Praedico Salvos: An Ensemble ML Framework for Predicting Survivability of Thyroid Cancer Patients

Fareeha Afzal[1], Bilal Wahid[1,2], Faria Anwar[3], Anoosha Tahir[1], Hafsa Rafique[4]
[1]Research & Development Division, Sabz Qalam, Lahore, Pakistan

[2]Dhanani School of Science and Engineering, Habib University, Karachi Pakistan

[3]Out-Patient Department, Mayo Hospital, Lahore, Pakistan.

[4]Department of Computer Science, University of Management & Technology, Lahore, Pakistan

Correspondence:

Fareeha Afzal: fareehaafzal2000@gmail.com

Article Link: https://journals.brainetwork.org/index.php/jcai/article/view/92

DOI: https://doi.org/10.69591/jcai.2.2.4

Volume 2, Issue 2, 2024

Funding
No

Citation: F. Afzal, B. Wajid, F. Anwar, A. Tahir, and H. Rafique, “Praedico–Salvos: An Ensemble ML Framework for Predicting Survivability of Thyroid Cancer Patients,” *Journal of Computing and Artificial Intelligence*, vol. 2, no. 2, 2024.

Conflict of Interest: Authors declared no Conflict of Interest
Acknowledgment: No administrative and technical support was taken for this research
Article History

Submitted: Sep 11, 2024
Last Revised: Oct 20, 2024
Accepted: Dec 15, 2024

An official Publication of

Beyond Research Advancement & Innovation Network, Islamabad, Pakistan

Praedico Salvos: An Ensemble ML Framework for Predicting Survivability of Thyroid Cancer Patients

Fareeha Afzal[*1], Bilal Wahid[1,2], Faria Anwar[3], Anoosha Tahir[1] , Hafsa Rafique[4]
[1] Research & Development Division, Sabz Qalam, Lahore, Pakistan
[2]Dhanani School of Science and Engineering, Habib University, Karachi Pakistan.
[3]Out-Patient Department, Mayo Hospital, Lahore, Pakistan.
[4]Department of Computer Science, University of Management & Technology, Lahore, Pakistan

Abstract

This paper introduces Praedico-Salvos, a novel machine-learning framework for predicting the survival of thyroid cancer patients. Praedico-Salvos offers a significant advancement over existing methods by predicting survival in four distinct time ranges, rather than a simple binary outcome. This fine-grained prognosis empowers oncologists to tailor treatment plans by considering factors like pain tolerance, financial limitations and predicted survival probability. The model leverages data from the well-established Surveillance, Epidemiology, and End Results (SEER) program, addressing the critical need for more nuanced prognoses in thyroid cancer treatment. Praedico-Salvos achieves a higher accuracy of 88% compared to previous models due to its unique capabilities: (a) handling missing data without imputation, (b) transcending binary classification limitations, and (c) categorizing survival into four distinct time bins. Future advancements could incorporate regression within these bins, further refining predictions for the month.

Keywords: Thyroid cancer, Machine learning, survivability, Classification, SEER.

Introduction

Although Thyroid cancer, being the fifth most common cancer in women [1], occurs differently for different genders, races, and ethnicities [2, 3].

The shape of thyroid cancer cells determines its four types: (a) medullary, (b) anaplastic, (c) papillary, and (d) follicular thyroid cancer cells. Together, they account for almost 98% of all thyroid malignancies [4]. Among these four types, papillary thyroid cancer accounts for 80%, followed by follicular thyroid cancer, which accounts for 10 – 20% of all thyroid cancers [5].

For treatment, most patients (96%) undergo surgery. Still, as treatment is complicated [6], information about a patient's survival can help choose the best option for treatment.

[*] Corresponding Author: fareehaafzal2000@gmail.com

The Surveillance Epidemiology and End Results (SEER) is an important dataset providing detailed statistics about cancer in the USA. For instance, SEER indicates that in 2018, some 893,094 thyroid cancer patients were residing in the USA.

