Respiratory Cancerous Cells Detection Using TRISS Model and Association Rule Mining

Ayyadurai.P[1], Kiruthiga.P[2], Valarmathi.S[3], Amritha.S[4]

Abstract- **Lung cancer is a disease of uncontrolled cell growth in Epithelium of the lung, Lung cancer is one of the most common and deadly diseases in the world. Early and accurate detection of cancer is critical to the well being of patients. We analyze the lung cancer data available from the SEER program with the aim of developing accurate survival Estradiol models for lung cancer using data mining techniques. The goal here is to identify characteristics of patient segments where average survival is significantly higher/ lower than average survival across the entire dataset. Several data mining classification techniques were used on the preprocessed data along with various data mining optimizations and validations.**

A subset of 13 patients attributes from the SEER data were recently linked with the survival outcome using Estradiol models, which is used in this study for segmentation. The resulting rules conform to existing biomedical knowledge and provide interesting insights into lung cancer survival.

Keywords**: Classification, Association rule, Lung cancer, SEER Database, Neural networks, Hotspot, Weka.**

I.INTRODUCTION

Lung cancer ranks second place in the list of popular common cancers and first in the list of most deadly cancers, with the survival rate being about 15% after 5 years of diagnosis [1].

The Surveillance, Epidemiology (Disease Outbreaks), and End Results (SEER) Program of the National Cancer Institute is an authoritative repository of cancer statistics in the United States. It is a population-based cancer registry which covers about 26% of the US population across several geographic regions and is the largest publicly available domestic cancer dataset. The data includes patient composition, cancer type and site, stage, first course of treatment, and follow-up vital information. The SEER program collects cancer data for all invasive and in situ cancers, except basal and squamous cell carcinomas of the skin and in situ carcinomas of the uterine cervix. The 'SEER limited-use data' is available from the SEER [1], [2] website on submitting a SEER limiteduse data agreement form.presents an overview study of the cancer data at all sites combined and on selected, frequently occurring cancers from the SEER data. The SEER data attributes can be broadly classified as composition attributes (e.g. age, gender, location), diagnosis attributes
(E.g. primary site, histology, grade, tumor size), treatment attributes (e.g. surgical procedure, radiation therapy), and Outcome attributes (e.g. survival time, cause of death), which makes the SEER data ideal for performing outcome analysis studies.

II.RELATED WORK

SEER data being available in the public or common domain, there is a mature literature on the statistics of SEER data, many of them using the the SEERStat software provided by SEER itself.

In addition, there also have been a few data mining applications, which have become a very significant component of cancer research and survivability analysis. A number of techniques based on data mining have been proposed for the survivability analysis of various cancers. Uses decision trees and artificial neural networks for survivability analysis of stomach cancer, breast cancer, diabetes and hepatitis.

Uses artificial neural networks on SEER data to predict breast cancer survival. Empirically compared three data mining techniques: neural networks, decision trees and logistic regression for the task of predicting 60 months breast cancer survival. They applied these techniques on 2000 version of SEER data. They found that decision trees performed the best with 93*:*6% accuracy, followed by neural networks. Found that the pre-classification process used by was not accurate in determining the records of the 'not survived' class. The authors of corrected this and investigated Naive bayes, the back-propagated neural networks, and the C4.5 decision tree algorithm using the data mining tool WEKA. Decision Trees and Neural networks performed the best with 86*:*7% and 86*:*5% accuracy respectively [3].

III.CLASSIFICATION SCHEMES

We used several morphology schemes resulting in identification of top 5 classification schemes, plus ensemble voting scheme to combine the prediction probabilities from the top5 (details presented in Experiments and Results section).

This section presents a brief theoretical of the classifiers and meta-classifiers used in the experiments reported in this paper.

1. **Support vector machines**: SVMs are based on the Structural Risk Minimization (SRM) principle from statistical learning theory. A detailed description of SVMs and SRM is available. In their basic form, SVMs attempt to perform classification by constructing hyperplanes in a multidimensional space that separates the cases of different class labels. It supports both classification and regression tasks and can handle multiple Continuous and nominal variables.

2. **Artificial neural networks**: ANNs are networks of interconnected artificial neurons, and are commonly used for non-linear statistical data modeling to model complex relationships between inputs and outputs. The network includes a hidden layer of multiple artificial
Neurons connected to the inputs and outputs with different edge weights.

