A Survey on Various Tasks in Healthcare Predictive Analytics

Dr. Shameem Kappan[1], Rasheed N K[2]
[1,2]Computer Science, Sullamussalam Science College, Areekode, Kalpetta

***Abstract* - Health care data related to a patient or community is very complex. As the health data is also rapidly increasing due to the increase in population and the technology which is capable of acquiring it, it is difficult for a healthcare professional like a doctor, epidemiologist, and policy makers to make good decisions based on this data. Thus, a healthcare professional faces a challenging task in predicting health-related decisions. However, advances in health care analytics using Machine e- learning predictive models have solved many real-world problems. This is where a predictive model can be used to predict health-related outcomes by identifying patterns in healthcare data. As such models can assist healthcare professionals in good decision making which can reduce healthcare costs. For the last decade, machine learning scientists, statisticians, data scientists and healthcare professionals are collaboratively working on creating predicative models for complex diseases and tasks. This paper contains a comprehensive survey of 92 papers on various tasks involved in predictive healthcare analytics from 2012 onwards. A summary table for each task with classifier performances is also given. Hence, this paper will help new researchers who aim to enter the health care predictive analytics area.**

***Index Terms* - Epidemiologist, Data scientists, Policy makers, Healthcare analytics, Machine learning.**

1.INTRODUCTION

Health is a crucial development index of a country. Most of the countries around the world are investing a lot of money and effort towards research in improving healthcare. Countries around the world are taking the utmost care in setting up hospitals and health centers for the people Loevinsohn and Harding ((2005)). Consequently, a large number of healthcare data is generated at a high velocity Patrick and Erickson ((1993)), Jee and Kim ((2013)). The health data is very complex. Since the reason for some of the deadly diseases is not yet known, health care professionals like doctors, policymakers are relying on the newest possibilities of analyzing healthcare data. As a result doctors, statisticians, data scientists, and Machine Learning (ML) scientists perform collaborative research on how an artificial intelligence- based model can be developed for future prediction. This will help in better decision making and thus, reduces healthcare costs and improves good planning in the health sector Shanthipriya and Prabavathi ((2018)), Raghupathi and Raghupathi ((2013)).

A predictive model creates a model that predicts the future based on already existing data. Predictive analytic in healthcare make use of already existing health data collected from various sensors and devices and create a model that predicts the future by making use of knowledge from the health domain, statistics, artificial intelligence, and Machine Learning (ML). It can provide a better insight into the future for a healthcare professional. This model can help in better health care planning in an efficient manner Shanthipriya and Prabavathi ((2018)), Bates et al. ((2014)), Raghupathi and Raghupathi ((2013)).

Physicians are skilled and trained to do tough medical problems. But, a physician will face challenges mainly in two ways. Firstly, when a large number of paients consult them and Secondly, decision making for a serious desease for which the exact reasons are not yet discovered Amadi-Obi et al. ((2014)), Steyerberg ((2008)). For the first challenge, a predictive model having an ML algorithm can handle and analyze the large data and efficiently draw conclusions. For the second challenge, knowledge outside the medical domain is required. This is where statisticians, ML scientists, and data scientists collaboratively work together for assisting a physician. A predictive model is not the final word; it is good assistance to a doctor in complicated situations. None of the predictive models can replace a doctor Miller and Brown ((2017)). Healthcare professionals like public policymakers, insurance companies, and hospital administrators are also benefited from a predictive healthcare model. The most benefited among all is the patients who enjoy the quality of good healthcare system Raghupathi and Raghupathi ((2013)).

Developed countries around the globe are investing and doing a lot of research in improving the health sector. Predictions on various health-related tasks are of great importance in such countries. A predictive healthcare model benefits doctors, patients, epidemiologists, public policymakers, government, hospital administrators, pharmaceutical companies, and insurance companies in many ways. Doctors can predict the diseases as well as prescribe the next medication strategies at a patient level. Epidemiologists monitor and observe the transmission of epidemics and can reduce its flow in advance using predictive models. Even, government and public policymakers can predict the effectiveness of health policy or plan before its implementation. Hospital ad ministrators can make use of better optimization strategies and resource utilization plans for administering hospitals. Pharma companies can tackle issues related to adverse drug events and drug side effects by using predictive models. Insurance companies can use the best strategies for reducing fraud and selection of better plans by exploring patient-related data. Palanisamy and Thirunavukarasu ((2017)). Early detection of diseases and epidemics helps in planning for the worst conditions. The outcome of early detection is crucial in decreasing the progression of deadly diseases. As far as epidemics are concerned, early detection is *very* much helpful in preventing further progression to new people and communities. Similarly, for some of the deadly diseases , early detection will result in providing adequate treatment that can cure or decrease the rate of progression of disease Gaser et al. ((2013)).

The survey contains 90 papers on various tasks related to healthcare analytics. The papers are chosen by searching words and phrases in well reputed search engines like Google and journal repositories like IEEE, Springer, Elseveir, PubMed , Bio med, and ArXi v. Since, the review is on health predictive analytic alone, the focus of searching is based on keywords and phrases related to healthcare. The figure-1 contains the related words and phrases used for searching the papers in the search engines. The figure- 2 contains the number of papers used for the survey from each year (2012 onwards).

Figure 1: Word cloud related to the keywords used in t he searching of pape rs in search engines.

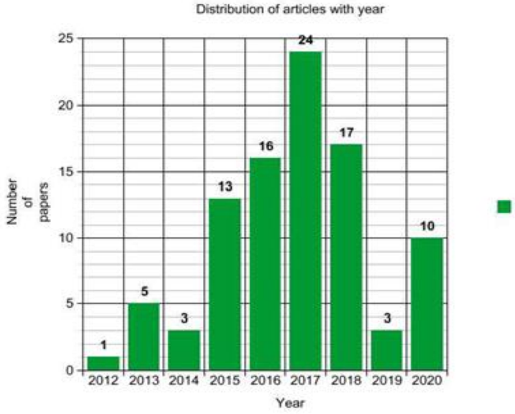

Figure 2: Distribution of papers with year.

2.BACKGROUND

This section contains the definition and related information about predictive health care models. A predictive model can be pretty useful in diagnosing complicated disease. Neither the exact reason nor treatment for some of the diseases is not yet discovered. In such situations, a predictive model can assist in judging a better accurate diagnosis Amadi-Obi et al. ((2014)). Doctors ca n take decisions on whether future treatment will be useful or not Kazd in ((2008)). Besides, doctors can also possible to reduce unnecessary hospital admissions. The behavior and response of patients to drugs and treatments can vary. To perform better treatment, patient-level information will help a doctor. Some drugs, treatment, or diagnosis can be effective for a patient or group of patients but not effective for others. This is because the human body and its functioning are complex. In such cases, a predictive model based on patient-level data can give more insight at a personal level. It will help a doctor to provide treatments and drugs specific to a patient or a group of patients Kazd in ((2008)).

The life cycle of a predictive healt h care model involves cohort selection, data acquisition, data pre processing, defining a model, deploying the model , and evaluating the model, and finally, improves the model by analyzing its performance. Cohort se lect ion involves identifying the study participants. Then, the required data of the study group is extracted from data sources followed by necessary preprocessing tasks. Finally, a predictive model is created by analyzing the data. Further, the performance of the model is checked by some experimental data until it gets ready for deployment Raghupathi and Raghupathi ((2013)). The main focus of this paper is to explain the various tasks namely survival rate prediction, medication prediction, health insurance related predictions, healthcare fraud detection, air quality prediction, hospital scheduling prediction, healthcare cost prediction, adverse drug reaction prediction, public health policy prediction, disease detection, and epidemic prediction. Hence, we conduct a survey of 90 papers from 2012 onwards and the health care prediction tasks involved in them. The figure 3 contains the various steps in healthcare predictive analytics.

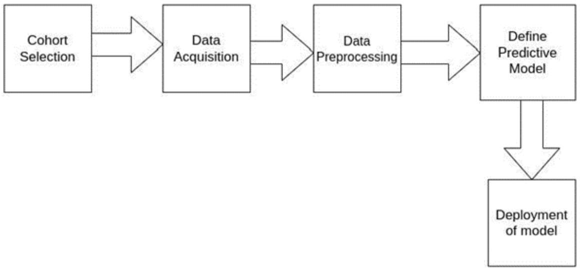

Figure 3: Predictive model as defined in Raghupathi and Raghupathi ((2013)).

2.0.1 Survival Rate Prediction

Survival rate prediction is the task of predicting whether a person or specific patient community will survive for a specific time after the treatment starts. This task is import ant in disease prognosis Xue and Chuah ((2017)), Song et al. ((2015)).

Xue and Chuah ((2017)) performed Amyotrophic Lateral Sclerosis (ALS) survival rate prediction using Electronic Health Record (EHR) data, proposed a Cox Regression model, and reported with an accuracy of 70% Xue and Chuah ((2017)). Further, they worked by adding synthetic data to the already existing crowd - sourced data reported with an accuracy of 73.4 % by using the Cox Regression model Xue a nd C huah ((2017)).

Song et al. ((2015)) experimented on survival prediction for Breast Cancer for binary Tumor subtypes with the Coxregression model. Authors predicted the prognostic values for the Tumor subtypes using combined clinical and genetic biomarkers, reported with an accuracy of 92% Song et al. ((2015)).

Taati et al. ((2014)) predicted survival rate among Bone Marrow transplant patients in Shariati Hospital at Tehran, Iran. Experimentation is done on three supervised learning approaches RF, Linea r-regression, and Support Vector Machine (SVM), reported an average Area under Curve (AUC) of 0. 9 Taati et al. ((2014)).