Using SEER, researchers have employed various mathematical tools to quantify cancer survival. For instance, Jajroudi et al. employed logistic regression (LR) and artificial neural networks to show an 80 - 90% success rate of patients surviving between 1 to 5 years after the diagnosis [7]. In addition, both univariate and multivariate Cox regression models have presented promising prediction rates for thyroid cancer [8 - 11]. Liu et al. reported Random Forest (RF) as the best tool, with an area under the curve (AUC) of 0.99 [12]. While Lin et al. employed the Kaplan-Meier survival curve for anaplastic thyroid cancer and reported an incident rate that remained stable for 30 long years [13]. Lastly, researchers have also used Multilayer perceptron (MLP) with the Kruskal Wallis (KW) test to report a 94.5% survival accuracy [14]. In essence, machine learning (ML) has shown promising results for predicting the survival of cancer patients.

As highlighted in Table I, prior work [7, 10, 15, 19, 20, 22] on the subject shows that survivability of thyroid cancer patients has historically been quantified in terms of 1-year, 3-year, 5-year, or 10-year survivability. In machine learning, each of these four measurements, say 5-year survivability, is simply a binary classification framework, wherein the patient either. survives years, or survives Years. Even though these measures are useful from a theoretical standpoint, for the clinical setting, there is a need for a finer-grained prognosis where the survivability of the patient is predicted over the continuous (non-binary) timeline.

This paper presents “Praedico – Salvos,” an ensemble ML framework that predicts the number of months a thyroid cancer patient can survive, based on their features at the time of diagnosis. Compared to earlier works, see Table I, where survivability predicts whether a patient will survive more than three or five years, Praedico – Salvos provides a fine-grained assessment of the survivability of the patient over a set of four classes as opposed to two classes in previous works.

Table 1: *Review of prior works (2014 – 2022) shows that previous models predict in terms of 1-year, 3-year, 5-year, or 10-year survivability. Herein below, (*) indicates that the paper was silent on certain matters.*

#	Year	Features	Duration	Models employed	Findings
1	2014	16	*	(1) MLP, (2) LR	− 1 year: MLP was optimum with ~93% accuracy.

					– 3 years: LR was best with 88.6% accuracy. – 5-years: LR was best with ~91% accuracy [7].
2	2018	*	04 – 12	(1) KM, (2) Cox Regression (CR).	Initial resection of patients suffering from Medullary thyroid cancer does not help in improving survival [16].
3	2018	12	98 – 12	(1) CR, (2) Optimal survival trees, (3) RF	Both 5-year and 10-year survival rates were high, i.e., 96%, and 94% respectively [19].
4	2019	9	86 – 15	(1) Join-point regression, (2) linear regression, (3) KM, (4) CR.	The incidence of anaplastic thyroid cancer remained stable from 1986 – 2015 [13].
5	2020	34	*	(1) Kruskal-Wallis' test, (2) MLP, (3) Relief-F, and (4) Fisher's discriminant ratio.	A survival accuracy of 94.5% was achieved using MLP Classifier [14].
6	2020	13	06 – 15	CR	The American Joint Committee on Cancer approved a framework for survivability with an AUC of 75.5% [18].
7	2021	7	10 – 15	(1) Univariate Cox, (2) Multivariate Cox analysis.	The 3- and 5-year survival rate predictive ability using nomogram presented a good Concordance – Index > 0.8 [10].
8	2021	8	95 – 16	CR	The survival rate in overall Primary Thyroid Lymphoma was found to be 81.5% for 5 years, and 51.4% for 15 years [17].

