International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

Efficient Feature Selection for Congenital Lungs Disorder in Human Fetus

K. Vimala, D. Usha

Abstract: ***Genetic disorders are one of major challenge in medical field, which is to be overcome in early stage, such that the patients can be diagnosed as soon as possible. This paper deals with fetal lungs disorder, as the initial process of the research proceeds with high dimensional dataset, where the fetal lungs dataset are preprocessed to check missing data or null criteria. Feature selection plays a major role here, where the denoised data are given to Principal component analysis, since the dataset was large in size; it was required to reduce the volume of data. Principal component analysis helps to reduce the redundancy in the data. The feature selection also provides minimum number of features, which is a pathway for performing the classification. Principal component analysis overcomes unrelated feature problem, increases the prediction accuracy level and decreases the computational overheads in classification. Efficiency of the feature selection is estimated using standard classification metrics.***

Keywords: Genetic disorder, Lungs development, Preprocessing, Principal Component Analysis.

I. INTRODUCTION

This paper deals with fetus lungs development, where most of the complications are diagnosed in the initial stage itself. Due to some heredity reasons the embryonic babies get affected, namely food habits, environmental changes and gene may be some of the reasons today. There are chances where parent may be alcoholic or smoker which may also effect the development of the fetal Lungs. Periodic clinical checkup will provide proper pathway for lungs development in embryonic stage itself. [2]The Lungs development is alienated into two phases, First phase deals with lungs growth and second phase is lungs maturation. [3]A lungs growth depends on physical accepts; such that it develops the structure of the lungs. [3]Lungs maturation involves functional development; these functionalities are achieved by biochemical process. Lungs growth is developed throughout the gestation period. As the final development of human fetus, a lung provides exchange of gases and increase in number of alveolar.

The full formed lungs organ is roughly 50-100m^2, the main function of the lungs provides exchange of oxygen and carbon dioxide.

II. IPHASE OF LUNGS DEVELOPMENT IN FETUS

The structural lungs development has [2]five gestation periods, that are given as Embryonic Phase, Pseudo-Glandular Phase, Canalicular Phase, Saccular Phase, Alveolar Phase.

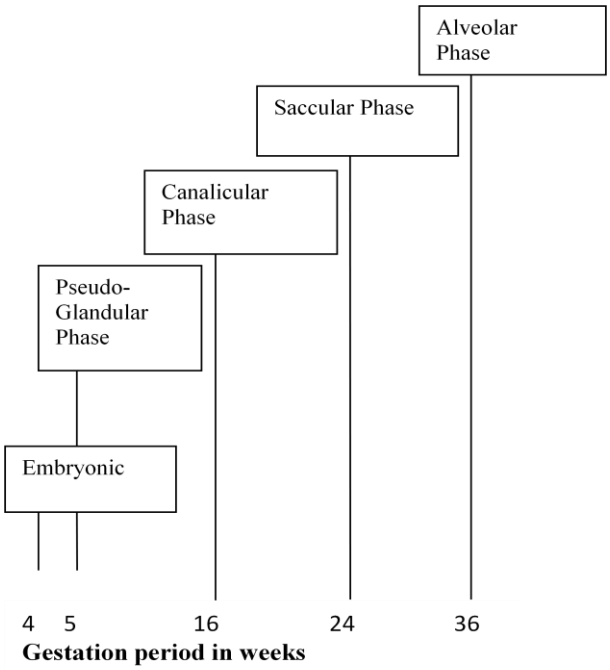

Fig 1. Lungs development phases

A. Embryonic Phase:

- The embryonic phase of the fetal lung develops at 4-5 weeks of conception period.
- The small formations of left and right lungs are seen.

B. Pseudo-Glandular Phase:

- The pseudo-glandular stage of fetal lung development begins at the 5- 17th week of gestational period,
- Branching have been continued to form terminal bronchioles

C. Canalicular Phase:

- The canalicular stage of fetal lung development begins approximately the 16-26th week of the conception period.
- During the canalicular phase, blood supply is given to respiratory system with the help of oxygen. Tissues are developed in this phase.