Roadknight et a l. ((2015)) proposed a supervised ensemble and anti- learning method for survival prediction of Tumor patients. 5- year survival rate prediction is performed to labels Will Survive or Will not Survive for Tumour , Node , Metastasis (TNM) stage 2 and 3 patients. They performed the selection of best and worst attributes using the SVM ranking method Roadknight et al. ((2015)) . ML algorithms used in the ensemble model are classification and regression trees, SVM, Logistic Regression (LR), and J48. Among these SVM and LR are applied on both best and worst features, J48 is applied on best features , and classification and a regression model is applied on TNM staging. Reported the best accuracy of 89% after combining all supervised learning algorithms Roadknight et al. ((2015)) . Better accuracy is obtained after using Convolutional Neural Network (CNN) for feature extraction for the task of predicting the survival rate of brain Tumor patients Chato and Latifi ((2017)). Chato and Latifi ((2017)) experimented with ML algorithms such as K Nearest Neighbor (KNN), SVM, and Linear Discriminant observed that the better accuracy of 91% is reported using CNN as a feature extractor and Linear Discriminant as the classifier Chato and Latifi ((2 017)). The study is conducted on the Glioma Brain Magnetic Resonance Image (GBMRI) dataset published as part of the Multimodal Brain Tumor Image Segmentation (BRATS) 2017 challenge dataset. The main classification is done into 3 categories namely short -term survivors between 0 to 10 months, mid- range 10 - 15 months, a long-range greaterthan 15 months. The prediction is done within 4 subgroups for brain tumor patients. In this paper, CNN is used as a feature extractor for brain Magnetic Resonance Image (MRI) Chato and Latifi ((2 017)). Experiments are again conducted on BRATS 2017 challenge dataset by Isensee et al. ((2 017)) using an ensemble of RF regressor and Multilayer Perceptron (MLP) reported with an accuracy of 52.6% for three tumor subtypes: the whole tumor, enhancing tumor, and core tumor using MRI Isensee et a l. ((2017)). Paul et al. ((2016)) proposed a VGG-net and CNN based transfer learning method for feature extraction and RF for classification on Computed Tomography (CT) images reported with an accuracy of 77% . The classification is done for Lung Adenocarcinoma into two classes: short survival and long survival Paul et al. ((2016)). Table 1 contains the summary of various survival rate predictions.

Article Reference	Disease	Data	Classifier	Result
Xue and Chuah ((2017))	ALS	EHR	Cox regression model	Accuracy 73%
Song et al. ((2015))	Breast Cancer	EHR	Cox proportional hazard model	Accuracy 92%
Taati et al. ((2014))	Bone Marrow Transplants	EHR	RF + SVM + LR	AUC – 0.69
Chato and Latifi ((2017))	Brain Tumor	MRI	CNN + Linear Discriminant	Accuracy 91%
Paul et al. ((2016))	Lung Cancer	CT	Transfer Learning with VGGnet and CNN	Accuracy 77%

Table 1: Accurate results on survival rate predictions for each disease.

2.0.2 Medication Prediction

Medication prediction is the task of predicting the best drug for a patient in order to cure or diagnose the disease . Certain drugs have side effects which affect the patient badly. In such cases medication prediction models can be used to choose the most optimal and good medication strategy Wright et al. ((2015)).

Some works are conducted on next prescribed medication prediction Wright et al. ((2015)). Next prescribed medication prediction is the task of predication the very next medication based on already existing medication data. Wright et al. ((2015)) performed this task at drug class and level using a Sequential Pattern Mining (SPM) algorithm with clinical, profile, and laboratory data reported with accuracies of 90% at drug class and 64% at drug level. They performed an integration of clinical, profile and laboratory data for predicting next medication Wright et al. ((2015)). Further, Song et al. ((2020)) implemented a modified Recurrent Neural Network (RNN) for predicting next prescribed medication using EHR data of MIMIC 3 dataset for Peneumonia, Mitra Valve Disorder, Prostate Cancer also reported with accuracies of 72%, 76%, and 67% respectively.

Optimal Drug Prediction is another task in which the better drugs for a patient is chosen on basis of clinical and other health data. Sheng et al. ((2015)) experimented on half maximal inhibitory concentration (IC50) data of 75 drugs on 624 cell lines from Genomics of Drug Sensitivity in Cancer (GDSC) and 24 drugs on 504 cell lines from Cancer Cell Line Encyclopedia (CCLE). Experiment is done on 50 types of cancers and 75 types of drugs. They classified a patient as either sensitive or resistant to a specific type of drug. The proposed algorithm is based on a similarity based Tanimoto Coefficient Sheng et al. ((2015)), where a drug cell line treat effect is calculated on the basis of drug, cell line simiarity and IC50 values implemented in R language. Further, the drugs with P-values < 0.05 is chosen and reported that model is capable of choosing first 10 optimal drugs. They would like to integrate the work with other data for an effective future research work Sheng et al. ((2015)). Table 2 contains the summary of medication predictions. Table 3 contains the tools used for drug name and structural compound extraction.

Article	Task	Disease	Level	Classifier	Results
Wright et al. ((2015))	Next medication prediction	NS	Drug class Drug level	SPM	Accuracy 90%
Song et al. ((2020))	Next medication prediction	Peneumonia, Mitra Valve Disorder, Prostate Cancer	Drug class	RNN	Accuracy 64%
Sheng et al. ((2015))	Optimal drug prediction	NS	Drug level	Similiarly based approach	10 best drugs

Table 2: Summary of medication prediction. NS- Not Specified.

Tools	Task
G DSC, CCLE	Extracting gene expression data of cell lines , and the I C50 data of drugs on the cell lines

Table 3: Tools used for drug name and structural compound extraction

2.0.3Health Insurance Related Predictions

Health insurance is an insurance that covers the whole or a part of the risk of a person incurring medical expenses, spreading the risk over a large number of persons, by estimating the overall risk of healthcare for availing some healthcare benefits Boodhun and Jayabalan ((2018)). The health insurance claims contain the information about patient's disease and insurance holder chooses the best plans Boodhun and Jayabalan ((2018)). The information about the health insurance helps to predict many tasks about the patients. Following are some of the health insurance related predictions.

Boodhun and Jayabalan ((2018)) proposed health insurance applicant's risk prediction scores using ensemble tree based approach. Authors experimented on multiple Linea r-regression , ANN , REPTree , and RF and reported that REPTree has obtained lowest Mean Absolute Error (MAE) of 1.5285 using corelation feature selection method Boodhun and Jayabalan ((2018)). In another work, Siswantining et a l. ((2018)) predicted the risk of hospitalization using o ut patient health insurance claims data. The study is conducted on six diseases coronary diseases as explained in Siswantining et al. ((2018)). Experiment is conducted on six variables like member ID , provider , total incurred , frequency, age , gender, claim type and proposed LR . Reported accuracies for 6 categories using LR reported with accuracies o f 81%, 85%, 80%, 77%, 95% and 93% respectively Siswantining et al. ((2018)). Researchers also conducted experiments for studying the patient's migration behaviour using the health insurance claim's data Cheng et a l. ((2020)). They conducted the experiments on Chinese hospital data for 21 diseases over 44 hositals and proposed a Recurrent Gated Unit (RGTJ) for predicting the migration behaviour with an accuracy of 80% Cheng et al. ((2020)). Table 4 contains the summary of various health insurance related predictions.

Article Reference	Task	Disease	Classifier	Result
Boodhun and Jayapalan ((2018))	Risk prediction	NS	LR + ANN + REPTree+ RF	MAE – 1.5285
Siswantining et al. ((2018))	Risk of hospitalization	Coronary artery	LR	Average accuracy – 86%
Cheng et al. ((2020))	Migration behaviour	NS	RGU	Accuracy – 80%

Table 4: Summary of health insurance related prediction. NS- Not specified.

2.0.4HealthCare Fraud Detection

Healthcare fraud is the one who deliberately gives misleading or false in formation in health related programs inorder to get the beneficaries that do not deserved for his or her. A health fraud can be a doctor, patient, insurance provider or any one who is related with a healthcare program Cui et al. ((2016)).

Graph based algorithms are used for healthcare fraud detection Cui et al. ((2016)). Cui et al. ((2016)) proposed Graph Mining Frequent Pattern (GMFP) approach from a medical insurance data set. Authors predicted three parameters for fraud detection: behavior patterns growth, local outlier fit, and frequent pattern outlier factor with a F l -score of 0.65, 0.50, 0.72 for Coronary artery, 0.62, 0.6, 0.69 for Hyper tension, and 0.75 , 0.76 and 0.79 for Diabetes respectively. In another study using Medicare part A, B, C, D dataset, Herland et al. ((2018)) proposed LR to predict health frauds reported an accuracy of 81%. They worked with a combined Medicare part A, B and C data and fraud labels are created by mapping to List of Excluded Individuals and Entities (LE IE) released by the officer of inspector general Herland et al. ((2018)). Further, Matloob et al. ((2020)) proposed a sequential mining algorithm for predicting the health care frauds using longitudinal transactional data with an accuracy of 85% Matloob et al. ((2020)). Table 5 contains the summary of healthcare fraud detection.

Article Reference	Disease	Classifier	Result
Cui et al. ((2016))	Coronary artery hypertension diabetes	GMFP	F1 – Score 0.65, 0.50, 0.72, 0.62, 0.6, 0.69, 0.75, 0.76, 0.79
Herlad et al. ((2018))	NS	LR	Accuracy – 81%
Matloob et al. ((2020))	NS	Sequential mining	Accuracy – 81%

Table 5: Results in Healthcare Fraud detection. NS – Not specified

2.0.5Air Quality Detection

Air quality detection is the task of detecting the concentration of air pollutants such as particulate matter, carbon monoxide, sulfurdioxide, and nitrous oxide present i n the atmosphere. The concentration of the pollutants at a higher level leads to air pollution. In this case, during air pollution nowcasting, the name for forecasting near future become relevant Jiang et al. ((2015)). Nowcasting the air quality parameters of a city is a challenging task. This requires analysis of air quality data, which may not be a strong indicator for prediction. With the increasing popularity of social media , users engage and express their issues in social media. Due to which researchers are using social media data to predict air pollution Jiang et al. ((2015)). Jiang et al. ((2015)) found out high association between Air Quality Index Data published by China's ministry of Environmental Protection and Tweets posted in Seina Weibo by analysing with Pearson's Correlation. Consequently, they predicted air quality index data using Gradient Boosting Treess (GBT) reported with a R-squared coefficient of 0 .57 for individual messages posted by the users which is higher than retweet and mobile app messages. Authors expressed their concern over switching of users from Seina Weibo to Wechat in China and urged the importance of choosing popular social media sites for this task Jiang et al. ((2015)). In another related work, the authorised government data is itself used for air quality detection at frequent time intervals. Nowcasting the air quality pollutants like Sulphur di oxide, Particle matter, and Ozone at hourly basis using meteorological and air pollutant dat a collected by weather station, and air quality system data of United System of America (USA) by Zhu et al. ((2018)). They proposed supervised transfer learning algorithm using RNN by which the information in one hour is used to predict the parameters of next hour using various optimization algorithms reported with a Root Mean Squared Error (RMSE) of 0.03 Zhu et al.((2018)) . The 8 hours advanced prediction for concen tration of ozone was proposed by the researchers using hourly air monitoring data Freeman et al. ((2018)). They reported a RMSE of 2.5 using the proposed RNN algorithm. Table 6 contains the summary of air quality predictions.

