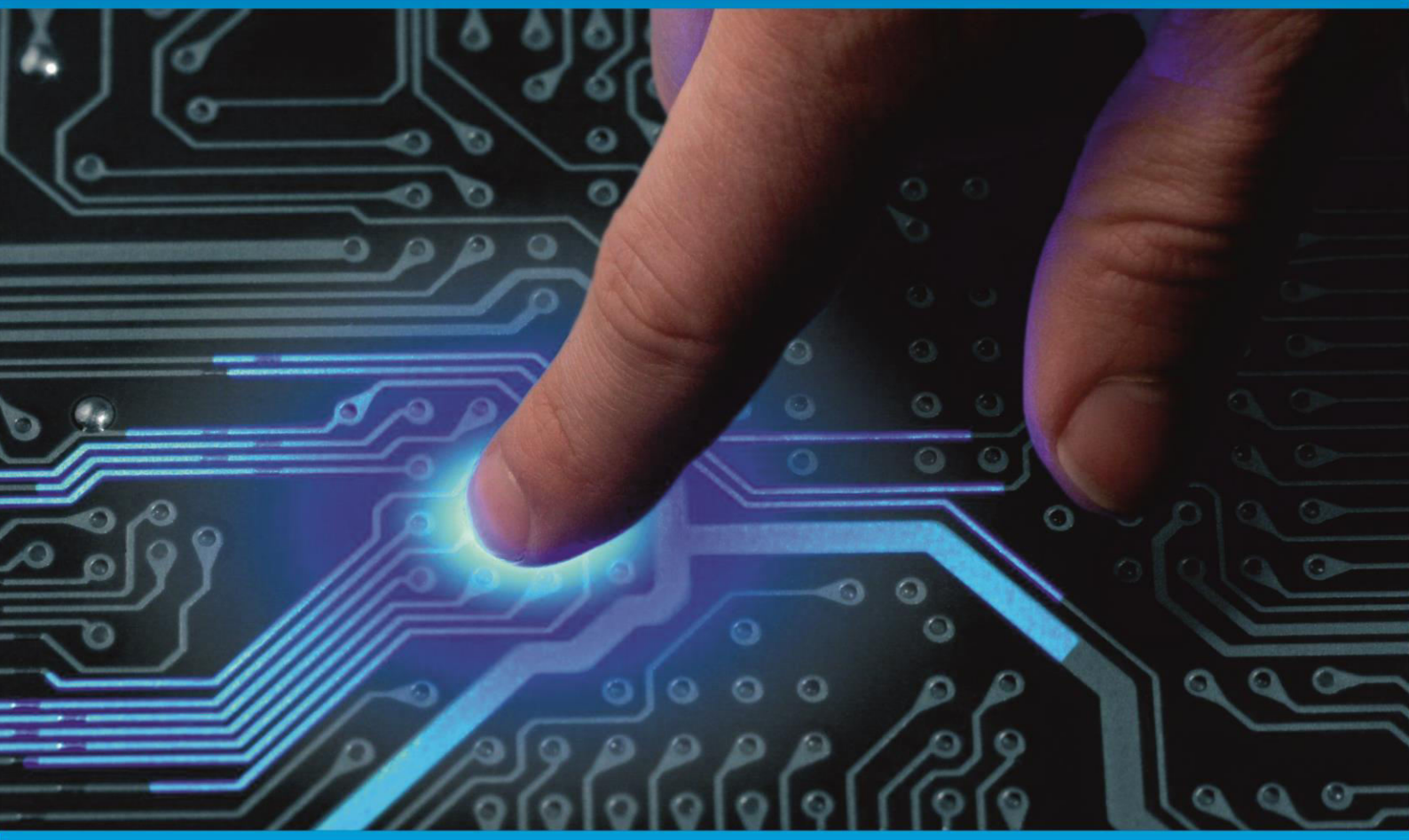
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 2, February 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

Effective Pattern Generation for Disease Prediction

Leena Patil[1], Kanchan Doke[2]

P.G. Student, Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India[1]

Associate Professor, Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India[2]

ABSTRACT: This system will give disease prediction for symptoms given by patient. The paper present system thatproposes a novel Discriminative Pattern-based Prediction framework (DPPred) to accomplish the prediction tasks by taking their advantages of both effectiveness and interpretability. Specifically, DPPred adopts the concise discriminative patterns that are on the prefix paths from the root to leaf nodes in the tree-based models. DPPred selects a limited number of the useful discriminative patterns by searching for the most effective pattern combination to fitgeneralized linear models.

