

Optimal Cancer Staging and Survival Analysis/Prognosis Using Machine Learning



By

Humna Mansoor

00000277623

Supervisor

Dr. Rafia Mumtaz

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters
of Science in Information Technology (MS IT)

In

School of Electrical Engineering & Computer Science (SEECS) ,

National University of Sciences and Technology (NUST),

Islamabad, Pakistan.

(April 2022)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Optimal cancer staging and survival analysis/prognosis using machine learning" written by HUMNA MANSOOR, (Registration No 00000277623), of SEecs has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____  _____

Name of Advisor: Dr. Rafia Mumtaz

Date: 14-Jun-2022

HoD/Associate Dean: _____

Date: _____

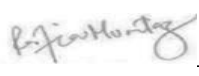
Signature (Dean/Principal): _____

Date: _____

Approval

It is certified that the contents and form of the thesis entitled "Optimal cancer staging and survival analysis/prognosis using machine learning" submitted by HUMNA MANSOOR have been found satisfactory for the requirement of the degree

Advisor: Dr. Rafia Mumtaz

Signature:  _____

Date: 14-Jun-2022

Committee Member 1: Dr. Farhan Khan

Signature:  _____


Date: 14-Jun-2022

Committee Member 2: Dr. Muhammad Moazam
Fraz

Signature:  _____

Date: 14-Jun-2022

Committee Member 3: Dr. Ahsan Saadat

Signature:  _____

Date: 14-Jun-2022

Dedication

This thesis is dedicated to my amazing parents Mr and Mrs Shahid Mansoor, whose support throughout my journey made me accomplish it. To my inspiring and guiding siblings, Huda, Abid and Rafia, who always provided me with strength, moral, emotional and financial support.

Thank You my dear family for being there when I was uncertain of myself for doing Masters.

As Harry says,


“ Working hard is important but there is something that matters even more; believing in yourself. ”

– Harry Potter

Certificate of Originality

I hereby declare that this submission titled "Optimal cancer staging and survival analysis/prognosis using machine learning" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEecs or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEecs or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: HUMNA MANSOOR

Student Signature: 

Acknowledgments

Glory be to Allah (S.W.A), the Creator, the Sustainer of the Universe. Who only has the power to honour whom He please, and to abase whom He please. Verily I can not do anything without His will.

I would like to express my gratitude to my Mentor, my Research Supervisor Dr. Rafia Mumtaz, whose guidance, immense support, mentor-ship, patience and confidence in me, have always encouraged me throughout the process. Her constant encouragement, freedom of research and understanding have helped me to learn, develop, enjoy, and make it a positive experience. I would like to extend my thanks to my committee members for being part of my thesis and providing me with necessary feedback.

I would like to thank my mother and sister for pushing me and motivating me throughout the process. A special thanks to my friends like family Aiman, Farwa and Zahra, who were the light in dark time. A special bow for their years of invaluable companionship, love, support and for making my journey in NUST wonderful. Without them this would not have been possible, " *For Them, a Thousand Times Over!* "

Humna Mansoor

Contents

1	Introduction and Motivation	1
1.1	Cancer	1
1.2	Role of Machine Learning	3
1.3	Problem Statement and Challenges	4
1.4	Research Statement and Objective	5
1.5	Contributions	6
1.6	Thesis Organization	7
2	Literature Review	8
2.1	Thyroid	8
2.2	Thyroid Cancer	9
2.2.1	Types of Thyroid Cancer	10
2.2.2	Mortality Rate	11
2.3	Machine Learning and its Evolution	12
2.4	Machine Learning Applications and Cancer	12
2.5	Comparative Analysis	17
3	Methodology	20
3.1	Proposed Methodology	20
3.1.1	System Architecture	21
3.2	Data Collection	21

CONTENTS

3.3	Data Description	24
3.4	Feature Selection	25
3.4.1	Renaming Attributes	25
3.5	Data Pre-processing and Encoding	25
3.5.1	Labelling and Encoding	26
3.5.2	Data Splitting	27
3.6	Class Balancing and Dimension Reduction Techniques	27
3.6.1	Principal Component Analysis (PCA)	27
3.6.2	Over-Sampling and Under-Sampling	29
3.7	K-fold cross Validation	29
3.8	Machine Learning Algorithms	30
3.8.1	Gaussian Naive Bayes	30
3.8.2	Decision Tree	31
3.8.3	Support Vector Machine	31
3.8.4	Random Forest	31
3.9	AdaBoost	32
3.9.1	K-Nearest Neighbours	32
3.9.2	K-Mode	32
3.9.3	Light Gradient Boosting	33
3.10	Hyper-parameter	34
3.11	Survival Analysis	34
3.11.1	Survival Function	36
3.11.2	Kaplan Meier Method	36
3.12	Evaluation Metric	37
3.13	Summary	39
4	Implementation and Results	41
4.1	Data Analysis	41

CONTENTS

4.2	Experiment Analysis	44
4.3	Results	45
4.3.1	Experimental Result with Decision Tree	45
4.3.2	Experimental Result with Gaussian Naive Bayes	46
4.3.3	Experimental Result with Support Vector Machine	46
4.3.4	Experimental Result with Random Forest	48
4.3.5	Experimental Result with AdaBoost	50
4.3.6	Experimental Result with K-Mode Clustering	51
4.3.7	Experimental Result with KNN	51
4.3.8	Experimental Result with Light Gradient Boosting Machine	53
4.3.9	Stage Predictions	54
4.3.10	Comparison of Machine Learning Classifiers	56
4.3.11	Survival Analysis	57
4.4	Summary	62
5	Future Work and Conclusion	63
5.1	Conclusion	63
5.2	Future Work	65

List of Figures

1.1	Cancer Deaths by Type around the World	2
1.2	Cancer Cases and Mortality by Type in Pakistan	3
2.1	A Tumorous Thyroid Gland.	9
2.2	A Brief History and Evolution of Machine Learning [24]	12
2.3	Architecture for the Proposed System in [28]	14
2.4	Proposed Architecture in [31]	16
3.1	System Flow Diagram	21
3.2	System Architecture Diagram	22
3.3	Renamed Data Columns	25
3.4	Branches for Machine Learning	30
3.5	Confusion Matrix Example for Multi-class Classification	38
3.6	Confusion Matrix for Multi-class Classification	38
4.1	Thyroid Cancer cases over years	41
4.2	Thyroid Dataset Correlation Heat-map	42
4.3	Thyroid Cancer Vital Status (Dead or Alive)	43
4.4	Patient count w.r.t Stages	43
4.5	Parameters passed for Decision Tree	45
4.6	Classification-Report Decision Tree	45
4.7	Classification-Report Naive Bayes	46

LIST OF FIGURES

4.8	Classification-Report SVM (linear)	47
4.9	Classification-Report SVM (poly)	47
4.10	Classification-Report SVM (rbf)	48
4.11	Classification-Report Random Forest using PCA	49
4.12	Classification-Report Random Forest using SMOTE	49
4.13	Parameters used with AdaBoost	50
4.14	Classification-Report AdaBoost	50
4.15	Optimal K values for Clusters	51
4.16	Accuracy plot for different values of K-NN	52
4.17	Classification-Report K-Nearest Neighbor	52
4.18	Hyper-parameters for Light Gradient Boosting	53
4.19	Classification-Report Light Gradient Boosting	53
4.20	Kaplan Meier Estimate for 5-year on Thyroid Cancer Data	57
4.21	Kaplan Meier curve for Thyroid Cancer Stage I and II	58
4.22	Kaplan Meier curve for Thyroid Cancer Stage II and III	58
4.23	Kaplan Meier curve for Thyroid Cancer Stage III and IVA	59
4.24	Kaplan Meier curve for Thyroid Cancer Stage IVA and IVB	60
4.25	Kaplan Meier curves for Thyroid Cancer Stages	61
4.26	Kaplan Meier curve for Thyroid Cancer With Respect to Age	62

List of Tables

2.1	A Comparative Discussion on Literature Work	18
3.1	Unprocessed Raw Sample Data from SEER Database	23
3.2	Variables and Definitions regarding the differentiated thyroid Cancer . .	24
3.3	Thyroid Cancer Stages Against T,N,M Values	28
3.4	Hyper-Parameters for the Machine Learning Algorithms	35
4.1	LightGBM and AJCC Grouping on thyroid cancer dataset generated from the SEER Database of thyroid cancer (papillary and follicular)	55
4.2	Results of different Models on Thyroid Dataset for Stage Prediction . .	56

Abstract

Early predictions and survivability analysis can often be a key to better treatment and accurate prognosis of Cancer. Changes in staging model are a requirement to understand the tumor behavior and its possible clinical outcomes. Different models of Machine learning are widely used in order to increase prognostic accuracy.

In this research, for prognosis and Stage prediction of thyroid cancer, the data was gathered from Cancer repository. National Cancer Institute has launched a program which holds a number of registries on almost every type of cancer. This disease specific dataset was fetched from the program's database known as Surveillance, Epidemiology, and End Results (SEER). The derived data model is similar to the American Joint Committee on Cancer (AJCC).

The data is pre-processed to achieve good outputs. After cleaning and encoding of data, the machine learning models are implemented. Models are tuned on hyper-parameters and trained using the training data. To enhance the overall performance of cancer stage prediction, class balancing strategies such as oversampling, undersampling, normalization techniques and principle component analysis were added into the models.

To achieve improved results and better understanding we used different machine learning classification models. The experimentation showed that the Gradient Boosting Machine Learning technique implemented on the data combination of Tumor, Nodes, Metastasis and Age (TNMA), generates best predictions for Stages. The evaluation measures used to compare the performance of the machine learning models showed that Light Gradient Boosting gave an accuracy of 91% while AdaBoost gave an accuracy of 88.5% but this value was enhanced to 96% when Decision Tree was used as the base for the Adaboost classifier. The results showed that adding class balancing approaches enhanced the models' performance greatly as well. The predicted stages are closely related to the standard for cancer staging.

