

A Novel Decision Model to Support the Prediction of Asthma Among the Big Data's of Various Different Patient's

Abhinav Hans¹, Navdeep Singh² and Sheetal Kalra³

^{1,2}Department of CSE, GNDU Regional Campus, Jalandhar
abhinavhans@gmail.com

³Department of CSE, GNDU Regional Campus, Jalandhar
sheetal.kalra@gmail.com

Abstract

The capability of distributed computing for overriding the requirements for sending different frameworks for running a server based administrations raised a progressive change in the way the conventional requests of the individuals utilization to be taken care of. Distributed computing gives the rental support of the client in which a client can utilize the specific programming by paying for that on the cloud server. Since the entire situation is useful for enormous businesses like facebook, google, orkut and so on, different fields are likewise getting reliant on distributed computing. Since huge amounts of information is transferring consistently to the cloud server does should be dug appropriately for productive information stockpiling. In this paper, we attempt to coordinate the information preprocessing strategy with an information grouping method to mine enormous information's of asthma based patients. We have utilized simulation tool called eclipse to run the Programming interface's of weak and cloudsim for setting up the trial environment.

Keywords: Data pre- processing, Data Classification, Random Fores, Big Data, Data Mining.

1. Introduction

The innovation requesting less assets and with preeminent yield dependably draws in the client. So the Distributed computing is the particular case that gives the same and lures the clients and dealing with one's own servers [1]. Despite the fact that there are numerous celebrated advancements like remote sensor systems, adhoc network, but distributed computing is the most acclaimed innovation amongst all. The property by giving the product's on lease permits the cloud to beat alternate advancements. The advantage of prepaid administration of distributed computing permitted the assets to be open by any kind of the client anyplace and at whatever time. Different components like versatility, low recuperation cost, less support, colossal information stockpiling procurements, quick arrangement and numerous more elements make cloud the most intense methodology. Since the cloud is related with the data innovation, as well as with numerous different fields that in the human wellbeing, deals and administration documents as well. By checking the current states of the patient by utilizing specific body sensor arrange the healing center costs can be evaded [2].

Separating a truly valuable information from different quantities of information sources that gives important data by isolating examples, images, traits and so forth is known as information mining. Information mining is a multi-disciplinary field of substantial database in which any obliged information can be brought and utilized by client's necessity. Various stages like Information comprehension, Information readiness, Demonstrating, Assessment, and Arrangement are the exceptional piece of information mining. Different online networking sites are absolutely reliant on the information mining process as the measure of information that gets transferred each and every hour is in huge amounts of

terabytes so it makes data mining process to be very much significant. Although not any social media websites, but many different sections like hospitals, IT industries, online shopping *etc.* are also bringing data mining as a significant approach in their data maintenance job.

1.1 Basic Models of Cloud Computing

The cloud system consists of three service models based on the basis of resource requirement, *i.e.* SaaS, PaaS and IaaS [14]. Various cloud computing models that provide the facility to the user as required are discussed below.

- ❖ *SAAS (Software as a service)*: Software as a service provides the software on lease *i.e.* pay per use service of software on a cloud server.
- ❖ *IAAS (Infrastructure as a service)*: it provides the particular infrastructure to cloud services by means of virtual machines and other hardware requirements.
- ❖ *PAAS (Platform as a service)*: To run the application on the cloud there must be an surroundings where this service must run. Therefore the marketers cater the platform where the operating system, web server, programming language execution environment is provided.
- ❖ *SECAAS (Security as a service)*: Many confidential data that user tries to hide from various internet threats, security as a service provides a huge help by providing a protocol based security on the cloud.

