Journal of Cardiovascular Disease Research

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

MACHINE LEARNING ALGORITHM FOR ANALYSIS AND PREDICTION OF BREAST CANCER

Subba Reddy Borra[1], B. Samyuktha[2], B. Aishwarya[2], Ch. Hima Bindhu[2], D. Sneha[2]

[1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

ABSTRACT

Cancer incidence and mortality have been increasing at an accelerated pace over the past 3 decades globally, making cancer the major public health problem. Among females, breast cancer is known as the most diagnosed cancer and the main cause of cancer deaths in more than 100 countries. In 2018, there are about 2.1 million newly diagnosed breast cancer cases around the world, responsible for nearly 1 in 4 cancer cases among females. However, the causes of breast cancer are still not clearly known to doctors. Early diagnosis of breast cancer can make the disease easier to treat. Several diagnosis techniques are commonly used to distinguish malignant breast tumors from benign ones. Fine Needle Aspiration (FNA) is a well-known procedure used to diagnose breast cancer, but it suffers from a lack of satisfactory diagnosis performance. For FNA, radiologist, oncologist, and pathologist are required to render final judgment together in breast cancer diagnosis, which is time-consuming. Also, there is higher possibility to give rise to errors due to exhaustion or inexperience, which panic patients when false-positive result happens or miss optimum treatment time when false-negative result appears. Therefore, developing an efficient diagnosis support system to assist doctors' diagnosis of cancer has great significance for medical diagnosis process.

Keywords: Breast cancer, FNA, Machine learning.

1. INTRODUCTION

The Surveillance, Epidemiology, 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 demographics, cancer type and site, stage, first course of treatment, and follow-up vital status. 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 website on submitting a SEER limited-use data agreement form.

Prostate is a small-sized gland located underneath the bladder and anterior to the rectum in the pelvis area. These glands are regarded as a part of male reproductive system that create and contain seminal fluid. The Prostate consists of several cell types. Nevertheless, most of the prostate cancers develop in the glandular cells. Prostate cancer is considered the second most frequently diagnosed malignancy in men and ranked as the fifth major cause of death internationally. This cancer is also the most commonly occurring type of cancer in 105 countries, including Australia and the United States. Moreover, it is common among African-Americans with a doubled death rate compared to white men. Maintaining certain diet criteria and physical activities are considered important factors for the progression of the cancer. Yet many unknown facts can influence the patients' treatment. That being the case, such studies facilitate the decision making of medical authorities and physicians regarding the most necessary treatments with reference to the patients' race or social status resulting in proper allocation of healthcare resources. Patients and medical experts also need to make confident decisions regarding the use of aggressive curative treatments or comfort-oriented palliative care. Making such

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

decisions can highly affect the survival risk and outcome of treatments. Therefore, using a treatment approach that is proved to be efficient for the survival of patients can improve the quality of care while reducing the costs of care and hospitalization. To find the high-risk patients, machine learning methods can be useful since these techniques can predict the survival status of patients given that historical data is available.

2. LITERATURE SURVEY

Gupta et. al [1] presents an analysis of prediction performance of multiple deep learning approaches. The performance of multiple deep learning models is thoroughly examined to discover which algorithm surpasses the others, followed by an investigation of the network's prediction accuracy. The simulation outcomes indicate that automated prediction models can predict colon cancer patient survival. Deep autoencoders displayed the best performance outcomes attaining 97% accuracy and 95% area under curve-receiver operating characteristic (AUC-ROC).

Sung Mo Ryu et. al [2] compiled spinal ependymoma cases diagnosed between 1973 and 2014 from the Surveillance, Epidemiology, and End Results (SEER) registry. To identify the factors influencing survival, statistical analyses were performed using the Kaplan-Meier method and Cox proportional hazards regression model. In addition, we implemented ML algorithms to predict the OS of patients with spinal ependymoma.

