Model Analysis for Predicting Stage and Survival of Prostate Cancer Patients

BY

Md. Shohidul Islam Polash
ID: 191-15-2523

AND

Shazzad Hossen
ID: 191-15-2420

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. S.M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFOFIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
December 2022

APPROVAL

This Project titled "**Model Analysis for Predicting Stage and Survival of Prostate Cancer Patients**", submitted by Md. Shohidul Islam Polash and Shazzad Hossen to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 06th December 2022.

BOARD OF EXAMINERS

Dr. S M Aminul Haque — **Chairman**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology Daffodil International University

(Name) — **Internal Examiner**
Designation
Department of Computer Science and Engineering
Faculty of Science & Information Technology Daffodil International University

(Name) — **External Examiner**
Designation
Department of Computer Science and Engineering
Faculty of Science & Information Technology Daffodil International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. S.M. Aminul Haque, Associate Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. S.M. Aminul Haque

Associate Professor

Department of CSE

Daffodil International University

Co-Supervised by:

Mohammad Jahangir Alam

Lecturer (Senior Scale)

Department of CSE

Daffodil International University

Submitted by:

Md. Shohidul Islam Polash

ID: 191-15-2523

Department of CSE

Daffodil International University

Shazzad Hossen

ID: 191-15-2420

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible for us to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to our supervisor **Dr. S.M. Aminul Haque**, **Associate Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge and keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Mohammad Jahangir Alam,** Lecturer (Senior Scale)**,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Prostate cancer is assumed to be the most familiar cancer and the principal cause of death in the world. For effective treatment to decrease mortality, an accurate staging and survival projection like Alive or dead, five-year life expectancy prediction is essential. A remedy plan can be scheme under the predicted stage and survival state. Machine Learning (ML) approaches have recently attracted significant attention, particularly in constructing data-driven prediction models. Prostate cancer staging and survival prediction have received little attention in research. In this research, we constructed models with the support of ML techniques to determine whether a patient with prostate cancer will survive or not and whether he/she will live five years or not. Also, the stage forecasting models have been constructed. Feature impact analysis, a good amount of data, and a distinctive track make our model's results better compared to previous research. The models have been created using data from the SEER (Surveillance, Epidemiology, End Results) database. SEER program collects and distributes cancer statistics to lessen the disease impact. Using around twelve ML algorithms, we assessed the survival, five-year life expectancy, and stage of prostate cancer patients. HGB, LGBM, XGBoost, Gradient Boosting, and Ada Boost are notable prediction models. Among them, the XGBoost contributes the most, with an accuracy of 89.57% in predicting survivorship (Alive or Dead) and being discovered to be the quickest model. The Gradient Boosting method, on the other hand, exceeds the others, with an accuracy of 88.45% in forecasting five-year life expectancy. Furthermore, LGBM attained the maximum accuracy of 96.27% in predicting the stage of prostate cancer patients.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

Prostate cancer is the second most common type of malignancy found in men and the fifth most common cause of death worldwide [1]. In terms of prevalence, prostate cancer ranks first, whereas mortality rates put it in third place; this is the most frequent cancer in 105 nations [1, 2].

Physicians might design a better treatment plan when treating prostate cancer patients if they know whether or not the patients will live for five years. Sometimes a doctor would try to diagnose a patient based on their physical state by comparing them with the prior patient; however, a doctor can only diagnose a few prostate cancer patients in their lifetime. In this article, we have constructed an artificial prediction model using data from over fifty thousand patients to evaluate the likelihood of patient survival assisting the doctors in preparing the prescription for thousands.

We have collected all of these essential factors from the SEER program, which are supported by the AJCC (American Joint Committee on Cancer) [3]. Through correlation analysis, we found features that had distinct effects and fed the features into the machine for learning. Machine learning (ML) are being used to make medical services more efficient in many sectors. Using machine learning techniques, we sought to forecast whether a patient would live five-year (sixty months) or not. Since the generated target characteristic contains two classes, 0 to 60 months and 61 to more months, it is a binary classification issue. To predict the survival of prostate cancer patients, we employed prominent ML classifiers such as Gradient Boosting Classifier (GBC), Light Gradient Boosting Machine (LGBM), AdaBoost (ABC), Decision Tree (DT), Random Forest (RF), and Extra Trees (ETC). Finally, hyperparameter optimization was used to enhance prediction outcomes. Furthermore, we interpreted our data using accuracy [1, 4], precision [1, 5, 6], sensitivity [4], specificity [4], AUC [5, 11], and ROC curves. The new feature added in interpretability is how fast our model predicts. All the boosting methods performed comparably; hence the best model was picked using performance measurements and prediction speed. The optimized GBC performs better than the other classifiers, with an accuracy of 88.45%.

The major contribution of our work is presented below:

1. A survival prediction model for prostate cancer patients with 89.56% accuracy, A stage prediction model which can detect 96.27% of the stages.
2. An optimized five-year life expectancy prediction model using the Gradient boosting technique for prostate cancer patients. Our model performed better compared to Wen et al.'s [1] prediction model.
3. Optimizing parameters for GBC to build survival prediction models.
4. Identified a fast forecast model for prostate cancer stage, survival, and five-year life expectancy. This technique can be helpful to conduct other cancer research.

In feature engineering, we employed label encoding to build prediction models with enormous data and characteristics with unique impacts. We improved the performance by optimizing the hyperparameters; also, we examined how quickly the models run on various platforms, which made the prostate cancer life expectancy prediction model superior to others. Additionally, to the best of our knowledge, only a few studies have predicted prostate cancer survival. Wen et al. [1] demonstrated prostate cancer five-year survival; moreover, their accuracy was 85.64%, whereas we achieved 88.45%.

The remains of the paper are categorized into five sections. Previous works concerning planning algorithms are described in section 2, and suitable methods are presented in section 3. The results acquired are shown and discussed in section 4. The key points are presented in section 5, which is the conclusion of the paper. The conflict of interest is presented in section 6 of the article.

1.2 Motivation

There are around 1,100,000 new instances of prostate cancer each year, which results in approximately 300,000 deaths. This makes prostate cancer one of the most common malignancies that affects men. The average age of diagnosis is far over 60 years old, and middle-aged and older males make up the majority of those affected. Because of its gradual progression, extended latency period, high morbidity, and high fatality rate, prostate cancer is a significant public health problem that requires immediate attention. It is crucial to have an accurate stage and survival estimate in order to provide effective treatments that will reduce mortality. A

treatment plan can be devised if the stage and survival status that is projected are taken into account. For this reason, we have chosen to devote our attention to this research-based endeavor.

1.3 Rational of the Study

Prostate cancer usually develops slowly, so there may be no signs for many years. Finding and treating it before symptoms occur may not improve men's health or help them live longer. Prostate cancer can spread to nearby organs, such as your bladder, or travel through your bloodstream or lymphatic system to your bones or other organs. Prostate cancer that spreads to the bones can cause pain and broken bones. Once prostate cancer has spread to other areas of the body, it may still respond to treatment and may be controlled, but it's unlikely to be cured. With improved therapies, prostate cancer death rates can be lowered. However, this is only doable if they are aware of the cancer's stage or the likelihood of the patient's survival. A doctor may try to make a diagnosis about a patient's health by drawing parallels between the patient's symptoms and those of a previous patient; however, there are very few cases of prostate cancer that a doctor will ever see. In this regard, we attempted to make predictions about both the stage at which prostate cancer patients are diagnosed and their chances of survival.

1.4 Research Questions

- What is the Prostate cancer status globally?
- What are the associated factors with prostate cancer?
- Which algorithm perform well and why?
- What is machine learning?

1.5 Research Objectives

- Determine what stage of cancer a patient suffering from prostate cancer is now in.
- Try to establish whether or not the patient will be able to live.
- After determining whether or not the patient will survive, the next stage is to predict whether or not the patient will be alive in five years.
- Providing patients with an effective treatment plan and ensuring its implementation.

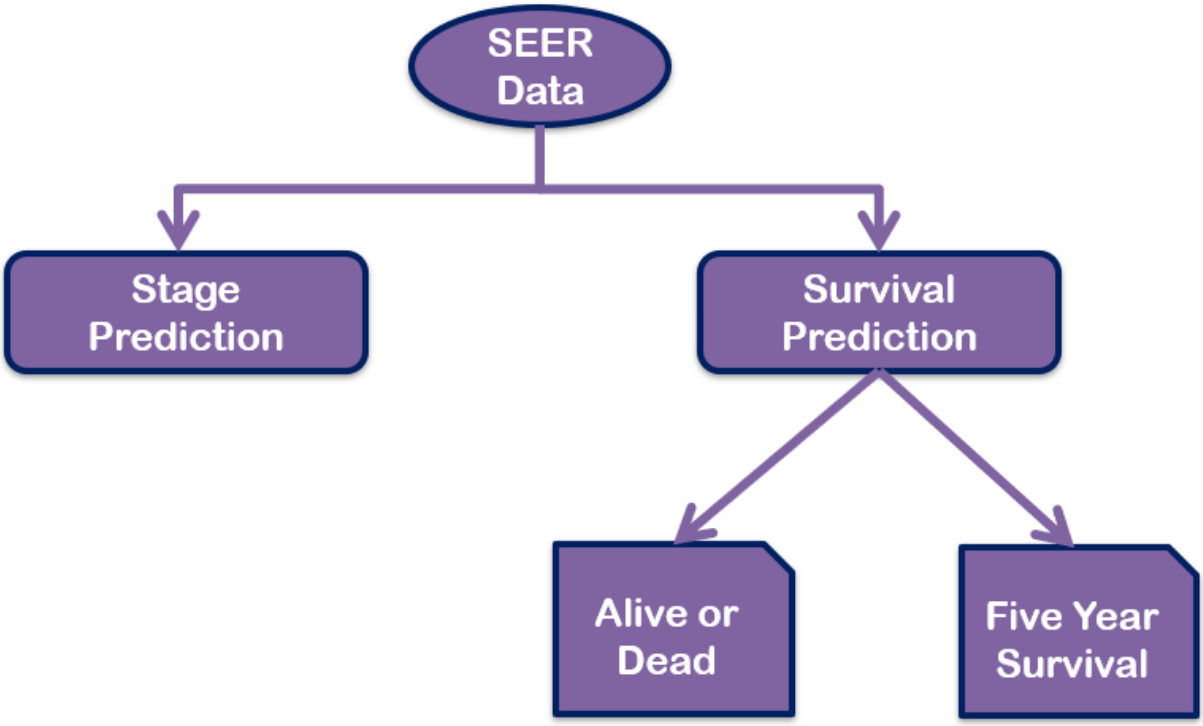

Fig. 2.5.1. Target classes

1.6 Expected Outcome

Determine the stage of prostate cancer in patients and their chances of survival. The survivability and stage of a man's prostate cancer are two of the most important factors to consider while weighing the available treatment choices. Sometimes a doctor would try to diagnose a patient based on their physical state by comparing them to the patient they just saw; nevertheless, in their whole career, a doctor can only diagnose a small number of prostate cancer patients. Patients will be able to receive the appropriate therapy in a shorter amount of time if the staging and prognosis of survival processes can be sped up. In this sense, our goal was to create artificial prediction models that could determine the stage of prostate cancer and the patient's likelihood of survival. We anticipated that the area of medicine would benefit significantly from the work that we did.

1.7 Report Layout

The paper is organized as follows:

i. Background
ii. Research Methodology
iii. Experimental Results and Discussion
iv. Summary, conclusion, Recommendation and implication for future Research
v. Reference.

CHAPTER 2

Background Study

2.1 Terminologies

Our main goal is to forecast the stage and the survivability of patients who have prostate cancer. We also tried to predict five-month survivability of prostate cancer patients. Prostate cancer is the second-leading cause of cancer deaths for men in the U.S. If the correct stage and survivability can be predicted, it will be easier for doctors to provide the right treatment. Machine Learning (ML) approaches have recently attracted significant attention, particularly in constructing data-driven prediction models. We constructed models with the support of ML techniques. We also tested how fast the models work on different platforms. On the later part, we did some comparative study with other works. We took inspiration from some notable researchers who are doing tremendous work in this field. The fact that we both have the same long-term objectives is what will make it possible for us to achieve greater success in the field of medicine.

2.2 Related Works

Machine learning methods have been used in a lot of cancer diagnostic research. Wen [1] researched prostate cancer prognosis. He applied standard preprocessing techniques, ANN, and ML methods such as Naive Bayes(NB), Decision Trees(DT, K Nearest Neighbors(KNN), and Support Vector Machines (SVM). His target survival categories are 60 months or greater. ANN's result is the best achievement with 85.6% accuracy. By taking the same prostate target attribute (five-year survivorship), Delen contributed a work[2] using data mining tools and SEER data in 2006. In their research, they used ANN, DT, and LR classifiers. With an accuracy of 91.07% , ANN beats all other classifiers. Data duplication removal and feature selection were not used in their work.

Cirkovic [4], Bellaachia [6], and Endo [7] predicted breast cancer survival with a similar target approach. Cirkovic [4] developed a ML model to predict breast cancer patients' chances of survival and recurrence. However, they only used 146 records and twenty qualities to predict 5-year survivability; the NB classifier was chosen as the best model. Bellaachia [6] used three data mining techniques on the SEER dataset: NB, back-propagation neural networks, and DT(C4.5) algorithms, where C4.5 performed better overall. By contrasting seven models (LR, ANN, NB,

Bayes Net, DT, ID3, and J48), Endo [7] attempted to create the five-year survival state. The logistic regression model shows the highest accuracy, 85.8% .

Montazeri [5] and Delen [8] survival indicate whether the patient will live or die. Montazeri [5] developed a rule-based classification approach for breast cancer. He employed a dataset of 900 patients, just 24 of whom were men or 2.7% of the whole patient population. He applied traditional preprocess techniques and ML algorithms, including NB, DT, Random Forest(RF), KNN, AdaBoost, SVM, RBF Network, and Multilayer Perceptron. He assessed the model using accuracy, precision, sensitivity, specificity, and area under the ROC curve and 10-cross fold validation. The RF with an accuracy of 96% was better than previous techniques. Delen [8] combined a widely utilized statistical technique, logistic regression, and decision trees to create the prediction models for breast cancer. For unique performance comparison, 10-fold cross-validation techniques were shown. The DT model exhibits the greatest performance with 93.6% accuracy.

The authors of [1],[2] focused on predicting whether a prostate cancer patient will live for sixty months or five years, referring to survival. However, since our objective attribute in this paper is different, we attempted to predict whether a patient would live or die in this experiment, known as the Vital Status Recode feature in the SEER database. This characteristic is used extensively to predict other cancer survival [5],[9],[8] and [10]. Fewer records are provided by Montazeri [5], Agrawal[11] and Lundin[12] to use in predicting cancer survival. It glances that Wen[1], Delen[8], and Pradeep[10] used a small 4 Model Analysis for Predicting Prostate Cancer Patient's Survival number (less than five) of algorithms for predicting breast, prostate, and lung cancer. Models for breast and lung cancer were constructed using a minimal set of characteristics in [11],[12], and [7] by Agrawal, Lundin, and Endo. They also did not mention the prediction time of any models, and there are differences in the number of attributes, which has allowed our efforts to go forward.

