

Predicting The Recurrence of Colorectal Cancer in Patients Using Ensemble Methods

Y. Jeevan Nagendra Kumar¹

¹Professor and Dean TIC,
Department of Information Technology,
Gokaraju Rangaraju Institute of Engineering
and Technology, JNTUH.
Hyderabad, India
jeevannagendra@griet.ac.in

²Ridhi Vyas, ³B. Hruthika, ⁴B. Seershika,
⁵T.Pranitha

^{2 3 4 5}Department of Information
Technology, Gokaraju Rangaraju Institute
of Engineering and Technology, JNTUH.
Hyderabad, India.

Abstract—Among the top ten death-causing diseases in the world, cancer is considered to be one. There are about 100 different forms of cancer, and cancer prediction is critical to the advancement of data mining applications. The development of adequate methods for locating cancer instances will go a long way toward improving cancer patient therapy diagnosis. The term "cancer" refers to a group of disorders. All malignancies are caused by cells that continue to divide without stopping and spread to surrounding tissues. The ensemble model can be effectively used to prognosticate the recurrence of colorectal cancer in clinical patients. According to current data from Globacancer India, over 27,000 cases were documented in the preceding year, with about 20,000 people dying. In this project, ensemble approaches are utilized to predict cancer recurrence. Basic classifier approaches such as SVM, and RF, as well as ensemble classifiers of bagging, were employed to construct classifier models. According to the findings, the constructed models' prediction performance was normally good. In terms of accuracy, ensemble models outperform when compared to Simple classifiers. In a model comparison, the ensemble voting technique outperformed all other models by a significant margin.

Keywords: *Colorectal Cancer, Ensemble methods*

I. INTRODUCTION

ML is increasingly being used in areas such as Electroencephalogram analysis and cancer detection/analysis by researchers. For instance, it is very easy to pick out the genes causing and the ones suppressing cancer expression just by viewing the biological data like Deoxyribonucleic Acid methylation and Ribonucleic Acid sequencing.

Machine Learning has found tremendous applications in healthcare (ML). It can be put into use to analyze various thousands of varying data points, make predictions, and provide timely risk rankings, direct resources, and so much more. This research analyses a few of the most prominent applications of machine learning in healthcare, as well as how such applications may be utilized to alter our perceptions of the business in 2020 and beyond. In the ensemble approach, several models are created in multiples and then combined to get better results. In most situations, ensemble techniques

produce more accurate findings than a single model, and the winning solutions in various machine learning competitions use ensemble techniques. Ensemble Methods are among the most effective and easy-to-use predictive analytics algorithms. Ensemble methods have higher predictive accuracy, compared to the individual models. With ensemble methods, bias/variance can be reduced and most of the time, the model is not under fitted/overfitted. The ensemble of models is always less noisy and more stable.

Among ensemble learning techniques, bagging also referred to as bootstrap aggregation, is often employed to minimize variance within noisy data sets. An individual data point may be selected many times from a training set by bagging - that is, taking data from the training set at random with replacement. Random forest is used for classification and regression issues and is a supervised ensemble learning technique. Given its outstanding performance across a broad variety of classification and regression predictive modeling tasks, it is possibly the most popular and commonly used machine learning method. A recurrence occurs when cancer has been treated, this can happen weeks, months, or even years later. Your doctor will not be able to accurately predict whether or not cancer will return. The likelihood of a recurrence is determined by the type of primary cancer. We took Colorectal cancer as our primary cancer. A recurrence of colorectal cancer can be local (limited to the location where the cancer was first discovered or neighboring tissues) or distant. Metastasis occurs when a recurrence occurs anywhere other than the colon or rectum.

Gene expression is a process that is employed by all known species and can be controlled and manipulated at multiple levels (transcription initiation, splicing, alternative splicing, mRNA stability, post-transcriptional regulation, and eventually translational and posttranslational regulation mechanisms). The recurrent disease typically manifests as distant metastases in the liver or lungs, as well as locoregional recurrence in the pelvic or peritoneum. Several prognostic variables related to the patient, tumor, and treatment regimen impact the occurrence of recurrence (radio preoperative chemotherapy, surgical method, and adjuvant

chemotherapy). These prognostic signs can be used to identify patients who are at high risk and require more frequent postoperative monitoring and adjuvant medicine. Because small cells of cancerous cells may go undetected in the body, cancer can recur. These cells may grow in quantity over time until they appear on test results or create signs or symptoms.

