Improving Disease Prediction by Machine Learning

Smriti Mukesh Singh[1], Dr. Dinesh B. Hanchate[2]

[1]*Student of Computer Engineering, Pune University,* VPKBIET, Baramati, India.
[2] *Professor of Computer Engineering, Pune University, VPKBIET, Baramati, India*

Abstract - *These days utilization of Big Data is expanding in biomedical and human services groups, exact investigation of medicinal information benefits early malady discovery, quiet care and group administrations. Fragmented therapeutic information lessens examination precision. The machine learning calculations are proposed for successful expectation of ceaseless infection. To beat the trouble of deficient information, Genetic algorithm will be utilized to remake the missing information. The dataset comprise of structured data and unstructured data. To extract features from unstructured data RNN algorithm will be utilized. Framework proposes SVM calculation and Naive Bayesian calculation for sickness expectation utilizing unstructured and structured information individually from hospital information. Community Question Answering (CQA) system is additionally proposed which will foresee the inquiry and answers and will give proper responses to the clients. For that, two calculations are proposed KNN and SVM. KNN algorithm will perform classification on answers and SVM calculation will perform classification on answers. It will help client to discover best inquiries and answers identified with infections.*

Key Words: **Big data analytics, Healthcare, Machine Learning**

1. INTRODUCTION

Almost 61% of deaths in India are presently ascribed to Non-Communicable Diseases (NCD), including heart issue, cancer and diabetes, as indicated by new information discharged by the World Health Organization on Monday. Very nearly 23% are in danger of premature death because of such ailments. In India, an aggregate of 58,17,000 deaths were evaluated from illnesses like cancer, diabetes and heart issues in 2016. Cardiovascular infections (coronary illness, stroke, and hypertension) add to 45% of all NCD deaths, trailed by chronic respiratory disease (22%), cancer (12%) and diabetes (3%). Cancer, Diabetes and heart disease alone record for 55% of the untimely mortality in India in the age gathering of 30-69 years. With the change of expectations for everyday comforts, the occurrence of chronic disease is expanding. It is basic to perform hazard appraisals for chronic disease. With the development in therapeutic information, gathering Electronic Health Records (EHR) is progressively advantageous. Proposed a healthcare system using smart clothing for sustainable health monitoring had thoroughly studied the heterogeneous systems and achieved the best results for cost minimization on tree and simple path cases for heterogeneous systems. Patients' factual data, test results and infection history are recorded in the EHR, empowering us to distinguish potential information driven answers for lessen the expenses of restorative contextual analyses. Proposed a proficient stream assessing calculation for the tele-health cloud framework and outlined an information soundness convention for the PHR (Personal Health Record)- based conveyed framework. Cloud framework and composed an information intelligence convention for the PHR (Personal Health Record) based conveyed framework. Proposed six applications of big data in the field of medicinal services yet these plans have attributes and imperfections moreover. The informational index is ordinarily little, for patients and ailments with particular conditions; the qualities are chosen through understanding. In any case, these pre-chosen attributes possibly not fulfill the adjustments in the ailment and its impacting factors. With the advancement of huge information examination innovation, more consideration has been paid to infection expectation from the point of view of huge information investigation, different explores have been directed by choosing the attributes naturally from an extensive number of information to enhance the precision of hazard order, instead of the beforehand chose qualities. Be that as it may, that current work generally thought to be organized information. For unstructured information, for instance, utilizing Convolutional Neural Network (CNN) to separate content qualities consequently has just pulled in wide consideration and furthermore accomplished great outcomes. Besides, there is a huge distinction between diseases in various districts, basically due to the assorted atmosphere and living propensities in the locale. In this manner, hazard arrangement in light of enormous information examination, the accompanying difficulties remain: How the missing information ought to be tended to? By what means should the principle constant maladies in a specific locale and the fundamental qualities of the sickness in the district be resolved? In what manner can enormous information investigation innovation be utilized to break down the ailment and make a superior model? To take care of these issues, proposed System joins the organized and unstructured information in human services field to evaluate the danger of sickness. To begin with, Genetic algorithm will be utilized to remake the missing information. To extract features from unstructured information RNN algorithm will be utilized. Finally, SVM algorithm and Naive Bayesian algorithm for disease prediction using unstructured and structured data, individually from healing center information. A conclusion of proposed System is that the execution of SVM and Naive Bayesian calculation is superior to other existing techniques. Paper proposes question answer framework i.e. Community Question Answering (CQA) system it perform include extraction and arrangement on the questions. It enables client to look through the inquiry here, first client will scan for the inquiry, based on that framework will perform highlight extraction and will show related inquiries. Then client will choose the inquiries after that framework performs include extraction on that inquiry and will show the related inquiries. To perform classification on questions it utilizes KNN algorithm. To perform classification on answers it utilizes SVM algorithm. It enables client to post the inquiries and answers related to all diseases. On the off chance that client needs to look through a specific inquiry, first client needs to enter the inquiries. After that RNN calculation will perform feature extraction.