Deep learning and machine learning approaches are being used to detect prostate cancer by Adeel Ahmed Abbasi [15] and Lal Hussain [16]. They employed CNN, Decision Tree, SVM, Bayes, RBF, and Gaussian machine learning algorithms in their research. GoogleNet employs a deep learning CNN technique to achieve 100% specificity, sensitivity, PPV, and TA and an AUC of 1.00 [15]. Based on single feature extraction algorithms, SVM Gaussian Kernel has the most

remarkable accuracy of 98.34% and an AUC of 0.999. The SVM Gaussian kernel with texture + morphological and EFDs + morphological features produces the maximum accuracy of 99.71% and AUC of 1.00 when utilizing a combination of feature extraction methodologies [16].

Pushpanjali Gupta investigated [17] using a Machine Learning Approach to forecast Colon Cancer Stages and Survival Periods. Their investigation used data from 4021 patients and roughly 28 characteristics. Random Forest, Support Vector Machines, Logistic Regression, Multilayer Perceptron, K-Nearest Neighbor, and Adaptive Boosting were among the machine learning techniques employed. The top-performing model, Random Forest, with an accuracy rate of 84%.

Predicting High-Risk Prostate Cancer Using Machine Learning was headed by Henry Barlow [18]. A total of 35,875 data points were gathered. The machine learning algorithms used were KN, SVM, DT, RF, MLPC, ADA, and QD. Their maximum level of accuracy was 91.5% found in ADA. Guanjin Wang published exploratory research [19] on using ml techniques to diagnose prostate cancer in a Chinese population. They employed a dataset of 1625 Chinese patient records, all of whom had undergone TRUS biopsy. SVM, LS-SVM, ANN, and RF are four standard machine learning algorithms used in this work to find PC linked with diagnostic indicators. With ANN, they were able to attain an accuracy of 95.27%.

Osama Hamzeh attempted[20] to determine tumour location in prostate ˙ cancer tissue using a machine learning system and gene expression data. SVM RBF, Naive Bayes, and Random Forest classifiers were utilized. With 99% accuracy, SVM RBF performs better. Ebru Erdem[21] attempted to compare several supervised ml approaches for prostate cancer prediction. They tried NB, LR, K-NN, SVM, Linear Regression, RF, LDA, MLP, and DNN. With the MLP classifier having the accuracy of 97%. Olivier Regnier-Couder led research [22] to enhance prostate cancer pathology staging utilizing ml. They conducted their investigation using the BAUS dataset. They employed K-NN, RF, LR, MLP, RBF, SVM, CHBN, TAN, and NB machine learning algorithms. BNs have a higher AUC of 67.9% than the other techniques when the number of variables increases. Mohammed Ismail B also does research [23] on utilizing ML Classification to predict prostate cancer. In this study, they employed DT, ANN, KNN, SVM, RF, LR, and a suggested Modified LR. The application of the proposed strategy has a maximum accuracy rate of 96.6%On the other hand, Jae Kwon ˙ Kim [24] compared the efficacy of Machine Learning

Classifiers in predicting prostate cancer pathology staging. About 944 patients' records are stored in SPCDB. The dataset was subjected to back BPN, SVM, NB, BNs, Classification and Regression Tree (CART), and RF. SVM has a 75% higher accuracy than the other approaches.

M.N. Doja used machine learning to study age-specific survival in prostate cancer [25]. A cancer hospital in India provided data on patients with metastatic prostate cancer. They employed DT, LR, SVM, Boosted trees, and Bagged trees as machine learning approaches. The best accuracy, 81.4%, was discovered using Bagged Trees. Sran Jovi used machine learning to predict the likelihood of prostate cancer [26]. They used three machine learning techniques. ELM, ANN, and GP are the three approaches. The coefficients of determination for the ELM, ANN, and GP techniques were 0.9976, 0.9647, and 0.9204, respectively, based on the results. ELM, ANN, and GP had a root mean square error of 0.0167, 0.0642, and 0.0964. The simulation results demonstrated that the ELM model could forecast prostate cancer likelihood favourably, resulting in the most accurate predictions. Raphael Lenain, on the other hand, attempted to derive phases from pathology reports on prostate cancer [27]. The T, N, and M classifications were used to define stage annotations at diagnosis. The classifiers SVM, DT, RF, and XGB were tested. They discovered that the N and M stages had the best results. Precision, recall, and F1-score were all 0.99 using SVM and Gradient Boosting classifier at the M stage.

Blazˇ Zupan analyzed the survival of prostate cancer patients with ml in a study paper[28]. They utilized two different datasets. DT, NB, and Cox's proportional hazards models are employed. The Naive Bayes classifier had an accuracy of 70.8% in the preoperative dataset. However, when all of the classifiers were applied to the postoperative dataset, Naive Bayes had the best accuracy of 78.4%. Furthermore, Felix D. Beacher conducted research [29] to predict the results of prostate cancer Phase III clinical trials. They employed classifiers such as XGBoost, Catboost, KNN, Logistic Regression, and Voting. XGBoost classifiers outperformed other classifiers with an accuracy of 84%.

Regarding survival, the authors of [1] concentrated on estimating whether a patient with prostate cancer will live for 60 months or five years. They built the model using data from 2004 to 2009, where the usage of 15 attributes was observed. No feature impact analysis information was found. Additionally, we also choose five years of survival as our objective quality. We also attempted to predict whether a patient would live or die in this experiment, known as the Vital

Status Recode feature in the SEER database. However, this trait is heavily utilized to forecast the survival of other cancers [4], [6], and [8]. In order to forecast cancer survival, Montazeri et al. [5], Agrawal et al. [11], and Lundin et al. [12] employed fewer records. For breast, prostate, and lung cancer survival prediction, it appears that Wen et al. [1], Delen [8], and Pradeep [10] utilized a few (fewer than five) different methods. Agrawal et al. [11], Lundin et al. [12], and Endo et al. [7] created models for breast and lung cancer using a limited set of traits. Additionally, none of the models' prediction times was mentioned, and the different quantity of characteristics permitted our attempts to move forward.

2.3 Research Summary

Although prostate cancer is the most common form of cancer in males, when detected in its early stages, it may frequently be effectively treated. The prostate gland, which may be found between the genitalia and the bladder, is the initial point of reference. Approximately one in eight men will be diagnosed with prostate cancer at some point in their lives. It is the second most common reason for men to pass away from cancer in the United States. Accurate staging and a reliable survival prediction are both essential components of treatment that minimizes mortality. When people go to their doctor's chamber, they frequently have to wait for many hours before obtaining any kind of consultation. This is an extremely common occurrence. A significant number of patients are unable to see their primary care physician due to the poor state of their health. If they have access to the results of all of their medical tests, those individuals may use our model to identify the stage of their prostate cancer and whether or not they will survive it. On the other hand, a physician engages in conversation with a large number of patients on a daily basis. As a consequence of this, it may be difficult to determine the specific stage of a prostate cancer patient and the patient's chance of survival. Our method will assist physicians in doing their duties by accurately predicting the stage at which a patient is now located as well as their likelihood of survival in a timely manner. It will be much simpler for medical professionals to provide the appropriate medication if they are able to rapidly detect the precise stage of the patient's prostate cancer, as well as whether or not the patient will survive the next 5 years, or if they do survive the next 5 years. Patients diagnosed with prostate cancer have a lower chance of passing away if they are given the appropriate therapy at the appropriate time. We have tried to predict the cancer stage and survival probability of a prostate cancer patient using artificial intelligence and machine learning and compared our work with that of others. Also, we have shown how our

models will perform on different platforms. As a result, we can easily choose the model that provides the fastest and most accurate predictions.

2.4 Scope and Challenges

There has been and will continue to be a great deal of research done on prostate cancer. In the course of this research, we intend to make use of statistical comparisons in order to appropriately evaluate the prognosis of prostate cancer patients, taking into account both the stage of their disease and the possibility that they will continue to live. In order to get a deeper comprehension of prostate cancer, we searched through the entirety of the SEER database as well as the research that was related to it. We preprocessed the data that was obtained, used feature engineering, performed correlation analysis, and constructed machine learning models in order to anticipate the appropriate stage and the actual survival. There is a variety of different models that may be used for prediction, and it is crucial that we choose the one that is the most dependable. It was not an easy task to create a model that was capable of providing the most accurate forecasts possible. In addition to this, we evaluate the overall performance of our models on a wide range of different platforms.

CHAPTER 3

Research Methodology

3.1 Introduction

Our desired models were built using data from the machine life cycle. Significant procedures have been carried out, including data collecting, feature elimination, data preparation, correlation analysis, data partitioning into train and test sets, model creation, hyperparameter tuning, cross-validation, and model testing.

3.2 Data Source

Data for this study has been acquired from the SEER database server. SEER is the U.S. Government's Official source for cancer statistics. The Surveillance, Epidemiology, and End Results (SEER) program collects data on cancer statistics with the goal of lowering the cancer burden in the United States and the whole world. The collection has 37 features and 187798 entries. The AJCC has found that these elements have a causal connection to cancer and can offer intelligence to machine reasoning inclusion in the SEER database. [3].

3.3 Data Collection Procedure

The data were taken from the SEER database in order to complete this investigation. Its objective is to bring the total number of sick persons residing in the United States down to a more manageable level. The identification of patients who have been diagnosed with cancer or who have received cancer treatment at hospitals, outpatient clinics, radiology departments, physician offices, laboratories, surgical centers, or other providers (such as pharmacists) who diagnose or treat cancer is the first step in the process of collecting data on cancer. All fifty states are required by law to report any new instances of cancer to a national register that is maintained by the government. Cancer registries look into these cases to see whether or not they are mandated to be reported to the appropriate authorities by law. If this is the situation, registries will access the patient's medical records in order to collect data on cancer in accordance with the NAACCR Data Standards. For the purpose of their analysis, they pre-processed data that they had acquired

from a variety of sources. Because of this, we went directly to the source and obtained it on our own.

3.4 Data Preparation

One of our study project's most difficult tasks is data collection. But the main challenge was to applied the algorithm. As we are not using framework, we have used the raw python coding to do that the data must need to prepare and specified folder to access the datasets. Here we prepare data for three types of prediction- General survivability, Five-year survivability, and Stage prediction.

Handling of missing value and data duplication:

The SEER statistics on prostate cancer are deficient in many different aspects. Due to the advancement of medical knowledge, new characteristics have emerged that did not previously exist. Consequently, prior patients' information on these current characteristics is unavailable. 10349 entries are entirely missing from our database, or around 5.5% of all records. This 10349 data point is negligible when compared to 187798 data points. Since prostate cancer patients expect accurate diagnostics for better treatment adding the missing information may lead to erroneous results. Therefore, the missing records have been eliminated. There is a possibility that the physiology of many patients is similar. For that, we discovered data duplication, which might have resulted in inaccurate conclusions; thus, we eliminated the redundant data. There are 54731 records in total for type 1 and 2, excluding duplicates. For type 3, 177449 records remain after deleting the records.

Feature Engineering: A prediction model can only be created by feeding the data into the numerical format. For prediction type 1 and 2, nineteen out of the twenty-seven characteristics in the dataset are nominal, while eight are numerical. Therefore, nineteen attributes must be transformed into a machine-readable format. The Label encoding module from the Python scikit-learn package was used to accomplish this goal. Label encoding is the process of taking textual labels and transforming them into numeric representations. Likewise, 15 of the 23 attributes in the dataset that make up prediction type 3 are nominal, while the remaining 8 are numerical. So, 15 attributes need to be converted to machine-readable format. For which we have used the get dummies module from the Python Pandas library. This is often referred to as" one-hot encoding." The get dummies module creates a unique column for each class of an attribute, and patients

have one if the class is upbeat and zero if it is negative. Target attribute contains a class unknown stage which is not in a machine learning format. The rest of the classes are in numerical format as usual, which exists from 1 to 4. So unknown stage has taken as 5 and sent for model building.

Correlation Analysis: Correlation coefficients show how connected attributes are. Strong correlation coefficients of +0.8 to +1 and -0.8 to -1 represent the same behavior [14].

$$x = \frac{\sum(xi-\overline{x})(yi-\overline{y})}{\sqrt{\sum(xi-\overline{x})^2(yi-\overline{y})^2}} \quad (1)$$

We found the correlation coefficients(r) using equation 1. Xi is the value of one feature, and the x bar is the average of all the values. Yi is the value of another feature, and the y bar is the average of all the values for that feature. Using Correlation analysis, we found some pairs of attributes which act in a same way and from those we selected one from each pair to build the prediction model.

3.5 Research Subject and Instrumentation

Our research topic is Model Analysis for Predicting Stage and Survival of Prostate Cancer Patients, which is mainly focused on Machine Learning Techniques. The Intel Xeon CPU from Google Colaboratory was utilized to generate models for the experiment. ML-based prediction model created using scikit-learn, pandas, NumPy, and seaborn library. Python programming language was used in general. For the purpose of model construction and simulation, we used Google Colaboratory. The models' prediction times were evaluated on many platforms and utilized a linear process with a single core of the following CPU: Intel Xeon, Intel Core i5-9300H, and Ryzen 7 3700X CPUs.

3.6 Used Machine Learning Models:

We build different models using machine learning methods for predicting survivorship, five-year life expectancy and cancer stage.

3.6.1 Models for Predicting Survivorship: The performance of twelve ML algorithms employed to predict survivorship. The ML methods include LGBM Classifier, Random Forest [11], Extra Trees, Logistic Regression [2, 7, 8], SGD Classifier, HGB, Gradient Boosting, XGBoost [14], KNN, Decision Tree [1, 4, 8, 11], AdaBoost classifier [13]. Our finest discovered predicting model is created using the XGboost classifier.

XGBoost Classifier: XGBoost is a decentralized gradient boosting toolkit that has been tuned for efficiency, flexibility, and portability. It incorporates machine learning methods inside the context of Gradient boosting. XGBoost is a parallel tree boosting algorithm that solves several data science challenges quickly and precisely. The same code operates on major distributed environments and is capable of solving issues that exceed billions of instances.

3.6.2 Models for Predicting Five-Year Survivorship: Performance of the six ML models and ANN used to forecast survival. LGBM, RF [11], ETC [7, 8], GBC [14], DT [1, 4, 8, 11], and AdaBoost classifier [13] are the machine learning techniques. The Gradient Boosting Classifier was used to generate our best forecasting model. Gradient Boosting: Gradient boosting classifiers combine weaker learning models to build a powerful predicting model. Gradient boosting techniques are used for identifying challenging datasets. This is used for regression and classification. It is an ensemble of weak prediction models that usually use decision trees.