II. LITERATURE SURVEY

[1] E. Osterman proposed an “Accurate population-based model for individual prediction of colon cancer recurrence” in the year 2021. When evaluating the recurrence rate and the requirement for adjuvant treatment in patients with colorectal cancer, a prediction model is helpful. In actuality, the most often used models are based on decades of patient data from treatment trials, which exaggerate the present probability of recurrence, particularly among low-risk persons, since outcomes have improved over time. The goal of this study was to create and verify a model based on routinely obtained clinicopathologic characteristics for predicting recurrence within 5 years following surgery.

[2] In 2016, Castellanos suggested "Multiple-View Multiple-Learner Supervised Learning for Predicting Colorectal Cancer Recurrence." To enhance recurrence prediction, an ensemble learning technique is employed to predict recurrence status 3 years following diagnosis. Using 10x10-fold cross-validation, the microarray dataset was split and diced into various perspectives and then used to train a varied selection of basic learner classes. A stacking generalization was used to create ensemble learning. As a consequence of our research, molecular data outperforms basic clinical data in predicting recurrence. On all performance criteria, we show that the performance of (MVML) multiple-view multiple learners supervised learning frameworks outperforms or matches that of the best base learners.

[3] Han-Ching Chan proposed the “Development of a Gene-Based Prediction Model for Recurrence of Colorectal Cancer Using Machine Learning Algorithm” in 2021. With the help of gene expression data, we attempted to determine a gene signature in the future, these techniques could be used to identify high-risk participants of recurrence among stage I and II colorectal cancer patients. Data with imbalances induced by uncommon recurrence occurrences were first analyzed using a synthetic minority oversampling approach. Then we used a three-step methodology (microarray significance analysis, logistic regression, and recursive feature elimination) to figure out which genes are differentially expressed between patients with and without recurrence. To stabilize the prediction algorithm, we repeated the aforementioned operations with 10 subsets of the training data and utilized support vector machine methods to generate prediction models. The Majority vote was used to determine the final forecasts.

[4] Nafizatus Salmi proposed “Naïve Bayes Classifier Models for Predicting Colon Cancer” in 2019. When the Naive Bayes Classifier model is used to identify whether patients have colon cancer or not, it is demonstrated that the model has excellent accuracy, precision, recall, and a decent f1 - score when used as a classification approach. The Naive Bayes Classifier, according to the authors, is a prediction approach based on fundamental probability and the Bayes theorem (or Bayes rule) with a high assumption of independence. As a result, this model may be utilized to produce more accurate classifications with less effort.

III. METHODOLOGY

Dataset Description: The dataset which we have referred to for this project has information about patients that had colorectal cancer. The attributes (columns) in the dataset include patients' age, id, location, gender, DFS, DFS Event, gene expressions, etc. The DFS is a disease-free survival attribute which is an integer value in months, it signifies the number of months the patient was cancer-free after his/her surgery. The location attribute has either left, right, rectum, or colon. The DFS Event is the attribute that determines whether there is a possibility of recurrence of cancer in a patient. Its values are either 0 or 1. The dukes stage attribute has four options which include A, B, C, and D. The gene expression levels of the patients are also included as attributes. **Working of Ensemble Methods:** The term "Ensemble Methods" refers to the process of creating a large number of slightly distinct prediction models and then merging them through voting or averaging to produce extremely high performance. Machine crowdsourcing has been dubbed ensemble approaches. Bagging, Boosting, and Random Forest all aim to enhance performance beyond what a binary decision tree can provide, but the algorithms employ various techniques to achieve this. Bagging and Random Forests were created to solve the problems of binary decision trees' variation and stability. Bootstrap Aggregator is the source of the phrase "bagging." Each random subset of the original dataset is fitted with the base learners (classifiers) (bootstrapping). Due to the parallel ensemble, all of the classifiers in a training set are independent of one another, resulting in somewhat distinct characteristics inherited by each model. To arrive at the final findings, bagging combines all of the learners' outcomes and adds (aggregates) their predictions by averaging (mean) their outputs.

