Synchronization of Machine Learning into Electronic Health Records

Meet N. Gandhi
Computer Engineering
MPSTME, NMIMS, Shirpur

Eshan Vatsa
Computer Engineering
MPSTME, NMIMS, Shirpur

Nitin S. Choubey, PhD
HOD IT Department
MPSTME, NMIMS, Shirpur

ABSTRACT
The introduction of EHR (Electronic Health Record), in the medical field has been under discussion for a while but due to a very low acceptance rate of this technology by physicians, it has proven to be a risky gamble in the successful implementation of EHR. EHR uses data accumulated on the subject's health to determine tests required, health analysis and real-time records to help the physician provide more accurate and detailed analysis on the subject. Due to Health Information Technology for Economic and Clinical Health (HITECH) there has been an increase in the amount of data accumulation by EHR. The data has great potential because of the large archive of information across the globe, but due to the random collection of data, it has resulted in the development of an unstructured record which has resulted to difficulty in transactions [1]. Even though there has been a large collection of data around the globe, the major issue has been making use of this data in a logical manner for purposeful implementation. **The intention behind this paper is how to proceed with the implementation of machine learning in EHR along with its steps in order to analyze the data [2, 3] so that one can understand the pattern generated by the data provided.** There are several machine learning algorithms for the interpretation of this data, but not all data are compactable with all the algorithms, thus in this paper the method of data gathering to applying machine learning algorithms on the data is explained and various ways to perform different steps are also discussed in detail.

Keywords
EHR, machine learning, data extraction, data mining tools, analysis of data, Naïve Bayes classifier

1. INTRODUCTION
The concept of EHR emerged in the 1960s and it focused mainly on clinical data management. Prior to 1960, the records such as report and prescriptions were stored as physical copies. After the introduction of Problem oriented medical records (POMR) by Lawrence Weed [39]. In this record system, the process of each medical decision could be clearly understood. But, the medical professionals refused to use it, due to lengthy procedures. Weed himself promoted an electronic version of Problem oriented medical records. The concept received tremendous attention and shifted the focus of health-care industries to pay attention in storing medical records electronically rather than on paper. Though the machines in that era were not even remotely as competent as the ones we use today, EHR did get the favour in medical industry. In the year 1972 the electronic health record system was introduced for the first time by the REGENSTRIEF Institute. Since it was one of its kinds, it obviously had its own set of flaws. The biggest issue being that it was not even closely cost effective due to the recent introduction of computers then, thus only majorly funded institutions could afford it. Though this problem was resolved by the 1990s when personal computers became affordable and with the availability of the internet, a lot of attention was redrawn towards EHR. In the year 1996 Health Insurance Portability and Accountability Act (HIPAA) was introduced in order to deal with health-related security and privacy concerns of the public [4]. Today, EHR is made available only to authenticated medical personnel only. The intention behind introducing EHR in medical field is to help provide a more accurate diagnosis to help create a better analytical report about subject's undergoing medical examinations. Introducing Electronic Health Records in medical field has always been a hot topic under discussion due to the high scale rejection of medical practitioners to use HER [32, 33]. Thus, EHR has been limited to very few institutions.

On trying to find out the reasons on why EHR was unable to attract the medical practitioners, certain issues were highlighted. The main issue encountered with the use of EHR was that that the subject's data expanded constantly. Also, the data grows into an unstructured data, due to the need of information from varied fields. Another concern that was found was the subject's privacy. Many subjects might not be willing to consent the use of their medical records. As, it deemed a threat to leakage of sensitive information about the subject's previous treatments, path that was followed to provide a cure and records that can help identify patients, for certain clinical trials for advancement in the field. Finally the last point that has been debatable, as the EHR can also help identify prospective people to participate in clinical trials as it would be easier to approach them and would reduce the time, effort and resources spent which in turn could further help in medical research. Since, this can have positive as well as negative impact; the medical field is sceptical about the use of HER

2. METHODS
As there has been a great increase in the data of EHR in last 5 years [5], it provides a great scope for research, development and increase in the efficiency of the risk prediction algorithms [6, 7]. Risk prediction algorithms have been generated on the basis of the data being collected from large cohort studies. This cohort data which is accumulated for research-based purposes maintains certain standards whereas on the other hand, the data accumulated by EHR is fairly rampant and un-categorized [8]. The data they collect basically comprises of contents deemed necessary by the clinicians or records of sick patients. This set of data tends to be unorganized leading to difficulties in analysis. There are certain factors that come into play while analysing these records. For instance, there might come situations where a person let's say 'A' is undergoing treatment for a disease, which might not necessarily mean that a person 'B' who is suffering from the same disease can

undergo similar treatment. In such cases, certain factors need to be taken into consideration, for example, the age of a patient, his allergies, and the previous records of his treatments and so on and so forth.