D. Saccular Phase:

- The fetus is in the saccular stage of the lung development at an approximately 36th week of conception period.
- Terminal bronchioles are formed here, where the passage of air flow is provided here.

E. Alveolar Phase:

- The alveolar phase is the last phase of fetal lung development,
- The fully developed lungs are grown for the baby.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

K.Vimala*, Research Scholar, Computer Science Department, Mother Teresa Women's University,vimala.arivalagan@gmail.com

Dr. D. Usha, Assistant Professor, Computer Science Department, Mother Teresa Women's University,

III. METHODOLOGY

High dimensionality fetal lungs dataset has been take for the work process. Initial process involves preprocessing phase, where missing data are identified. Since the data are, high in dimensionality, it is required to reduce the dataset and provide desired feature selection for the lungs development in the fetus dataset. The selected features are useful to detect any abnormalities in the fetal lungs which may be using for the doctors to provide diagnosis to the patient.

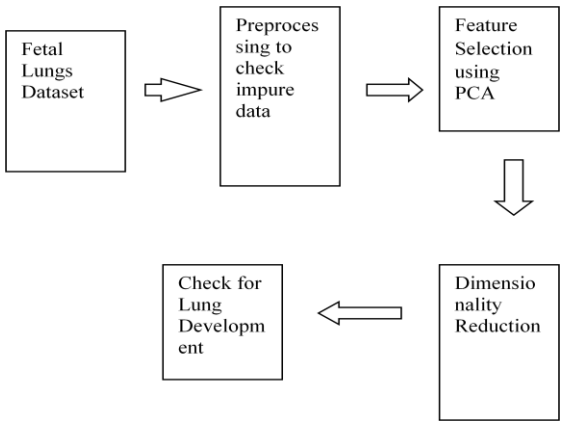

Fig 2. Work flow of Methodology

IV. DATABASE

High dimensionality dataset has been chosen for the research work. Gene expression omnibus is referred for fetal lungs dataset in homosapiens. Gene expression omnibus is a curated and publicly available dataset; it also supports minimum information about a micro array experiment. It consists of 54675 records and 38 samples of fetal lungs development.

V. PREPROCESSING

- The raw data involves noise in it, which is to be removed, before data to be processed
- Here in this data, noise detected is missing values in the gene information which is to be avoided
- Null criteria are overcome, and particular recorded is not taking for further processing.

VI. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis, or PCA, helps in feature selection process, in this paper the PCA overcomes the redundancy in the gene dataset and removes the irrelevant features in the dataset. This algorithm helps in improving accuracy and reduces large dataset to smaller one; with minimum features, Principal component analysis is easier to explore and fast to analysis for machine learning algorithm.

A. PCA PROCESS FLOW

Initial process involves preprocessed data. The extracted pure data are given to principal component analysis. Standard matrix is found by two steps, first step involves finding mean value for gene value and second step is to subtract mean value from each gene value.

Standard matrix = Gene value(X) – Mean Gene value (X') ..(1)

Covariance matrix is performed to check how y value changes with respect to x value. If x value increases then y value also increases automatically. Resultant matrix is a square matrix.

$$Cov\,(X,Y) = \sum_{i=1}^{n} \frac{(X-X')(Y-Y')}{n-1} \quad \ldots (2)$$

Eigen Values can be calculated only for square matrix

$$Eigen\ values = Cov - \lambda I \quad ..(3)$$

Cov is the covariance matrix
I is the identity matrix
λ is Eigen values

Eigen vector is calculated for Eigen value. Highest Eigen value of Eigen vector are known as PCA (Data are reduced here)