The machine learning model is implemented using a jupyter notebook. We have used a jupyter notebook for python coding. The information of the cancer patients is collected in a dataset for predicting the recurrence of cancer in those patients using a machine learning model. The dataset collected is first preprocessed. In this stage first, the dataset is checked to remove null values, then all the categorical data is converted to numeric data. After preprocessing the dataset is divided into test and train data. For the model, we have used a bagging ensemble method. In bagging two classifiers are

used: decision tree classifier and random forest classifier. Training of the model is done by using the training dataset. For checking the accuracy of the model confusion matrix is used which determines the support, recall, precision, and f-score of an algorithm. Dataset loading: The dataset to be used for the machine learning model is loaded into the jupyter notebook by using the pandas library. `read_csv()` in the library is used to read a CSV file. Data Pre-processing: The procedure followed in machine learning for preparing raw data is known as Data preprocessing. It is the most important stage in the evolution of a machine-learning model. The most important aspects of data preprocessing are still data cleaning, integration, reduction, and transformation. The practice of removing erroneous, incomplete, and inaccurate data from datasets, as well as replacing missing information, is known as data cleaning. Creating a single dataset using data from several sources. The data integration process is one of the most crucial parts of data management. There are a few factors to consider during data integration. There are a few things to bear in mind while creating a machine-learning project. By preparing data, we make it easier to interpret and use. This approach helps reduce data volume, making analysis easier while maintaining the same or almost the same findings. This process eliminates data inconsistencies or duplicates that may otherwise reduce a model's accuracy. As a result of human errors or faults, data preparation ensures that no wrong or missing values are included. It is critical to eliminate any noisy and superfluous data from the dataset before utilizing it. The data in the dataset we utilized was first validated for null values, then all categorical data was transformed into numeric data.

Splitting the data: The data is split to test and train data. This approach can be used to find the model hyper-parameter and estimate the generalization performance. Ideally, the data should be separated into three sets: train, test, and holdout cross-validation or development (dev). To split the dataset sci-kit learn library in python is used. The method `train_test_split` is used and the dataset is divided into 20% test and 80% training data. The data should be separated so that neither is excessively large, which is more dependent on the amount of data you have. The dataset must be segmented in charge of conducting an unbiased estimate of predictive accuracy. When splitting a dataset there are two competing concerns: Your parameter estimates will have more volatility if you have less training data. And if you have less testing data, your performance statistics will have greater variance.

Training of the model: The model has been trained by using training data. The following classifiers are utilized in this research. To train an ML model, we must provide the following information: Input the source of your training data. The name of the data attribute that holds the projected target, Instructions for data transformation is required. The learning algorithm is controlled by training parameters. For binary classification tasks, machine learning models anticipate a binary outcome (one of two possible classes).

Gaussian Naïve Bayes: Gaussian Naïve Bayes is a probabilistic classification algorithm based on applying Bayes' theorem with strong independence assumptions. It is a simple classification technique but has high functionality. They are useful when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naïve Bayes Classifier.

Bagging classifier: A bagging classifier is an ensemble meta-estimator that fits individual base classifiers on random subsets of the original dataset before combining their predictions to provide a final prediction. It also aids in the reduction of individual estimator variance. The model is initially trained using this algorithm. To utilize it, we first import the bagging classifier technique from sklearn. ensemble package, and then we fit the training data into the model.

Random forest classifier: Random forest is a machine-learning method that works in groups. Keeping in mind its outstanding performance across a wide range of classification and regression predictive modeling tasks, it is possibly the most popular and commonly used machine learning method. A random forest ensemble is a natural extension of bagging that consists of a group of decision trees. A random forest classifier is another approach used to prepare the model.

Testing the model: Once the model is prepared using a particular algorithm to check the accuracy of the prediction made by the model we use the following method. The testing data is used to first make the prediction and then the accuracy is checked.

Confusion matrix: Confusion matrices are tables that depict how well classification models (or "classifiers") perform on test data whose real values are known. It also reports the right and wrong number count values. The performance of a classification model is evaluated using an NxN matrix, where N is the number of target classes. We import `confusion_matrix` from sklearn.metrics package.

Precision: Precision which is one of the terms in the confusion matrix is calculated as the ratio of tp to (tp + fp), where tp is the number of true positives, fp is the number of false positives. It's critical to make sure we're receiving the right results. A good classifier should have a precision of 1 (high).

Recall: A recall is calculated as TP by (TP + FN), where TP stands for true positives, and FN stands for false positives. All positive samples must be located by the classifier to be included in the recall calculation. 1 is the best value, while 0 is the worst.

F1-score: The proportional mean of reliability and recall is mostly used for the F1 score. The metric to check how accurately a model is working when a dataset is provided to it is considered as f1- score. The greater the f1-score, the better the outcome.