On reviewing the various methods in different research papers, it was observed that there was a particular procedure that was followed. The same procedure to obtain the results here was used.

2.1 Procedure to be followed

A major problem faced in operating EHR has been the large accumulation of unorganized data. The sheer amount of data that was being accumulated seemed unfit to be used and to filter the necessary data; a procedure was brought into action [9].The procedure that is follow here processes the data first. It brings the data into a particular format to analyze frequent patterns and then uses predictive tools. These predictive tools are based on the historic data already feed to the tool. So, in nutshell, the procedure that was followed is as below:

1) Focus on gathering information from authenticated website/hospital or pathology reports.

2) Perform data extraction and cleansing using suitable tools to extract all possible data sets.

3) Provide a thorough scan and analysis of data using data mining tools.

4) Acquire the data and analyze the results using machine learning algorithms.

2.1.1 Step 1: Gathering information from authenticated websites/hospital or pathology reports.

One of the most convenient ways to gather information and data regarding any patient is to directly ask them, which can be feasible and practical in certain situation by doctors while treating then and entering them into their records. While, in the current scenario, various clinicians and institutions that wish to utilize a subject's medical records into application have to follow a certain process. First, they need to obtain the patient's consent, in case that is not possible then they need to obtain an ordinance of informed consent from their Institutional Review Boards (IRB), or another possibility to go about it, is to only use that particular data which doesn't have any identifiers. [11]

There are certain methods that focus on parameters that are taken into consideration to perform the above-mentioned procedures on particular systems [10]. Below given are the various methods that were mentioned in works of different researchers.

1) To achieve de-identification of medical records in order to avoid personal health information of patients being exposed, Uzuner developed a method in which certain set of characters cannot be fragmented or processed. He developed it on the state-id system, the program is not openly available it was developed using LIBSVM and is used to read discharge summaries [21].

2) One other software was developed to identify names, UMLS Metathesaurus terms in surgical pathology reports, Thomas developed a software in Regenstrief Institute system using Java and XSL to de-identify certain data. Also, this system is not publicly available. [18]

3) Another software was developed by Taira for de-identification from UCLA system, this software is used to deduce names and drugs for multiple client documents, thus making it easy and autonomous to perform the entire task without the help of any extra man hours. This system is publicly unavailable [19].

4) Mostly un-grammatical and fragmented clinical records showed a need for a system. The i2b2 de-identification challenge system was created by Prof. EijiAramaki, it mainly focuses on capturing information beyond a sentence and following the same word sequence for the same label. It utilizes CRF++ dataset which is unavailable publicly [12].

5) EHR, pathological reports are required in researches. HMS Scrubber which is a licensed open source (GNU LGPL v2) using JAVA , JDOM and MySQL was created by Beckwith BA to be used as a de-identifying tool for pathology reports as per HIPAA norms.[14]

6) For de-identification purposes to get components like list of names, location and medical terminology, Fredlin developed a system MedDS system using Java. This system is not publicly available this reads documents like HL7 messages [16].
7) Unlike SVM, grammar-based or rule-based techniques [13] James Gardner and Li Xiong used Conditional Random Fields-based named entity recognizer (NER), for mining sensitive attributes from the data set alone with it they have developed HIDE system which is a method that helps in keeping anonymity and prioritizing the attributes which should be considered within the framework. It currently works on Perl, Java, and this is the unavailability Mallet which is licensed as Open source (Common Public License v1). [10] [11].