9	2021	10	04 – 15	(1) KM, (2) CR	The authors noticed that unmarried older patients presented lower overall survival and lower cancer-specific survival, compared to married patients, indicating the need for moral and psychological support [21].
10	2022	*	04 – 15	(1) LR, (2) CR.	Incidence trends indicate the rate of increase of thyroid cancer (i) remained consistent among Native Hawaiians, (ii) slowed among Caucasians, & (iii) remained constant for Asians [8].
11	2022	*	04 – 15	*	The 10-year disease-specific survival rates of patients in stages I, II, III, and IV were 97.9%, 77.9%, 35.3%, and 12.1%, respectively [15].
12	2022	9	10 – 15	(1) Support vector machine (SVM), (2) LR, (3) XGBoost, (4) Decision tree, (5) RF, and (6) KNN rule	RF showed the highest accuracy on 2-year survival with low precision [12].
13	2022	5	04 – 16	CR	The proposed risk classification framework employs a nomogram with (i) age, (ii) tumor size, (iii) extent of surgery, (iv) T stage, and (v) M stage as risk factors and presents good results [11].
14	2022	9	04 – 15	(1) KM, (2) CR	The proposed framework presented an AUC of 0.878 for 5-year, and 0.811 for 10-year survival [20].

15	2022	7	04 – 15	(1) Fine-grey model, (2) CR	The 10-year Thyroid-specific cancer survival and overall survival rates of patients without Prophylactic Central Lymph node dissection were 99.53% and 92.77%, respectively [22].

Methodology

Praedico Salvos is developed using the following steps:

SEER Database and Preprocessing

The SEER database is a valuable resource for this study as it offers a wealth of patient data, including demographics (age, sex, race), diagnosis details (year of diagnosis), and even geographic location. This comprehensive data is updated annually, ensuring we have access to the latest information. We downloaded the SEER data from its software which allowed us to calculate survival rates based on factors that we are considering in the model, like stage at diagnosis and age. Moreover, since SEER collects data from multiple registries, it provides a robust and generalizable patient cohort, strengthening the validity of your findings.

We employed the SEER database as it contains details of 72,116 thyroid cancer patients from 1975 to 2018 with 250 attributes. The patient cohort for this analysis was restricted to cases identified as primary thyroid cancer within the SEER database. This ensures the focus is solely on patients with the initial development of thyroid cancer, excluding any secondary or metastatic occurrences. As SEER presents multiple types of cancers, we selected features relevant to thyroid cancer. Moreover, we removed entire entries containing null, blank, missing, unknown, or zero values. Moreover, categorical, and non-numeric entries were encoded to numerical values via a label encoder. The resulting dataset contained 2,325 entries with 17 features, as shown in Table 2

Normalization and data splitting

We used a min-max scalar to restrict feature values within [0, 1]. We reserved 90% (2092 entries) of the dataset for training and testing while the remaining 10% (233 entries) was used for validation. Moreover, we divided the data into training, test, and validation sets by random spitting. The split division is shown in Figure 1 (b).

Binning

Our approach to predicting thyroid cancer patient survival takes a layered classification route, offering a more nuanced picture compared to a simple alive/deceased binary model. We achieve this by stacking the target variable, survival months, into four distinct bins. Here's a breakdown of the binning strategy and the reasoning behind the chosen intervals:

1. **Bin 1: 0-3 months** - This bin captures very short-term survival, potentially indicating aggressive cancer or immediate post-surgical complications.
2. **Bin 2: 4-6 months** - This bin encompasses a slightly longer timeframe, possibly representing patients with a more advanced stage of cancer or those requiring additional treatment soon after diagnosis.
3. **Bin 3: 7-60 months** - This broader bin covers a significant period, potentially indicating patients with a good prognosis who may respond well to treatment and have a moderate to long-term survival expectancy.
4. **Bin 4**: More than 60 months (5 years+) - This bin identifies patients with a long-term survival exceeding 5 years, suggesting a potentially favorable prognosis and potentially lower risk of recurrence.

Feature Selection

We used a Boruta random forest classifier to quantify the relative importance of each of the 17 features with respect to the target bins to obtain the top 10 features. The Random Forest Classifier provides a built-in measure of feature importance, revealing which features admit strong influence on predicting a patient's survival (in terms of months) [23-25]. Together, these top 10 features amount to a relative score > 90%, as shown in Table 3

Modeling

We applied several classification frameworks to choose the optimum. Specifically, we tested (a) Linear Regressor, (b) Random Forest Regressor, (c) Gradient Boosting Regressor, (d) MLP Regressor, (e) Ridge Regressor, (f) XGB Regressor, (g) KNN rule, (h) Logistic Regression, (i) Support Vector machines, (j) Decision Tree, and (k) Ada Boost for classification. We concluded that the optimal framework was an ensemble machine learning model (‘Praedico – Salvos’) to predict the survivability of thyroid cancer patients, shown in Figure 1.