This metric is calculated as:

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 1. Formula

Support: It appears to imply that the number of instances of each class in the true replies is the support (responses in your test set). It may be calculated by adding the rows of the confusion matrix together. **Classification Report:** It's a text report that summarizes the most important categorization metrics.

IV. PROPOSED ARCHITECTURE

Clinically, it could be tough to analyze and conclude the high-risk patients of different stages of colorectal cancer and further subject them to chemotherapy. The importance of this study was to identify a gene signature that could be used to identify colorectal cancer patients with a high risk of recurrence using gene expression data. We proposed an ensemble-based model in this study. Bagging mechanisms are used in ensemble models. The bagging approach is one of the most powerful ensemble methods for reducing variance and, by extension, preventing overfitting. A bagging classifier is a type of ensemble meta-estimator that uses randomized subsets of the original dataset to train base classifiers, then aggregates their predictions (through voting or averaging) to provide a final prediction. The two classifiers which are used under bagging are the decision tree classifier and the random forest classifier. There might be particular features of the data that significantly impact our resulting set of differentially expressed genes and subsequent validated results. The mechanism by which a gene in a cell is turned on to synthesize Ribonucleic Acid and proteins. Gene expression can be evaluated by examining the Ribonucleic Acid, the protein produced first from Ribonucleic Acid, or the function of the protein in the cell. In order to achieve a more stable and efficient prediction performance, ensemble classifiers can be utilized to find out about the final prediction.

Architecture

A system's behavior, structure, and viewpoints are defined by the system architecture, which is a conceptual model. An architectural description is a formal description or representation of a system arranged in such a way that reasoning about the system's structures and behaviors is facilitated.

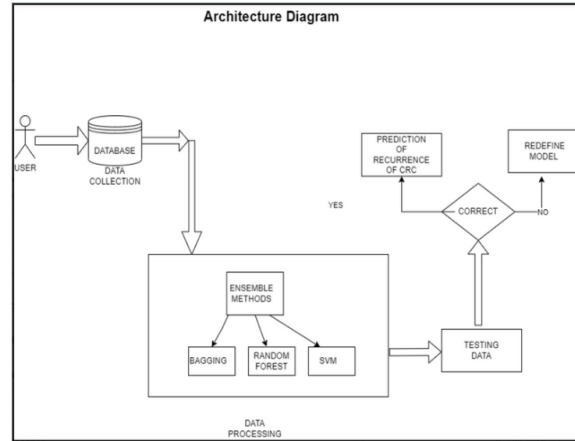


Figure 2. System Architecture

Fig 2 shows the architecture of our system which describes all the processes, methods, functions, and many more which are involved in our model from the input to the output. First, we collect our data and create a database. Next, we pre-process our data to remove noisy or irrelevant data. The dataset is then divided into two parts: testing and training data. To prepare a model for prediction we use ensemble methods. The ensemble methods we are using for our model include bagging and random forest. After the machine learning model is coded we move on to testing the data. If the data is not accurate we need to redefine the model else, our prediction of the recurrence of the colorectal cancer model is ready.

V. RESULTS

The model predicts the recurrence of colorectal cancer in patients using the gene expression data available in the dataset as attributes. A classification algorithm's performance is influenced by the confusion matrix. Using a confusion matrix, one can visualize and analyze the results of a classification algorithm. An F1 score is a superior measurement of accuracy since it is a harmonic mean of precision and recall. In conclusion, identification of the classifier's capacity to detect all of the positive samples. Two classifiers under the bagging ensemble method are used for preparing the model. Each model made using the two classifiers is tested for checking the accuracy model. To check the accuracy along with support, recall, precision, and f1-score of the model confusion matrix are used.

