Survival Analysis of Thyroid Cancer Patients Using Machine Learning Algorithms

Article *in* IEEE Access · January 2024

DOI: 10.1109/ACCESS.2024.3392275

Saadat Alhashmi
University of Sharjah
114 PUBLICATIONS **683** CITATIONS

Fazley Rabbe
Daffodil International University
1 PUBLICATION **4** CITATIONS

Shazzad Hossen
Daffodil International University
4 PUBLICATIONS **13** CITATIONS

Nuruzzaman Faruqui
Daffodil International University
39 PUBLICATIONS **618** CITATIONS

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 December 2023, accepted 9 April 2024, date of publication 22 April 2024, date of current version 7 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3392275

RESEARCH ARTICLE

Survival Analysis of Thyroid Cancer Patients Using Machine Learning Algorithms

SAADAT M. ALHASHMI[1], **MD. SHOHIDUL ISLAM POLASH**[2], **AMINUL HAQUE**[2], **FAZLEY RABBE**[3], **SHAZZAD HOSSEN**[2], **NURUZZAMAN FARUQUI**[4], **IBRAHIM ABAKER TARGIO HASHEM**[5], **AND NIRASE FATHIMA ABUBACKER**[6]

[1]Department of Information Systems, University of Sharjah, Sharjah, United Arab Emirates
[2]Department of Computer Science and Engineering, Daffodil International University, Daffodil Smart City, Savar, Dhaka 1216, Bangladesh
[3]Department of Information Technology, Frankfurt University of Applied Sciences, 60318 Frankfurt am Main, Germany
[4]Department of Software Engineering, Daffodil International University, Daffodil Smart City, Savar, Dhaka 1216, Bangladesh
[5]Department of Computer Science, University of Sharjah, Sharjah, United Arab Emirates
[6]School of Computing, Asia Pacific University, Kuala Lumpur 57000, Malaysia

Corresponding author: Aminul Haque (aminul.cse@daffodilvarsity.edu.bd)

ABSTRACT The medical community strives continually to improve the quality of care patients receive. Predictions of prognosis are essential for doctors and patients to choose a course of treatment. Recent years have witnessed the development of numerous new cancer survival prediction models. Most attempts to predict the prognosis of people with malignant growth rely on classification techniques. We could experiment with significantly different results using only a subset of SEER (Surveillance, Epidemiology, and End Results) data. These models were created using machine learning techniques by selecting univariate features and calculating correlations. We illustrated the variation in results and discrepancy of impurity that can result from varying data quantities and critical factors. Seventeen crucial factors were identified, and a group of classification algorithms were trained to evaluate the effectiveness of an estimation technique. In the display mode, the accuracy of these computations ranges from 97% to 99%Ȧlong with accuracy, the models are further evaluated regarding the F1 score, precision, recall, and the AUC score. Compared to earlier studies, a more accurate model has been developed, and, to the best of our knowledge, our prediction model is superior to the models studied in the previous works.

INDEX TERMS Logistic regression, machine learning, random forest, thyroid survivability.

I. INTRODUCTION

Of the various forms of cancer, thyroid carcinoma is the most prevalent endocrine cancer, with a constant increase in prevalence globally [1]. The Surveillance, Epidemiology, and End Results Program (SEER) of the National Cancer Institute, USA, contains much information about thyroid cancer. Over the past decade, progress has been made in cancer studies. This way, categorizing malignant growth patients into danger classes is an incredibly lively subject of study with obvious therapeutic applications. More research on the treatment's intricacy and difficulties is required. This study may be done on how medical organs, such as data mining approaches, are discussed [2]. The data mining approach is a procedure that collects information from complex data through clever tactics. Disclosure can enhance the quality of health and treatment management. Computer science and data mining techniques are used in decision-making systems to consider all relevant factors [2]. Because cancer therapy takes so long and costs so much money, accurate survival projections are a must for providing affordable healthcare. The clinical importance of thyroid nodules varies from 7% to −15% depending on the requirement to rule out thyroid cancer, which in turn varies based on age, sex, radiation exposure date, and family history, among other factors [3]. Above 90% of all thyroid tumors [4] are classified as distinct thyroid cancer (DTC), which includes papillary and pituitary tumors. Around 63,000 new cases of thyroid illness were identified in the United States in 2014 [5], up from 37,200 in 2009 when the most recent ATA recommendations were

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

issued. The number of new cases per year has climbed from 4,9 per 100,000 in 1975 to 100,000 in 2009 (100) 14.3. This trend has been studied using various statistical approaches, including Cox regression, log logistics, log-normal, and Kaplan-Meyer experimental models [6]. New, cutting-edge data mining techniques outperform their more established counterparts in terms of versatility and efficiency. Our investigation began by identifying the factors that affected the objective characteristic most. The Sklearn tools of Select K-Best and Chi-Square have been employed here. The survival of the patient is what we are interested in seeing. Many machine-learning techniques were initially conceived to solve this grouping problem. The novelty of our work:

(i) Improved Accuracy: Provided a 99.30% accurate prediction model using the Random Forest classifier, which is superior to Mourad's work [7].

(ii) Data Cleaning: Introduced sophisticated data preprocessing techniques tailored to tackle the dataset's specific impurities, ensuring the results' quality and reliability.

(iii) Robustness Under Impurity: The Logistic Regression model retains high accuracy even with significant dataset impurities. This robustness could set a benchmark for future models in similar contexts.

(iv) Feature Sensitivity Analysis: Delved deep into understanding which features are most sensitive to impurities and how they influence the Logistic Regression model's performance.

(v) Cross-Validation Techniques: Implemented robust cross-validation strategies, ensuring that the reported 98.77% accuracy is consistent across different dataset splits and is not a result of overfitting.

(vi) Scalable Solutions: Provided methodologies that can scale to larger datasets with similar impurity challenges, ensuring the applicability of our findings to broader contexts.

These methods helped us to get closer to the mark while making predictions—the effects of both a balanced and unbalanced outcome on the evaluation. According to what we know, we can confidently assert that our work is superior to that of others, particularly when compared to Mourad's work [7].

II. LITERATURE REVIEW

Accurate cancer survival prediction should aid doctors in making wise choices and creating successful treatment plans [8]. Meanwhile, it can save many individuals from obtaining unnecessary treatment and the high medical expenditures that accompany it [9]. Malignant cells can develop in the thyroid gland's tissues, a condition known as thyroid cancer. In recent decades, the incidence of thyroid cancer [10] has increased in several countries, including the US. Based on past patient treatment, the doctor sometimes mispredicts longevity. Doctors and patients need survival estimates to choose the best medicine. Many research experts have attempted to solve the task of predicting cancer prognosis using machine learning that aims to estimate entirely accurately.

Mourad et al. [7] predicted thyroid cancer prognoses using machine learning, feature selection, and the SEER dataset. Thirty-four clinical variables and 61,362 items make up the dataset. MPL1 was the most accurate of his ANN-based MLP models at 94.5%Ṡonuc [11] used machine learning to classify thyroid illnesses as usual, hypothyroidism and hyperthyroidism. He used SVM, RF, DT, NB, LR, K-NN, MLP, and linear discriminant analysis on a 14-attribute dataset of Iraqi nationals. MLP has 96.4% accuracy. Wu et al. [12] utilized machine learning to predict central lymph node metastases. This study used 22 unique characteristics. 5-fold cross-validation is employed. The 7-variable gradient boosting decision tree model has the greatest ROC (AUC = 0.731) and decision curves. It was chosen as the top model. Park and Lee [13] used machine learning to predict illness recurrence and analyzed 1040 patients with 12 characteristics. They evaluated sex, tumor size, and disease recurrence. The Decision Tree model had 95% accuracy. Duggal and Shukla [14] predicted thyroid conditions using machine learning. He diagnosed this thyroid disease using feature selection and classification methods. Tree-based, recursive, and univariate feature selections are recommended. Naive Bayes, Support vector machines, and Random Forest classified thyroid illnesses into four classes: Hypothyroid, Hyperthyroid, Sick Euthyroid, and Euthyroid. The SVM classifier was best, with 96.92% accuracy.

Deep learning has also been used extensively to predict prostate cancer survival. Wen et al. [15] studied prostate cancer prognoses and employed an artificial neural network as a form of deep learning. He used Naive Bayes, Decision Trees, K Nearest Neighbors, and Support Vector Machines (SVM). They split survival into fewer than 60 months and more than 60 months. The highest success rate, 85.64% is shown by ANN. Montazeri and Beigzadeh [16] created a rule-based survival classification system for breast cancer. They used Naive Bayes (NB), Trees Random Forest (TRF), 1-Nearest Neighbor (1NN), AdaBoost (AD), Support Vector Machine (SVM), RBF Network (RBFN), and Multilayer Perceptron (MLP) with 10-cross fold approach on a small dataset of 900 patients. They measured model accuracy, precision, sensitivity, specificity, and area under the ROC curve. The Trees Random Forest technique was more accurate (96%).