3. **J48 decision tree**: In a decision tree classifier, the internal nodes denote the different attributes whose values would be used to decide on the classification Path, and the branches denote the split depending on the attribute values, while the leaf nodes denote the final value (classification) of the dependent variable.

4. **Random forest**: The Random Forest classifier consists of multiple decision trees. The final class of an instance in a Random Forest is assigned by outputting the class that is the mode of the outputs of individual trees, which can produce robust and accurate classification, and ability to handle a very large number of input variables. It is relatively robust to overfitting and can handle datasets with highly imbalance class distributions.

5. **LogitBoost**: Boosting is a technique that can dramatically improve the performance of several classification techniques by sequentially applying them repeatedly to re-weighted versions of the input data, and taking a weighted majority of the sequence of classifiers thereby produced [4], [5].

IV.TRISS MODEL

TRISS Stands for Traumatic Injury Survival Prediction System. Understanding and cleaning data to prepare it for a data mining analysis is one of the most important steps in the data mining approaches. The proposed respiratory cancer survival prediction system consists of four stages:

1. **SEER-related preprocessing**: This is the first stage preprocessing designed according to the way SEER program records, documents, and releases the data.
2. **Problem-specific preprocessing**: This is the second stage preprocessing which is specific to the problem of survival prediction.
3. **Predictive modeling**: This is where data mining classifiers are employed to construct predictive models for cancer-specific survival, on the preprocessing data.
4. **Evaluation**: In this stage, the predictive model is evaluated on the testing data.

V.ASSOCIATION RULE

Association rule mining is often stated as follows Let *I* be a set of *n* binary attributes called items. Let *T* be a set of transactions data. Each transaction data in *T* contains a subset of the items in *I*. Association rule mining is popularly done with flag attributes, indicating the presence/absence of the item in the transaction.

Hotspot Technique

This is an association rule mining algorithm which is directed by a target attribute, which means that the consequent is fixed to the target attribute. It can be used for segmentation with both nominal and numeric targets. It uses a greedy approach to construct the tree of rules in a depth-first fashion, where the search is constrained by the following parameters [6],
1) Maximum branching factor: The number of children nodes to consider at each node. This parameter controls the amount of search performance and comparison, since the algorithm uses a greedy search.
2) Minimum improvement in target value: This is the minimum improvement in the target value of the resulting segment in order to consider adding a new branch.
3) Minimum segment size: The size of the resulting segment must be at least this much in order to add a new branch.
It uses the following 13 Patient attributes:
1) Age at diagnosis: Numeric Random age of the patient at the time of diagnosis for lung cancer.

2) Birth place: The native place of birth of the patient.

3) Cancer grade: A descriptor or influence of how the cancer cells appear and how fast they may grow and spread.

4) Diagnostic result confirmation: The best method used to confirm the presence of lung cancer.

5) Farthest extension of tumor: The farthest documented extension of tumor away from the lung, either by contiguous extension (regional growth or population highest of cells data) [7].

6) Lymph node involvement or decision making process: The highest specific lymph node chain that is involved by the tumor.

7) Type of surgery performed: The surgical procedure that removes and/or destroys cancerous tissue of the lung, performed as part of the initial work-up or first course of therapy.

8) Reason for no surgery in patient: The reason why surgery was not performed (if not).

9) Order of surgery and radiation therapy or practice: The order in which surgery and radiation therapies were administered for those patients who had both surgery and radiation.

10) Scope of regional lymph node surgery: It describes the removal, biopsy, or aspiration of regional lymph node(s) at the time of surgery of the primary site or during a separate surgical event.

11) Cancer stage: A descriptor or influence of the extent the cancer has spread, taking into account the size of the tumor, depth of penetration, metastasis, etc.

12) Number of malignant or affected tumors in the past: An integer denoting the number of malignant tumors in the patient’s lifetime.

13) Total regional lymph nodes examined: An integer denoting the total number of regional lymph nodes that were removed and examined by the pathologist or specialist [1], [8].