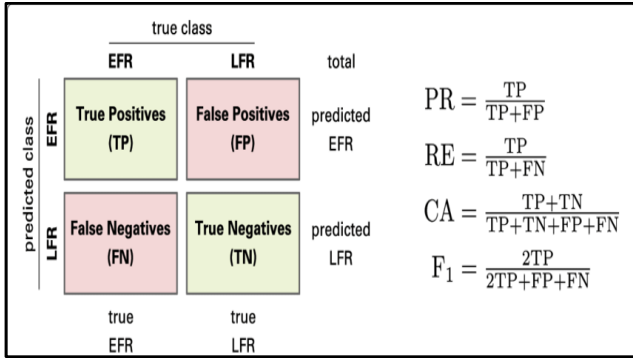


Figure 3. Confusion Matrix Formulas

Fig 3 shows the formulas used in calculating the recall, support, f1-score, and precision of the model.

```

TRAINING RESULTS:
=====
CONFUSION MATRIX:
[[20 0]
 [ 0 29]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
      0   1 accuracy macro avg weighted avg
precision 1.00 1.00    1.00    1.00    1.00
recall    1.00 1.00    1.00    1.00    1.00
f1-score  1.00 1.00    1.00    1.00    1.00
support   20.00 29.00    1.00    49.00    49.00
TESTING RESULTS:
=====
CONFUSION MATRIX:
[[2 3]
 [2 6]]
ACCURACY SCORE:
0.6154
CLASSIFICATION REPORT:
      0   1 accuracy macro avg weighted avg
precision 0.50 0.67    0.62    0.58    0.60
recall    0.40 0.75    0.62    0.57    0.62
f1-score  0.44 0.71    0.62    0.58    0.61
support   5.00 8.00    0.62    13.00    13.00

```

Figure 4. Result of BaggingClassifier.

Fig 4 shows the results of the model made using a decision tree classifier. A confusion matrix is used to check the model's correctness. In the confusion matrix support, recall, f1-score, and support for the model are calculated.

```

TRAINING RESULTS:
=====
CONFUSION MATRIX:
[[20 0]
 [ 0 29]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
      0   1 accuracy macro avg weighted avg
precision 1.00 1.00    1.00    1.00    1.00
recall    1.00 1.00    1.00    1.00    1.00
f1-score  1.00 1.00    1.00    1.00    1.00
support   20.00 29.00    1.00    49.00    49.00
TESTING RESULTS:
=====
CONFUSION MATRIX:
[[0 5]
 [1 7]]
ACCURACY SCORE:
0.5385
CLASSIFICATION REPORT:
      0   1 accuracy macro avg weighted avg
precision 0.00 0.58    0.54    0.29    0.36
recall    0.00 0.88    0.54    0.44    0.54
f1-score  0.00 0.70    0.54    0.35    0.43
support   5.00 8.00    0.54    13.00    13.00

```

Figure 5. Result of RandomForestClassifier

Fig 5 shows the results acquired on testing the model using testing data. The model uses a random forest classifier. To check the performance of the model we use a confusion matrix.

```

TRAINING RESULTS:
=====
CONFUSION MATRIX:
[[20 0]
 [ 0 29]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
      0   1 accuracy macro avg weighted avg
precision 1.00 1.00    1.00    1.00    1.00
recall    1.00 1.00    1.00    1.00    1.00
f1-score  1.00 1.00    1.00    1.00    1.00
support   20.00 29.00    1.00    49.00    49.00
TESTING RESULTS:
=====
CONFUSION MATRIX:
[[0 5]
 [2 6]]
ACCURACY SCORE:
0.4615
CLASSIFICATION REPORT:
      0   1 accuracy macro avg weighted avg
precision 0.00 0.55    0.46    0.27    0.34
recall    0.00 0.75    0.46    0.38    0.46
f1-score  0.00 0.63    0.46    0.32    0.39
support   5.00 8.00    0.46    13.00    13.00

```

Figure 6. Result of Gaussian Naïve Bayes

Fig 6 shows the results acquired on testing the model using testing data. The model is based on Gaussian Naïve Bayes. To check the performance of the model we use a confusion matrix.

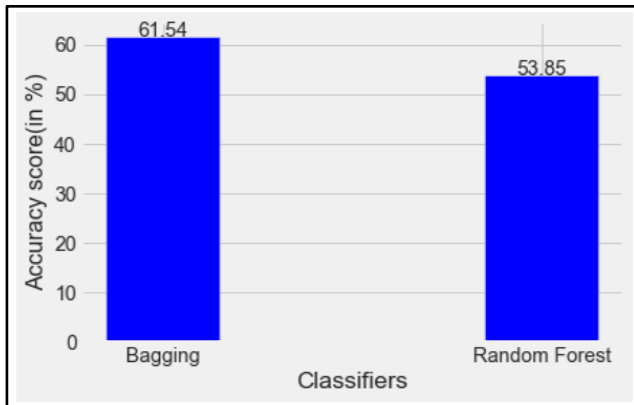


Figure 7. Bar graph representing the accuracy score of the classifiers.