Article Reference	Data	Interval	Region	Classifier	Performance
Jiang et al. ((2015))	Scina Wcibo	Daily basis	Beijing	Pearson correlation	R-Squared – 0.57
Zhu et al. ((2018))	Government	Hourly basis	2 Villages 1 airport, 1 university Zhu et al. ((2015))	Transfer learning RNN	RMSE – 0.03
Freeman et al. ((2018))	Government	8 Hour bassis	City	RNN	RMSE-2. 5

Table 6: Distribution of articles on data, interval and region for nowcasting air quality

2.0.6Hospital Scheduling Prediction

Hospital scheduling helps in managing staffs, nurses, resources and other facilities of a hospital in an appropriate manner Hendri and Sulaiman ((2017)).

A sub task of hospital scheduling is the Length of Stay (LOS) prediction which is defined as the number of days a patient admitted to the day of discharge. Hendri and Sulaiman ((2017)) proposed a regression model for predicting LOC of Dengue patients into three categories : >4 days, 4-20 days, and <30 from a Malaysian hospital. They predicted 80% of the patients have > 4 days and rest have stay beyond 30 days Hendri and Sulaiman ((2017)). Ensemble methods are also used by researchers Kumar and Anjomshoa ((2018)), Xie et al. ((2016)), Maharlou et al.((2018)). Researchers Kumar and Anjomshoa ((2018)), proposed a Classification And Regression Tree (CART) model for predicting LOC for surgical patients from Elective Surgery Information System (ESIS), Victoria Admitted Episode Dataset (VAED), and theatre dataset in USA . They performed a two stage classification followed by a regression model for predicting length of stays for general, Urology, Orthopedic, and Gynaecology departments. They reported an average LOS of 0.97 days for 3805 patients, 13 days for 15 patients and 0.062 days for 2527 patients. Experiments at a departmental level predictions showed that the error rates are low for Gynaecology department. Further, prediction model for whole departments yielded more error compared to each department level prediction Kumar and Anjomshoa ((2018)) . In a different approach using health insurance claims data , Xie et al. ((2016)) proposed a Bagged Decision Tree (BDT) for labelling into two categories no hospitalization and at least one day in hospital. Predictions are performed for yearly, half year, quarter and two months durations. They have designed a temporal dataset of insurance claims data comprising of patients demographics, hospital admission, and insurance procedure data. Authors reported a Mathews Correlation Coefficient of 0.426 for yearly model Xie et al. ((2016)). However, authors expressed their concern over the models inability to predict length of stays beyond 250 days and they are planning to increase the size of dataset Mazhar et al. ((2017)). A supervised approach is used for predicting Intensive Care Unit (ICU) LOS after Cardiac surgery by Maharlou et al. ((2018)). They used MLP and CART reported with an R value of 0.607 using MLP and 0.88 using CART Maharlou et al. ((2018)). Statistical based approaches using Logarithmic Transformation (LT) is employed for patients having Spinal Cord injury reported with an accuracy of 81% by Mazhar et al. ((2017)). Table 7 contains the summary of various LOS predictions.

Article Reference	Disease	Time span	Classifier	Result
Hendri and Sulaiman ((2017))	Dengue	4- 20 days	LR	80% < 4
Kumar and Anjomshoa ((2018))	Surgical patients	1-13 days	CART	Mostly 1day
Xie et al. ((2016))	NS	2 months to 1 year	BDT	MCC-0.426
Maharlou et al. ((2018))	Cardiac surgery	1-2 weeks	MLP+CART	R value 0.607
Mazhar et.al((2017))	Spinal Cord	NS	LT	Accuracy 81 %

Table 7: Summary of articles on LOS. NS-Not Specified.

Intensive Care Unit (ICU) predictions are another task in hospital scheduling. ICU is one such place where the critical and serious patients are admitted in a hospital. Predictions at ICU level helps doctors and other health care stakeholders in decision making Rouzbahman et al. ((2017)). Two types of ICU predictions are mentioned here mortality and LOS. Supervised approach is used by Rouzbahman et al. ((2017)) experimented on both ICU death and LOS prediction for Cancer patients in the ICU of the hospitals on MIMIC 2 dataset Rouzbahman et al. ((2017)). They performed experiments using LR and reported an accuracy of 72% and 75% respectively. Futher, they reported a MAE of 7474 hours for LOS using Linear- regression Rouzbahman et al. ((2017)). Further, Meadows et al. ((2018)) proposed a LR model for predicting LOS in ICTJ's , reported with an accuracy of 79.7 %. They found out the ICU related LOS for cardiac patients Meadows et al. ((2018)). Table 8 contains the summary of various ICU tasks.

Article Reference	Task	Disease	Data	Classifier	Result
Rouzbahman et al. ((2017))	LOS	NS	MIMIC 2	LH	MAE – 7474 hours
Meadows et al. ((2018))	LOS	Cardiac Surgery	Private Hospital	LH	Accuracy – 79.7%

Table 8: Summary of various ICU tasks. NS-Not Specified.

Daily discharge prediction is a type of hospital scheduling prediction. Predictive models are also created for forecasting the number of daily discharged patients in hospitals Zhu et al. ((2017)). Experiments are conducted on three forecasting modesl Seasonal Autoregressive Moving Average (SARIMA), Multiplicative Seasonal ARIMA (MS ARIMA) and Weighted Markov Chain (WMC) Model as explained in Zhu et al. ((2017)) on Women's and Childrens Hospital data in Australia. A combined model of Weighted Markov Chain and MS ARIMA model outperformed other two models Zhu et al. ((2017)) reported with a R-Squared value of 0.93. However, the authors pointed out the following drawbacks in their work,

1. It is required to consider multiple days ahead data for prediction rather than considering a single day data Zhu et al. ((2017)). In another study, Cho et al. ((2017)) predicted the discharge status of a Stroke patient into one of the 16 categories as mentioned in Cho et al. ((2017)) at the time of admission in a hospital They experimented on both supervised learning using a LR model and unsupervised learning using Stacked Auto Encoders (SAE). The reported accuracies for LR, SAE with one encoder layer and two encoded layers are 68%, 68% and 69% respectively Cho et al. ((2017)). McCoy et al. ((2018)) proposed SARIMA model for daily discharge prediction on England hospital reported with a R-squared value of 0.843 McCoy et a l. ((2018)). Table 9 contains the summary of daily discharge predictions.

Article Reference	Data	Disease	Classifier	Result
Zhu et al. ((2017))	HER	NS	WMC + MS ARIMA	R – Squared 0.93
Cho et al. ((2017))	HER	Stroke	LR, SAE	Accuracy 68%
McCoy et al. ((2018))	EHR	NS	SARIMA	R – Squared 0.843

Table 9: Summary of reults in daily discharge predictions.

Hospital Emergency Department (ED) Visit prediction is the task related to the prediction of cases which are related to ED of a hospital. As the ED of a hospital is the first entrance for any patient, better scheduling tasks needs is required . Some of ED related prediction tasks are ED visits, retriage, and overflow predictions. ED visit prediction is the task of finding the number of patients who are likely to visit ED Ram et al. ((2015)). Ram et al. ((2015)) classified Asthma visits into three categories namely high, low, and medieum. They performed experiments on DT, ANN and concluded that a combined model of DT and ANN for predicting Asthma related for a hospital in USA has given a precision of 72%, 71% and 75% for high, low, and medium respectively. Further, authors like to extend the work to more diseases with a regional variability and temporal dataset Ram et al. ((2015)). ED re- triage is the process of accessing patients on the basis of their severity Rahmat et al. ((2013)). Graham et al. ((2018)) experimented this task using three predictive models namely Ir, DT and GBT. Out of all the classifiers, GBT performed well with an accuracy of 80%, Reciever Operator Characteristic Curve (ROC) of 0.824. The cohort is selected from a hospital in Northern Ireland. Following features like triage category, visit time, arrival mode, week, month, care group, admission history are extracted Graham et al. ((2018)). Rahmat et al. ((2013)) conducted a study using historical EHR data of a hospital in Malaysia for ED re- triage. Authors classified patients into three levels of severity namely red, yellow, and green. Authors proposed an agent based emergency department re- triage scheduling model Rahmat et al. ((2013)), where there are nine agent roles and six simulation factors as in Rahmat et al. (2013)). Experiment is conducted on examining how waiting time is reduced for patients in each label and various deteriooration patterns as explained Rahmat et al. ((2013)). Model acheived a significant reduction in waiting time for Green to Yellow deterioration pattern patients by 7.25% of time and on holidays waiting time is reduced by 9.51% for all patients Rahmat et al. ((2013)). Table 10 contains the summary of various ED predictions.

Article Reference	Task	Disease	Classification level	Classifier	Result
Ram et al.((2015))	ED visit	Asthrna	3 way	DT + ANN	Precision – 72% high, 71% low, 72% Medium
Rahmat et al.((2013))	ED triage	NS	3 way	GBT	Accuracy 81%
Rahmat et al.((2013))	ED triage	NS	NS	Agen Model	Waiting reduction 9.51%

Table 10: Summary of resul ts in daily diacharge predictions. NS- Not Specified.

Hospital read mission rate prediction is the task of predicting whether a lready admitted patient will be readmitted again in a hospital Shameer et al. ((2017)). Shameer et al. ((2017)) proposed a Naive Bayes (NB) classifier for the prediction of 30 days hospital readmission rate for Heart failure patients at Mount Sinaihospital in USA. Authors report ed an accuracy of 83% which is better than RF, Adaboost. However, genomic data of patients are not used in this study. Researchers Jamei et a l. ((2017)) also implemented a Neural Network model in Kern's framework using Google's Tensor flow library for 30 days hospital readmission rates which is not specific to a disease by collecting data from 15 hospital located in Sutter. They have created the model by including variety of features that comes under many categories such as hospital encounters, hospital problems, procedures, medications, discharge, socio economic, admission, lab results, co - morbidities, demographics, family history, paying history. A total of 1667 features are extracted. They reported AUC of 0.78 in classifying patients into either admitted or non admitted categories. Further, the training time is also reduced significantly using the model. Authors worked on real time hospital data and felt the need for working on social determinants of health data as the future work Jamei et al. ((2017)). In a different approach, Craig et al.((2017)) experimented on doctor's prescription notes collected at Sarosata Memorial Hospital, USA during 2004 to 2014 for classifying planned and unplanned hospital read missions, reported better performance for CNN implemented with Python's Kerns framework rnnning Tensor Flow with a C-statistic of 0.70. The drawback of the study is that out of the total hospital data 40% belongs to doctor's prescription notes Craig et a l. ((2017)). Jiang et al. ((2018)) proposed a Particle Swarm Optimization (PSO) and DNN model for hospital readmission rate prediction in Chinese hospital using a large EHR reported accuracy, AUC of 84% and 89% respectively Jiang et al. ((2018)). Koola et al. ((2020)) proposed a 30 day hospital read mission rate prediction using LR Koola et al. ((2020)). They used Cirrhosis patients EHR, reported with an AUC of 0.67 Koo la et al. ((2020)).