2. REVIEW OF LITERATURE

A new Convolutional Neural Network based Multimodal Disease Risk Prediction (CNN-MDRP) [1] has been indicated through which high danger of disease is being anticipated. A recurrent structure to capture contextual information. The new deep learning architecture Bi-CNN-MI Paraphrase Identification (PI) [2].The PI contemplates two sentences on various levels of granularity. They choose if rephrase by and large has a similar importance. The parameters of the considerable number of models are updated for PI. Usage of vernacular showing task is to address the nonattendance of planning data. They have analyzed machine learning calculations like Decision Tree, Bayes algorithm, Support Vector Machine (SVM) and Nearest Neighbor [3]. These figuring are used all together generally. They are used for anticipating group enlistment for data illustrations. They give a relative examination of various calculations. In information mining they remove the covered insightful data from the sweeping database. The capacity of Electronic Health Record (EHR) [4] is for setting up the new patients by revealing the dark sickness connection. In EHR and its mining a sweeping extent of good, honest to goodness and particular reasons may keep the systematic declaration. The tele-health administrations are being used which are known as the tele-health cautioning organizations. They are generally used as a piece of metropolitan urban communities. Due to tele-health organizations the patients can get help effortlessly [5].

A quick incremental in the tele-health structure has become diverse strategies like distributed computing and enormous information. They have a dynamic programming to make perfect game plans with the objective that data sharing frameworks can be dealt with. In this it contemplates the transmission probabilities, the arranging objectives, and moreover increasing as far as possible. For a content conclusion examination with jointed Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) engineering [6], taking the upsides of both like course grained neighborhood highlights features which are made by CNN and long-separate conditions learned by methods for the RNN. The provincial perpetual infection has been engaged. Consideration has been paid on both organized and unstructured information. It utilizes a maximum pooling layer that consequently judges, which words assume an essential part in content arrangement to catch the key segments in writings [7]. Their strategy and results could be capable to update the perception of sickness specific settings and moreover to improve the insightful execution in mortality showing in intense healing center care. In [9] they join illness particular settings into mortality displaying by detailing the mortality forecast issue as a multi-errand learning issue in which an undertaking relates to an ailment. Our technique viably coordinates restorative area information relating to the similitude among illnesses and the likeness among Electronic Health Records (EHRs) into information driven approach by joining chart Laplacians into the regularization term to encode these likenesses. The test comes about on a genuine dataset from a healing facility support the viability of the proposed strategy. The Acute Hospital Care (AUC) of a few baselines was enhanced, including calculated relapse without multi-errand learning and a few multi-undertaking learning strategies that don't consolidate the area information. Moreover, we show some fascinating outcomes relating to disease specific prescient highlights, some of which are not just steady with existing medicinal area learning, yet in addition contain suggestive theories that could be approved by facilitate examinations in the medicinal area. Finding comparative inquiries [10] from chronicled files has been connected to address replying, with well hypothetical underpinnings and extraordinary functional achievement. All things considered, each inquiry in the returned competitor pool regularly connects with various answers, and thus clients need to meticulously peruse a considerable measure before finding the right one. To lighten such issue, we exhibit a novel plan to rank answer applicants by means of pairwise examinations. Specifically, it comprises of one disconnected learning part and one online inquiry segment. In the disconnected learning segment, we first consequently set up the positive, negative, and impartial preparing tests as far as inclination sets guided by our information driven perceptions. We at that point exhibit a novel model to together fuse these three sorts of preparing tests. The shut frame arrangement of this model is determined. In the online inquiry segment, we first gather a pool of answer possibility for the given inquiry by means of discovering its comparative inquiries. We at that point sort the appropriate response applicants by utilizing the disconnected prepared model to judge the inclination orders. Broad examinations on this present reality vertical and general group based inquiry noting datasets have nearly shown its heartiness and promising execution. Additionally, we have discharged the codes and information to encourage different scientists. In [13] the Natural Language Processing (NLP) is a path for PCs to break down, comprehend, and get importance from human dialect smartly. Recurrent Neural Networks (RNN) has upset the field of NLP. RNNs are utilized at demonstrating units in arrangement. Not at all like nourish forward neural systems, RNNs have cyclic associations making them all the more effective for demonstrating contributions of groupings. They have been effectively utilized for succession marking and arrangement forecast undertakings, for example, penmanship acknowledgment, dialect displaying, machine interpretation, phonetic naming of acoustic edges and so on. This paper gives a review of how RNNs are being utilized and fit for managing Natural Language Processing. In this it additionally outlines LSTM based RNNs models.