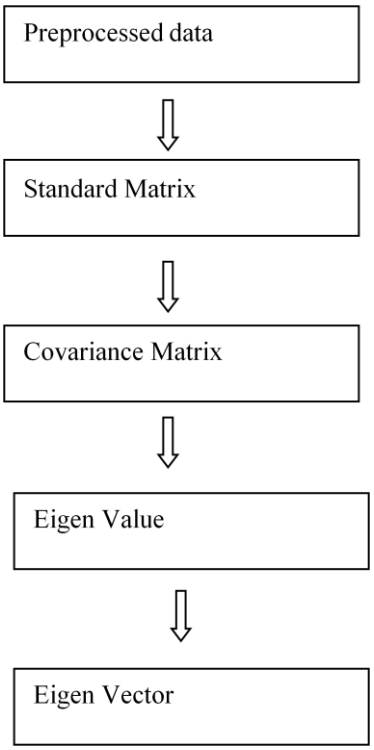

Fig 3. PCA flow

B. ALGORITHM

Input: Extracted data (Ext_{dt})
Output: Set of selected Features.
#Applying PCA to the Data
Step 1: Initialize all the variables.
 # Find the mean of gene data
Step 2: Mean→
Step 3: Reading file
Variable d, out.
Step 4: For (double d: entry)
Out += d/entry.len.
Step 5: End For
Step 6: return out
#Find the standard matrix with help of mean value
Step 7: Standard Matrix→

Step 8: Variable sum=0, aM=mean (a), bM=mean (b), dv=a.len-1.
Step 9: For →i to out.len
Step 10: For → j to out.len
sum += (a (i)-aM) * (b(i)-bM)
Step 11: If sum=0.0 then sum = val
Step 12: End If, For (j)
Step 14: return sum / dv
#Find the covariance matrix(square matrix)
Step 14: Covariance Matrix→
Step 15: Variable out = [mat.leng][mat.leng]
Step 16: For → I to out.len
Step 17: For → J to out.len
Variable dtA = mat[i], dtB = mat[j]
out[i][j]=cov(dtA,dtB)
Step 18: End For (i) (j)
Step 19: return out.
#Find Eigen Value and Eigen set
Step 20: Eigen Set→
Variable cpy = input, q = [cpy.leng][cpy.leng]
Step 21: For → I to q.leng
q[i][j]=1
Step 22: End For
Step 23: bolean done = false
Step 24: Reading Done
Variable nMat=mul(q[i],q[j])
Step 25: If nMat – copy > 0.0000001
copy = nMat.
Step 26: End If, For,While
Variable dat = covMat();
Step 27: return eigen
Variable val = {eigen.values}
 # Find PCA
Step 28: Eigen vectors and Values
Step 29: For I to vals[0].length
Variable j
Step 30: If double.isNAN(vals[i][j])
Variable vals[j][i] = 1.0
k = transporse *inttransporse.
lin=substring of k
Step 31: For I to length(k)
tt = substring (k)
print tt.
Step 32: End For
Step 33: End While
Step 34:Featured Data Achieved.

VII. RESULTS AND DISCUSSION

The high dimensionality dataset consist of fetal lungs development where it consists of 26177 normal data and 28498 abnormal dataset. The total features of 54675 X 40 are reduced to 54675 X 4. Principal component analysis gives best dimensionality reduction.

Table I. Fetal Lungs Information

Dataset	Sample	Class	Type	Total records	Gestation period
Fetal Lungs Dataset	38	2	txt	54675	53-154 days

The table I, gives the details of the dataset based on sample, class, record type, total records and conception period of the lungs fetal dataset.

Table II. Experimental design and value distribution

Algorithm	Accuracy
Rough PCA	85.38
Emprical Distribution	83.78
USQR	79.33
CMIM	86.44
Hybrid Feature Selection	87.37
Proposed	90.85

The table II, gives the experiment result of the dataset used, where the accuracy level is increased in the proposed method. Total of four features are selected for classification process. The Fig 4, shows the representation of algorithm efficiency.