Desikan et al. ((2012)) experimented on early prediction of preventable events in ambulatory care sensitive admissions using clinical data from USA hospitals. A total of 2000 features are extracted for each patient as explained in Desikan et al. ((2012)) and trained with a rule based classifier as in Desikan et al.((2012)), reported with a sensitivity of 0.3and precision for the Preventable Events is 0.887. They made a successful attempt for prediction of possibly preventable events using EHR data Desikan et al. ((2012)). Table 11 conatins the summary of various hospital readmission rates.

Article Reference	Task	Data	Disease	Classifier	Result
Shameer et al.((2017))	Readmission	EHR	Heart failure	NB	Accuracy - 83%
Jamei et al. ((2017))	Readmission	8 HR	NS	DNN	ACC- 0.78
Craig et a l. ((2017))	Readmission	Prescription notes	NS	DNN	C-statistic 0.70
Jiang et al. ((2018))	Readmission	EH R	NS	DNN	Accuracy 84% AUC0.89
Koola et al. ((2020))	Readmission	EHR	Cirrhosis	LR	AUC 0.67
Desikan et al. ((2012))	Preventable events in ambulatory care sensitive admissions	EHR	NS	Rule based classifier	Precision 0.87

Table 11: Summary of results from on each task in hospital readmission rates. NS - Not Specified.

2.0.7Health Care Cost Prediction

As the medical costs is increasing rapidly, it is imperative to predict the health care costs in advance for better health planning Sushmita et al. ((2015)). There are three sub tasks related to health care costs prediction namely high-cost patients prediction, medical price prediction, and patients whose expenditure likely to increase prediction are done.

High cost patients are those patients who are likely to use the expensive hospital facilities and resources in future. Identifying high cost patients will aid in taking good decisions and helps in optimizing the health care costs efficiently McWilliams and Schwartz ((2017)). Supervised regression models are proposed for predicting high cost patients Boscardin et al. ((2015)), Seng et al. ((2016)). Boscardin et al. ((2015)) they proposed a multivariable logistic regression model. The model is created using demographics , emergency department visits , hospital admissionsi, depressions, chronic pain and diabetes information of each patients achieved a C-statistic 0.65. The novelty of the paper is that predictive model is developed using a self reported health data Boscardin et al. ((2015)). In a similar work, T. Seng et al. ((2016)) proposed a multiple Linear- regression to predict the high cost patie nts among Type 2 Diabetes Mellitus at National University Hospital, Singapore using demographics, utilization of resources, length of stay, number of admissions/ readmissions and many other features as explained in Seng et al.((2016)) reported an AUC of 0.708 Seng et al. ((2016)). Graph based data is also experimented with RF for predicting high cost patients by Srini vasan et al. ((2017)). The proposed model creates a disease co-occurence network for a patient's disease data. Then, a community detection algorithm is applied to the network to find various communities of high cost patients. Further, they found out top most diseases affecting different communities. The model is evauated by baseline and network reported a high sensitivity for random forest classifier 84% for base line, 80% for baseline and baseline network data respectively Srinivasan et al. ((2017)).

Researchers also experimented using publicly available Medicare Payment dataset Gittelman et al. ((2015)), Guo et al. ((2015)) for medical price predictions. Medical price prediction is the task of predicting the overall medical expenses in future. The market price of various health instruments, drugs changes from time to time. In such a scenario, this prediction helps hospital administrators plan smartly Tike ((2018)). A. Tike and S. Tavarageri Gittelman et al. ((2015)) predicted medical prices specifically average total payments using Medicare payment dataset. They proposed a hierarchical decision tree model reported with a good accuracy for GBT with error percentages of 13, 10 and 10 at 3, 8, and 12 depths Gittelman et al. ((2015)). Researchers also studied health care expenditure increase for a patient, bywhich they predict the patients whose medical expenses are likely to increase incoming years.

Prediction of healthcare expenditure increase for an individual from 2008 to 2010 is studied by Guo et al. ((2015)) using Medicare Payment dataset. Features used for the study are demographic, prescription drugs and medical conditions patients undergo because the variations in those factors can change health care expenditure. They proposed a ensemble of six classification algorithms: GBT, SVM, NN and conditional inference trees (CIF) with stacked generalization to the algorithms reported with an accuracy of 77% in classifying a patient's expenditure will increase or not Guo et al. ((2015)). Table 12 contains the summary of various health care costs predictions.

Article Reference	Disease	Classifier	Results
Boscardin et al. ((2015))	NS	LR	C-Statistic 0.65
Seng et al. ((2016))	Diabetes	RF	AUC – 84%
Srinivasan et al. ((2017))	NS	RF	Accuracy – 84%

Table 12: Distribution of articles on healthcare cost predictions. NS- Not Specified.

2.0.8 Adverse Drug Reaction Prediction

Adverse Drug Reaction prediction is the task of predicting whether the usage of a drug or certain mix of drugs results in adverse events to a patient. Adverse events refer to the occurrence of abnormal reactions in body; say a person may allergic towards a disease which leads to a red skin rash. Some adverse events can even lead to death. As a result, pharmaceutical companies take utmost care in designing drugs for a disease and patient. Early prediction of adverse drng events can assist them in designing better safety drugs[1].

Social media texts in biogs and online health forums are used for ADR event detection Liu and Chen ((2015)), Sampathkumar et al. ((2014)), Yang et al. ((2013)), A semi supervised approach has shown good result of using social media text data in classifying Adverse Drug Reactions (ADR). Liuand Chen ((2015)) predicted adverse drug event for Diabetes patients from a social media forum known as American Diabetes Online Community (ADOC) in USA, whose data can be accessed by researchers. They used already available tools like FAERS, UMLS, and CHV Li u and Chen ((2015)) to extract drug names and adverse event names. They proposed transductive SVM, a semi supervised approach implemented using SVMlight for identifying medical events, patient report extraction, ADR reported with a F-measure of 86%, 84% and 69% respectively Liu and Chen ((2015)). In another approach using social media texts, Sampathkumar et al. ((2014)) proposed a Hidden Markov Model (HMM) for the detection of ADR mentions from the publicly available manually annotated medications[2] database. The appearance of certain keywords and phrases as mentioned in Sampathkumar et al. ((2014)) are observed, whether that comes after or before the mentions of side affects and drugs Sampathkumar et al. ((2014)). Moreover, authors created a dictionary of drug names containing in SIDER database for checking whether the mentions of those drugs are in the discussions of medication's forum. They reported an Fl -measure of 0.76 Sampathkumar et al. ((2014)). Supervised learning approaches are also applied on ADR detection using web forum texts which are extracted through a web crawler Yang et al. ((2013)). A domain expert is utilized to set the ground truth to two labels positive ADR and negative ADR. The syntactic, linguistic and semantic features are extracted from the texts as explained in Yang et al. ((2013)). They proposed a SVM model experimented on two health related web forums namely ProzacAwaraness and SSRISrx of Yahoo reported with an accuracy of 88% for ProzacAwaraness and 89% for SSRISrx Yang et al. ((2013)). A supervised learning approach is proposed by Jamal et al. ((2017)) for predicting whether a person will have neurological ADR for a drug using phenotypic, biological and chemical properties of drugs obtained from PubChem, SIDER, and DrugBank data which are publicly available. PADEL software is used to collect chemical sub structure information of the drugs. They proposed a sequential minimization a lgorithm (SMO) implemented as a black box in Weka reported with an average accuracy of 93% for ADR events Jamal et al. ((2017)). The model is validated by Anti Alzheimer's Drugs dataset that is publicly available. Much better accuracy is reported after using a drug knowledge graph based algorithm. The graph is created using the available data of drugs, indications, and possible ADR Bean et al. ((2017)). The model is validated using South London and Maudsley NHS Foundation Trust reported with an AUC of 0.92% which out performed other supervised approaches Bean et al. ((2017)). Moreover, Dai and Wang ((2019)) reported an average precision of 64% using minority over sampling technique on Twitter (AMIA SMM4H) imbalanced data. Further, Chapman et al. ((2019)) proposed a Natural Language Processing (NLP) model consists of Conditional Random Field (CRF) and RF reported with a Fl score of 61.2%. They used the standard dataset of Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE 1.0) competition Chapman et al. ((2019)). Table 13 contains the summary of various ADR events.

Article Reference	Disease of ADR	Data	Classifier	Result
Liu and Chen ((2015))	Diabetes	ADOC post	SVM	F1-measure 0.69%
Sampathkumar et al. ((2014))	NS	Medication post	HMM	F1-measure 0.76
Yang et al. ((2013)).	Health, wellbeing, sexual disorders	Web forum post	SVM	Accuracy 89%
Jamal et al. ((2017))	Neurological	PubChem, SIDER, DrugBank	SMO	Accuracy 93%
Bean et al. ((2017))	NS	NS	Graph Algorithm	AUC 0.92
Dai and Wang ((2019))	NS	Twitter posts	SMOTE	Precision 64%
Chapman et al. ((2019))	NS	MADE 1.0	CRF+RF	F1-score61%

Table 13: Distribution of ADR articles on basis of type of disease ADR mentioned and data used. NS- Not Specified.

1. https://en.wikipedia.org/wiki/Adverse_drug_reaction
2. www.medications.com

2.0.9Public Health Policy Prediction

Advanced prediction of public health policies are important for good decision making at a government level. Infact, predictive models that analyze longitudinal data of people who used government programs can be utilized for these Metcalf et al. ((2015)).