3. SYSTEM ARCITECTURE

To help foresee whether a patient is experiencing chronic disease or not as indicated by his/her medical history. The input esteem is the attribute value of the patient, which incorporates the patient's close to home data, for example, age, sex, the pervasiveness of side effects, and living propensities (smoking or not) and other structured information and unstructured information. The yield esteem shows whether the patient is experiencing chronic disease or not. For disease hazard, demonstrating the precision of risk expectation relies upon the assorted variety highlight of the doctor's facility information, i.e., the better is the element depiction of the disease, the higher the exactness will be. For some straightforward sickness, e.g., hyperlipidemia, just a couple of highlights of organized information can get a decent depiction of the illness, bringing about genuinely great impact of disease expectation. Be that as it may, for an unpredictable disease, for example, cerebral infarction, diabetes, hypertension and asthma just utilizing highlights of structured data isn't a decent method to depict the disease. In this way, use the structured data as well as the content information of patients in view of the Support Vector Machine and Naive Bayes (NB) algorithms.

In fig. 1, the dataset contains patient's information related to chronic disease. The dataset is been collected from the hospital. With the help of dataset, the accurate prediction of disease can be done. In structured data the prediction of disease is done with the help of symptoms of each chronic disease. The disease prediction is done by NB algorithm. The NB algorithm is useful for predicting the probability of multiple classes based on various attributes. In this the prediction of disease is by 96% based on the symptoms of chronic diseases like hypertension, diabetes, cerebral infraction and asthma. For Structured data, the system uses a traditional machine learning algorithm, i.e., NB algorithm to predict the disease.

NB classification is a simple probabilistic classifier. It requires calculating the probability of feature attributes. For Structured information, framework utilizes conventional machine learning calculation, i.e., NB calculation to anticipate the sickness. NB characterization is a straight forward probabilistic classifier. It requires figuring the likelihood of highlight properties. A NB classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independent assumption. A more descriptive term for the underlying probability model would be the self-determining feature model. In basic terms, a NB classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature. The NB classifier performs reasonably well, even if the underlying assumption is not true. The advantage of the NB classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. In order to train a Naive Bayes (NB) model for text classification, there is a need to prepare data set. Genetic Algorithm includes process of initialization, and then it improves with a repetitive application of mutation, crossover, inversion and selection operations. It requires a genetic representation and fitness function. When some user's data is missing then it is been recovered by genetic algorithm. In unstructured data, if there is missing data which is caused by patient's mistake. Then missing data is been recovered with the genetic algorithm. The unstructured data mainly focuses on the case study and interrogation which are given by doctors. The Recurrent Neural Network (RNN) algorithm is used to extract features of the text. The stop words are been removed from the text data and the features are extracted successfully. After text feature extraction, SVM Classier performs classification on the data; it will predict whether the patient is suffering from chronic disease or not. With the help of RNN, unstructured data is been converted into structured and the prediction of chronic disease is been done. In a traditional neural system it is expected that all inputs (and outputs) are autonomous of each other. On the off chance that you need to foresee the following word in a sentence you better know which words preceded it. RNNs are called recurrent on the grounds that they play out a similar undertaking for each component of a sequence, with the yield being relied upon the past calculations. Another approach to consider RNNs is that they have a "memory" which catches data about what has been figured up until now. In principle RNNs can make utilization of data in subjectively long arrangements. The textual features are extracted by RNN. In Fig 1. Xt is the input; Ht is the hidden state which is calculated based on the previous hidden state and input of the current state. V, W and U are weight matrices, gh is the activation function, bh is the bias function and Ot is the output.

