shows the most popular understanding of Machine Learning paradigms from research papers (Shinde and Rajeswari, 2018).

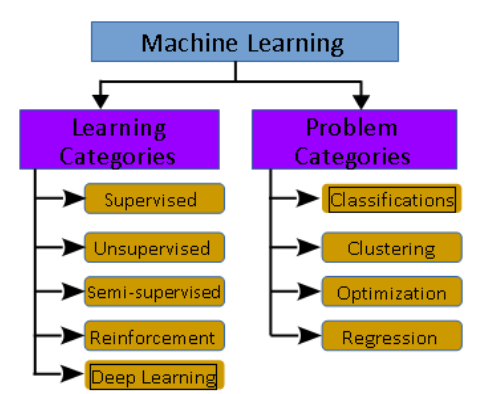


Figure 1.1 - Machine Learning Taxonomy (Shinde and Rajeswari, 2018)

**Supervised Learning:** Supervised learning techniques work on known assumptions on the marked info dataset.

**Unsupervised Learning:** These techniques focus to break down the design of information in the provided input unlabeled dataset and construct planning between the info and yield credits, while yield credits are obscure before the examination.

**Semi-supervised Learning:** These strategies utilized both named and unlabeled datasets to create models for insight surmising.

**Reinforcement Learning:** This technique's objective is to maximize the awards from the outcome. That is support learning technique creates a grouping of choices that help to secure the most noteworthy prizes.

**Deep Learning:** These techniques center on bringing together fake knowledge with AI. It chips away at regular information to give significant bits of knowledge. It deals with an input dataset that has less marked information and addresses issues grouped under semi-administered figuring out how to fabricate complex neural organization models.

AI calculations or methods are additionally characterized utilizing learning issues as Classification, Clustering, Optimization, and Regression (Shinde and Rajeswari, 2018).

**Classification:** It is a gathering method that relies upon the given estimation of the target and dataset. As per the provided target esteem, it measures and orders the dataset.

**Clustering:** It is a method that takes just the dataset as an input and recognizes fascinating examples to infer knowledge. When contrasted with the order, in the grouping objective worth isn't given as a piece of information or it is an obscure boundary.

**Regression:** In this strategy knowledge or data is gotten from the past learning experience. A condition is inferred that matches with the vast majority of the information focus and the situations where information doesn't fit with the bend, those information focuses we dispose of. This strategy is known as relapse.

**Optimization:** It is a technique to improve the exhibition of the framework as far as different ascribes.

Choy et al., (2018) Machine learning is a process of data discipline which gives the skill to computers for learning deprived of being encoded with obvious rules. The design of procedures that can absorb and make accurate and relevant guesses are enabled by Machine learning. Machine learning is taking benefit of bigger exposure to bulky and novel data sets and can progress and acquire with experiences, instead of rule-based algorithms. Fresh developments within machine learning provide potential in many fields and applications, with computed tomography. Machine learning is a course of techniques and area of research which is allowing computers to understand like hominids and making them able to outline or organize configurations. Technologies may be more reliable to further analyze data and snippet features from data that may be humans are not adequate to do.

It is known that there are types of machine learning, and these types work with multiple techniques to solve numerous scientific problems. Types of machine learning are named supervised machine learning, semi-supervised machine learning, unsupervised machine learning, and reinforcement learning.

Kushwaha et al., (2020) Nowadays, machine learning is becoming an innovative area of research with its application for analysis and prediction purposes. Machine learning is an approach to analyzing data for updating computers. Appropriate selection of actions is a must for measuring the accuracy of the system. Machine learning techniques help to understand and recognize the situation of risk. Machine learning is also playing an important role in the development of appropriate cures to control this outbreak and also in finding the root cause of this virus.

Supervised Machine learning

Lötsch et al., (2018) Supervised machine learning means when algorithms are trained to learn which is the accurate assignment from the following parameters for an accurate diagnose from scientific and laboratory data, its success is supervised as the information about the accurate detection is existing. Then, the expert algorithm will certainly complete the accurate assignment, e.g., to make the accurate detection in fresh and formerly unseen cases, if it is provided with clinical or laboratory data on which it had been trained.

Uddin et al., (2019) when using supervised learning systems, firstly a classified training data is utilized to train the basic algorithm. After training the algorithm it is applied to unclassified test data to classify them into relevant groups. Supervised learning algorithms work smoothly with regression and classification issues. In the classification process, the elementary variables for output are separate. This output variable is written off as into multiple groups or classes, for example, ‘black’ or ‘white’, or maybe ‘diabetic’ and ‘non-diabetic. Applications of Supervised Machine learning

Halgamuge**,** (2020) Classification algorithm is mostly used by practical applications for prediction and forecasting purposes. Supervised ML proceeds a well-known set of input variables, the known comebacks to the statistics or output variable, and a process that learns the mapping task or trains a model after the input to the output variables. All algorithms and labeled data are used to understand the process of predicting output from the given data.

Muhammad et al., (2020) Years before, ML has resolved numerous complicated and tough real-world issues in the application regions like marketing, business applications and retails applications, natural language processing, health care, autonomous vehicle system, Intellectual robots, image processing software, and gaming among others. ML techniques had been used for the estimate and diagnosis of numerous, illnesses such as typhoid, malaria, and diabetes. Supervised ML programs are classically based on the trial and error method which is exactly reverse to conventional programs that follow the programming order based on if or else decision.

Shinde et al., (2020) Supervised machine learning (ML) is widely used to expose new units on the way to attain support for disease. Numerous scholars are applying machine learning (ml) to get new drugs and medications for the treatment, by some computer science scholars aiming to identify infected victims through medical picture processing and CT scans. Machine learning (ml) is promoting the development of drugs and virus vaccines using expert machines. Machine learning (ml) is also introducing some tracing Software e.g., monitoring wristlets which helps to find peoples who breach the quarantine law.

Hyde et al., (2019) A Support Vector Machine is a process for supervised machine learning that turns an optimal hyperplane in a dimensional space to accurately classify the expected result by the independent variables in the dataset. A Support Vector Machine is an extreme boundary classifier that means it exploits the separation among specific classes of data efficiently in a high dimensional area. SVMs are exclusively useful when the limit between sets is non-linear as points can be transformed to a space in which the limit is linear. For this feature, SVMs are mostly used in classification issues in which the distinction between groups is non-linear.

Liu et al., (2018) many classification methods have been applied and played an important role in the medications of breast cancer. The Logistic Regression algorithm of the Sklearn machine learning library is utilized to classify the data sets of breast cancer for diagnosis. The breast cancer data is used for classification methods namely decision tree algorithm, support vector machine, and naive Bayes algorithm to build, investigate and match the execution of every model, and the performance of SVM is better. Support vector machine can efficiently adjust the contradiction among algorithm difficulty and simplification capability, and it has better simplification capability than traditional pattern recognition method in the small sample learning field. The classification outcomes show that when the two features of maximum texture and maximum perimeter are selected, the classification accurateness is 96.5%, which is upgraded and matched with the earlier methods.

Murali et al., (2020) Supervised Machine learning could be applied for quick and reliable calculations and to find the original form of data effectively. Supervised machine learning programs have been a demonstrative system in the data mining domain. Infection prediction by applying health data has recently revealed a potential application domain for these purposes.

Rustam et al., (2020) said that machine learning has verified itself as a prominent field of study over the last decade by solving many difficult and sophisticated glitches. Lots of researches are performed for the prediction of multiple diseases using machine learning methods. Using these prediction techniques is very helpful for taking decisions to handle the present situation to guide early mediations to manage these diseases very effectively.

Computer Technology Barman et al., (2014) Information technology is playing a vital part in every single field of life, specifically in the medical field. The medical diagnosis system is developing rapidly with the help of computer technology. Computer-based medical expert systems are used for the diagnosis and treatment of many diseases. Medical experts often have a different view in the diagnosis of disease because of the ambiguity in the diagnosis of diseases.

Dramsch, (2020) scientific instruments were not open access and could be difficult to operate due to limited accompanying qualification, such as MATLAB® with the Neural Networks Toolbox and Wolfram Mathematica, or autonomous university assignments like Stuttgart Neural Network Simulator (SNNS). WEKA is a graphical user interface for machine learning and data mining projects and is one of the many open-source projects available today. Shortly after, LibSVM, an open-source software (FOSS) that can implement support vector machines, was released. With SVMs, single-layer neural networks, and Random forests, as well as utility features like cross-validation, Scikit-learn implements a wide range of machine learning algorithms.

Mehta, (2020) since many Asian countries are positioned near China and share borders with it, they are more responsive to the spread of this virus. In this situation, Singapore was the second Asian country after China to record instances of patients, but it also set an example of the appropriate use of digitalization and technologies to fight the virus. They used an application called name as Trace Together, which used Bluetooth signals to trace down infected instances. In Singapore, around 70% of the population has downloaded this application, which has served the government in effectively monitoring infected persons. Likewise, Hong Kong announced the outbreak as an emergency from the beginning, citing the dreadful statistics of fatalities associated with the SARS-CoV-1 epidemic in 2002–2004. Past experiences with viruses have rendered made them more proficient at taking the necessary steps to control the virus by the use of technologies. The nation has used the Stay Home Safe application, which is paired with a smartwatch wristband, to inform the government of those diagnosed sufferers who have disrupted their quarantine time. Mongolia and South Korea, meanwhile, have been discovered to be engaged in population digital testing, including contact tracking and recording of corona-prone areas.

Kummitha, (2020) China has also upgraded its monitoring system by installing security cameras in its towns. In addition, drones embedded with cameras were used to gives instructions and alerts to anyone who won't wear face masks or did not obey safety protocols. Moreover, several hospitals in this area have engaged in the 5G Internet of Things (IoT), making it possible to create a fast and reliable infrastructure for diagnosing affected persons.

Kumar et al., (2020) Moreover, in their clinics, these countries have actively used robotics as front-line personnel for sanitizing the premises and making and delivering meals. They've also been discovered to be active in diagnosing their patients and doing radiation scanning.

Vandenberg et al., (2020) the affordability, specificity, toughness, and swiftness, being free of apparatus, and easily deliverable to end users are the key drivers headed for diagnostic readiness in problematic situations. Moreover, the medical authentication in fewer resources situations requires to launch of diagnostic abilities in the objective population with its co-endemic infections, which may be periodic or geologically disperse.

Nanshan et al.,(2020) The researcher write a review about the single-center study of numerous patient’s information collected from Jinyintan Hospital and described short term and long term exposure of epicenters, symptoms and signs, CT Findings, and clinical outcomes. The research gives an idea of improved clinical results.

Chaudhry et al. (2020) Forecasting means forecasting the upcoming situations using previous and present facts. The most used approaches for forecasting purposes are graphic techniques, qualitative methods, and time series algorithms. A time series is a set of discrete-time data points indexed in a time scale. Several programs of time-series can be used for forecasting purposes, such as moving average, single exponential smoothing, and weighted moving average. A moving average is usually applied on a technical report indicator in which a series of data points are close to analyzing the data points.

Siddiqui et al., (2019) quickly rising and the emerging field of technology is IoT (Internet of Things). Most of the intelligent devices which we use in our daily life like laundry machines blood pressure checking devices and smartwatches can transfer data over the networks. The assembling of all this information from billions of devices is introducing huge security threats to public data.