KEYWORDS: Discriminative Pattern, Generalized Linear Model, Tree-based Models, Classification

I.INTRODUCTION

I.I BACKGROUND:

In a hospital health care monitoring system, it is necessary to constantly monitor the patient's physiologicalparameters. For example, a Brain tumor parameter such as headaches and weakness and vomiting and loss of awareness or a partial or total loss of consciousness. Pattern-based models have been given in the last decade toconstruct high-order patterns from the large set of features, including association rule-based methods on categorical data and frequent pattern-based algorithms on text data. To address the above challenges, in this paper, the systemproposes a novel **discriminative patterns-based learning framework** (**DPPred**) that extracts a concise set of discriminative patterns from high-order interactions among features for accurate classification and regression.

I.II. MOTIVATION:

Many pattern-based models have been proposed in the last decade to construct high-order patterns from the large setof features, including association rule-based methods on categorical data and frequent pattern-based algorithms on textdata, and graph data. Recently, a novel series of models, the discriminative pattern-based models have demonstratedtheir advantages over the traditional models. They prune non-discriminative patterns from the whole set of frequent patterns, however, the number of discriminative patterns used in their classification or regression models are still huge.

Traditional frequent pattern mining works on categorical data and item set data, in which discretization is required to deal with continuous variables. Instead of roughly discretizing the numerical values, this system adopts the thresholdingBoolean function in DPPred.

II. RELATED WORK

Efficient approximate inference techniques based on variation methods and an EM algorithm for empirical Bayesparameter estimation. System report results in document modelling, text classification, and collaborative filtering,comparing to a mixture of unigrams model and the probabilistic LSI model. There are two ways to deal with the balanced data classification problem using random forest. One is based on cost sensitive learning, and the other is based on a sampling technique. This system conducts a systematic exploration of frequent pattern-based classification,and provides solid reasons supporting this methodology. It was well known that feature combinations

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

(patterns) couldcapture more underlying semantics than single features. However, inclusion of infrequent patterns may not significantlyimprove the accuracy due to their limited predictive power. The algorithm discovers the top -k covering rule groups foreach row of gene expression profiles. Several experiments on real bioinformatics datasets show that the new top-k covering rule mining algorithm is orders of magnitude faster than previous association rule mining algorithms.
Furthermore, system proposes a new classification method RCBT. The algorithm discovers the top -k covering rulegroups for each row of gene expression profiles. Several experiments on real bioinformatics datasets show that the new
top-k covering rule mining algorithm is orders of magnitude faster than previous association rule mining algorithms.
Furthermore, system proposes a new classification method RCBT.

The following table shows related work on the basis of advantages, disadvantages, and methodology:

Table 1: Related work

Sr.No	Author, Title and Journal Name	Advantages	Disadvantage	Propose Points
1	D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. JMLR, 3:993–1022, 2003.	Useful in document modelling, text classification.	Need Exploration of such that involve Dirichlet-multinomial over trigrams instead of unigrams	In this paper Present efficient approximate inference techniques based on variation methods and an EM algorithm for empirical Bayes parameter estimation.
2	H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification,2007	Advantage is that during classification model construction, all relevant substructures are available allowing the classifier to intelligently select the most discriminating ones.8	feature selection algorithms can be developed by taking this information into account so that to ensure that the entire (or most of) molecule is covered by the selected features.	In this paper presents a substructure-based classification algorithm that decouples the substructure discovery process from the classification model construction and uses frequent subgraph discovery algorithms to find all topological and geometric substructures present in the dataset.
3	W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. Yu, and O. Verscheure. Direct mining of discriminative and essential frequent patterns via model-based search tree,2008	The discovered feature vectors are more accurate on some of the most difficult graph as well as frequent item set problems than most recently proposed algorithms.	Still find good features even when the training and testing data follow significantly different prior class distribution. More studies are being conducted	In this paper it builds a decision tree that partitions the data onto different nodes. Then at each node, it directly discovers a discriminative pattern to further divide its examples into purer sub-sets
4	M. Kobet ski and J. Sullivan. Discriminative tree-based feature Mapping,2011	linear SVM classifier is able to achieve much higher performance than in the original feature space,	It would be interesting to see how the results translate to detection when used together with a complex linear model	In this paper introducing an intermediate mapping step where examples are mapped from a given feature space to one where they are easier to separate using a linear classifier focused mainly on exploring and justifying our tree-based mapping algorithms.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