Liu et al. [17] investigated the SEER dataset of 107,114 thyroid cancer records to see if ETE affected cancer prognosis and survival. Liu et al. [18] created a machine learning-based random forest to predict poor thyroid cancer quality of life. Two hundred sixty thyroidectomy-receiving thyroid cancer individuals were studied. Training and validation courts had 0.834 and 0.897 areas under the curve. Kukar et al. [19] predicted anaplastic thyroid cancer survival using machine learning. They enrolled 126 patients and compared machine learning to statistical studies.

Agrawal et al. [20] used SEER data to forecast lung cancer patient survival. Two of the 11 derived traits they found were highly predictive. Preprocessing, data mining optimizations, and dataset validations commence. They selected 13 attributes using multiple methods and attribute selection methodologies. Ensemble voting of five decision tree-based classifiers and meta-classifiers enhanced prediction. Lundin et al. [21] predicted breast cancer survival using an artificial neural network. The area under the ROC curve (AUC) measured how effectively prediction models predicted patient survival rates. The neural network models' 5-, 10-, and 15-year breast cancer-specific survival AUCs were 0.909, 0.886, and 0.883. Logistic regression AUC values were 0.897, 0.862, and 0.858. Neural network accurately predicts breast cancer survival after 5, 10, and 15 years. Delen et al. [28] predicted breast cancer survival using data mining. They used logistic regression and two data mining approaches to develop prediction models (artificial neural networks and decision trees)—performance comparison made using 10-fold cross-validation. The decision tree (C5) model has 93.6% accuracy. Jajroudi et al. [2] used Logistic Regression with MLP as the optimal neural network ANN for survival prediction of thyroid cancer patients. He used evidence and radiation oncologists to choose important SEER dataset properties. They investigated 16 attributes and 7706 data points. He estimated 1-, 3-, and 5-year survival. He measured model accuracy, sensitivity, and specificity. Their work suggested MLP for thyroid cancer survival prediction. Although approaches and algorithms are well-developed and have an adequate logical foundation, they frequently encounter difficulties because of the size and properties of the underlying data [22].

Yet, the benefits of a large sample size for interpreting significant results include that it permits more accurate estimation of the treatment impact and often makes it simpler to judge the sample's representativeness and generalize the findings [23]. Many treatments classified as "no difference from control" in studies with insufficient samples were unfairly examined. Planning clinical studies should pay more attention to the possibility of missing a critical therapeutic advancement due to limited sample numbers [24].

Since it is clear based on [23] and [24], it is clear that the large sample size in the medical field results in a more precise estimate of the treatment effect, so we have considered collecting extensive data for the prediction of accurate survival analysis. After reviewing the research works by Jajroudi et al. [2], Park and Lee [13], Duggal and Shukla [14], Montazeri and Beigzadeh [16], and Liu et al. [17], it is noticed that they all have used smaller datasets of 7706, 1040, 7200, 900, and 286 records, respectively, hence to obtain the benefit of a large sample size for interpreting significant results, this paper aims to use a dataset with more records.

The other major issue in datasets that contain unevenly distributed data, also known as imbalanced data, is what gives rise to the problem of class imbalance. The class imbalance problem is where data points with class labels have one class instance devalued by the other instances. Class imbalance distribution is typical for real-world medical data, particularly cancer data [25]. The unequal quality of majority and minority classes is one of the critical issues with employing data analysis for diagnosis and therapy [26].

In addition to the significance of data diversity, it is essential to use more methods to narrow down the best algorithms that suit the data and the business requirements. This paper has experimented with 14 different ways. At the same time, Jajroudi et al. [2], Mourad et al. [7], Sonuç et al. [11], Yijun Wu et al. [12], Park and Lee [13], Duggal and Shukla [14], Wen et al. [15], Montazeri and Beigzadeh [16], Liu et al. [17], Lundin et al. [21] just used a few (less than 10) different methods for thyroid, prostate, and breast cancer. Table 1 presents some more recent works related to ours.

In addition, the works mentioned above have yet to deal with the problem of imbalanced data. Our work addresses the data imbalance problem by employing the SMOTE approach based on the work by [27]. A heavy class imbalance is examined as a critical problem for their dataset, and multiple balancing techniques, such as weight balancing and data augmentation, were considered. In addition to this, our proposed work also investigates a few efficient feature selection methods that select only the prominent features to attain greater accuracy in determining whether the patient would survive compared to the recent research work conducted by Lee et al. [8], where only seven elements were employed for the prediction. Likewise, Delen et al. [28] and Thongkam et al. [29] also predicted survival for breast cancer patients. The high-performing model is Logistic regression with 98.77% accuracy in the low degree of impurity. Also, the best-performing model is the Random Forest classifier, with 99.30 correctness for the high degree of impurity.

III. METHODS

Essential procedures include data collection, preparation, feature selection, model creation, cross-validation, and model testing (Fig.1).

A. DATA COLLECTION AND TRANSFORMATION

This paper aims to use a dataset with a higher number of records to obtain the benefit of a large sample size for interpreting significant results. The SEER database from the National Cancer Institute's SEER program is vital for understanding cancer patterns, trends, and outcomes in the United States. Its comprehensive nature, longitudinal data collection, and broad coverage of cancer types make it a valuable tool for research, policy development, and cancer control efforts. We retrieved 57155 records on thyroid cancer from the SEER database; however, we had to drop out a majority of the records due to a significant number of records with missing values identified. Finally, we consider 25217 records in total for the analysis. To further set up the data for analysis, other pre-processing techniques such as feature selection, class imbalance problem, and class encoding are applied to the dataset.

Our study attempted to determine the survivability of thyroid cancer patients. To do this, we have employed several well-known machine learning methods. There were

TABLE 1. Related works.

Paper	Cancer Type	Prediction Type	Year	Dataset Used	Proposed Algorithm	Evaluation Parameter
Deep learning-based multifeature integration robustly predicts central lymph node metastasis in papillary thyroid cancer [38]	Thyroid	Lymph node metastasis	2023	488 patients data collected by them self from Zhuzhou Hospital Xiangya Medical College (2019-2021)	CNN	AUC, in train data 0.87 and in test data 0.76
Diagnosis of Metastatic Lymph Nodes in Patients With Papillary Thyroid Cancer [39]	Thyroid	Metastatic Lymph Nodes	2022	Self Collected	ClymphNet -Deep Learning Based Model	Accuracy 93.0% AUC 0.948, Sensitivity 93.27% Specificity 92.71%
The Role of Machine Learning in Thyroid Cancer Diagnosis [40]	Thyroid	Cancer Detection	2023	Review Paper	ANN, Support Vector Machine, Decision Tree	Accuracy
Model Analysis for Predicting Prostate Cancer Patient's Survival: A SEER Case Study [41]	Prostate	Survivability	2023	SEER Database	XGBoost	Accuracy 89.57%
Survivability Prediction for Patients with Tonsil Cancer Utilizing Machine Learning Algorithms [42]	Tonsil Cancer	Survival	2022	SEER Database (2004-2018)	Random Forest	Accuracy 93.88%
Machine learning–based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy [18]	Thyroid	Quality of life after thyroidectomy	2022	Self Collection	Random Forest	Training AUC 0.834 and Validation AUC 0.897
Transfer Learning Based Breast cancer Classification using One-Hot Encoding Technique [30]	Breast	Cancer Detection	2021	BreakHis dataset	VGG-16	Accuracy 98% for training and 95% for validation

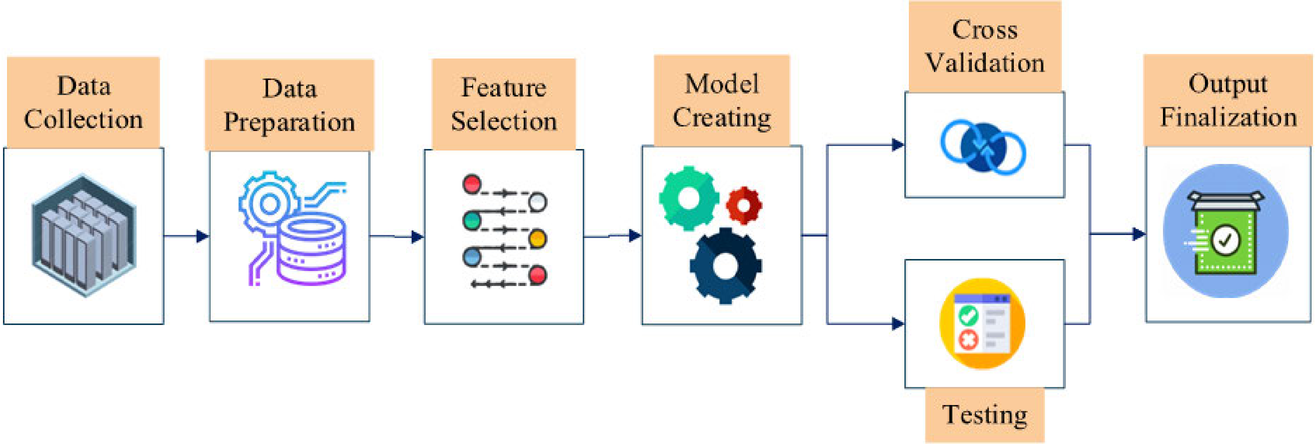


FIGURE 1. Procedural architecture.