1.2 Types of Asthma

Asthma maladies analyze is absolutely subject to the sound that is made by the patient amid cough. A non asthmatic patient when hacks delivers a sound of recurrence (206(14) Hz) where as an asthmatic patient on hack creates a sound with recurrence (239(19)Hz). There are different sound recording gadgets now a days that records the sound and demonstrates the recurrence of the sound from which it can be unmistakably analyze whether the sound is of a typical individual or of an asthmatic patient. On the premise of bronchial excessive touchiness the asthma has the accompanying sorts: hypersensitive asthma (atopic, extraneous, brought on by immunologic jolt of an antigen), natural (non-unfavourably susceptible, impelled by contamination, physically or artificially), practice affected, medication instigated asthma, occupative asthma and asthmatic bronchitis [3]. There are distinctive sort of sounds as indicated by diverse hypotheses. As indicated by the before American Thoracic Culture, sounds are viewed as "persistent" if their length of time is longer than 250 ms; else they are viewed as "intermittent" [10]. Sharp persistent sounds (predominant recurrence over 400 Hz) and rhonchi as low-pitched consistent sounds (overwhelming recurrence of 200 Hz or less) is considered as wheeze as indicated by the ATS.

Be that as it may, as per the new meaning of CORSA (Automated Respiratory Sound Examination) guide- lines, the prevailing recurrence of wheeze is as a rule over 100 Hz and the span more noteworthy than 100ms [10]. Wheezes are nonstop extrinsic sounds, which are superimposed on typical breath sounds and regularly connected with bronchial aviation route impediment. There are numerous circumstances prompting wheezing which incorporate all instruments narrowing aviation route bore, for example, bronchospasm, mucosal edema, outer pressure by a tumour mass, or element aviation route obstacle [14]. Asthmas unfavourable impact is as per the side effects a patient endures. Despite the fact that asthma can be characterized into four stages on the premise of indications:

1.2.1 Intermittent

Patient suffers light cough and wheezing for less than twice per week and at night less than twice per month.

1.2.2 Mild Persistent

Patient gets an asthma attack at least once in a week. Shortening of breath, heavy cough, wheezing, chest tightness occurs.

1.2.3 Moderate Persistence

The big air passageway of the lungs is affected by this, heavy coughing in time slots and wheezing.

1.2.4 Severe Persistent

Continues episodes occur all day and night time for several days, persistent cough and wheeze.

2. Related Work

Asthma is the most climb malady now days in which age variable doesn't matters *i.e.* it can be in Individuals of any age bunch. In [11] creator has displayed stepwise the foundation of asthma in therapeutic terms took after by data about the Pathology and manifestations later. After that creator has highlighted some the downsides of the current methods for overseeing asthma by underlining on demonstrating the critical ailment administration procedures in the conventional way. A tele-checking method on float ways to asthma is finished.

Asthma is a serious infection and can be exceptionally hurtful impacts on the off chance that its not considered important. By taking its not kidding effect on wellbeing, a nonstop observing is must to check the body and breath conduct of the patient. The most countable variable is nature in which they relax. So In [12] creator proposes an improvement of a standard based asthma framework. So as per it,the patients are given different proposals on the conceivable outcomes of happening an asthma assault ,as indicated by understanding's present body conditions and nature in which they relax. The framework is in view of the scrutinizing procedure to the patient and answer given by patients characterizes the quiet's available wellbeing condition and the natural condition in which they are living with.

This examination work puts light on the information mining strategy in which diabetes infection can be anticipated on the premise of the medicinal record history of the patient. Diabetes is an exceptionally basic ailment that can happen in any age bunch. It is a genuine ailment that has a genuine effect on heart, kidneys, sensory system, bloodline and vessels. However, mining the information of diabetes patient in a productive way is a basic issue. The creator had gathered the information from different patients either having diabetes or diabetes free. For information mining method adjusted J48 so its exactness rate can be expanded. Creator utilized MATLAB for performing the reproduction work with weka as a Programming interface to concentrate the different exploratory results [10].

In [7], prescient model of soil fruitfulness has been clarified in diverse steps. In this paper A strategy of the choice trees calculation in information mining is utilized to anticipate the fruitfulness of soil took after by execution tuning of J48 choice tree calculation with the assistance of meta-procedures, for example, property determination and boosting.

For arrangement of information and items, the strategy of choice tree in used to get profitable results. Furthermore, these outcomes can be utilized for examination and future forecast. In [8] paper the creator made a target to present the improved choice tree calculation that groups the information. In his work ID3, J48, NBTree are utilized as the tree classifiers.Then the comparative examination is done on the premise of parameters like effectiveness and execution new improved choice tree calculation (NEDTA).

