SURVIVAL ANALYSIS OF THYROID CANCER PATIENTS USING MACHINE LEARNING

A PROJECT REPORT

***Submitted by***

MANISHA SHARMI M(210701146)

***in partial fulfillment for the award of***

***the degree of***

**BACHELOR OF ENGINEERING**

***in***

**COMPUTER SCIENCE AND ENGINEERING**

**RAJALAKSHMI ENGINEERING**

**COLLEGE ANNA UNIVERSITY,**

**CHENNAI**

**MAY 2024**

**RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI**

**BONAFIDE CERTIFICATE**

Certified that this Thesis titled **“Survival Analysis of Throid cancer patients using machine learning**” is the bonafide work of “MANISHA SHARMI M **(2116210701146)”** who carried out the work under my supervision. Certified

further that to the best of my knowledge the work reported herein does not form part

of any other thesis or dissertation on the basis of which a degree or award was

conferred on an earlier occasion on this or any other candidate.

**SIGNATURE**

Dr . S Senthil Pandi M.E.,Ph.D.,

**PROJECT COORDINATOR**

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Internal Examiner External Examiner**

**ABSTRACT**

Prognostic forecasts are an essential tool for patients and healthcare professionals to use when making treatment decisions, since the medical community is always working to improve the standard of patient care. The creation of novel cancer survival models has exploded in recent years, and the prognosis of patients with malignant illnesses is mostly estimated through the use of categorization approaches. With the use of a portion of the SEER (Surveillance, Epidemiology, and End Results) database, our research investigates many possible outcomes. We create prediction models by utilizing machine learning techniques, particularly by choosing characteristics that are only variable and doing correlation analysis. We examine the effects of different data sets and important factors on discrepancies and irregularities in the results. We train a set of algorithms by identifying seventeen critical variables.

**INTRODUCTION**

Of all the cancer forms, thyroid carcinoma is the most common type that is endocrine cancer, and its frequency is increasing worldwide. Data about thyroid cancer may be found in the Surveillance, Epidemiology, and End Results Program (SEER) of the US National Cancer Institute. Research on cancer has come a long way in the last ten years, especially in the area of risk classification for patients with cancer, which offers exciting new treatment directions. Still, there is an urgent need for more research into the difficulties and complexity of therapy. Examining debates about medical organs and using advanced data mining methods can provide insightful information in this field. Data mining is the process of utilizing clever techniques to extract detailed information from large, complicated databases.

Distinct thyroid cancer (DTC), which includes pituitary and papillary tumors, accounts for over 90% of thyroid cancer cases. 2009 had 37,200 new instances of thyroid illness in the United States, which coincided with the release of the American Thyroid Association's (ATA) most recent guidelines. This figure shot up to more than 63,000 by 2014. The yearly incidence rate increased significantly over time, rising from 4.9 per 100,000 in 1975 to 100,000 in 2009. Many statistical methods have been used to study this escalation, including log logistics, log-normal, Cox regression, and experimental models of Kaplan-Meyer. Notably, when it comes to efficiency and flexibility, contemporary data mining techniques have shown to perform better than conventional ones. Finding the factors that had the most effects on the goal was the first step in our investigation.

Several unique factors are introduced by our study:   
(i) Greater Accuracy: By employing the Random Forest classifier, we were able to beat Moustafa Mourad's work by obtaining a prediction accuracy of 99.30% [7].   
(ii) Improved Data Cleaning: Sophisticated data preparation methods were used to guarantee the accuracy and caliber of our findings, and they were especially designed to deal with the particular flaws in the dataset.   
(iii) Robustness Against Defects: The Logistic Regression model exhibits an impressive degree of accuracy even when substantial dataset impurities are present. Its adaptability to comparable situations may serve as a standard for next versions.   
(iv) characteristics Sensitivity study: To find out which characteristics are most susceptible to contaminants and how they affect the Logistic Regression model's performance, we carried out a thorough sensitivity study.

(v) Validation Integrity: To prevent overfitting and guarantee that our claimed accuracy of 98.77% held true over a variety of dataset partitions, we used cross-validation techniques with rigor.   
(vi) Scalability Responses: Our suggested methods ensure that our findings are broadly applicable by scaling up to bigger datasets with comparable data impurity concerns.   
We were able to analyze balanced and unbalanced results with greater accuracy by utilizing these techniques. Our methodology is firmly said to be superior to current approaches, especially when compared to Mostafa Mourad's work [7].