5	T. Kudo, E. Maeda, and Y. Matsumoto. An application of boosting to graph classification,2004	Method achieves comparable or even better performance than SVMs with convolution kernels as well as improves the testing efficiency.	Takes more time to execute	In this paper discuss the relation between our algorithm and SVMs with convolution kernels. Two experiments using natural language data and chemical compounds
6	W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. Yu, and O. Verscheure. Direct mining of discriminative and essential frequent patterns via model-based search tree,2008	The instance elimination effectively reduces the problem size iteratively and expedites the mining process	Still need two find more effectively.	In this paper propose a direct discriminative pattern mining approach, DPMine, to tackle the efficiency issue arising from the two-step approach. DDP Mine performs a branch-and-bound search for directly mining discriminative patterns without generating the complete pattern set. Instead of selecting best patterns in a batch.
7	M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent substructure-based approaches for classifying chemical compounds. TKDE, 17(8):1036–1050, 2005	During classification model construction, all relevant sub-structures are available allowing the classifier to intelligently select the most discriminating ones	It takes time to execute	In this paper present a sub-structure-based classification algorithm that decouples the sub-structure discovery process from the classification model construction and uses frequent sub graph discovery algorithms to find all topological and geometric sub-structures present in the dataset. The advantage of our approach is that during classification model construction, all relevant sub-structures are available allowing the classifier to intelligently select the most discriminating ones.
8	H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. JMLR, 2:419–444, 2002.	capture semantic information, to the extent that it can outperform state of the art systems on some data.	No longer so important in large datasets where there is enough data to learn the relevance of the terms.	In this paper presented a novel kernel and its approximation for text analysis. The performance of the string subsequence kernel was empirically tested by applying it to a text categorization task. This kernel can be used with any kernel-based learning system, for example in clustering, categorization, ranking,
9	B. L. W. H. Y. Ma. Integrating classification and association rule mining. In Proceedings of the fourth international conference on knowledge discovery and data mining, 1998.	solve a number of problems that exist in current classification systems.	Need accurate classifiers by using more sophisticated techniques and to mine CARs without pre-discretization.	In this paper proposes a framework to integrate classification and association rule mining. An algorithm is presented to generate all class association rules (CARs) and to build an accurate classifier. The new framework not only gives anew way to construct classifiers, but also helps to solve a number of problems that exist in current classifications systems.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

10	Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 623–631. ACM, 2013.	It efficiently measures the strength of all potential pair wise interactions	Worked on one-dimensional features data only	In this paper present a framework called GA2M for building intelligible models with pairwise interactions. Adding pairwise interactions to traditional GAMs retains intelligibility, while substantially increasing model accuracy. To scale up pair-wise interaction detection

III. IMPLEMENTING THE SYSTEM

The architecture of the system is as follows:

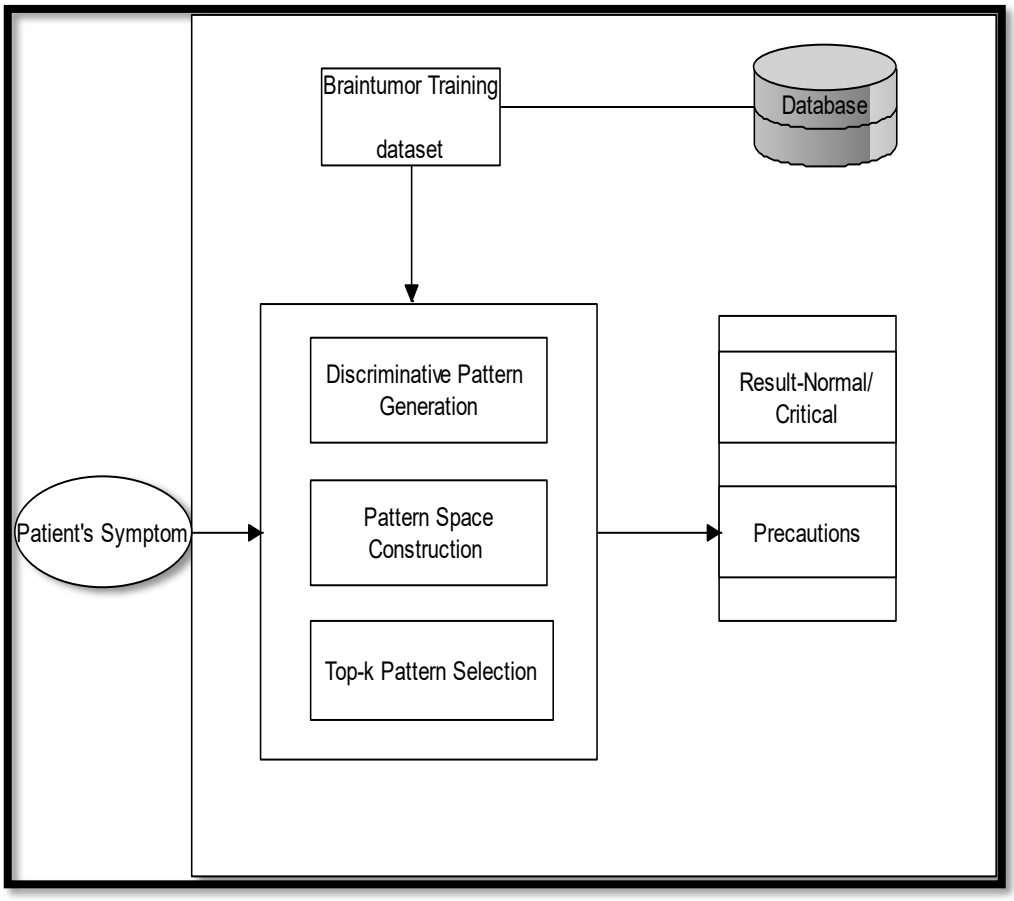


Figure 1: Architecture

The preceding architecture shows systemcan diagnose the disease using some rules based on this information, it can be verified whether the topdiscriminative patterns selected by DPPred are consistent with the actual diagnosing rules. The system will work ondisease prediction for user entered data feature. Propose a novel discriminative patterns-based learning framework (DPPred) that extracts a concise set of discriminative patterns from high-order interactions among features for accurateclassification and regression. In DPPred, first System train tree-based models to generate a large set of high-orderpatterns. Second, System explores all prefix paths from root nodes to leaf nodes in the tree-based models as ourdiscriminative patterns. Third, System compresses the number of discriminative patterns by selecting the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

most effectivepattern combinations that fit into a generalized linear model with high classification accuracy or small regression error.
This component of fast and effective pattern extraction enables the strong predictability and interpretability of DPPred.

IV. METHDOLOGY

1. **Discriminative Pattern Generation**:
In our implementation, once the splitting feature is selected, all mid-points between any two consecutive feature values after sorting are considered as potential splitting points. T random decision trees are generated, and for each tree,all prefix paths from its root to non-leaf nodes are treated as discriminative patterns. Due to the predictivity of decision trees, so-generated patterns are highly effective in the specific prediction task.

Algorithm 1: Discriminative Pattern Generation
Require: n training instances (x_i,,y_i), the number of trees T, the depth threshold D, and minimum tree size α
Return: a set of discriminative patterns for further selection
P<— θ
fort = 1 to T **do**
Build a random decision tree [2] with maximum depth D and minimum tree bag size α.
foreach *non-leaf node u***do**
P<—P ∪{root —>u}
returnP

For each discriminative pattern, there is one corresponding binary dimension describing whether the instances satisfy the pattern or not. Because the dimension of the pattern space is equal to the number of dis criminative patterns which is a very large number after the generation phase, we need to further select a limited number ofpatterns and thus make the pattern space small and efficient. It is also worth a mention that this mapping process isable to be fully parallelized for speedup.