Table-I Lung Cancer Dataset Attributes

Attribute	Type
Age at Diagnosis	Numeric
Birth Place	Nominal
Cancer grade	Nominal
Diagnostic Confirmation	Nominal
Extension of Tumor	Nominal
Lymph node	Nominal
Type of Surgery	Nominal
Reason for Surgery	Nominal
Order of Surgery	Nominal
Regional Lymph node	Nominal
Cancer Stage	Nominal
Malignant tumor	Numeric
Total regional lymph node	Numeric
Survival time	Numeric

Table II
Reason for No Surgery

Code	Description
0	Surgery Performed
1	Surgery not Recommended
2	Constrained Due to other Conditions
6	Unknown Reason for No Surgery
7	Patient Guardian Refused
8	Unknown done
9	Surgery Performed

Table III
Codes for Cancer Stage

Code	Description
0	In Situ
1	Localized
2	Regional
7	Distant
9	Upstaged

VI.EXPERIMENTS AND RESULTS

WEKA Tool is a popular machine learning software and free software written in java. Thus tools to provide various classifiers included to analyze real applications.

The distribution of survival time across all the patients is shown in Figure 1. Before performing Hotspot analysis, we would like to study the influence of each of the individual 13 attributes on survival time. For this purpose, we plotted the average survival time for different possible values of each Input attributes.

We performed two independent analyses to find segments in which average survival time was higher and lower than overall average survival. Several combinations of algorithm Parameters (maximum branching factor, minimum improvement in target value, and minimum segment size) were tried.

Here we report the results with the following parameters: maximum branching factor = 3, minimum improvement in Target value = 1%, and minimum segment size = 100.

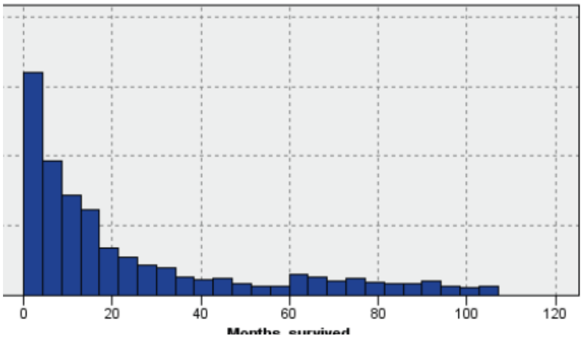


Figure 1.Distribution of Survival time

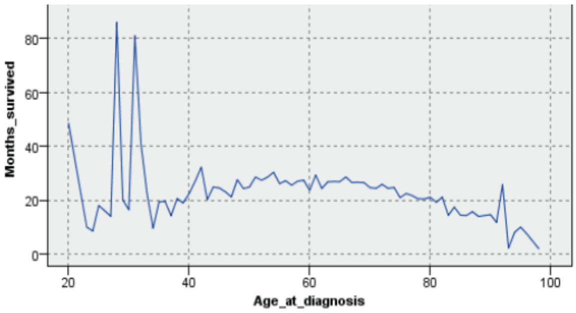


Figure 2. Survival time vs. Age Diagnosis

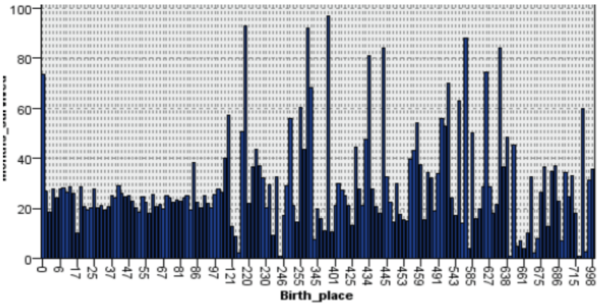


Figure 3. Survival time vs. Birth place.

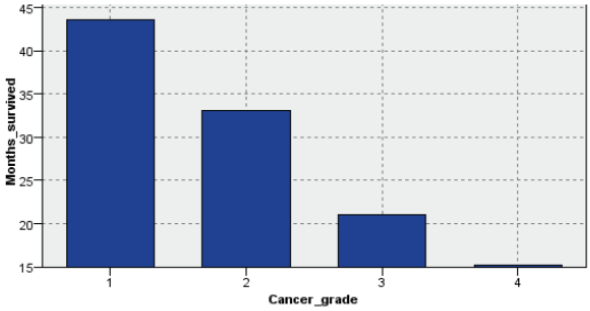


Figure 4. Survival time vs. Cancer grade.

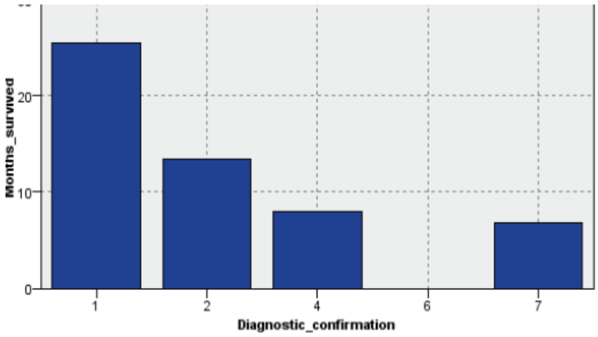


Figure 5. Survival time vs. Diagnostic confirmation.