**LITERATURE REVIEW**

In order to make educated treatment decisions and maybe spare patients from needless medical procedures and the ensuing financial burden, doctors rely on reliable cancer survival estimates. Over the past several decades, there has been an increase in the incidence of thyroid cancer, which is defined as the malignant proliferation of cells in the tissues of the thyroid gland as well as in the US. Patient life expectancy estimates that are too high due to previous treatment errors highlight the need for reliable survival estimates, which are essential for doctors and patients to make the best decisions on courses of action and drugs. The problem of cancer prognosis prediction has drawn interest, and machine learning approaches have emerged as a potential solution. To predict thyroid cancer prognoses, for instance, Moustafa Mourad used machine learning techniques that included a variety of characteristics

Deep learning algorithms have been used to study prostate cancer survival rates in great detail. Using techniques such as K Nearest Neighbors, Support Vector Machines (SVM), Decision Trees, and Naive Bayes, Huaiyu Wen investigated prognostication in prostate cancer using an artificial neural network (ANN). The survival spans were classified as "less than 60 months" and "more than 60 months," with an 85.64% success rate for ANN being the greatest success rate.Mitra Montazeri also created a rule-based framework for classifying breast cancer survivorship. In their work, 900 patients' worth of data were subjected to a 10-fold cross-validation procedure. Numerous algorithms were used by them, such as Support Vector Machine (SVM), RBF Network (RBFN), 1-Nearest Neighbor (1NN), AdaBoost (AD), Trees Random Forest (TRF), Naive Bayes (NB), and Multilayer Perceptron (MLP).With 96% accuracy, the Trees Random Forest method was superior.Zeming Liu [17] looked at how ETE affects cancer prognosis and survival using the SEER dataset of 107,114 thyroid cancer records. A machine learning-based random forest was developed by Yong Hong Liu [18] to forecast a low quality of life following thyroid cancer. A study of 266 patients with thyroid cancer who had thyroidectomies was conducted. AUCs for training and validation courts were 0.834 and 0.897, respectively. Matja Kukar [19] used machine learning to predict the survival of anaplastic thyroid carcinoma. They conducted 126 patient enrollments and contrasted statistical research with machine learning.

Ankit Agrawal [20] identified 11 derived characteristics, two of which had considerable predictive value, using SEER data to predict the survival of lung cancer patients. They started the data mining optimizations, dataset validations, and preprocessing. Using a range of strategies and attribute selection methodologies, they identified 13 pertinent traits. Predictive performance was improved by combining the conclusions of five decision tree-based classifiers and using meta-classifiers to further refine predictions.

In the meanwhile, M. Lundin [21] predicted breast cancer survival using an artificial neural network. The area under the ROC curve (AUC) was used to evaluate how well prediction models estimated patient survival rates.Neural network research has improved the reliability of predicting breast cancer survival over several time horizons, such as five, ten, and fifteen years.

To predict breast cancer survival rates, Dursun Delen [28] used data mining techniques. They used logistic regression with two different data mining approaches to create prediction models, such as artificial neural networks and decision trees. Ten-fold cross-validation was used to assess the performance of these models, and the decision tree (C5) model showed an impressive 93.6% accuracy.   
 M. Jajroudi [2] investigated the best artificial neural network (ANN) in a different study, utilizing Logistic Regression with MLP to predict the survival of patients with thyroid cancer. Using evidence and expert discussions as a guide, this method entailed carefully selecting key characteristics from the SEER dataset. 7706 data points and 16 characteristics were examined, and survival rates

Numerous links have been found in the dataset in terms of correlations. In particular, there are favorable connections between "CS lymph nodes" and "Derived AJCC M," "Derived AJCC T," "CS extension," and "CS extension/EOD extension." It shows, on the other hand, inverse relationships with both "Radiation" and "Radiation sequence with surgery." Furthermore, there is little or no link with "Radiation" and none with "Reason no cancer-directed surgery."

The difficulty of handling unbalanced data has not received much attention in earlier studies. Building on the ideas presented in [27], we use the SMOTE approach in our work to solve this problem. We identify the significant class imbalance in our dataset and investigate ways to mitigate it, such as weight balancing and data augmentation. Furthermore, our suggested method incorporates many reliable feature selection strategies, in contrast to the recent work by Mostafa Murad [8], which depended on a restricted collection of seven components for prediction. By identifying and ranking the most important characteristics, these approaches hope to improve patient survival prediction accuracy. Similar studies on predicting survival for patients with breast cancer have been carried out by Thongkam [29] and Delen [28].