8) There are many de-identification frameworks, many uses NLP techniques usually they don't provide structured data from textual data but by replacing sensitive or personal information from the records this is done using pattern recognition, conditional random field and support vector machines which have shown some of the best results. Author Morrison FP [15] has found a way to improve existing NLP system for de-identification and better utility.

Thus as discussed above there are certain software's which are designed to gather information for the pathological reports, this software can make it easy and quick to collect information from any reports as they are supported by advanced algorithms like support vector machine and Natural Language Processing. These algorithms not only just predict the output from the given data but also understand the grammar of the language to predict the output.

2.1.2 Step 2: Data extraction

The data extraction can be done from the various data fed by the medical practitioners. This process becomes easy, if the data is put in a particular format, but not otherwise. While the process of data gathering is going on, there are many factors that affect the symmetry of data arrangement. There are various factors such as, whether the provider considers the factors which he has observed important enough to enter in the patient's records. Information which has been added by the provider would be added in free text format rather than any defined list nor be picked from any particular structured data list. Lack of standardization is also going to affect a lot in building the dataset and lack of detailed diagnosis of the patient at the time of check-up are some of the factors which can make a huge impact on data integrity. While, there may be many factors that do not affect the integrity of the data, but some factors tend to affect the results of prediction. Some of them are as follows. [20]

1) When a family Physician is about to make a diagnosis, he takes help of a predefined set of parameters and creates a record according to it. In such cases, the format and the data entry in the EHR is questionable. As, there may be a time where family physician does not enter the required data in EHR because of the different structure of the records. Or it might be the case that the family doctor forgets to enter the minute details about the patient as they are already treating them for a long time.

2) EHR users choose to use free text rather than limiting their options to selected items within the drop-down list for information, that give away the problems faced, the medication recommended and the relevance. It requires queries to analyse information from the free text entered into the dataset into specified keywords. A drawback faced by of certain text within its set of prescribed keywords. Also free text is not reliable as there might be a case where the written handwriting of the doctor is not eligible to the person making entry in the Electrical Health Record and thus harming the integrity of the record collected.

3) While providers feed data into the data set there may come a time when they use different terms for the same disease. For an example when a physician is diagnosing a type 1 Diabetic patient, 'type 1 diabetes mellitus' can be referred as 'DM1' or 'diabetes' etc. with the synonymous text. This can cause a problem in the search for the predefined keywords. Not only the difference in name but also a case should be considered where local name of the symptom might differ from the actual scientific name and thus increase in ambiguity of the record is observed.

4) Medical records are generally stored as descriptive texts these are generally used as memos to be revisited. Since these are stored as free text without coded fields this becomes fairly difficult to analyse without a certain degree of manipulation which makes it time consuming and works rather like a chart audit. Since these texts are stored digitally the images are not recognizable but the text is somewhat searchable this is what makes it practically difficult.

2.1.3 Step 3: Using Data Mining Tools in EHR.

In this section Data Mining Tools for data analysis are discussed. There are several leading data mining tools in practice. Here, few frequently used open source tools such as RapidMiner, Knime and Weka which are recommended for big data analysis [22, 23] are considered. Gartner has noticed RapidMiner and Knime as leaders in advanced analytics platforms [24]. In this section, the open technologies and machine learning techniques that can be used to give a predictive analysis of big data which can be supported by visual tools are described. This is followed by an overview of the various visual tools that do not require code for big data analytics.[25,27]