Koçak et. al [3] presented a novel network-based feature extraction method and tested on two cancer cases, namely (1) lung and bronchus cancer and (2) pancreatic cancer. Named as Signed Maximal Frequent Itemset Network, the proposed method uses maximal frequent item sets as actors in a network and extracts features by considering their co-occurrence and structure of the sub-graph. To investigate patterns on prediction, the top ten maximal item sets are selected with the recursive feature elimination method and their distributions are analyzed. In conclusion, survival months are low when the information on the disease was unknown or blank, and higher in case chemotherapy was given and the primary site was labelled, such as head of the pancreas

Du et. al [4] presented the use of popular tree-based machine learning algorithms and compare them to the standard Cox regression as an aim to predict OPCs survival. The predictive models discussed here are based on a large cancer registry dataset incorporating various prognosis factors and different forms of bias. The comparable predictive performance between Cox and tree-based models suggested that these machine learning algorithms provide non-parametric alternatives to Cox regression and are of clinical use for estimating the survival probability of OPCs patients.

Momenzadeh et. al [5] proposed a hybrid methodology for predicting the survivability of patients suffering from prostate cancer by applying the Factor Analysis of Mixed Data (FAMD) algorithm, along with under-sampling methods for the SEER dataset as the pre-processing step prior to the main models, namely XGBoost, random forest (RF), support vector machine (SVM), and logistic regression (LR) with a cross-validation technique for parameter tuning to predict both binary labeled and multi-class labeled (including other causes of death) cases, which has been rarely investigated in other related studies. The sensitivity analysis was done by cluster centroid as an under-sampling method by which the different proportions of the majority and minority classes were examined for training the binary classification.

Venkatesh et. al [6] evaluated different ensemble learning methods for lung cancer survival prediction on the Surveillance, Epidemiology and End Results (SEER) dataset. Data were preprocessed in several steps before applying classification models. The popular ensemble methods Bagging, Adaboost and three classification algorithms, K-Nearest Neighbours, Decision Tree and Neural Networks as base classifiers were evaluated for lung cancer survival prediction. The results

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

empirically showed that ensemble methods are able to evaluate the performance of their base classifiers and they are appropriate methods for analysis of cancer survival.

Huang et. al [7] enrolled 4,696 patients in SEER Database who were 70 years or older, diagnosed with primary early TNBC (larger than 5 mm), from 2010 to 2016. The propensity-score matched method was utilized to reduce covariable imbalance. Univariable and multivariable analyses were used to compare breast cancer-specific survival (BCSS) and overall survival (OS). Nine models were developed by machine learning to predict the 5-year OS and BCSS for patients received chemotherapy.

Kwak et. al [8] used data from the Surveillance Epidemiology and End Results Database to develop and validate the predictive models for LNM in patients with T1, T2 OSCC. Using simple clinical and histopathological data, we developed six ML algorithms to predict LNM. The predictive performance of models was compared.

Yan et. al [9] propose a priori knowledge- and stability-based feature selection (PKSFS) method and develop a novel two-stage heterogeneous stacked ensemble learning model (BQAXR) to predict the survival status of cancer patients. Specifically, PKSFS first obtains the optimal feature subsets from the high-dimensional cancer datasets to guide the subsequent model construction. Then, BQAXR seeks to generate five high-quality heterogeneous learners, among which the shortcomings of the learners are overcome by using improved methods, and integrate them in two stages through the stacked generalization strategy based on optimal feature subsets. To verify the merits of PKSFS and BQAXR, this paper collected the real survival datasets of gastric cancer and skin cancer from the Surveillance, Epidemiology, and End Results (SEER) database of the National Cancer Institute, and conducted extensive numerical experiments from different perspectives based on these two datasets. The accuracy and AUC of the proposed method are 0.8209 and 0.8203 in the gastric cancer dataset, and 0.8336 and 0.8214 in the skin cancer dataset.

Yin et. al [10] aims to explore a deep learning (DL) algorithm for developing a prognostic model and perform survival analyses in SBT patients. Methods The demographic and clinical features of patients with SBTs were extracted from the Surveillance, Epidemiology and End Results (SEER) database. They randomly split the samples into the training set and the validation set at 7:3. Cox proportional hazards (Cox-PH) analysis and the DeepSurv algorithm were used to develop models. The performance of the Cox-PH and DeepSurv models was evaluated using receiver operating characteristic curves, calibration curves, C-statistics and decision-curve analysis (DCA). A Kaplan–Meier (K–M) survival analysis was performed for further explanation on prognostic effect of the Cox-PH model. Results The multivariate analysis demonstrated that seven variables were associated with cancer-specific survival (CSS) (all $p < 0.05$).