Proceeding to "Derived AJCC T," it exhibits a negligible positive correlation with "Derived AJCC M," "CS extension/EOD extension," "Radiation," and "Radiation sequence with surgery." It also exhibits a low positive correlation with "CS lymph nodes," a moderate positive correlation with "CS extension," and a negligible positive correlation with "Derived AJCC T." Furthermore, there is a negligible negative link between it and "Reason no cancer-directed surgery." With respect to "Derived AJCC N," it has a negligible positive correlation with "CS extension/EOD."

We see a weak positive association between "Radiation" and "CS extension" as well as "CS extension/EOD extension." Also showing a favorable association are "CS extension" and "CS extension/EOD extension." Interestingly, there is a substantial positive correlation coefficient of 0.84 between "CS lymph node" and "Derived AJCC N." During the model building stage, we will select one of these variables. Extent of Disease (EOD), Collaborative Stage (CS), and American Joint Committee on Cancer (AJCC) are represented by these acronyms. Furthermore, TNM, an acronym for Thyroid Nodule that is used in the stage of thyroid malignancy, may be shortened as follows:   
- Tumor size (T): Describe the cancer's size. Does it now affect buildings nearby?   
Is there evidence of cancer spreading to nearby lymph nodes? - Lymph node involvement (N)

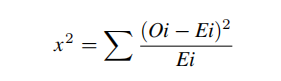
**METHODS**

1. **TRANSFORMATION AND COLLECTION OF DATA**The goal of this study is to use a dataset with a big number of entries in order to take advantage of a substantial sample size for meaningful findings. A thorough grasp of the patterns, trends, and outcomes related to cancer in the United States requires access to The National Cancer Institute's SEER database. Because it collects longitudinal data and covers a wide range of cancer types, this database is an invaluable tool for research, policy development, and cancer control programs. After pulling out 57,155 thyroid cancer-related records from the SEER database, a large number of them had to be disregarded because of missing information. Consequently, a total of 25,217 records are the subject of the analysis.

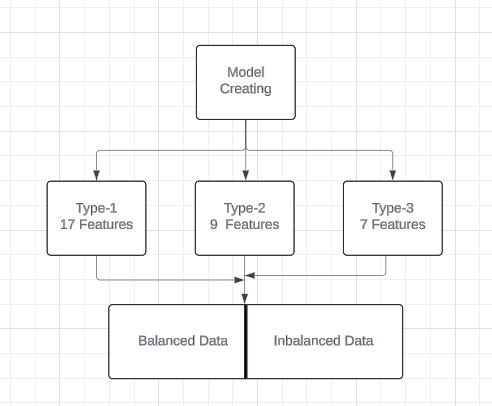
The purpose of our research was to find out how many people with thyroid cancer survive. We used a range of tried-and-true machine learning methods to accomplish this goal. The analysis was carried out in three separate sessions, each of which was further split into two sections in order to fully evaluate the effects of the balanced and unbalanced data sets. We found that the findings differed considerably based on how much weight was given to each component. Because of the intrinsic asymmetry of the data, logistic regression was shown to be the most effective methodology, outperforming other approaches in terms of performance. Impressive results were obtained with logistic regression: 96.77% accuracy rate, 0.76 F1 score, and 0.96 area under the curve (AUC). But notable

1. **FEATURE APPOPRIATION**

The first difficulty is from the diversity of our characteristics, as not all of them have an equal impact on the results we want. Determining the particular factors that matter is critical. We used chi-square analysis, correlation measures, and the pick k-best approach, among other methods, to address this. We also converted the dataset into a numerical representation for these studies using Label Encoding. A column's values are converted into a numeric representation, a process known as label encoding, which is necessary to make labels understandable to computers. For example, have a look at a dataset that includes details on many kinds of bridges, including cantilever, beam, arch, suspension, linked arch, and cable. The textual data are transformed into numerical values by using a sequential encoding technique.