Alshurafa et al. ((2017)) implemented a model for predicting the success of Cardiovascular monitoring system for individual young black women using sensors that measure certain physical variables. In addition, data of the women during baseline, 3 and 6 months are also collected which include questionnaires, family history, percieved social support scale and soon as explained in Alshurafa et al. ((2017)). They reported a F-measure of 83% using RF Alshurafa et al. ((2017)). Further, Chandir et al. ((2018)) predicted the childrens who are likely to be frequent utilizers of an immunization program for Measles 2 in advance which is a binary classification problem Yes or No. Authors proposed a Recursive Partioning (RP) which is an ensemble tree based model reported with an accuracy of 79%.

Further, the prediction is done for the validation set after analyzing the longitudinal data at the baseline itself Chandir et al. ((2018)).Table 14 contains the summary of various health policy predictions.

Dataset	Disease	Classifier	Result
Alshurafa et al.((2017))	Cardiovascular	RF	Accuracy – 83%
Metcalf et al. ((2015)).	Measles 2	RP	Accuracy – 79%

TaTable 14: Results for health policy prediction.

2.0.10 Disease Detection

There are two types of disease detections namely infectious and other diseases which are not caused due to infectious agents. In factious diseases are transmitted from one person to another person or living being through an agent. Non infectious diseases are not caused by an infectious agent and caused due to many other reasons[3]. Hence, it will not transmit from one body to another. Rather, the disease will develop inside the body of an individual and progress slowly[4] • in this section, the review will discuss various infectious and non-infectious diseases detected.

Stroke: Ensemble methods reported good performance in detection of Stroke related diseases from brain MRI and EHR Ojha and Mathur ((2016)), Chen et al. ((2017)). Chen et al. ((2017)) worked on both structured and unstructured data of Chinese hospital data for identifying patients who are in risk for Cerebral Infarction in future. They worked on NB, DT, and KNN models for structured data. They used CNN for unstructured data. After combining all the classifiers, they obtained an overall accuracy of 94.8%. The novelty of the paper is that it has also used unstructured data for training the model Chen et al. ((2017)). Further,Ojha and Mathur ((2016)) proposed ANN for the detection of Isochemic Stroke using MRI collected from Maharaja MRI and PET centre and PBM hospital at Bikaner in India. They collected MRI data of patients and extracted grey level co-occurence matrix features for each image and fed into an ANN reported with an accuracy of 98% Ojha and Mathur ((2016)). CNN is used for brain lesion segmentation tasks which are crucial for Stroke detection Kamnitsas et al. ((2017)). Kamnitsas et al.((2017)) proposed a 3D CNN and CRF for the detection of Brain lesions in (ISLE 2015) challenge reported with a Dice score of 63 on brain MRI images. 3-D CNN is implemented using Pylearnlibrary Kamnitsas et al.((2017)). In other work, Menze et al. ((2016)) proposed a Probabilistic Generative model for Brainlesion segmentation on MRI images for three types of tumor namely edema, enhancing, and non enhancing Tumor. They acquired an Dice score of 0.80 for segmenting lesion in edema and 0.50 to 0.60 for enhancing tumor Menze et al. ((2016)). An MRI image for the work is collected from ISLE 2015 challenge data set. The model is evaluated using BRATS dataset Menze et al. ((2016)).

Parkinson: Following are the works related to the detection of Parkinson using (PPMI) dataset. Experiments are conducted by Dinov et al. ((2016)) for classifying Parkinsons and Healthy Controls, Healthy controls and Parkinsons. Adaboost classifier reported with best accuracy, sensitivity and specificity of 98%, 97% and 99% respectively for classifying Parkinsons and healthy controls, 98%, 97% and 98% for classifying healthy controls and Parkinsons with scans without evidence for dopaminergic deficit Dinov et al. ((2016)). Supervised MLalgorithms are proposed by Amoroso et al. ((2018)), Martinez- Murcia et al. ((2013)). Amoroso et al. ((2018)) proposed a SVM algorithm for binary classification of Parkinsons and Normal Controls. They reported with an accuracy, sensitivity and specificity of 93%, 92% and 92% respectively using network measures obtained from MRI and clinical scores Amoroso et al. ((2018)). Again binary classification of Parkisons into PD or Without SWEDD is experimented by Martinez-Murcia et al. ((2013)) using Haralick texture features which are extracted from the grey level co-occurence matrix corresponding to the SPECT images of a patient. They proposed SVM reported with an accuracy and sensitivity of 95% and 97% respectively Martinez - Murcia et al. ((2013)). Further, Hirschauer et al. ((20 15)) classified Parkinson's patients into either SWEDD or Without SWEDD using motor, non-motor and neuro imaging features reported accuracy of 92% using probabilistic neural network and 98% for binary classification of Parkinsons Hirschauer et al. ((2015)). Further, Alhussein ((2017)) proposed SVM for predicting Parkinson's patients in Saudi Arabia with an accuracy of 97.2%.

Alzheimer's: Deep learning based CNN proved highly capable in Alzheimers detection Islam and Zhang ((2017)), Billones et al. ((2016)), Sarraf and Tofighi ((2016)).Islam and Zhang ((2017)) performed an experiment on classification of patients into mild, moderate, very mild, and non demented categories on common dataset Open Access Series of Imaging Studies (OASIS). Ensembles of three CNN are reported with an accuracy of 93.8% Islam and Zhang ((2017)). Tree ensemble based classifier is used for the multi class classification of Alzheimers disease by Wehenkel et al. ((2017)) on Positron Emission Tomography (PET) images and OASIS dataset MRI images using group selection procedure is implemented on random forest and extremely randomized trees for choosing best set of features implemented in Matlab with 10 fold cross validation for tuning parameters. The proposed model reported an accuracy of 81% and 82% for random forest and extremely randomized trees respectively Wehenkel et al. ((2017)). Following classification tasks on Alzheimer's are done in Alzheimer's Disease Neurolmaging Initiative Database (ADNI). Billones et al. ((2016)) proposed a CNN model named DemNet which is based on VGG Net for 3 way classification of patients into healthy controls, Mild Cognitive Impairment, and healthy patients. DemNet has achieved an accuracy of 91.85% for 3 way classification. Authors highlighted that 17 coronal slices from the middle part of brain helped in distinguishing three classes. The proposed algorithm outperformed already existing CNN architectures Billones et al. ((2016)). Researcher also experimented using unsupervised deep learning algorithms on medical images Liu et al. ((2015)). Liu et al. ((2015)) experimented an Stacked Auto Encoder (SAE) based deep learning model on MRI and PET and they proposed algorithm has obtained an accuracy of 90% for classifying a patient into either Alzheimers or healthy controls which outperformed already existing

SVM. For binary classification of MCI versus normals classification, an accuracy of 82% is reported. They also experimented on multi class classification Alzheimers versus MCI versus Non MCI's versus healthy controls using SAE reported with an accuracy of 53%. They also implemented a novel visualization method to show high level biomarkers Liu et al. ((2015)). Using a similar approach of SAE. Sarraf and Tofighi ((2016)) experimented on already existing CNN architecture LeNet-5 for binary classification of Alzheimer's using ADNI reported an accuracy of 96%. In another approach, Thung et al. ((2017)) proposed a multi task deep learning algorithm implemented in Keras framework of python for multi class classification of Alzheimers into MCI, Alzheimers and healthy subject categories using both MRI and PET images reported an accuracy of 65% Thung et al. ((2017)). Authors also worked with structured OASIS Alzheimer's data Khan and Zubair ((2020)), Battineni et al. ((2019)). Khan and Zubair ((2020)) proposed a RF classifier for classifying demented and non demented patients, reported with sensitivity and specificity of 80%. Further, G. Battineni proposed SVM for dementia classification on OASIS dataset with accuracy and precision of 69% and 64% respectively Battineni et al. ((2019)).

Heart Disease: Wang et al. ((2015)) proposed RF clasiifier for early prediction of heart disease using structured and unstructured data from EHR. The highest accuracy is obtained after combining structured and unstructured data by observing 720 days prediction with an accuracy of 83%. However, authors considered the prediction of early detection of heart failure subtype as their future work and observed that increase in prediction window is directly proportional to accuracy Wang et al. ((2015)). Buchan et al. ((2017)) classified Coronary artery and non-coronary artery disease patients using narrative medical histories text data. Study was conducted on Heart Disease Risk Factors Challenge dataset and they used an APACHE based natural language processing tool for feature extraction from the text data Buchan et al. ((2017)) and experiment is conducted on NB, SVM, and maximum entropy classifiers. Among all SVM obtained an Fl - score of 0.76. Authors used text data for detecting Coronary Artery disease Buchan et al. ((2017)). Further, a neonatal heart defect architecure developed by Reddy et al. ((2020)) using OMNeT++ network architecture also.

Breast Cancer: All the breast cancer detection tasks explained below are done on Fine Needle Aspirate (FNA) data of Wisconsin Breast Cancer (WBC) dataset. Experiments are also done using supervised learning approaches for the detection of Breast Cancer Carvalho et al. ((2016)), Agarap ((2018)). Carvalho et al. ((2016)) experimented a Bayesian network for detecting whether a person is having Breast Cancer or not. The experiment is conducted on WBC dataset which is publicly available, reported an accuracy of 96%. The authors also proposed a multi criteria decision making model to take alternative solution in a difficult situation while diagnosing Breast Cancer Carvalho et al. ((2016)). Automatic detection of nuclei from histopathology images are of great importance Xu et al. ((2016)). Xu et al. ((2016)) proposed an unsupervised SAE based approach for achieving this. The model reported a Fl -measure of 84% and precision of 88% Xu et al. ((2016)). Agarap ((2018)) achieved higher accuracy of 99% on Wisconsin breast cancer dataset. An ensemble of deep learning methods using CNN named FCLF-CNN are proposed by Liu et al. ((2018)) for the classification of Breast Cancer patients into either benign or malignant, reported with highest sensitivity, specificity and accuracy of 99%, 97% and 98% respectively.