Sharma and Saxena (2017) consulted the World Health Organization (WHO) and said that 23.6 million people have suffered from cardiovascular disease. Therefore, it is necessary to address this issue. Cardiovascular diseases should be predicted timely so that heart diseases can be minimized. In this regard, the difficult and complex issue is to identify its ranges/level of illness. Medical tests of heart disease patients are used for cardiovascular disease prediction. The researchers have proposed an application of fuzzy logic and genetic algorithm for heart disease risk level prediction. In this study, data mining is used for the prediction of heart disease. The proposed system has two steps. In the first step, it develops a methodology that is based on the weighted fuzzy rules, and in the second step, fuzzy rules are based on heart disease risk level prediction by using a genetic algorithm. After that, the fuzzy framework is ready for its testing. The proposed model provides an accuracy of up to 88.11 % and can be compared with other existing models, it’s results are better. The system can handle the risk level of the patients for the available variables. The basic objective of the proposed method is to help medical practitioners who don’t have too much knowledge of cardiovascular disease in the context of technological intervention for the diagnosis.

Wagh and Paygude (2016) have defined that the Clinical Decision Support System (CDSS) is a very attractive method that helps medical experts to make a decision. The researchers have proposed a new model Clinical Decision Support System (CDSS) for Heart Disease Prediction Using Risk Factors. The proposed model has used risk factors like ECG, heart rate, diabetes, slope, hypertension, high cholesterol, or physical inactivity, etc. These risk factors are needed to be diagnosis at the proper time. Different applications of artificial intelligence are used to diagnose different diseases and helpful for medical experts. These types of applications are also useful for the patients to diagnosis their diseases. The proposed model has used the Neuro-Fuzzy model (NFS) for the prediction of cardiovascular disease. The proposed model diagnoses cardiovascular disease by using neural network adaptive capabilities and the fuzzy logic reasoning approach. The neuro-fuzzy system evaluates the available data and produces an effective and intelligent system with correctness.

Abdar et. al., (2015) elaborated that cardiovascular diseases are the main reason for death in the nation. Many techniques have been used for the prediction of heart diseases. Data mining is also used for the prediction of heart disease. The researchers have shown a comparison between different algorithms for the prediction of heart diseases in their work. The proposed work has evaluated different five algorithms i.e. C5.0, Neural network, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN), and Logistic Regression in data mining techniques to predict cardiovascular diseases. In the comparison between algorithms of data mining, the C5.0 decision tree algorithm is capable of getting effective accuracy of 93.02% when compared with other algorithms. The uniform results have shown that C5.0 is the better and medical experts can easily understand.

Meraj et al., (2021) Many interactive platforms of IoT for health care sectors are available in the market. Most of the health care sectors are using this technology for the safety of patients and their information. The IoT-based health care system encourages the patients by permitting them to communicate with their physician for medications and also helps a doctor in promoting his career. The most significant help to doctors is that they can start extra clinical work with help of IoT developments and applications.

Ye, (2020) classified and evaluated several health technologies, like cloud computing, big data, AI, and mobile health, to combat the epidemic. Certain technologies and an extensive range of data types, including omics, radiological images, omics, drug databases, data from social media, and public health firms, have been used for disease detection. Several AI methods have been used in the health care field on various scales ranging from the forecast of disease spread trajectory to the advancement of predictive and diagnostic models. These applications mobile health, telemedicine or telehealth, 5G, big data, artificial intelligence, IoT, etc, and many others, have become a dominant tool to fight against the epidemic and also give great support in epidemic restriction and control.

Affonso et al., (2017) explained the human inspection is used by a large variety of companies to conceptually categorize the efficiency of goods and the natural resources used throughout the manufacturing system, however, using image processing techniques, the whole process can be performed automatically. Companies aren't always involved in the most precise solution to a problem, but rather in the one that is best suited to the expected outcomes; therefore, a balance between performance and time cost. The categorization of wood panels depending on their images is examined in the study. It needs to compare the use of CNN to a mixture of structure-based extracting features approaches and traditional methods like Nearest Neighbors, Decision Tree Induction Algorithms, CNN, and SVM. While deploying to image processing applications, Deep Learning methods are outperformed conventional classification techniques in terms of predictive efficiency, particularly in high-complex scenarios. Deep learning methods extract relevant features from a given image dataset by directly recognizing these features.

Rathod et al., (2018) highlighted due to its complex nature, dermatology is among the most unexpected and incredibly hard areas to diagnose. Enormous examinations were often performed in the field of dermatology to decide if the patient's skin condition may be affected. The period varies from one doctor to the next. As a consequence, a system is required that can identifies skin diseases without any of these barriers. Using machine learning classification, an automated image-based method is proposed for identifying skin diseases. This framework uses computational methods to analyze, process, and relegate image data associated with different characteristics of the images. To eliminate unnecessary distortion and enhance the quality of images, skin images are processed and filtered. Features are extracted with advanced techniques including Convolutional Neural Networks, image classification using Softmax classification algorithm, and prognosis review as an outcome. The framework provides greater precision and faster results making it a reliable and accurate method of detecting dermatological infection. In addition, it is a real-time teaching tool and used to teach dermatological students.

The medical domain data is needed to be mined by mining techniques. The medical domain is a very rich data that can be helpful for research purposes, in the prediction of diseases particularly tuberculosis. The authors have presented a system for the prediction of tuberculosis by using data mining classification techniques. The proposed system has used 15 input important factors. Decision Trees, Naïve Bayes, and Neural Networks are the types of data mining techniques that are applied to the heart disease data set. The accuracy is an important part of the proposed system and results have shown that the Neural Networks are found 100% which is better accuracy than other classification i.e. Decision Tree and Naïve Bayes (30).

Singhal et al., (2018) discussed that in the medical field, decisions are taken based on stored data and the doctors' experience. There is a risk of errors, a longer time to diagnose, and a rise in the cost of treating the vital organ heart. Hospitals have a large amount of data in existing database systems that can be used to predict heart health. This data can be converted into valuable information that can be used to build an intelligent decision-making system that can predict the chance of disease occurrence. The model gives the ability to predict the occurrence of heart disease in a person. It uses convolutional neural networks to make predictions. It can predict the chances of a person developing heart disease based on important factors such as age, sex, cholesterol, and ECG slope. These characteristics are nothing more than the patients' clinical.

Abiodun et al., (2018) described that ANN is another model of machine learning (ML) the development of Artificial Neural Network is inspired by the biological neuron system to simulate the structure and base of functionally like a human brain. As per the name of Artificial Neural Network is a combination of three different words first is Artificial which is defined as it a presenting of a real objector some time called human-made and its function is close to that original object. Second is Neural, neural is an adjective of neurons this was originally taken from the brain of human in our brain there are billions of cells which called neuron and fundamentally work like biological neurons many researcher and developer are used many other alternative words like connection base network, parallel distributed processing network, etc.

Artificial Neural Network is the processing based on an algorithm that can be build complex patterns to predict the problems or provide the solution of that problem. The similarity between Artificial Neural Network and Biological Brain or Neuron System, to know about the functionality of ANN must be needed to know how BNS work because the idea of ANN techniques is originated from Biological Brain/Neuron. Many problems which we solved on ANN but Also can have solved previous or alternative methods so why we move on ANN, basically ANN is very suitable and efficient for those problem solvers who want to get significant advantages such as cost, ease of debugging/maintenance, accuracy, time and many more. There is another advantage of ANN to solve the problem by using the lookup table approach. Fundamentally lookup table is used to store all the information for gaining the appropriate result and reference of upcoming events and through lookup table approach to generalize the data and in this generalization, ANN will be trained to provide an appropriate solution to the required problem and ANN train through many inputs according to the problem which we are facing after the training section solution maybe not satisfied if the given query is not matched with training section. Another major advantage of ANN is the memory distributed for large problem or component which is used within the network.

ANN is applied in various applications such as Social media Analytics, the medical field, image processing, pattern recognition, speech recognition, radar recognition, air traffic system, power supply system, cybersecurity, and many more. In the 1800s researcher builds an idea that is possible to make a technique or model that works like the human brain like thinking, self-learning, act according to behavior, and the ability of many other functions. A simple and basic ANN was developed between the 1960s and 1970s but fall due to lack of accuracy and efficiency, after in 1986 improve the enhancement of ANN. The proposed system of this research which is used it consists of four layers: the first one is the sensor layer, the second layer is the object layer, the third phase of layer used the preprocessing layer, and the last layer is used in this proposed system is the application layer. The functionality of these mention layers is describing as sensor layers are also called the data acquisition layer and it consists of n-sensors for the purpose as input variables that are used to sense the environment and send the data to the object layer through a medium which is linked. After collecting data from sensors it will be saved in the form of raw material then data will be passed in the layer of preprocessing layer where analyze the data, mitigate the noise, normalize the data and make it in proper form for gaining the maximum accurate result because the collecting data may contain noise, irrelevant or missing information. After data arrange it will be exceeding into the next layer of the application layer. Where the application layer is divided into two main layers: the prediction layer and the performance evaluation layer.

The prediction layer is further divided into three layers of the input layer, hidden layer, and output layer and these three layers are perform on the base of Artificial Neural Network (ANN). ANN is a model whose function provides computational results with desire accuracy and efficiency. Three major qualities of ANN make it better than other computational methods and these qualities or advantages are ANN use parallel operation which faster than other computational methods those are used serial operation the reason behind this logic is that the parallelism methods has a high degree of sharing information or i/o signals because of distributed memory. Second is ANN learn from Data and store this information into the lookup table and third it uses nonlinear processing function to manage the complex nonlinear problems. ANN has very excellent properties such as self-learning, progressively in input to an output arranging, adaptively, and nonlinear problem that’s the reason ANN is mostly used for universal calculation in numerical standards.

Bitter et al. (2017) describe the importance of security not only in the pure IT department but also in those areas involved with the internet or computer. They defined that when we talk or concern with cybersecurity, we must accomplish three major qualities and these qualities are availability, confidentially and integrity and these qualities are satisfying with regulation, policies, and authorities. ANN successfully applied on a broad range of problem such as medical areas (heart, HIV aids, hepatitis, diabetes, etc.), finance (credit card, transactions, etc.), engineering (machine monitoring, robot, automobiles, etc.), Science (biology, chemistry, physics, nuclear physics, etc.).

Abiodun et al. (2018) describe the survey paper about the application of ANN which is used in real-world scenarios. These coming days ANN very popular and useful model for different disciplines like prediction, clustering, classification, pattern recognition, etc. This review study author furthermore presents the ANN application challenges, contribution, compare performance, evaluation methods.

Lek (2018) described that ANN is a powerful tool for prediction or solution of any problems and it has a very strong function and different types of algorithms or models also it shows universal and flexibility for any data. Prediction of modeling ANN is having a well-known approach use name ‘black box’ and this word black box means that all characters have an unknown situation, so these characters are identifying through training of ANN.