The survival probability for the Thyroid Cancer Stages showed that patients in earlier stages can survive longer than the patients in the higher stages. The number of patients reaching the final stages is found to be low. The demonstrated approaches in the thesis can aid in the patient's treatment decision making and can be utilized in making prognostic systems.

Introduction and Motivation

The healthcare industry has been utilising IT for a few decades to store patient visit records, costs, insurance information, and more. Data warehousing and knowledge management strategies can help decision support systems in healthcare since healthcare data is enormous. Data gathering and storage has vastly improved, not just in terms of quantity but also in terms of quality. Analyzing such a large volume of data would necessitate the use of specialist technologies, as manual data analysis would be time-consuming. Medical informatics addresses these issues by employing statistical pattern recognition, machine learning, and visualisation techniques to aid in the interpretation of data [1]. The sections below explain the impact of cancer and machine learning role in different fields. Pinpointing the research problem in this domain and the major goals to achieve through this study.

1.1 Cancer

Cancer is a category of disorders characterised by the uncontrolled division of faulty cells that develop and spread uncontrollably throughout the body. Normal cells in the human body develop, grow, and divide as they ought to. When cells become old or get damaged, they are replaced, but if they multiply and increase uncontrolled, they are classified as tumor. These tumors can spread throughout the body and interact with other organs. It may spread to the major body organs and cause damage such as to the neurological system, digestive system, or circulatory system. The release of hormones by such diseased regions of the body produces changes in the body. Cancer is a term

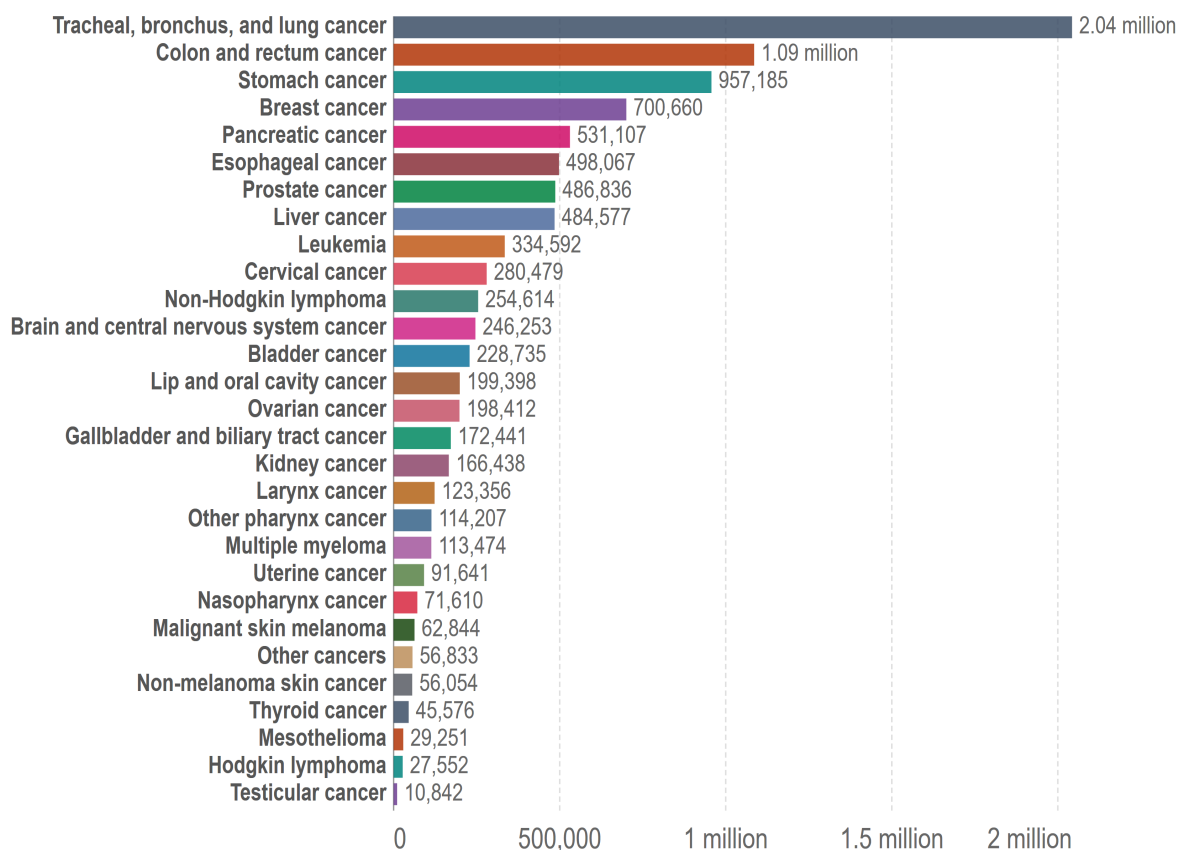
that encompasses more than 100 different diseases [2].

Cancer is one of the world’s most serious issues, claiming the lives of millions of people as reported by the Institute of Health Metrics and Evaluation (IHME).[3] Cancer is the second largest cause of mortality in the globe. It has always been considered to be one of the most heterogeneous disease, which keeps on growing and consists of multiple sub-types. The statistics of total yearly cancer deaths across all ages and both sexes are shown in figure 1.1 from IHME’s Global Burden of Disease (GBD).

Cancer deaths by type, World, 2019



Total annual number of deaths from cancers across all ages and both sexes, broken down by cancer type.



Source: IHME, Global Burden of Disease (GBD)

CC BY

Figure 1.1: Cancer Deaths by Type around the World

One of the most life threatening and recurring disease to be known till date. The early diagnosis and prognosis of cancer and its type have become a necessity in cancer domain research, as it can facilitate the subsequent clinical management of patients. Early stage prediction and accurate treatment selection take up days and such delay cause health

and sometimes life threats. Its global burden is increasing year by year, and varies from country to country. For example in China in 2018, an estimated 4.3 million new cancer diagnoses and 2.9 million new cancer deaths. In comparison to the United States and the United Kingdom, China has a lower incidence of cancer but a 30 percent and 40 percent higher cancer mortality, respectively, with 36.4 percent of cancer-related fatalities [4]. Similarly in Pakistan, each year the number grows irrespective of the type of cancer. Figure shows the estimated mortality rate and number of cancer cases and for different cancer types in Pakistan for the year 2020. These estimations are made by GLOBOCAN [5].

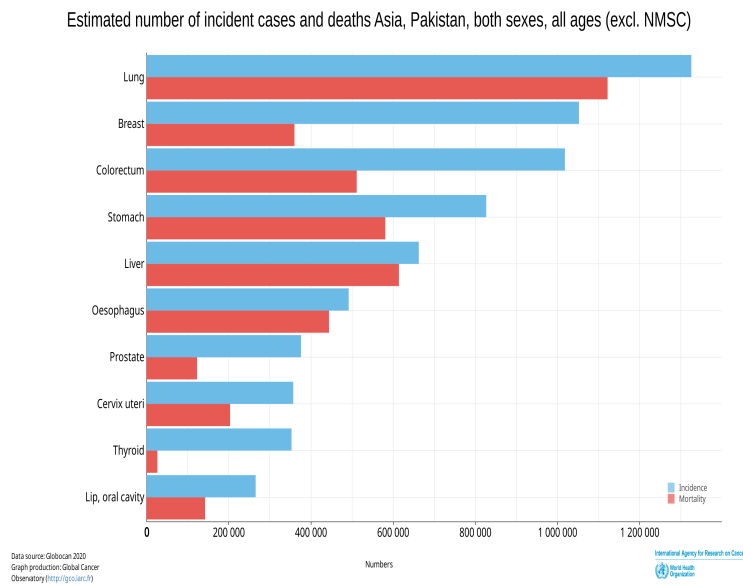


Figure 1.2: Cancer Cases and Mortality by Type in Pakistan

1.2 Role of Machine Learning

The objective of machine learning is to learn from provided data, examples and experiences. It learns from the information that is feed into it and produces an output which is set as the goal. Instead of designing complicated algorithms to solve specific issues, computers may make logical predictions based on the facts they are given. It is necessary to provide both training and test data to an algorithm in order for it to generate such predictions. A machine learning model is then trained using the algorithm [6]. Machine learning is now widely employed for a variety of reasons, and today there are

several real-world applications. The techniques are utilised in image recognition, voice recognition, traffic pattern analysis, for robotics, weather prediction, stock analysis, and many other applications.

Machine learning has also grown in healthcare domain. It offers a number of approaches, strategies, and tools for diagnosing and predicting issues in a variety of medical fields. It is quite obvious that the Artificial intelligence can be very helpful in the field of medicine [7].

The use of Machine Learning (ML) in medicine is getting more importance as researchers have turned towards the ML applications. In the last decades, large scale work has been done in machine learning for cancer, ranging from diagnosis and classification of tumor or a proceeding predictive analysis. In cancer prognosis there are three different focus points; cancer recurrence, cancer risk assessments and cancer survivability. Although cancer prognosis only comes into consideration, once the diagnosis is completed but can not be left or relied simply upon the diagnosis. Applying Machine Learning Techniques can boost the survival probability prediction and prognostic accuracy [8].

1.3 Problem Statement and Challenges

Cancer is one of the hardest diseases to detect and survive. Thyroid cancer is becoming more common, and it is expected to overtake lung cancer as the fourth most common cancer worldwide. Despite the fact that thyroid illness is spreading widely, the diagnosis technique in clinics has not changed during the twentieth century. Unfortunately, medical imaging cannot fully support differentiating between benign and malignant thyroid nodules[9].