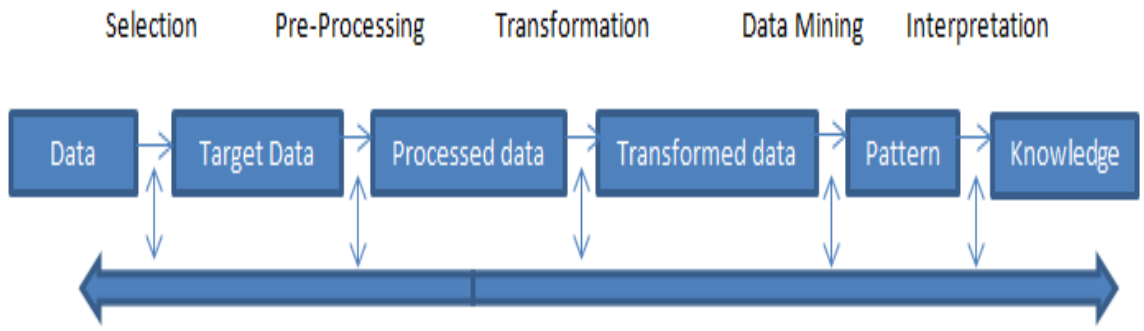

Figure 1 Steps for data mining

Data Mining is just another word for knowledge discovery in databases, referring to the above figure [1] data processes, data mining starts from collecting raw data and converting it into target data by applying selection process onto it. Selection process is just a simple filter to remove unwanted data or data which is not ready to use from the database, on which data mining is to be performed. In this case the raw data is the clinical records which need to be filtered and only important information in to be extracted. The process of selection is done using various software discussed in the above section, thus software are usually using Machine Learning algorithms or Natural Language Processing algorithms which helps the filtering process and the final output of them is targeted data. The Targeted data is then pre-processed which can give us the results of missing values or identification of outlier data. In Electronic Health Record due to many parameters the case of missing information can be easily observed as the parameters which are important for person A might not be important enough to consider for the person B thus increasing the case of missing information. Also, there is a case of outlier data which is defined as information which is abundantly different from the rest of the observations. Pre-processed data is then transformed using normalization which filters out the data. For example, normalization will filter out the data which is in 0 or 1 from the format that consist of textual data or by arranging the data in such a way that it can be easily comparable with another set of data present in database. Data mining is performed on this refined data which can be done by several algorithms as K-near, Decision tree, SVM and many others. These algorithms are used to find the patterns in the data, these patterns are then interpreted by the user in order to cross check any irrelevant data or any redundancy. This interpreted data from the patterns obtain using algorithms is termed as knowledge [26]. In this paper a Health Record dataset is taken from Kaggle, which is an open source platform that provides datasets for practice. The dataset was not pre-processed thus the method of pre-processing was applied on that and then the processed data was used for prediction, the Naive Bayes algorithm was applied on the dataset.

Data mining tools helps us to do the entire process with/without writing the code. A brief about some data mining tools such as MapReduce, Hadoop, YARN, Mahout, Spark, ML lib, SparkR&PySpark, RapidMiner, Weka (Waikato Environment for Knowledge Analysis) and KNIMW (Konstan Information Miner) is show in table[1].

Table 1 Different data mining tools

Sr No.	Data Mining Tools	Creator	Programming Languages Supported
1	MapReduce	Google	JAVA
2	Hadoop	Yahoo	JAVA, SCALA, PYTHON
3	Mahout	Apache Software Foundation	JAVA, SCALA
4	Spark	Apache	Java, Python, Scala, SQL, R
5	MLlib	Apache Spark's scalable machine learning library	Scala, Python, JAVA, R
6	SparkR&PySpark	Apache	Python and R
7	RapidMiner	Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer	R and Python

The table 1 consists of different software for data mining. Working of some of the most used data mining tools such as MapReduce, Apache Hadoop and Apache Spark are explained in this paper.

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. The implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on thousands of machines. [37]

Apache Hadoop was developed across clusters of computers using simple programming models, Apache Hadoop software library is a framework allowing distributed processing of large datasets using clusters. [30, 31]It is intended to scale up from single servers to a great many machines, each offering nearby calculation and capacity. Instead of depending on equipment to convey high-accessibility, the library itself is intended to identify and deal with issues at the application layer, hence conveying a profoundly accessible assistance over a group of computers, every one of which might be inclined to failures.[28, 17]

The committee and PMC members of Apache Spark, RaynoldXin and Aaron Davidson developed a framework named Databricks to mine big data in Spark. As Apache Spark works on the principal of Cluster compute system that makes data analytics fast with respect to both fast to run and fast to write. Spark written programs are much faster with respect to MapReduce which is up to 100X; also they are almost 10X times shorter in length and complexity. Also there are library present in Spark which can be used for streaming, machine learning, query execution and graph computation. [36]

2.1.4 Step 4: Acquiring and Analysis of data.

To predict the disease using the previous records Naïve Bayes Classifier was used. Bayesian Classifiers are statistical classifiers [34]. They are used to determine classes of tuples. This algorithm was chosen, since; it exhibits high accuracy and speed even on large datasets. When compared with decision tree and some neural networks, they deliver almost the same result.