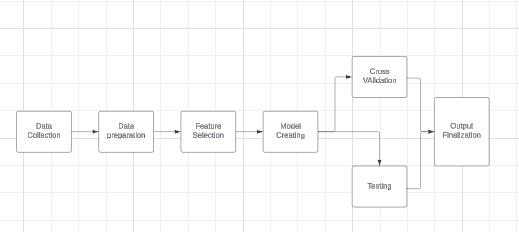


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Race Recode | After on-hot encoding | Race Recode-White | Race Recode-black | Race recode-other |
| White Black Other |  | 1  0  0 | 0  1  0 | 0  0  1 |



1. ***CLASS ENCODING***

The next step involves building a prediction model using machine learning techniques. One-hot encoding is required to make it easier to use nominal characteristics in the model. Utilizing this method improves predictions by converting variables in categorical data into numerical representations. A binary column with the values 1 and 0 represents the existence and absence of each category that makes up a nominal characteristic. Think of a dataset, for example, that has a column called "Race Recode," and it has categories like "black," "white," and other. Setting a value of 1 or 0 to denote the existence or absence of a category in a new column is the first step in implementing one-hot encoding.



**FIGURE 1:** Procedural architecture

1. GENERATING MODELS AND ANALYZING RESULTS   
    A thorough analysis of the results produced by 14 different models was done. A variety of algorithms are included in these models, such as Random Forest, Extra Tree, Ada Boost, K-Nearest Neighbor, Naive Bayes, Logistic Regression, Bagging Classifier, Multilayer Perceptron, Voting Classifier, Gradient Descent, Stochastic Gradient Descent, Hist Gradient Boost, and Light Gradient Boost. The efficacy of every model was evaluated using performance parameters including accuracy score, F1 score, recall, precision, and AUC score, among others. Twenty percent of the dataset's test data were used to calculate these performance indicators. In order to guarantee strong accuracy, 10-fold cross-validation was also used.

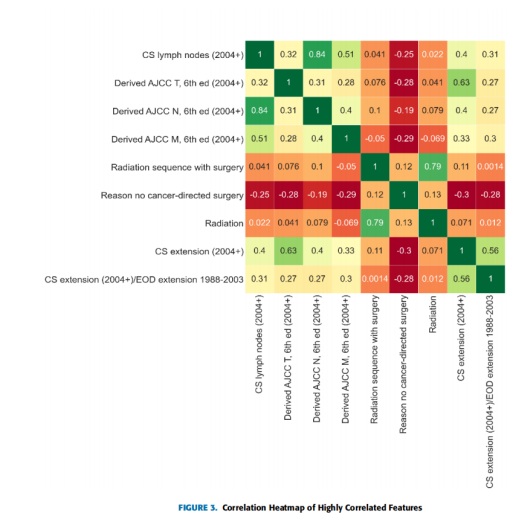
**IV. RESULTS AND DISCUSSION**   
A. ESSENTIAL FEATURE   
 To ascertain the characteristics used in building the model, the Select K-best approach and the chi-square library are applied. To further improve comprehension, the correlation between the characteristics is also calculated. When the correlation coefficient falls between 0.7 and 0.89, it shows a significant positive connection; a correlation value between 0.9 and 1.0 suggests a very strong positive link. The intervals corresponding to moderate positive, low positive, and insignificant correlations are 0.5 to 0.69, 0.3 to 0.49, and 0.0 to 0.29, respectively. A negative correlation coefficient, on the other hand, denotes an inverse link between the qualities.

Correlation research reveals a variety of relationships between the characteristics. "Reason for no cancer-directed surgery" shows a negligible negative association; on the other hand, "Radiation" and "Radiation sequence with surgery" show insignificant positive pairings. On the other hand, there is a somewhat favorable connection with "CS lymph nodes". While "CS extension" and "CS extension/EOD extension" show little positive correlations, "derived AJCC M" has a modest positive association. For "Derived AJCC T," the data show that there is a minimal positive connection with "Derived AJCC M," "CS extension/EOD extension," "Radiation," and "Radiation sequence with surgery." There is a low positive correlation with "CS lymph nodes," a strong positive correlation with "CS extension," and nothing at all with "Derived AJCC T." Additionally, there is a negligible negative connection between "Reason for no cancer-directed surgery" and "Derived AJCC N." Furthermore, there is a negligible positive correlation shown by "Derived AJCC N".