3.6.3 Models for Predicting Cancer Stage: Thirteen machine learning algorithms have been used for stage prediction, and their performance has been tested at the end. The algorithms are LGBM Classifier, Hist Gradient Boosting Classifier, Gradient Boosting Classifier [25,26], XGB Classifier [26,29], Kneighbors Classifier [18,29], Decision Tree Classifier [18,24-25], Random Forest Classifier [18,20,24,27], Extra Trees Classifier, Logistic Regression [17,25,29], SGD Classifier, AdaBoost Classifier[18]. LGBMClassifier, Hist Gradient Boosting Classifier and Gradient Boosting Classifier.

LGBM Classifier: LightGBM is a framework for gradient boosting that makes use of tree-based learning techniques. It is dispersed and efficient, with the benefits of increasing training pace and efficiency, reducing memory uses, increasing precision, parallel process and GPU-accelerated learning are supported. Capable of managing massive amounts of data Many Machine Learning Algorithms may be used to assess the dataset's correctness. However, the LGBM Algorithm has been deemed the most accurate [citation Prediction of Type-2]. Light Gradient Boosting Machine is the acronym for the LGBM Algorithm. There are two basic principles in the LGBM Algorithm. GBDT (Gradient Boosting Decision Tree) and GOSS are two examples (Gradient-based one-sided sampling) and using GBDT we reached our destination. In this path learning rate= parameter was 0.1 and n estimators was 100. Moreover for parallel processing n jobs was -1.

Hist Gradient Boosting Classifier: This algorithm's central principle is the Histogram-based Gradient Boosting Classification Tree. The method can be built using scikit-learn. For large datasets with far more than 10000 data points, this estimator is much quicker than Gradient Boosting Classifier. This estimator comes equipped with built-in support for missing data. If no missing values are found during training for a particular feature, samples with missing values are mapped to the child with the most samples. Input samples X into integer-valued bins, which are typically 256 bins in size, significantly reducing the number of splitting points to evaluate and allowing the algorithm to construct the trees using integer-based data structures (histograms) rather than sorted continuous values. This implementation is based on the LightGBM standard.

3.7 Implementation Procedure:

I At the very beginning of this process, a comprehensive diagram is developed to represent the general workflow that is going to be followed. As the picture makes evident, the first step in our methodology is to collect data from the SEER program, which adheres to a distinct and unmistakable chronological order of occurrences. After we had obtained the datasets, it was important to do preprocessing on them by cleaning and filtering them before utilizing the algorithms. This was required before we could use the datasets. After that, we began looking into the possible connections between the two events. When we did this, we subsequently produced two distinct sets of data: one for use in evaluation, and the other for use in instruction. The majority of the data, around 80%, will be used toward the purpose of training, while the

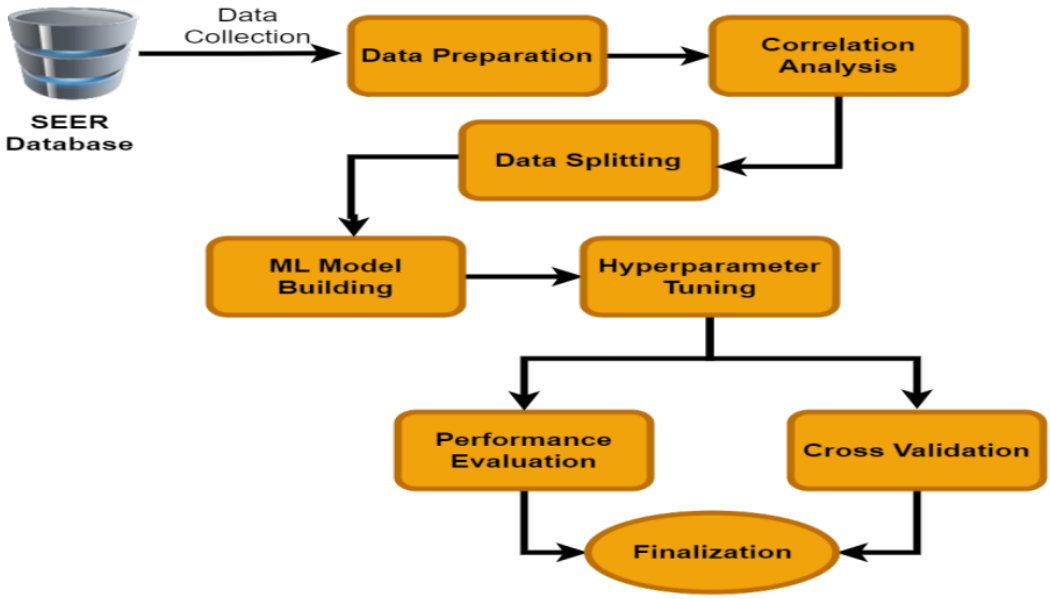

Fig. 3.7.1. Procedural Framework

remaining data, 20%, will be put toward the purpose of testing. Now that we've completed the process, we are able to declare that we created the models using methods that are associated with machine learning. In order to evaluate the outcomes, we carried out a cross-validation procedure. Following the initial model development step, we employ hyperparameter tuning for stage prediction. The information presented in Figure 3.7.1 helps us understand the entire process

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

We discussed the dataset, dataset processing techniques, and machine learning models in the prior section. This section will describe the evaluation approach as well as the results of the models that employ the processed data. Several machine learning algorithms were applied, and the results are being examined to see which method delivers the best accuracy.

4.2 Performance Evaluation Parameter

Following the development of the machine learning models, specific tests are run to determine their viability. Our model's accuracy [15-25,28,29], F1 score [17-19,21,27,29], Precision [17,20,21], Recall [17,21,27], Cross-Validation [9,14,18-19], and time interpretability have all been assessed. A confusion matrix is a required component for measuring these metrics. A confusion matrix, also known as an error matrix, is a table structure used in machine learning to show the effectiveness of a supervised learning system. We get True Positive (TP), False Positive (FP), True Negative (TN), and False Negative using a confusion matrix (FN).

Equations that need to calculate ML performance measurements:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

$$Precision = TP/(TP + FP) \quad (4)$$

$$F1Score = 2 * (Recall * Precision)/(Recall + Precision) \quad (5)$$

AUC: Area Under the Curve (AUC) It has used to measure performance over many criteria. It measures how far apart result is. It also shows how successfully the model classifies data.

ROC Curve: The total classification levels of a categorization model are represented graphically by a receiver operating characteristic (ROC) curve. This curve depicts two variables: True Positive and False Positive Rates.

Sensitivity: Sensitivity is another word for actual positive rate, which is the percentage of positive samples that give a positive result when a specific test is added to a model and does not change the samples.

$$Sensitivity = TP/(TP + FN) \quad (6)$$

Specificity: In the context of an unaffectedly negative model, the true negative rate, sometimes referred to as specificity, is the percentage of samples that test negative when the test is employed.

$$Specificity = TN/(TN + FP) \quad (7)$$

Cross-Validation: Cross-validation is crucial after model construction. As a result, we employed stratified K-fold cross-validation. Each time, the data set was split into ten equal folds, and a model was generated using nine folds and evaluated with one-fold. Thus, the average value has been compared to the model accuracy after ten iterations. The rationale for utilizing the Stratified K-fold is that it evenly distributes the target attribute's classes across all folds. We have experimented with a variety of approaches. We go into great depth on how we were able to achieve our goals.

4.3 Experimental Results & Analysis

We sought to create a model that could predict the stage, survivability, and five-year life expectancy of a patient's prostate cancer using machine learning techniques. Many prediction models have developed, and the mandatory job is to pick the most effective one for each type of prediction.

Findings of Correlation Analysis: For prediction model type 1 and 2, we can see that four pairs of attributes behave similarly in Figure 3.4.1. The manners of" Histology recode broad groupings" and" ICD-O- 3 Hist or behav" have the same impact because their correlation coefficient is 0.98. In the same way," RX Summ Surg Prim Site" and" Reason no cancer directed surgery" have 0.9, "CS Mets at dx" and "Derived AJCC M" have 0.95; finally, "CS lymph node"

and ”Derived AJCC N” have 0.94 coefficients. These pair’s impacts are the same in the dataset. Models will be created with one attribute from each of these two pairs. ’ICD-O- 3 Hist or behav’, ’RX Summ Surg Prim Site’, ’CS lymph nodes’, and ’CS mets at dx’; these four attributes have been omitted. Moreover, the rest of the attributes have their own individuality.

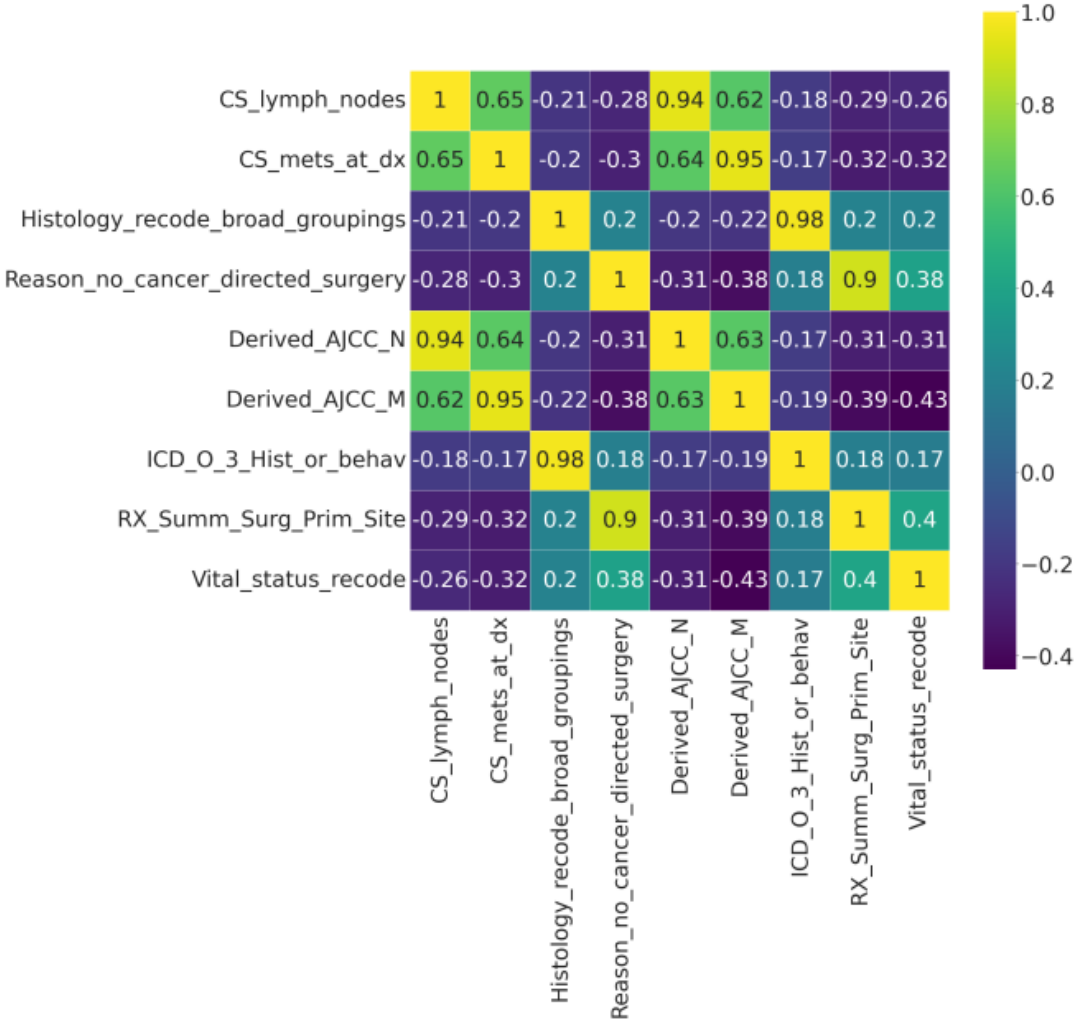

Fig. 4.3.1. Zoomed Correlation Heatmap (Alive or Dead and Five-Year Survival)

For prediction type 3, the manners of “Histology recode broad groupings” and “ICDO-3 Hist or behav” are similar because their correlation coefficient is 0.98. In the same way, “RX Summ Surg Prim Site” and “Reason no cancer directed surgery” behave the same religion, and their coefficient is 0.89. We can clearly notice it from Figure 3.4.2. Models will be created with one attribute from each of these two pairs. So ’Histology recode broad groupings’, and ’Reason no

cancer directed surgery' these two attributes have been omitted. Moreover, the rest of the attributes have their own individuality.

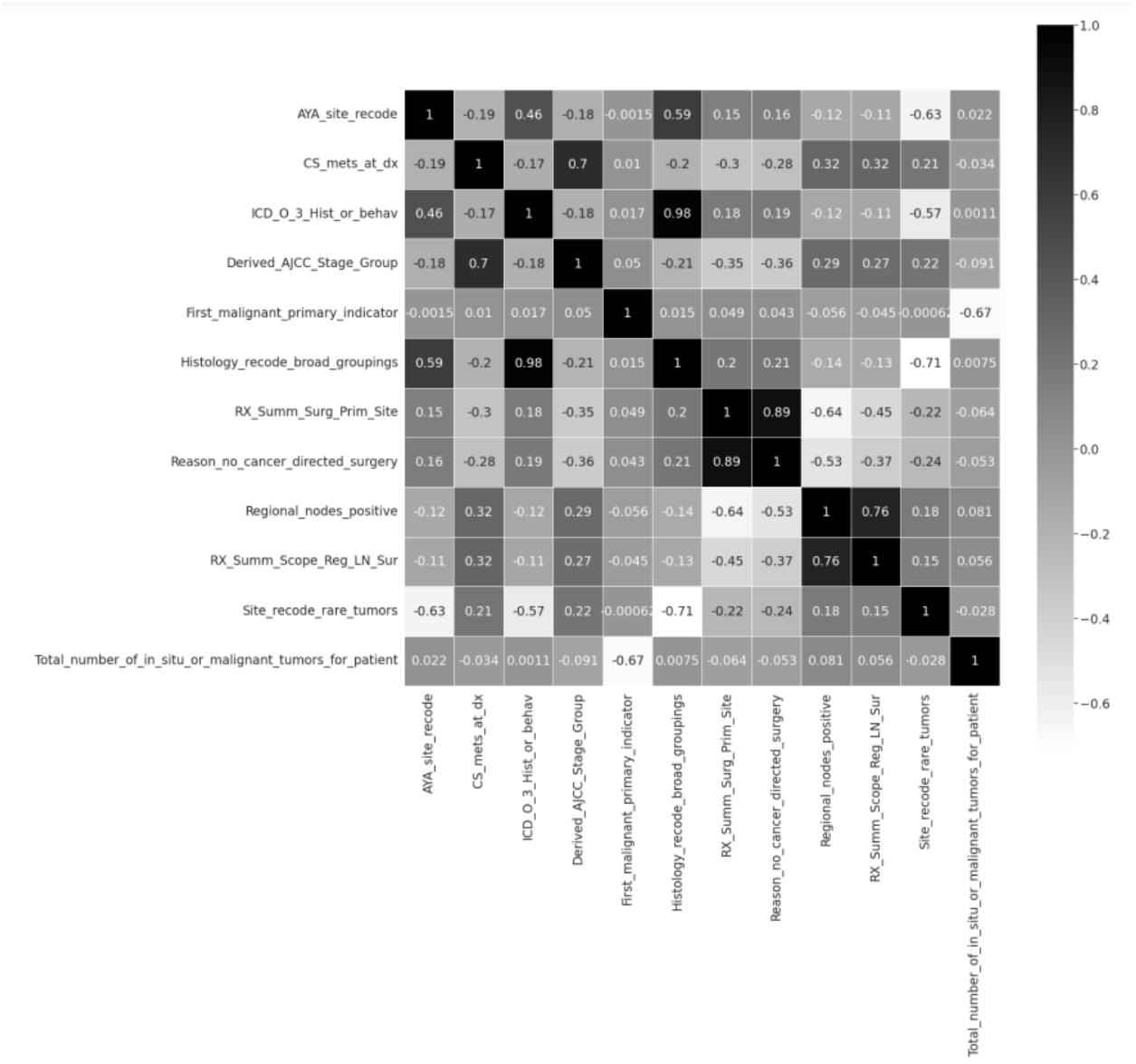

Fig. 4.3.2. Zoomed Correlation Heatmap (Stage)

Feature Impact Analysis: The idea of "feature impact" is what's utilized to figure out which features in a dataset have the most significant impact on the conclusions that a machine learning model draws from its analysis of the data. In addition, feature impact is used in feature selection,

which is one of the best ways to increase the accuracy of the models, and in target leakage identification, which is one of the best ways to prevent severely incorrect models. Both of these methods make use of one of the most effective ways to prevent severely incorrect models. Indicative of target leakage is when a single variable has an influence on the findings of the model that is disproportionately larger than what would be expected from that characteristic alone. Figures 3.4.1 through 3.4.4 show that "CS_mets_at_dx", "CS_lymph_nodes", and "CS_tumor_size" have a strong influence on five-year survival prediction, and "Laterality", "AYA_site_record", and "Race record" have a lower impact.