three separate sessions when we oversaw the examinations. Each meeting is divided into halves once again, with each half exploring the results of both evenness and imbalance. We have observed outcomes shift in both attributes to varying degrees, depending on their relative relevance. Due to the asymmetry of the data, we discovered that logistic regression outperformed other methods. With an area under the curve (AUC) of 0.96, an F1 score of 0.76, and an accuracy of 98.77%Ṫhe results, however, were enhanced when we normalized the two objective categories. The accuracy was calculated using a random forest classifier to be nearly 100%Ȯur calculations show the results achieved by using the 17 best qualities Using the two scenarios as examples. Moreover, our research indicates that the SMOTE approach is the most advantageous inequality provision.

B. FEATURE SELECTION

The first problem is that not all of our features will have the same impact on our desired characteristics. It's essential to detail the factors that will have an effect. Chi-square and pick k-best were employed, as well as a correlation measure, for this purpose. These two tests have been conducted using

TABLE 2. One-hot encoding transformation.

Race Recode	After one-hot encoding	Race-Recode _white	Race-Recode _black	Race-Recode _other
White		1	0	0
Black		0	1	0
Other		0	0	1

Label Encoding, transforming the dataset into a numerical representation. Label encoding is giving labels a numerical expression that computers can understand. This method is relatively straightforward and requires turning each value in a column into a number. Consider a dataset of bridges with the following column values for "bridge types": arch, beam, truss, cantilever, tied arch, suspension, and cable. We encode the text values by inserting a running sequence for each text value: 0, 1, 2, 3, 4, 5, 6, and 7.

$$x^2 = \sum \frac{(Oi - Ei)^2}{Ei} \tag{1}$$

In equation 1, x^2 indicates the value of chi-squared. Oi and Ei refer to the observed value and expected value, respectively.

C. CLASS ENCODING

The second step is to construct a prediction model using machine learning techniques. We must use one-hot encoding [30] to turn our nominal qualities into numbers. The one-hot encoding technique encodes categorical data variables to improve predictions with machine learning algorithms. Each category of the nominal attribute is encoded as a binary column using a single hot encoding. A 1 is entered into the column if that feature is present, and a 0 otherwise. Take a dataset with the column "Race recode" as an example. "Race recode" has white, black, and other classes. Assigning a 1 or 0 (the true/false notation) to each category value in a brand-new column is one way to implement this tactic (Table 2). If the first column in a row has the value "1" (meaning true), then all subsequent columns in that row will have the value "0" (indicating false); similarly, for any additional rows where the value in that row corresponds to the value in that column.

D. HANDLING IMBALANCED DATA

Fourteen machine learning techniques are used to create prediction models from each group's data. There are a total of two sets of information in each category. One kind has been represented with an unbalanced proportion of the target two classes, while the other type has been designed with a balanced distribution of the two classes. The ratio of 24605:412 is exceptionally skewed and indicative of a low level of impurities. The Synthetic Minority Over-sampling Method (SMOTE) has been implemented to address the issue [31]. To create a new sample, SMOTE picks samples that are near together in the feature space, draws a line between them, and picks a point on the line. This method helps assemble cohesive yet evenly distributed teams. It's a standard method for making datasets fairer. The exact process was used to develop models for each category. Fig. 2 shows how predictive models are made.

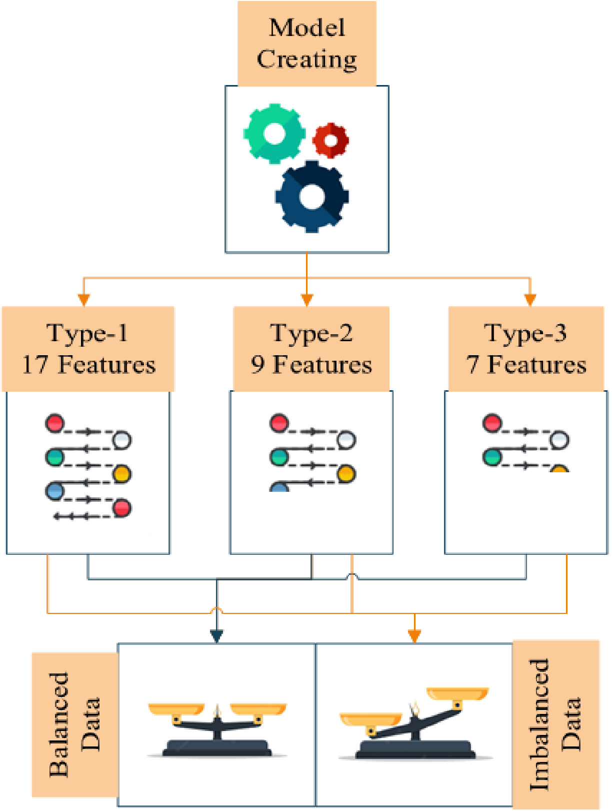


FIGURE 2. Types of predictive models.

E. MODELS CREATION AND RESULT ANALYSIS

A total of 14 models have been created, and their results have been reviewed. Algorithms are used: Decision Tree [16], [18], [20], [23], Random Forest [18], [43], Extra Tree [33], Ada Boost [32], Gradient descent, Stochastic Gradient Descent, Hist Gradient Boost, Light Gradient Boost, K- Nearest Neighbor, Naive Bayes, Logistic Regression, Bagging classifier, Multilayer Perceptron, Voting Classifier. The results of accuracy score, F1 score, recall, precision, AUC score [8], [18], [19], [25], [34], [35], [36], [37], etc., measurement have been used to analyze the model's performance. These performance measurements have been calculated with 20% of test data. In addition, we employ 10-fold cross-validation for obtaining justified accuracies.

IV. RESULTS AND DISCUSSION

A. FEATURE IMPORTANCE

The features with which the model is built are determined with the help of the Select K-best and chi-square library. For better understanding, the correlation among attributes also has been calculated. We already knew that if the correlation coefficient is between 0.9 and 1.0, it has a very high positive