Table 2: *List of 17 features retained after preprocessing.*

#	Feature	#	Feature	#	Feature
1	Patient id	**2**	Sex	**3**	Year of diagnosis
4	Race and origin	**5**	Primary Site	**6**	AYA site recode 2020 Revision.

7	Histologic Type ICD-O-3	**8**	Behavior recode	**9**	Site recode – rare tumors
10	SEER historic stage A (73 – 15)	**11**	Site specific surgery	**12**	Survival months
13	Vital status recode	**14**	SEER other cause of death	**15**	Total number of in situ/malignant tumors
16	Age recode	**17**	Race/ethnicity		

Figure 1(a): *As shown above, Praedico–Salvos is an ensemble machine learning framework comprising three finely-tuned SVMs collectively reporting an accuracy of 88%.*

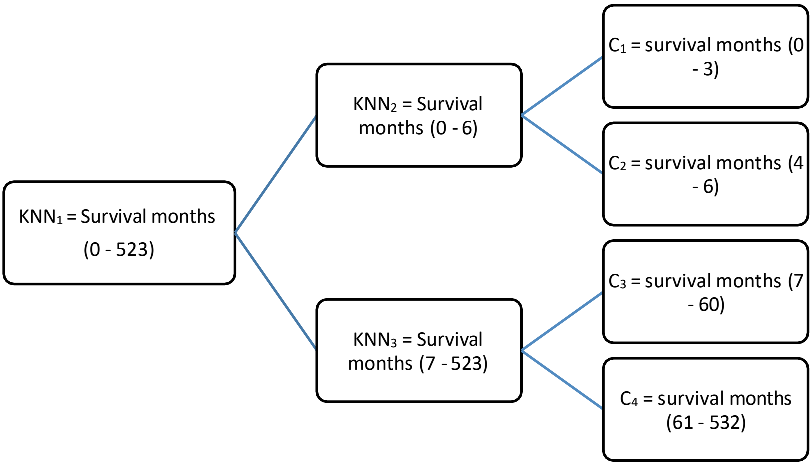


Figure 1(b):S*hows data sampling at each tier.*

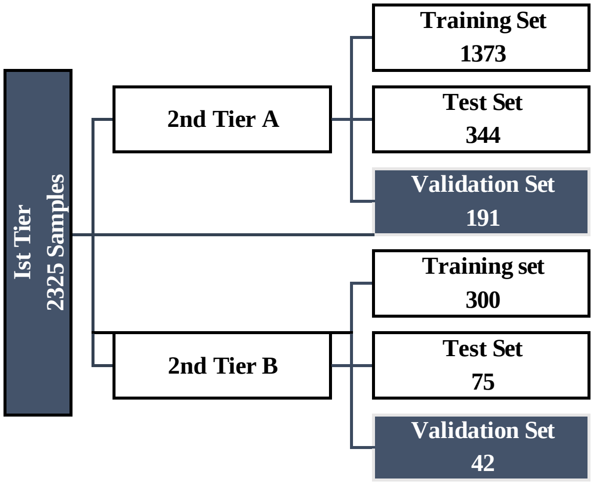


Table 3: *Relative feature score (in ascending order) for top features*

#	Feature	Relative Score (100%)
1	Age recode with single ages and 85+	23.3
2	Year of diagnosis	18.1
3	Site-specific surgery	9.0
4	SEER historic stage A	8.5
5	AYA site recode 2020 revision	8.2
6	Histologic Type ICD-0-3	8.1
7	Sex	4.5
8	Site recode-rare tumors	4.3
9	Race and origin recode	3.7
10	Total number of in situ/malignant tumors	3.2
	Total	91.9