Potluri et al., (2017) described that Artificial Neural Network has a lot of advantages including requiring less formal statistical training to ANN for predict or finding the solution of a particular problem, ANN can point out all possible relation between predictor variables, there is a variety of algorithm and model are used in it for problem-solving and these algorithms and software/model and packages are easily available in the simple range, user-friendly, mostly graphical user interface (GUI), detect the non-linear relationship between dependent and independent variables, multiple algorithms are used to raining of an ANN, no need to know deep knowledge for developing/training of ANN just know about the basic structure of ANN model and data or parameters that can be adjusted with the model.

Dreossi et al. (2019) describe real-time data mining have the most attention on this problem. Authors have mentioned that three types of issues that face are accuracy, efficiency, and usability. The first problem is accuracy and it’s difficult to define how intrusion detects or performance or accuracy because naturally, the data mining has different definitions. Accuracy improves by analyzing the data and extracts data by using a data mining program and artificial anomalies make more improvement to detect the misuse or irregularity detection model by applying different algorithms and techniques. For improving efficiency in this paper use cost analysis and multiple models approach through this methodology proposed low cost and high efficiency they also present a distributed architecture with real-time for evaluating cost-sensitive. The third issue mentions in this paper was usability because is based on data mining is more complex compare to a traditional system and this issue improving by an adaptive learning algorithm to develop the model construction and increasing the update for this purpose used unsupervised anomaly detection algorithm. In this research paper, they discuss and present an architecture which is consisting of detectors, model generation, sensors, data warehouse, and implementing this architecture improves the efficiency and scalability of the IDSs.

Tarik & Singh (2016) describe that how data mining is improving intrusion detection in the field of cybersecurity because in present days the internet is more open and everyone is hanging on the internet for saving the information that’s why the security risks have been increased in the area of cybercrimes. In 2005 the pentagon publishes a paper which mentions that over 79000 effort to intrusion and 1300 are successful one so watching all these circumferences need to introduce new intrusion detection systems (IDSs) because older techniques are not enough for controlling the attack on cyber life. In this paper, the authors present the join of classifiers use with feature selection and multi-boosting techniques based on data mining for intrusion detection.

CNN's deep architecture has been effective in providing fine and excellent efficiency to trained models by learning patterns through raw images. A proposed deep neural network architecture is called Inception which attains the ILSVRC2014 (Image Net Large-Scale Visual Recognition Challenge 2014) assessment and the new architecture for identification. The basic goal of the design is to make better use of computing resources within the network. Through careful design, the depth and breadth of the network are increased when maintaining a computational and financial plan (Szegedy et al., 2015).

Deep learning has supported the purpose of Computer Vision in recognizing and classifying images, and it is an important tool for automating tasks in daily lives. Object identification, classification, and segmentation have been developed using convolutional networks. Because of its ability to learn to represent data, the use of the convolutional neural network (CNN) on social media images has greatly helped social media analytics (Lecun et al., 2015).

The (SML) Supervised Machine Learning is the exploration toward algorithms that cause after outwardly delivered cases to make overall assumptions, which then make calculations about upcoming cases. Most intelligent systems frequently use supervised classification. Algorithms like Linear Classifiers, (LR) Logistical Regression, Perceptron, NB Classifier, Support Vector Machine; Quadratic Classifiers, Boosting, Random Forest; and Neural networks algorithm are used where supervised machine learning deals with more classification (Osisanwo et al., 2017). Supervised Machine Learning methods need human beings to give essential inputs and outputs correspondingly, moreover, it gives estimation about the correctness of the calculation in the training procedure. Using the decision tree technique, the supervised ML algorithm is used to develop a prediction model using a labeled dataset to calculate social media analysis by the number of user responses (Muhammad et al., 2020).

The Support Vector Machines executes classification by creating an N-dimensional hyperplane that split up the data as classes. Concerning SVM, the analyst variable is called an element and the converted element is called a feature. Using the selection feature, the best and appropriate symbolic data is selected. Recounting a case in a set of features is known as a vector. Discovering the best hyperplane to split the clusters is one of the final objectives of this modeling. At the one side of the plane, there is a target variable whereas on the other side of the plane there is another group. Support vectors are the vectors that are near the hyperplane. The modules of support vector machines are nearly similar to classical multilayer perceptron neural networks (Oladipupo, 2016).

The Naïve Bayes approach is executed and the result plainly shows the high precision in many research papers (Venkatesh, Balasubramanian, and Kaliappan, 2019). The Machine learning approach (Naïve Bayes) is incorporated with sparkle climate and demonstrated to be a suitable foreseeing arrangement when managing wellbeing boundaries. The proposed BPA-NB plot involved two phases including the bunching and forecast stage. The bunching stage includes cleaning the acquired information and gathering the information dependent on the illnesses (Venkatesh, Balasubramanian, and Kaliappan, 2019). The prediction stage finds the class marks incorporate Yes\_p and No\_p utilizing the Naïve Bayes method.

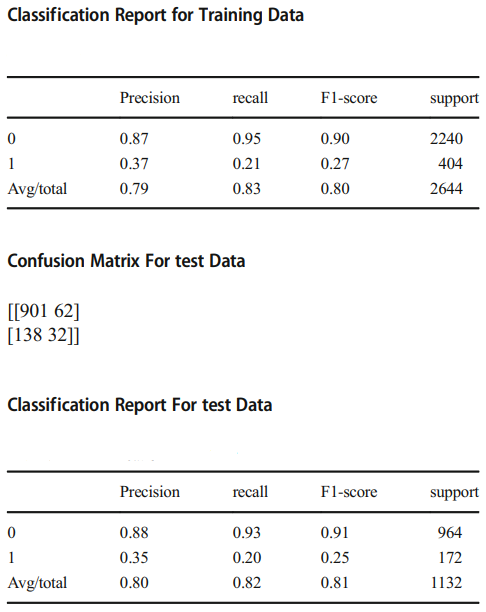


Figure 2.2: Results (Venkatesh, Balasubramanian, and Kaliappan, 2019)

The authors in this domain have tried the chance of AI models to anticipate future frequency of Alzheimer's sickness (AD) utilizing enormous scope authoritative wellbeing information (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020). From the Korean Public Health Insurance Service data set somewhere in the range of 2002 and 2010, the authors got de-recognized wellbeing information in seniors over 65 years (N = 40,736) containing 4,894 interesting clinical highlights including ICD-10 codes, drug codes, research center qualities, history of individual and family disease and socio-socioeconomics. To characterize episode AD, the authors thought about two operational definitions: "unmistakable AD" with demonstrative codes and dementia prescription (n = 614) and "plausible AD" with just finding (n = 2026).

The authors prepared and approved irregular woodland, support vector machine and strategic relapse to anticipate occurrence AD in 1, 2, 3, and 4 resulting years (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020). For foreseeing future rate of AD in adjusted examples (bootstrapping), the AI models showed sensible execution in the 1-year forecast with AUC of 0.775 and 0.759, because of "unequivocal AD" and "likely AD" results, separately; in 2-year, 0.730 and 0.693; in 3-year, 0.677 and 0.644; in 4-year, 0.725 and 0.683. The outcomes were comparative when the whole (uneven) tests were utilized. Significant clinical highlights chose in strategic relapse included hemoglobin level, age, and pee protein level. This investigation may reveal an insight into the utility of the information-driven AI model-dependent for enormous scope authoritative wellbeing information in AD hazard forecast, which may empower better choice of people in danger for AD in clinical preliminaries or early discovery in clinical settings (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020).

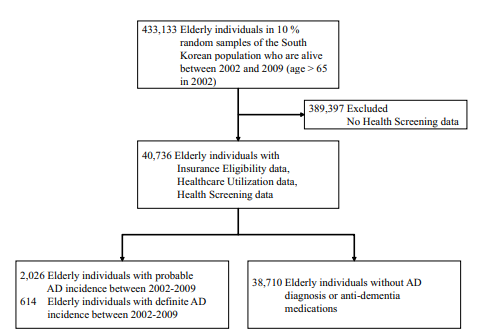


Figure 2.3: Data Statistics (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020)

Of 40,736 people with age over 65 years in 2002, the authors distinguished 614 one-of-a-kind people with AD frequency utilizing the unmistakable AD result, 2026 with AD frequency utilizing the likely Promotion definition, and 38,710 seniors with no AD frequency. The pace of AD in this companion was 1.56% utilizing the positive AD definition, and 4.97% utilizing the plausible AD definition. Demographic attributes showed huge contrasts in age between both AD gatherings and non-AD gatherings and non-significant contrasts in pay and sex.

The outcomes were comparable when the authors utilized the whole, lopsided examples for model preparing and assessment, RF showed the best execution in anticipating a 0-year rate of AD with AUC of 0.887 when utilizing the positive AD definition and AUC of 0.805 when utilizing the likely AD definition. Order execution diminished as the anticipating time frame getting longer; utilizing the unequivocal AD definition, AUC of 0.781 (1 year), 0.739 (long term), 0.686 (long term), and 0.662 (long term); utilizing the likely AD definition, AUC of 0.730 (1 year), 0.645 (long term), 0.575 (long term), and 0.602 (long term). Quantities of highlights and think-back periods likewise diminished in the latter year.

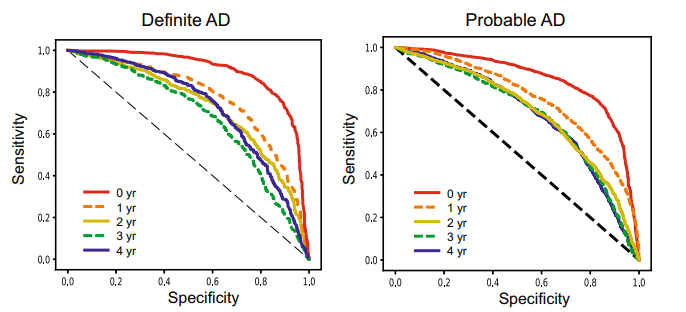


Figure 2.4: Performances for Models Trained (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020)

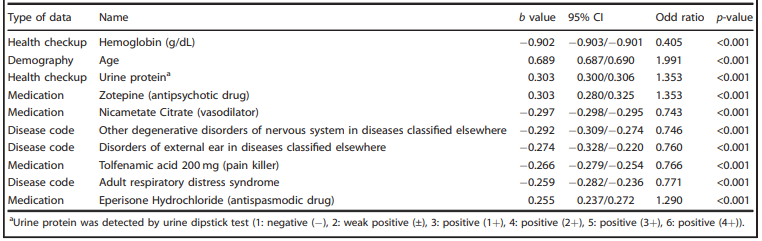


Figure 2.5: Top Performing Features (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020)

