International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

Good Physician Prediction using Data Mining and Multidisciplinary Approach

Prof. Bhosale R. S[1], Prajakta Kulkarni[2], Shruti Hadawale[3], Varsha Gorde[4], Kalyani Jadhav[5]

Professor, Department of Information Technology, Amrutvahini Engineering College, Sangamner, Maharashtra, India[1]

Student, Department of Information Technology, Amrutvahini Engineering College, Sangamner, Maharashtra, India[2,3,4,5]

ABSTRACT*:* The fast growing field of big data analysis has started to play a fundamental role in the evolution of healthcare research. The huge data growth in biomedical and healthcare businesses, accurate analysis of medical data benefits from early detection, patient care, and community services .It has provided tools to analyze, manage and assimilate large volumes of big data produced by current healthcare systems. Data analysis has recently been applied to help the disease delivery and treatment process. This proposed system rationalizes machine learning algorithms to effectively predict the physician outbreaks in disease-frequent communities we are experimenting with the modified predictive models with real hospital data.

KEYWORDS: Big data, Open data, Decision Tree, Multidisciplinary diagnosis.

I. INTRODUCTION

Big Data is the art of finding information or knowledge in a large amount of data. Like statistics, Big Data is becoming increasingly common in companies and organizations that want to extract relevant information from their databases, which they can use for their own needs. Big Data tasks can, in general, be classified as tasks of description and prediction. To understand the discovery goal, it is vital to understand the difference between descriptive and predictive tasks. Big Data technology is applied in an increasing variety of fields. Prediction should be done to reduce the risk of Disease. Diagnosis is usually based on signs, symptoms and physical examination of patient. All most all the doctors are predicting diseases by learning and experience. The Diagnosis of disease is difficult and tedious task in medical field. Predicting disease from various factors and symptoms is a multi-layered issue which may lead to false presumptions and unpredictable effects. Health care industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. The large amount data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. We use Machine Learning algorithms should aim to predict disease and Physician.

II. LITERATURE SURVEY

L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua [1]. Using the novel scheme to code the medical records, it jointly utilizes local mining and global learning approaches, which are tightly linked mutually reinforced. For extracting medical concepts from medical record, local mining attempts to code the individual medical record and then mapping them to authenticated terminologies. Sometimes, local mining approach, suffer from data loss and lower precision, which are caused by the absence of key medical concepts and the presence of irrelevant medical concepts.

L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S [2]. Initially report a user have the information of health seekers in terms of questions and then select those that ask for possible diseases of their manifested symptoms for further analytic. And then propose a novel deep learning scheme to interpret the possible diseases given the questions of health seekers. The proposed scheme includes the two key components. The first globally mines the discriminant medical signatures from raw features. The second holds the raw features and their signatures as input nodes in one layer and hidden nodes in the subsequent layer, respectively. It learns the inter-relations between these two layers via pre-training with pseudo labeled data.

X. Yin, J. Han, and P. S. Yu [3]. In that propose a new problem, called Veracity, i.e., conformity to truth, and in that they study about how to find true facts from a large amount of incompatible information on many subjects that is

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

provided by various websites. They design a general framework for the Veracity problem and invent an algorithm, called TRUTHFINDER, which utilizes the relationships between websites and their information, i.e., a website is trustworthy if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy websites. An iterative method is used to infer the trustworthiness of websites and the correctness of information from each other.

X. L. Dong, L. Berti-Equille, and D. Srivastava [4]. A novel approach that considers dependence between data sources in truth discovery. Suddenly, if two data sources provide a large number of common values and many of these values are rarely provided by other sources (e.g., particular false values), it is very likely that one copies from the other. They apply Bayesian analysis to decide dependence between sources and design an algorithm that iteratively detects dependence and discovers truth from conflicting information.

J. Pasternack and D. Roth [5]. In that they find the different algorithm for the prior knowledge as framework. A framework expresses both general "common-sense" reasoning and specific facts already known to the user as first-order logic and translating this into a tractable linear program. As our results show, this approach scales well to even large problems, both reducing error and allowing the system to determine truth respective to the user rather than the majority. Additionally, they introduce three new fact-finding algorithms capable of outperforming existing fact-finders in many of our experiments.

Q. Li, Y. Li, J. Gao, B. Zhao,W. Fan, and J. Han [6]. To resolves quarrel among multiple sources of heterogeneous data types. They model the difficulty using an optimization framework where truths and source reliability are defined as two sets of unknown variables. The objectives minimize the overall weighted deviation between the truth and the multi-source observations where each source is weighted by its reliability. Different loss functions can be incorporated into this framework to recognize the characteristics of various data types, and efficient computation approaches are developed.

Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han [7]. Truth discovery approaches have different assumptions about input data, relations between sources and objects, identified truths, etc. Applications in some types of domains also have their unique characteristics that should be taken into account. These motivate the needs for diverse truth discovery techniques.

