Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach

Article *in* International Journal of Innovative Technology and Exploring Engineering · August 2021

DOI: 10.35940/ijitee.I9364.0710921

5 authors, including:

Bhanuteja Talasila
VIT University
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Saipoornachand Kolli
VIT University
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Poonati Anudeep
VIT University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-9, July 2021

Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach

Talasila Bhanuteja, Kilaru Venkata Narendra Kumar, Kolli Sai Poornachand, Chennupati Ashish, Poonati Anudeep.

Abstract*: The turn of events and misuse of a few noticeable Data mining strategies in various genuine application regions (for example Trade, Medical management and Natural science) has induced the usage of such methods in Machine Learning (ML) constrains, to distinct helpful snippets of information of the predefined information in medical services networks, biomedical fields and so forth The exact examination of clinical data set advantages in early illness expectation, patient consideration and local area administrations. The methodology of Machine Learning (ML) has been effectively utilized in grouped technologies including Disease forecast. The objective of generating classifier framework utilizing Machine Learning (ML) models is to massively assist with addressing the well-being related issues by helping the doctors to foresee and analyze illnesses at a beginning phase. Sample information of 4920 patient's records determined to have 41 illnesses was chosen for examination. A reliant variable was made out of 41 sicknesses. 95 of 132 autonomous variables (symptoms) firmly identified with infections were chosen and advanced. This examination work completed shows the illness expectation framework created utilizing Machine learning calculations like Random Forest, Decision Tree Classifier and LightGBM. The paper confers the relative investigation of the consequences of the above-mentioned algorithms are utilized efficiently.

Keywords: Decision Tree model algorithm, Data Mining, Random Forest model Algorithm, Patient, Illness, LightGBM.

I. INTRODUCTION

Our medical care area every day gathers an enormous information worried about patients including clinical assessment, imperative boundaries, examination reports, therapy subsequent meet-ups, and drug choices and so forth Yet, tragically it isn't examine and mine in a suitable manner. It is taken care of either in record room as bunches of paper sheet or devouring hard circle space. The experts similarly as investigators are hasty stressed over this huge data. In government just as in certain associations the information is handling mostly by analysts at proficient level. The improvement of mechanized structures and their precision will oversee us in future. It will supportive in different illnesses the executives including viability of surgeries, clinical trials, drug, and the disclosure of connections among clinical and determination information to utilize Data Mining systems [3]. The medical care and clinical area are more needing data mining today. At the point when certain information mining techniques are utilized in a correct manner, significant data can be removed from enormous data set and which could guide the clinical professional to draw rapid choice and upgrade wellbeing administrations. The soul point is utilizing the grouping so that it can help doctor. Illnesses also good fitness associated issues such as intestinal sickness, Chickenpox, Migraine, Diabetes, Impetigo, Jaundice, dengue and so forth, tend to critical impact on person's wellbeing and at times may likewise prompt passing whenever disregarded. The medical services industry can settle on a powerful dynamic by "mining" the vast information base they have for example by removing the secret examples also, connections in the data set. Data mining models like Random Forest, Decision Tree and LightGBM, Algorithms or models can give a solution for the present circumstance. Subsequently, we have fostered a robotized framework that can find and separate secret information related with the infections from a historical (diseases-side effects) data set by the standard set of the individual Algorithms and models. The main objective of the paper is

•To characterize the illnesses by utilizing different calculations like Random Forest, LightGBM and Decision Tree.

•To track down the most affecting danger factors causing these illnesses.

•Comparison of different arrangement procedures and tracking down the best characterization strategy for the given information.

•To investigate the impact of change of one danger factor by another during the characterization (e.g., diabetes by hypertension, coronary illness, or smoking).

Manuscript received on July 17, 2021.
Revised Manuscript received on July 21, 2021.
Manuscript published on July 30, 2021.
* Correspondence Author
Talasila Bhanuteja*, School of Computer Science And Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: bhanuteja0018@gmail.com
Kilaru Venkata Narendra Kumar, School of Computer Science And Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: Kilaru.narendra@gmail.com
Kolli Sai Poornachand, School of Computer Science And Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: ksaipoornachand@gmail.com
Chennupati Ashish, School of Computer Science And Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: chennupatiashish@gmail.com
Poonati Anudeep, School of Computer Science And Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: iamanudeep6@gmail.com