$$H_i = g_h(UX^i + VH^{(t-1)} + b_{h)}$$

Algorithm 1: Algorithm of finding particular risk of classes

1. The basic task of algorithm is to compute the posterior probability for a query variable given an observed event.

2. Let X =patients X1, X2.....Xn be the records of patient. A1,A2 ..Am be the attributes of that particular class of the diseases, where diseases D. l is the number of diseases, where i= 1 to l.

3. And where the numbers of diseases l of D are represented as the classes C.

4. The classifier needs to predict X belongs to the class C with the highest a probability, i.e., X is predicted to lie in the class Ci if and only if there exists Cj, such that:

```
5. For all $1 \leq j \leq l$ and $j \neq i$
        a. $P(Ci \mid X) > P(Cj \mid X)$
6. By Bayess theorem,
        b. $P(C_i) = P(X|C_j)P(C_j)/P(X)$   ......(1)
Where,
        $P(X|C_j) = \prod_{K=1}^{n} P(X_k|C_j)$
8. End.
```

here, we will get probability of particular class C of disease D out of patients X. With the help of symptoms of patients the accurate disease is predicted.

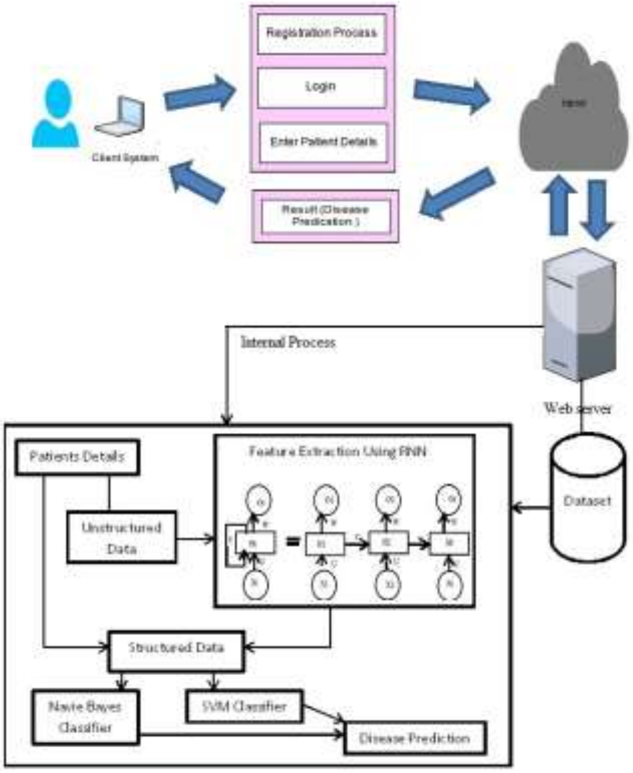

Fig -1: System Architecture

4. OUR CONTRIBUTION

Inquire the questions in the past made questions (recorded questions) to find the significant solutions. The pursuit question have in excess of answers, implies that it can find hopeful pool with numerous solutions. As a result of that physical torment to peruse all the hopeful pool answers and request to look one by one and pick the best one. To take care of this sort of issue, we give the rank answer competitors as match astute correlations. Specifically, it comprises of one disconnected learning segment and one online hunt segment. As disconnected we can discover the opinion in positive, negative and neutral to locate the correct rank of the appropriate responses and recommend best one in that. It gives the three sorts of preparing tests. In the online inquiry part, first gather a pool of answer contender for the given inquiry by means of discovering its comparative inquiries. At that point sort the appropriate response competitors by utilizing the disconnected prepared model to judge the inclination orders. The real time datasets are been used on the online and offline methodology.