As the Covid-19 outbreak happened, its impact on the Thyroid glands were taken into consideration. The adverse effect of Covid-19 over the thyroid glands in some cases may fall into severe cases like cancer [10]. Hence, knowing the survival situation of a patient has gained much importance than it had before.

There exists a direct correlation between mortality rate and early prediction of cancer. With this rapid growth of disease, the clinical methods of prediction and prognosis are not enough, there is a need to find new methods as each individual is different from another [11]. Using machine learning techniques with such clinical and different data is difficult, as the data is very sophisticated in terms of any slight change may cause a

complete different output where a patients life depends on it. Although medical data is present in abundance, but it frequently contains huge amounts of missing, incomplete, biased, skewed, or incorrectly categorised data. Dealing with these prior to the Machine learning model training is crucial, else the model may produce bias results. This may disrupt the health care merger with Technology sector.

In order to address these points of concern, the dataset for Thyroid is gathered using an authorized association and cleaned considering the standards of Cancer association. Different Machine Learning techniques are explored to have the best possible results. Techniques are applied to boost the Stage prediction for the patients already diagnosed and also to determine the possible survival over the years, aiding the prognosis.

1.4 Research Statement and Objective

Thyroid cancer prognosis and medications or therapy treatments are only determined, if the kind of tumor and its stage is correctly known at the time of diagnosis. Using machine learning and traditional methods, we can not only predict the disease early but also take a look on the survival probability of an individual that can drive this disease treatment into right direction in early stage.

In this research, we aim to use Machine learning concepts with the cancer data. The idea revolves around using Machine Learning techniques with the cancer data in order to have timely and correct Stage predictions for the cancer type and its analysis. Our main focus is to overcome the delays caused in the prognosis due to time taking process of determining the cancer situation and stage. Also finding the patients 5-year survival as it helps in understanding of the patients condition. Training the machine learning model with such large historical data can help in producing best suitable outcome.

The thesis' main purpose is to conduct predictive analysis using various models and classifying the provided datasets into particular cancer stage. Furthermore, a vast number of research have looked into thyroid cancer early detection but quite a few for its early stage predictions and prognosis. The main aims of the study are:

- Using machine-learning algorithms, providing a new way for predicting the stages and survival of thyroid cancer patients.
- To tackle this problem, employ descriptive profiling, which involves identifying

trends and patterns in data.

- Predictive modelling or supervised learning techniques are used. In this approach two set of prominent variables, input and target are used. Inputs taking all predictors, features, and explanatory factors from data. While the target which are outcome and dependent variables. This study will use different classification models to achieve the target.
- In addition, evaluating the performances for all machine learning models.

Currently it is found that the main focus is on cancer disease detection using machine learning, which itself face high computing and expense, as well as low accuracy. We wish to create a system that can predict stages with high accuracy and non-invasive characteristics. The findings are likely to aid in determining optimal techniques for avoiding delays and aiding in treating cancer. It will, in particular, make it easier to choose relevant models and methodologies for analysing not only thyroid cancer but can be utilized for the other types of cancer too for future studies.

1.5 Contributions

The research work in the field of medicine with machine learning and artificial techniques is still under process. It faces challenges and lack of acceptance. In the presented research work, the main focus is on attaining efficient results with the clinical dataset.

The practical contributions by this research are :

- The merge of the medical dataset with the Machine Learning techniques.
- The usage of Surveillance, Epidemiology, and End Results (SEER) for obtaining cancer dataset for prediction.
- Creating a combination of features i.e. Tumor, Nodes, Metastasis, Age as TNMA
- Utilizing the idea of Machine learning algorithms to construct and evaluate a cancer specific predictions.
- The robustness in predicting stages.
- Avoiding the prognostic delays caused by the time-consuming procedure of evaluating the cancer situation and than determining the stage for the cancer.

- Analyzing the survival of the patients for the stage
- Opening new paths for Cancer prognosis with ML/DL techniques

1.6 Thesis Organization

This thesis is divided into five chapters. The first chapter Introduction includes the details for cancer and machine learning, research problem statement with the limitations of the study, research aims and objectives, and lastly the thesis structure.

The second chapter contains descriptive information on thyroid cancer and machine learning. It includes a review and critical analysis of chosen literature work that is important and relevant to the research subject, which will aid in determining the present status of cancer research in field of machine learning.

The third chapter of Methodology, highlights the major research work conducted. From Data collection to its pre-processing, providing the details on feature selection and data normalization. Machine learning models used in the study including the Survival functions, the analysis techniques used and the evaluation matrices, all are explained in the chapter.

Fourth chapter gives the findings for applied machine learning models on the data. The comparative metrics and the survival analysis with the graphical representation, these are all defined and in this chapter.

The fifth Chapter wraps up this thesis with a brief review of the findings and suggestions for further research.

CHAPTER 2

Literature Review

A vast number of research is being conducted to look into the identification and prognosis of cancer. Several studies have found that signs of cancer may not appear until the latter stages of the disease is reached. The technological growth has its mark on every field including the healthcare. Using Machine Learning and Artificial Intelligence ideas in the domain of cancer is a milestone to achieve. To obtain a better knowledge of the subject, which is about machine learning, thyroid cancer, and ethical growth of machine learning techniques in healthcare, this chapter presents the project's key theory.

2.1 Thyroid

Human body consists of multiple types of glands, each gland associated with some different kind of functionality and purpose. Thyroid is one of the vital gland that is located at the front base of the neck under the voice box [12]. Thyroid gland is responsible for regulating metabolism, growth, and a variety of other body activities. Out of all, the main function of thyroid glands in the body is to generate different types of hormones. That is why known as an important hormone gland, playing a major rule in human body growth and development.

This butterfly shaped gland as shown in the figure 2.1 plays a crucial role in maintaining body temperatures, heart rate, body weight and hormonal balance. Thyroid hormones are continually released into the circulation, which serves to control various physiological processes [13]. The thyroid gland generates extra hormones when the body need more energy in particular conditions, such as when it is growing, cold, or pregnant.

When any abnormality appears in thyroid gland [14], it disturbs the whole body functionality since it plays a vital role in body maintenance. There are multiple diseases associated with thyroid out of them cancer in thyroid glands is the worst. Thyroid cancer generally manifests as a single thyroid nodule or a growing goitre. It was considered that thyroid nodules are prevalent, but thyroid cancer is uncommon. But since 2001, it is seen the incidences has grown around the globe. The Figure shows the tumors on the glands, if stayed untreated and undetected may turn into metastasis.

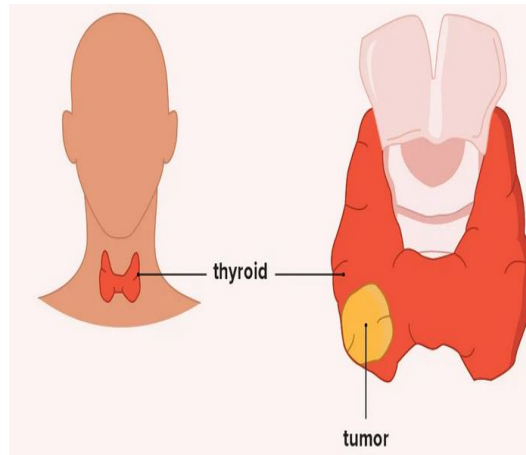


Figure 2.1: A Tumorous Thyroid Gland.

2.2 Thyroid Cancer

The most commonly occurring malignancy in the endocrine system is Thyroid cancer. In Thyroid Cancer [15], the glands out grow there original size. The glands grow out of control and enlarge to the size which needs to be treated surgically or via medications. Not only the size but also disabling it's functionality. In 2015 [16], it was found that there has been a widespread of thyroid cancer over past 3 decades. Since the increase, this organ has taken up a lot of attention. By the year 2018, it was surveyed that around 300 million people are suffering from thyroid and keeps growing year by year[17]. There are multiple reasons that thyroid cancer happens, out of which the deficiency for iodine is considered the prominent one. Many countries face such issues and are unable to fulfil the iodine requirements through edible ways. Many complications rise due to this endocrine disease, it is a severe health issue among those who suffer from iodine deficiency. Although the research today shows that not just this but other contributing

factors add up too in thyroid cancer [18].

It was seen as the new technology emerged the detection and diagnosis has made it possible, earlier considered as the nodules in thyroid. Malignant nodule diagnosis has mostly relied on surgeons' and radiologists' clinical expertise. But Human judgement is time-consuming and error-prone in many circumstances. Delaying the immediate medications and therapies [19].

2.2.1 Types of Thyroid Cancer

Thyroid cancer [15] is of different types, each having different conditions and impact rates. The types of thyroid cancer:

- Differentiated cancer (papillary, follicular and Hürthle cell)
- Medullary cancer
- Anaplastic cancer

There are total 5 kinds of Thyroid Cancer out of them 4 are main and commonly found. Most of the Cancers are the differentiated cancers, when looked upon in labs they seem to be quite similar to the normal tissues of thyroid.

Papillary thyroid cancer : This kind of thyroid cancer grows slowly and usually only affects one lobe of the thyroid gland. It is one of the most common type of cancer. Papillary malignancies frequently spread to nearby lymph in the neck, despite their slow growth. Even when these tumours have progressed to the lymph nodes, they are usually treatable and seldom deadly.

Follicular cancer : It is second most common type of thyroid cancer. It is largely found in places where individuals do not consume enough iodine. Although these malignancies do not usually move to lymph nodes, they can expand to other regions of the body, such as the lungs or bones. Follicular cancer has a less favourable prognosis than papillary cancer, while it is still suitable and favourable in the majority of instances.