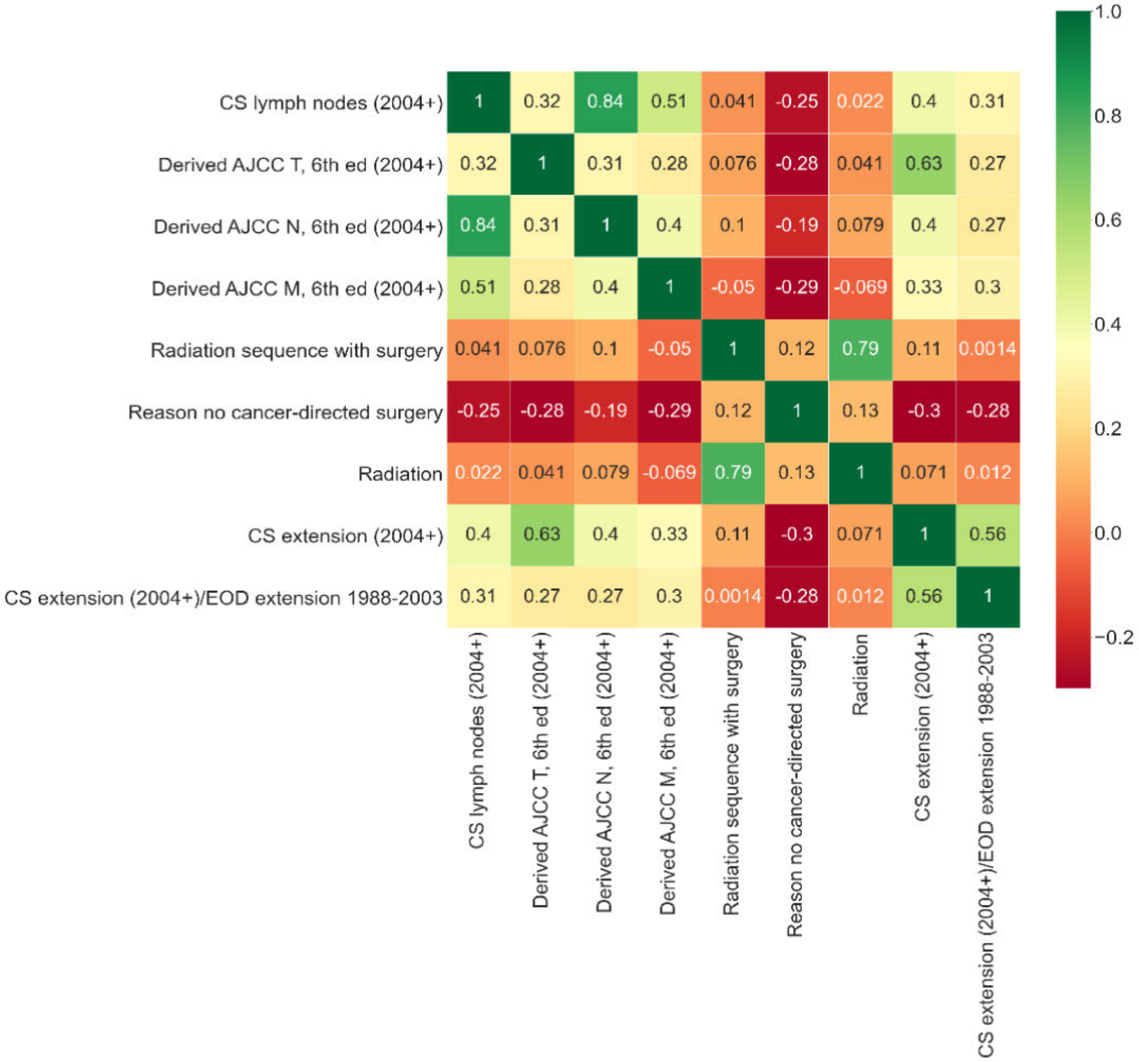


FIGURE 3. Correlation heatmap of highly correlated features.

correlation [35]. If it is 0.7 to 0.89, it has a high positive correlation. Similarly, if the correlation coefficient ranges from 0.5 to 0.69, 0.3 to 0.49, and 0.0 to 0.29, there has been a moderately positive, low positive, and negligible correlation, respectively. On the other hand, the negative value of the correlation coefficient refers to the negative association. The Correlation Heatmap of Highly Correlated Features is shown in Fig. 3.

"CS lymph nodes" has a moderate positive correlation with "Derived AJCC M," a low positive correlation with "Derived AJCC T," "CS extension," and "CS extension/EOD extension," and a negligible positive correlation with "Radiation," and "Radiation sequence with surgery," and insignificant negative Correlation with "Reason no cancer-directed surgery." Therefore, "Derived AJCC T" has a moderate positive correlation with "CS extension," low positive Correlation with "CS lymph nodes," and "Derived AJCC N" has a negligible positive correlation with "Derived AJCC M," "CS extension/EOD extension," "Radiation", and "Radiation sequence with surgery" insignificant negative Correlation with "Reason no cancer-directed surgery." "Derived AJCC N" has a low positive correlation with "Derived AJCC M," "CS extension," and "Derived AJCC T," negligible positive Correlation with "CS extension/EOD extension," "Radiation," and "Radiation sequence with surgery," and insignificant negative Correlation with "Reason no cancer-directed surgery." "Derived AJCC M" has a low positive correlation with "CS extension/EOD extension," and "CS extension," and negligible negative Correlation with "Radiation," "Reason no cancer-directed surgery," and "Radiation sequence with surgery". "Radiation sequence with surgery" has a high positive correlation with "Radiation" and negligible positive Correlation with "Reason no cancer-directed surgery," "CS extension," and "CS extension/EOD extension." "Reason no cancer-directed surgery" has a negligible positive and negative correlation with “Radiation”, and “CS extension/EOD extension,” respectively, and a low negative correlation with “CS extension.” “Radiation” has a negligible positive correlation with “CS extension” and “CS extension/EOD extension.” “CS extension” positively correlates with “CS extension/EOD extension.” Correlation shows that “CS lymph node” and “Derived AJCC N” are highly correlated, with 0.84 scores having a high positive correlation. During the model creation time, we will take one of these two. The abbreviation is AJCC- American Joint Committee on Cancer, CS - Collaborative Stage, and EOD - Extent of the disease. The AJCC-TNM system is used to understand the staging of Thyroid cancer, where TNM is abbreviated as:

- The extent (size) of the tumor (T): How large is the cancer? Has it grown into nearby structures?

- The spread to nearby lymph nodes (N): Has the cancer spread to nearby lymph nodes?

- The spread (metastasis) to distant sites (M): Has the cancer spread to distant organs such as the lungs or liver?

TABLE 3. Feature scores.

SL.	Features	Scores
1.	CS tumour size	8460.206655
2.	Derived AJCC Stage Group	4089.969175
3.	CS extension	2755.952590
4.	Age	1392.608139
5.	Derived AJCC T	1109.025427
6.	Derived AJCC M	1008.956050
7.	Derived AJCC N	919.134665
8.	Regional nodes examined	633.112340
9.	CS lymph nodes	587.726986
10.	Reason no cancer-directed surgery	71.831886
11.	Regional nodes positive	56.755035
12.	CS Mets at dx (2no metastasis no metastasis4+)	47.504682
13.	Radiation	31.515411
14.	Sex	22.153289
15.	ICD-O-3 Hist/Behav	17.326269
16.	RX Summ–Scope Reg LN Sur	15.426450
17.	Grade	11.132441

In Table 3, features with a score higher than ten are reported. The scores are obtained using the select K-best and chi-squared process. Seventeen (17) attributes scored above 10, 9 highly important scored above 500. The top 5 essential features are:’ CS tumor size,” Derived AJCC Stage Group,” CS extension,” Age,’ Derived AJCC T.’ The tumor size is at the top, meaning that the patient’s survival mostly depends on the tumor’s size. Gradually, other essential features are available in Table 3.

B. MODELS PERFORMANCE EVALUATION

We have compared our study with some existing studies that used similar datasets. Our results outperformed the existing studies (Table 5). In our future research study, we will consider exploring the performances of the algorithms for additional datasets. However, to obtain better accuracy, we have applied 10-fold cross-validation, for which the results are presented in the last column of Table 4. The results of the models are discussed below:

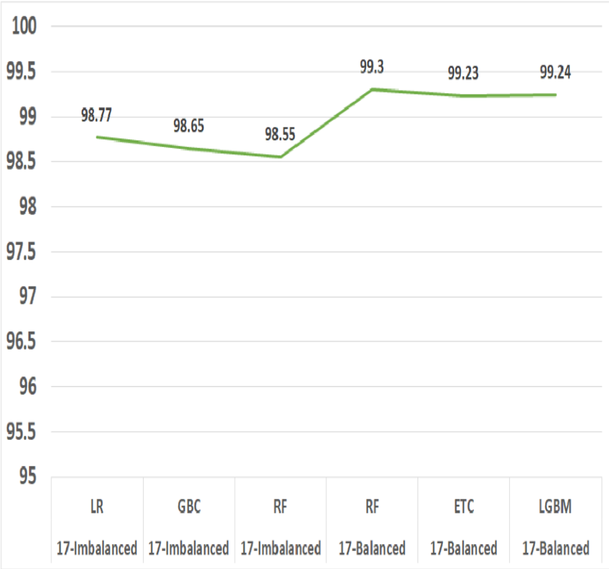


FIGURE 4. Accuracy of prediction models using 17 features.

1) TYPE 1

In Type 1, 17 features are taken by using feature selection. In each type, two data groups are available: balanced and imbalanced. Imbalance data has been balanced using SMOTE. In each case, 14 algorithms have been used, and the best of the three algorithms results have been mentioned here. Table 4 shows that logistic regression performed the best with imbalanced data. Then, GBC and RF both turn in excellent performances. In such cases, their accuracy rates were 98.55% 98.65% and 98.07%Ẇhen the data are balanced, the random forest classifier demonstrates maximum performance with enhanced accuracy. Each method achieved near-perfect accuracy in the balanced dataset, with RF at 99.3% ETC at 99.23% and LGBM at 99.24%Ṫhe accuracy, F1 Score, Precision, Recall, and Area Under the Curve (AUC) of the LR model applied to the imbalanced data were respectively 98.77, 0.76, 0.87, 0.70, and 0.96. Which in Random Forest (Fig. 4) grows to 99.30, 0.99, 0.99, 0.99, and 1 with balanced data. The findings make it abundantly evident that having a balanced dataset resulted in the best possible outcomes, even when the highest number of 17 attributes were considered.