Table 1. Definition and Procedure of Existing Data Mining Algorithms

Algorithms	Definition and procedure
J48	J48 target variable prediction rules are formed by the algorithm. <ul style="list-style-type: none"> • Uses top down approach by divide and conquer technique and form tree. • Test attributes are selected by some measures and divide and conquer is applied until no sample leaf is left
ID3	One of the decision tree algorithm called iterative dichotomiser. <ul style="list-style-type: none"> • Uses greedy approach for creating tree by top down approach. • Selection of attributes which classifies the data at its best on each node and keeps on following this at every node till the tree is not formed is done.
NBTree	Naïve Baseyan classication and decision tree algorithm learning together forms NBtree. <ul style="list-style-type: none"> • Each node is selected and naïve baseyan algorithm is applied,which classifies the instances. • The naïve baseyan tree is constructed for each leaf.

3. Proposed Approach

There are different quantities of methodologies that chips away at asthma patients yet are not sufficiently competent to conquer the different issues of information missing qualities and order issues. So in our proposed methodology we attempt to coordinate the information preprocessing methodology and the arrangement way to deal with fabricated an effective setup for information mining on the asthma patients. So as to complete the examination of information mining on a dataset of asthma patients, we have gotten the information from different wellbeing assets that give the information on the premise of the reviews they had made. In our information set the quantities of properties utilized are 11 and the aggregate quantities of cases are 1776 of distinctive asthmatic and non-asthmatic patients.

Table 2. Various Health Attributes with Description and Domain Used for Building Asthma Dataset

S.No.	Attribute	Description
1	Age	Age of the patients in years.
2	Gender	Gender of the patient whether Male or Female(M/F) (0/1)
3	Begin	Start of the patient record,according to the hospital records
4	End	Ending of the patient record,according to the hospital records
5	Current wheeze	Level of the wheezes (%)
6	Symptoms of severe wheezes	Level of severe wheezes after the current wheezes situation (%)
7	Alcohol consumption	Whether a patient consume alcohol or not(scales 1-5)
8	Drug intake	Whether a person is on any drugs or not (Y/N)(0/1)
9	Physical activity status	Whether a person is physically active or not(Y/N) (0/1)
10	Smoking habits	Whether a person smokes or not(scales 1-5)
11	Hereditary	Whether a person has any previous family asthmatic symptoms or not(Y/N)(0/1)

Data pre-processing: At the point when an information is transferred to the server there are ordinarily the information having missing values in them, which can come about into bogus presumptions of different asthma patients. In this manner the missing qualities should be dealt with for which we attempt to present the missing quality calculation for an information preprocessing procedure known as resampling. Resampling is the procedure in which the given dataset is standardized separated from the class property in standardized interims of time. Albeit in our exploratory setup there are no missing qualities among 1776 properties of dataset, but still when we apply information preprocessing

Calculation the different results come that shrewdly sets the information into different graphical representations. Figure 1 demonstrates the discovery of different asthmatic and non-asthmatic patients, taking into account distinctive examples. The blue line symbolizes the asthmatic patient and the red one the non-asthmatic patient. In every chart the mix of red and blue segment tells the number of patients, these are asthmatic on the premise of the single quality. Therefore, every chart demonstrates its own worth because of the distinctive quality esteem in it. Information preprocessing permits the clients to preprocess the information before characterization which is done above. According to these graphical representations of different traits and occasions, taking after table can be inferred which further mines the information shrewdly.

Data classification: The significant part of the proposed methodology is the information grouping system which arranges the information through information characterization calculation into different diverse classes makes it simple to concentrate different valuable data out of it. The order calculation that we are utilizing here is the Irregular woodland arrangement calculation. The proposed calculation is in view of developing the tree of irregular measure of estimations of the occasions.