As commonly known, association rule analysis invariably leads to discovery of a large number of redundant rules, which need to be removed. We used a 2-stage semi-manual procedure to remove redundant rules:

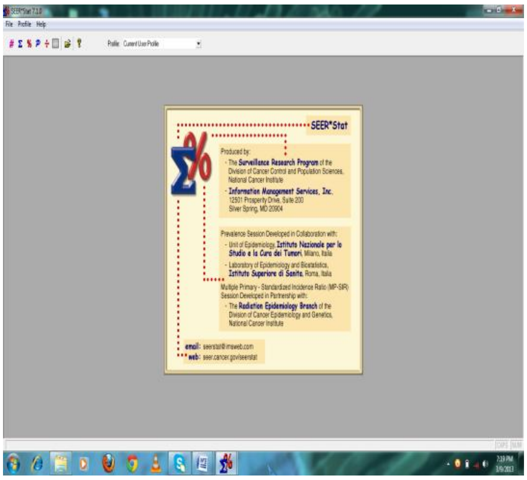
1) Stage I: Since the Hotspot algorithm tries to go deeper into the data as long as it is able to improve the target value, the leaf nodes would have the best target value compared to all the nodes on its path. So, we discard all the rules corresponding to the non-leaf nodes, and Retain only the rules corresponding to the leaf nodes. This stage does not require manual intervention.

2) Stage II: Even after Stage I, there still remain quite a few redundant rules, the removal of which require domain expertise. The redundant rules at this stage were manually removed.

Lift of a rule is the relative improvement in the target (here survival time) as compared to the average value of the target across the entire dataset.

Sample Experimental –Demo section

Step 1: Installs the seer stat* software for any operating system used.



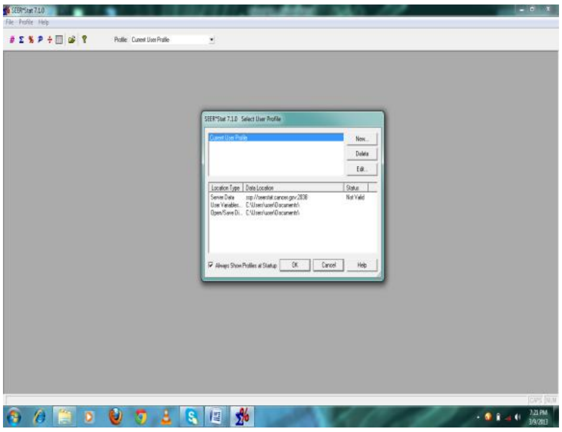
Produced by:
- The Surveillance Research Program of the Division of Cancer Control and Population Sciences, National Cancer Institute
- Information Management Services, Inc.
12501 Prosperity Drive, Suite 200
Silver Spring, MD 20904

Prevalence Session Developed in Collaboration with:
- Unit of Epidemiology, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano, Italia
- Laboratory of Epidemiology and Biostatistics, Istituto Superiore di Sanita, Roma, Italia

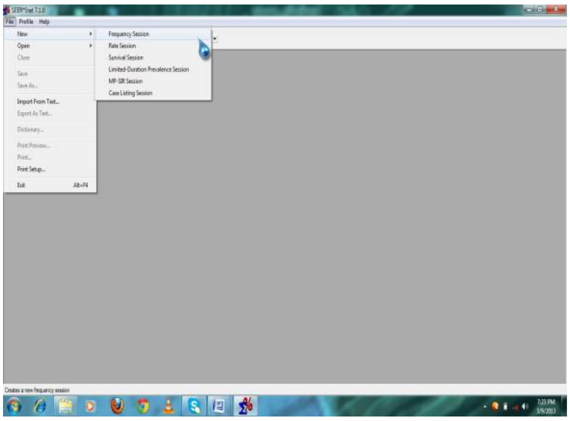
Multiple Primary - Standardized Incidence Ratio (MP-SIR) Session Developed in Partnership with:
- The Radiation Epidemiology Branch of the Division of Cancer Epidemiology and Genetics, National Cancer Institute