II. LITERATURE SURVEY

As per a description by McKinsey [1], half of Americans have at least one of the ongoing illnesses, and 80% of American clinical consideration charge is exhausted on persistent illness treatment. With the improvement of expectations for everyday comforts, the rate of persistent sickness is expanding. The United States has spent a normal of 2.7 trillion USD yearly on persistent illness treatment. This sum involves 18% of the whole yearly GDP of the United States. The medical services issue of ongoing infections is additionally very significant in numerous different nations. In China, persistent sicknesses are the primary driver of death, as per a Chinese report on sustenance and ongoing infections in 2015, 86.6% of passing's are brought about by persistent sicknesses. In this manner, it is fundamental to perform hazard evaluations for persistent illnesses. With the development in clinical information [2], gathering electronic wellbeing records (EHR) is progressively helpful [3]. Moreover, [4] first introduced a bio-inspired superior heterogeneous vehicular telematics worldview, with the end goal that the assortment of versatile clients' health related continuous large information can be accomplished with the arrangement of cutting edge heterogeneous vehicular organizations. Chen et.al [5]– [7] suggested a clinical benefits system utilizing smart clothing for acceptable prosperity noticing. Qiu et al. [8] had inside and out examined the miscellaneous systems and attained the best results for cost reduction on tree and primitive path cases for heterogeneous structures. Patient's important genuine Data's, results of various tests and illness history are securely stored in the EHR, appealing us to find the more possible data based driven approach mechanism for decline the costs of clinically relevant examinations. Qiu et al. [9] described an effective flow surveying estimation for the tele-wellbeing cloud structure and arranged a information knowledge show for the PHR (Personal Health Record)- related spread system. Bates et al. [10] described six utilizes of huge amount of data in the field of clinical benefits. Qiu et al. [11] described a flawless huge information sharing calculation to deal with the muddle informational index in tele-health with cloud procedures. Here one of the main approaches is to find most-danger patients with the help of the information, so that it can be utilized to minimize the clinical price since most-peril patient's routinely need to do have expensive clinical benefits besides, in the first paper proposing medical services digital actual framework [12], it inventively presented the idea of forecast based medical services applications, including wellbeing hazard evaluation. Figure using standard disorder danger models customarily incorporates an AI estimation (e.g., determined backslide and backslide examination, etc), and mainly an oversaw learning computation by the utilization of planning information with names to set up the model [13], [14]. In the test dataset, patients can be broadly differentiated into social events of either more-peril or for the most part protected. These models are critical in clinical conditions and are by and large considered [15], [16]. Regardless, these plans have the going with qualities and deformations. The instructive assortment is pretty much nothing, for patients and ailments with certain conditions [17], the attributes are picked through knowledge. Nonetheless, these pre-chosen qualities possibly not fulfill the progressions in the sickness and its influencing factors. With the evolution of enormous information examination revolution, more consideration has been paid to infection expectation from the viewpoint of enormous information examination, distinct investigates have been led by selecting the attributes naturally from a large number of information to improve the precision of hazard classification [18], [19], instead of the recently selected attributes. Be that as it may, those progress work for the most part seen as organized information. For unstructured information, for instance, utilizing convolutional neural organization (CNN) to remove text attributes consequently has effectively drawn in wide consideration and furthermore proficient awesome conclusion.

III. METHODOLOGY

The dataset we have considered comprises of 132 indications, the blend or stages of which leads to 41 illnesses. In light of the 4920 documents of various patient samples, mainly to point foster a forecast algorithm that considers in the side effects of various client and forecasts the sickness that the person is bound to be affected.