Table 3. Mathematical Values Based Result Derived After Data Mining

Attribute	Minimum value	Maximum value	Mean	StdDev	Distinct values
Age	9	81	33.214	11.705	54
Gender	0	1	0.576	0.494	2
Begin	0	599	266.207	189.578	546
End	1	600	340.15	188.133	56
Current wheeze	1.1	32	9.512	6.497	75
Symptoms of severe wheezes	1.1	28	6.806	4.673	65
Alcohol consumption	0	5	2.139	1.718	6
Drug intake	0	1	0.495	0.5	2
Physical activity status	0	1	0.878	0.328	2
Smoking habits	0	5	1.718	1.589	6
Hereditary	0	1	0.131	0.337	2

The random forest algorithm is tested on the data set of 1776 instances which perform much better than the other classification algorithms. The performance result of the random forest algorithm is 100%, which is the highest among all other classification algorithms. The confusion matrix of the algorithm tells us about the performance and classification capability of the random forest.

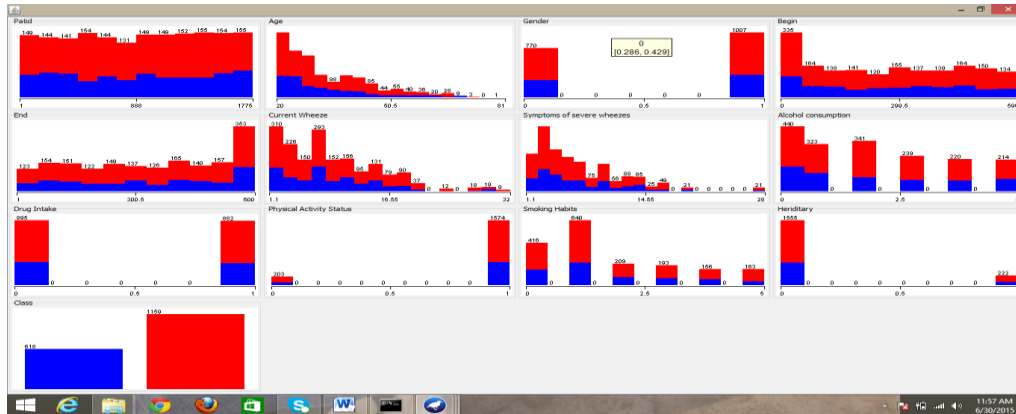


Table 4. Confusion Matrix of the Amount of Instances Correctly and Incorrectly Classified

a	b	←Classified as
628	0	a=Asthmatic
0	1148	b=Non-Asthmatic

The confusion matrix above of random forest classification algorithm tells that out of 628 values of 'a' *i.e.* asthmatic patients 0 are the wrongly classified values and out of 1148 values of 'b' *i.e.* non-asthmatic patients 0 are the wrongly classified instances, therefore the performance of the random forest algorithm is quite high. The precision of the calculation is numbered with the execution rate,

as well as with different variables too should be in any way excluded and of which the expense/advantage qualities and edge qualities play an essential role. To be a successful calculation the two qualities must be conversely corresponding to one another that implies the limit quality is higher than the expense esteem which must be sufficiently low to suit the entire qualities. In figure 2 a graphical correlation is being finished with the limit values and expense/advantage values by looking at an edge cure and expense/advantage cure with one another which Cleary demonstrates the connection between them. The bends demonstrate that among the aggregate of 1776 (100%) estimations of the datasets 1148(64.64%) are asthmatic and 628(35.36%) are thought to be non-asthmatic patients.

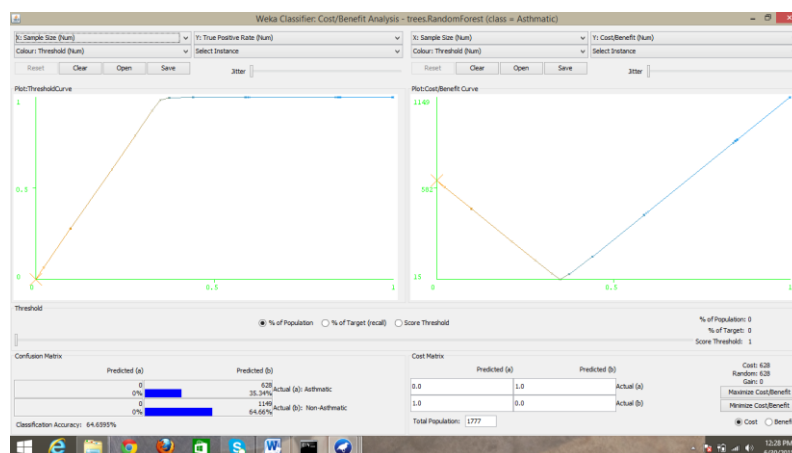


Figure 2 Comparison between threshold Curve and Cost/Benefit Curve of Asthmatic and Non-Asthmatic Patients