Algorithm 2: Pattern Space Construction
Require: n instances (x_i), a discriminative pattern set P
Return: n instances in pattern space (x'_i)
fori = 1 ton**do**
x'_i<— 0
for jth pattern P_j in P **do**
if x_i satisfies patient P_j**then**
$x'_{i,j}$<— 1
return (x'_i)

3. **Top-k Pattern Selection**:
After a large pool of discriminative patterns is generated, further top-k selection needs to be done to identify themost informative and interpretable patterns. A naïve way is to use heuristic functions, such as information gain andGini index, to evaluate the significance of different patterns on the prediction task and choose the top rankedpatterns. However, the effects of top ranked patterns based on the simple heuristic scores may have a large portionof overlaps and thus their combination does not work optimally.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

Algorithm 3: Top-k Pattern Selection: Forward
Require: n training examples $(x_i,,y_i)$, a set of discriminative patterns P and k
Return: top-k discriminative patterns set P_k and a generalized linear model f(.)
P_k <— θ
for t=l to k **do**
for each pattern p in P **do**
x' <— construct pattern space (x, P_k U {p}) using Algorithm 2
g(.)<— a generalized linear model [32] on (x'_i,y_i)
per_p <— g(.)'s training performance
P_k <— P_k U {arg max_p per_p}
x' <— construct pattern space (x, P_k)
f(.) <— a generalized linear model on (x'_i,y_i)
return P_k, f(.)

The following diagram shows how to select top k discriminative patterns:

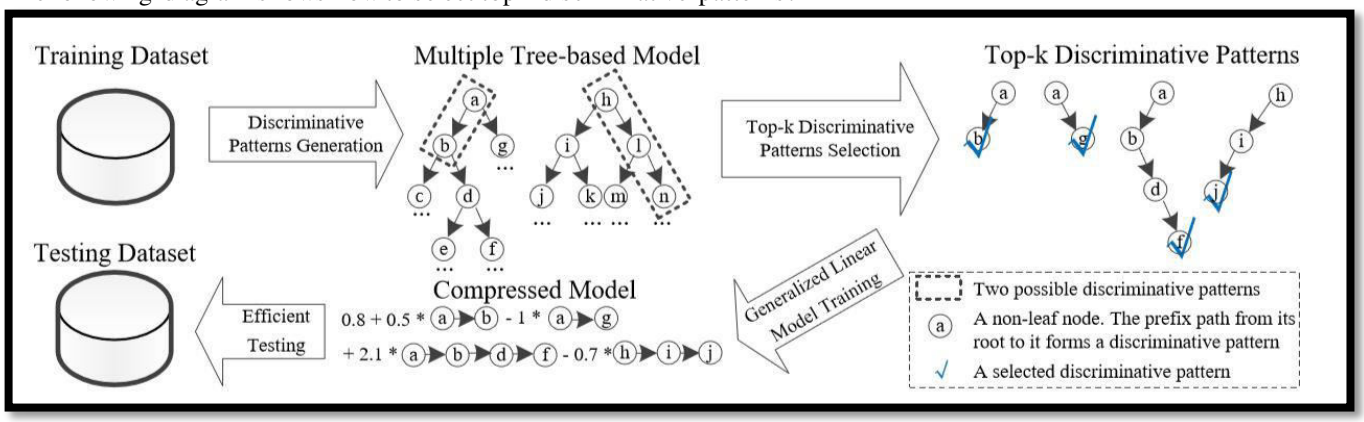


Figure 2: Selecting top k discriminative patterns

V. SIMULATION RESULTS

The patient can select the symptoms as follows:

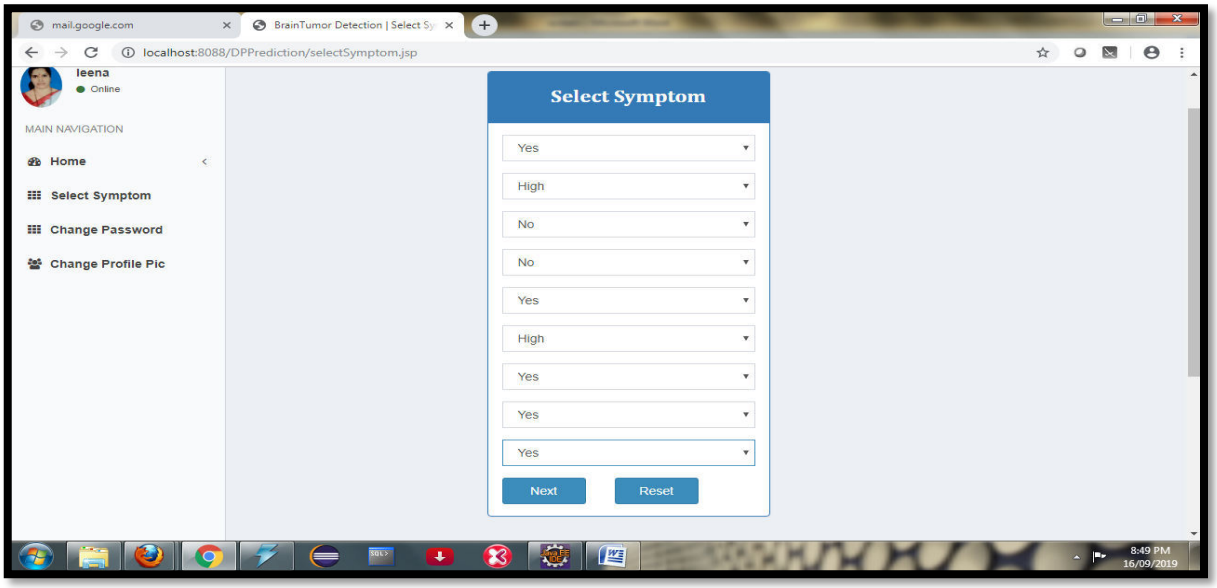


Figure 3: Selecting symptoms

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

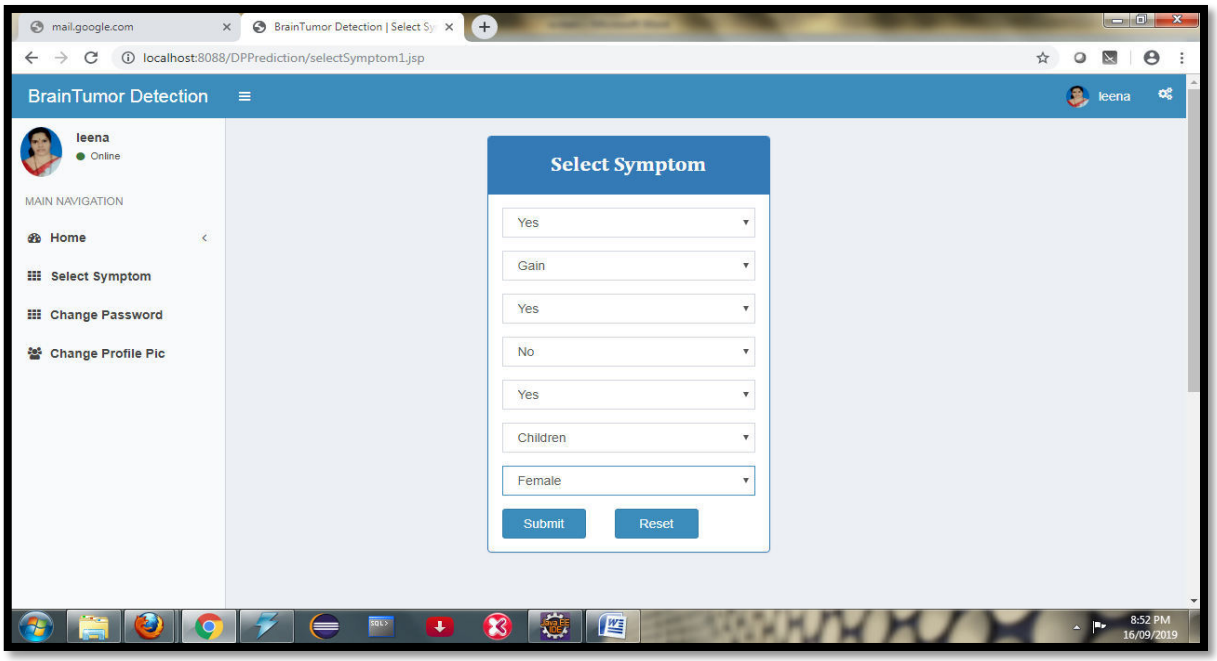


Figure 4: Selecting symptoms (cont....)