Algorithm 2: Algorithm of Question Answering System

```
1. Order the training samples by the value of K(xi, x1)
   2 * K (xi , x) in ascending order.   // QA pair Dataset
2. MinErr = 1000; Value = 0.        // Threshold value
3. If the first k1 value are all from the same class c then checking Similarity of each question in data set.
4. Return c
5. End if
6. For all k do
7. Train SVM model on the first k training samples in the ordered list .
8. Classify x using the SVM model with equal error costs, get the result yp
9. Classify the same training samples using this model.
10. Fit the parameters A and B for the estimation of P( y == 1|yp ) where decision yp (-1,1)
11. ErrorNegative = P (y = 1 | yp );
      ErrorPositive = 1 - P(y == 1 | yp )
12. If ( ErrorPositive < MinErr then
13. MinErr = ErrorPositive ; Value = 1
14. End if
      If ErrorNegative _ C < MinErr then
      MinErr = ErrorNegative _ C; Value = - 1
15. End if
16. End for
17. Return best answer
```

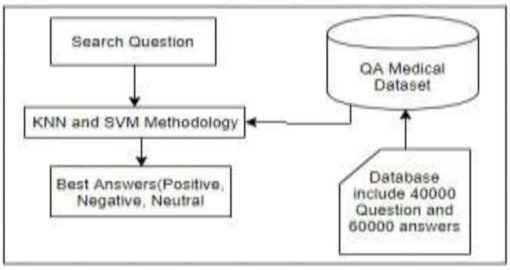
Search Question
QA Medical Dataset
KNN and SVM Methodology
Best Answers(Positive, Negative, Neutral
Database include 40000 Question and 60000 answers

Fig -2: Our Contribution

Table -1: Dataset

Hospital Data		
Category of Data	Items	Description
Structured Data	Details of Patients	Patient age, name, gender, etc.
	Habits	Whether patient drinks, smokes, etc.
	Examination Items	Symptoms, vision, etc.
	Diseases	Diabetes, Asthma, Cerebral Infraction and Hypertension.
Unstructured Data	Patients Details	Blood Pressure, Symptoms, etc.
		Prescriptions, Inquiry records.

Dataset from hospital which will be used for accurate prediction of diseases. This includes all attributes which will support in training and testing for predicting the accurate chronic disease of patients.

5. METHOD

Recurrent Neural Network (RNN) a traditional text classification works fundamentally center around three subjects: feature engineering, feature selection and using different types of machine learning algorithms. For feature engineering, the most generally utilized component is the sack of-words highlight. Likewise, some more mind boggling highlights have been outlined, for example, grammatical feature labels, thing phrases and tree portions. Feature selection goes for erasing noisy features and enhancing the order execution. The most well-known feature selection method is evacuating the stop words (e.g., "the"). Progressed approaches utilize data increase, shared data, or L1 regularization to choose valuable highlights. Machine learning algorithms frequently utilize classifiers, for example, Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machine (SVM) [6]. Be that as it may, these techniques have the information scarcity issue. Another model, which only exhibits a time complexity O(n), is the Recurrent Neural Network (RNN). This model investigates content word by word and stores the semantics of all the past content in a settled measured concealed layer. The benefit of RNN is the capacity to better catch the logical data. This could be gainful to catch semantics of long messages. Be that as it may, the RNN is a one-sided display, where later words are more predominant than prior words. In this way, it could decrease the adequacy when it is utilized to catch the semantics of an entirety report, since key segments could show up anyplace in a report as opposed to toward the end.