Hürthle cell cancer : It is also known as oxyphil cell carcinoma. This form of thyroid cancer accounts for around 3% of all thyroid cancers. It is more difficult to locate and even cure.

Medullary cancer : Thyroid cancer of this sort is extremely rare. It begins in the C-cells, a kind of thyroid cell. C-cells produce a hormone that helps regulate calcium

levels in the blood hence destroying the whole body system. Even before the diagnosis or the thyroid nodule is found, this cancer can progress to lymph nodes, the lungs, or the liver. It is hard to discover and treat this type of cancer.

There are further 2 types of this cancer. One of them is inherited further in family and happens in young age and have higher possibility of having multiple other tumours in body. While the other type is not inherited and happens only in adulthood.

Anaplastic cancer : Thyroid cancer of this sort is extremely rare. It is difficult to cure since it spreads swiftly into the neck and other regions of the body. The cancer cells in it, do not resemble normal thyroid cells [20].

2.2.2 Mortality Rate

Incidence of thyroid cancer is rapidly increasing worldwide. It is on constant rise, according to the National Institutes of Health's Surveillance, Epidemiology, and End Results Program (SEER) it is growing annually by 5.5%. Between 2014 and 2018, the yearly incidence rate in the United States was approximately 14.1 per 100,000. Cancer is spreading in both genders equally, but according to the Cancer Registry it is the second most prevalent cancer among women. Despite thyroid cancer accounting for just 3.5 percent of all malignancies. The incidence-based mortality rate is rising, owing to changes in thyroid cancer biology. This might be related to any unknown cause, such as environmental or etiologic factors [21].

Mortality rates [22] based on incidence continue to rise. These data show that changes in thyroid nodule treatment may have resulted in a decrease in the identification of tiny indolent tumours, but not of more advanced tumours.

In every 5 year duration, the yearly death rate was 0.5 per 100,000 and has now rose by 0.8% annually. Thyroid cancer is expected to cause 43,800 new cases and 2,230 deaths in 2022, according to the American Cancer Society [23]. Thyroid cancer still has a 1.1 percent lifetime risk, which in comparison to other cancers, but its increase is alarming and the 5-year survival rate is approximately 97.8 percent, which is going down every year.

It is quite evident from the data that the overall death rate is not that high but as the spread grows, diagnosis delays and wrong prognosis, the possibilities for survival gradually decreases.

2.3 Machine Learning and its Evolution

The figure 2.2 below shows a brief history and the evolution of machine learning over a period of time. From just trying to solve basic problems to a complete new era of solving real world things. Machine learning has revolutionised the technological world.[24]

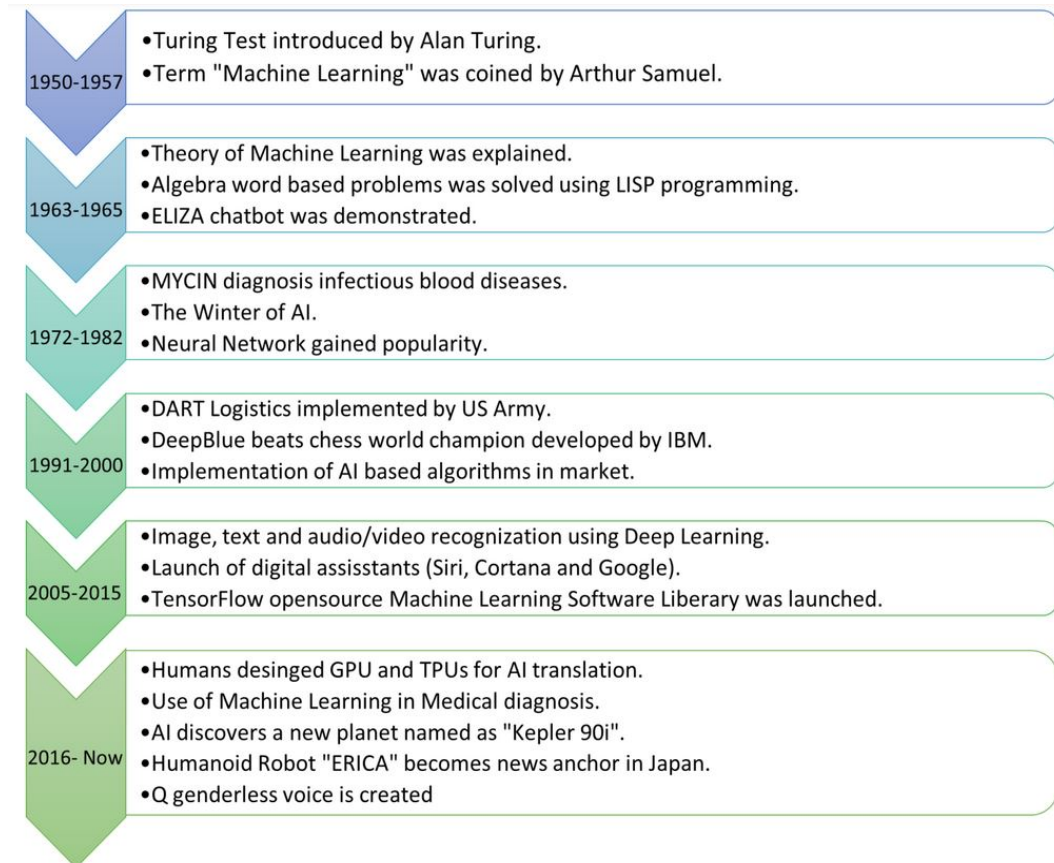


Figure 2.2: A Brief History and Evolution of Machine Learning [24]

Future trends of Machine Learning : Quantum Machine Learning (QML) is a new theoretical topic that studies and researches on interaction between the quantum computing and machine learning approaches. Several investigations have recently proven that quantum computing offers advantages for machine learning.

2.4 Machine Learning Applications and Cancer

There are several ideas and techniques that till date have been presented and studied involving machine learning with Cancer data. Apart from the diagnosis cancer survival

predictions and prognosis have been of much importance. Since cancer is not just confined to one part of body, there are multiple sub-types, which expands the domain of research and work. Looking at the diversity, the research presented in [25] have hence used the idea of using modern ML Techniques and their applications for cancer. They presented a detailed review of ML methods Bayesian Network, Decision Tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN), their application in cancer related diagnosis and predictions. It was concluded that out of all techniques studied, it was ANN and SVM classifiers which performed better and were so used widely. Every method did have some limits as the data-sets were small and may cause classification errors but can be improved when combined with such heterogeneous and huge data, with feature selection and ML modelling can make big difference in cancer prognosis and analysis. The authors suggested for new techniques and methods to be taken into consideration for cancer predictions as it can prove to be a promising tool for cancer domain.

Leili.et al emphasised on the timely identification of Cancer and its importance for improving the survival rate and lowering down the mortality among cancer patients. 50% of the patients, despite having early diagnosis still develop metastases in later stages of their follow-ups. It has been proven that timely identification of the tumors can increase the survival rate up to 86%. The research demonstrated in [26], have created a ML model using multiple classifiers including Naïve Bayes (NB), Random Forest (RF), SVM, Least Square SVM, Adabag, Logistic Regression (LR) and Linear Discriminant Analysis (LDA) for the prediction of Breast Cancer survival. They have measured the performance of these models against sensitivity, specificity, likelihood ratio and accuracy. Their dataset was based on 550 breast cancer patients and their results indicate that SVM performs best which is closely followed by LDA.

Here in [27], the research work on the prediction of tumor stage and survival in the colon cancer is presented. They used the idea of TNM staging that is T for tumor, N for Node, M for metastasis and determined the tumor stage and 5 year survival of the patient. By applying Machine Learning algorithms, the results were generated, considering the tumor aggression score (TAS) as a prognostic factor. This new attribute TAS was introduced to check the authentication of predicted stage, since in colon cancer it's the tumor size and tumor grade which is important to determine. The highest accuracy was

achieved by algorithms in both cases that is with TNM staging with TAS and without TAS. Out of all it was Random Forest which gave an accuracy of 84% with an AUC of 0.82 which is the area under curve for the five year disease free period. It was found that experiments with more datasets is required which must be more diverse.

Ryu et.al [28] have created a prediction model using deep survival neural network machine learning algorithm using the SEER database. They have applied the Risk Estimate Survival Neural Network (RE-SNN) on a dataset comprising of 1088 subjects. 80% data is used for training while 20% data is used for training. For randomization of the raining data, they have used a 5-fold cross validation method. The hyper parameters are fine tuned to increase the model accuracy. The authors have set the hyper parameters to risk value: 10, time window: 2 months and two learning epochs to achieve the optimized results. After being fine-tuned, the Area under the Receiver Operating Characteristic curve (AUC-ROC) gives a value of 0.85. The figure 2.3 below is the representation of their proposed system architecture.