The classifier is known as naïve, since; it assumes that an attribute is independent to the other one in a given class. This classifier is based on Bayes Theorem which depends on the conditional probability. The classifier predicts the class label of a tuple X is class Ci if and only if,

$$P(X|C_i)P(C_i)>P(X|C_j)P(C_j),$$

$$\text{for} i\leq j \text{ and } j\neq i$$

And P(x) is the probability of x

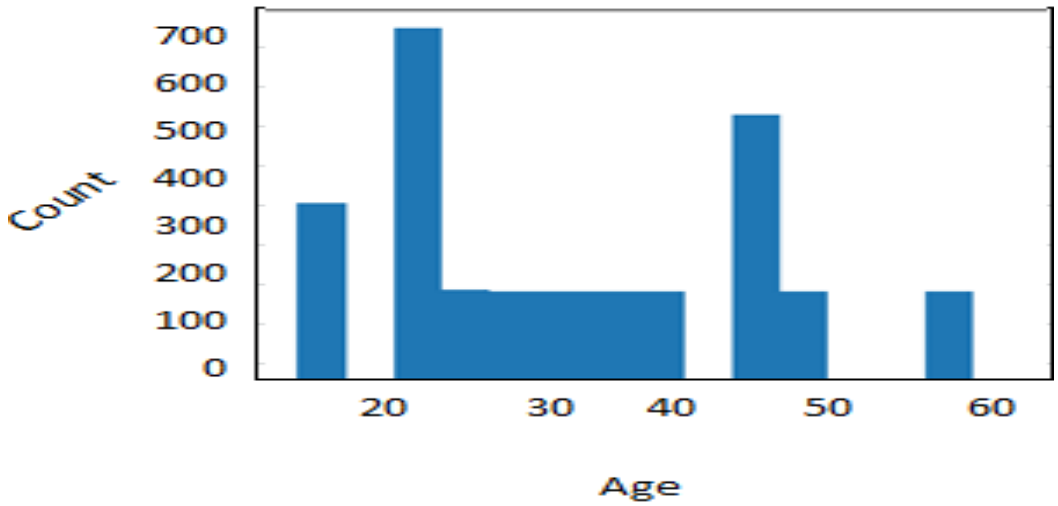

Figure 2 Age v/s count for people in data set

2.2 Algorithm to be followed

The following algorithm was used to implement Naïve Bayes classifier on the dataset taken from Kaggle [35]

1. Import the libraries.
2. Convert the column names according to one's preference.
3. Convert all the numeric data into integer type.
4. Normalize the attributes which are out of range to make their range same as other attributes.
5. Use One Hot Encoding to deal with the string values of gender.
6. Drop the diseases predicted and patient id column from the dataset.
7. Use all the attributes for x-coordinates, while labels of diseases are used as y-coordinates.
8. Divide the data into training and testing dataset along with randomizing the dataset.
9. Lastly, fit the model with the train data and later on predict using the testing dataset.
10. High accuracy was observed in this data set as shown in confusion matrix (figure 3) and the disease predicted is shown in table 3.

```
1  from sklearn.metrics import confusion_matrix
2  confusion_matrix(y_test, y_pred, labels=None, sample_weight=None)
```

```
array([[41,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0, 43,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0, 55,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0, 41,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0, 50,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0, 42,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0, 54,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0, 48,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0, 53,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0, 49,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 39,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 48,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 36,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 42,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 49,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 31]],
      dtype=int64)
```

Figure 3 Confusion matrix on Naïve Bayes classification

After receiving the entire data, what needs to be done is de-identification of data. Using the appropriate data mining tool, the necessary information is extracted followed by the necessary manipulations, after identifying certain keywords to create an organised data set which then can be analyzed and dealt with, as a seemed fit by clinicians. Figure 2 shows the relation between the numbers of people in the dataset to their age. The figure signifies the varied set of age groups involved in the analysis of the given data to signify its accuracy. Table 2 represents an instance of the data set that was used.