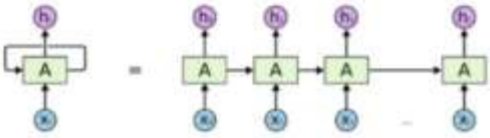

Fig -3: Block Diagram of RNN

6. EXPERIMENTAL RESULTS

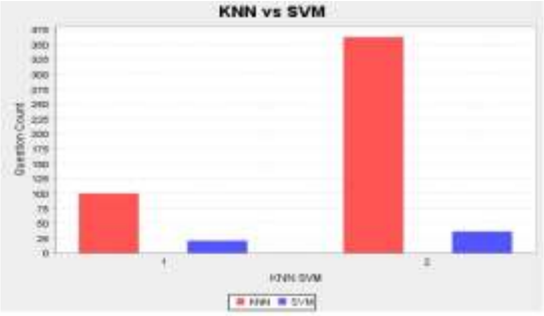

Fig -4: Bar chart of KNN vs SVM

Table -2: KNN vs SVM

Sr. No	KNN	SVM
1	100 Questions	21 Questions
2	362 Questions	36 Questions

Table -3: Accuracy Table

Sr. No	Accuracy
1	99
2	86
3	88
4	92
5	96
6	92

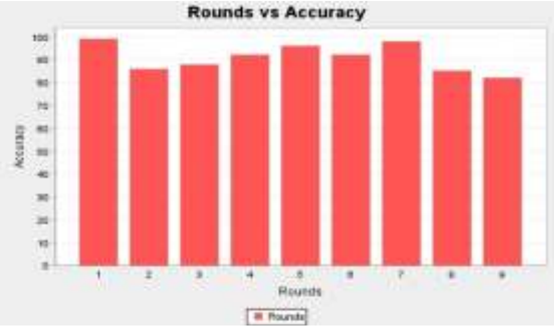

Fig -5: Bar chart of Rounds vs Accuracy

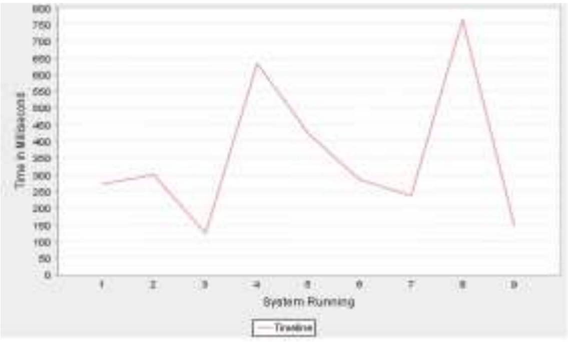

Fig -6: Structure Data disease prediction time graph

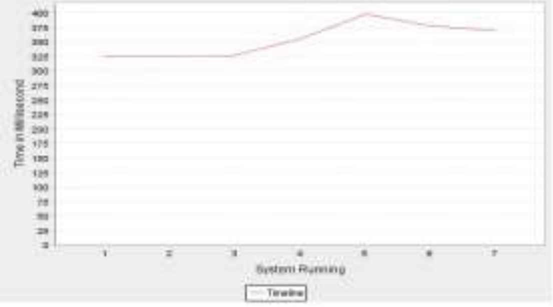

Fig -7: Unstructured Data disease prediction time graph

7. CONCLUSIONS

As chronic disease has increased, a new conventional neural network based multimodal disease risk prediction (CNNMDRP) algorithm in which structured and unstructured data from hospital is being used. In this structured and unstructured data, the personal information and detail history of the patient is being stored. In this CNN-MDRP both data are being used for predicting the chronic disease in that particular patient. In unstructured data patients may have missing data. So, the missing data of that particular patient can also retrieve through the genetic algorithm. The featured from unstructured data are been extracted correctly. Then the extracted features are structured data. Both Structured data and extracted structured data are used for predicting the exact chronic disease with Naive Bayes classifier and the SVM classifier. Community question answering system (CQA) is also proposed to help user to post the questions and answers related to the disease. To propose CQA system, KNN and SVM algorithms are used.