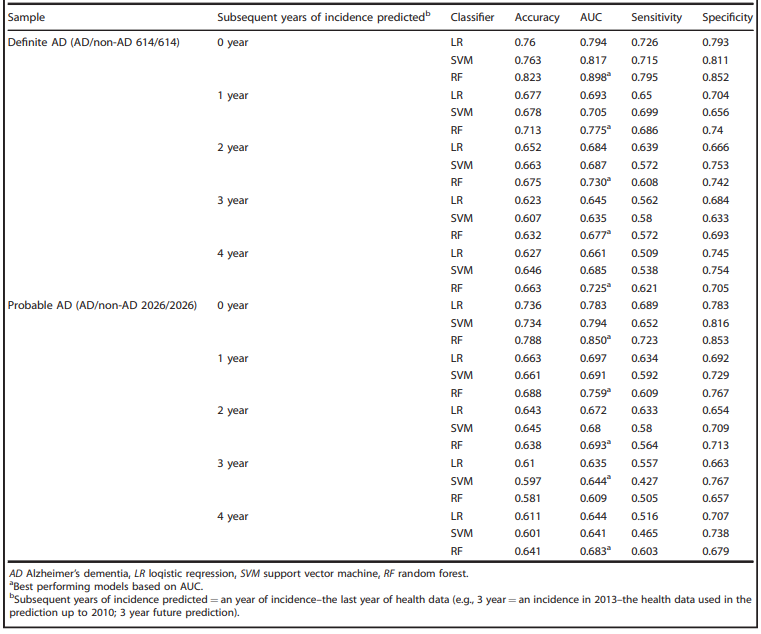


Figure 2.5 & 2.6: Performance of AD predictive models trained on NHIS-NSC by using balanced samples (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020)

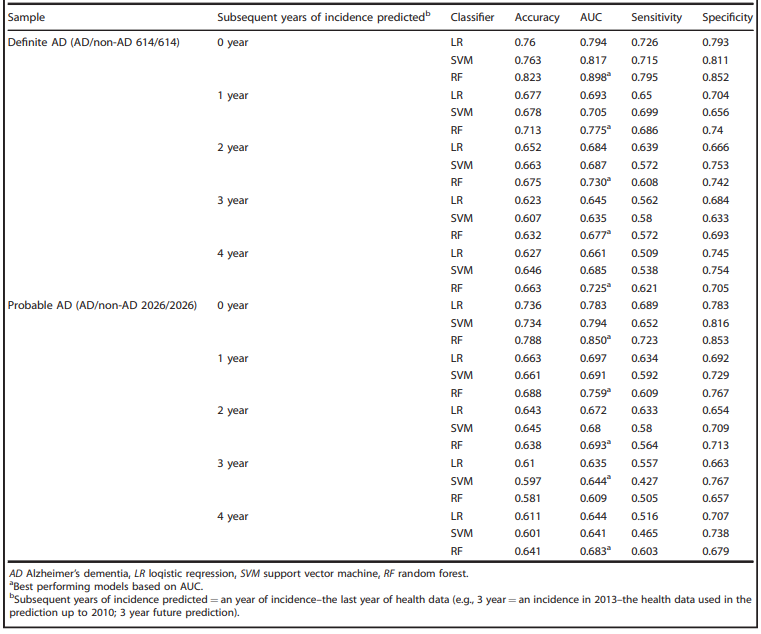


Figure 2.7: Performance of AD predictive models trained on NHIS-NSC by using balanced samples (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020).

The results from the above literature review show huge jumps and strides in implementing Machine Learning for the medical domain. The popular methods used in Machine Learning for the medical domain have shown great results and promise regarding the accuracy of their prediction. Moreover, it can be deduced that the only limitation of Machine Learning from the Literature Review is the limitation of data. As long as there exists data to feed to Machine Learning algorithms, there will be better results.

**CHAPTER 3**

1. **Project Design and Methodology**

The present study is carried out in the domain of artificial intelligence. Artificial intelligence has done remarkable contributions in almost every field by sensitizing the course of accuracy and efficiency. The main purpose of this study is to render out the limits of Machine Learning in achieving prognosis for patient health using Machine Learning. For this purpose, the researcher has gone through a thorough literature review and has identified various machine learning algorithms along with their strength and weakness. After getting acquainted with the decision support systems, the research approached the different domain specialists, clinical experts, and concerned lab technicians for the identification of relevant variables in the light of our living habits as well as the notion of the elements or variables that become a cause of disease. Even some conversations are held with patients informally to discuss the specific elements of their ailments. Such efforts have placed the researcher in a position to identify its important and general variables in the application for causing disease among the patients.

In research methodology, at the first stage, we determined the factors which can impact upon diagnosis of disease. If these factors were not relevant then we discarded them. Other relevant factors were selected for the next step. The evolution of new technologies for organizing and analyzing data with efficient computing has increased our ability to precisely predict any incident and Machine Learning (ML) can be used to make more accurate and reliable predictions, as well as to help understand the underlying pattern of the results. Machine learning includes prediction and classification of data and for this purpose, we use numerous machine learning techniques consistent with the given dataset.

There are many possible techniques and paradigms of machine used. Classification algorithms like Multi-layer Perceptron, Decision trees, SVM, or Regression algorithms like Logistic Regression. The Support Vector Machine is the best option for classification problems because it can resolve linear and non-linear complications and also working fine for numerous real-world problems. Important principals of SVM are Support Vectors, margin, and hyperplane. Support vectors are the nearest data points to a hyperplane. Whereas, a hyperplane is the kind of decision that is divided by a different set of objects and these objects are classified. The Margins are the gaps in-between the two lines on the nearest data points of classified objects. The Margins are said to be as good as large and considered bad as small. The SVM model divides the dataset into classes to figure out the extreme marginal hyperplane which can be done by creating hyperplanes iteratively which set aside the classes in a good way. After creating a hyperplane it will choose the hyperplane that splits the classes in a correct sequence.

Different Sensors from the environment taking constantly conservation data. It’s affecting a physical measure in quantity. With the board of sensors in the shape of topology, many types of data are collected. Every sensing node gets a subsection of the gathered to nearby compressing & brief since the unsystematic sign.

**CHAPTER 4**

1. **Results and Discussion**

There are many possible machine learning techniques and paradigms used in this research. Classification algorithms like Multi-layer Perceptron, Decision trees, SVM, or Regression algorithms like Logistic Regression.

* 1. **Dataset**

The dataset used is the RAPID dataset. RAPID is an abbreviation for Rapid Preliminary In-patient Data. This dataset has the basic diagnostic information of a patient admitted in a hospital ward. The features in the dataset are as follows:

* Timestamp (GMT): The date and time at which the reading was taken.
* Lifetouch Heart Rate: The real-time heart rate of the patient.
* Lifetouch Respiration Rate: The real-time respiration rate of the patient.
* Oximeter SpO2: The real-time Oxygen levels of the patient.
* Oximeter Pulse: The real-time pulse of the patient.

The source of the dataset was the student’s supervisor. The supervisor provided the dataset. Although there is limited information available on the dataset regarding the dataset.

* 1. **Machine Learning**

The algorithm used is the SARIMAX algorithm. The objective of this code is to predict separately using the **Regression** paradigm each of the features present in the dataset. To serve this purpose, the SARIMAX algorithm from statsmodels library for Python has been applied.

* 1. **Data Preprocessing**

This data has been recorded in real-time from the patients. On observation, the dataset has issues like missing values, values that don’t follow the pattern i.e. these issues have to be checked to increase the quality of the dataset and the results of the Machine Learning algorithm. Therefore, data pre-processing is mandatory for this project. Firstly, after acquiring the data, it is pre-processed and organized using SVM techniques like normalization and moving average values.

The number of features isn’t that many. Also, from observing the dataset, the paradigm of this machine learning project can be selected as regression. The dataset is time-series data providing diagnosis on the patient. This data used to predict the upcoming values for these features in advance. This system helps to predict if the patient’s health is going to decline or not. Regression algorithms were implemented for this research project. Machine Learning libraries for Python like sci-kit-learn provide many Regression algorithms like ANN Regressor, SVM Regressor, ARIMA, and many more. These algorithms are used to train and predict this dataset.

The dataset is certifiably not an enormous dataset with only four features to work with. The example line chart above is of the dataset. It shows how the four features in the dataset return to time. Every component is exceptional and fundamental for the dataset. The avocation of this is that there are as of now fewer features to work with. The solitary possible data cleaning for this dataset will be to fill the missing characteristics. That done by the settled ML outfit made in this assessment.

The code contains the following graphs that harbor the following purposes:

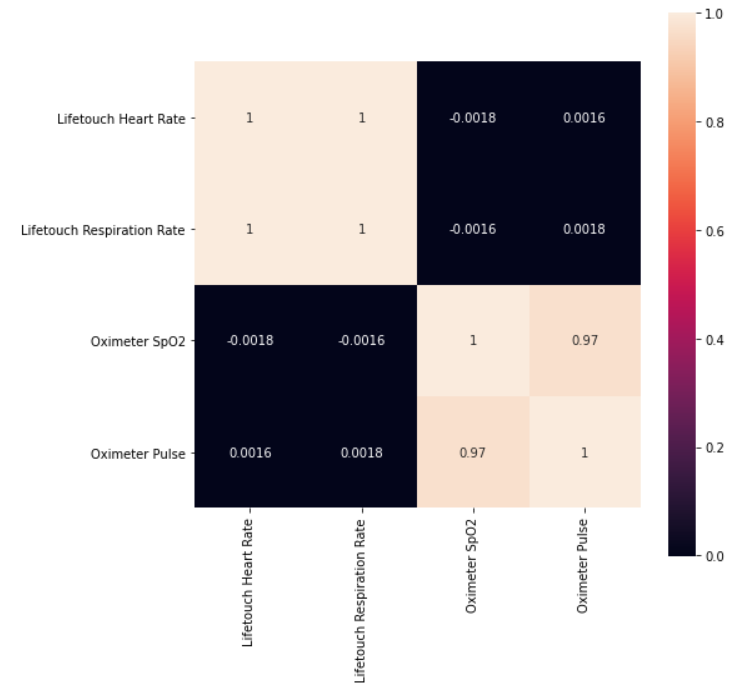


Figure 4.1: Correlation Matrix Explanation

Figure 4.1 explains the correlation between the features in the dataset by calculating the correlation value for each of the features. The higher the correlation value, the better the correlation between the features in the dataset and higher the dataset quality.

The figure 4.2 (A to D) are simply the data points plotted from the dataset to see the shape of the feature set. Whereas, a hyperplane is the kind of decision that is divided by a different set of objects and these objects are classified. The Margins are the gaps in-between the two lines on the nearest data points of classified objects. The Margins are said to be as good as large and considered bad as small.

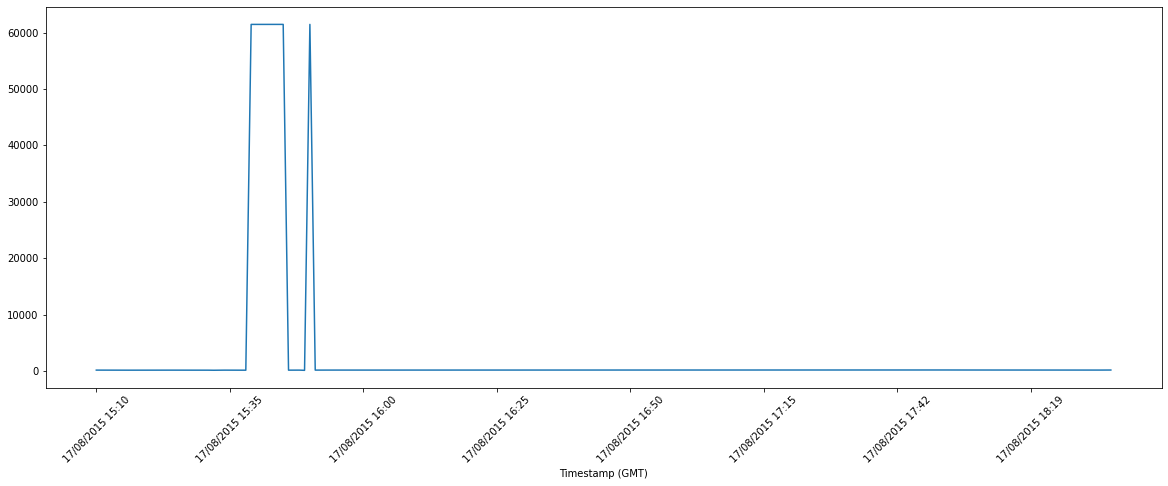


Figure 4.2 (A): Graphical representation for the “Lifetouch Heartrate” Feature of the dataset. This is the same graphical reading that an ECG will provide.