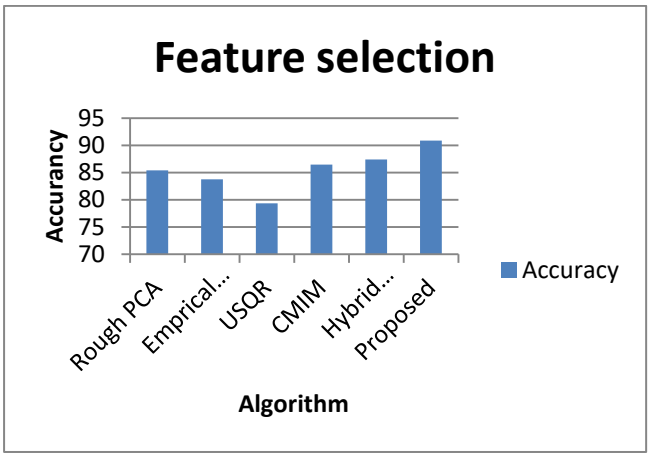

Fig 4. Accuracy of Feature Selection

The rank and value are evaluated for the each sample of 38 along with gestation period in Fig 5.

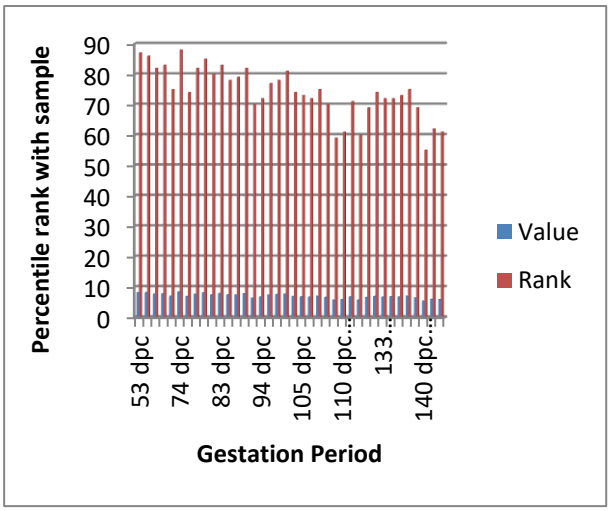

Fig 5. Rank and Value for Gestation period

VIII. CONCLUSION

- The feature selection is successfully performed in high dimensional dataset.
- Accuracy around 90.85% was been achieved, while compared to other algorithms, the proposed algorithm gives better result.
- Further process includes classification using machine learning algorithm such as support vector machine and random forest algorithm.
- The accuracy level of feature selection [7] can be improved by using different other methods based on filter, wrapped and embedded feature selection methods.