2) TYPE 2

In this type, we have gleaned nine distinguishing features—specifics with a score greater than or equal to 500 (Table 3). In Fig. 5, the outputs of models created from balanced and unbalanced data are displayed using the nine most essential attributes. When imbalanced data is considered, the LR, GBC, and ABC models perform better than any other models with nine characteristics chosen. Their degrees of accuracy came in at 98.73, 98.69, and 98.67% respectively. When we took into account the balanced data, on the other hand, we found that HGB, LR, and ABC attained greater height accuracy than the other models. Their accuracy degrees were 99.2% 98.87% and 98.14%. The LR and HGB models rank

TABLE 4. Performance results of the prediction models.

Type	Algorithm	Accuracy (%)	F1 Score	Precision	Recall	AUC	10 Fold CV (Average)
17-Imbalanced	Logistic Regression(LR)	98.77	0.76	0.87	0.7	0.96	0.985
17-Imbalanced	Gradient Boosting Classifier(GBC)	98.65	0.75	0.83	0.69	0.96	0.985
17-Imbalanced	Random Forest (RF)	98.55	0.69	0.84	0.643	0.95	0.984
17-Balanced	Random Forest	99.3	0.99	0.99	0.99	1	0.985
17-Balanced	Extra Trees Classifier (ETC)	99.23	0.99	0.99	0.99	0.99	0.983
17-Balanced	Light Gradient Boost (LGBM)	99.24	0.99	0.99	0.99	1	0.984
9-Imbalanced	Logistic Regression (LR)	98.73	0.73	0.89	0.67	0.96	0.985
9-Imbalanced	Gradient Boosting Classifier (GBC)	98.69	0.74	0.86	0.68	0.95	0.985
9-Imbalanced	Ada Boost Classifier	98.67	0.73	0.86	0.67	0.94	0.985
9-Balanced	Hist Gradient Boosting Classifier (HGB)	99.2	0.99	0.99	0.99	0.99	0.984
9-Balanced	Logistic Regression	98.87	0.98	0.98	0.98	0.99	0.985
9-Balanced	Ada Boost Classifier (ABC)	98.14	0.98	0.98	0.98	0.99	0.985
7-Imbalanced	Logistic Regression	97.75	0.7	0.84	0.65	0.95	0.975
7-Imbalanced	Ada Boost Classifier	97.58	0.68	0.79	0.63	0.943	0.975
7-Imbalanced	Hist Gradient Boosting Classifier	97.34	0.68	0.74	0.64	0.925	0.973

highest. Once more, the models designed with balanced data yield the best results. In contrast, the model's accuracy with 17 features considerably surpasses the one with only nine characteristics. In this particular scenario, the LR and HGB models achieved an accuracy rating of 99.20 percent for their predictions. This model does not have the same level of accuracy as Type 1 models.

3) TYPE 3

The paper's author [7] selected features using Fisher's discriminant ratio, Kruskal-Wallis analysis, and Relief-F. He suggests seven qualities for the model training process. We tried to develop models using the same data depending on their preferred characteristics. Their best model achieved an accuracy of 94.49% an F1 score of 0.431, and an area under the curve (AUC) of 0.988. They employed a model called multilayer perceptron with 19 neurons. We applied several machine learning techniques, and Fig. 5 displays the most successful findings. Here, seven selected features [7] have been taken. The total data points are (8256 Alive/ 221 COD-TC) [7]. From this, we can observe that our model's performance is comparable to that of Paper 7. In this instance, we investigated imbalanced data consisting of seven characteristics. The LR, ABC, and HGB results were superior to those of the other models. Their accuracy rates came up at 97.75% 97.58% and 97.34% respectively. The LR model performs noticeably better than the competition. Fig. 9 demonstrates that our accuracy was 2% higher than that achieved by the author of the paper [7]. Additionally, the F1 score was 0.27 points higher, and he achieved an AUC of 0.98, whereas we achieved 0.95. Two of our measurements are more significant than theirs, indicating that our models' performance is comparable.

We can observe that LR works better in this case with imbalanced data. It exhibits 98.77% accuracy with 17 features, 98.73% with nine features, and 97.75% with

TABLE 5. Comparing to previous works' findings.

Paper	Cancer Type	Prediction Type	Year	Dataset Used	Proposed Algorithm	Evaluation Parameter
Deep learning-based multifeature integration robustly predicts central lymph node metastasis in papillary thyroid cancer [38]	Thyroid	Lymph node metastasis	2023	488 patients data collected from Zhuzhou Hospital Xiangya Medical College (2019-2021)	CNN	AUC, in train data 0.87 and in test data 0.76
Model Analysis for Predicting Prostate Cancer Patient's Survival: A SEER Case Study [41]	Prostate	Survival	2023	SEER Database	XGBoost	Accuracy 89.57%
Diagnosis of Metastatic Lymph Nodes in Patients With Papillary Thyroid Cancer [39]	Thyroid	Metastatic Lymph Nodes	2022	Self Collected	ClymphNet -Deep Learning Based Model	Accuracy 93.0% AUC 0.948, Sensitivity 93.27% Specificity 92.71%
Thyroid Disease Classification Using Machine Learning Algorithms.[11]	Thyroid	Types of Thyroid Disease	2021	External Hospitals and Laboratories.	MLP	Accuracy, 96.4%
Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence.[13]	Thyroid	Cancer Detection	2021	Clinico-Pathologic. (2003-2009)	DT	Accuracy, 95%
Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis[7]	Thyroid	Survivability	2020	SEER Database. (1988-2007)	ANN-based MLP model	Accuracy, 94.5%
Machine learning algorithms for the prediction of central lymph node metastasis in patients with papillary thyroid cancer.[12]	Thyroid	Cancer Detection	2020	Peking Union Medical College Hospital. (2018-2019)	Gradient Boosting Decision tree model	AUC, 0.731
Prediction of thyroid disorders using advanced machine learning techniques.[14]	Thyroid	Cancer Detection	2020	UCI Machine Learning Repository.	SVM	Accuracy, 96.92%
Comparision of four machine learning techniques for the prediction of prostate cancer survivability.[15]	Prostate	Cancer Detection	2018	SEER Database. (2004-2009)	ANN	Accuracy, 85.64%
Machine learning models in breast cancer survival prediction.[16]	Breast	Survivability	2016	Cancer Registry Organization of Kerman Province, Iran. (1999–2007)	Trees Random Forest	Accuracy, 96%

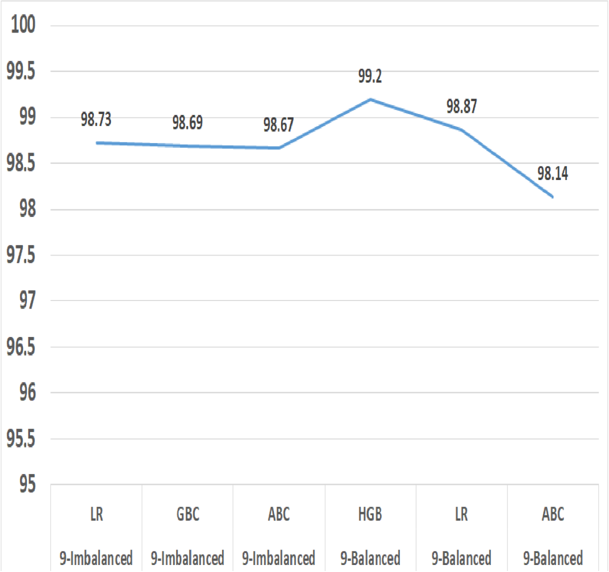


FIGURE 5. Accuracy of prediction models using 9 features.

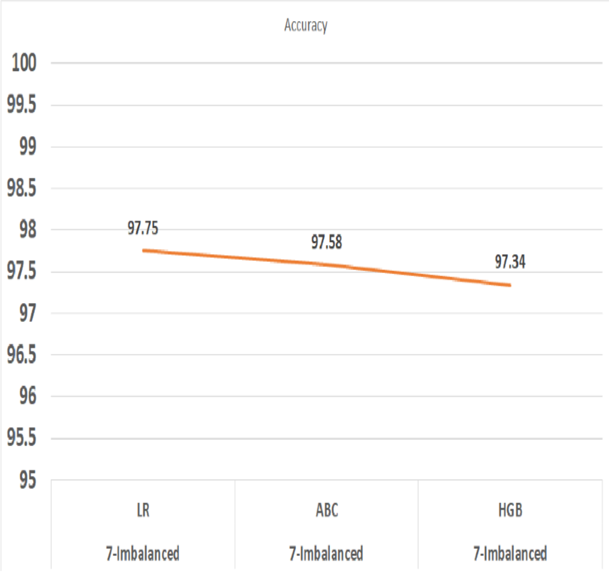


FIGURE 6. Accuracy of prediction models using 7 features.

seven features. When we examine the balanced data, we find that RF and LGBM exhibit the highest accuracy, with respective values of 99.3% and 99.24% and 17 characteristics. HGB and LR, respectively, demonstrate 99.2% and 98.98% accuracy, with nine characteristics. However, compared to all three categories, our 17-feature prediction model performs better.