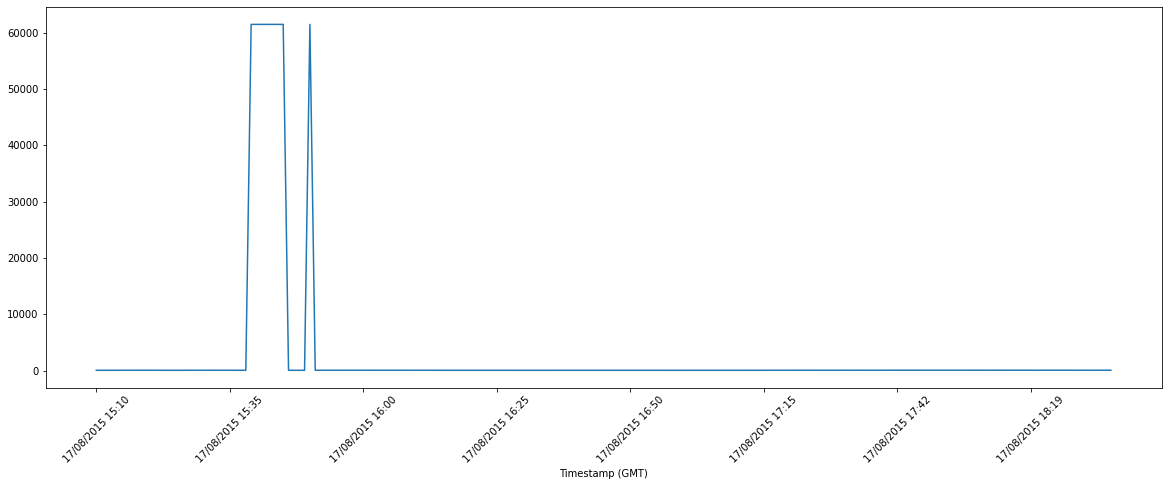


Figure 4.2 (B): Graphical representation for the “Lifetouch Respiration” Feature of the dataset. This is the same graphical reading that an Respiration measuring machine will provide.

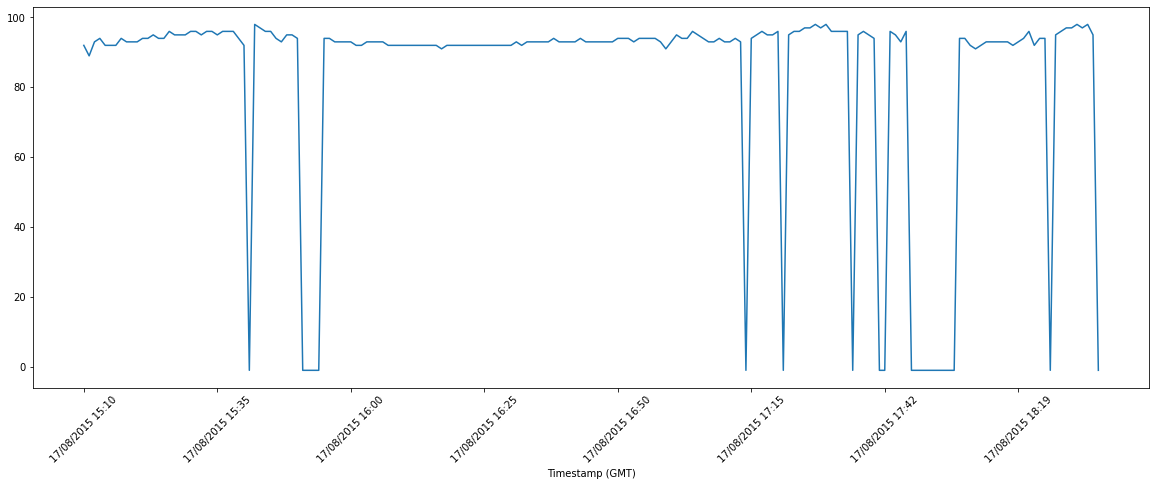


Figure 4.2 (C): Graphical representation for the “Oximeter SpO2” Feature of the dataset. This shows the oxygen consumption levels of the patient as a function of time.

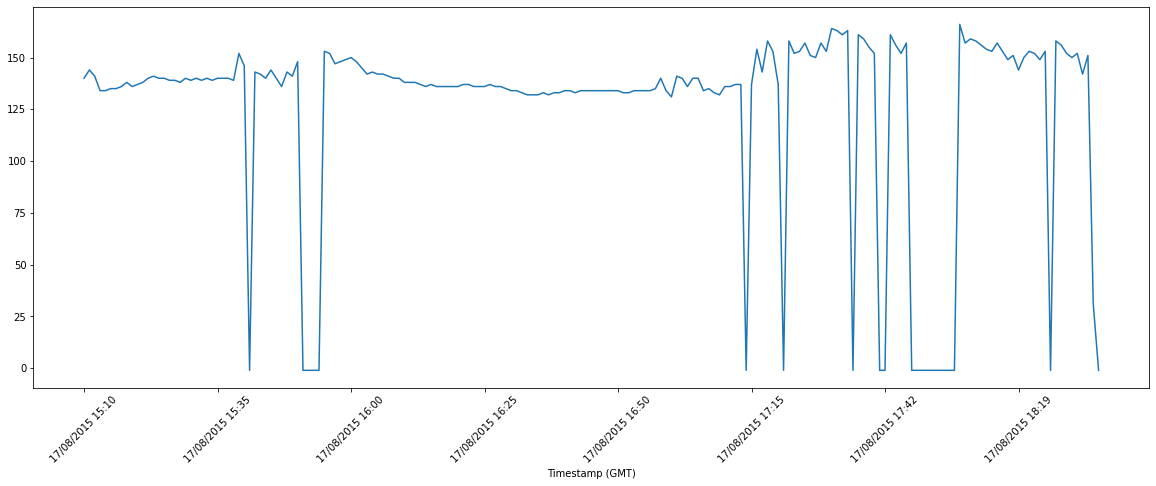


Figure 4.2 (D): Graphical representation for the “Oximeter Pulse” Feature of the dataset. This shows the pulse levels of the patient as a function of time.

As you can see, the graphs for each feature look like the output on the screens of the machines that the hospitals use to measure the vitals of an admitted patient. For the implementation of SARIMAX, there are a few conditions need to be met to perform Regression with it. These conditions are as follows:

* The data must be stationary
* The parameters of the SARIMAX algorithm including the seasonal ones must be determined
* The first step in this regard is to handle the seasonality of the feature sets. The graphs below show the process.

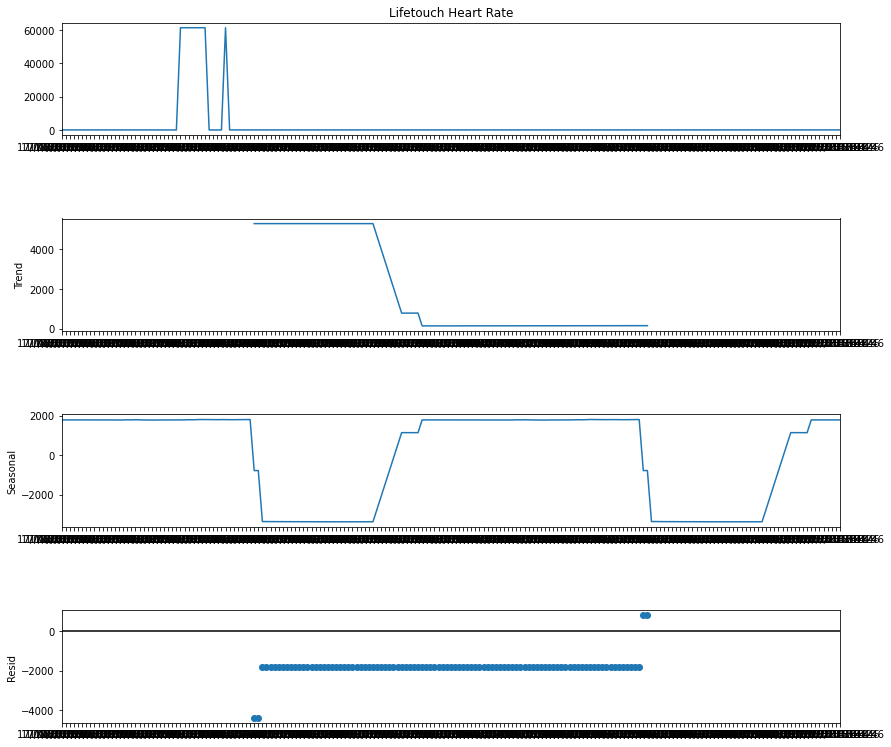


Figure 4.3: Life Touch Heart Rate seasonal decomposition. There are four graphs shown in this Figure. The first graph shows the graphical description of the Lifetouch heart rate feature. The second graph shows the Trend analysis of the Lifetouch heart rate feature. The third graph shows the Seasonality trend of the feature i.e., the repetitive short-term cycles in the feature. The fourth graph shows the residual seasonality of the Lifetouch heart rate feature i.e., the seasonal element in the feature that cannot be removed and is left as residue.