We see a weak positive association between "Radiation" and "CS extension" as well as "CS extension/EOD extension." Also showing a favorable association are "CS extension" and "CS extension/EOD extension." Interestingly, there is a substantial positive correlation coefficient of 0.84 between "CS lymph node" and "Derived AJCC N." During the model building stage, we will select one of these variables. Extent of Disease (EOD), Collaborative Stage (CS), and American Joint Committee on Cancer (AJCC) are represented by these acronyms. Furthermore, TNM, an acronym for Thyroid Nodule that is used in the stage of thyroid malignancy, may be shortened as follows:   
- Tumor size (T): Describe the cancer's size. Does it now affect buildings nearby?   
Is there evidence of cancer spreading to nearby lymph nodes? - Lymph node involvement (N)

|  |  |  |
| --- | --- | --- |
| SL. | Features | Scores |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17 | CS tumour size  Derived AJCC  CS extension  Age  Derived AJCC T  Derived AJCC M  Derived AJCC N  Regional nodes examined  CS Lymph nodes  Reason no cancer directed  Regional nodes positive  CS Meta at dx  radiation  Sex  ICD-O-3 Hist/Behav  RX summ-scope reg LN Sur  Grade | 1738.35364873  4042.34546563  2756.74648978  6739.43248929  1234.34353443  1123.43453556  937.48920878  633.1123230  587.726908  71.894759  56.878964  47.344678  31.823969  22.153566  17.327886  15.647488  11.576456 |

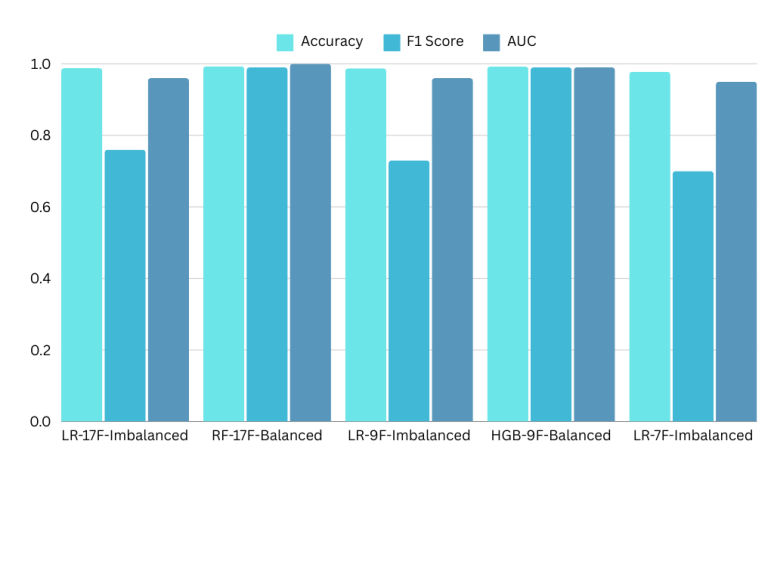
1. PERFORMANCE EVALUATION OF MODELS   
    We have contrasted our work with a number of other research projects that made use of comparable datasets, and our results exceed those of earlier studies (see Table 5). In further research projects, we hope to investigate the functionality of algorithms on more datasets. We used 10-fold cross-validation to improve accuracy; the results are shown in Table 4's last column. We present a description of the model results below:   
     
   1) Class 1: We found 17 characteristics in Type 1 using feature selection. For each kind, two sets of data were created: balanced and unbalanced, with the latter using SMOTE to correct for data imbalances. In every case, fourteen different algorithms were used. The best three algorithmic results are as follows:



98.73%, 98.69%, and 98.67% accuracy percentages were attained by the LR, GBC, and ABC models, respectively, after nine criteria were chosen. These models performed better than any other. Conversely, we found that the HGB, LR, and ABC models had greater accuracy rates—99.2%, 98.87%, and 98.14%, respectively—when balanced data was taken into account. The two best-performing models were LR and HGB. More results were obtained with models constructed with balanced data. Significantly better accuracy was shown by the model with 17 characteristics than by the one with just 9. Under one particular set of conditions, 99.20% accuracy was attained by the LR and HGB models.

1. Class 2: Nine unique feature features with scores of 500 or above have been discovered in this section (Table 3). Figure 5 shows the outcomes of models built with these nine essential characteristics that took into account both balanced and unbalanced datasets. When dealing with uneven data, the LR, GBC, and ABC models perform better than the others. Their respective accuracy percentages were 98.73%, 98.69%, and 98.67%. Nevertheless, HGB, LR, and ABC models outperformed the others in terms of accuracy when considering balanced data, with scores of 99.2%, 98.87%, and 98.14%, respectively. The two most popular models were LR and HGB.