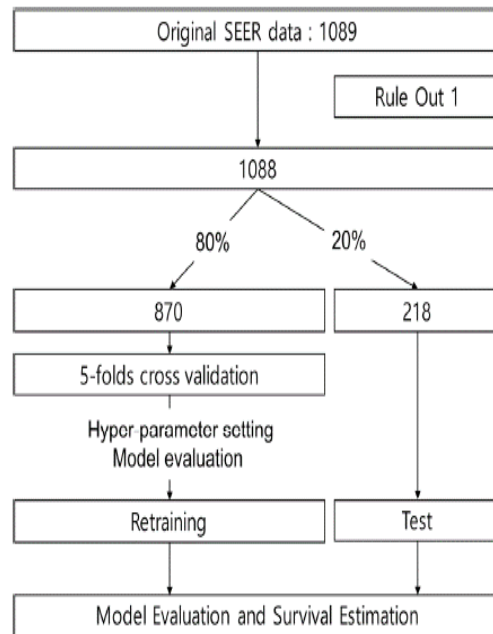


Figure 2.3: Architecture for the Proposed System in [28]

In John W. et.al[29] compiled a detailed contrast of Tree based ML models with the traditional Cox Regression model, to predict survival of an individual. Data was gathered from the seer database specifically for the oral and pharyngeal cancer (OPC). The

time series dataset included a total of 21 thousand patients who were diagnosed with OPC cancer over the period of 5 years from 2004 to 2009. Dataset was pre-processed, removing all the records of patients having any missing value of examined variable while if any eligible patient data still had any missing value, then imputation techniques were performed. Three tree-based ML algorithms (Random Forest, Conditional Inference Forest, Survival Tree) including the cox proportional hazard model (cox) were used together to give a survival prediction model. It was concluded that the traditional cox regression method is best suited for survival prediction but only if the data set is small and is known. But the other tree-based methods deals easily with imputed data and handle large datasets, but for such time series data they may cause biasness. For which more predictors should be added in dataset and other machine learning methods must be explored.

Dealing with the clinical data, specifically cancer data has always been an important topic. Liu et.al[30] found another way to validate and analyze it. The collected Breast Cancer data from Clinical Research Center for Breast (CRCB) of around 12119 patients and created a boosting algorithm based on XGBOOST and Cox proportional hazard model used for survival analysis. They named it as EXSA, which helps in keeping the follow-up and delivers prognosis of the disease. For a 5 year follow up an AUC of 0.83851 was attained. The data set was used with the Random survival forest, simple XGBoost and Cox survival analysis method, the proposed EXSA was relatively better in performance. The model proposed the stages for the cancer data and grouped them as: high risk, mid-high risk, mid-low risk and low risk,. Although, the method used was better but needed to be worked on with larger scale of clinical data from hospitals along with the follow-up data to validate and enhance the model.

Breast cancer is a leading cause of death in women worldwide making it the most common type of cancer in women. In this paper[31], Khourdifi et.al, have created a breast cancer prediction model using different machine learning classifiers including RF, NB, SVM and K Nearest Neighbors (KNN). The main objective of this paper is to create a classification model which can correctly identify the tumor being labelled as malignant or benign. Apart from these classifiers they have also used Weka Data mining tool for prediction. Used the publically available data set of Breast Cancer. The dataset has 699

instance which are measured against 11 different attributes. The performance of used classification techniques is evaluated against multiple measures: Accuracy, Precision, F1 score, Recall and AUC-ROC curve analysis. The authors have also used 10-fold cross validation technique to avoid any case of over fitting. Their results indicate that support vector machine performs best in this case with the by correctly identifying 569 instances out of 699 instances.

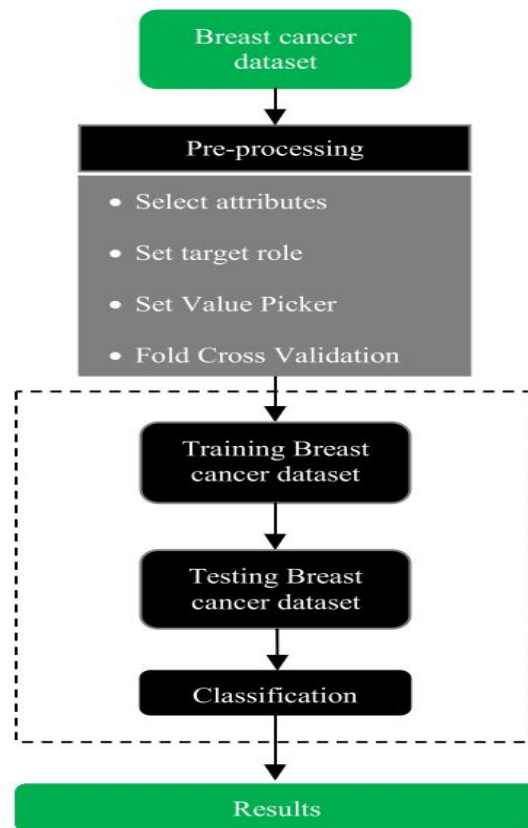


Figure 2.4: Proposed Architecture in [31]

Different survival analysis and prediction methods have been widely utilized out of which hazard proportion model is most popular but the model depends on the ratios between hazards which are constant over time, hence gives a restricted assumption.[32] Since, the treatment's time is never known, usually cancer patients have a long-term treatment period, and in this case the Hazard proportion prediction model lacks. There are several ideas and techniques that till date have been presented and studied involving ML with Cancer data. Apart from the diagnosis cancer survival predictions and prognosis have been of much importance. Since cancer is not just confined to one part of body, there

are multiple sub-types, which expands the domain of research and work. Looking at the diversity, instant changes in the domain are required.

The authors in [33], present here the idea of multiple machine learning techniques merged in to work on thyroid cancer dataset. The data consisted of 3 different types of images obtained from ultrasound, CT scans and X-rays. They aimed to achieve low morbidity and death rates, in order to have thyroid illness diagnosed early. Using the neural network techniques created a model with CNN for early prediction. The Xception neural network model serves as the foundation for the framework that includes three customizable multi-channel architectures. The framework is tested utilising real-world data sets of thyroid cancer. With ultrasound pictures, the proposed architectures outperformed existing approaches, achieving a diagnosis accuracy rate of 0.989.

The research [34], describe the issue that might be faced as the American Joint Committee for Cancer (AJCC) announces the new Staging system. It is seen that they have downgraded a few stages for cancer, such that the predicted disease stage earlier if was of Stage III now might be in Stage II or I. In an effort to prevent over staging and, as a result, over treatment of the illness. The authors discuss the concerning impact it might cause.

Wang et.al in [35], try finding lungs metastasis in thyroid cancer. Gathered the data from SEER and used six ML classifiers that included Random Forest, support vector machine, logistic regression, eXtreme gradient boosting (XGBoost), decision tree and KNN. The authors evaluated the models and it was evident from the evaluation results of all models that the precision results were very low. But out of all random forest model produced most accurate predictive results with 0.99 accuracy and precision at 0.61 with recall 0.88. The imbalance caused the models to gave random predictive outcomes.

2.5 Comparative Analysis

The above discussed research work is put down for a critical analysis to understand the gaps and scope for the research being conducted. From the mentioned literature it is found that majority of the work is done with a small datasets as a trial. Many important

features are missed to avoid the biasness in the work. A large amount of work on the tumor existence, tumor being benign or malignant and diagnosis of cancer is done, but not much of the work is conducted for the prediction of stages. The table below 2.1, gives a comparison of a few literature work discussed above.

Table 2.1: A Comparative Discussion on Literature Work

Literature Work	Data	Techniques	Limitation
Kourou et.al [25]	Breast Cancer	DT, NB, SVM, ANN	Dataset with limited Features
Leili et.al [26]	Breast Cancer	SVM, RF, NB, LR	Small Dataset, Biased outcomes
Gupta et.al [27]	Colon Cancer	TAS attribute with ML classifiers	large computation time, poor accuracy
Ryu [28]	spinal, pelvic chondrosarcoma	RE-SNN, survival analysis	missing factors, testing required
John W. et.al [29]	oral and pharyngeal cancers from SEER	Cox hazard proportion, Tree based models	missing information, biased results
Liu [30]	Breast Cancer from CRCB	XGBOOST, created EXSA	not applied on clinical data,
Khourdifi [31]	publicly available Breast Cancer data	RF, NB, SVM, KNN	small dataset, incorrectly classified instances
Zhang [33]	Image data ultrasound, CT scans and X-rays	Xception NN	poor accuracy with un-tuned images, low performance on CT scans Xrays
Wang et.al [35]	Thyroid Cancer SEER for lungs	RF, SVM, XGBOOST, KNN, LR	Poor precision, skewed model

Despite evaluating the literature on Cancer with Machine Learning in many different perspectives, there have been no particular studies that expressly include Stage prediction and Stage based survival analyses for Thyroid cancer using SEER repositories.

This significant point is addressed in this thesis. These research gaps are covered through the idea presented above. The procedure and methodology is explained in detail in next chapters.

Methodology

3.1 Proposed Methodology

The main idea of the research is to predict the thyroid cancer stage of the patients and their survival over the period of 5 years. The dataset was collected from the SEER database, which is one of the most prominent Cancer repositories. The data contained all the major to minor information regarding the patient including all the factors involved in its survival and all the other factors which may have an impact on the patient's condition.

It was found best to make heat-maps for finding co-relation between different data attributes. Following feature selection, the data is divided in a 70/30 ratio. Keeping 70% of the data for the training purpose and the rest 30% of the data for the testing phase. In addition, a 10-fold cross-validation approach is also used. We employed a variety of assessment measures for the classifiers to evaluate the effectiveness of the model. The proposed system works through following main stages:

- Data gathering
- Data pre-processing
- Feature Selection
- Data splitting and cross validation
- Model fit on the data
- Evaluation

The data pre-processing and feature selection process are most important part as whole system rely on it. It is basically the process of cleaning data and picking important features out of it. Data cleaning involves removal of any noisiness, anomalies in the data, filling in the missing or unknown data, normalizing and encoding of the data. After the pre-processing it is essential to select the main attributes required for the model. Graphical representation of the data points and a detailed analysis of the data attributes can help in selection of important required features more effectively. Figure 3.1 shows the flow of the system that is proposed.