Yu et. al [11] collected patients with rectal adenocarcinoma in the United States and older than 20 years who had been added to the SEER database from 2004 to 2015. They divided these patients into training and test cohorts at a ratio of 7:3. The training cohort was used to develop a seven-layer neural network based on the analysis method established by Katzman and colleagues to construct a DeepSurv prediction model. They then used the C-index and calibration plots to evaluate the prediction performance of the DeepSurv model.

3. PROPOSED SYSTEM

3.1 Data Preprocessing in Machine learning

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

Why do we need Data Pre-processing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

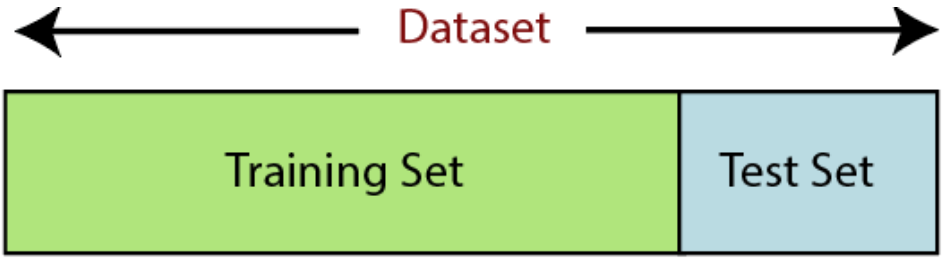


Fig. 1: Splitting of dataset.

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

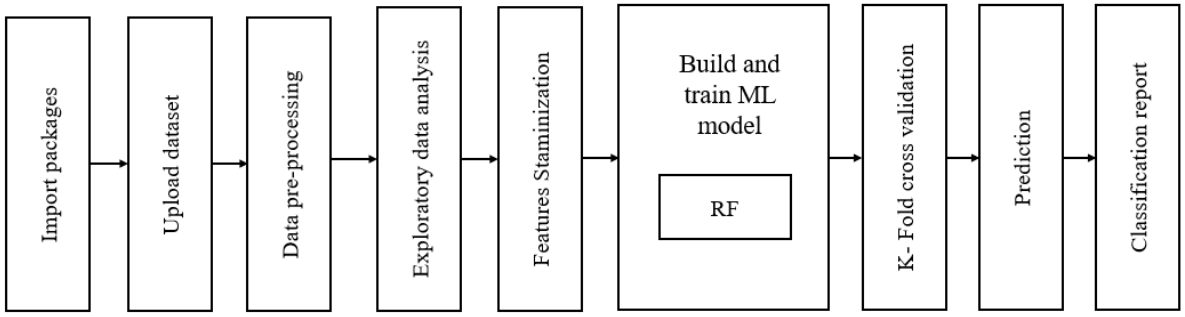


Fig. 2: Block diagram of proposed system.

3.2 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

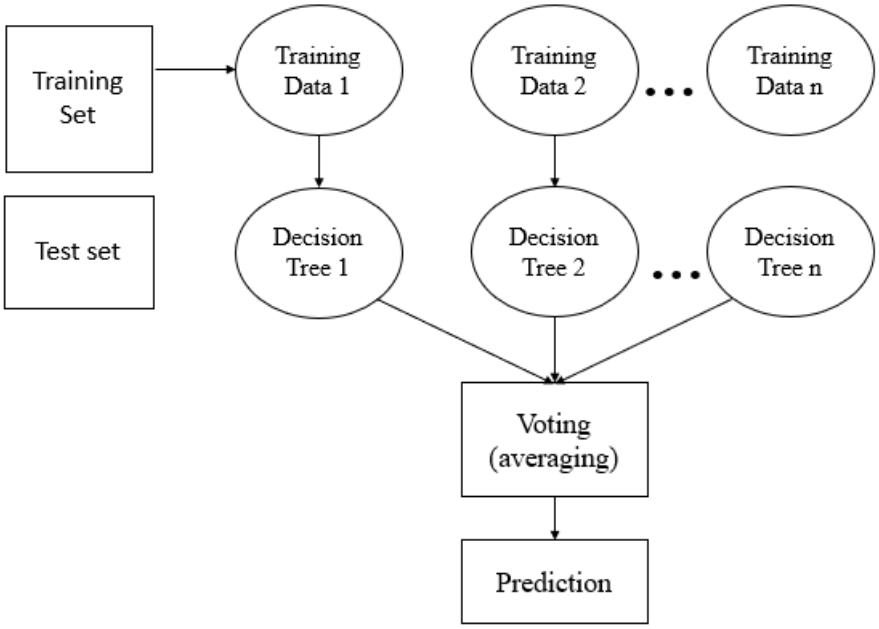


Fig. 3: Random Forest algorithm.