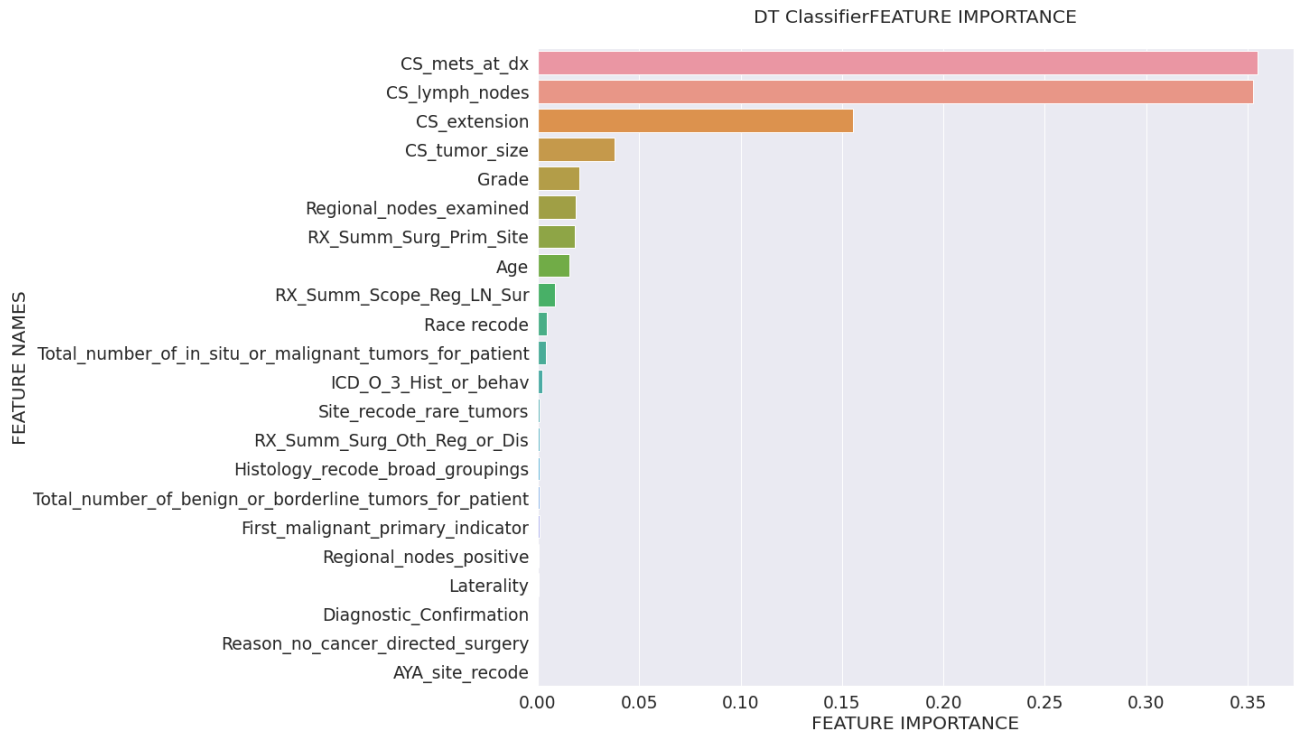

Fig. 4.3.3. DT Classifier Feature Importance (Five year Survival)

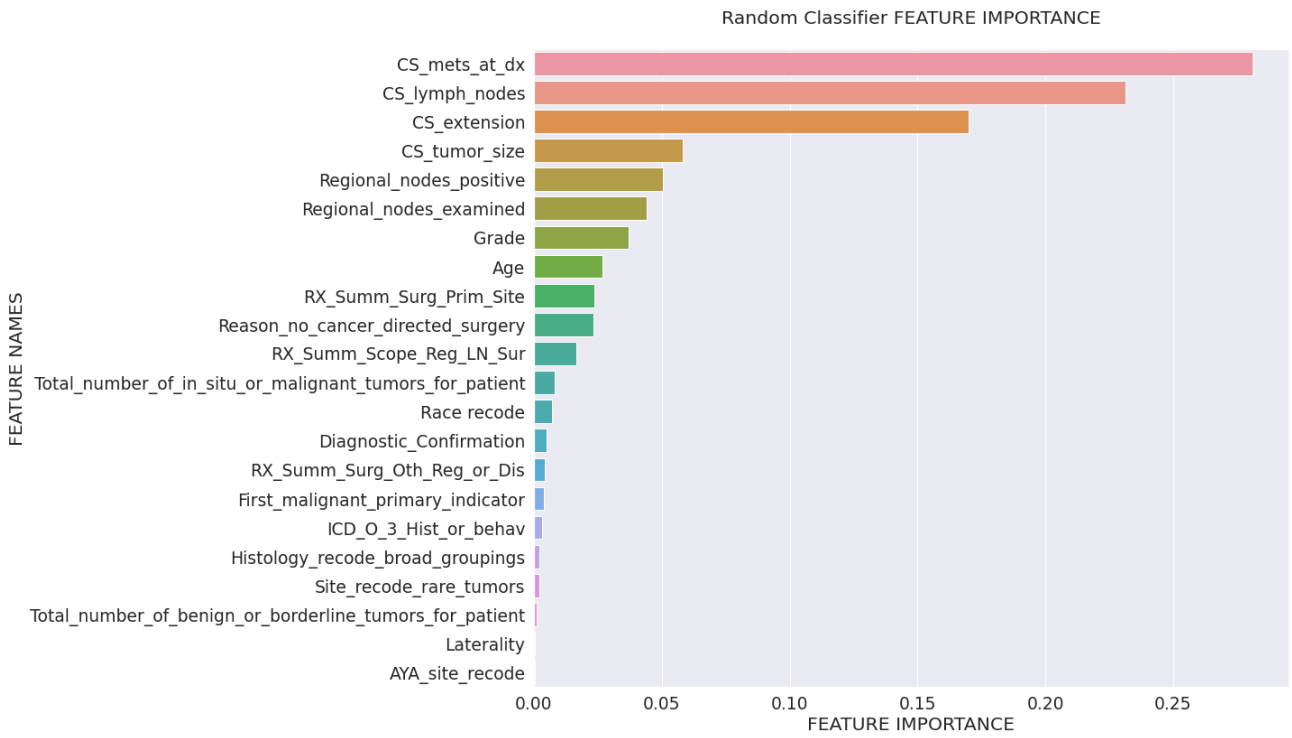

Fig. 4.3.4. Random Classifier Feature Importance (Five year Survival)

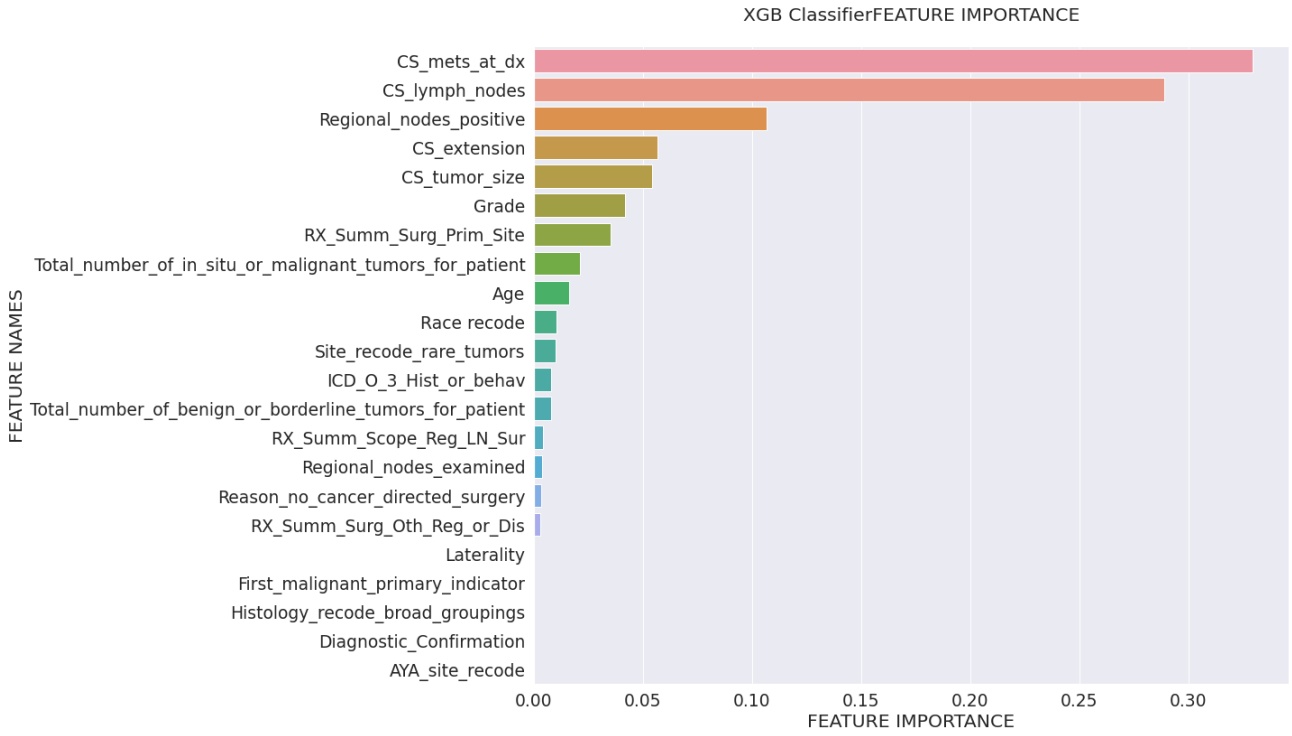

Fig. 4.3.5. XGB Classifier Feature Importance (Five year Survival)

	Specs	Score
19	CS_lymph_nodes	21875084.15
18	CS_extension	2328863.91
17	CS_tumor_size	1829309.70
20	CS_mets_at_dx	1711872.59
21	Regional_nodes_positive	232319.46
16	Regional_nodes_examined	216693.95
9	RX_Summ_Surg_Prim_Site	50578.63
4	Site_recode_rare_tumors	9999.41
7	Reason_no_cancer_directed_surgery	9664.44
12	Grade	5274.90
15	RX_Summ_Scope_Reg_LN_Sur	4773.50
11	Total_number_of_in_situ_or_malignant_tumors_fo...	3159.55
2	ICD_O_3_Hist_or_behav	2348.17
0	Age	1337.70
13	Histology_recode_broad_groupings	559.98
14	Diagnostic_Confirmation	142.06
10	Total_number_of_benign_or_borderline_tumors_fo...	119.10
5	AYA_site_recode	36.74
8	First_malignant_primary_indicator	35.97
6	RX_Summ_Surg_Oth_Reg_or_Dis	27.48
1	Race recode	26.64
3	Laterality	0.19

Fig. 4.3.6. Chi- Square (Five year Survival)

Figure 3.4.5 to 3.4.8 show that “Age”, “CS_mets_at_dx”, and “CS_tumor_size” have a greater influence on the Alive and Death survival forecast. The effects of “Laterality”, “AYA_site_record”, and “Total_number _of_bening_or_borderline_tumors_for_patient” are, on the other hand, less significant.