Fig 7 gives a graphical depiction of the classifiers' accuracy scores. On the x-axis, we have the classifier names used in this project, and on the y-axis, we have the accuracy score in percentages.

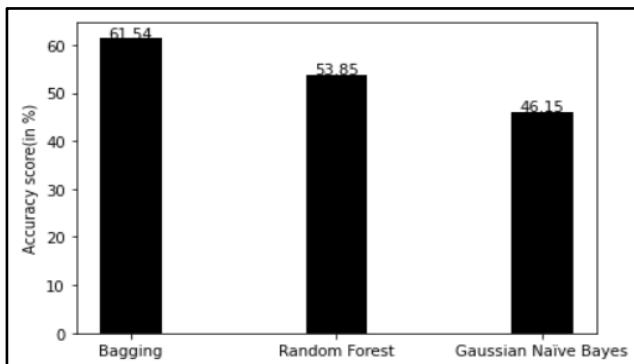


Figure 8. Bar graph representing the accuracy score of different machine-learning models.

Fig 8 gives a graphical depiction of the model's accuracy scores. On the x-axis, we have the algorithm names used in this project, and on the y-axis, we have the accuracy score in percentages.

VI. CONCLUSION

Based on the obtained results, the developed models displayed a high predictive performance. Compared to basic classifiers, ensemble models provided better results. In the process of checking and comparing various classifiers, the result seemed to favor the usage of ensemble methods. Boosting reduces the model's bias while bagging decreases model variance. An algorithm for assigning a class label to data input is known as a classifier in data science. In this project, two ensemble methods are used to train a machine-learning model. Bagging and random forest classifiers are the two classifiers. The use of a decision tree classifier as the base estimator in the construction of the bagging classifier is done. According to the results, the bagging classifier produced higher accuracy than the random forest classifier.

V. FURTHER ENHANCEMENTS

Compared to basic classifiers, ensemble methods produce more accurate results. The dataset that we have referred to in this project is comparatively of less size than required, so we can further improve the accuracy of the existing model by including a larger dataset with more related attributes with less noisy data. Under ensemble methods, we have explored only bagging methods in further improvements the other methods under ensemble that are boosting and stacking can be used to check whether that increases the accuracy or not.

REFERENCES

- [1] Han-Ching Chan, Amrita Chattopadhyay, Eric Y. Chuang, Tzu-Pin Lu, 2021. Development of a Gene-Based Prediction Model for Recurrence of Colorectal Cancer Using an Ensemble Learning Algorithm.
- [2] Mohamad Amin Pourhoseingholi, Sedigheh Kheirian, & Mohammad Reza Zali, 2017. Comparison of Basic and Ensemble Data Mining Methods in Predicting 5-Year Survival of Colorectal Cancer Patients.
- [3] Amit Kumar, Adesh Kumar, G.S. Tomar "Hardware Chip Performance of CORDIC Based OFDM Transceiver for Wireless Communication", Computer Systems Science & Engineering, Vol.40, No.2, pp.645-659, 2022. Global trends in colorectal cancer mortality: projections to the year 2035. *Int J Cancer* (2019) 144(12):2992–3000.
- [4] Varghese A. Chemotherapy for Stage II Colon Cancer. *Clinics Colon Rectal Surg* (2015) 28(4):256–61. DOI: 10.1055/s-0035-1564430
- [5] Ensemble Methods in Machine Learning: What are They and Why Use Them? 2017
- [6] Taiwan Cancer Registry Annual Report of 2016. E. Osterman, J. Ekström, T. Sjöblom, 2021. Accurate population-based model for individual prediction of colon cancer recurrence.
- [7] R. Gary, J. Barnwell, C. McConkey, R.K. Hills, N.S. Williams, and D.J. Kerr. A randomized trial compared the adjuvant treatment to observation in patients with colorectal cancer. *The Lancet* 370(9604):2020–9.
- [8] Fathy SK. A prediction survival model for colorectal cancer. Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications. 2011:36–42
- [9] Using an Ensemble Learning Algorithm, researchers developed a gene-based prediction model for colorectal cancer recurrence. <https://www.frontiersin.org/articles/10.3389/fonc.2021.631056/full>
- [10] Development of a Gene-Based Prediction Model for Recurrence of Colorectal Cancer Using an Ensemble Learning Algorithm <https://www.frontiersin.org/articles/10.3389/fonc.2021.631056/full>
- [11] Comparison of Basic and Ensemble Data Mining Methods in Predicting 5-Year Survival of Colorectal Cancer Patients: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5723205/>
- [12] PB Snow, DJ Kerr, JM Brandt, and DM Rodvold. 5-year survival following colon cancer treatment was predicted using a neural network and regression. *Cancer*, 91(S8), 1673–8. [PubMed] [Source: Google Scholar]
- [13] Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI, Kourou K. Applications of machine learning in cancer prognosis and prediction *Journal of computational and structural biotechnology*, 13:8-17, 2015. [Free article from PMC] [PubMed] [Source: Google Scholar]
- [14] A 16-year study of factors predicting rectal cancer recurrence after curative resection <https://wjso.biomedcentral.com/articles/10.1186/s12957-019-1718-1>
- [15] Atef Zaguia, Vikram Raju, Y. Jeevan Nagendra Kumar, Umashankar Rawat, "Secure Vertical Handover to NEMO using Hybrid Cryptosystem" Hindawi, Article ID 6751423, Hindawi.