D.Wang, L. Kaplan, H. Le, and T. Abdelzaher [8]. In that they focus on binary measurements. Optimality, in the sense of maximum likelihood estimation, is attained by solving an expectation maximization problem that returns the best guess regarding the correctness of each measurement. The approach is shown to outperform the state of the art fact-finding heuristics, as well as simple baselines such as majority voting.

X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava, "Truth finding on the deep web: Is the problem solved?" year-2012[9]. In that the information about Truthfulness, Web data in two domains where they believed data are fairly clean and data quality is important to people's lives: Stock and Flight, the huge measure of conflicting information from the distinctive sources and also some sources with quite low accuracy. Apply two data sets state-of-the-art data fusion methods that aim at resolving conflicts and finding the truth, analyzed their strengths and limitations,
and suggested promising research directions.

S. Mukherjee, G. Weikum, and C. Danisco-Niculescu-Mizil, "People on drugs: credibility of user statements in health communities," year- 2014 [10]. A method for automatically establishes the credibility of user-generated medical statements and the trustworthiness by exploiting linguistic cues and distant supervision from expert sources. To this end they introduce a probabilistic graphical model that jointly learns user trustworthiness, statement credibility, and language objectivity. This methodology to the task of extracting rare or unknown side-effects of medical drugs and this being one of the problems where large scale non-expert data has the potential to complement expert medical knowledge.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

III. BACKGROUND

Data mining

Data mining is the art of finding information or knowledge in a large amount of data. Like statistics, data mining is becoming increasingly common in companies and organizations that want to extract relevant information from their databases, which they can use for their

own needs . Data mining tasks can, in general, be classified as tasks of description and prediction .To understand the discovery goal, it is vital to understand the difference between descriptive and predictive tasks.

Fig 1 : The problem of selecting Physician

IV. PROPOSED SYSTEM

Architecture

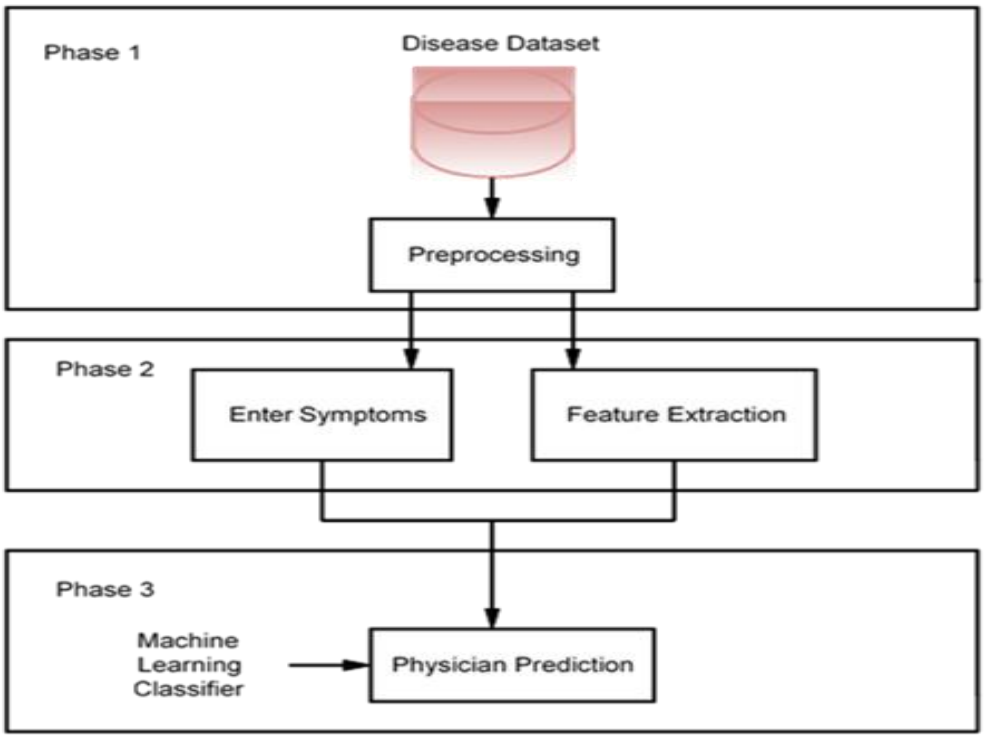

Fig 2 : Architecture

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

Health care industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. The large amount data is a key resource to be processed and analyzed for knowledge extraction that enables support for decision making. We use machine learning algorithms should aim to predict good physician.