Results and Discussion

Praedico–Salvos presents an ensemble SVM model showcasing a two-layered classification model for finer-grained prognosis of thyroid cancer patients, as shown in Fig. 1. Praedico-Salvos prioritizes clinical action ability. While regression offers continuous survival time prediction, it presents challenges in translating this to concrete treatment plans. A layered approach with defined bins provides more relevant information for oncologists, allowing for targeted interventions and resource allocation. Additionally, high RMSE was observed previously in our experiment when we used regression. It highlighted the limitations of this approach for survival prediction where small deviations significantly impacted treatment decisions as shown in Table 4.

Table 4: *The table shows the results of different regression models on test and validation data. Herein below, the best results are shown in bold and underlined. As evident, regression does not admit good results, hence the authors proceed with an alternate route.*

MODEL	**RMSE (TEST SET)**	**RMSE (VALIDATION SET)**
Linear Regressor	62.13	58.82
Random Forest Regressor	55.36	55.04
Gradient Boosting Regressor	52.58	51.98
<u>MLP Regressor</u>	**<u>63.88</u>**	**<u>65.52</u>**
Ridge Regressor	62.07	58.84
XGB Regressor	57.81	58.45

Hence,we divided the target variable into four classes 0 – 3 months, 4 – 6 months, 7 – 60, and > 60 months. This assymmetric division was done to ensure (an almost) equal distribution of representatives per class. Rather than looking for the best classifier that optimally divided the data into four classes, we opted to form two layers. Here each layer employed a binary classifier, such that with 2 layers of binary classification, we obtained the needed 4 classes.

The rationale behind the binning intervals is as follows:

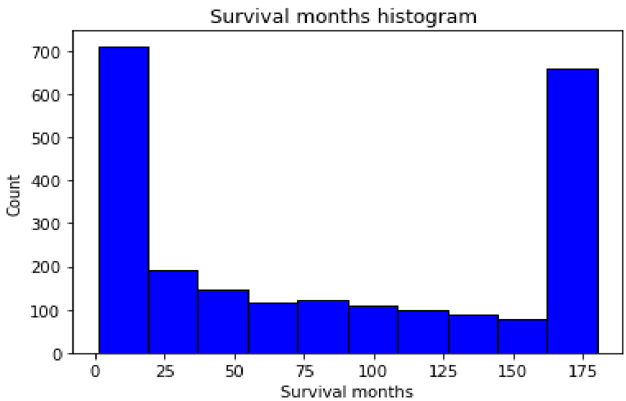
Early Mortality: The first two bins (0-3 months and 4-6 months) capture patients with very short-term survival. This could be due to factors like highly aggressive cancer, complications arising from the initial surgery, or pre-existing health conditions.

Mid-Term Survival: The third bin (7-60 months) represents a broader range, encompassing patients with a moderate prognosis who may undergo additional treatment and have a fair chance of surviving several years.

Long-Term Survival: The final bin (more than 60 months) identifies patients with a very positive outlook, exceeding the traditional 5-year survival benchmark often used in cancer studies.

It's important to acknowledge the seemingly inconsistent division between the first two bins (3 months) and the broader range of the third bin (7-60 months). This choice is because the initial months after diagnosis are often critical, with a higher risk of complications. Separating this period allows for a clearer understanding of very short-term survival outcomes. Moreover, the distribution of survival data as shown in Figure 2 has shown a significant incline in the initial months post-diagnosis, followed by a more gradual decline. Capturing this pattern with narrower bins in the early timeframe can be informative.

Figure 2:*Distribution of survival months: As per SEER data, shown above, most patients survive between 0 – 20 months. Thereafter, the survival of thyroid cancer patients reduces consistently. The last bar is high only because all other patients surviving from 165 to 532 months are binned together for the sake of brevity.* .