From the ROC curve, we can see that RF covers more area than LR with 17 features. In this case, RF fire forms better than LR. Similarly, for nine features, HGB outperforms LR with an AUC of 0.999, where the AUC of LR is 0.959. On the other hand, LR performs better with seven features with an AUC of 0.948. We have seen that the model built with 17 features has shown the best results. Figures 7 and 8 help

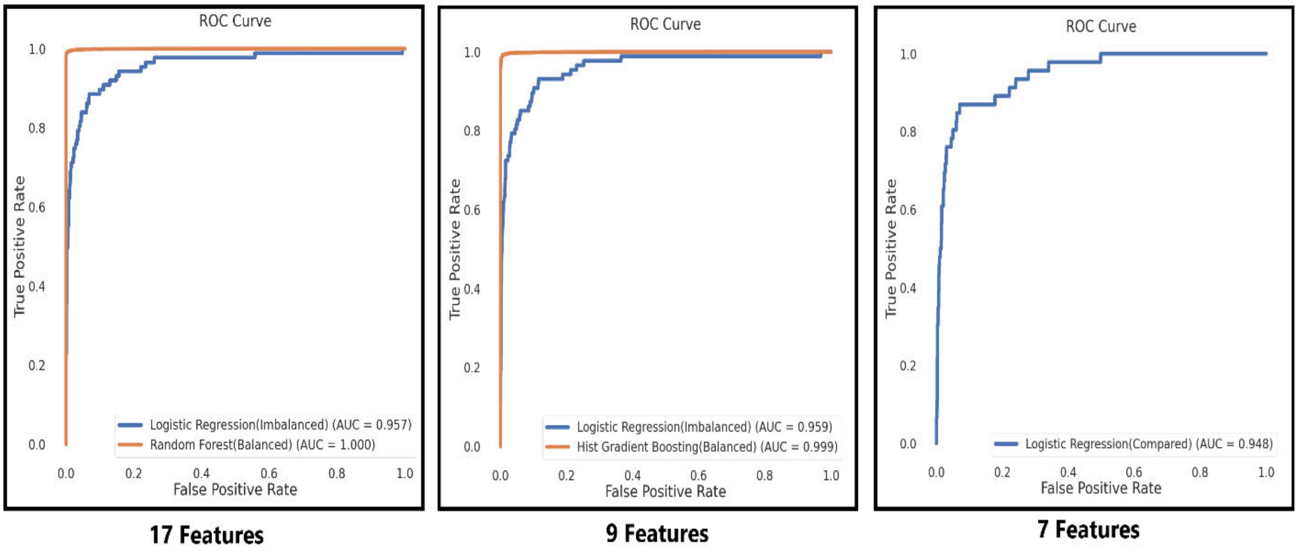


FIGURE 7. ROC curves of 3 groups of features.

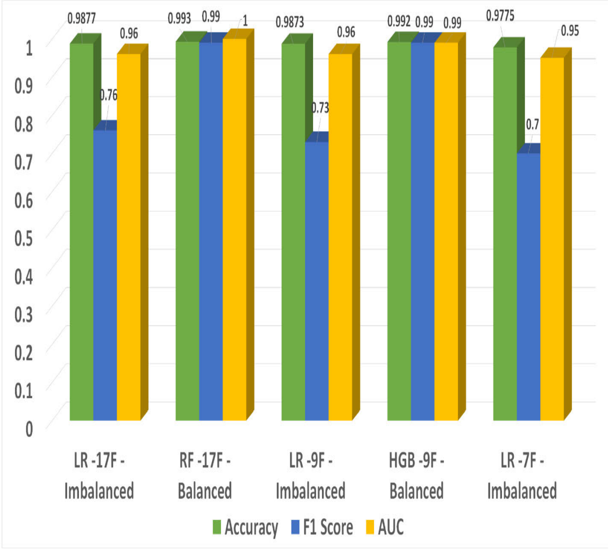


FIGURE 8. Results comparison between top models.

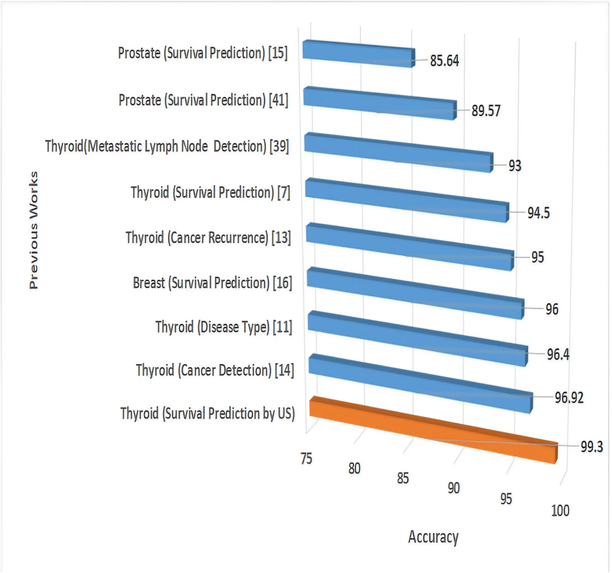


FIGURE 9. Results comparison between other author's models.

us select the best model more clearly. In Fig. 7, we can see that the AUC score of the RF model made with 17 attributes is 1, which means that the ROC curve can cover 100% of the data, ahead of models made with 9 and 7 features. Finally, Fig. 8 shows that the RF (RF-17F-Balanced) model, built with 17 features, is at the highest position, with an accuracy of 99.3% an F1 score of 0.99, and an AUC of 1.

Moreover, our element choice framework worked finely. For this kind of SEER data, Logistic regression and Ada boost performed very well with different features and different amounts of data. Most importantly, if we look at the results of the balanced dataset, we can see that (type 1) with 17 attributes, three models are giving 99% accuracy; not only the accuracy, but other measurements also reached 99%Ẇith 17 attributes, two models' AUC scores were found: 1. But with nine attributes, results are so near but downward than type 1. More features are giving more performance in this case.

Table 5 compares some recent works. From the table, we can learn about their works, the algorithm they used, and the accuracy they achieved.

Fig. 9 is a Results Comparison Between Other Authors's Models. We employed a dataset of 25217 records and employed 14 methods where Jajroudi et al. [2], Park and Lee [13], Duggal and Shukla [14], Montazeri and Beigzadeh [16], and Liu et al. [17] used smaller datasets of

7706, 1040, 7200, 900, and 286 records, respectively and employed less than ten methods. If we compare type 3, type 1 (imbalanced), and type 2(imbalanced), it is clear that the SMOTE technique helps increase performance. The author of the paper [7] applied machine learning and feature selection techniques for thyroid cancer prediction, but he did not work on imbalanced data handling. However, we resolved the problem by using the SMOTE technique. Moreover, he used only seven features, whereas we used 17 features, and we achieved the highest accuracy with our balance data. Previously, Delen et al. [28] and Thongkam et al. [29] also tried to predict the survivability of other cancers, but their models' accuracy could have been more satisfactory. Nevertheless, our proposed models give 99% accuracy. Our best-proposed model is a Random Forest classifier with an accuracy of 99.30% with 17 attributes and balanced data.

V. CONCLUSION AND FUTURE WORK

Our research aims to estimate the survivability of thyroid cancer patients. To do this, we have employed several well-known machine learning methods. The three best-looking computations are used in nearly all the ML algorithms we have tried. Also, we have determined which characteristics will play crucial roles. We recommend selecting the k-best and the Chi-squared test for this circumstance. There were three separate sessions when we oversaw the examinations. Each meeting is divided into halves once again, with each half exploring the results of both evenness and imbalance. We have observed outcomes shift in both attributes to varying degrees, depending on their relative relevance. Due to the asymmetry of the data, we discovered that the Random Forest prediction model outperformed other models. With an area under the curve (AUC) of 1, an F1 score of 0.99, and an accuracy of 99.30%.The results were enhanced by utilizing two effective techniques. Robust cross-validation strategies were implemented to ensure that the reported 98.77% accuracy remains consistent across different dataset splits, guarding against overfitting. Our calculations show the results of the 17 best qualities and the two scenarios. They are introducing sophisticated data preprocessing techniques tailored to tackle the dataset's specific impurities, ensuring the quality and reliability of the results. Moreover, our research indicates that the SMOTE approach is advantageous in balancing an imbalanced dataset. Notably, the Logistic Regression model demonstrates high accuracy despite significant dataset impurities, potentially setting a benchmark for future models in similar contexts.

Future research could concentrate on integrating data categories, including genomics, clinical, and lifestyle data across diverse populations [44], [45]. Furthermore, methodologies were provided to scale these solutions to larger datasets with similar impurity challenges, ensuring the broad applicability of the findings. Moreover, it is possible to do hyperparameter tuning by incorporating additional datasets and neural network technologies, thereby augmenting the precision of the obtained outcomes. This can provide a comprehensive view of a patient's health and identify novel prognostic factors.