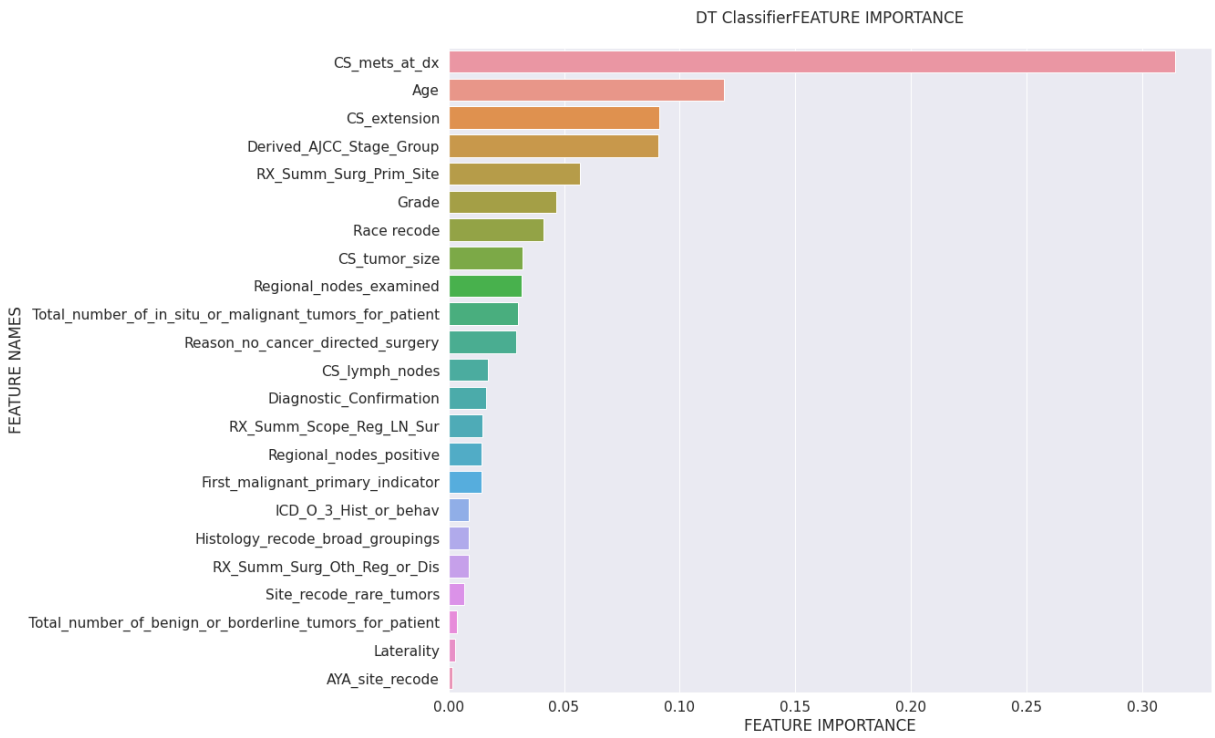

Fig. 4.3.7. DT Classifier Feature Importance (Death or Alive Survival)

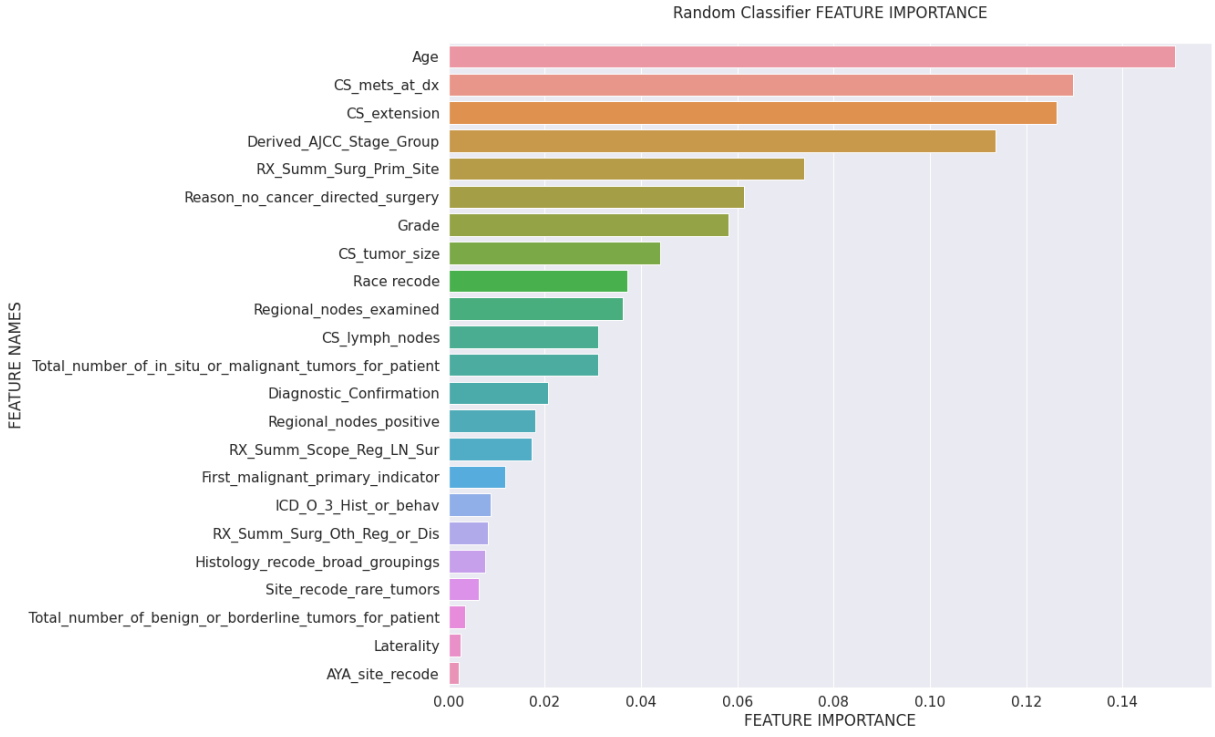

Fig. 4.3.8. Random Classifier Feature Importance (Death or Alive Survival)

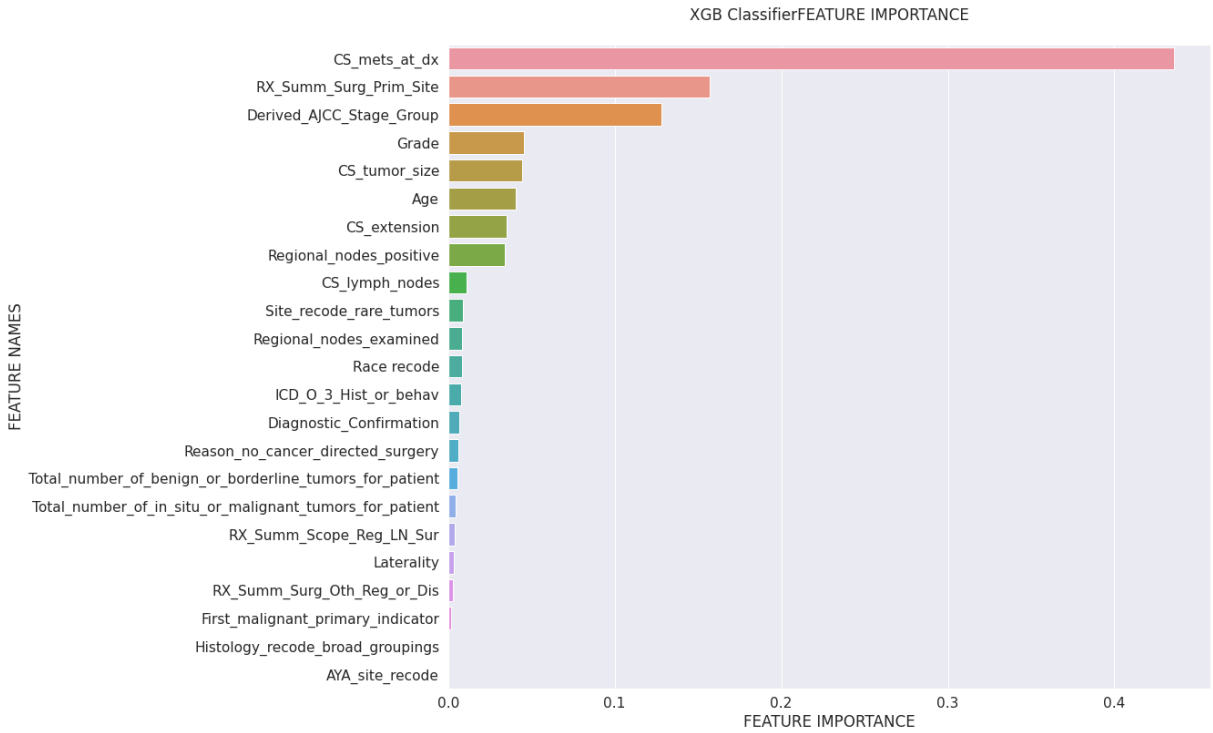

Fig. 4.3.9. XGB Classifier Feature Importance (Death or Alive Survival)

	Specs	Score
20	CS_lymph_nodes	2580918.66
18	CS_tumor_size	1200957.69
19	CS_extension	1154193.50
21	CS_mets_at_dx	289676.87
22	Regional_nodes_positive	170777.02
9	RX_Summ_Surg_Prim_Site	49294.76
4	Site_recode_rare_tumors	10740.84
7	Reason_no_cancer_directed_surgery	8070.75
16	Derived_AJCC_Stage_Group	5038.83
12	Grade	2948.42
0	Age	2434.66
15	RX_Summ_Scope_Reg_LN_Sur	2190.10
2	ICD_O_3_Hist_or_behav	1869.70
13	Histology_recode_broad_groupings	483.41
17	Regional_nodes_examined	344.51
11	Total_number_of_in_situ_or_malignant_tumors_fo...	144.37
14	Diagnostic_Confirmation	105.10
10	Total_number_of_benign_or_borderline_tumors_fo...	41.40
5	AYA_site_recode	28.36
6	RX_Summ_Surg_Oth_Reg_or_Dis	7.04
1	Race recode	3.99
8	First_malignant_primary_indicator	0.25
3	Laterality	0.01

Fig. 4.3.10. Chi- Square (Death or Alive Survival)

Last but not least, when it comes to stage prediction, “CS_lymph_nodes”, “CS_extension”, and “CS_tumor_size” have a bigger influence, whereas “Laterality”, “AYA_site_recode”, and “First_malignant_primary_indicator” have a smaller impact compared to other variables, as can be shown in Figure 3.4.9 to 3.4.12.

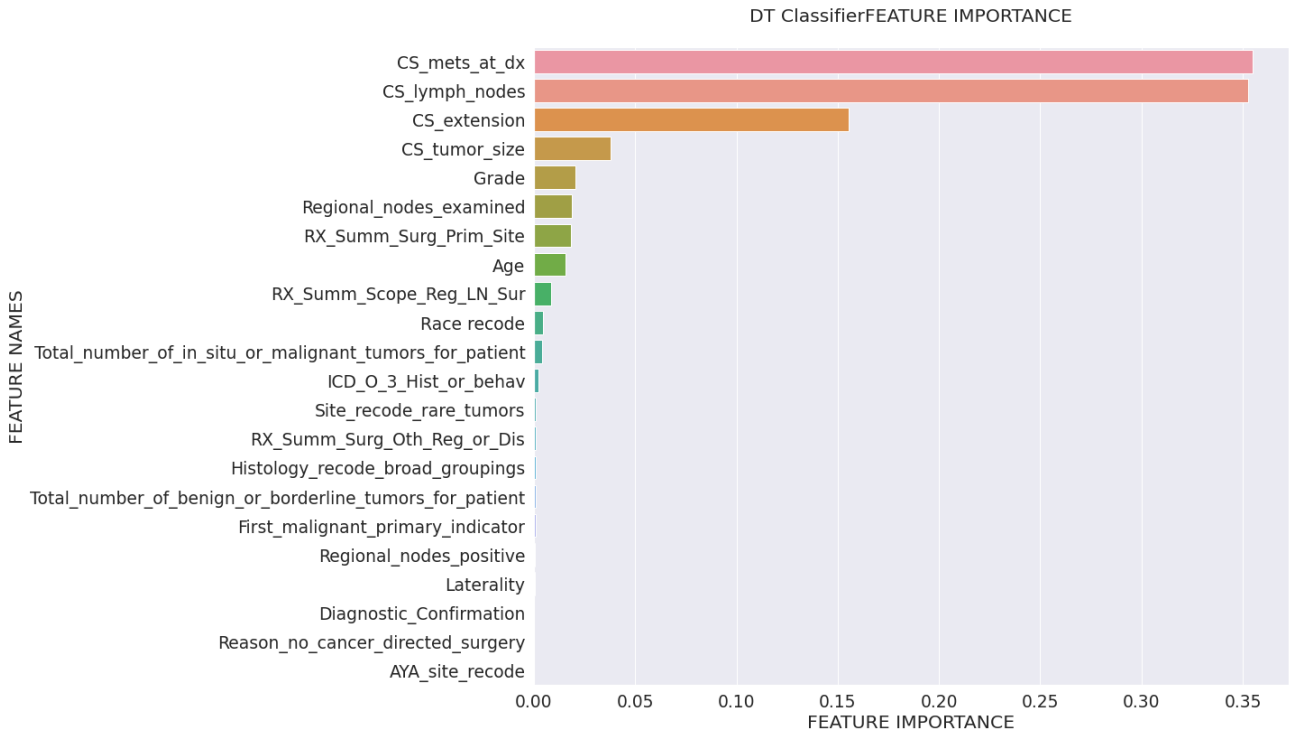

Fig. 4.3.11. DT Classifier Feature Importance (Stage)

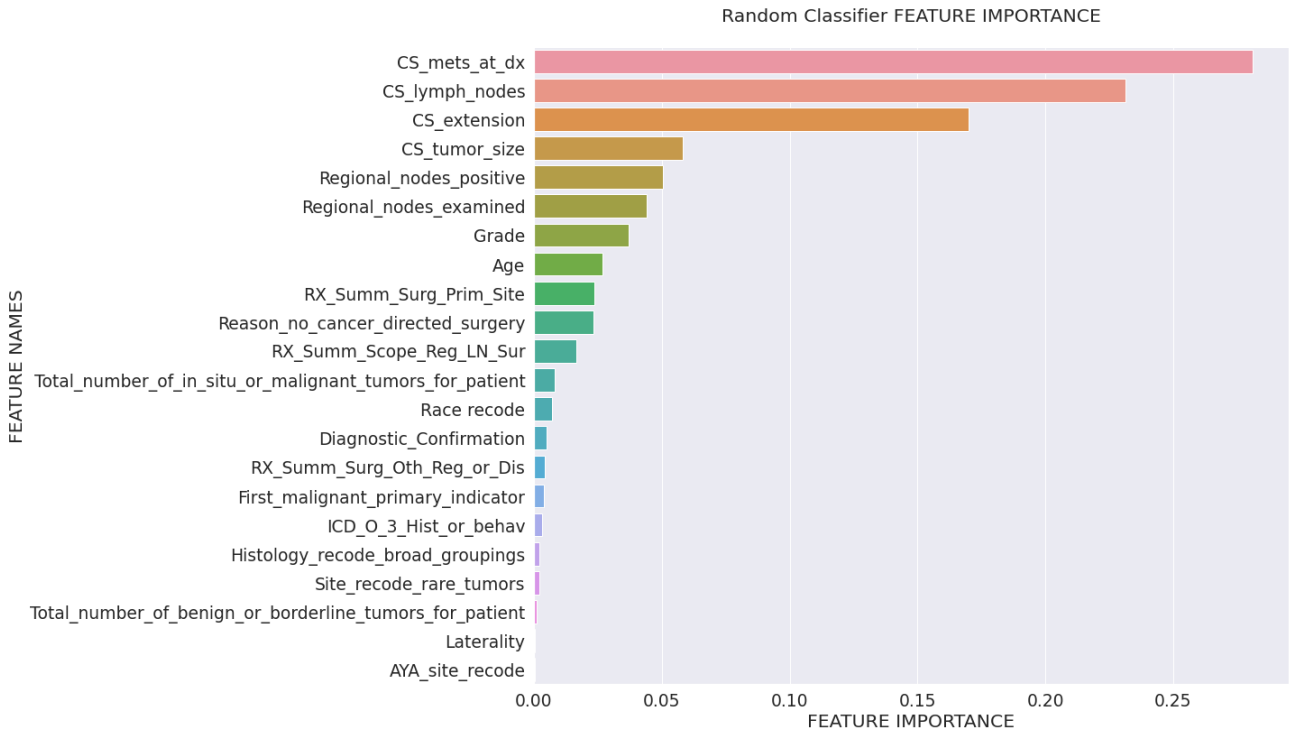

Fig. 4.3.12. Random Classifier Feature Importance (Stage)