Tumor: For prediction of brain Tumor from MRI, a deep learning algorithm is used in Hussain et al. ((2017)). Hussain et al. ((2017)) proposed a CNN for training flair MRI brain Tumor images for the task of tumor segmentation. The proposed model is built using Keras library in Python. Tumor classification is done into three categories namely complete, core and enhanced reported with an accuracy, specificity and sensitivity of 0.82, 0.63 and 0.83 respectively Hussain et al. ((2017)). In another work, Min and Kyu ((2017)) proposed an image processing based segmentation method using BRATS 2015 challenge dataset. They experimented on K-means clustering and morphological operations as well as K- means clustering alone for this. The accuracy is high for K-means clustering and morphological operator reported with 98% Min and Kyu ((2017)). Doyle et al. ((2013)) proposed a HMM for identifying high grade and low grade tumors in the complete, enhancing, and core

Tumors. The experiment is conducted on BRATS challenge dataset reported with Dice scores of 0.84, 0.67 and 0.54 for high grade tumors and 0.81, 0.11 and 0.54 for low grade tumors in each category Doyle et al. ((2013)). Better results are reported using Residual Neural Networks (RES-net) by Shehab et al. ((2020)) on BRATS 2015 data. They reported an accuracy of 83%, 90%, and 85% for the complete, core, and enhancing regions respectively. Table 15 contains the good performance results for disease detection found in the previous works.

Reference	Disease	Data type	Classifier	Performance
Kamnitsas et al.((2017))	Stroke	MRI	CNN	Dice Score 0.83
Martinez-Murciaet al.((2013))	Parkinsons	MRI + HER	SVM	Accuracy 95%
Sarraf and Tofighi ((2016))	Alzheimer's	MRI	CNN (LeNet)	Accuracy 96%
Buchan et al. ((2017))	Heart disease	Clinical notes	SVM	F1 Score 0.76
Carvalho et al.((2016))	Breast Cancer	FNA	Bayesian network	Accuracy 96%
Shehab et al.((2020))	Tumor	MRI	Res-Net	Average Accuracy 85%

Table 15: Standard results found in the literature for disease detection.

2.1 Epidemic prediction

Epidemic prediction is the task of predicting the outbrea k of an epidemic Woo et al. ((2016)). Researchers used social media and other health data for predicting the outbreak of an epidemic in advance Woo et al. ((2016)). Social media data is used by the researchers for predicting the outbreak of epidemics Woo et al. ((2016)), van de Belt et al. ((2018)). Woo et al. ((2016)) predicted the outbreak of influenza using the Korean search engine Daum. They predicted the forecasting of Influenza outbreak using Support Vector Regression (SVR) with a r value of 0.956. van de Belt et al. ((2018)) pred icted the outbreak of Staphylococcus aureus (MRSA) in Netherlands using the Google trends data and ARIMA. They reported with an AUC of 0.61. Further, Saba and Elsheikh ((2020)) forecasted the Corona virus cases in Egypt using ANN reported with a MAE of 7.752 Saba and Elsheikh ((2020)). Further, an online malaria prediction system was developed by Pattanaik et al. ((2020)) using patient's image datasets. They used CNN for identifying the infected patients with reported accuracy of 98%. Table 16 contains the summary of epidemic predictions.

Reference	Disease	Data	Classifier	Result
Woo et al.((2016))	Influenza	Daum serach engine	SVR	r-value 0.956
Van de Belt et al. ((2018))	MRSA	Google trends	ARIMA	AUC 0.61
Saba and Elsheikh ((2020))	Corona	NS	ANN	MAE 7.752
Pattanaik et al.((2020))	CNN	Malaria	CNN	Accuracy 98%

Table 16: Summary of results on epidemic prediction.

3.SUMMARY

Figure 4 contains the percentage of ML and Deep Learning from the selected papers. Figure 5 contains the percentage of ensemble algorithms from the selected papers. Table 17 contains the standard datasets found in the literature that can be used by a new researcher in future. We observed that out of the surveyed papers (n=92), most of them implemented ML algorithms (n= 57). Also, a significant fraction of papers also implemented the ensemble models (n= 25) (see figures 4, 5). CNN is proving to be effective for finding the survival rate prediction task from unstructured data (see table 1). The health insurance data is playing a crucial role in the prediction of migration behavior and risk of hospitalization (see table 4). Social media posts by internet users are pretty useful for nowcasting air quality and ADR predictions (see tables 6, 13). Community detection algorithms are widely used by the researchers for healthcare fraud detection (see table 5). Hospital scheduling tasks and health care cost predictions are emerging research areas in health care analytics. The performance of the predictive models is increased using the advanced deep learning models for disease detection. Researchers are also using feature extraction techniques from unstructured data for better predictions (see table 15).

This survey paper identifies various tasks in healthcare predictive analytics in a broader way. The summary table that includes the classifier use for each task is also clearly given in the paper. Thus, a new researche raiming to enter the predictive health care analytics will get sufficient information about the area**.**

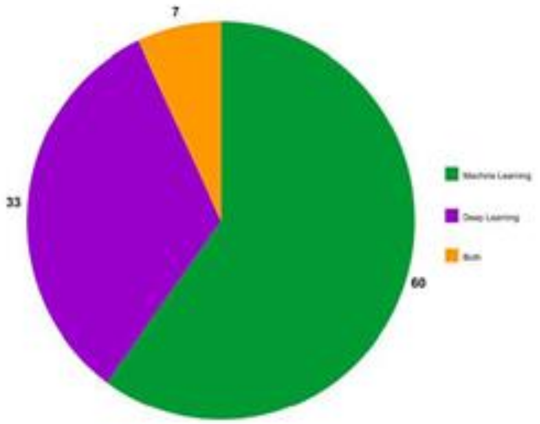

Figure 4: Percentage of ML, DL papers. DL-Deep Learning

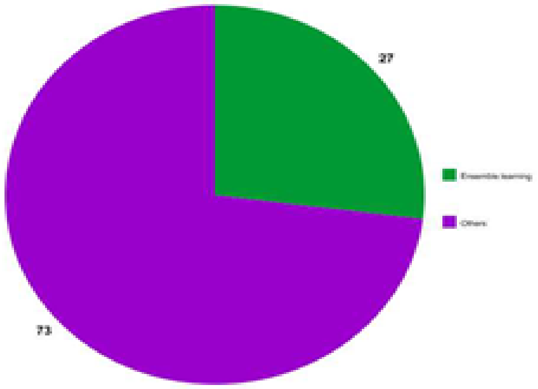

Figure 5: Percentage of ensemble algorithms.

4.FUTURE DIRECTIONS FOR RESEARCH

The following are some of the observed future directions for research.

- Text data can experiment for the prediction of heart disease. A new researcher can experiment on social media and other internet data for this.
- Researchers experimented on multi modalities of data when it comes to the multi-label classification of disease subtypes. In such situations, researchers have to And out the strong parameters which can predict the disease. But, this task is challenging because the exact or reliable parameters are not found out in the case of many diseases. Consequently, creating a model with available data types explaining various dimensions will be a good task.
- Nowcasting at smaller intervals may be less than an hour for air quality and epidemics are a well challenging task. As the nowcasting interval reduces, it is possible to capture minute epidemic transmission patterns.
- Due to the unavailability of government data for epidemic monitoring tasks, researchers can utilize available internet and social media data for this. Further, a predictive model created by social media data can also be evaluated by using previous available social media data will be interesting. This is due to the reason that the validation of predictive models using social media data requires original data of the epidemics , which concerned higher authorities of a country to release is mostly a time lagging process .
- Researchers can make use of publicly available air quality dataset of various countries and can validate the results of their predictive models . For example, www.data.gov.in[5], index.okfn .org[6], www.eea.europa.eu[7] which are publicly available.
- Multiple social media post data includes text, images will be a good experiment for ADR detection. Because, in health forums people will be engaging in health discussion involving disease , drug effects posting in the form of text, video, image, or audio.
- New researchers can experiment on more granular levels like an intersection of disease, hospital, resources, and infrastructures required for LOS predicions in hospital scheduling tasks.
- The potential of deep learning can be explored in the hospital scheduling tasks. Incorporating a large volume of inter – hospital data and applying various deep neural networks will be exciting.
- It is better to go for multi-social media data while forecasting multiple diseases. Selection of relevant online social forum which is popular among the targeted population is vital also.

  As the privacy and ethical issues prevent researchers from accessing individual- level data of a user. It is preferable to design tools and methods in online forums that can extract the required information without affecting privacy and security. Say, for example, D. Bell et. al Trotzek et al. ((2018)) created an online questionnaire for collecting information about diabetes, 2. The questionnaire is designed in such a way that health information is shared by only interested users without giving their identity.
- Transfer learning would be interesting for nowcasting prediction. As the previously acquired

knowledge can be transferred to the existing problem in a more dynamic way that increases the speed time of predictions, which is the key factor in nowcasting tasks.

- Researchers can build their own neural network and transfer other neural network model that is already implemented as black boxes for their task.

5https://data.gov.in/dataset-group-name/air-quality
6https://index.okfn.org/dataset/emissions/
7https://www.eea.europa.eu/themes/air/links/data-sources/airbase-public-air-quality-database

Reference	Dataset
Medicaion prediction	IC50
Healthacare fraud prediction	Medicare
Hospital scheduling prediction	MIMIC 1,2,3
Healthcare cost prediction	Medicare
ADR	SIDER, PubChem, DrugBank, MADE 1.0
Disease detection	Stroke – ISLE 2015 Challenge , Parkinson – PPMI Alzheimer's – ADNI, OASIS, Heart Disease – Heart Disease Risk factor Challenge Breast Cancer – WBC, TB – Tbneat, Schizophrenia – MLPS 2014 Challenge dataset Tumor- BRATS challenge (2015,2017)

Table 17: Standard datasets found in the li terature.

5.CONCLUSION

This paper surveyed 84 papers related to health care predictive analytics. The various tasks associated with the 85 papers are explained in detail. This paper also contains the summary tables for each task which helps a new researcher in exploring the area. We conducted a comprehensive survey more broadly so that a new researcher could find a new area of research and proceed in – depth in that area. The paper also discussed the various ML algorithms that are used for solving the various tasks in health- care predictive analytics. Further, the recommendations for future research from the topic are also addressed for giving more insights to new researchers.

REFERENCES

[1] A.F. M. Agarap. On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. In *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing,* pages 5- 9. ACM, 2018.

[2] M. Alhussein. Monitoring parkinson's disease in smart cities. *IEEE Access,* 5:19835- 19841, 2017.

[3] N. Alshurafa, C. Sideris, M. Pourhomayoun, H. Kalantarian, M. Sarrafzadeh, and J. - A. Eastwood. Remote health monitoring outcome success prediction using baseline and first month in tervention data. *IEEE journal of biomedical and health informatics,* 21(2): 507-514, 2017.