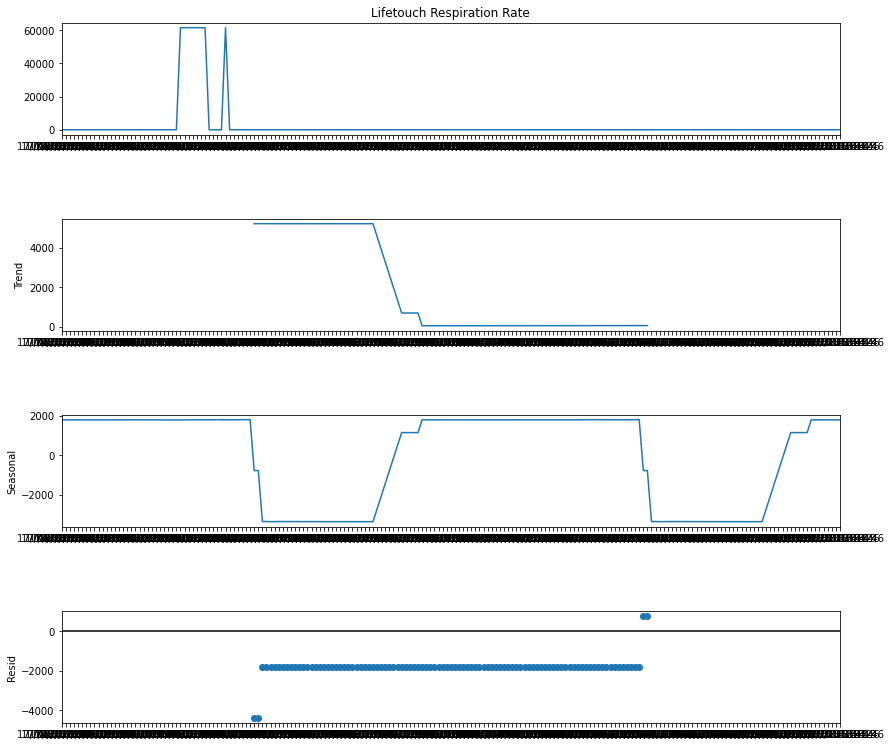


Figure 4.3: Life Touch Respiration Rate seasonal decomposition. There are four graphs shown in this Figure. The first graph shows the graphical description of the Lifetouch Respiration rate feature. The second graph shows the Trend analysis of the Lifetouch Respiration rate feature. The third graph shows the Seasonality trend of the feature i.e., the repetitive short-term cycles in the feature. The fourth graph shows the residual seasonality of the Lifetouch Respiration rate feature i.e., the seasonal element in the feature that cannot be removed and is left as residue.

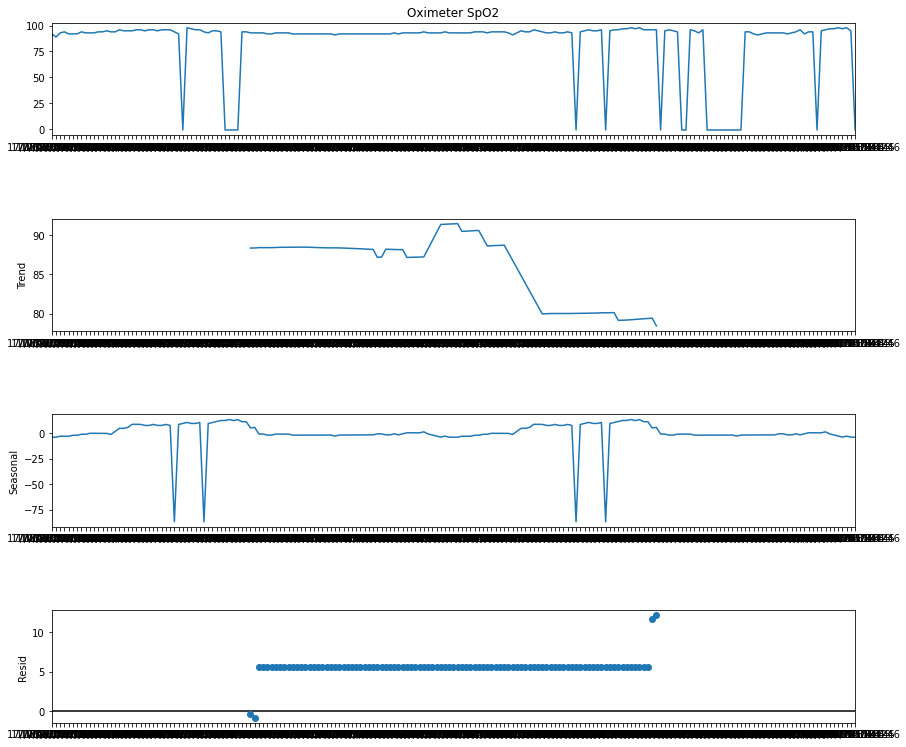


Figure 4.4: Oximeter SpO2 seasonal decomposition. There are four graphs shown in this Figure. The first graph shows the graphical description of the Oximeter SpO2 feature. The second graph shows the Trend analysis of the Oximeter SpO2 feature. The third graph shows the Seasonality trend of the feature i.e., the repetitive short-term cycles in the feature. The fourth graph shows the residual seasonality of the Oximeter SpO2 feature i.e., the seasonal element in the feature that cannot be removed and is left as residue.

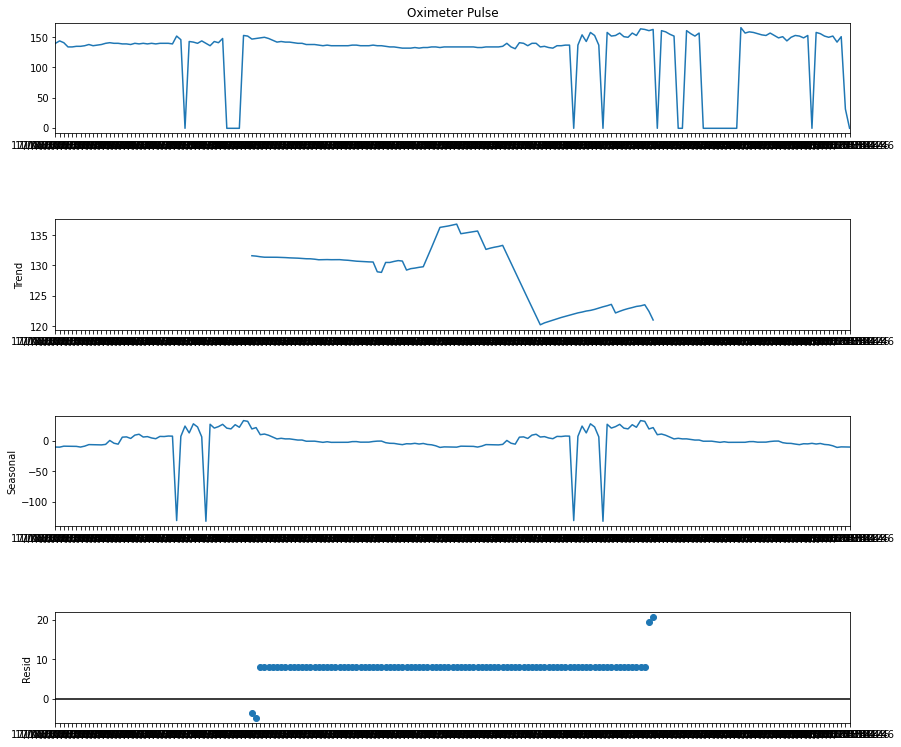


Figure 4.5: Oximeter Pulse seasonal decomposition. There are four graphs shown in this Figure. The first graph shows the graphical description of the Oximeter Pulse feature. The second graph shows the Trend analysis of the Oximeter Pulse feature. The third graph shows the Seasonality trend of the feature i.e., the repetitive short-term cycles in the feature. The fourth graph shows the residual seasonality of the Oximeter Pulse feature i.e., the seasonal element in the feature that cannot be removed and is left as residue.

The following graphs 4.3, 4.4, 4.5 are for testing the stationarity of each of the features in the dataset after applying the requisite data preprocessing on the features. These tests are performed using the Dickey-Fuller test. The graphs below associated with each feature represent graphically the stationarity of each of the features and also graphically represent the rolling mean and standard deviation readings for the feature. The p-value can be determined using these values which can then be used to verify the stationarity hypothesis for each feature.

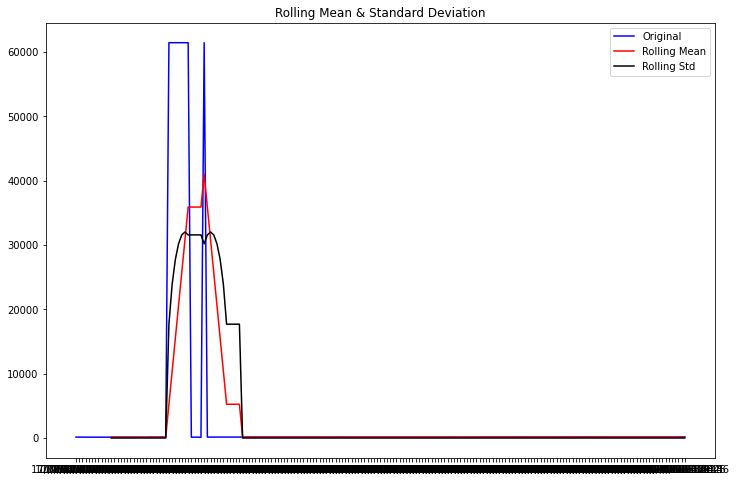


Figure 4.6 (A): Rolling Mean and Standard Deviation Graph

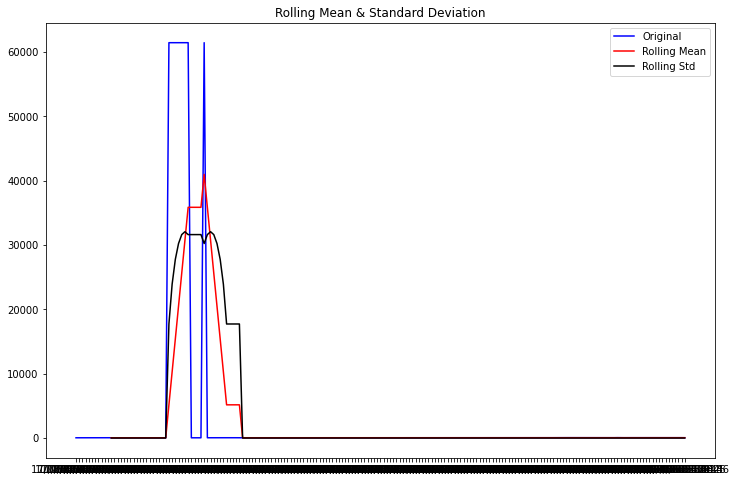


Figure 4.6 (B): Rolling Mean and Standard Deviation Graph

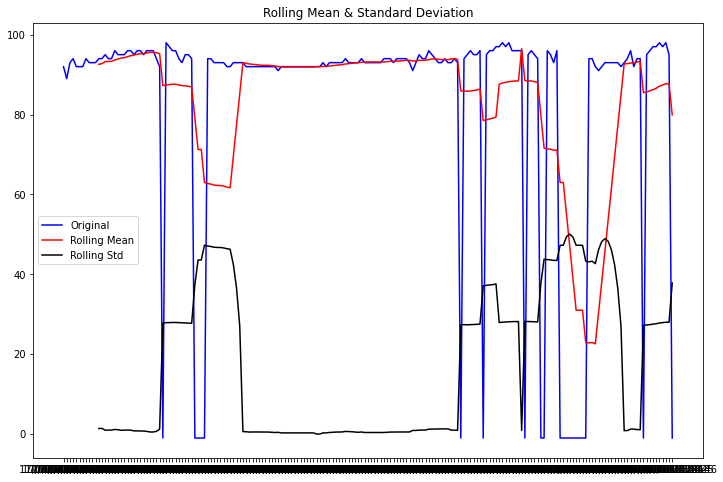


Figure 4.6(C) : Rolling Mean and Standard Deviation Graph

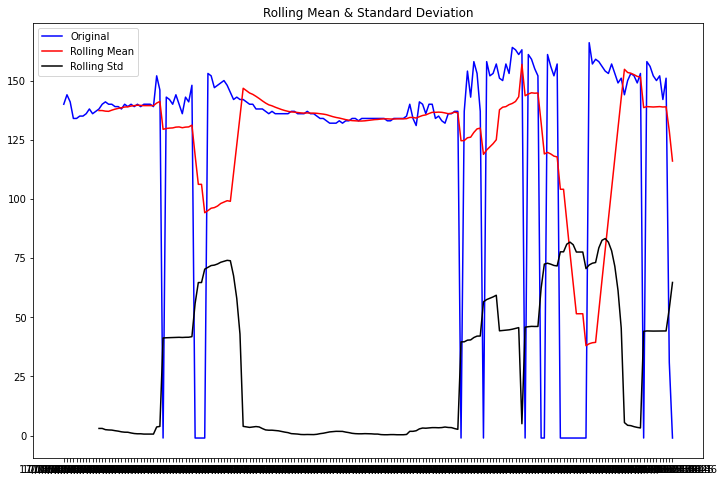


Figure 4.6(D): Rolling Mean and Standard Deviation Graph

The next step is the determination of the hyper parameters of the SARIMAX models. I would like to make clear once again that the objective of this code is to predict using Regression the future values of each of the features in the dataset separately. Therefore, any procedure done is performed four times repeatedly to accommodate the four features in the dataset.