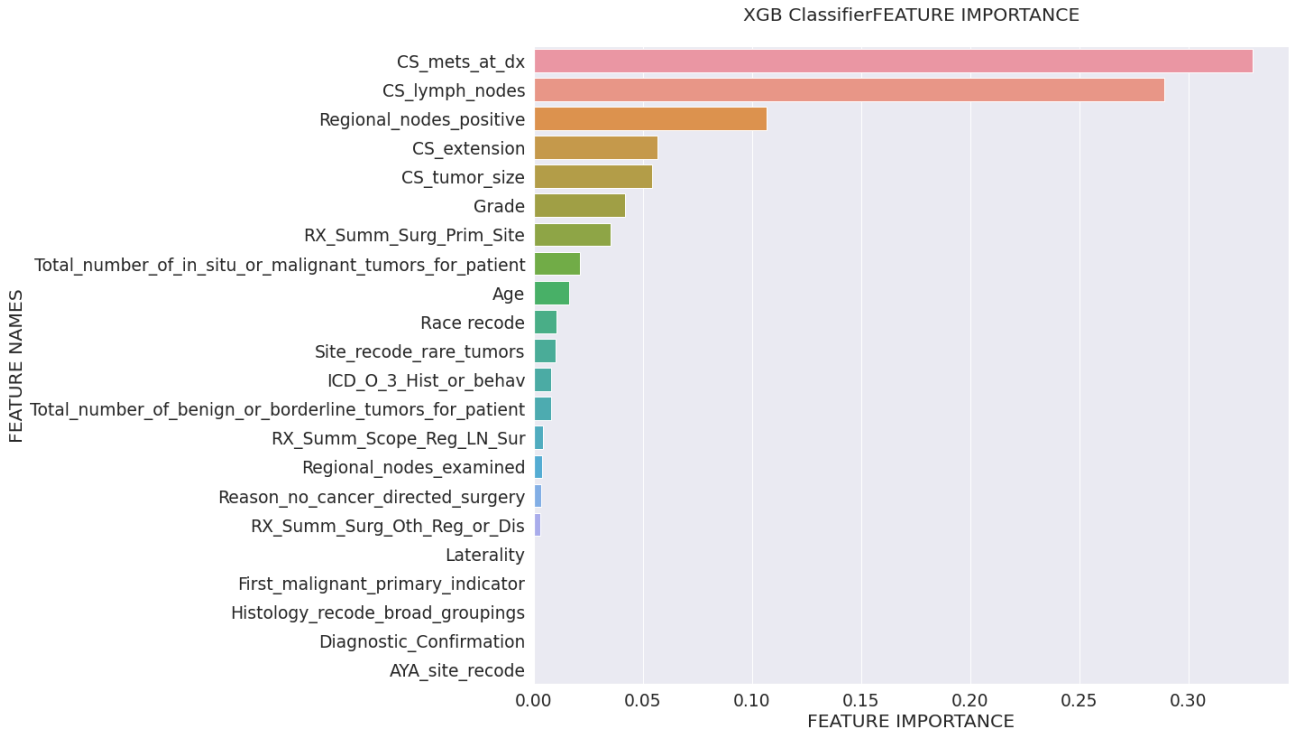

Fig. 4.3.13. XGB Classifier Feature Importance (Stage)

	Specs	Score
19	CS_lymph_nodes	21875084.15
18	CS_extension	2328863.91
17	CS_tumor_size	1829309.70
20	CS_mets_at_dx	1711872.59
21	Regional_nodes_positive	232319.46
16	Regional_nodes_examined	216693.95
9	RX_Summ_Surg_Prim_Site	50578.63
4	Site_recode_rare_tumors	9999.41
7	Reason_no_cancer_directed_surgery	9664.44
12	Grade	5274.90
15	RX_Summ_Scope_Reg_LN_Sur	4773.50
11	Total_number_of_in_situ_or_malignant_tumors_fo...	3159.55
2	ICD_O_3_Hist_or_behav	2348.17
0	Age	1337.70
13	Histology_recode_broad_groupings	559.98
14	Diagnostic_Confirmation	142.06
10	Total_number_of_benign_or_borderline_tumors_fo...	119.10
5	AYA_site_recode	36.74
8	First_malignant_primary_indicator	35.97
6	RX_Summ_Surg_Oth_Reg_or_Dis	27.48
1	Race recode	26.64
3	Laterality	0.19

Fig. 4.3.14. Chi- Square (Stage)

Prediction Model for Prostate Cancer Survival (Alive or Dead): We used ML methods to build a model that could predict a patient's prognosis for prostate cancer survival. Table 4.3.1 shows the models' prediction performance. Consequently, we can observe that the tree-based boosting techniques performed exceptionally well on our dataset. Each of them has an accuracy of little more than 89%. Now, we must choose the optimal algorithm from among these candidates. In Table 4.3.1, we can see that Hist Gradient Boost performs best based on accuracy, and its accuracy is 89.61%.

Table 4.3.1 Performance Measurements of Survivability (Alive or Dead) Models

Algorithms	Accuracy	F1 Score	Precision	Recall	AUC	Avg. CV
HGB Classifier	89.6136	0.8187	0.8453	0.7984	0.9243	0.8874
LGBM	89.5862	0.8176	0.8458	0.7964	0.9249	0.8872

Classifier						
XGB Classifier	89.5679	0.8143	0.8501	0.7891	0.923	0.8883
Gradient Boosting Classifier	89.5588	0.8164	0.8464	0.7942	0.9229	0.8881
Ada Boost Classifier	89.4857	0.8148	0.8454	0.7923	0.9192	0.8864
Logistic Regression	87.7592	0.7748	0.8224	0.7459	0.8862	0.8717
ANN	89.2116	0.8111	0.8386	0.7904	0.913	-

The AUC score from Table 4.3.1 and the ROC curve in Figure 4.3.1 shows that the best four algorithms, HGB, LGBM, XGBoost, and Gradient Boosting Classifier, have covered almost 92% of the data correctly. However, we have to think about which algorithm has been able to identify the target two classes more accurately. The sensitivity of Figure 4.3.1 clearly distinguishes the algorithms. Because the sensitivity of the XGBoost has achieved 0.7851, which

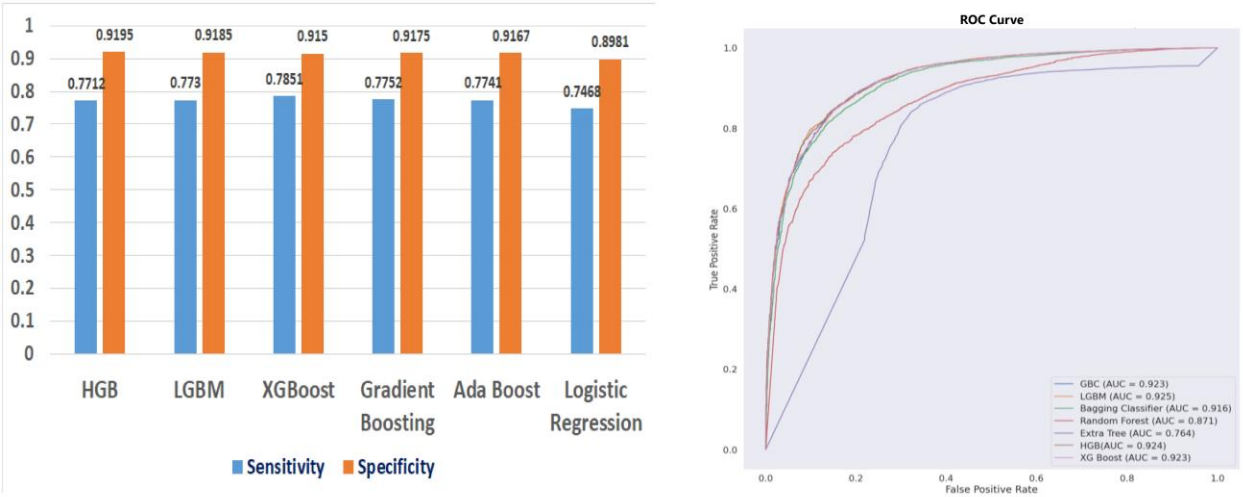

Fig. 4.3.15 Sensitivity and Specificity of Models and ROC Curve (Alive or Dead)

is more than others. Nevertheless, the sensitivity and specificity of the algorithms do not show much difference. For which the prediction time of the algorithms has been calculated. Different sets of data have used to check the prediction time; test data, train data, and single-person data were on the list. As the test records are unseen for the models, we represented prediction time with test data in Figure 4.3.2. Till now, Hist Gradient Boosting Classifier, LGBM Classifier, and

XGB Classifier algorithm accuracy are ahead in the race, but among them, XGB Classifier can predict in the shortest time.

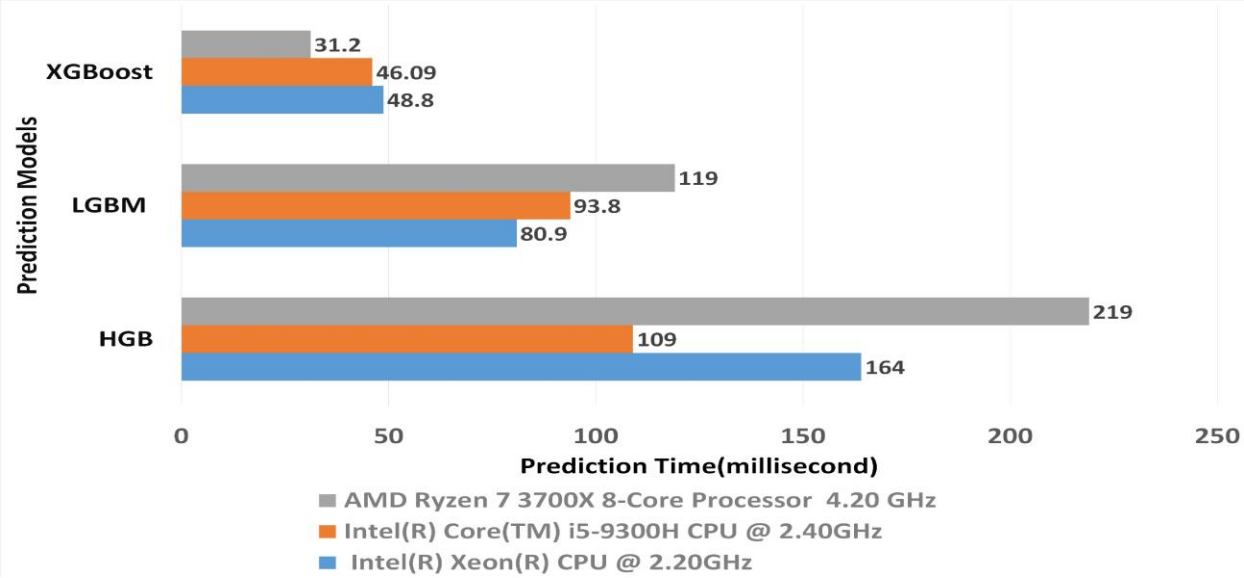

Fig. 4.3.16 Prediction time for test data with best three models (Alive or Dead).

We can see that in three different platforms, on average the XGB Classifier takes 42.03 milliseconds, the LGBM Classifier takes 97.9 milliseconds, and the HGB classifier takes 164 milliseconds. It indicates that the prediction model using the XGboost model is nearly two times and four times faster than LGBM and HGB models, respectively. In other sets of data, we also received the same ratio of time difference. Although we can see from Figure 4 that the XGBoost can identify the dead class 0.02% less than LGBM, which is a tiny difference, but XGBoost model is two times faster than LGBM. Therefore, we came to the conclusion that the predictive model using XGB Classifier would be our proposed model. Let us look at the in-depth analysis of the XGBoost model.

Our accuracy score in the XGBoost Classifier is 89.56% which is very close to the LGBM classifier (Table 4.3.1). The report also said that the macro f1 score is 0.8143, the recall score is 0.7891, and the precision score is 0.8501, which is so near to the LGBM classifier. The average cross-validation score was 88.83% in the 10 folds cross-validation. That means this model perform better than LGBM(88.72%) and HGB(88.74%) in total data. This model is not overfitted either, since the training error is within acceptable bounds. Overall the XGBoost prediction model outperforms all models and is our recommended model.

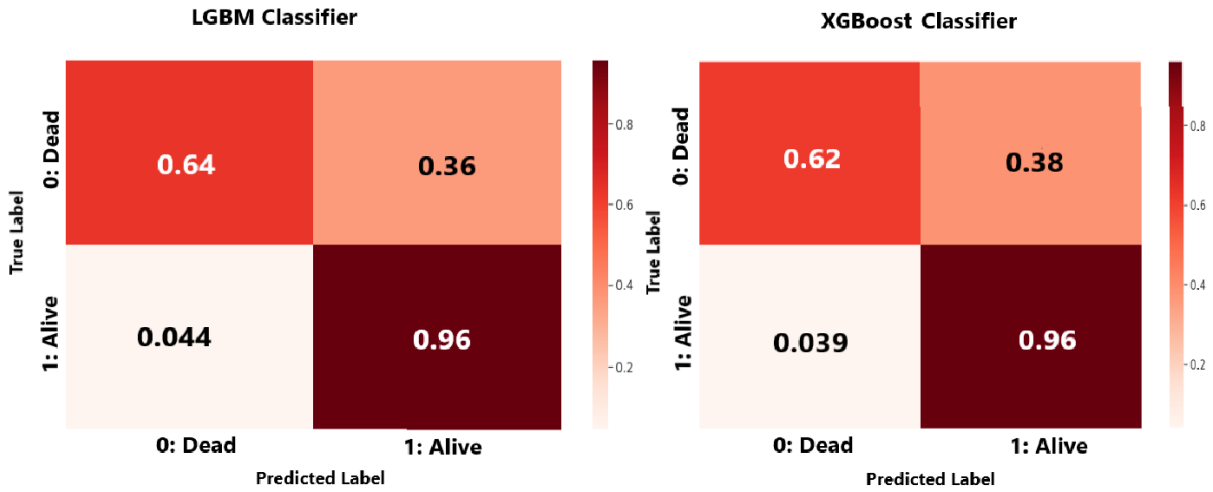

Fig. 4.3.17 Normalised Confusion Matrix of LGBM and XGBoost classifier for survivability prediction (Alive or Dead).

Model for Prostate Cancer Five-Year Survival: We also tried to predict five-year survivorship by building models through machine learning approaches. The predictive efficacy of the models is displayed in Table 4.3.2. Thus, it is clear that tree-based boosting methods outperformed when applied to our dataset. Each of them is around 86% to 88% accurate. Now we have many potential models, need to choose the best one.