[4] A.Amadi-Obi, P. Gilligan, N. Owens, and C. O'Donnell. Telemedicine in pre -hospital care: a review of telemedicine applications in the pre-hospital environment. *International journal of emergency medicine,* 7(1):29, 2014.

[5] N. Amoroso, M. La Rocca, A. Monaco, R. Bellotti, and S. Tangaro . Complex networks reveal early mri markers of parkinson's disease. *Medical image analysis,* 48:12- 24, 2018.

[6] D. W. Bates, S. Saria, L . Ohno- Machado, A. Shah, and G. Escobar. Big data in healthcare: using analytics to identify and manage high-risk and high-cost patients. *Health Af fa'irs,* 33(7): 1123- 1131, 2014.

[7] G. Battineni, N. C hintalapudi, and F. Amenta. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (svm). *Informatics in Medicine Unlocked,* 16:100200, 2019.

[8] D. M. Bean, H. Wu, O. Dzahini. M. Broadbent, R. Stewart, and R. J. Dobson. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Scientific reports,* 7(1):16416, 2017.

[9] A.D. Billones, O. J. L . D. Demetria, D. E. D. Hostallero, and P. C. Naval. Demnet : A convolutional neural network for the detection of alzheimer's disease and mild cognitive impairment. In *Region 1D Conference (TENCON), 2016 IEEE,* pages 3724-3727. IEEE, 2016.

[10] N. Boodhun and M. Jayabalan. Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems,* 4(2):145- 154, 2018.

[11] A.K. Boscardin, R. Gonzales, K. L. Bradley, and M. C. Raven. Predicting cost of care using self-

reported health status data. *BMC health services research,*15(1):406, 2015.

[12] K. Buchan, M. Filannino, and O. Uzuner. Automatic prediction of coronary artery disease from clinical narratives. *Journal of biomedical in formatics,* 72:23-32, 21117.

[13] Y. Cao, L . L. Garcia, W. H. Curioso, C. Liu, B. Liu, M. J. Brunette, N. Zhang, T. Sun, P. Zhang, J. Peinado, et al. Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor and marginalized communities. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE),* pages 274- 281. IEEE, 2016.

[14] A.Carvalho, P. R. Pinheiro, and M. C. D. Pinheiro. A hybrid model to support the early diagnosis of breast cancer. *Procedia Computer Science,* 91:927- 934, 2016.

[15] S. Chandir, D. Siddiqi, O. Hussain, T. Niazi, M. Shah, V. Dharma, A. Habib, and A. Khan. Using predictive analytics to identify children at high risk of defaulting from a routine immunization program: Feasibility study. *JMIR public health and surveillance,* 4 (3):e63, 2018.

[16] A.B. Chapman, K. S. Peterson, P. R. Alba, S. L. DuVall, and O. V. Patterson. Detecting adverse drug events with rapidly trained classification models. *Drug safety,* 42(1):147- 156, 2019.

[17] L. Chato and S. Latifi. Machine learning and deep learning techniques to predict overall survival of brain tumor patients using mri images . In *Bioinformatics and Bioengineering (BIBE), 2011 IEEE 11th International Conference on,* pages 9- 14. IEEE, 2017.

[18] S. P. Chatrati, G. Hossain, A. Goyal, A. Bhan, S. Bhattacharya, D. Gaurav, and S. M. Tiwari. Smart home health monitoring system for predicting type 2 diabetes and hypertension. *Journal of King Saud University- Computer and Information Sciences,* 2020.

[19] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access,* 5:8869-8879, 2017.

[20] L. Cheng , Y. Shi , and K. Zhang. Medical treatment migration behavior prediction and recommendation based on health insurance data. *World Wide Web,* pages 1- 20, 2020.

[21] J. Cho, Z. Hu and M. Sartipi. Post-stroke discharge disposition prediction using deep learning. In *SoutheastCon, 2011,* pages 1- 2. IE EE, 2017.

[22] L. Chu, R. Qiu, H . Liu, Z. Ling, and X. Shi. Individual recognition in schizophrenia using deep learning methods with random forest and voting classifiers: Insights from resting state eeg streams. *arXiv preprint arXiv:1707.03467,* 2017.

[23] A.Craig, C. Arias, and D. Gillman. Predicting read mission risk from doctors' notes. *arXiv preprint arXiv:1711.10663,* 2017.

[24] H. Cui, Q. Li, H. Li, and Z. Yan. Healthcare fraud detection based on trustworthiness of doctors. In *Trustcom / Bi gDataSE/I SPA, 2016 IEEE,* pages 74- 81. IEEE, 2016.

[25] H.-J. Dai and C. -K. Wang. Classifying adverse drug reactions from imbalanced twitter data. *International journal of medical informatics,* 129:122-132, 2019.

[26] P. Desikan, N. Srivasta va, T. Winden, T. Lindquist , H . Britt , and J . Srivastava. Early prediction of potenti lly preventable events in ambulatory care sensitive admissions from clinical data. In *Healthcare In formatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on* , pages 124- 124. IEEE, 2012.

[27] I. D. Dinov, B. Heavner, M. Tang, G. Glusman, K. Chard, M. Darcy, R. Madduri, J. Pa, C. Spino, C. Kesselman, et al. Predictive big data analytics: a study of parkinson's disease using large, complex, heterogeneous, incongruent , multi- source and incomplete observat ions. *PloS one,* 11 (8):e0157077, 2016.

[28] S. Doyle, F. Vasseur, M. Dojat, and F. Forbes. Fully automatic brain tumor segmentation from multiple mr sequences using hidden markov fields and variational em. *Procs. NC I- M ICCAI BraTS,* pages 18- 22, 2013.

[29] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. The. Forecasting air quality time series using deep learning. *Journal of the Air &Waste Management Association,* 68(8):866-886, 2018.

[30] C. Gaser, K. Fra nk e, S. Kloppel , N. Koutsouleris, H . Sauer, A. D. N. Initiative , et al. Brainage in mild cognitive impaired patients :

predicting the conversion to a lzheimer' s disease . *PloS on e,* 8(6):e67346, 2013.

[31] S. Gittelman, V. Lange, C. A. G. Crawford, C. A. Okoro, E. Lieb, S.S. Dhingra, and E. Trmnarchi. A new source of data for public health surveillance: Facebook likes . *Journal of medical Internet research,* 17(4), 2015.

[32] B. Graham, R. Bond, M. Quinn, and M. Mulvenna. Using data mining to predict hospital admissions from the emergency department. *IEEE Access,* 6:10458-10469, 2018.

[33] X. Guo, W. Gandy, C. Coberley, J. Pope, E. Rula, and A. Wells. Predicting healthcare cost transitions using a multidimensional adaptive prediction process. *Population health management,* 18 (4):290-299, 2015.

[34] H. J. M. Hendri and H. S ulaiman. Predictive modeling for dengue patient's length of stay (los) using big data analytics (bda). In *International Conference of Reliable Information and Communication Technology,* pages 12-19. Springer, 2017.

[35] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder. Big data fraud detection using multiple medicare data sources. *Journal of Big Data,* 5(1):29, 2018.

[36] T. J. Hirschauer, H. Adeli, and J. A. Buford. Computer -aided diagnosis of parkinson 's disease using enhanced probabilistic neural network. *Journal of medicalsystems,* 39(11):179, 2015.

[37] S. Hussain, S. M. Anwar, and M. Majid. Brain tumor segmentation using cascaded deep convolutional neural network. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE,* pages 1998- 2001. IEEE, 2017.

[38] A.Isensee, P . Kickingereder, W. Wick, M. Bendszus, and K. H. Maier- Hein. Brain tumor segmentation nd radiomics survival prediction: Contribution to the brats 2017 challenge. *2017 International MICCAI BraTS Challenge ,* 2017.

[39] J. Islam and Y. Zhang. An ensemble of deep convolutional neural networks for alzheimer's disease detection and classification. *arXiv preprint arXiv: 1712.01675,* 2017.

[40] S. Jamal , S. Goyal, A. Shanker, and A. Grover. Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models. *Scientific Reports,* 7(1):872, 2017.

[41] M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, and E. Liu. Predicting all-cause risk of 311- day hospital readmission using artificial neural networks. *PloS one,* 12 (7): e0181173, 2017.

[42] K. Jee and G.-H. Kim. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. *Healthcare in formatics research,* 19(2):79- 85, 21113.

[43] S.Jiang, K.-S. Chin, G. Qu, and K. L. Tsui. An integrated machine learning framework for hospital readmission prediction. *Knowledge-Based Systems,* 146:73- 911, 21118.

[44] W.Jiang, Y. Wang, M.- H. Tsou, and X. Fu. Using social media to detect outdoor air pollution and monitor air

[45] quality index (aqi): ageo- targeted spatiotemporal analysis framework with sin a weibo (chinese twitter). *PloS one,* 10(10): e0141185, 2015.

[46] J. D. Koola, S. B. Ho, A. Cao, G. Chen, A. M. Perkins, S. E. Davis, and M. E. Matheny. Predicting 30-day hospital readmission risk in a national cohort of patients with cirrhosis. *Digestive Diseases and Sciences,* 65(4): 1003-1031, 2020.

[47] A.Kumar and H. Anjomshoa. A two -stage model to predict surgical patients' lengths of stay from an electronic patient database. *IEEE Jo·urnal of Biomedical and Health Informatics,*2018.

[48] K. Liu, G. Kang, N. Zhang, and B. Hou. Breast cancer classification based on fully-connected layer first convolutional neural networks. *IEEE Access,* 6:23722- 23732, 2018.

[49] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, et al. Multimodal neuro imaging feature learning for multiclass diagnosis of alzheimer's disease . *IEEE Transactions on Biomedical Engineering,* 62(4):1132- 11411, 2015.

[50] X.Liu, H. Chen. Identifying adverse drug events from patient social media: a case study for diabetes. *IEEE intelligent systems,* 311(3):44-51, 2015.

[51] B. Loevinsohn and A. Harding . Buying results? Contracting for health service delivery in developing countries. *The Lancet,* 366(9486): 676- 681, 2005.

[52] A.Maharlou, S. R. Niakan Kalhori , S. Shahbazi , and R. Ravangard. Predicting length of stay in intensive care units after cardiac surgery: Comparison of artificial neural networks and adaptive neuro- fuzzy system. *Health care informatics research,* 24(2) :109- 117, 2018.