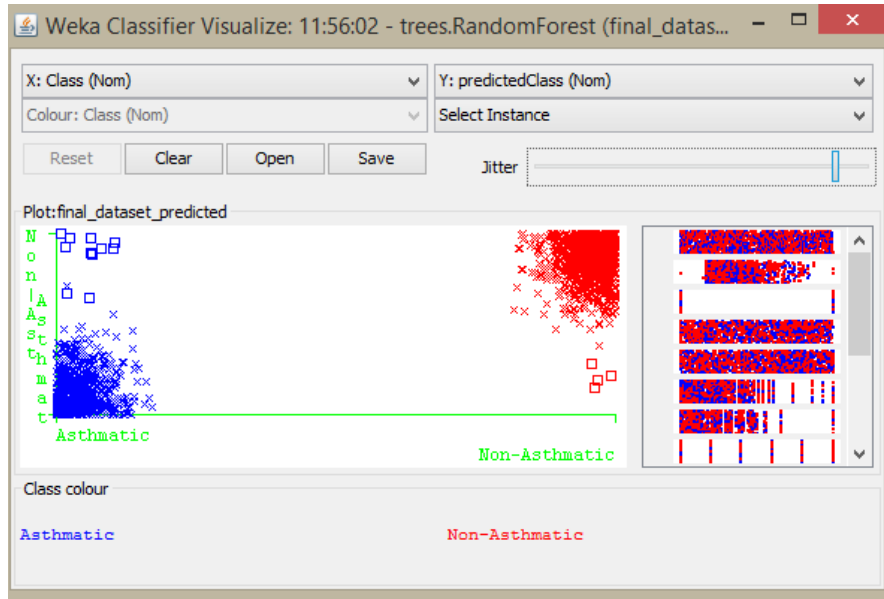


Figure 3. Graphical View of Number of Total Patients Classified as Asthmatic and Non-Asthmatic

To ascertain the execution, the nature of the calculation, the calculation depends on different QOS parameters that make a nearer examination with different other existing arrangement calculations to make the proposed approach as predominant among every one of them.

Table 5. QOS Based Comparative Analysis of Various Data Classification Algorithms with the Proposed Approach

Algorithms	Correctly Classified Instances %	Incorrectly Classified Instances %	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error (in%)	Root relative squared error(in %)	Total Number of Instance	Time to build
AD Tree	66.3851	33.6149	0.1196	0.4639	0.4732	101.4724	98.9847	1776	0.11
BF tree	64.6396	35.3604	0	0.4571	0.4781	99.9895	100	1776	0.93
Decision Stump	64.6959	35.3041	0.0021	0.4556	0.4773	99.6622	99.8362	1776	0.03
J48	69.1441	30.8559	0.1837	0.4114	0.4535	89.9855	94.8657	1776	0.07
J48graft	69.1441	30.8559	0.1837	0.4114	0.4535	89.9855	94.8657	1776	0.15
LAD Tree	66.1599	33.8401	0.0669	0.44	0.4678	96.234	97.8579	1776	0.19

NB Tree	64.6396	35.3604	0	0.4572	0.478 1	100	100	1776	0.18
Random Forest	99.1554	0.8446	0.981 5	0.1672	0.211 2	36.564 2	44.1744	1776	0.23
REP Tree	69.1441	30.8559	0.194 9	0.4085	0.451 9	89.354 3	94.5284	1776	0.07
Simple Cart	64.6396	35.3604	0	0.4571	0.478 1	99.989 5	100	1776	0.05

4. Conclusion

In this paper, we have proposed an incorporated methodology for mining the information of asthma patients for which standardize methodology is utilized for information preprocessing and irregular woodland approach for information grouping. We have made a trial setup for our calculation to execute and distinctive execution charts and information mined qualities have been taken after information preprocessing and grouping of information. Towards the end, we made a relative examination of random forest tree with numerous other grouping methodologies and presume that the random forest tree is the main calculation whose ability of ordering the cases accurately is 100% with no wrongly grouped occurrences and zero lapse rates. The CPU use is likewise low as the time taken to mine the entire information set is less *i.e.* 0.02 seconds which is minimum among all calculations.