For the case of feature selection, Boruta is a wrapper method built around the Random Forest algorithm. It essentially creates "shadow features" by shuffling the values within each existing feature column. Therefore, the interpretability, built-in feature importance calculation, and good overall accuracy make Boruta a strong choice for understanding which features are most critical in predicting survival bin classification for thyroid cancer patients.

In tier 1, the KNN classifier with an accuracy of 87% performed the best, dividing the data into months, and months. For tier 2, again the KNN rule performed best, exhibiting an accuracy of 76% in dividing. into two disjoint sets , and an accuracy of 72% in dividing into classes months, as shown in Fig. 4 – 6, and Table 5.

KNN and Random Forests perform well with moderate-sized datasets like the one we have used (72,116 patients with 17 features). Additionally, if the data has clear underlying relationships between features and survival outcomes, these algorithms are simple and effective in capturing those patterns.

Collectively, the accuracy of the proposed ensemble model ("Praedico – Salvos") comes out to be 88%. This ensemble approach effectively breaks down the classification task into simpler sub-tasks, allowing the KNN rule to achieve high accuracy at each level. The model operates hierarchically, refining its predictions step by step. Even if some steps have lower accuracy, the combined process can still yield high overall performance as each tier builds on the previous one. The high accuracy in the initial broad classification (87% for Tier 1) means that subsequent classifications are working with more reliably partitioned data, leading to a robust outcome. Even if some tiers have lower accuracy, these tiers are specialized sub-tasks. The errors in these sub-tasks may not drastically impact the final application if the broader classification is accurate. This combined accuracy matters more as it reflects the real-world performance of the model in categorizing data through multiple stages, ensuring robustness despite some intermediate steps having lower accuracy.

Figure 3: *The figure shows KNN to perform well (>80%) for 1st tier classification, i.e., vs. .*

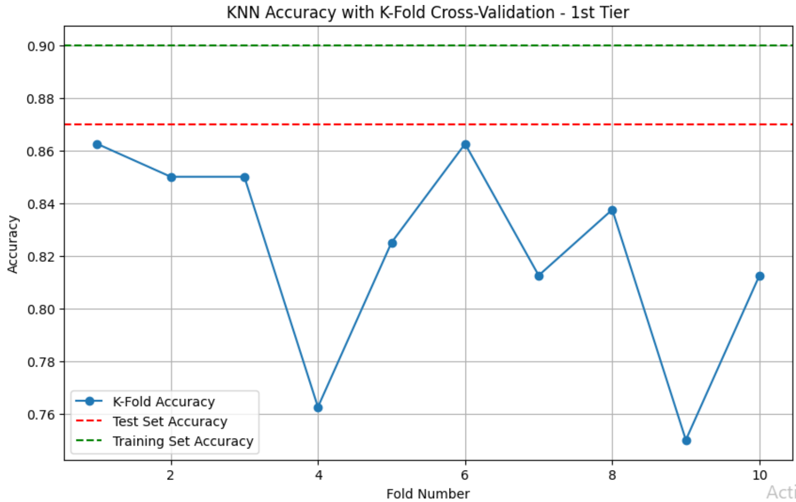


Figure 4: *The figure shows KNN performance for tier 2 – part A classification vs. months.*