3)Class 3: Using Fisher's discriminant ratio, Relief-F, and Kruskal-Wallis analysis, the researcher [7] determined particular features. Based on their importance, they chose seven characteristics for model training. Using a 19-neuron multilayer perceptron model, their best model produced an accuracy of 94.49%, an F1 score of 0.431, and an area under the curve (AUC) of 0.988. The same seven characteristics were taken into account and a range of machine learning algorithms were tested in our investigation. With an accuracy percentage of 97.75%, the LR, ABC, and HGB models fared better than the others.



V. CONCLUSION AND FUTURE RESEARCH   
 Finding the survival rate for people with thyroid cancer is the goal of our research. Several well-researched machine learning approaches have been used to accomplish this goal. Three of the best machine learning algorithms have been determined by testing almost all of the methods. Additionally, the traits that are most important for this study have been determined. Using k-best and Chi-squared testing is advised in this situation. To evaluate the effects of imbalances and balance in the data, the tests were separated into three distinct stages, each consisting of two halves. For each attribute, we saw different outcomes depending on how important it was in relation to the others. The Random Forest prediction model is the most appropriate, as determined by the asymmetry of the data. Models get more and more

Subsequent studies might concentrate on combining other types of data, such lifestyle, clinical, and genetic data, from different groups of people [44], [45]. To guarantee that the outcomes might be applied more broadly, strategies for expanding these solutions to bigger datasets with comparable impurity problems were also proposed. To increase the accuracy of the results, more datasets might be used to fine-tune the hyperparameters. This thorough method can reveal new prognostic indicators and offer in-depth insights into a patient's health.

When it comes to temporal modeling, the results obtained with random forests were compared to those of other machine learning methods, such as support vector machines, neural networks, and gradient boosting, which are frequently employed in survival prediction tasks. After the first prognosis for survival, we will broaden our research to include long-term survival and prognosis. Our goal is to investigate how models like the Random Forest can forecast survival rates over long periods of time, like five or ten years.

**REFERENCES**

1. X. Wu and colleagues, "LncRNA LINC01061 upregulation mediated by DNA copy number gain predicts poor prognosis and promotes papillary thyroid cancer progression," Biochemical and biophysical research communications, vol. 503, no. 3, pp. 1247–1253, 2018.   
   In 2014, Technology in Cancer Research & Treatment published a paper titled "Prediction of survival in thyroid cancer using data mining technique," written by

[2] M. Jajroudi, T. Baniasadi, L. Kamkar, F. Arbabi, M. Sanei, and M. Ah madzade.   
In Jama, vol. 292, no. 21, pp. 2632–2642, 2004,

[3]S. J. Mandel wrote about a 64-year-old lady who had a thyroid nodule.

[4] A 2003 study paper by S. I. Sherma titled "Thyroid carcinoma," which was published in The Lancet, explores the complexities of thyroid cancer and provides insightful information on both diagnosis and therapy.   
[5]In 2014, the field received a contribution from R. Siegel, J. Ma, Z. Zou, and A. Jemal's paper "Cancer statistics, 2014," which was published in CA: a cancer journal for clinicians. The publication contained extensive statistical information and analysis about cancer incidence and death rates.

[6] A research by J. Llobera et al. titled "Terminal cancer: length and forecasting survival time," which was published in The European Analysis of Cancer in 2000, clarifies the difficult task of determining survival time in patients with terminal cancer and advances knowledge of end-of-life care.   
[7] To improve the precision of the disease's prognosis prediction, M. Mourad et al.'s research article, "Determining thyroid cancer prognosis with validity using machine learning and choosing features applied with SEER data," published in Scientific Reports in 2020, uses cutting-edge machine learning techniques and data from the SEER database.

[8] Research by S. Lee, S. Lim, T. Lee, I. Sung, and S. Kim is presented in a joint effort named "Cancer subtype identification and modeling by route attention and propagation," which was published in Bioinformatics in 2020.   
[9]D. Sun, M. Wang, and A. Li's 2018 paper in IEEE/ACM Transactions, titled "A modal deep neural network model for human breast cancer prognostic prediction by integrating multiple dimensions data," made significant contributions to the fields of biological computation and bioinformatics.   
[10] Reference C. M. Kitahara and J. A. Sosa's paper, "The changing rates of thyroid cancer," which was published in Nature Reviews Endocrinology in 2016, provides insight into the changing patterns in thyroid cancer incidence.