[53] F. J. Martinez - Murcia, J. M. Gorriz, J. Ramirez, I. A. Illan, and C. G. Punt onet. Texture features based detection of parkinson's disease on datscan images. In *International Work-Conference on the Interplay Between Natural and Artificial Computation* , pages 266- 277. Springer, 2013.

[54] I. Matloob, S. A. Khan, and H. U. Rahman. Sequence mining and prediction - based healthcare fraud detection methodology. *IEEE Access,* 8:143256- 14327 3, 2020.

[55] T. I. Mazhar, N. J. Suha, D. Chaki, and M. H. Ali. Spinal cord injured (sci) patients' length of stay (los) prediction based on hospital admission data . In *Electrical Information and Communication Technology (EICT), 2017 3rd International Conference on,* pages 1- 6. IEEE, 2017.

[56] T . H . McCoy, A. M. Pellegrini, and R. H**.** Perlis. Assessment of time-series machine learning methods for forecasting hospital discharge volume. *JAMA network open,* 1(7):e184087 - e184 087, 2018.

[57] J. M. McWilliams and A. L.Schwartz. Focusing on high- cost patients: the key to addressing high costs? *The New England jo'Urnal of medicine,* 376(9):807, 2017.

[58] K. Meadows, R. G ibbens, C. Gerrard, and A. Vuylsteke. Prediction of patient length of stay on the intensive care unit following cardiac surgery: a logistic regression analysis based on the cardiac operative mortality risk calculator, euro score. *Journal of cardiothoracic and vascular anesthesia,* 32(6):2676- 2682, 2018.

[59] H. Menze, K. Van Leemput, D. Lashkari, T. Riklin- Raviv, E. Geremia, E. Alberts, P. Gruber, S. Wegener, M.-A. Weber, G. Szekely, et al. A generative probabilistic model and discriminative extensions for brain lesion segmentation- with application to tumor and stroke. *IEEE transactions on medical imaging,* 35(4):933- 946, 2016.

[60] Metcalf, W. Edmunds, and J. Lessler. Six challenges in modelling for public health policy. *Epidemics,* 10:93- 96, 2015.

[61] D. Miller and E. W. Brown. Artificial intelligence in medical practice: the question to the answer? *The American journal of medicine*, 2017.

[62] A.Min and Z. M. Kyu. Mri images enhancement and tumor segmentation for brain. In *Parallel and Distributed C'omput·ing, Applications and Technologies (PDC'AT), 2017 18th International Conference on*, pages 270-275. IEEE, 21117,

[63] M. Ojha and K. Mathur, Proposed application of big data analytics in healthcare at maharaja yeshwantrao hospital, In *Big Data and Smart City(ICBDSC), 2016 3rd MEC International Conference on* , pages 1-7. IEEE, 2016.

[64] V. Palanisamy and R. Thirunavukarasu. Implications of big data analytics in developing healthcare frameworks - a review. *Journal of King Saud University-Computer and Information Sciences,* 2017,

[65] D. L, Patrick and P, Erickson, Health status and health policy: quality of life in healthcare evaluation and resource allocation. 1993,

[66] R. Paul, S, H. Hawkins, Y, Balagurunathan, M, B, Schabath, R. J, Gillies, L.O. Hall, and D. B. Goldgof. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma, *Tomography: a journal for imaging research,* 2(4): 388 , 2016,

[67] J, Qi and J. Tejedor. Deep multi-view representation learning for multi - modal features of the schizophrenia and schizo-affective disorder. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on,* pages 952- 956. IEEE, 2016,

[68] W. Raghupathi and V, Raghupathi, An overview of health analytics . *J Health Med Informat,* 4(132): 2, 2013

[69] M. H. Rahmat, M. Annamalai, S. A. Halim, and R. Ahmad. Agent-based modelling and simulation of emergency department re- triage. In *Business Engineering and Industrial Applications Colloquium (BEIAC), 2 013 IEEE,* pages 219- 22, 4 IEEE, 2013.

[70] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze. Predicting asthma-related emergency department visits using big data, *IEEE J . Biomedical and Health Informatics,*19 (4): 1216-1223, 2015.

[71] S. Reddy, D. Kumar, et al. Early congenital heart defect diagnosis in neonates using novel wban based three - tier network architecture. *Journal of King Saud University- Computer and Information Sciences*, 2020.

[72] C. Roadknight, D. Suryanarayanan, U. Aickelin, J. Scholefield, and L. Durrant. An ensemble of machine learning and anti-learning methods for predicting tumour patient survival rates, In *Data Science and Advanced Analytics (DSAA), 2015, 36678 2015, IEEE International Conference on*, pages 1- 8. IEEE, 2015,

[73] M, Rouzbahman, A, Jovicic, and M. Chignell. Can cluster-boosted regression improve prediction of death and length of stay in the icu ? *IEEE journal of biomedical and health informatics*, 21 (3):851- 858, 2017.

[74] A.I. Saba and A. H. Elsheikh Forecasting the prevalence of covid - 19 out break in egypt using non linear autoregressive artificial neural networks. *Process Safety and Environmental Protection*, 2020.

[75] H**,** Sampathkumar, X.-w Chen, and B. Luo. Mining adverse drug reactions from online healthcare forums using hidden markov model. *BM C medical informatics and decision making*, 14(1):91, 2014.

[76] T. C. Seng, D. Xiao dong, T. E. Shyong, K. Y. H. Eric, T. E. Shiow, S. M. Kelvin, K. T. Wee, W . Thomas, and W. **H.** Lin. Predicting high cost patients with type 2 diabetes mellitus using hospital databases in a multi- ethnic asian population, In *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*, pages 240-243. IEEE, 2016.

[77] K. Shameer, K. W. Johnson, A. Yahi, R. Miotto, L. Li, D. Ricks, J. Jebakaran, P. KOVATCH, P. P. Sengupta, S. GELIJNS, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using mount sinai heart failure cohort . In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, pages 276-287. World Scientific, 2017.

[78] M. Shanthipriya and G, Prabhavathi Healthcare predictive analytics. 2018

[79] T. Siswantining, Y. Windesia, S. M. Soemartojo, M. M. Ariyanto, and M.R. Shahab. Predicting the risk of hospitalization to six diagnoses with highest costs based on outpatient claims. In *AIP Conference Proceedings*, volume 2014, page 02006, AIP Publishing, 2018.

[80] J. Song, Y. Wang, S. Tang, Y. Zhang, Z. Chen, Z. Zhang, T. Zhang, and F. Wu. Local- global memory neural network for medication prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[81] N. Song, J.-Y. Choi, H. Sung, S. Jeon, S. Chung, S. K. Park, W. Han, J. W. Lee, M. K. Kim, J .-Y. Lee, et al. Prediction of breast cancer survival using clinical and genetic markers by tumor subtypes. *PloSone*, 10(4):e0122413, 2015.

[82] K. Srinivasan, F. Currim, and S. Ram. Predicting high cost patients at point of admission using network science. *IEEE Journal of Biomedical and Health Informatics*, 2017.

[83] E. W. Steyerberg. *Clinical prediction models: a practical approach to development, validation, and updating.* Springer Science&Business Media, 2008.

[84] S. Sushmita , S. Newman, J. Marquardt, P . Ram, V. Prasad, M. D. Cock, and A. Teredesai. Population cost prediction on public healthcare datasets. In *Proceedings of the 5th International Conference on Digital Health 2015*, pages 87-94. ACM, 2015.

[85] B. Taati, J. Snoek, D. Aleman, and A. Ghavamzadeh. Data mining in bone marrow transplant records to identify patients with high odds of survival. *IEEE journal of biomedical and health informatics*, 18(1):21- 27, 2014.

[86] A.Tenev, S. Markovska-Simoska, L. Kocarev, J. Pop-Jordanov, A. Muller, and G. Candrian. Machine learning approach for classification of adhd adults. *International Journal of Psycho physiology*, 93(1):16 2- 166, 2014.

[87] T. H. van de Belt, P. T. van Stockum, L. J. Engelen, J. Lancee, R. Schrijver, J. Rodriguez-Bano, E. Tacconelli, K. Saris, M. M. van Gelder, and A. Voss. Social media posts and online search behaviour as early- warning system for mrsa out breaks. *Antimicrobial Resistance & Infection Control*, 7(1):1-10, 2018.

[88] Y. Wang, K. Ng, R. J. Byrd, J. Hu, S. Ebadollahi, Z. Daar, S. R. Steinhubl, W. F Stewart, et al. Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records. In

Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, pages 2530- 2533. IEEE, 2015.

[89] M. Wehenkel, C. Bastin, C. Phillips, and P. Geurts. Tree ensemble methods and parcelling to identify brain areas related to alzheimer's disease. In *Pattern Recognition in Neuroimaging (PRNI) , 2017 International Workshop on* , pages 1- 4. IEEE, 2017.

[90] A.Woo, Y. Cho, E. Shim, J.-K. Lee, C.-G. Lee, and S. H. Kim. Estimating influenza out breaks using both search engine query data and social media data in south korea. *Journal of medical Internet research*, 18(7):el77, 2016.

[91] A.P. Wright, A. T. Wright, A. B. McCoy, and D. F. Sittig. The use of sequential pattern mining to predict next prescribed medications. *Journal of biomedical informatics*, 53:73-80, 2015.

[92] A.Wu, S. Yang, Z. Huang, J. He, and X. Wang. Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Un locked*, 10 :100- 107, 2018.

[93] Y. Xie, G. Schreier, M. Hoy, Y. Liu, S. Neubauer, D. C. Chang, S. J. Redmond, and N. **H.** Lovell. Analyzing health insurance claims on different timescales to predict days in hospital. *Journal of biomedical informatics*, 60:187- 196, 2016.

[94] A.Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119-130, 2016.

[95] Q. Xue and M. C. Chuah. Incentivising high quality crowd sourcing clinical data for disease prediction. In *Proceedings of the Second IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies*, pages 185- 194. IEEE Press, 2017.

[96] M. Yang, X. Wang, and M. Y. Kiang. Identification of consumer adverse drug reaction messages on social media. In *PACIS*, page 193, 2013.

[97] D. Zhu, C. Cai, T. Yang, and X. Zhou. A machine learning approach for air quality prediction: Model regularization and optimization. *Big Data and Cognitive Computing*, 2(1):5, 2018.

[98] T. Zhu, L. Luo, X. Zhang, Y. Shi and W. Shen. Time-series approaches for forecasting the number of hospital daily discharged in patients. *IEEE journal of biomedical and health informatics*, 21(2):515-526, 2017.