email: seerstat@imsweb.com
web: seer.cancer.gov/seerstat

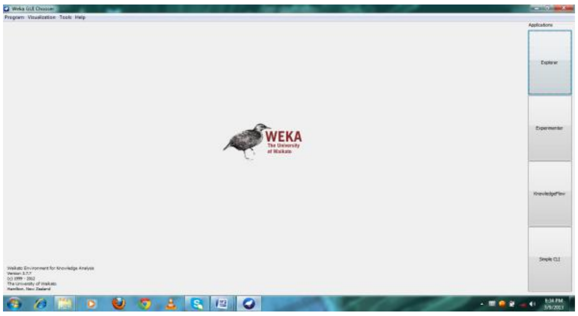
Step 2: Create one user profile stored patient records, and adding many new profile created.



Step 3: Calculates the cancer affected person Rate session, frequency session and Survival session categories on individual section, each file updating the own sessions.

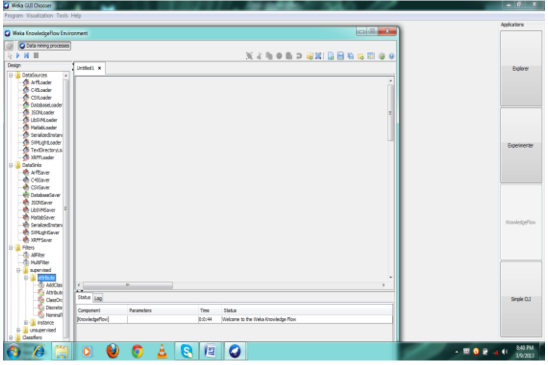


Step 4: Weka tool installed next step, weka is one of the open access machine learning software.

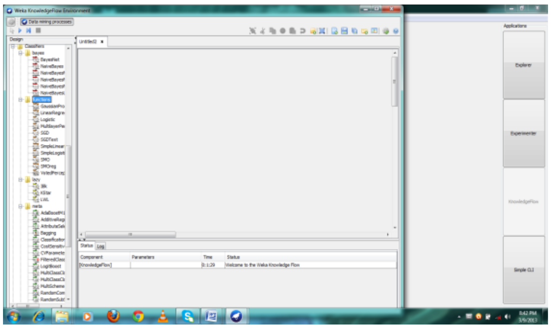


Waikato Environment for Knowledge Analysis
Version 3.7.7
(c) 1999 - 2012
The University of Waikato
Hamilton, New Zealand

Step 5: Weka provides different types of classifiers used for many applications.

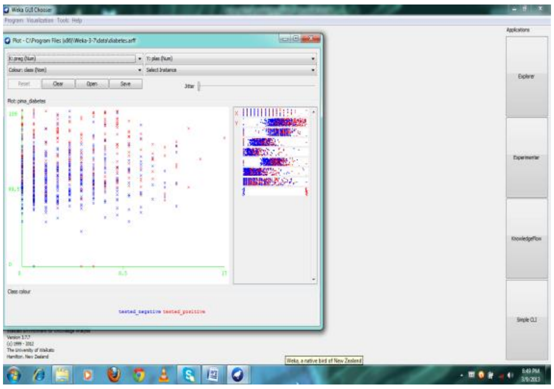


[KnowledgeFlow] 0:0:44 Welcome to the Weka Knowledge Flow



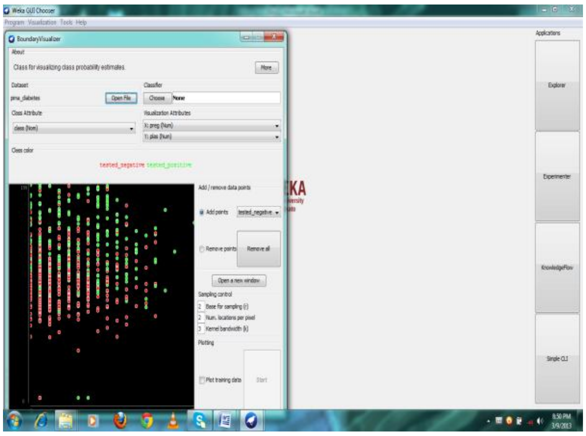
[KnowledgeFlow] 0:1:28 Welcome to the Weka Knowledge Flow

Step 6: Calculates the approximate result for cancer person survival time to plot the graphs on simulated or boundary output.