	headache	fever	running nose	sneezing	shivering	eye itch	dizziness	sweating	earache	cough	...	leg color change	weight loss	loss of appetite	bodyache	abnormal weight gain
1	1	5	0	0	2	0	1	2	0	1	...	0	0	3	2	0
2	0	1	0	0	0	0	2	3	0	0	...	0	0	0	4	0
3	5	2	0	0	0	0	4	3	0	0	...	0	0	0	0	0
4	1	0	5	5	3	0	0	0	0	3	...	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	3	0	0	0	0

Table 2 Data set with parameters

3. OUTPUT

The data collection was done from a trusted place, such as pathological reports and scrapped them to obtain information. Then, the data obtained by certain tools such as Hadoop or Apache Spark was processed for applying machine learning algorithms and found the correct algorithm suitable for the data. The entire process is done for prediction of diseases as output. The dataset, after pre-processing is shown in table 2. Naïve Bayes algorithm was implemented on the dataset using the steps shown above.

The external libraries should be imported first, and then the column names should be modified as per preference. In the next step all the data should be converted to a single format; in this case, it was numerical data format in integer data type. Next, there might be some attributes that are out of the desired limit such as outliers that need to be normalized to make the range the same as other attributes. Also, one-hot encoding is to be implemented to deal with string values of gender so that the integer 1 or 0 assigned to any gender doesn't bias the decision of the algorithm. Furthermore to pre-process the dataset, there is no use of disease predicted column and the patient ID column in the dataset so it is better to drop those columns. The next step is to use all the attributes as x-axis labels, while name of diseases are used for y-axis labels. To get training and testing data set, the dataset was divided into 30 - 70 ratios, in which 30 percent is for testing and 70 percent of data from the dataset is for training the model. Once the model is trained testing dataset was fitted in the trained model to obtain the predicted output. The output is shown in table 3; also the confusion matrix of the Naive Bayes algorithm is shown in figure 3. On studying the confusion matrix in figure 3 it can be observed that by following the prescribed steps, high accuracy is obtained, and the predicted results of disease obtained are quite close to the expected output.

	Disease predicted
1	Typhoid
2	Heart Attack
3	Migraine
4	Common Cold
5	PAD

Table 3 Predicted Desease for test data

4. CONCLUSION

Machine Learning algorithms are widely used in various fields, Electronic Health Record which is a highly accumulated unstructured data that consists collection of patient symptoms mapped to their disease along with their diagnosed reports. The implementation of any machine learning algorithm to a pre-processed data set is not a thing to worry about, but how the data is been pre-processed determines the output and the accuracy of the prediction. In this paper the steps of data mining is explained in detail which can be categorized as Data Gathering, converting that data to target data, Processing of the target data, Transforming the processed data, finding patterns in transformed data and obtaining knowledge from the transformed data. The process of obtaining information for data is illustrated and how the specialized software can be used to make the work easy is considered too. Also how these data mining steps can be automated using data mining tools like Hadoop and Apache Spark and many more are discussed. Lastly Naïve Bayes algorithm is implemented on the pre-processed data using the

proposed steps and the predicted output of the parent's disease was similar to the diagnosed one.

5. REFERENCES

[1] Ohio-Machado L. 2011. Realizing the full potential of electronic health records: the role of natural language processing. J. Am. Med. Inform. Assoc. 18, 539 (doi:10.1136/amiajnl-2011-000501) [PMCfree article] [PubMed]

[2] BruijnBd, Cherry C, Kiritchenko S, Martin J, Zhu X. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J. Am. Med. Inform. Assoc. 18, 557–562. (doi:10.1136/amiajnl-2011-000150) [PMC free article] [PubMed]

[3] Opportunities and obstacles for deep learning in biology and medicine_doi: 10.1098/rsif.2017.0387

[4] HIPAA act is available here:https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf)

[5] Opportunities and challenges in developing risk prediction models with electronic health record data: a systematic reviewJ Am Med Inform Assoc. 2017 Jan; 24(1): 198–208.Published online 2016 May 17. doi: 10.1093/jamia/ocw042

[6] Future of electronic health records: implications for decision support. Rothman B, Leonard JC,Vigoda MM Mt Sinai J Med. 2012 Nov-Dec; 79(6):757-68. [PubMed] [Ref list]

[7] Prediction of coronary heart disease using risk factor categories. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB Circulation. 1998 May 12; 97(18):1837-47.[PubMed] [Ref list]

[8] Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl3):S30–S37. [PMC free article][PubMed]

[9] Knowledge Acquisition for Electronic Health Records on clouddoi.org/10.1016/j.procs.2017.08.031

[10] Automatic de-identification of textual documents in the electronic health record: a review of recent research doi:10.1186/1471-2288-10-70

[11] Gardner J, Xiong L: HIDE: An Integrated System for Health Information De-identification. Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems 2008, 254-9.