ACKNOWLEDGEMENT

I avail this opportunity to express my deep sense of gratitude and whole hearted thanks to my supervisor D. B. Hanchate for giving his valuable guidance, inspiration and encouragement to embark this paper.

The hospital management who helped with the details of patients and information of chronic diseases.

REFERENCES

[1] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang, ”Disease Prediction by Machine Learning over Big Data from Healthcare Communities”, IEEE transaction, 2017, pp 8869-8879.

[2] W. Yin and H. Schutze, ”Convolutional neural network for paraphrase identification”, in HLTNAACL, 2015, pp. 901-911.

[3] Seema sharma, Jitendra Agarwal, Shikha Agarwal, Sanjeev Sharma, “Machine Learning Techniques for Data Mining: A Survey, in Computational Intelligence and Computing Research”, IEEE International Conference on. IEEE, 2013, pp.1-6.

[4] Jensen PB, Jensen LJ, Brunak S, “Mining electronic health records: towards better research applications and clinical care”, Nat Rev Genet.2013 Jan; 14(1):75.

[5] L. Qiu, K. Gai, and M. Qiu, “Optimal big data sharing approach for tele-health in cloud computing, in Smart Cloud (Smart Cloud)”, IEEE International Conference on. IEEE, 2016, pp. 184-189.

[6] Siwei Lai,Xu Kang Liu,Jun Zhao, “Recurrent Convolutional Neural Networks for Text Classification”, in proceeding of the twenty-ninth AAAI Conference on Artificial Intelligence 2015.

[7] Xingyou Wang, Weijie Jiang, Zhiyong Luo, “Combination of Convolutional and Recurrent Neural Network for Sentimental Analysis of Short Texts”, International Conference on Computational Linguistics: technical papers, 2016, pg 2428-2437

[8] Dipak V.Patil, R.S. Bichkar, “Multiple Imputation of Missing Data with Genetic Algorithm based Techniques”, IJCA Special issue on Evolutionary Computation for Optimization Technique, 2010.

[9] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, “Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care”, in Proceedings of the 21th ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining. ACM, 2015.

[10] Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang, “Data-driven Answer Selection in Community QA Systems”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING JUNE 2016

[11] Kanchan M. Tarwani, Swathi Edem, “Survey on Recurrent Neural Network in Natural Language Processing”, International Journal of EngineeringTrends and Technology(IJETT), Volume 48 Number 6, June 2017.

[12] https://en.wikipedia.org/wiki/Recurrent neural network.

[13] Networks Dimitrios H. Mantzaris, George C. Anastassopoulos and Dimitrios K. Lymberopoulos, "Medical Disease Prediction Using Artificial Neural, BioInformatics and BioEngineering", 2008. BIBE 2008. 8th IEEE International Conference on 08 December 2008.

[14] Youn-Jung Son, Hong-Gee Kim, Eung-Hee Kim, Sangsup Choi, and Soo-Kyoung Lee, "Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients", Health Information Research, v.16, 2010.

[15] Alaa Elsayad and Mahmoud Fakhr, "Diagnosis of Cardiovascular Diseases with Bayesian Classifiers", Department of Computers and Systems, Electronics Research Institute, 12622 Bohoth St., Dokki, Geza, Egypt,2015.

BIOGRAPHIES

Smriti Singh received the B.E. degree in Computer Engineering from Pune University in 2015. Pursuing M.E degree in Computer Engineering from VPKBIET, Baramati- 413102 from Pune University.

Dr. Dinesh B Hanchate received degree of B.E. Comp. from Walchand College of Engg., Sangli (India), M. Tech. Computer from Dr. Babasaheb Ambedkar Technological University, Lonere (India). Ph.D. from Comp. Engg. Faculty at SGGSIET, Nanded and SRTMU, Nanded (India). Was HOD of Comp. and IT. Did STTP, QIP programs sponsored by IIT, Kanpur, AICTE, ISTE, SPPU and UG. Interest in Machine Learning, S/w Engineering, AI. , IR, Math Modelling, Usability Engg.