A. Algorithm

1. Naive Bayes

Steps:

1. Given training dataset D which consists of documents belonging to different class say Class A and Class B
2. Calculate the prior probability of class A=number of objects of class A/total number of objects. Calculate the prior probability of class.
3. B=number of objects of class B/total number of objects
4. Find NI, the total no of frequency of each class

Na=the total no of frequency of class A
Nb=the total no of frequency of class B

5. Find conditional probability of keyword occurrence given a class:
   P (value 1/Class A) =count/ni (A)
   P (value 1/Class B) =count/ni (B)
   P (value 2/Class A) =count/ni (A)
   P (value 2/Class B) =count/ni (B)
   ……..
   ……..
   ……..
   P (value n/Class B) =count/ni (B)
6. Avoid zero frequency problems by applying uniform distribution
7. Classify Document C based on the probability p(C/W)
   a. Find P (A/W) =P (A)*P (value 1/Class A)* P (value 2/Class A)……. P(value n /Class A)
   b. Find P (B/W) =P (B)*P (value 1/Class B)* P (value 2/Class B)……. P(value n /Class B)
8. Assign document to class that has higher probability.

4.2. Block diagram

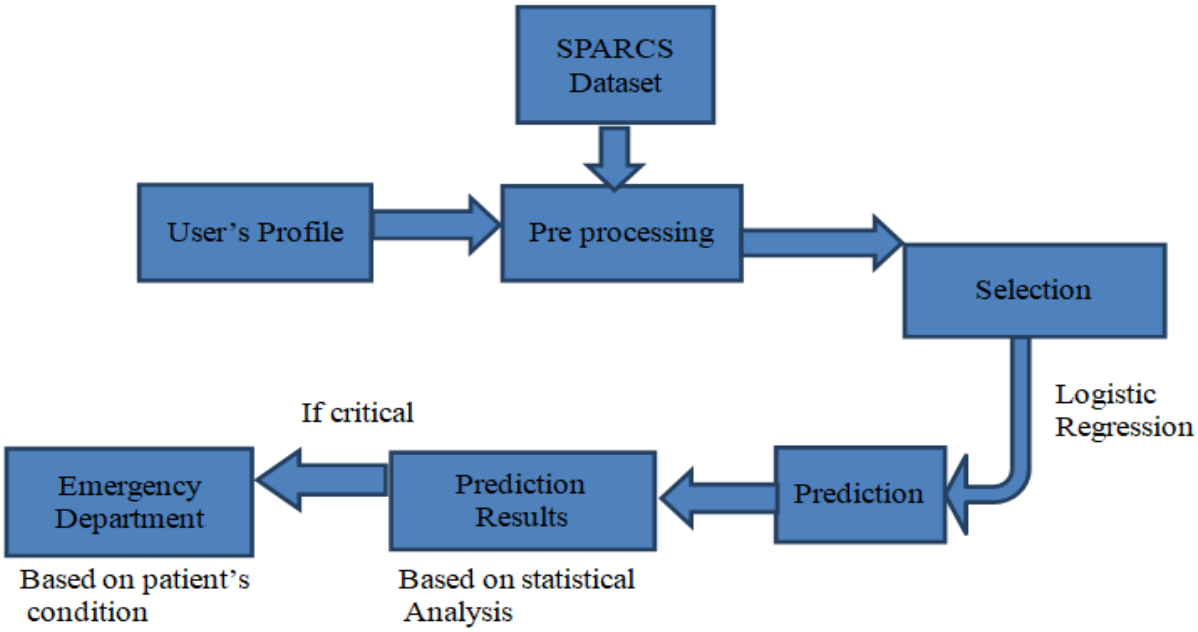

Fig 3: Physician-Patient Relationship Model

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

V. EXPERIMENTAL RESULTS

A. Comparison Graph :

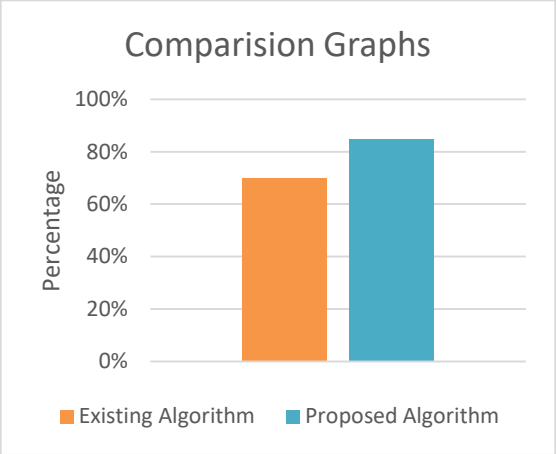

Fig 4 : Comparison Graph

B. Comparison Table :

Sr. No	Existing Algorithm	Proposed Algorithm
1	65%	86%

Table 1: Comparitive Results.

VI. CONCLUSION

In this paper, Probability models are powerful tools with which to predict probabilities in the medical field. They have not been widely used for making predictions on individual physicians but, they can be used in this way. The big data analytics technology adopted in this study is decision tree. Based on our experimental results, we believe that choosing an appropriate physician will be possible with an appropriate predictive model.

REFERENCES

[1] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, “Bridging the vocabulary gap between health seekers and healthcare knowledge,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 2, pp. 396–409, 2015.
[2] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, “Disease inference from health-related questions via sparse deep learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 8, pp. 2107–2119, 2015.
[3] X. Yin, J. Han, and P. S. Yu, “Truth discovery with multiple conflicting information providers on the web,” in Proc.