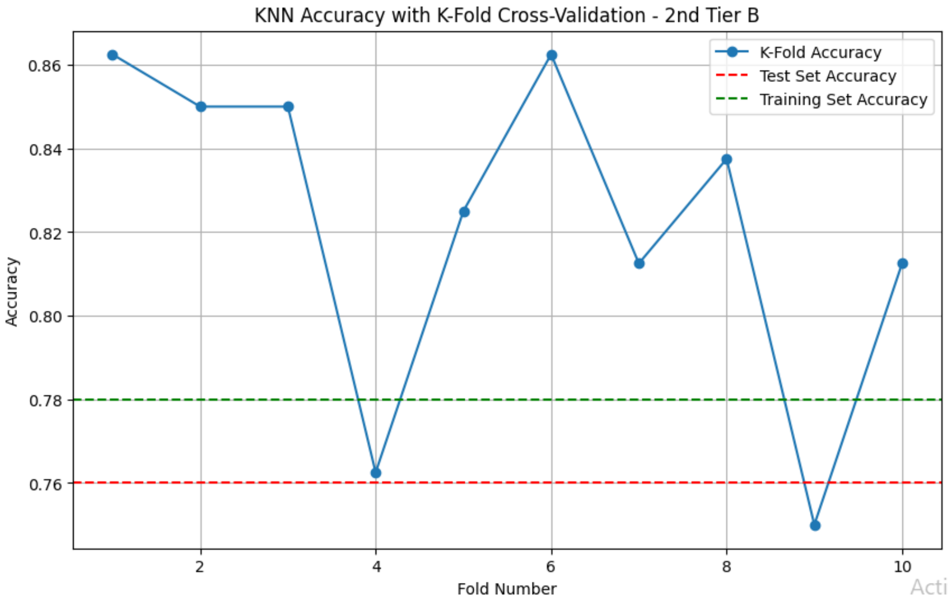


Table 5: *Comparison of classifiers for each classification tier.*

Tier Classes (month)	1st Tier (0 – 523) $C_a: (0-6)$ vs. $C_b: > 6$		2nd Tier – part A $C_1: (0-3)$ vs. $C_2: (4-6)$		2nd Tier – part B $C_3: (7-60)$ vs. $C_4: > 60$	
Models (accuracies %)	**Train set**	**Test set**	**Train set**	**Test set**	**Train set**	**Test set**
KNN rule	90	**87**	81	**78**	78	**76**
Logistic Regression	85	82	63	60	77	75
Support Vector machine	87	77	73	65	67	65
Decision Tree	82	80	69	63	65	60
Random Forest	83	82	70	62	71	69
Ada Boost	85	81	65	61	63	62

Conclusion

Cancer treatment is expensive. It is a branch of medicine that does not follow the 'survival of the fittest,' rather it follows the 'survival of the richest.' Here, Praedico–Salvos presents the state-of-the-art framework for predicting the survival of thyroid cancer patients. Compared to previous works which were binary, Praedico–Salvos predicts survival over four time periods, thereby improving the overall framework. As cancer treatment is both painful and expensive, Praedico–Salvos could help oncologists determine the likelihood of survival, deciding the best course of treatment, based on capacity to endure pain, expected chances of survival, and available finances.

Compared to existing methods, shown in Table 1, Praedico–Salvos is better as it (a) does not impute missing values, (b) is not restricted to binary classification, and (c) classifies the survival of the patient into 4 separate bins, each highlighting the likelihood of the patient's survival.

Looking ahead, Praedico-Salvos holds immense potential for further refinement. Integrating regression within each of the four classes, for determining an exact survival month presents a compelling avenue for future work. This could enhance the model's resolution, potentially predicting survival down to individual months. Additionally, exploring the incorporation of factors like treatment response and emerging therapies could broaden the scope of Praedico–Salvos, making it an even more valuable tool in the fight against thyroid cancer.

Lastly, the phrase "Praedico – Salvos" is a combination of two Latin words 'praedico' meaning to predict or foretell, while 'salvos' translates as survival. Hence, we combined the two words 'predict' and 'survival' into Latin as 'Praedico – Salvos.'

Refrences

[1] Cabanillas, Maria E., David G. McFadden, and Cosimo Durante. "Thyroid cancer." The Lancet 388.10061 (2016): 2783-2795.

[2] Races, All, and Males White Males Black Males. "SEER cancer statistics review 1975-2017." National Cancer Institute, (2020).

[3] Thun, Michael, et al., eds. Cancer epidemiology and prevention. Oxford University Press, 2017.

[4] Carling, Tobias, and Robert Udelsman. "Thyroid cancer." Annual review of medicine 65.1 (2014): 125-137.

[5] Gimm, Oliver. "Thyroid cancer." Cancer Letters 163.2 (2001): 143-156.