[12] Aramaki E, et al: Automatic Deidentification by using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC 2006.

[13] R. M. B. A. Beckwith, U. J. Balis, and F. Kuo. Development and evaluation of an open source software tool for deidentification of pathology reports. B

[14] Beckwith BA, et al: Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform DecisMak 2006, 12.

[15] Morrison FP, et al: Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? J Am Med Inform Assoc 2009, 16(1):37-9

[16] Friedlin FJ, McDonald CJ: A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc 2008, 15(5):601-10.

[17] Hadoop Development Available: https://metadesignsolutions.com/hadoop-development

[18] Thomas SM, et al: A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp 2002, 777-81.

[19] Taira RK, Bui AA, Kangarloo H: Identification of patient name referenceswithin medical documents using semantic selectional restrictions. Proc AMIA Symp 2002, 757-61.

[20] Using your electronic medical record for research: a primer for avoiding pitfalls https://doi.org/10.1093/fampra/cmp068

[21] Uzuner O, et al: A de-identifier for medical discharge summaries. ArtifIntell Med. 2008, 42 (1): 13-35. 10.1016/j.artmed.2007.10.001.

[22] Wimmer H, Powell LM. A comparison of open source tools for sentiment analysis. 2015;1–9. Available:http://fotiad.is/blog/sentiment-analysis-comparison/.

[23] Jovic, A, Brkic K, Bogunovic N. An overview of free software tools for general data mining. Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on.IEEE. 2014: 1112–1117.

[24] Herschel G, Linden A, Kart L. Magic quadrant for advanced analytics platforms. Available:http://www.gartner.com/technology/reprints.do?id=1-2A881DN&ct=150219&st=sb.

[25] Landset S, Khoshgoftaar TM, Richter AN, Hasanin T. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J Big Data [Internet]. Springer International Publishing; 2015;2(1):24. Available:http://www.journalofbigdata.com/content/2/1/24.

[26] Fayyad, Piatetsky-Shapiro, Smyth Communications of the ACM,1996.

[27] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. Commun ACM [Internet]. 2008;51(1):1–13. Available:http://www.usenix.org/events/osdi04/tech/full_papers/dean/dean_html/.

[28] ApacheHadoop. Available:http://hadoop.apache.org/.

[29] ApacheMahout. Available:http://mahout.apache.org/.

[30] Zaharia M, Chowdhury M, Das T, Dave A. Fast and interactive analytics over Hadoop data with Spark. USENIX Login. 2012;37(4):45–51.

[31] Landset S, Khoshgoftaar TM, Richter AN, Hasanin T. A survey of open source tools for machine learning with

big data in the Hadoop ecosystem. J Big Data [Internet]. Springer International Publishing; 2015;2(1):24. Available:http://www.journalofbigdata.com/content/2/1/24.

[32] https://www.hindawi.com/journals/jhe/2018/4302425/

[33] https://onlinelibrary.wiley.com/doi/full/10.1111/acem.12876

[34] Data Mining: Concepts and Techniques by Jiawei Han and MichelineKamber.

[35] DatasetAvailable:https://www.kaggle.com/asaumya/healthcare-data

[36] Big data mining using Apache Spark Available: https://insidebigdata.com/2014/10/27/data-science-101-mining-big-data-apache-spark/

[37] MapReduce: Simplified Data Processing on Large Clusters Jeffrey Dean and Sanjay Ghemawat

[38] A micropartitioning technique for massive data analysis using MapReduce S. Mohanapriya ; P. Natesan.https://www.icanotes.com/2019/04/16/a-history-of-ehr-through-the-years/