References

- [1] A. Rosenthal, P. Mork, M. H. Li, J. Stanford, D. Koester and P. Reynolds, "Cloud computing: A new business paradigm for biomedical information sharing", *Journal of Biomedical Informatics*, vol. 43, (2010), pp. 342–353.
- [2] H. Xia, I. Asif and X. Zhao, "Cloud-ECG for real time ECG monitoring and analysis" computer methods and programs in biomedicine, vol. 110, (2013), pp. 253-259.
- [3] J. W. Dexheimer, T. J. Abramo, D. H. Arnold, M. P. H. Kevin Johnson, M. S. Y. S. Fei Ye Kang-Hsien Fan Neal Patel and M. S. D. Aronsky, "Implementation and Evaluation of an Integrated Computerized Asthma Management System in a Pediatric Emergency Department: A Randomized Clinical Trial", *International Journal of Medical Informatics*.
- [4] S. Pandeya, W. Voorsluys, S. Niua, A. Khandokerb and R. Buyyaa, "An autonomic cloud environment for hosting ECG data analysis services", *Future Generation Computer Systems*, vol. 28, (2012), pp. 147–154.
- [5] V. Vaithyanathan, K. Rajeswari, K. Tajane and R. Pitale, "Comparison Of Different Classification Techniques Using Different Datasets", *International Journal of Advances in Engineering & Technology*, May (2013).
- [6] A. Goyal and R. Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", *International Journal of Applied Engineering Research*, ISSN 0973-4562.
- [7] J. Gholap, "Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility".
- [8] H. Kaur and H. Kaur, "Classification of data using New Enhanced Decision Tree Algorithm", *International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)*.
- [9] T. R. Patil and S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", *International Journal Of Computer Science And Applications*, ISSN: 0974-1011, vol. 6, no. 2, April (2013).
- [10] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", *International Journal of Computer Applications (0975 – 8887)*, vol. 98, no. 22, July (2014).
- [11] D. Oletic, "Wireless sensor networks in monitoring of asthma", *International convention mipro* vol. 34, (2011).
- [12] G. K. Nee, M. A. Syafiq, S. K. Sugathan, C. Y. Yie and E. A. P. Akhir, "The Development of a Rule-based Asthma System", *Information Technology (ITSim)*, 2010 International Symposium in Date, June 15-17, (2010).