Table 4.3.2. Performance Measurements of the ML Algorithms for Five-Year Life Expectancy

Algorithms	Accuracy (%)	F1 Score	Precision	Recall	AUC	Avg. Cross Validation
Tuned Gradient Boosting	88.45	0.844	0.842	0.8470	0.905	0.8811
Gradient Boosting	88.386	0.8443	0.8416	0.8471	0.9044	0.8798
LGBM	88.0351	0.8371	0.8396	0.8346	0.8989	0.8794
Ada Boost	86.9825	0.8254	0.8229	0.8281	0.8989	0.8758
Random Forest	86.5965	0.817	0.8202	0.814	0.8776	0.8596
Extra Trees	85.0877	0.7947	0.8005	0.7895	0.8522	0.851
Artificial Neural Network (ANN)	84.98	0.8033	0.7949	0.8134	0.852	0.824
Decision Tree	82.8421	0.767	0.7682	0.7659	0.7914	0.8238

In table 4.3.2, Accuracy, Recall, Precision, and F1 score were calculated using equations 2, 3, 4, and 5, respectively. There we can see that Gradient Boost performs best based on accuracy, and its accuracy is 88.386%. The accuracy difference between the top two models is 0.27%, which implies GBC can accurately predict nearly ten more people than LGBM. The AUC score from Table 4.3.2 and the ROC curve in Figure 4.3.4 shows that the top performed model GBC covered almost 90.044% of the data correctly. The rest of the algorithms are also covered near of the GBC model, but the blue curve indicates GBC model covers more data than other models. However, we must consider which algorithm can identify the two target classes more accurately.

Thus, the sensitivity (green pillars) of Figure 4.3.5 clearly distinguishes the algorithms. The sensitivity of the GBC model has achieved 0.9265, which is more than others. Nevertheless, the specificity (blue pillars) is slightly down from the LGBM model, but the value is in an acceptable range. Using equations 6 and 7, the measurements have calculated. Moreover, we also tried to improve the performance of gradient boosting and found that the model's accuracy improved from 88.38 to 88.45 (Table 4.3.2) by tuning the hyperparameters. A 0.07% increase in accuracy resulted in 3 new patients correctly predicting life expectancy. The parameters that made this improvement are: min-samples-split is 10 and n-estimators is 200 rest of the parameters will be the default.

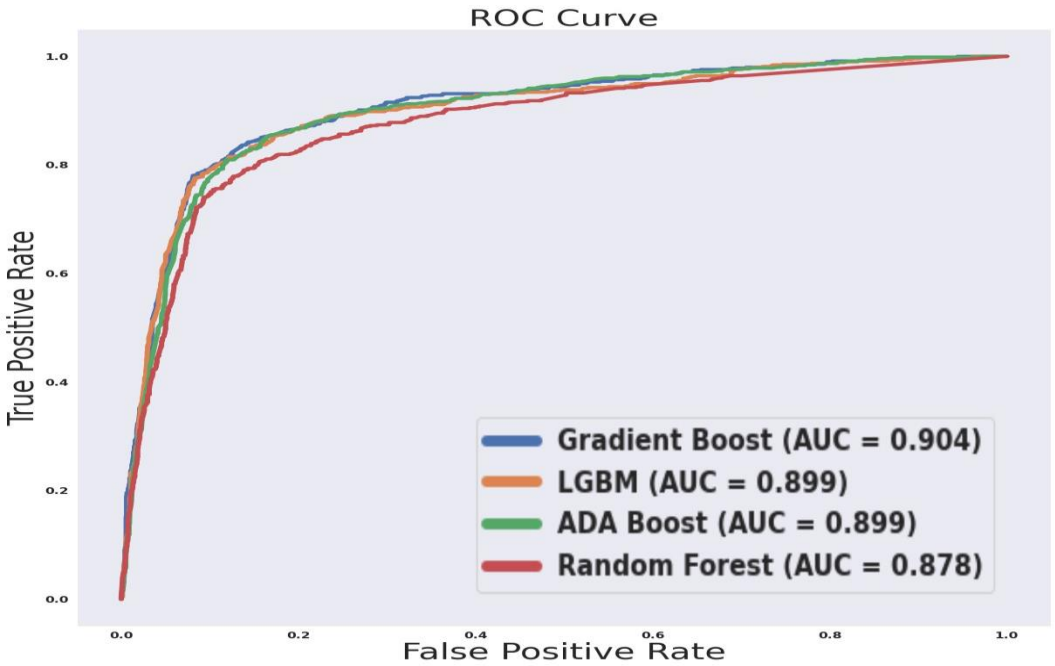

Fig.4.3.18. ROC curve (Five Year Life Expectancy)

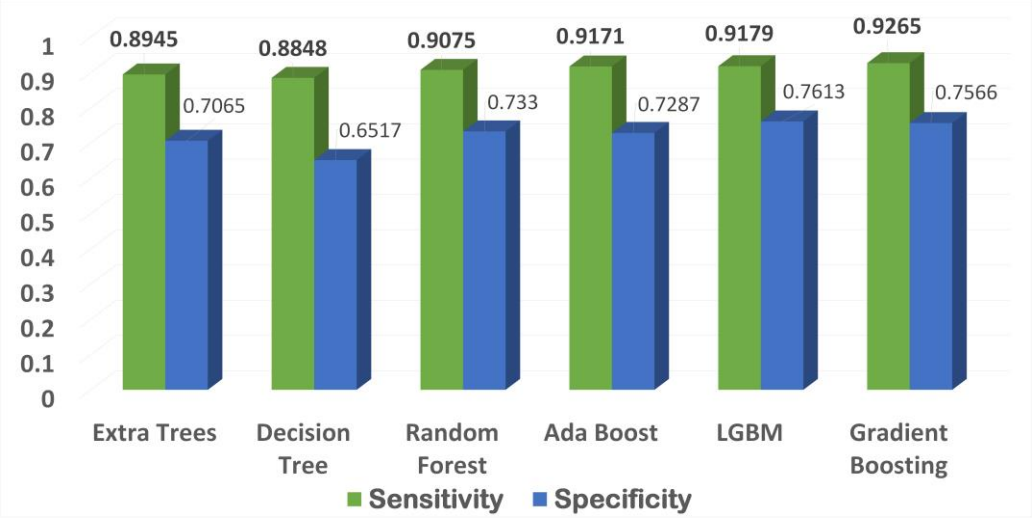

Fig.4.3.19. Sensitivity and Specificity of Algorithms (Five Year Life Expectancy).

time of each algorithm, and it specifies our preferred algorithm. Various ratios of the dataset, including test data, train data, and single-person data, were utilized to examine the prediction time. It is visualized that Tuned GBC has taken 14.7 milliseconds on average on three different platforms. The AMD Ryzen 7 and Intel Core i5 processors took 15.6 milliseconds, and the Intel Xeon processor from colab took 12 milliseconds to predict the test data. Compared with other prediction models, we can see LGBM took an average of 33.73 milliseconds, which is 2.29 times slower than the Tuned GBC model, and the ABC model took an average of 38.8 milliseconds it

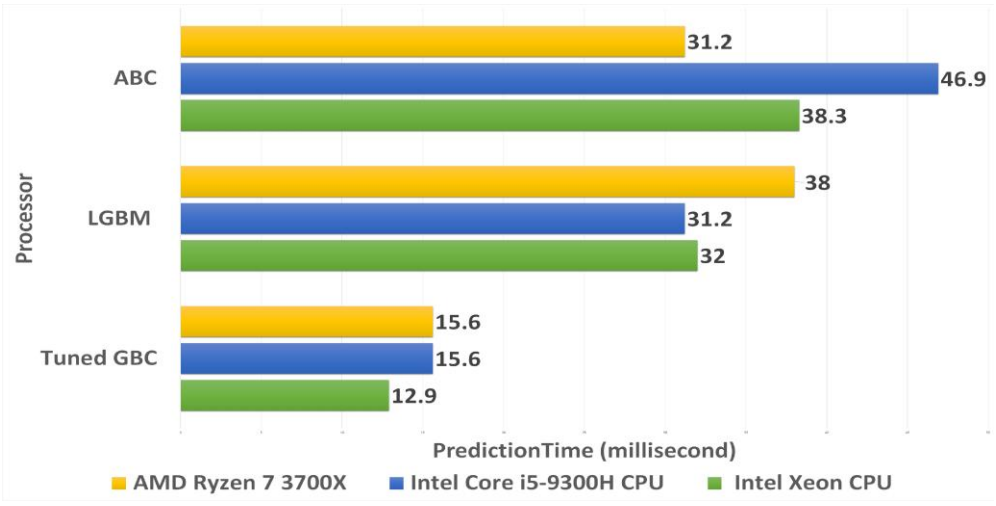

Fig.4.3.20. Prediction time for different models on test data (Five Year Life Expectancy).

is 2.6 times slower than the Tuned GBC model. Here clearly visualized that the tuned GBC model is faster than other models. As other authors did not mention cancer survival forecast models' prediction time, we could not compare them.

Till now, we can see that the gradient boosting model performs the best.

The GBC model gave us an F1 score of 0.8443 (Table 4.3.2). We were able to get an accuracy rate of 88.38%. The categorization report indicated that the recall was 0.847, and the precision was 0.841. Clearly can observe that each class identified very well through the measurements. This model was evaluated using stratified cross-validation. The 87.98 % average accuracy of 10 folds was discovered. Moreover, the model's training error was within acceptable limits. Both models are evaluated with test data; we first set aside the test data, then after training the model, test with unseen test data and get the results. Cross-validation yielded a positive outcome, which is a way to eliminate future data leakage; we may assume that the model will also do well with new data. Consequently, it is evident that this model does not exhibit any overfitting. Therefore, this is a good demonstration. Similarly, tuned GBC increased the accuracy from 88.386% to 88.45%, AUC 0.9044 to 0.905, cross-validation 0.8798 to 0.8811, and the rest of the values are nearly the regular GBC model. Comparing the normalized confusion matrix of Figure 4.3.7, we can see that the value of "61 to more months" has increased from 0.77 to 0.78. As we already know, three more patients' five-year survival becomes accurate for this increase of results, so the tuned gradient boosting model will be our proposed model.

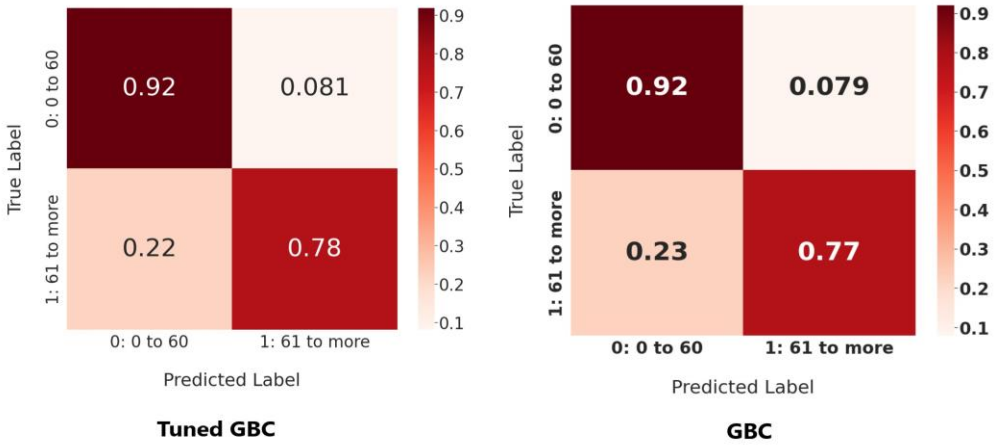

Fig.4.3.21. Normalised Confusion Matrix of Tuned GBC and GBC Model (Five Year Life Expectancy).

Model for Prostate Cancer Stage Prediction: ML algorithms have been used to develop a variety of prophecy models. Raphael Lenain has developed different classification models for stage T, N, M of prostate cancer [27]. But our model's target attribute indicates the overall stage of the cancer patient. A model's optimal performance must be determined based on its outcomes. Table 4.3.3 shows the findings of the experiment.

Table 4.3.3. Performance Measurements of the ML Algorithms for Stage Prediction

Serial	Algorithms	Accuracy	F1 Score	Precision	Recall	Avg. Cross Validation
1	LGBM Classifier	96.2708	0.9419	0.9693	0.9219	0.9612
2	Hist Gradient Boosting Classifier	96.2143	0.9418	0.965	0.9239	0.9585
3	Gradient Boosting Classifier	96.0335	0.9366	0.975	0.9119	0.9599
4	XGB Classifier	95.5588	0.9265	0.975	0.8987	0.9561
5	Kneighbors Classifier	93.9315	0.8968	0.9269	0.8718	0.9384
6	Decision Tree Classifier	93.3326	0.9096	0.9132	0.9061	0.9311
7	Random Forest Classifier	93.1292	0.8964	0.9061	0.8874	0.9295
8	Extra Trees Classifier	91.7053	0.8343	0.8447	0.8252	0.9114
9	Logistic Regression	85.1396	0.5082	0.5207	0.5128	0.8527
10	SGD Classifier	81.7154	0.534	0.5358	0.5423	0.7385
11	AdaBoost Classifier	33.0207	0.4686	0.5611	0.6516	0.315

The algorithms are sorted based on accuracy in Table 4.3.3. Moreover, the values of Accuracy, F1 Score, Precision, Recall and Cross-Validation of each algorithm are displayed. Which are calculated using equations 1,2,3, and 4. The best performing algorithm is LGBM Classifier, and the worst-performing algorithm is AdaBoost Classifier. There were also experiments with more algorithms, but their results were not acceptable. The rest of the algorithms present can accurately predict 80% to 95%However, due to the large size of the data, the accuracy˙ of many

records varies by a 1% difference. So, we have to choose the appropriate algorithm for the stage prediction of prostate cancer. Based on the accuracy, the two algorithms that showed the best performance are,

LGBM Classifier: We discovered that the F1 scores of classes 1, 2, 4, and 5 are more than 0.96 in the LightGBM Classifier, whereas class 3 earned 0.76. This implies that the algorithm is capable of accurately classifying each category. We achieved a score of 96.27% accuracy. According to the classification report, the macro f1 score is 0.94, the recall is 0.92, and the precision is 0.96. Moreover, the model's training error was 3.28 %. As a result, it is evident that this model does not exhibit any overfitting. To improve the testability of this model, stratified cross-validation was applied. The average precision of ten folds was determined to be 96.12 %. We can observe from the normalised confusion matrix in Fig. 4.3.8. that the model can adequately predict between 97 and 100 % of the test data for classes 1, 2, 4, and 5. Additionally, the model properly predicts 66% of class 3 events. Thus, we can say it is an outstanding model.