The condition of patient will be displayed on screen as follows:

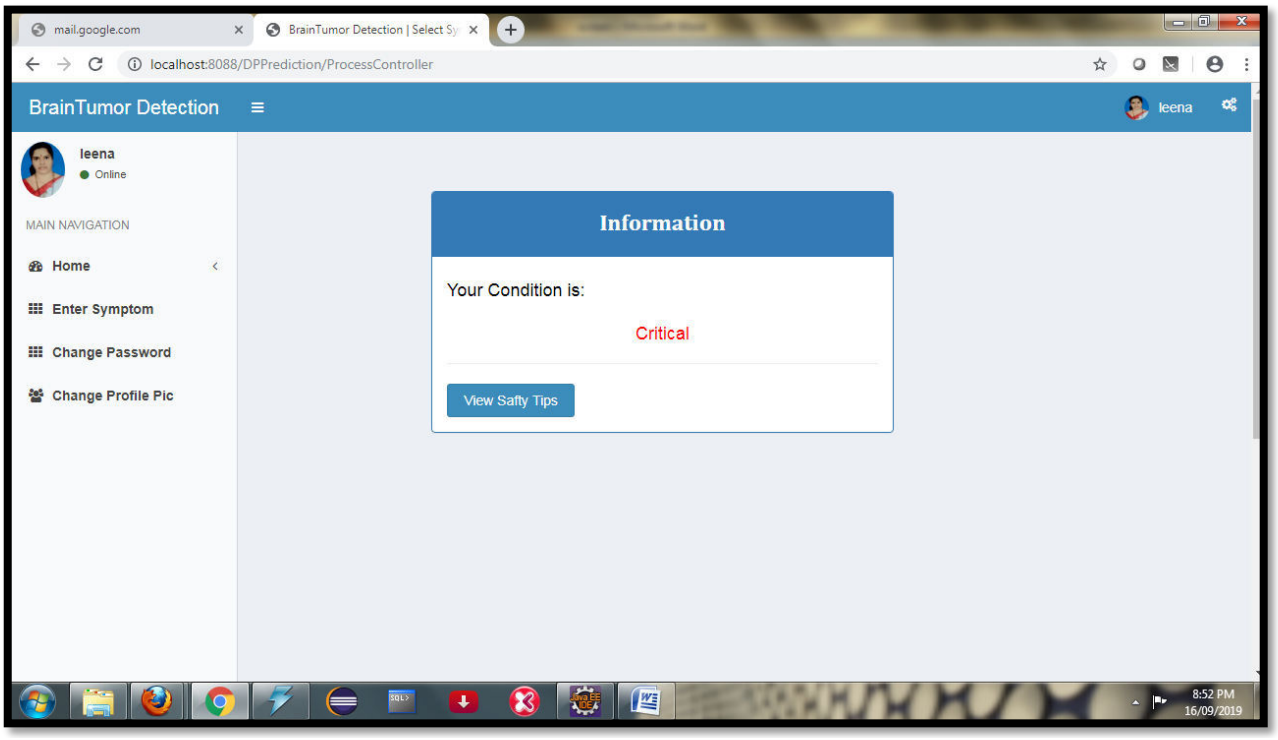


Figure 5: Condition of patient

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

The final output of result is displays on admin's window as follows:

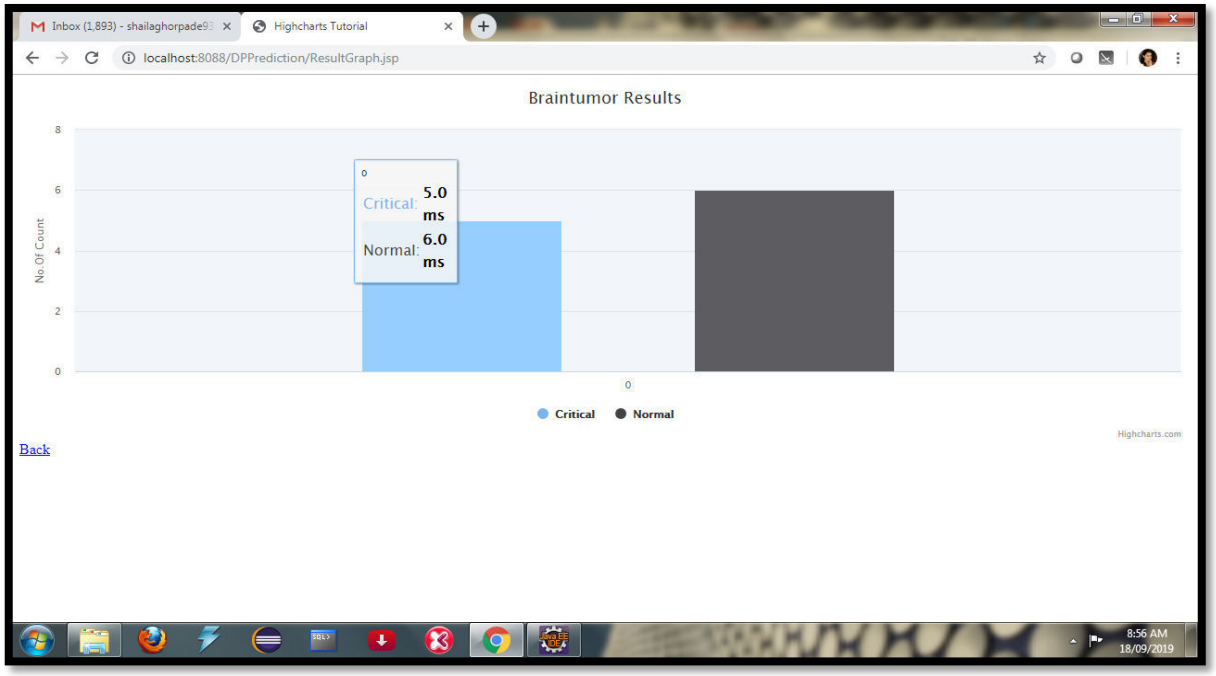


Figure 6: Final result

VII. CONCLUSION

This is an effective and concise discriminative pattern-based prediction framework (DPPred) to address the classification problems and provide high interpretability with a small number of discriminative patterns. The system work on disease prediction based on different feature testing at testing time. This system gives disease prediction on data given by patient.

REFERENCES

[1] Jingbo Shang, Meng Jiang, Wenzhu Tong, Jinfeng Xiao, JianPeng, Jiawei Han,the Pooled Resource Open-Access ALS Clinical Trials Consortium, DPPred: An Effective Prediction Framework with Concise Discriminative Patterns, 2017 IEEE.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. JMLR, 3:993–1022, 2003.

[3] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. University of California, Berkeley, 2004.

[4] H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, pages 716– 725. IEEE, 2007.

[5] H. Cheng, X. Yan, J. Han, and P. S. Yu. Direct discriminative pattern mining for effective classification. In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, pages 169–178. IEEE, 2008.

[6] Q. Cheng, J. Shang, J. Juen, J. Han, and B. Schatz. Mining discriminative patterns to predict health status for cardiopulmonary patients. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '16, pages 41–49, New York, NY, USA, 2016. ACM.

[7] G. Cong, K.-L. Tan, A. K. Tung, and X. Xu. Mining top-k covering rule groups for gene expression data. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 670–681. ACM, 2005.

[8] S. Derksen and H. Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology, 45(2):265–282, 1992.

[9] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent substructure-based approaches for classifying chemical compounds. TKDE, 17(8):1036–1050, 2005.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 2, February 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0902022 |

[10] G. Dong and V. Taslimitehrani. Pattern aided classification. In Proceedings of 2016 SIAM international conference on Data Mining, 2016.
[11] G. Dong, X. Zhang, L. Wong, and J. Li. Caep: Classification by aggregating emerging patterns. In International Conference on Discovery Science, pages 30–42. Springer, 1999.
[12] M. Kobetski and J. Sullivan. Discriminative tree-based feature mapping. Intelligence, 34(3), 2011.
[13] W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. Yu, and O. Verscheure. Direct mining of discriminative and essentialfrequent patterns via model-based search tree. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 230–238. ACM, 2008.
[14] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent substructure-based approaches for classifying chemical compounds. TKDE, 17(8):1036–1050, 2005.
[15] T. Kudo, E. Maeda, and Y. Matsumoto. An application of boosting to graph classification. In Advances in neural information processing systems, pages 729–736, 2004.
[16] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. JMLR, 2:419– 444, 2002.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details