- [15] Y. Jeevan Nagendra Kumar, Dr. T. V. Rajini Kanth, "GIS-MAP Based Spatial Analysis of Rainfall Data of Andhra Pradesh and Telangana States Using R", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol 7, No 1, February 2017, Scopus Indexed Journal, ISSN: 2088-8708
- [16] B Sankara Babu, A Suneetha, G. Charles Babu, Y. Jeevan Nagendra Kumar, G. Karuna, "Medical Disease Prediction using Grey Wolf optimization and Auto Encoder based Recurrent Neural Network" *Periodicals of Engineering and Natural Sciences*, Vol 6 Issue 1 Pg: 229-240 ISSN 2303-4521 Jun 2018
- [17] Y. Jeevan Nagendra Kumar, Guntreddi Sai Kiran, Partapu Preetham, Chila Lohith, Guntha Sai Roshik, G. Vijendar Reddy, "A Data Science View on Effects of Agriculture & Industry Sector on the GDP of India" *International Journal of Recent Technology and Engineering*, Volume-8, Issue-1, May 2019, ISBN: 2277-3878.
- [18] Y. Jeevan Nagendra Kumar, Guntreddi Sai Kiran, Partapu Preetham, Chila Lohith, Guntha Sai Roshik, G. Vijendar Reddy, "A Data Science View on Effects of Agriculture & Industry Sector on the GDP of India" *International Journal of Recent Technology and Engineering*, Volume-8, Issue-1, May 2019, ISSN: 2277-3878
- [19] Y. Jeevan Nagendra Kumar, N. Kameswari Shalini, P.K. Abhilash, K. Sandeep, D. Indira, "Prediction of Diabetes using Machine Learning" *International Journal of Innovative Technology and Exploring Engineering*, ISSN: 2278-3075, Volume-8 Issue-7 May 2019
- [20] Y. Jeevan Nagendra Kumar, B. Mani Sai, Varagiri Shailaja, Singanamalli Renuka, Bharathi Panduri, "Python NLTK Sentiment Inspection using Naïve Bayes Classifier" *International Journal of Recent Technology and Engineering*, ISSN: 2277-3878, Volume-8, Issue-2S11, Sep 2019
- [21] Srikanth Bethu, V Sowmya, B Sankara Babu, G Charles Babu, Y. Jeevan Nagendra Kumar, "Data Science: Identifying influencers in Social Networks", *Periodicals of Engineering and Natural Sciences*, ISSN 2303-4521 Vol.6, No.1, pp. 215~228
- [22] Y Jeevan Nagendra Kumar, V Spandana, VS Vaishnavi, K Neha, VGRR Devi, "Supervised Machine Learning approach for Crop Prediction in Agriculture Sector", *IEEE - 5th International Conference on Communication and Electronics Systems (ICCES)*, ISBN: 978-1-7281-5370-4 pg: 736-741
- [23] D. Srinivasa Rao, Ch. Ramesh Babu, Y. J. Nagendra Kumar, N. Rajasekhar, T. Ravi, "Medical Image Fusion Using Transform Based Fusion Techniques", *International Journal of Recent Technology and Engineering*, Volume-8 Issue-2 ISSN: 2277-3878
- [24] Atef Zaguia, Vikram Raju, Y. Jeevan Nagendra Kumar, Umashankar Rawat, "Secure Vertical Handover to NEMO using Hybrid Cryptosystem" *Hindawi*, Article ID 6751423, Hindawi