[6] Debela, Dejene Tolossa, et al. "New approaches and procedures for cancer treatment: Current perspectives." SAGE open medicine 9 (2021): 20503121211034366.

[7] Jajroudi, M., et al. "Prediction of survival in thyroid cancer using data mining technique." Technology in cancer research & treatment 13.4 (2014): 353-359.

[8] Moon, Peter K., et al. "Thyroid cancer incidence, clinical presentation, and survival among Native Hawaiian and other Pacific islanders." Otolaryngology–Head and Neck Surgery 169.1 (2023): 86-96.

[9] Sun, W., et al. "Newly proposed survival staging system for poorly differentiated thyroid cancer: a SEER-based study." Journal of Endocrinological Investigation 46.5 (2023): 947-955.

[10] C. Wang, L. Dai, X. Wu, and Z. Wang, "A nomogram for predicting overall-specific survival in thyroid cancer patients with total thyroidectomy: a SEER database analysis," (in eng), Gland surgery, Aug 2021, vol. 10, no. 8, pp. 2546-2556.

[11] Wang, Cheng, et al. "A nomogram for predicting overall-specific survival in thyroid cancer patients with total thyroidectomy: a SEER database analysis." Gland Surgery 10.8 (2021): 2546.

[12] W. Liu, S. Wang, Z. Ye, P. Xu, X. Xia, and M. Guo, "Prediction of lung metastases in thyroid cancer using machine learning based on SEER database," 2022, vol. 11, no. 12, pp. 2503-2515.

[13] Lin, Bo, et al. "The incidence and survival analysis for anaplastic thyroid cancer: a SEER database analysis." American journal of translational research 11.9 (2019): 5888.

[14] Mourad, Moustafa, et al. "Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis." Scientific Reports 10.1 (2020): 5176.

[15] Sun, W., et al. "Newly proposed survival staging system for poorly differentiated thyroid cancer: a SEER-based study." Journal of Endocrinological Investigation 46.5 (2023): 947-955.

[16] Randle, Reese W., et al. "Survival in patients with medullary thyroid cancer after less than the recommended initial operation." Journal of Surgical Oncology 117.6 (2018): 1211-1216.

[17] Florindez, Jorge A., et al. "Primary thyroid lymphoma: survival analysis of SEER database (1995–2016)." Leukemia & lymphoma 62.11 (2021): 2796-2799.

[18] Liu, Xiangxiang, et al. "The impact of radioactive iodine treatment on survival among papillary thyroid cancer patients according to the 7th and 8th editions of the AJCC/TNM staging system: a SEER-based study." Updates in Surgery 72 (2020): 871-884.

[19] Banerjee, Mousumi, David Reyes-Gastelum, and Megan R. Haymart. "Treatment-free survival in patients with differentiated thyroid cancer." The Journal of Clinical Endocrinology & Metabolism 103.7 (2018): 2720-2727.

[20] Jin, Shuai, et al. "Development and validation of a nomogram model for cancer-specific survival of patients with poorly differentiated thyroid carcinoma: A SEER database analysis." Frontiers in Endocrinology 13 (2022): 882279.

[21] Ai, Lei, et al. "Effects of marital status on survival of medullary thyroid cancer stratified by age." Cancer medicine 10.24 (2021): 8829-8837.

[22] Song, Jun Long, et al. "Long-term survival in patients with papillary thyroid cancer who did not undergo prophylactic central lymph node dissection: a SEER-based study." World journal of oncology 13.3 (2022): 136.

[23] Rudnicki, Witold R., Mariusz Wrzesień, and Wiesław Paja. "All relevant feature selection methods and applications." Feature Selection for Data and Pattern Recognition (2015): 11-28.

[24] Chen, Rung-Ching, et al. "Selecting critical features for data classification based on machine learning methods." Journal of Big Data 7.1 (2020): 52.

[25] Degenhardt, Frauke, Stephan Seifert, and Silke Szymczak. "Evaluation of variable selection methods for random forests and omics data sets." Briefings in Bioinformatics 20.2 (2019): 492-503.