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

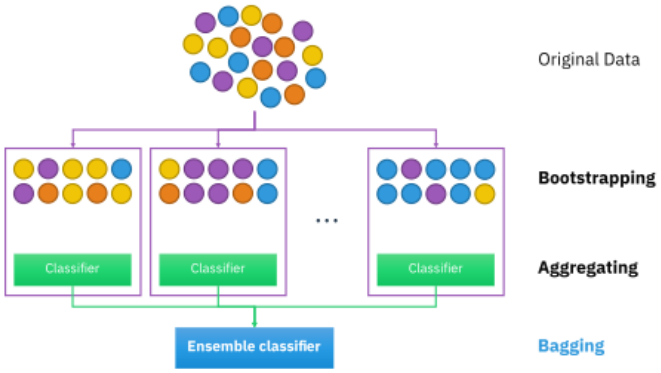


Fig. 4: RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

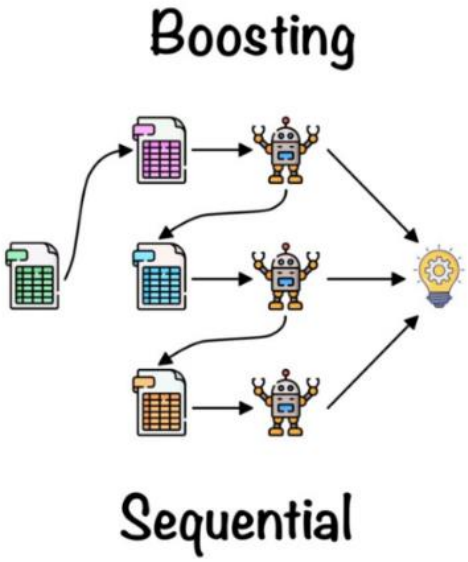


Fig. 5: Boosting RF Classifier.

3.3 Advantages of proposed system

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

Applications of Random Forest: There are mainly four sectors where Random Forest mostly used:

- Banking: Banking sector mostly uses this algorithm for the identification of loan risk.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

- Medicine: With the help of this algorithm, disease trends and risks of the disease scan be identified.
- Land Use: We can identify the areas of similar land use by this algorithm.
- Marketing: Marketing trends can be identified using this algorithm.

4. RESULTS AND DISCUSSION

```
In [2]: seer = pd.read_csv("F:\\21projects\\M.L\\CANCER DETECTION\\CANCER DETECTION\\CODE\\seercancer\\data.csv",header = 0)
        seer.head()
```

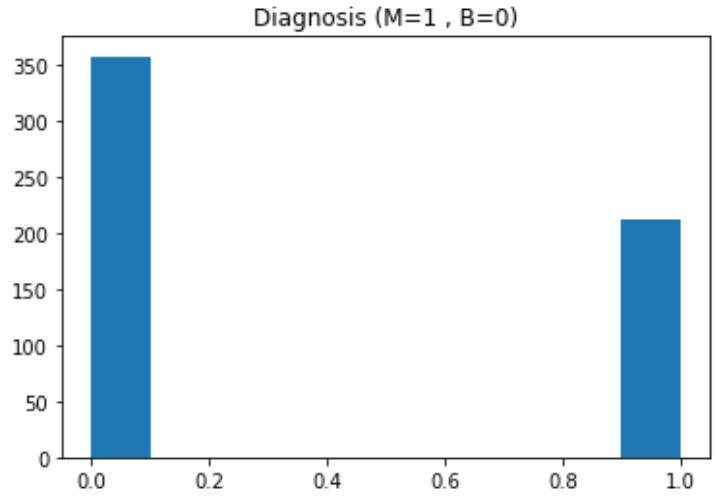
Out[2]:

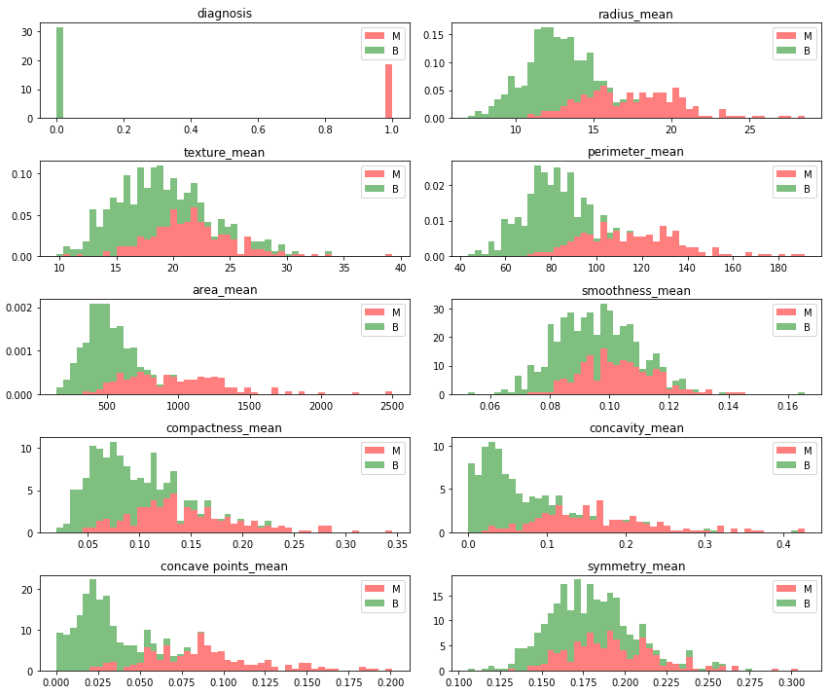
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...

5 rows × 33 columns

```
In [3]: seer.shape
```

Out[3]: (569, 33)





ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

RandomForestClassifier

```
[27]: from sklearn.ensemble import RandomForestClassifier
      RF= RandomForestClassifier(max_depth=5,n_estimators=100)
```

```
[28]: RF.fit(x_train,y_train)
```

```
[28]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                  max_depth=5, max_features='auto', max_leaf_nodes=None,
                  min_impurity_decrease=0.0, min_impurity_split=None,
                  min_samples_leaf=1, min_samples_split=2,
                  min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=1,
                  oob_score=False, random_state=None, verbose=0,
                  warm_start=False)
```

```
[29]: RF.score(x_train,y_train)
```

```
[29]: 0.9925558312655087
```

```
[30]: RF.score(x_test,y_test)
```

```
[30]: 0.9578313253012049
```

```
[31]: RF.score(x,y)
```

```
[31]: 0.9824253075571178
```

```
In [36]: from sklearn.metrics import confusion_matrix, classification_report

         cm_df = pd.DataFrame(confusion_matrix(y_test, y_predict).T, index=RF.classes_,columns=RF.classes_)
         cm_df.index.name = 'Predicted'
         cm_df.columns.name = 'True'
         print(cm_df)
```

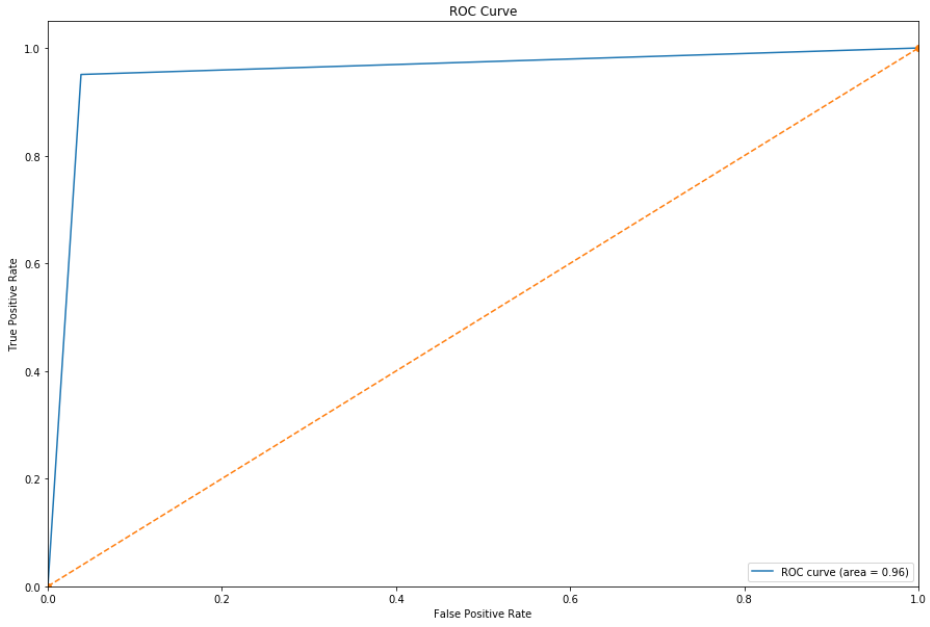
```
True         0   1
Predicted
0          101   3
1            4  58
```

```
In [37]: print(classification_report(y_test, y_predict))
```

```
             precision    recall  f1-score   support

          0       0.97      0.96      0.97       105
          1       0.94      0.95      0.94        61

avg / total       0.96      0.96      0.96       166
```



ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

5. CONCLUSION

Cancer prediction and diagnosis is a complicated issue that has aroused international attention due to the disease's high morbidity and fatality rates. Early accurate prognosis is critical for successful treatment and can improve cancer outcomes. Early identification and strict implementation of curative procedures have been two of the most successful approaches to treating cancer. Also, predicting the outcomes of treatments post cancer like therapies (chemotherapy, immunotherapy, and other related therapies) is very important for estimating the survivability of cancer patients. In this work, we evaluated colon cancer data from the SEER programmed to generate reliable colon cancer survival prediction models. We compared several categorization methods to determine the risk of death five years following diagnosis. Our study discovered that the deep autoencoder model provided the most significant prediction performance in terms of accuracy and area under the receiver operating characteristic curve.

REFERENCES

[1] Gupta, S., Kalaivani, S., Rajasundaram, A., Ameta, G.K., Oleiwi, A.K. and Dugbakie, B.N., 2022. Prediction Performance of Deep Learning for Colon Cancer Survival Prediction on SEER Data. BioMed Research International, 2022.

[2] Sung Mo Ryu, Sun-Ho Lee, Eun-Sang Kim, Whan Eoh, Predicting Survival of Patients with Spinal Ependymoma Using Machine Learning Algorithms with the SEER Database, World Neurosurgery, Volume 124, 2019, Pages e331-e339, ISSN 1878-8750, https://doi.org/10.1016/j.wneu.2018.12.091.

[3] Koçak, Y. and Özyer, T., 2021. Analysing SEER cancer data using signed maximal frequent itemset networks. International Journal of Data Mining and Bioinformatics, 26(1-2), pp.20-58.

[4] Du, M., Haag, D.G., Lynch, J.W. and Mittinty, M.N., 2020. Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: analyses based on SEER database. Cancers, 12(10), p.2802.

[5] N. Momenzadeh, H. Hafezalseheh, M.R. Nayebpour, M. Fathian, R. Noorossana, A hybrid machine learning approach for predicting survival of patients with prostate cancer: A SEER-based population study, Informatics in Medicine Unlocked, Volume 27, 2021, 100763, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100763.

[6] Venkatesh, S.P. and Raamesh, L., 2022. Predicting Lung Cancer Survivability: a Machine Learning Ensemble Method on Seer Data.

[7] Huang, K., Zhang, J., Yu, Y. et al. The impact of chemotherapy and survival prediction by machine learning in early Elderly Triple Negative Breast Cancer (eTNBC): a population-based study from the SEER database. BMC Geriatr 22, 268 (2022). https://doi.org/10.1186/s12877-022-02936-5

[8] Kwak, M.S., Eun, Y.G., Lee, J.W. and Lee, Y.C., 2021. Development of a machine learning model for the prediction of nodal metastasis in early T classification oral squamous cell carcinoma: SEER□based population study. Head & Neck, 43(8), pp.2316-2324.

[9] Yan, F., Feng, Y. A two-stage stacked-based heterogeneous ensemble learning for cancer survival prediction. Complex Intell. Syst. 8, 4619–4639 (2022). https://doi.org/10.1007/s40747-022-00791-w

[10] Yin, M., Lin, J., Liu, L., Gao, J., Xu, W., Yu, C., Qu, S., Liu, X., Qian, L., Xu, C. and Zhu, J., 2022. Development of a Deep Learning Model for Malignant Small Bowel Tumors Survival: A SEER-Based Study. Diagnostics, 12(5), p.1247.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

[11] Yu, H., Huang, T., Feng, B. and Lyu, J., 2021. Deep-learning Model for Predicting the Survival of Rectal Adenocarcinoma Patients based on the SEER Database.