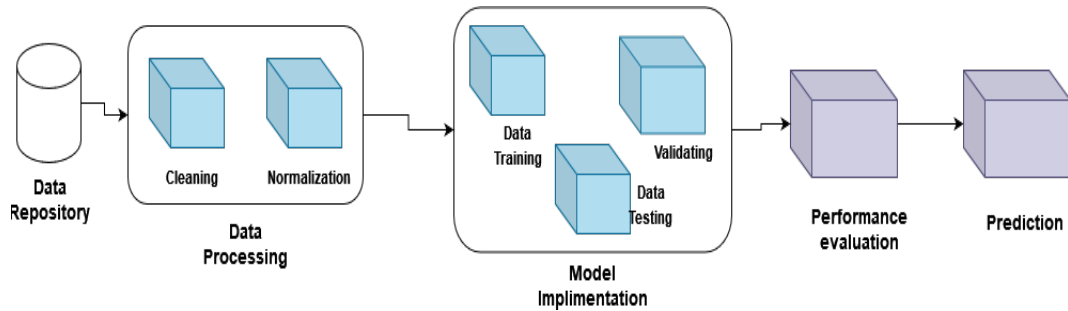


Figure 3.1: System Flow Diagram

Data is divided and trained in the post processing part. The trained data is further tested and validated in order to achieve the better output. Patterns selection and evaluation, creating model, model prediction all are conducted on the data.

In this research work, Supervised learning algorithms are used for the prediction purpose, whereas other statistical methods are used for the survival analysis.

3.1.1 System Architecture

The below given figure 3.2 shows the proposed system architecture.

3.2 Data Collection

The SEER database is used to fetch the data on thyroid cancer. SEER is the program by National Cancer Institute, which collects cancer related data from the cancer registries of U.S population, covering approximately 35% of the population. The database contains a wide range of data including individuals demographics, gender, time of diagnosis, follow up time, site of tumor and vital sign.

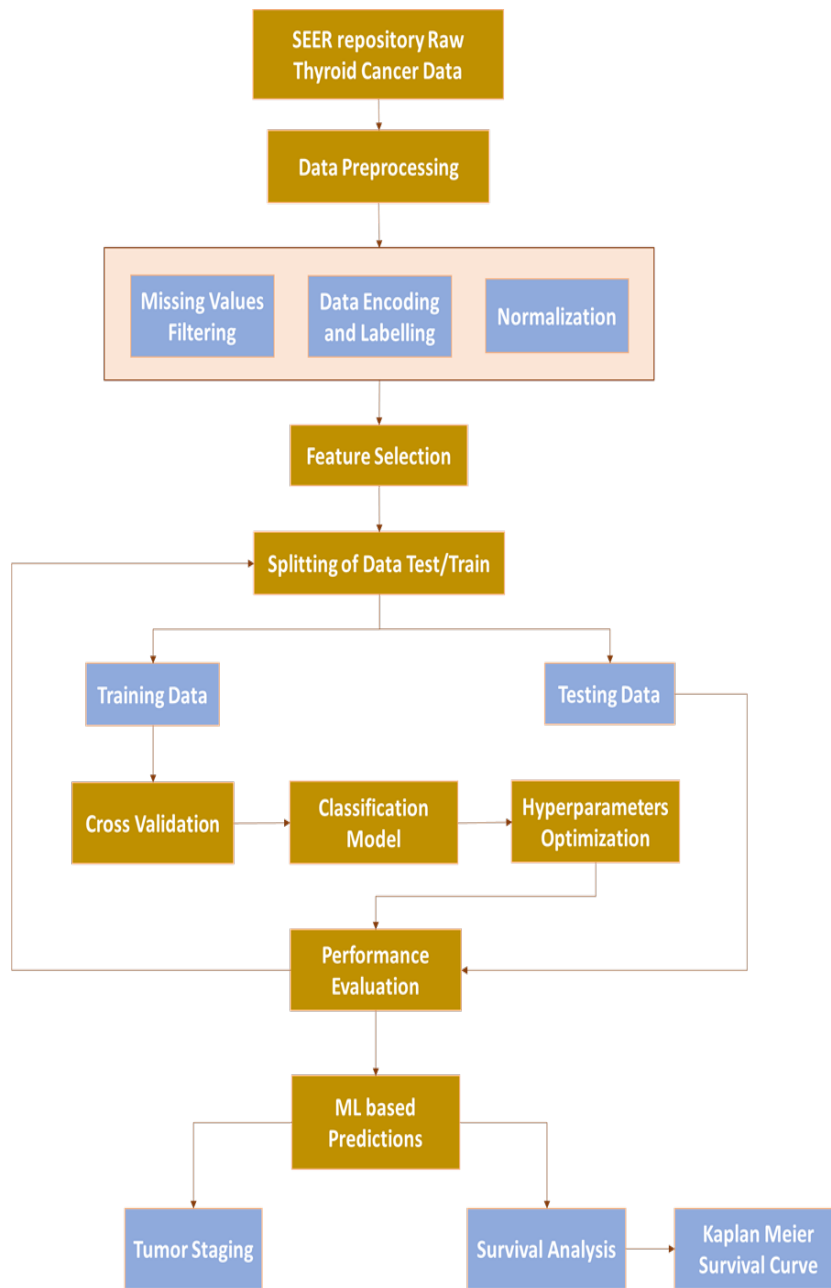


Figure 3.2: System Architecture Diagram

We particularly picked thyroid cancer diagnosed over the year 2004 to 2010, from the updated SEER registry submission of November 2017. Thyroid cancer has two most common differentiated types; follicular and papillary thyroid cancer. For our case, we selected these two from the SEER 18 Databases with the restrictions of ICD-O-3 equal to Thyroid. Specified the histologic type by selecting all available and exact ICD-O-3 for papillary and follicular cancer. Along with them Hurtle cell thyroid carcinoma was also included as follicular. The constraints ICDO-3/WHO 2008 = Thyroid and Histologic Type ICD-O-3 = 8050, 8260, 8340–8344, 8350, 8450–8460 (for papillary cancer) or Histologic Type ICD-O-3 = 8290, 8330–8335 (for follicular cancer) are being used to select cases of papillary and follicular thyroid cancer from the SEER 18 databases. Whereas, Hurthle cell cancer (ICDO-3=8290) was also classified as a follicular carcinoma. All other Thyroid based carcinomas were excluded as some are extremely rare and fatal while one is also staged differently, So including them for a population based analysis will be of less utility. The initially obtained raw dataset contained around 250 hundred thousand patients data 3.1.

Table 3.1: Unprocessed Raw Sample Data from SEER Database

Patient ID	2137	3375	355350	44872298
Survival months	48	0		75		51
SEER cause-specific death classification	Alive or dead of other cause	Dead (attributable to this cancer dx)		Alive or dead of other cause		Alive or dead of other cause
EOD 10 - size (1988-2003)	Blank(s)	Blank(s)		Blank(s)		Blank(s)
CS tumor size (2004-2015)	35	28		20		1
EOD 10 - extent (1988-2003)	Blank(s)	Blank(s)		Blank(s)		Blank(s)
CS extension (2004-2015)	200	550		200		100
Regional nodes positive (1988+)	1	98		98		98
CS mets at dx (2004-2015)	0	0		0		0
Age recode with <1 year olds	50-54 years	75-79 years		65-69 years		60-64 years
Year of diagnosis	2013	2013		2008		2010
Histology ICD-O-2	8260	8260		8290		8260
Histologic Type ICD-O-3	8260	8260		8290		8260
Site recode ICD-O-3/WHO 2008	Thyroid	Thyroid		Thyroid		Thyroid
COD to site recode	Alive	Thyroid		Alive		Diseases of Heart
Vital status recode (study cutoff used)	Alive	Dead		Alive		Dead
Gender	Male	Female		Female		Male
Derived AJCC T, 6th ed (2004-2015)	T2	T4a		T1		T1
Derived AJCC T, 7th ed (2010-2015)	T2(m)	T4a(s)		Blank(s)		T1a(s)
Derived AJCC N, 6th ed (2004-2015)	N1a	N0		N0		N0
Derived AJCC N, 7th ed (2010-2015)	N1a	N0		Blank(s)		N0
Derived AJCC M, 6th ed (2004-2015)	M0	M0		M0		M0
Derived AJCC M, 7th ed (2010-2015)	M0	M0		Blank(s)		M0

3.3 Data Description

The data collected consisted of a large number of variables each having a related information regarding the patient. Including the traditional values that were recorded until 2003 EOD Tumor Size, EOD Extension, EOD Lymph Nodes and hence multiple missing values. It included the same data under different label as CS Tumor Size, CS Extension, and CS Lymph Node. The SEER database collected data of tumor (T), lymph nodes (N) and metastasis (M) and started to record the derived T, N, M according to the 6th AJCC manual [36]. The dataset contained diagnosis starting from 2004 till 2010 including all the categories that match the current AJCC staging categories. The survival time was included, measured by months. Along with it, Seer cause specific death classification variable was added which records the deaths caused by the thyroid cancer. Whereas, age data was in the form of ranges but was converted into two bigger range categories as A1 (0–54), and A2 (55+). The important factors in data, included tumor size, regional lymph nodes, status of distant metastasis, and age (T,N,M,A). The descriptions related to T,N,M,A variables are mentioned below in the Table 3.2:

Table 3.2: Variables and Definitions regarding the differentiated thyroid Cancer

Factors	Levels	Definition
Tumor	T0	No evidence of primary tumor
	T1	Tumor less or equal to 2cm (dimension limit for thyroid)
	T2	Tumor size 2cm > 4 cm
	T3	Tumor greater than 4 cm
	T4a	Tumor of any size, exceeding the thyroid capsule and invading side organs including soft tissues, trachea, larynx, laryngeal nerves, or esophagus
	T4b	Tumor invades prevertebral fascia or the artery or vessels
Reginal Nodes	N0	No reginal lymph nodes
	N1a	Metastasis reached to level VI
	N1b	Metastasis to unilateral, bilateral, or contralateral cervical or superior mediastinal lymph nodes
Metastasis	M0	No occurrence of distant metastasis
	M1	Distant metastasis
Age	A1	Ages from 0 – < 55
	A2	Ages from 55 – 85+