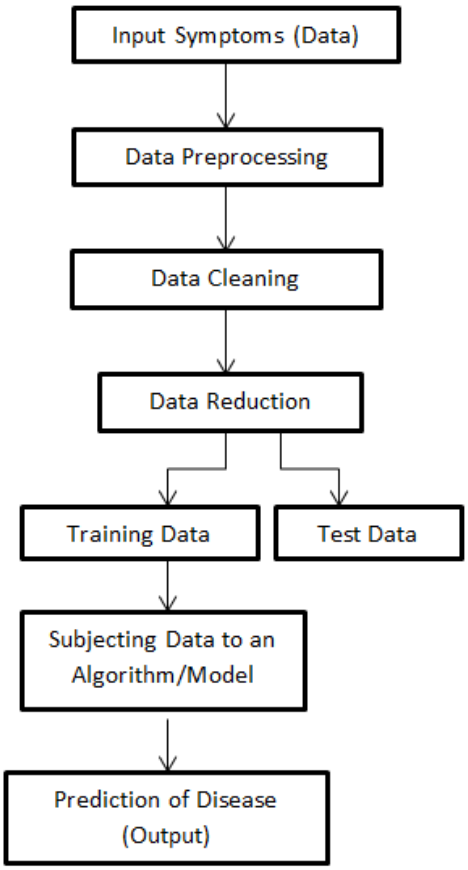

Fig1. Methodology Flowchart

A. Inputs (Patient Symptoms): When planning the algorithm, we have expected to be the client can has an unmistakable thought regarding the indications he is encountering. The Prediction created considers 95 manifestations in the midst of which the client can permit the indications his preparing as the input.

	Disease	Count of Disease Occurrence	Symptom
0	UMLS:C0020538_hypertensive disease	3363.0	UMLS:C0008031_pain chest
1	UMLS:C0020538_hypertensive disease	3363.0	UMLS:C0392680_shortness of breath
2	UMLS:C0020538_hypertensive disease	3363.0	UMLS:C0012833_dizziness
3	UMLS:C0020538_hypertensive disease	3363.0	UMLS:C0004093_asthenia
4	UMLS:C0020538_hypertensive disease	3363.0	UMLS:C0085639_fall

Fig2. Normal Input data

B. Data pre-processing: The mining of the data's approaches that changes the crude information or then again encrypts the information to form a structure so that it can be effectively deciphered with the help of calculation is known as information pre-processing. The information pre-processing strategies utilized in the introduced work which listed as follows:

Data Purification: Data is purified using certain measures like stuffing in lost worth, along these lines settling the irregularities in the information.

Data Reduction: The examination turns out to be hard when managing gigantic information base. Thus, we kill those autonomous variables (symptoms) which may not affect the objective variables (diseases). So that in the progress task, which of around 95 of 132 side effects firmly identified with the illnesses are chosen.

	disease	symptom	occurence_count
0	hypertensive disease	shortness of breath	3363.0
1	hypertensive disease	dizziness	3363.0
2	hypertensive disease	asthenia	3363.0
3	hypertensive disease	fall	3363.0
4	hypertensive disease	syncope	3363.0

Fig3. Preprocessed Data

C. Models: The entire system is designed in such a way to predict the diseases by utilizing the three Algorithms i.e., *Decision* Tree model, LightGBM model and Random Forest classifier model, so that the predictive analysis study is proposed at the end of the study by exploring its speed, efficiency and performance of the various algorithms for the input dataset.

D. Output(diseases): While a framework is made with the preparation set utilizing the validated calculations standard datasets are shaped and whenever the client indications are provided as a contribution as input of the algorithm, and the side effects are composed agreeing as the standard dataset created, accordingly creating arrangements and foreseeing the high probable infection.

IV. IMPLEMENTATION

The Disease forecast framework is executed utilizing the three information mining calculations for example Random Forest, Decision tree classifier and LightGBM. The portrayal and implementation of the calculations are provided beneath.

A. Decision Tree Model:

The order of the algorithm worked as Decision tree look like the model of many branches in a tree. So, by analyzing the arrangement of unequivocal assuming at that point rules on highlight esteems (manifestations for our situation), it classifies down the dataset into more modest and more modest subsets that outcomes in anticipating an objective value (disease). A tree comprises of the mainly a decision Node and a Leaf Node.

Decision Node: It has a minimum of 2 branches. In our analysis we introduced, every one of the manifestations are taken as decision node.

Leaf Node: Constitutes the order which denotes that, the decision may of any of the branch. So that the diseases here represent to a leaf node.

1. ID3 Technique:

The core center calculations are utilized in the project is mainly ID3 calculation created by Quinlan J.R. ID3 utilizes an up down, voracious pursuit based on the provided sections, where each column (Attributes = Symptoms) at each hub is tried and finds the variable (symptoms) that is core for grouping of a provided dataset. So, to pick the particular side effect is perfect to construct a Decision tree model, the technique ID3 utilizes information gain and entropy. In this way, continuing similarly the whole Decision tree model will be built prompting malaria and dengue at the result.ID3 adheres to a standard: A division which has entropy of zero as a leaf hub. A branch that's entropy is more prominent than zero necessities parting. So, in the event that has absurd to expect to accomplish zero entropy, finally the core Decision is done by the strategy for a basic dominant part.