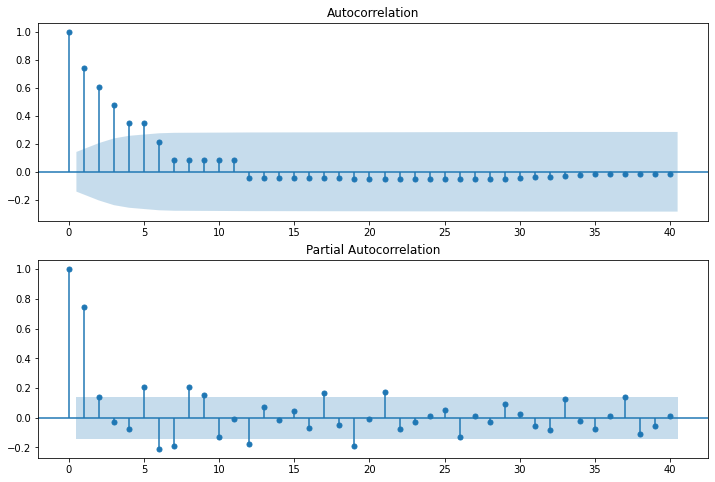


Figure 4.7 (A): ACF and PACF graphs

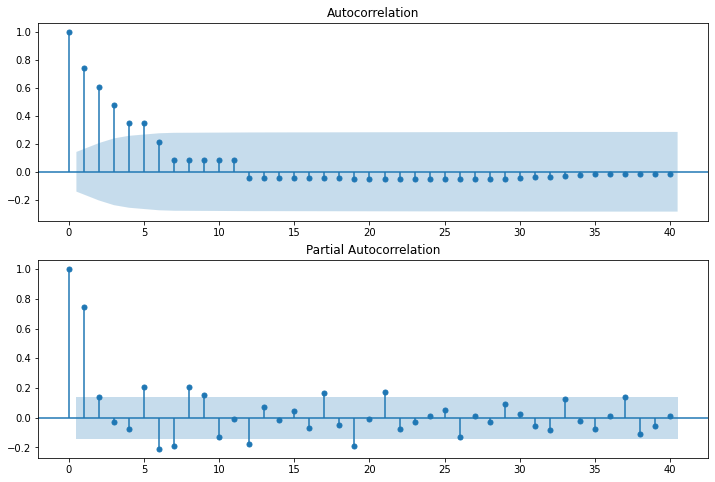


Figure 4.7 (B): ACF and PACF graphs

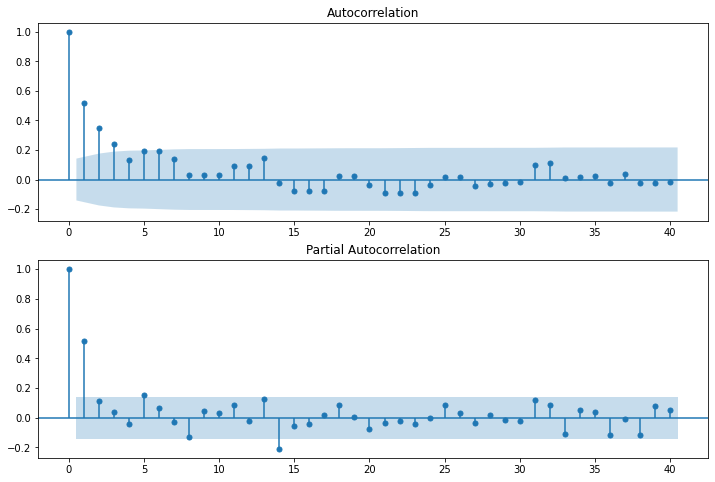


Figure 4.7 (C): ACF and PACF graphs

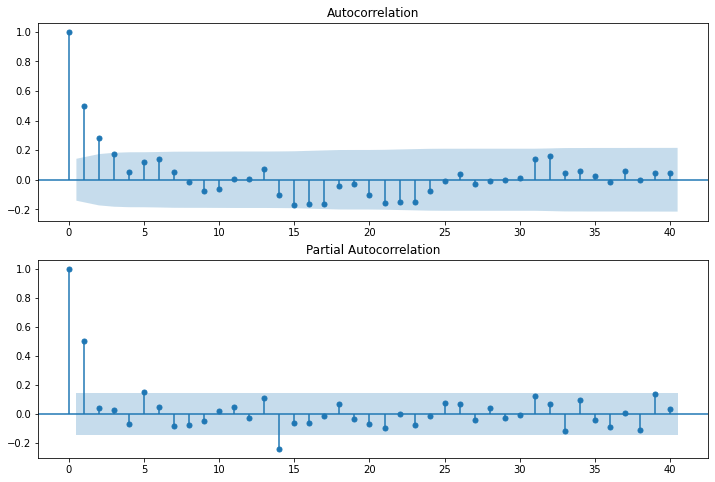


Figure 4.7 (D): ACF and PACF graphs

The data preprocessing is complete and the algorithm has been implemented. The following graphs are for the results generated for each feature in the dataset.

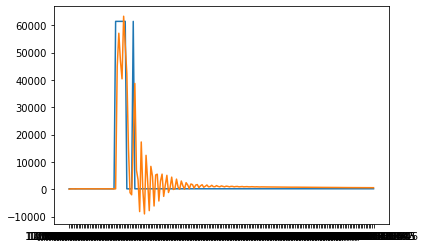


Figure 4.8 (A): Regression results for the Lifetouch heart rate feature. The observation that can be made is that the prediction cycle follows a damped oscillatory movement. The issue with this is that this dampens out to an asymptotic graph no matter what the reading from the actual information source is. This can also result in overfitting on the data. This is harmful to the prediction under a long time period.

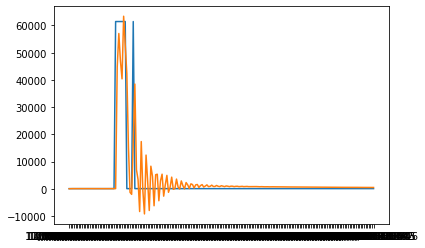


Figure 4.8 (B): Regression results for the Lifetouch respiration rate feature. The observation that can be made is that the prediction cycle follows a damped oscillatory movement. The issue with this is that this dampens out to an asymptotic graph no matter what the reading from the actual information source is. This can also result in overfitting on the data. This is harmful to the prediction under a long time period.

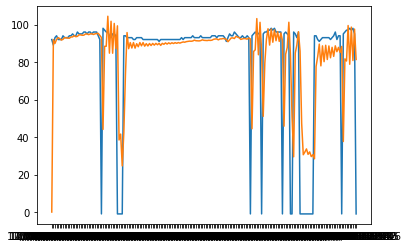


Figure 4.8 (C): Regression graph for Oximeter SpO2 pulse

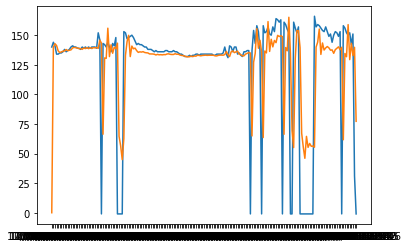


Figure 4.8 (D): Regression graph for Oximeter pulse

In the above graphs, the orange line represents the prediction from the algorithm and the blue line represents the feature set itself. The proposed research is using a classification method for the limitation of machine learning in predicting infected and not infected patients. The model is trained based on the classification method. With a trained model, predicting how to classify new data will make a decision boundary. Decision statements decide if the learning criteria meets, then place the results in big Data and if not so, then send it back to retrain. Now the validation phase will test the SVM and analyses the results. After analyzing the results performance evaluation phase is activated and evaluates the performance of the model using some parameters like the accuracy of detecting infection, miss rate. After evaluating the performance the information is updated on big data for future use.

**CHAPTER 5**

**Conclusion**

Enormous information examination assumes an urgent part to foresee future status of wellbeing and offers superior well-being results to individuals. The advent of technology has brought together medicine and machine learning. Researchers are trying to find better ways and hence develop the technology for the automation of patient prognosis. Various techniques exist but their limitations are not a well-known fact. This research provided the limits of Machine Learning in achieving prognosis for patient health using Machine Learning. There are many possible techniques and paradigms used in this research. Classification algorithms like Multi-layer Perceptron, Decision trees, SVM, or Regression algorithms like Logistic Regression.