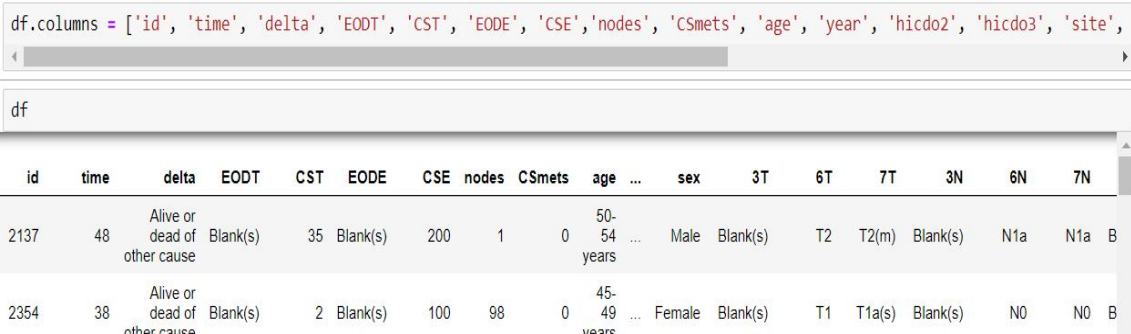
* For definitions refer to SEER Research Data Record Description [37]

3.4 Feature Selection

Feature selection strategies decrease the amount of traits and pick the most important input variables with the strongest link to the output variables. It is the process of extracting important characteristics from a dataset that may be utilised as input variables in the model that predict target values. The important features selected included Tumor, Lymph Nodes, Metastasis, Age, survival time, diagnosis years and vital sign which is a binary variable stating the patient's current state as dead or alive.

3.4.1 Renaming Attributes

The columns names are altered in the dataset into some relevant names to make things a little bit easier and comfortable for the study purpose, as the features in the dataset are identified by a certain keyword. An example of a few columns from the dataset can be explained. The columns named as Age recode with <1 year olds, Histologic Type ICD-O-3, Histologic Type ICD-O-2, Site recode ICD-O-3/WHO 2008 and Cause of death to site recode, were renamed as Age, hcode3, hcode2, site and COD respectively. The figure 3.3 below shows a few renamed columns:



The screenshot shows a Jupyter Notebook interface. At the top, a code cell displays the list of columns for a DataFrame named 'df':

```
df.columns = ['id', 'time', 'delta', 'EODT', 'CST', 'EODE', 'CSE', 'nodes', 'CSmets', 'age', 'year', 'hicdo2', 'hicdo3', 'site',
```

Below the code cell, a table preview of the DataFrame 'df' is shown. The table has 17 columns: id, time, delta, EODT, CST, EODE, CSE, nodes, CSmets, age, sex, 3T, 6T, 7T, 3N, 6N, 7N. The first two rows of data are visible:

id	time	delta	EODT	CST	EODE	CSE	nodes	CSmets	age	sex	3T	6T	7T	3N	6N	7N
2137	48	Alive or dead of other cause	Blank(s)	35	Blank(s)	200	1	0	50-54 years	Male	Blank(s)	T2	T2(m)	Blank(s)	N1a	N1a
2354	38	Alive or dead of other cause	Blank(s)	2	Blank(s)	100	98	0	45-49 years	Female	Blank(s)	T1	T1a(s)	Blank(s)	N0	N0

Figure 3.3: Renamed Data Columns

3.5 Data Pre-processing and Encoding

The dataset included multiple missing or NaN values, which if imputed using any imputing technique will produce biased results as its a real time, time-to-event data. Removing any of such missing or unknown data is more efficient way to have better results.

The data pre-processing is immediately taken into action by excluding any patient data

with missing, blank or unknown value in any of the main factors T, N, M, A, SEER cause-specific death and Survival time. It was noticeable that if a patient had a missing or unknown value of one variable from any of the important variables, it was more likely to have multiple other missing or unknown values. So, We eliminated patients with "T4NOS", around 3000 patients with unknown T values, 2500 patients with "N1NOS", 1800 patients with unknown N values, approximately 1500 patients with unknown M values, patients with unknown age and unknown survival time were also removed, 200+ patients with "Dead (missing/unknown COD)," and 8030 patients with "N/A not first tumor."

For the prediction and analysis, we decided to make combinations of our data such that the columns are merged into one. It is more likely to deal with the combination of data rather than the individual patients attributes. Basically, a combination formed of the prognostic factors is a subset of the data. To keep it more robust, the combinations formed are based on T, N, M, A. For example; T1N0M0A1 is a combination for the patients condition, in which T1 represents the tumor size, N0 is the spread of lymph nodes, where M0 is the metastasis site and A1 shows the age of the patient. The threshold is kept 25 such that each of the combination containing at least 25 patients. Therefore, we have 39 rare combinations, a total of 55137 unique cases with the diagnosis year 2004 to 2010.

3.5.1 Labelling and Encoding

The data is later, one hot-encoded and label encoded as well since it is a categorical data so for applying classification techniques it is adequate to encode the data. the features such as vital sign is encoded to a binary value as 0 and 1, where 0 represents the alive patients and 0 represents the dead patients. TNMA feature when one hot-encoded the dataset increased in size since there where around 39 rare combinations and so the number of columns increased.

After all cleaning and pre-processing of our data, we moved to the next step of labelling the data. Since the data is unlabelled so it is necessary to label the data for applying classification techniques. To learn the Stage of any T, N, M of a patient we followed the AJCC Staging. It is crucial to understand this that how the cancer data is labelled for its stages. There are several main points that needed to be kept in mind:

- The tumor's expanse (size) (T): What is the size of the cancer? Is it now encroaching on surrounding structures?
- The spread of the infection to neighbouring lymph nodes (N): Is there any evidence that the malignancy has migrated to neighbouring lymph nodes?
- The spread of cancer to distant places (metastasis): Is the cancer spreading to other organs, such as the lungs or the liver?

Hence, we labelled the data against the standard AJCC staging system [38]. The following Table 3.3 displays the AJCC based stages and the T,N, M stage against each stage and their description:

Since we made the combinations of Tumor, Nodes and Metastasis (TNM) with Age such that each combination represented the cancer condition of the patient including the factor of age. We had a combination of TNMA which is then one-hot encoded such that none of the categorical data is left in the final dataset.

3.5.2 Data Splitting

The data is divided into two parts, one is for training and the other for testing. Applying training and testing on same dataset does not provide with the appropriate results and might result in over fitting. In order to avoid such occurrences it is the best approach to split the dataset into train and test datasets. Out of the total 80% of the data is used for the training purpose while 20% of the data is used for the testing phase.

3.6 Class Balancing and Dimension Reduction Techniques

3.6.1 Principal Component Analysis (PCA)

PCA is a traditional method for dimensionality reduction or in other words feature reduction. It involves converting observed data into a new collection of variables that best preserves data diversity. It also accepts data with mixed-signs as input. As a result, PCA is one of the most commonly utilised algorithms.

The Thyroid cancer dataset is quite large even after all cleaning and processing. It will take large computational power and may cause trouble in outputs. It is required that all important features stay intact and performs robustly and efficiently.

Table 3.3: Thyroid Cancer Stages Against T,N,M Values

AJCC STAGE	STAGE GROUPING	DESCRIPTION FOR THYROID CANCER
I	T1	The cancer here is 2cm or smaller, confined in the thyroid (T1)
	N0	Cancer has not spread to lymph nodes (N0) or any distant site (M0).
	M0	
II	T2	The cancer is larger in size from 2cm but is not more than 4cm and is still confined to the thyroid. (T2)
	N0	Cancer has not spread to lymph nodes (N0) or any distant site (M0).
	M0	
	T3	Cancer is larger than 4cm and has confined thyroid or any size and is growing outside of the thyroid but has not touched other structures around (T3).
	N0	
	M0	Cancer has not spread to lymph nodes (N0) or any distant site (M0).
III	T1, T2 or T3	The cancer is of any size and can possibly grow outside of thyroid but is not involving nearby structures (T1, T2, T3).
	N1a	It has spread to the lymph nodes in the neck but not to other lymph nodes (N1a) or other distant sites (M0).
	M0	
IVA	T4a	Cancer is of any size and has grown beyond thyroid gland into nearby tissues in the neck including, larynx, trachea, esophagus, or the nerve to larynx (T4a).
	Any N	
	M0	It might or might not have spread to nearby lymph nodes (any N), and it has not spread to any distant site (M0).
	T1, T2 or T3	The cancer is of any size and may grow outside of thyroid but is not involving nearby structures (T1, T2, T3).
	N1b	Cancer has spread to certain lymph nodes in the neck such as cervical or jugular nodes (N1b). It has not spread to distant sites (M0).
	M0	
IVB	Any T,	Cancer is of any size and may have grown into the nearby structures in neck (any T).
	Any N,	
	M1	It might or might not have spread to nearby lymph nodes (any N). It has spread to other distant sites such as lungs, bones, liver or even brain (M1).

It is used with my dataset, By lowering the dimensionality of the datasets, we employed principal component analysis (PCA) to speed up the algorithms. For the sake of experiment the models were prepared in both ways, with PCA and without PCA.

3.6.2 Over-Sampling and Under-Sampling

Machine learning models perform better when the datasets are balanced. Models developed on unbalanced datasets tend to favour the majority and this results in biasness in the outputs. It leads to increased miss-classification of classes less in count and so a poor model performance. Balancing the classes improves the models' prediction performance and decreases the likelihood of miss-classification.