Weka GUI Chooser
Program Visualization Tools Help
Plot - C:\Program Files (x86)\Weka-3-7\data\diabetes.arff
Reset Clear Open Save
Jitter
Plot: pima_diabetes
Class colour
tested_negative tested_positive
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI
Weka, a native bird of New Zealand

X: preg (Num)
Y: plas (Num)
Colour: class (Nom)
Select Instance

Version 3.7.7
(c) 1999 - 2012
The University of Waikato
Hamilton, New Zealand



Class for visualizing class probability estimates.

Dataset: pima_diabetes
Classifier: None
Class Attribute: class (Nom)
Visualization Attributes: X: preg (Num), Y: plas (Num)
● Add points: tested_negative
○ Remove points
2 Base for sampling (r)
2 Num. locations per pixel
3 Kernel bandwidth (k)
□ Plot training data

VII. CONCLUSION & FUTURE WORK

In this paper, we performance association rule mining analysis on lung cancer data from SEER to identify hotspots in the cancer data, where the patient survival time is significantly higher than and lower than the average survival time across the entire dataset.

We believe that such analysis can be very useful to identify the factors affecting survival, and aid doctors and Patients in avoiding the conditions which are known to reduce survival time, and encourage the conditions which are known to increase the survival time, whenever possible.

It can also aid doctors in decision making and improve informed patient consent by providing a better understanding of the risks involved in a particular treatment procedure. Similar analysis can also be done for other cancers.

ACKNOWLEDGEMENT

We thank the SEER program to make their limited-use data available for this work. This work is supported in part by NSF award numbers. My sincere thanks to Adyar Cancer Institute, for providing facility data collection Work.

REFERENCES

[1] P.Ramachandran, Dr.N.Girija, Dr.T.Buvaneswari; Health service sector: Classifying and Finding Cancer Spread Pattern in Southern India Using Data Mining Techniques; International Journal on Computer Science and Engineering (IJCSE); Vol 4. No.5, May2012.

[2] Der-Chiang Li and Chiao-Wen Liu; Extending Attribute Information for Small Data Set Classification; IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 3, March 2012.

[3] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "A lung cancer outcome calculator using ensemble data mining on seer data," in Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics, ser. BIOKDD '11, 2011, pp. 5:1–5:9.

[4] "Overview of the seer program," surveillance Epidemiology and End Results, URL: http://seer.cancer.gov/about/Accessed: Aug 2, 2011.

[5] Zakaria Suliman Zubi, Rema Ashibani Saad; Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer; Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases; April2010.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," SIGKDD Explorations, vol. 11, no. 1, June2009.

[7] Shital Shah, Andrew Kusiak; Cancer gene search with data-mining and genetic algorithms; Computers in Biology and Medicine 37 (2007) 251 – 261.

[8] Developed by the National Collaborating Centre for Acute Care, Lung cancer; the diagnosis and treatment of lung cancer, ISBN: 1-84257-920-7, Published by the National Institute for Clinical Excellence February 2005.

AUTHORS BIBLIOGRAPHY

P.Ayyadurai received the B.Tech Degree in Information Technology from Dr.Navalar Nedunchezhiyan College of Engineering, Anna University in 2011 and now he is an M.E student in the Department of Computer Science & Engineering from Srinivasan Engineering College – Dhanalakshmi Srinivasan Group of Institutions, Perambalur, and Tamilnadu, India. His research interest includes Data mining, Pattern Recognition, and Image Processing.

P.Kiruthiga received the B.Tech Degree in Information Technology from Dhanalakshmi Srinivasan Engineering College, Anna University in 2011 and now she is an M.E student in the Department of Computer Science & Engineering from Srinivasan Engineering College – Dhanalakshmi Srinivasan Group of Institutions, Perambalur, and Tamilnadu, India. Her research interest includes Data mining, Genetic algorithms and Image Processing.

S.Valarmathi received the B.Tech Degree in Information Technology from Dhanalakshmi Srinivasan Engineering College, Anna University in 2011 and now she is an M.E student in the Department of Computer Science & Engineering from Srinivasan Engineering College – Dhanalakshmi Srinivasan Group of Institutions, Perambalur, and Tamilnadu, India. Her research interest includes Network security, and Image Processing.

S.Amritha received the B.E Degree in CSE from Srinivasan Engineering College, Anna University in 2011 and now she is an M.E student in the Department of Computer Science & Engineering from Srinivasan Engineering College – Dhanalakshmi Srinivasan Group of Institutions, Perambalur, and Tamilnadu, India. Her research interest includes Data mining, Cloud Computing and Image Processing.