Limitation: At the point when every one of the 132 indications was considered from the first dataset rather than 95 side effects, it prompted Overfitting. For example, the tree appears to remember the dataset given and thus neglects to characterize the new information. Thus just 95 manifestations were considered picking the upgraded ones during information cleaning step.

B. LightGBM:

LightGBM [25] is a rapid, dispersed, higher impulse boosting technique which is dependent on Decision tree model output, which is used for pointing the position, arrangement and various other ML techniques. Basically, LightGBM is a group technique that consolidates the expectations of different Decision trees (by adding them together) to make the last forecast that sums up well. Significantly, LightGBM trains the numerous tree models in an added substance way, with each new tree model being prepared to foresee the residuals (i.e., mistakes) of the earlier models. LightGBM is an execution of slope boosting Decision tree (GBDT) [26].

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

When preparing every individual Decision tree (f) and parting the information, there are two select procedures LightGBM utilized: inclination based one-side examining (GOSS) [25] and leaf-wise development. GOSS intends to handle the computational intricacy issue of traditional executions of GBDT, which need to go through each element of each information moment that figuring the data acquire for every one of the potential parts. The significant perception behind GOSS is that information cases with bigger slopes assume larger parts in data acquire calculation. Consequently, while assessing the best split, GOSS keeps information cases with huge angles and arbitrarily tests information with little inclinations. This technique has been demonstrated viable and work quicker than regular ones. Leaf-wise development is an effective technique for developing trees. It discovers the leaf with the biggest parting acquire from every one of the current leaves each time, and afterward parts the leaf, and circles this interaction. At the end of the day, it will pick the leaf with max delta misfortune to develop. Contrasted and level-wise development methodology, leaf-wise one can decrease more mistakes and get better exactness under similar dividing times. The hindrance of leaf-wise technique is that it might develop trees profoundly and lead to over fitting.

C. Random Forest Classifier:

The Random Forest classifier is an adaptable, simple to utilize AI calculation that gives remarkable outcomes more often than not applied without any hyper tuning. So, as validated in the Decision tree model, the notable restriction of tree calculation is overfitting. So, it shows up as though the decision tree has remembered core of the information.

This model forestalls this issue: That It's a form of troupe investigation. Troupe investigation alludes to utilizing different calculations or identical calculation on numerous occasions. Random Forest model is a group of many decision trees. Also, more noteworthy the quantity of these trees in this model is the fitter of the speculation.

All the more decisively, the random Forest fills in as listed below:

1. Fix the k side effects from data (clinical data) the sum of m manifestations arbitrarily (here k << m). At that point, it assembles a Decision tree model with the help of side effects of k.

2. Rehashes n number of times with the goal that we have n number of tree model worked from various Random mixes of indications which is denoted as 'k' (or an alternate irregular example of information, known as bootstrap test).

3. Consider every one of the various n-constructed trees and proceeds a variable which is random to foresee the illness. Here it Store the anticipated illness, so that we can have a sum of n illness anticipated from n number of the decision tree model.

4. Computes the decisions in favor of each anticipated illness and consider the mode (which is most continuous illness anticipated) as last expectation from the Random Forest model calculation.

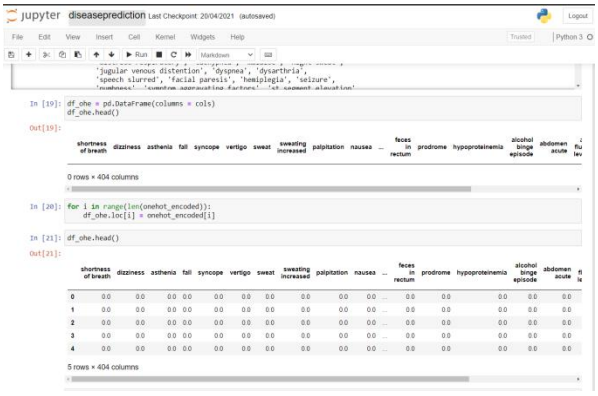

Fig4. Coding Snapshot

```
Pred: Alzheimer's disease
Actual: Alzheimer's disease

Pred: HIV
Actual: HIV

Pred: Pneumocystis carinii pneumonia
Actual: Pneumocystis carinii pneumonia

Pred: accident cerebrovascular
Actual: accident cerebrovascular

Pred: acquired immuno-deficiency syndrome
Actual: acquired immuno-deficiency syndrome

Pred: adenocarcinoma
Actual: adenocarcinoma
```

Fig5. Prediction Vs Actual

V. RESULTS

The framework was prepared on clinical record of 4920 patients inclined to 41 illnesses that was due to the of the mix of different indications. We have scrutinized 95 indications out of 132 side effects to keep away from overfitting.