Temporal Modeling: We compared the performance obtained on random forests with other machine learning algorithms commonly used in survival prediction tasks, such as support vector machines, gradient boosting, or neural networks. We will expand the analysis to include long-term survival and prognosis beyond the initial survivability prediction. We will investigate how the Random Forest model or other algorithms predict survival outcomes over extended periods, such as 5 or 10 years. This investigation would provide insights into the long-term prognosis and guide treatment planning for thyroid cancer patients. Machine learning models that consider temporal patterns in patient data could provide more accurate survival predictions. These may consist of recurrent neural networks and temporal convolutional networks.

As the complexity of models increases, it becomes progressively imperative to understand their decision-making process.

External Validation: To affirm their generalizability and reliability, models should be validated using multiple data sets.

As machine learning in healthcare uses sensitive patient data, future research must continue to address ethical considerations and data privacy concerns.

ACKNOWLEDGMENT

Saadat M. Alhashmi, Aminul Haque, and Ibrahim Abaker Targio Hashem: conceptualization, review, and editing; and Md. Shohidul Islam Polash, Fazley Rabbe, Shazzad Hossen, Nuruzzaman Faruqui, and Nirase Fathima Abubacker: literature review, experimentation, and writing.

REFERENCES

[1] X. Wu, Y. Yan, H. Li, N. Ji, T. Yu, Y. Huang, W. Shi, L. Gao, L. Ma, and Y. Hu, "DNA copy number gain-mediated lncRNA LINC01061 upregulation predicts poor prognosis and promotes papillary thyroid cancer progression," *Biochem. Biophysical Res. Commun.*, vol. 503, no. 3, pp. 1247–1253, Sep. 2018.

[2] M. Jajroudi, T. Baniasadi, L. Kamkar, F. Arbabi, M. Sanei, and M. Ahmadzade, "Prediction of survival in thyroid cancer using data mining technique," *Technol. Cancer Res. Treatment*, vol. 13, no. 4, pp. 353–359, Aug. 2014.

[3] S. J. Mandel, "A 64-Year-Old woman with a thyroid nodule," *JAMA*, vol. 292, no. 21, pp. 2632–2642, Dec. 2004.

[4] S. I. Sherma, "Thyroid carcinoma," *Lancet*, vol. 361, no. 9356, pp. 501–511, 2003.

[5] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics, 2014," *CA, A Cancer J. Clinicians*, vol. 64, no. 1, pp. 9–29, 2014.

[6] J. Llobera, M. Esteva, J. Rifa, E. Benito, J. Terrasa, C. Rojas, O. Pons, G. Catalan, and A. Avella, "Terminal cancer: Duration and prediction of survival time," *Eur. J. Cancer*, vol. 36, no. 16, pp. 2036–2043, 2000.

[7] M. Mourad, S. Moubayed, A. Dezube, Y. Mourad, K. Park, A. Torreblanca-Zanca, J. S. Torrecilla, J. C. Cancilla, and J. Wang, "Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis," *Sci. Rep.*, vol. 10, no. 1, p. 5176, Mar. 2020.

[8] S. Lee, S. Lim, T. Lee, I. Sung, and S. Kim, "Cancer subtype classification and modeling by pathway attention and propagation," *Bioinformatics*, vol. 36, no. 12, pp. 3818–3824, Jun. 2020.

[9] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 841–850, May 2019.

[10] C. M. Kitahara and J. A. Sosa, "The changing incidence of thyroid cancer," *Nature Rev. Endocrinology*, vol. 12, no. 11, pp. 646–653, Nov. 2016.

[11] K. Salman and E. Sonuç, "Thyroid disease classification using machine learning algorithms," *J. Phys., Conf.*, vol. 1963, no. 1, Jul. 2021, Art. no. 012140.

[12] Y. Wu, K. Rao, J. Liu, C. Han, L. Gong, Y. Chong, Z. Liu, and X. Xu, "Machine learning algorithms for the prediction of central lymph node metastasis in patients with papillary thyroid cancer," *Frontiers Endocrinol.*, vol. 11, Oct. 2020, Art. no. 577537.

[13] Y. M. Park and B.-J. Lee, "Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence," *Sci. Rep.*, vol. 11, no. 1, pp. 1–7, Mar. 2021.

[14] P. Duggal and S. Shukla, "Prediction of thyroid disorders using advanced machine learning techniques," in *Proc. 10th Int. Conf. Cloud Comput., Data Sci. Eng.*, Jan. 2020, pp. 670–675.

[15] H. Wen, S. Li, W. Li, J. Li, and C. Yin, "Comparision of four machine learning techniques for the prediction of prostate cancer survivability," in *Proc. 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2018, pp. 112–116.

[16] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," *Technol. Health Care*, vol. 24, no. 1, pp. 31–42, Jan. 2016.

[17] Z. Liu, Y. Huang, S. Chen, D. Hu, M. Wang, L. Zhou, W. Zhou, D. Chen, H. Feng, W. Wei, C. Zhang, W. Zeng, and L. Guo, "Minimal extrathyroidal extension affects the prognosis of differentiated thyroid cancer: Is there a need for change in the AJCC classification system?" *PLoS ONE*, vol. 14, no. 6, Jun. 2019, Art. no. e0218171.

[18] Y. H. Liu, J. Jin, and Y. J. Liu, "Machine learning-based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy," *Supportive Care Cancer*, pp. 1–7, Mar. 2022.

[19] M. Kukar, N. Besic, I. Kononenko, M. Auersperg, and M. Robnik-Sikonja, "Prognosing the survival time of patients with anaplastic thyroid carcinoma using machine learning," in *Intelligent Data Analysis in Medicine and Pharmacology* (International Series in Engineering and Computer Science), vol. 414. Boston, MA, USA: Springer, 1997, pp. 115–129.

[20] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on seer data," *Sci. Program.*, vol. 20, no. 1, pp. 29–42, 2012.

[21] M. Lundin, J. Lundin, H. B. Burke, S. Toikkanen, L. Pylkkänen, and H. Joensuu, "Artificial neural networks applied to survival prediction in breast cancer," *Oncology*, vol. 57, no. 4, pp. 281–286, 1999.

[22] D. Devi, S. K. Biswas, and B. Purkayastha, "Redundancy-driven modified tomek-link based undersampling: A solution to class imbalance," *Pattern Recognit. Lett.*, vol. 93, pp. 3–12, Jul. 2017.

[23] D. J. Biau, S. Kernéis, and R. Porcher, "Statistics in brief: The importance of sample size in the planning and interpretation of medical research," *Clin. Orthopaedics Rel. Res.*, vol. 466, no. 9, pp. 2282–2288, Sep. 2008.

[24] J. A. Freiman, T. C. Chalmers, H. A. Smith, and R. R. Kuebler, "The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial," in *Medical Uses of Statistics*. Boca Raton, FL, USA: CRC Press, 2019, pp. 357–389.

[25] D. A. Dablain, C. Bellinger, B. Krawczyk, D. W. Aha, and N. V. Chawla, "Interpretable ML for imbalanced data," 2022, *arXiv:2212.07743*.

[26] S. E. Roshan and S. Asadi, "Improvement of bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization," *Eng. Appl. Artif. Intell.*, vol. 87, Jan. 2020, Art. no. 103319.

[27] A. Aldwgeri and N. F. Abubacker, "Ensemble of deep convolutional neural network for skin lesion classification in dermoscopy images," in *Proc. Int. Vis. Inform. Conf.*, Bangi, Malaysia, 2019, pp. 214–226.

[28] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jun. 2005.

[29] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Breast cancer survivability via AdaBoost algorithms," in *Proc. 2nd Australasian Workshop Health Data Knowl. Manag.*, vol. 80, 2008, pp. 55–64.

[30] R. Karthiga, G. Usha, N. Raju, and K. Narasimhan, "Transfer learning based breast cancer classification using one-hot encoding technique," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 115–120.

[31] W. Satriaji and R. Kusumaningrum, "Effect of synthetic minority oversampling technique (SMOTE), feature representation, and classification algorithm on imbalanced sentiment analysis," in *Proc. 2nd Int. Conf. Informat. Comput. Sci. (ICICoS)*, Oct. 2018, pp. 1–5.

[32] B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, and N. D. Filipovic, "Prediction models for estimation of survival rate and relapse for breast cancer patients," in *Proc. IEEE 15th Int. Conf. Bioinf. Bioengineering (BIBE)*, Nov. 2015, pp. 1–6.