- [13] M. A. khassaweneh, S. B. Mustafa and F. A. Ekteish, "Asthma Attack Monitoring and Diagnosis: A Proposed System", Biomedical Engineering and Sciences (IECBES), 2012 IEEE EMBS Conference on Date, December 17-19, (2012).
- [14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing and Management, vol. 45, (2009), pp. 427–437.
- [15] S. Rietveld, M. Oud, E.H. Dooijes "Classification of asthmatic breath sounds: preliminary results of the classifying capacity of human examiners versus artificial neural network" Computers and Biomedical Research 32 (1999) 440–448.
- [16] R. Bozorgmanesh a_, M. Otadi b, A. A. Safe Kordi, F. Zabihi and M. B. Ahmadi, "Lagrange Two-Dimensional Interpolation Method for Modeling Nanoparticle Formation During RESS Process", Int. J. Industrial Mathematics, vol. 1, no. 2, (2009), pp. 175-181.
- [17] C. E. Sabel, W. Kihal, D. Bard and C. Weber, "Creation of synthetic homogeneous neighborhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France", Social Science & Medicine.
- [18] T. Pham and M. Wagner, "Ambiguity reduction in speaker identification by the relaxation labeling process", Pattern Recognition, vol. 32, no. 7, (1999), pp. 1249–1254.
- [19] A. Rosenthal, P. Mork, M. H. Li, J. Stanford, D. Koester and P. Reynolds, "Cloud computing: A new business paradigm for biomedical information sharing", Journal of Biomedical Informatics, vol. 43, (2010), pp. 342–353.
- [20] H. Xia, I. Asif and X. Zhao, "Cloud-ECG for real time ECG monitoring and analysis", computer methods and programs in biomedicine, vol. 110, (2013), pp. 253-259.
- [21] J. W. Dexheimer, T. J. Abramo, D. H. Arnold, M. P. H. K. Johnson, M. S. Y. S. F. Ye Kang-Hsien Fan Neal Patel and M. S. D. Aronsky, "Implementation and Evaluation of an Integrated Computerized Asthma Management System in a Pediatric Emergency Department: A Randomized Clinical Trial", International Journal of Medical Informatics.
- [22] S. Pandeya, W. Voorsluys, S. Niua, A. Khandokerb and R. Buyyaa, "An autonomic cloud environment for hosting ECG data analysis services", Future Generation Computer Systems, vol. 28, (2012), pp. 147–154.
- [23] V. Vaithyanathan, K. Rajeswari, K. Tajane and R. Pitale, "Comparison Of Different Classification Techniques Using Different Datasets", International Journal of Advances in Engineering & Technology, May (2013).
- [24] A. Goyal and R. Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", International Journal of Applied Engineering Research, ISSN 0973-4562.
- [25] J. Gholap, "Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility".
- [26] H. Kaur and H. Kaur, "Classification of data using New Enhanced Decision Tree Algorithm", International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS).
- [27] T. R. Patil and S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal Of Computer Science And Applications, ISSN: 0974-1011, vol. 6, no. 2, April (2013).
- [28] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887), vol. 98, no. 22, July (2014).
- [29] D. Oletic, "Wireless sensor networks in monitoring of asthma", International convention mipro, vol. 34, (2011).
- [30] G. K. Nee, M. A. Syafiq, S. K. Sugathan, C. Y. Yie and E. A. P. Akhir, "The Development of a Rule-based Asthma System", Information Technology (ITSim), 2010 International Symposium in Date, June 15-17, (2010).
- [31] M. A. khassaweneh, S. B. Mustafa and F. A. Ekteish, "Asthma Attack Monitoring and Diagnosis: A Proposed System", Biomedical Engineering and Sciences (IECBES), 2012 IEEE EMBS Conference on Date, December 17-19, (2012).
- [32] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing and Management, vol. 45, (2009), pp. 427–437.
- [33] S. Rietveld, M. Oud and E. H. Dooijes, "Classification of asthmatic breath sounds: preliminary results of the classifying capacity of human examiners versus artificial neural network", Computers and Biomedical Research, vol. 32, (1999), pp. 440–448.
- [34] R. Bozorgmanesh, M. Otadi, A. A. Safe Kordi, F. Zabihi and M. B. Ahmadi, "Lagrange Two-Dimensional Interpolation Method for Modeling Nanoparticle Formation During RESS Process", Int. J. Industrial Mathematics, vol. 1, no. 2, (2009), pp. 175-181.
- [35] C. E. Sabel, W. Kihal, D. Bard and C. Weber, "Creation of synthetic homogeneous neighborhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France", Social Science & Medicine.
- [36] T. Pham and M. Wagner, "Ambiguity reduction in speaker identification by the relaxation labeling process", Pattern Recognition, vol. 32, no. 7, (1999), pp. 1249–1254.

Authors



Abhinav Hans, He was born on 15-09-1990. He completed his B.Tech in Computer science and engineering from Lovely Professional University, Phagwara, Punjab, India in the year of 2013. He is pursuing his M. Tech in Computer Science and Engineering from Guru Nanak Dev University, RC, Jalandhar. His field of interest are cloud computing, biomedical, big data, wireless sensor network, body area network. Till now he has various number of publications out of which much of them are in IEEEXPLORE and International Journals.