From these outcomes, we can construe that every one of the three Models function admirably on the dataset. Be that as it may that the Random Forest is working somewhat better when contrasted with the other two Models. The Accuracy score of every model is given below:

Algorithm used	Accuracy score
Decision Tree	0.973154
LightGBM	0.973154
Random Forest	0.98315

When the indications are given, the Models are to be chosen. As the Models are chosen, the indications are handled, and the infection is looked through dependent on the standard set that has been characterized in the Methodology segment.

Fig6. UI Home page for Disease input

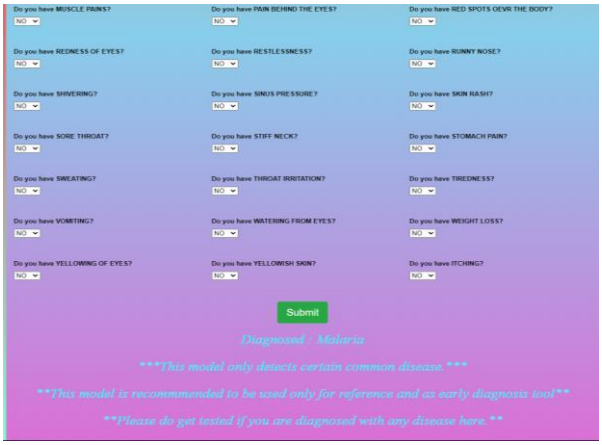

Fig7. Malaria detection

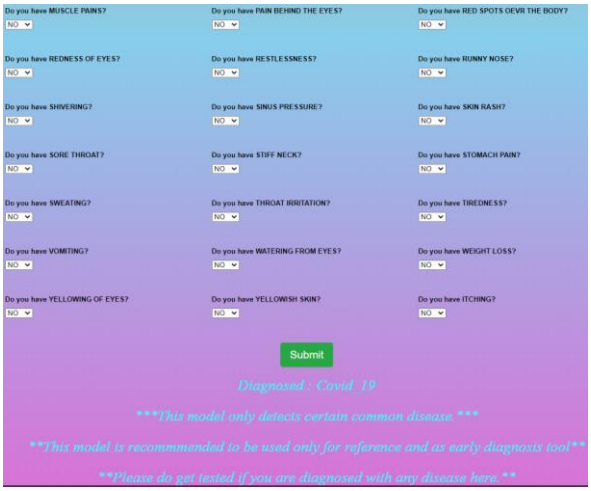

Fig8. Covid19 detection

VI. CONCLUSION

Finally, from the recorded advancement of (ML) Machine Learning technique and the approaches in clinical area, it very well may be shown that frameworks and systems have been arisen that has empowered refined information investigation by basic and direct utilization of Machine Learning (ML) models. This paper brings an extensive near investigation of three models' execution of a clinical record with each of the obtaining accuracy score up to 98 %. Finally, the paper is investigated with disarray lattice & precision value. Man-made consciousness will be assumed significantly more significant part in information investigation later on because of the accessibility of gigantic information created and put away by the cutting-edge innovation.