[33] A. Endo, T. Shibata, and H. Tanaka, "Comparison of seven algorithms to predict breast cancer survival (<special issue> contribution to 21 century intelligent technologies and bioinformatics)," *Int. J. Biomed. Comput. Hum. Sci., Off. J. Biomed. Fuzzy Syst. Assoc.*, vol. 13, no. 2, pp. 11–16, 2008.

[34] K. R. Pradeep and N. C. Naveen, "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4.5 and naive Bayes algorithms for healthcare analytics," *Proc. Comput. Sci.*, vol. 132, pp. 412–420, Jan. 2018.

[35] Md. S. I. Polash, S. Hossen, R. K. R. Sarker, Md. A. Bhuiyan, and A. Taher, "Functionality testing of machine learning algorithms to anticipate life expectancy of stomach cancer patients," in *Proc. Int. Conf. Advancement Electr. Electron. Eng. (ICAEEE)*, Feb. 2022, pp. 1–6.

[36] C.-H. Yang, S.-H. Moi, F. Ou-Yang, L.-Y. Chuang, M.-F. Hou, and Y.-D. Lin, "Identifying risk stratification associated with a cancer for overall survival by deep learning-based CoxPH," *IEEE Access*, vol. 7, pp. 67708–67717, 2019.

[37] M. S. I. Polash, S. Hossen, and A. Haque, "Five-year life expectancy prediction of prostate cancer patients using machine learning algorithms," in *Soft Computing and Its Engineering Applications* (Communications in Computer and Information Science), vol. 1788. Cham, Switzerland: Springer, 2023, pp. 314–326.

[38] Z. Wang, L. Qu, Q. Chen, Y. Zhou, H. Duan, B. Li, Y. Weng, J. Su, and W. Yi, "Deep learning-based multifeature integration robustly predicts central lymph node metastasis in papillary thyroid cancer," *BMC Cancer*, vol. 23, no. 1, Feb. 2023, doi: 10.1186/s12885-023-10598-8.

[39] A. Abbasian Ardakani, A. Mohammadi, M. Mirza-Aghazadeh-Attari, F. Faeghi, T. J. Vogl, and U. R. Acharya, "Diagnosis of metastatic lymph nodes in patients with papillary thyroid cancer," *J. Ultrasound Med.*, vol. 42, no. 6, pp. 1211–1221, Jun. 2023, doi: 10.1002/jum.16131.

[40] M. D. Kate and V. Kale, "The role of machine learning in thyroid cancer diagnosis," in *Advances in Computer Science Research*. The Netherlands: Atlantis Press, 2023, pp. 276–287, doi: 10.2991/978-94-6463-136-4_25.

[41] M. S. I. Polash, S. Hossen, and A. Haque, "Model analysis for predicting prostate cancer patient's survival: A seer case study," in *Proc. 4th Int. Conf. Trends Comput. Cogn. Eng.*, 2023, pp. 279–290, doi: 10.1007/978-981-19-9483-8_24.

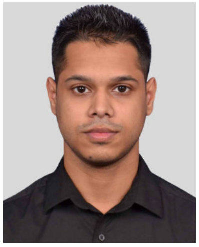
[42] R. H. Nobin, M. Rahman, and M. J. Alam, "Survivability prediction for patients with tonsil cancer utilizing machine learning algorithms," in *Proc. 2nd Int. Conf. Intell. Cybern. Technol. Appl. (ICICyTA)*, Dec. 2022, pp. 210–215, doi: 10.1109/ICICyTA57421.2022.10038122.

[43] H. Torkey, M. Atlam, N. El-Fishawy, and H. Salem, "Machine learning model for cancer diagnosis based on RNAseq microarray," *Menoufia J. Electron. Eng. Res.*, vol. 30, no. 1, pp. 65–75, Jan. 2021, doi: 10.21608/mjeer.2021.146277.

[44] H. Salem, G. Attiya, and N. El-Fishawy, "Intelligent decision support system for breast cancer diagnosis by gene expression profiles," in *Proc. 33rd Nat. Radio Sci. Conf. (NRSC)*, Feb. 2016, pp. 421–430, doi: 10.1109/NRSC.2016.7450870.

[45] M. Atlam, H. Torkey, H. Salem, and N. El-Fishawy, "A new feature selection method for enhancing cancer diagnosis based on DNA microarray," in *Proc. 37th Nat. Radio Sci. Conf. (NRSC)*, Sep. 2020, pp. 285–295.

SAADAT M. ALHASHMI received the Ph.D. degree from Sheffield Hallam University, Sheffield, U.K. He is currently an Associate Professor of information systems with the University of Sharjah, Sharjah, United Arab Emirates. He has supervised several Ph.D. students and published extensively in various high-impact journals and conferences.



MD. SHOHIDUL ISLAM POLASH received the B.Sc. degree in CSE from the Computer Science and Engineering Department, Daffodil International University, Dhaka, Bangladesh, in 2023, with a focus on 3.97/4.00 CGPA. He is currently a Lecturer with the Computer Science and Engineering Department, Daffodil International University. Several worldwide peer-reviewed conferences have published his scientific contributions. His academic research interests include machine learning, deep learning, and computer vision.

AMINUL HAQUE received the B.Sc. degree from the Shahjalal University of Science and Technology, Bangladesh, and the Ph.D. degree from MONASH University. He is currently a Professor with the Department of Computer Science and Engineering, Daffodil International University (DIU), Daffodil Smart City, Dhaka, Bangladesh. He has published his research outputs in several international peer-reviewed journals and conferences. He also contributed data science-related courses to online platforms, such as International Online University (IOU). Recently, he contributed to developing a skill-based national curriculum on big data and data science-related courses. His research interests include data mining, machine learning, and distributed computing.

FAZLEY RABBE received the bachelor's degree in computer science and engineering from Daffodil International University, Bangladesh, in 2021. He is currently pursuing the Master of Engineering degree in information technology with Frankfurt University of Applied Sciences. His current research interests include data mining, the Internet of Things, cyber security, and mobile application authentication.

SHAZZAD HOSSEN received the Bachelor of Science degree in computer science and engineering from Daffodil International University, Daffodil Smart City, Ashulia, Dhaka, Bangladesh, in 2022. He is currently a Software Engineer, leveraging his expertise in computer science and engineering. Within the domain of web3 technologies, he has delved into blockchain development, smart contracts, and the creation of decentralized applications (dApps). His research interests include across blockchain technology, machine learning, deep learning, and reinforcement learning, reflecting his passion for cutting-edge advancements in the tech industry.

NURUZZAMAN FARUQUI received the B.Sc. degree in electrical and electronics engineering from North South University and the master's degree in information technology from the Institute of Information Technology (IIT), Jahangirnagar University (JU), Bangladesh, in 2018, with a focus on 4/4 CGPA.

He is currently an Assistant Professor with the Department of Software Engineering (SWE), Daffodil International University, Bangladesh. He is a Research Coordinator with the Department of Software Engineering. He is also a YouTuber and an Author. He is globally recognized for his educational video content on MATLAB neural networks. He has authored three books. His research interests include artificial intelligence, machine learning, deep learning, cloud computing, and image processing. He is a member of The Institution of Engineers (IEB), Bangladesh, and Bangladesh Society for Private University Academics (BSPUA).

IBRAHIM ABAKER TARGIO HASHEM received the master's degree in computer science from the University of Wales, Newport, and the Ph.D. degree in computer science from the University of Malaya, Kula Lumpur, Malaysia. He is currently an Assistant Professor of computer science with the University of Sharjah, United Arab Emirates. He is an Active Member of the Center for Mobile Cloud Computing Research (C4MCCR), University of Malaya. His numerous research articles are famous and among the most downloaded in top journals. He has published several research articles in refereed international journals and magazines. His areas of research interests include big data, cloud computing, distributed computing, and machine learning. He obtained professional certificates from CISCO (CCNP, CCNA, and CCNA Security) and the APMG Group (PRINCE2 Foundation, ITIL v3 Foundation, and OBASHI Foundation).

NIRASE FATHIMA ABUBACKER was an Associate Professor with Dublin City University (DCU), Ireland, and taught a joint M.Sc. degree in computing (data analytics) with Princess Noura University, Riyadh, through collaboration with DCU. She is currently an Active Supervisor for data analytics master's students capstone projects. She has good hands-on experience in teaching universities, such as Dublin City University, the University of East London, U.K., Staffordshire University, U.K., and London School of Commerce, U.K., modules for both degrees and master's international students from all over the country more than the past 20 years in India, Malaysia, and Saudi Arabia. She has developed sound teaching and research skills with an excellent grasp of the subject material covered by IT and computer science, specifically in data analytics/data science courses. She has excellent experience in developing materials for data science courses with good hands-on experience in teaching data science modules for master's degree students, such as data mining and data analytics, applied machine learning, data management and visualization, big data analytics and technologies, artificial intelligence, data mining, and predictive modeling.