REFERENCE

1. Alan H. Jobe, "An Unknown Lung Growth and Development after Very Preterm Birth", American Journal of Respiratory and Critical Care Medicine. Vol 166. No. 12, pp 1529–1536, 2002
2. S. Ranganathan, "Lung development, lung growth and the future of respiratory medicine**"**, European Respiratory Journal, Vol. 36 No. 4, pp 716-717, 2010
3. Maryanne E. Ardini-Poleske, "LungMAP: The Molecular Atlas of Lung Development Program", American journal of physiology,Vol.313, pp.733-740,2017
4. Linnette D. Miff, "Extra-Alveolar Vessels and Edema Development in Excised Dog Lungs", Circulation Research Vol.28, No.5, pp. 524-532, 2019
5. A E Bonner, W J Lemon, M You, "Gene expression signatures identify novel regulatory pathways during murine lung development: implications for lung tumorigenesis", Journal of Medical Genetics, Vol.40, No.6, pp.408-417,2019
6. Sayak Ganguli, Manoj Kumar Gupta," Feature Extraction from Gene Expression Data Files", International Journal of Engineering Inventions, Vol. 1, No. 2, pp:15-16, 2012.
7. S. Karthik, M. Sudha,"A Survey on Machine Learning Approaches in Gene Expression Classification in Modelling Computational Diagnostic System for Complex Diseases", International Journal of Engineering and Advanced Technology,Vol.8, No.2, pp.182-191, 2018
8. Suyan Tian, Chi Wang, **"**Feature Selection for Longitudinal Data by Using Sign Averages to Summarize Gene Expression Values over Time", Hindawi BioMed Research International, Vol. 2019, Article ID 1724898, 2019.
9. Sen Lianga, Anjun Mab,c, Sen Yanga, Yan Wanga,QinMa "A Review of Matched-pairs Feature Selection Methods for GeneExpression Data Analysis", Computational and Structural Biotechnology Journal, Vol.16, pp.88-97, 2018.
10. Megan Crow, Nathaniel Lim, Sara Ballouz, "Predictability of human differential gene expression", Proceeding of National Academy of Sciences of the United States of America, Vol. 116,No. 13,6491–6500,2019.
11. F. Amiri, M. RezaeiYousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," Journal of Network and Computer. Application., vol. 34, no. 4, pp. 1184– 1199,2011.
12. C. R. Marshall, A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, D. Pinto, Y. Ren et al., "Structural variation of chromosomes in autism spectrum disorder," The American Journal of Human Genetics, vol. 82, no. 2, pp. 477–488, 2008
13. L. Yu, H. Liu, "Efficient feature selection via analysis of relevance and redundancy", Journal of Machine Learning Research,pp.1205–1224, 2015.
14. M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," Machine learning, vol. 53, no. 1-2, pp. 23–69, 2003.
15. S. Maldonado, R. Weber, "A wrapper method for feature selection using support vector machines", Informatics and Computer Science Intelligent Systems Applications, 179, 2208–2217, 2009.
16. Jianzhong Wang, Shuang Zhou, "An Improved Feature Selection Based on Effective Range for Classification", Scientific World Journal, Vol. 2014, Article ID 972125, pp.1-8.
17. Jerzy Krawczuk, Tomasz,"The feature selection bias problem in relation to high-dimensional gene data,Artificial Intelligence in Medicine, Vol.66, pp.63-71,2016
18. Elyasigomari, V., Mirjafari, M.S., Screen, H.R. and Shaheed, M.H. "Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization. Applied Soft Computing",vol. 35, pp.43-51,2015.
19. Gentleman, R., Carey, V., Huber, W., Irizarry, R. and Dudoit, S. eds., ." Bioinformatics and computational biology solutions using R and Bioconductor. Springer Science & Business Media, 2006
20. Gour, D.K., Jain, Y.K. and Pandey, G.S.. The Classification of Cancer Gene using Hybrid Method of Machine Learning. International Journal of Advanced Research in Computer Science, Vol 2, pp.149-153, 2011
21. Whitsett JA, Weaver TE. Hydrophobic surfactant proteins in lung function and disease. "The New England Journal of Medicine" 347:2141-48, 2002.
22. Harding R, Hooper SB. Regulation of lung expansion and lung growth before birth. Journal of Applied Physiology 81:209-24, 1996.
23. Gregory GA, Kitterman JA, Phibbs RH, Tooley WH, Hamilton WK. Treatment of the idiopathic respiratory distress syndrome with continuous positive airway pressure. "The New England Journal of Medicine. 284:1333-40, 1971.
24. P.K Kumarsan**,** Feature Selection Clustering for Gene Data, International Journal of Emerging Research in Management and Technology, Vol.6,No.9, pp.183-188, 2017
25. JenniferG.Dy, CarlaE.Brodley,"Feature Selection for Unsupervised Learning", Journal of Machine Learning Research Vol.5, pp.845–889, 2004

AUTHORS PROFILE

Mrs. K.Vimala, received M.E degree in Computer Science at National Engineering College, Tuticorin. Currently, Research Scholar, pursuing Ph.D at Mother Teresa Women's University. Areas of interest are Datamining, Machine Learning. Have 7 years of teaching experience, Lifetime membership of ISTE and CSI

Dr. D. Usha, is currently as Assistant Professor in the department of Computer Science, Mother Teresa Women's University. Have 14 years of teaching cum research experience. Research expertise at Wireless sensor network, Datamining, Network security