REFERENCES

1. S. Mitra, S.K.Pal & Mitra , P., Data mining in soft computing framework: A survey, IEEE transactions on neural networks, 13(1), 314,2018.
2. Krzysztof J. Cios, G.William Moore, Uniqueness of medical data mining, Artificial Intelligence in Medicine 26, 1–24, 2017.
3. Parvez Ahmad, Saqib Qamar, Syed QasimAfser Rizvi, Techniques of Data Mining in Healthcare : A Review, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15, June 2017.
4. Hsinchun Chen, Sherrilynne, S. Fuller, Carol Friedman and William Hersh, Knowledge Management, Data Mining and text mining inmedical informatics.
5. V. krishnaiah, G. Narsimha, & N. Subhash Chandra, A study on clinical prediction using Data Mining techniques, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1, 239 248, March 2017.
6. Divya Tomar and Sonali Agarwal , A survey on data mining approaches for healthcare, International Journal of Bio-Science and Bio-Technology Vol.No.5, pp. 241-266, 2017.7. Mohammed Abdul Khalid, Sateesh kumar Pradhan, G.N.Dash, F.A.Mazarbhuiya, A survey of data mining techniques on medical data for finding temporally frequent diseases", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2018.
7. S.D.Gheware, A.S.Kejkar, S.M.Tondare, Data Mining: Task, Tools, Techniques and Applications, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2017.
8. Yongjian Fu , Data Mining : Tasks, Techniques and Applications http://academic.csuohio.edu/fuy/Pub/pot97.pdf
9. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2017.
10. G. Beller, J. Nucl. Cardiol. "The rising cost of health care in the United States: is it making the United States globally noncompetitive?" vol. 15, no. 4, pp. 481-482, 2018.
11. Pang-Ning Tan, Michael Steinbach ,Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2016.
12. Gosain, A.; Kumar, A., "Analysis of health care data using different data mining techniques," Intelligent Agent & Multi-Agent Systems, 2017. IAMA 2009, International Conference on, vol. no., pp.1,6, 22-24 July 2018.
13. Dr. M.H.Dunham, "Data Mining, Introductory and Advanced Topics", Prentice Hall, 2017.
14. A. S. Elmaghraby, et al. Data Mining from multimedia patient records. 6, 2017.
15. Nada Lavrac, BlažZupan, "Data Mining in Medicine" in Data Mining and Knowledge Discovery Handbook, 2018.
16. Soni J, Ansari U, Sharma D, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2018.
17. Naren Ramakrishnan, David Hanauer, Benjamin J. Keller, Mining Electronic Health Records, IEEE Computer 43(10): 77-81, 2018.
18. O. Mary K, Mat, "Applications of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, August 2017.
19. Hian Chye K, Gerald T, Data mining applications in healthcare, Journal of healthcare information management: JHIM.19 (2): 64-72, (2016).
20. A. Milley, "Healthcare and data mining", Health Management Technology, vol. 21, no. 8, pp. 44-47, 2017.

21. Gaynes R, Richards C, Edwards J, et al. Feeding back surveillance data to prevent hospital-acquired infections. Emerg Infect Dis 2001;7:2 95298, 2017.
22. Brosette SE, Spragre AP, Jones WT, Moser SA. A data mining system for infection control surveillance. Methods Inf Med,39: 303-310, 2018.
23. M. Ridinger, “American Healthways uses SAS to improve patient care”, DM Review, vol. 12, no.139, 2018.
24. M. Durairaj, V.Ranjani, Data mining applications in healthcare sector: A Study, International Journal Of Scientific & Technology Research Volume 2, Issue 10, ISSN 2277-8616, October 2016.
25. Anonymous. Texas Medicaid Fraud and Abuse Detection System recovers $2.2 million, wins national award. Health Management Technology, vol. 20, no. 10, 2017.

AUTHORS PROFILE

Talasila Bhanuteja, pursued his Bachelor of Technology (B. Tech) in computer science Engineering from Vellore Institute of Technology, Vellore. He is currently working as Software Engineer at Bank of America continuum India Pvt. Ltd. He previously presented a paper at the 3rd International Conference on Intelligent Computing, Information and Control Systems (ICICCS 2021) Organized by CARE College of Engineering, Trichy, India. His area of research lies in the domain of Machine Learning, Cloud Computing and Internet of Things.

Kilaru Venkata Narendra Kumar, pursued his Bachelor of Technology (B. Tech) in computer science Engineering from Vellore Institute of Technology, Vellore.

Kolli Sai Poornachand, pursued his Bachelor of Technology (B. Tech) in computer science Engineering from Vellore Institute of Technology, Vellore. He is currently working as Android Developer at Digital Horizons technology and media services Pvt. Ltd

Chennupati Ashish, pursued his Bachelor of Technology (B. Tech) in computer science Engineering from Vellore Institute of Technology, Vellore. He is currently working as Associate consultant at Ernst & Young LLP. He has wide range of experience in machine learning, mobile app development and completed internships in those fields. Apart from these he also worked on Internet of things and built prototype models at the time of his graduation.

Poonati Anudeep, pursued his Bachelor of Technology (B. Tech) in computer science Engineering from Vellore Institute of Technology, Vellore. He is currently working as Project Engineer at Wipro Limited