**REFERENCES**

1. Venkatesh, R., Balasubramanian, C. and Kaliappan, M., 2019. Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique. Journal of Medical Systems, 43(8).
2. Santosh A. Shinde, M. and P. Raja Rajeswari, D., 2018. Intelligent health risk prediction systems using machine learning: a review. International Journal of Engineering & Technology, 7(3), p.1019.
3. Park, J., Cho, H., Kim, J., Wall, M., Stern, Y., Lim, H., Yoo, S., Kim, H. and Cha, J., 2020. Machine-learning prediction of the incidence of Alzheimer’s disease using large-scale administrative health data. npj Digital Medicine, 3(1).
4. Bengtsson, F. and Lindblad, K., 2021. Methods for handling missing values: A simulation study comparing imputation methods for missing values on a Poisson distributed explanatory variable. Uppsala University, [online] Available at: <https://www.diva-portal.org/smash/get/diva2:1520218/FULLTEXT01.pdf> [Accessed 5 February 2021].
5. Jadhav, A., Pramod, D. and Ramanathan, K., 2019. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence, 33(10), pp.913-933.
6. Jason Anastasopoulos, L. and G.Hunter, S., 2020. Fast and Robust Missing Data Imputation. researchgate.net, [online] Available at: <https://www.researchgate.net/publication/347741781\_Fast\_and\_Robust\_Missing\_Data\_Imputation> [Accessed 5 February 2021].
7. Zhang, Z., 2016. Missing data imputation: focusing on single imputation. Annals of Translational Medicine, [online] 4(1). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4716933/> [Accessed 5 February 2021].
8. Amiri, M. and Jensen, R., 2016. Missing data imputation using fuzzy-rough methods. Neurocomputing, 205, pp.152-164.
9. Bertsimas, D., Pawlowski, C. and Daisy Zhuo, Y., 2018. From Predictive Methods to Missing Data Imputation: An Optimization Approach. Journal of Machine Learning Research, pp.1-39.
10. Duan, Y., Lv, Y., Liu, Y. and Wang, F., 2016. An efficient realization of deep learning for traffic data imputation. Transportation Research Part C: Emerging Technologies, 72, pp.168-181.
11. Yoon, J., Jordon, J. and van der Schaar, M., 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. Proceedings of the 35th International Conference on Machine Learning.
12. Dirks, S., & Keeling, M. (2009). A vision of smarter cities: How cities can lead the way into a prosperous and sustainable future. IBM Institute for business Value, 8.
13. Albrecht, K., & Michael, K. (2013). Connected: to everyone and everything [guest editorial: special section on sensors]. IEEE Technology and Society Magazine, 32(4), 31-34.
14. Luxton, D. D., Mishkind, M. C., Crumpton, R. M., Ayers, T. D., & Mysliwiec, V. (2012). Usability and feasibility of smartphone video capabilities for telehealth care in the US military. Telemedicine and e-Health, 18(6), 409-412.
15. Anzanpour, A., Rahmani, A. M., Liljeberg, P., & Tenhunen, H. (2015, December). Internet of things enabled in-home health monitoring system using early warning score. In Proceedings of the 5th EAI international conference on wireless mobile communication and healthcare (pp. 174-177).
16. Jara, A. J., Belchi, F. J., Alcolea, A. F., Santa, J., Zamora-Izquierdo, M. A., & Gómez-Skarmeta, A. F. (2010, March). A Pharmaceutical Intelligent Information System to detect allergies and Adverse Drugs Reactions based on internet of things. In 2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops) (pp. 809-812). IEEE.
17. Alvear, O., Calafate, C. T., Cano, J. C., & Manzoni, P. (2018). Crowdsensing in smart cities: Overview, platforms, and environment sensing issues. Sensors, 18(2), 460.
18. Viani, F., Robol, F., Polo, A., Rocca, P., Oliveri, G., & Massa, A. (2013). Wireless architectures for heterogeneous sensing in smart home applications: Concepts and real implementation. *Proceedings of the IEEE*, *101*(11), 2381-2396.
19. Bisio, I., Lavagetto, F., Marchese, M., & Sciarrone, A. (2015). Smartphone-centric ambient assisted living platform for patients suffering from co-morbidities monitoring. *IEEE Communications Magazine*, *53*(1), 34-41.
20. Corno, F., & Razzak, F. (2012). Intelligent energy optimization for user intelligible goals in smart home environments. *IEEE transactions on Smart Grid*, *3*(4), 2128-2135.
21. Chen, L., Nugent, C., & Okeyo, G. (2013). An ontology-based hybrid approach to activity modeling for smart homes. *IEEE Transactions on human-machine systems*, *44*(1), 92-105.
22. Kung, H. Y., Chaisit, S., & Phuong, N. T. M. (2015). Optimization of an RFID location identification scheme based on the neural network. *International Journal of Communication Systems*, *28*(4), 625-644.
23. Perera, C., Zaslavsky, A., Christen, P., & Georgakopoulos, D. (2014). Sensing as a service model for smart cities supported by internet of things. *Transactions on emerging telecommunications technologies*, *25*(1), 81-93.
24. Gil, D., Ferrández, A., Mora-Mora, H., & Peral, J. (2016). Internet of things: A review of surveys based on context aware intelligent services. *Sensors*, *16*(7), 1069.
25. He, X., Wang, K., Huang, H., & Liu, B. (2018). QoE-driven big data architecture for smart city. *IEEE Communications Magazine*, *56*(2), 88-93.
26. Rasch, K. (2014). An unsupervised recommender system for smart homes. *Journal of Ambient Intelligence and Smart Environments*, *6*(1), 21-37.
27. Rafferty, J., Nugent, C. D., Liu, J., & Chen, L. (2017). From activity recognition to intention recognition for assisted living within smart homes. *IEEE Transactions on Human-Machine Systems*, *47*(3), 368-379.
28. Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. *International Journal of Computer Applications*, *115*(9).
29. Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, *1*(1), 4-20.
30. Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á. L., Heredia, I., ... & Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, *52*(1), 77-124.
31. Sibi, P., Jones, S. A., & Siddarth, P. (2013). Analysis of different activation functions using back propagation neural networks. *Journal of theoretical and applied information technology*, *47*(3), 1264-1268.
32. Soltani, Z., & Jafarian, A. (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. *Int J Adv Comput Sci Appl*, *7*, 89-94.
33. Zebardast, B., Ghaffari, A., & Masdari, M. (2013). A new generalized regression artificial neural networks approach for diagnosing heart disease. *International Journal of Innovation and Applied Studies*, *4*(4), 679.
34. Ahmad, G., Khan, M. A., Abbas, S., Athar, A., Khan, B. S., & Aslam, M. S. (2019). Automated diagnosis of hepatitis b using multilayer mamdani fuzzy inference system. *Journal of healthcare engineering*, *2019*.
35. Hernandez-Ramos, J. L., Pawlowski, M. P., Jara, A. J., Skarmeta, A. F., & Ladid, L. (2015). Toward a lightweight authentication and authorization framework for smart objects. *IEEE Journal on Selected Areas in Communications*, *33*(4), 690-702.
36. Xu, Q., Ren, P., Song, H., & Du, Q. (2016). Security enhancement for IoT communications exposed to eavesdroppers with uncertain locations. *IEEE Access*, *4*, 2840-2853.
37. Notra, S., Siddiqi, M., Gharakheili, H. H., Sivaraman, V., & Boreli, R. (2014, October). An experimental study of security and privacy risks with emerging household appliances. In *2014 IEEE conference on communications and network security* (pp. 79-84). IEEE.
38. Abomhara, M., & Køien, G. M. (2014, May). Security and privacy in the Internet of Things: Current status and open issues. In *2014 international conference on privacy and security in mobile systems (PRISMS)* (pp. 1-8). IEEE.
39. Lu, J., Hu, J., & Tan, Y. P. (2017). Discriminative deep metric learning for face and kinship verification. *IEEE Transactions on Image Processing*, *26*(9), 4269-4282.
40. Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert systems with applications*, *35*(1-2), 82-89.
41. Begum, S. A., & Devi, O. M. (2011). Fuzzy algorithms for pattern recognition in medical diagnosis. *Assam University Journal of Science and Technology*, *7*(2), 1-12.
42. Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR)*, *6*(2), 1-12.
43. Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.
44. Almorsy, M., Grundy, J., & Müller, I. (2016). An analysis of the cloud computing security problem. *arXiv preprint arXiv:1609.01107*.
45. Fowler, G., & Worthen, B. (2009). The internet industry is on a cloud–whatever that may mean. *The Wall Street Journal, March*, *26*.
46. Devaraj, S., & Paulraj, S. (2015). An efficient feature subset selection algorithm for classification of multidimensional dataset. *The Scientific World Journal*, *2015*.
47. Mwema, J., Kimani, S., & Kimwele, M. (2015). A study of approaches and measures aimed at securing biometric fingerprint templates in verification and identification systems. *International Journal of Computer Applications Technology and Research*, *4*(2), 108-119.
48. Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer networks*, *54*(15), 2787-2805.
49. Shelby, Z., Chakrabarti, S., Nordmark, E., & Bormann, C. (2012, November). Neighbor discovery optimization for IPv6 over low-power wireless personal area networks (6LoWPANs). In *RFC 6775*.
50. Atzori, L., Iera, A., & Morabito, G. (2011). Siot: Giving a social structure to the internet of things. *IEEE communications letters*, *15*(11), 1193-1195.
51. Abu Ghali, M. J., Mukhaimer, M. N., Abu Yousef, M. K., & Abu-Naser, S. S. (2017). Expert System for Problems of Teeth and Gums.
52. Affonso, C., Rossi, A. L. D., Vieira, F. H. A., & de Leon Ferreira, A. C. P. (2017). Deep learning for biological image classification. *Expert Systems with Applications*, *85*, 114-122.
53. Ayon, S. I., & Islam, M. (2019). Diabetes Prediction: A Deep Learning Approach. *International Journal of Information Engineering & Electronic Business*, *11*(2).
54. Mique Jr, E. L., & Palaoag, T. D. (2018, April). Rice pest and disease detection using convolutional neural network. In *Proceedings of the 2018 International Conference on Information Science and System* (pp. 147-151).
55. SİRİPONGDEE, K., PIMDEE, P., & TUNGWONGWANİCH, S. (2020). A blended learning model with IoT-based technology. *Journal For The Education Of Gifted Young Scientists*, 905-917. doi: 10.17478/jegys.698869
56. Siddiqui, F., Beley, J., Zeadally, S., & Braught, G. (2019). Secure and lightweight communicationin heterogeneous IoT environments. *Internet Of Things*, 100093. doi: 10.1016/j.iot.2019.100093.
57. Murali, N., Kucukkaya, A., Petukhova, A., Onofrey, J., & Chapiro, J. (2020). Supervised Machine Learning in Oncology: A Clinician's Guide. *Digestive Disease Interventions*, *04*(01), 073-081. doi: 10.1055/s-0040-1705097
58. Meraj, M., Singh, S. P., Johri, P., & Quasim, M. T. (2021). Detection and Prediction of Infectious Diseases Using IoT Sensors: A Review. *arXiv preprint arXiv:2101.02029*.Retrived from <https://arxiv.org/abs/2101.02029>
59. Lötsch, J., Kringel, D., & Hummel, T. (2018). Machine Learning in Human Olfactory Research. *Chemical Senses*, *44*(1), 11-22. doi: 10.1093/chemse/bjy067.
60. Allahverdi, N., Torun, S., & Saritas, I. (2007). Design of a fuzzy expert system for determination of coronary heart disease risk. In *Proceedings of the 2007 international conference on Computer systems and technologies* (pp. 36-39). ACM.
61. Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on* (pp. 108-115). IEEE
62. Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. *Infection Control & Hospital Epidemiology*, *25*(8), 690-695.
63. Avci, E. (2009). A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier. *Expert Systems with Applications*, *36*(7), 10618-10626.
64. Wu, R., Peters, W., & Morgan, M. W. (2002). The next generation of clinical decision support: linking evidence to best practice. *Journal of healthcare information management: JHIM*, *16*(4), 50-55.
65. Sharma, P., & Saxena, K. (2017). Application of fuzzy logic and genetic algorithm in heart disease risk level prediction. *International Journal of System Assurance Engineering and Management*, *8*(2), 1109-1125.
66. Wagh, R., & Paygude, S. S. (2016). CDSS for Heart Disease Prediction Using Risk Factors. *International Journal of Innovative Research in Computer and Communication Engineering*, *4*(6), 12082-12089.
67. Abdar, M., Kalhori, S. R. N., Sutikno, T., Subroto, I. M. I., & Arji, G. (2015). Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering (IJECE)*, *5*(6), 1569-1576.