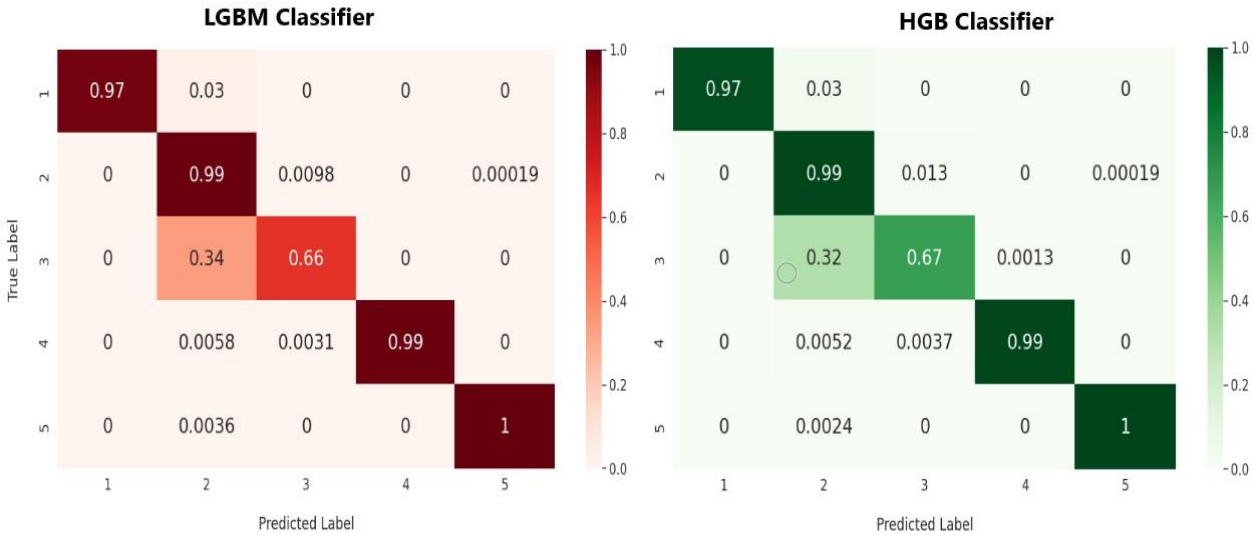

Fig.4.3.22. Normalised Confusion Matrix of LGBM and HGB classifier (Stage Prediction).

Hist Gradient Boosting Classifier: Our accuracy score in the Hist Gradient Boosting(HGB) Classifier was 96.21% which is very close to the LGBM classifier. The report also said that the macro f1 score is 0.94, the recall score is 0.92, and the precision score is 0.96, which are all the same as the LGBM classifier. The average cross-validation score was 95.85 % in the 10 folds crossvalidation. Compared to the LGBM classifier, this score is lower. This model is not overfitted either, since the training error is within acceptable bounds. The prediction rate of

classes 1, 2, 4, and 5 in the HGB classifier's normalised confusion matrix (Fig. 4.3.8) is similar to the LGBM classifier. However, this model predicts 67% of class 3 correctly, which is 1% better than the LGBM classifier.

As we have seen, the performance measurement of the two best algorithms is almost the same. So, we have to decide which model is more suitable. In this case, we have looked at how much time algorithms take to predict the stages. We found that the LGBM and Hist gradient boost classifiers are the top performers predicting the stages. From Fig. 4.3.9, we can see the LGBM model predicted the test data in 246 milliseconds and the Hist gradient boost model took 646 milliseconds to anticipate the test data. The Hist Gradient Boost algorithm takes about 2.6 times longer to predict, so it is slower than the LGBM algorithm. The figure also shows that other algorithms take less time, but their performance is not comparatively good. For example, the decision tree algorithm takes 13.7 milliseconds, but its accuracy is 93.33%. So, we are not proposing a decision tree. Above all, we propose the LGBM algorithm for stage prediction because it took less time than hist and predicted 0.06% more accurately.

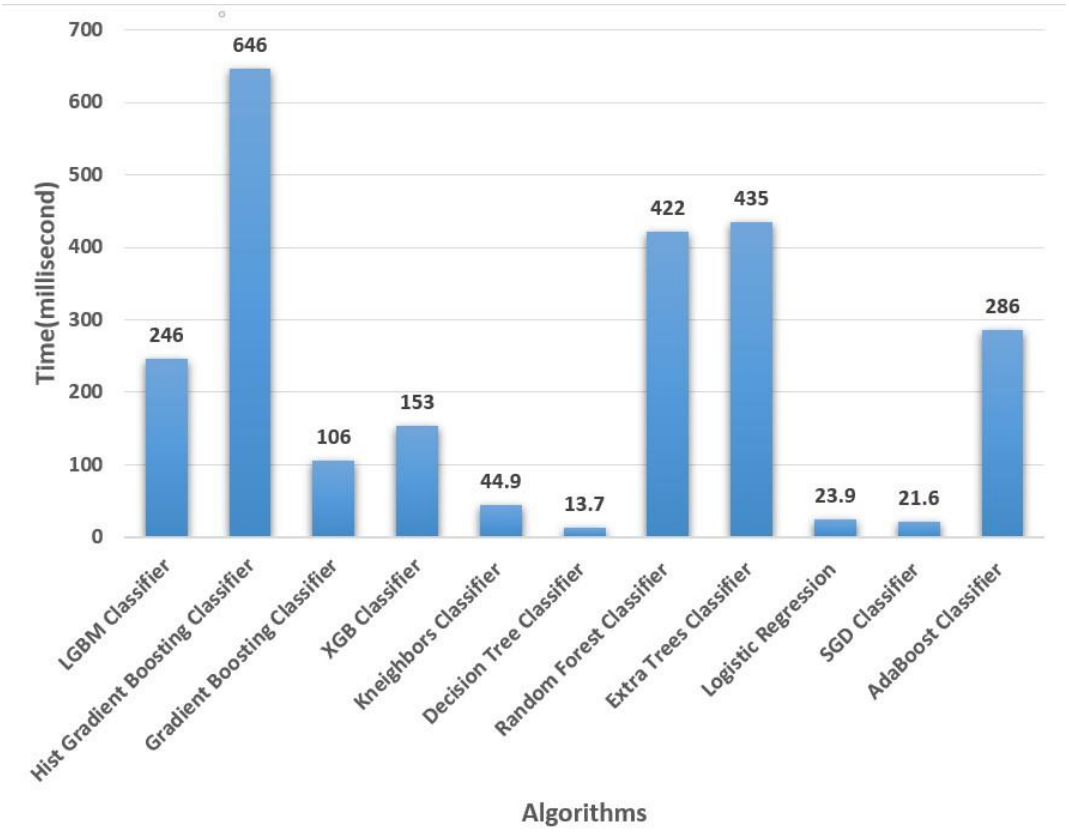

Fig.4.3.23. Model's time for the prediction (Stage Prediction).

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Prostate cancer is the second most often diagnosed illness and the fifth greatest cause of cancer mortality in males worldwide, with an estimated 1,414,000 new cases and 375,304 deaths in 2020. Prostate cancer is the most often diagnosed cancer in 112 countries, and it is also the leading cause of cancer mortality in 48 countries. In Bangladesh, most people are not aware of prostate cancer. Again, sometimes doctors make mistakes in detecting the proper disease and patients do not get proper treatment. That is why many patients die at an immature age. In this regard, we made a model with the help of machine learning techniques that can identify patients' cancer stage and survivability with 96.27% and 89.56% accuracy, respectively. Our model will help doctors identify cancer stages and survivability. We hope this will help to reduce the death rate of prostate cancer patients.

5.2 Empact on Environment

In the current situation of the world, we can see that when a patient gets affected by some disease, they should suffer more to detect their disease and get treatments. They need to go to the doctor's chambers and wait a long time to meet the doctor. This kills the patient's valuable time. Sometimes it causes patients' deaths, as patients don't get proper treatment in time. But if people use this kind of AI-based system to identify their disease, they can easily recover. Again, it helps doctors identify patients' diseases in a short time, and patients can get proper suggestions and treatment in time. Again, this kind of system reduces the cost of diagnosis. Every day, the pathology laboratory produces huge amounts of waste. This technology can reduce waste produced by pathology laboratories. It saves patients' time. In our country, we can see that the environment of government hospitals and doctors' chambers is overcrowded. In this crowded situation, patients feel illness. That is harmful for patients. Sometimes patients feel sicker because of the dirty environment of government hospitals. If people use this kind of machine learning based system, they can test their disease at home through the internet.

5.3 Ethical Aspects

When we do research, we should be honest. Some researchers used fake datasets or they manipulated data to get better results in their research. It is totally against ethics. Because in medical science, this kind of research can destroy patients' lives. Patients can die from faulty treatment. For our study, we gathered data from the SEER Institute. SEER is a trustworthy source of cancer statistics in the United States. The SEER (Surveillance, End Results, and Epidemiology) Program gathers cancer statistics in order to reduce the country's cancer burden. The SRP in NCI's DCCPS supports SEER. We used a dataset of 10946 patients and 187798 patients for predicting survivability and cancer stage, respectively. We used a huge amount of data to build our models so that they could predict more accurately.

5.4 Sustainability Plan

When we want to do something, we have to think about the sustainability of that work. In our research, we tried to build models using machine learning approaches that can identify prostate cancer patients' cancer stages and survivability. This kind of research can have a significant impact on human beings. That's why we thought about the sustainability plan for our research. The symptoms of all diseases have been changing day by day. This is also applicable to prostate cancer. That's why we have a plan regarding this problem. In the near future, we will use reinforcement learning in our model. That will collect new types of data and train our model, and it will deploy new models continuously. In this process, there will be no problem if the symptoms of prostate cancer change.

CHAPTER 6

Conclusions and Future work

6.1 Conclusions

In this work, we seek to assess the possibility of a prostate cancer patient's survival, five-year life expectancy, and the stage of a prostate cancer patients with computational intelligence. Several strategies are examined based on characteristics with distinct effects. Our XGBoost classifier-based prediction model is proposed to predict the survival of prostate cancer patients by analyzing how fast algorithms can successfully predict. We have shown the suitability of our model compared to others. Our system would play a revolutionary role in the digitalization of medical diagnosis for prostate cancer. Our customized Gradient Boosting prediction model is proposed to estimate the five-year survival of patients with prostate cancer. This model is the top performer in terms of prediction speed, accuracy, AUC score, and sensitivity. It is also demonstrated that our model is superior to others in terms of performance. We are proposing our LGBM classifier to predict the stage of a prostate cancer patient by reviewing how quickly algorithms can be accurately identified. No one has ever developed this type of intelligence to detect the stage of prostate cancer. So, this model of ours will play a groundbreaking role in the work of automation in the medical sector. With the aid of artificial intelligence, physicians can predict the patient's likelihood of survival, five-year life expectancy, and stage, enabling them to develop a more effective treatment plan

6.2 Future Work

There are none placed on any work that may be done in the future. There are a great number of works that deal with this topic. We endeavored to make predictions and locate the most effective algorithms that may produce the best results. At this point, we are working on a classification problem. In the not-too-distant future, one of our goals is to develop a prediction model that is more comprehensive in order to attain a better level of accuracy.

Reference

[1]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 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2018, pp. 112–116.

[2]D. Delen and N. Patil, "Knowledge extraction from prostate cancer data," in Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), 2006, vol. 5, pp. 92b–92b.

[3]C. M. Lynch and J. D. BehnazAbdollahi, "Fuqua, Alexandra R. de Carlo, James A. Bartholomai, Rayeanne N. Balgemann, Victor H. van Berkel, Hermann B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques," International Journal of Medical Informatics, vol. 108, pp. 1–8, 2017.

[4]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 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE), 2015, pp. 1–6.

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

[6]A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," Age, vol. 58, no. 13, pp. 10–110, 2006.

[7]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)," International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association, vol. 13, no. 2, pp. 11–16, 2008.

[8]D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial intelligence in medicine, vol. 34, no. 2, pp. 113–127, 2005.

[9]M. Mourad et al., "Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis," Scientific Reports, vol. 10, no. 1, pp. 1–11, 2020.

[10]K. Pradeep and N. Naveen, "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and Naive Bayes algorithms for healthcare analytics," Procedia computer science, vol. 132, pp. 412–420, 2018.

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

[12]M. Lundin, J. Lundin, H. 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.

[13]J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Breast cancer survivability via AdaBoost algorithms," in Proceedings of the second Australasian workshop on Health data and knowledge management-Volume 80, 2008, pp. 55–64.

[14]M. S. I. Polash, S. Hossen, R. K. R. Sarker, M. A. Bhuiyan, and A. Taher, "Functionality Testing of Machine Learning Algorithms to Anticipate Life Expectancy of Stomach Cancer Patients," in 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), 2022, pp. 1–6.

[15]A. A. Abbasi et al., "Detecting prostate cancer using deep learning convolution neural network with transfer learning approach," Cognitive Neurodynamics, vol. 14, no. 4, pp. 523–533, 2020.

[16]L. Hussain et al., "Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies," Cancer Biomarkers, vol. 21, no. 2, pp. 393–413, 2018.

[17]P. Gupta et al., "Prediction of colon cancer stages and survival period with machine learning approach," Cancers, vol. 11, no. 12, p. 2007, 2019.

[18]H. Barlow, S. Mao, and M. Khushi, “Predicting high-risk prostate cancer using machine learning methods,” Data, vol. 4, no. 3, p. 129, 2019.

[19]G. Wang, J. Y.-C. Teoh, and K.-S. Choi, “Diagnosis of prostate cancer in a Chinese population by using machine learning methods,” in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 1–4.

[20]O. Hamzeh, A. Alkhateeb, J. Zheng, S. Kandalam, and L. Rueda, “Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data,” BMC bioinformatics, vol. 21, no. 2, pp. 1–10, 2020.

[21]E. Erdem and F. Bozkurt, “A comparison of various supervised machine learning techniques for prostate cancer prediction,” Avrupa Bilim ve Teknoloji Dergisi, no. 21, pp. 610–620, 2021.

[22]O. Regnier-Coudert, J. McCall, R. Lothian, T. Lam, S. McClinton, and J. N’Dow, “Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers,” Artificial intelligence in medicine, vol. 55, no. 1, pp. 25–35, 2012.

[23]M. Alam, M. Tahernezhadi, H. K. Vege, P. Rajesh, and others, “A machine learning classification technique for predicting prostate cancer,” in 2020 IEEE International Conference on Electro Information Technology (EIT), 2020, pp. 228–232.

[24]J. K. Kim et al., “A performance comparison on the machine learning classifiers in predictive pathology staging of prostate cancer,” in MEDINFO 2017: Precision Healthcare through Informatics, IOS Press, 2017, pp. 1273–1273.

[25]M. Doja, I. Kaur, and T. Ahmad, “Age-specific survival in prostate cancer using machine learning,” Data Technologies and Applications, 2020.

[26]S. Jović, M. Miljković, M. Ivanović, M. Šaranović, and M. Arsić, “Prostate cancer probability prediction by machine learning technique,” Cancer investigation, vol. 35, no. 10, pp. 647–651, 2017.

[27]R. Lenain, M. G. Seneviratne, S. Bozkurt, D. W. Blayney, J. D. Brooks, and T. Hernandez-Boussard, “Machine learning approaches for extracting stage from pathology reports in prostate cancer,” Studies in health technology and informatics, vol. 264, p. 1522, 2019.

[28]B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck, and I. Bratko, “Machine learning for survival analysis: a case study on recurrence of prostate cancer,” in Joint European conference on artificial intelligence in medicine and medical decision making, 1999, pp. 346–355.

[29]F. D. Beacher, L. R. Mujica-Parodi, S. Gupta, and L. A. Ancora, “Machine Learning Predicts Outcomes of Phase III Clinical Trials for Prostate Cancer,” Algorithms, vol. 14, no. 5, p. 147, 2021.