Though applying it on whole data may overfit the results, so its always better not to apply balancing techniques on whole dataset.

For our dataset, the idea of Oversampling was used for the classes less in count. Oversampling raises the number of the minority class to the majority class until the distribution is balanced. There are multiple techniques for this purpose, I used SMOTE technique. It is a robust technique in terms of oversampling. Not all the models were prepared with this data but for the purpose of experiment and analysis a few classifiers were trained with SMOTE generated data.

3.7 K-fold cross Validation

Once the model is trained, the separated data is used for testing and validation. Cross Validation often known as re-sampling of procedure for data. It is an unknown fact that how well does a trained model performs on an unseen dataset. There are multiple validation techniques and we used K-fold cross validation technique.

To perform K fold cross validation, the data is divided into k subsets. Each k-1 subset is used for training and rest one subset is used for testing and validation. This process keeps on working in iterations such that each subset is used for training and testing. It allows the model to learn at each data point. We kept k as 10 and so 9 out of the total subsets of data are used for the training purpose, changing frequently on every iteration.

3.8 Machine Learning Algorithms

Machine learning algorithms can be categorised into few important branches, Supervised, Unsupervised and Reinforcement learning. Each learning method has a different processing for the input dataset generating related output. As the fresh data is fed into these algorithms, they train and optimise their processes in order to increase performance, thereby acquiring intelligence.

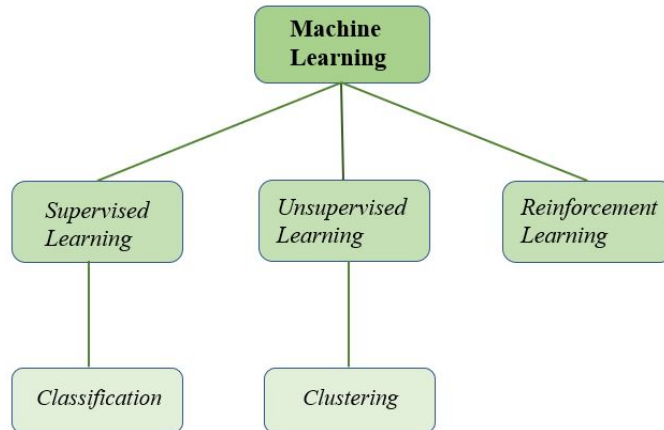


Figure 3.4: Branches for Machine Learning

Supervised learning algorithms are widely used for the cancer diagnosis and prognosis. Supervised learning methods produce a function that can translate an input to an output based on the supplied sample input-output combinations. These algorithms require labelled and organised datasets for learning and training in order to provide an output. It's tough to manually link the traits and estimate a patient's result and indications of a tumor grade or disease-free survival time. Hence, for our research, we used multiple machine learning algorithms and ensemble methods to achieve the accuracy for such real time data.

3.8.1 Gaussian Naive Bayes

This section covers supervised learning method classification using naive Bayes. It is one of the simplest machine learning classification algorithm. It uses the Naive Bayes Theorem 3.8.1, where the probability of one with respect to the other is true, if the event happened earlier is true.

$$P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A}) \frac{P(\mathbf{B}|\mathbf{A})}{P(\mathbf{B})}, \quad (3.8.1)$$

In most cases, classification begins with a set of vectors that may be used as a blueprint for each data point in the dataset. We sought to create a classifier called Y that could predict cancer Stages for the dataset provided. To obtain a prediction, the algorithm uses the likelihood of each attribute belonging to the specified class/stage. The qualities of the baseline data collection, in this case the SEER database, were the foundation for this development.

3.8.2 Decision Tree

The decision tree (DT) algorithms are a tree based structure, in which the data points are classified in a tree form [25]. This approach is popular because it depicts a conceptual thought process that can begin at the root and end at the leaves.

The dataset is passed through the model such that the branches reflect the combination of all the main characteristics from the feature vectors in the dataset, while the leaves represent classes such as the stages.

3.8.3 Support Vector Machine

Support Vector Machines (SVM) are more recent and different approach then the previous ones in the field of cancer prognosis [31]. Initially it transfer the input vector into a higher-dimensional feature space and find the hyperplane that divides the data points into the given output test set which are the stages in our dataset. The formed hyperplane may be viewed as a decision line between the trained five clusters. By this it is understandable that SVM defines the decision boundary, it determines the best place to draw the separation lines. These lines divide the space into sub-spaces. Indeed, the presence of a decision boundary enables for the discovery of any method-induced wrong classification of the data points.

3.8.4 Random Forest

Random Forest (RF) is one of the ensemble method used for classification. It is considered as one of the most flexible and advance ML technique [39]. RF relies on generating number of random trees. Due to this reason, It is found one of the plus points for the RF that multiple decision tress generation boosted our results, whereas a single decision tree has a higher probability for flawed outputs. That is why considered a strong learner.

Keeping this in mind that the number of estimators, used to build the number of decision trees and split the selection criteria, which is used to assess the quality of splits must be taken into account in Random Forest. Else the results obtained are not interpretable. So, the two main parameters required for the random forest algorithm to work effectively on our dataset are: \mathbf{t} is the number of decision trees that are to be formed and \mathbf{m} is the number of input parameters to be included, required when formed decision tree is split on each node.

3.9 AdaBoost

AdaBoost is a machine learning approach that may be thought of as a meta-algorithm. It simply enhances the model performance when used in conjunction with other learning strategies. AdaBoost focuses on using weak classifiers consecutively in a classification issue. It produces results using the weighted ideology. As a result, the method is applied to the updated data many times. This improves the outcome a double times, this is why, AdaBoost is preferred over many other classifiers. It is widely used in the medical field research due to its robustness and better outcomes.

3.9.1 K-Nearest Neighbours

Instance-based learning, often known as lazy learning, is a technique for classification and regression. K-NN is instance based learning. The new instance predictions are created by searching the full training set for the K closest examples. For new instances, distance is calculated, and weights are applied to the closest neighbours rather than the furthest. KNN uses the Euclidean distance formula to determine the distance between data points [40]. It is easy to implement and hence widely used in almost all fields.

It uses the brute force computation method, works well when datasets are not so large. Else with large datasets it becomes inefficient.

3.9.2 K-Mode

One of the widely used unsupervised learning method for categorical data, K-mode Clustering is used with our dataset. K-modes cluster techniques, works well when there are huge datasets, commonly called as partitioning clustering.

The K-Modes approach has been frequently used to substitute k-means in categorical data clustering since K-mean can only deal with numerical data clustering. Because of its uncomplicated implementation and minimal number of repetitions, K-mode clustering is a popular approach.

It forms clusters on the basis of the nearest centroid, which the mode of the data points in a group and so is called as K-mode [41]. It uses a randomly determined starting cluster centre (modes). Where k is the number of clusters formed out of the given data points in the hyperplane.

In the described dataset, the distance between each data point is calculated and the centre of each group is quickly determined, resulting in an optimal clustering of data in which the data points within each cluster are near to the it's centre whereas, the distance between two differently clustered data points have distance equal to maximum.

3.9.3 Light Gradient Boosting

Gradient Boosting Machines are the Decision Trees-based ensemble algorithms (DT). It uses the concept of forward distribution algorithms. The residual from each tree iteration is fit by a negative gradient in each iteration to learn of the decision tree [42]. Boosting is a fundamental and naive machine learning approach. By each iteration, the weight of each weak learner is modified based on the learning performance.

The Gradient boosting machine works in a very effective way. Every new tree is formed after each iteration which helps to improve its learning, such that every previous tree residual is used to form the new tree and ultimately produces a more accurate and reliable output prediction. The Light Gradient Boosting Machine (LightGBM) [43] is more robust that rather than growing horizontally it grows vertically, consuming less system power, lesser time and maintaining high accuracy. Its leaf-wise growth lower downs the loss. The trees formed are more complex, they split leaf-wise rather than the level-wise approach of splitting.

Typically, other gradient boosting machines are used for diagnosis only. In this research, LightGBM was picked for the prognosis purpose. Provided a faster training speed with our large dataset and higher efficiency, better than the many other algorithms. It consumed less memory area of the system and provided higher accuracy rate. Due to its compatibility with the larger dataset, the model handled the dataset quite effeciently.

3.10 Hyper-parameter

Machine Learning Algorithms have option to add both parameter and hyper-parameter. Other than the parameters, hyper-parameters are entered in the model as they are not learned during the training phase.

We used the GridSearch CV for the model tuning with hyper-parameters as it uses all the best possible options given. It uses cross-validation method to select the best parameter for the model. Different hyper-parameters are selected for every model. The Hyper-parameters used for the machine learning algorithm in this research are as follow in the table [3.4](#):

3.11 Survival Analysis

Survival analysis is basically an estimated duration over time until specific event occurs. It is the study of time-to-event (survival times) data. In our research survival can be defined as the, patients life duration during his disease until event such death may occur. In an idealized study, it will be optimal to have all patients diagnosed at the same time and enrolled in the research until a result (perhaps death) was obtained, and follow-ups would be conducted in a continuous pattern.

Finding a large sample of patients in optimum settings, on the other hand, is nearly difficult. Some patients had already been diagnosed and were on treatment in the data set. Others dropped out of the research and never followed up. For such scenarios statistical procedures are used to acquire useful information from partial datasets since we do not always have a perfect dataset. This is why we are compelled to employ survival analysis approaches.

In the observed survival data, the death of a patient can not only due the the specific thyroid cancer but can be from multiple reasons too, such as heart attack, accidental death or other underlying medical condition etc. But in the relative survival we take the thyroid cancer specific deaths into account. Since the other causes of death will not be having any impact on the survival of the patient due to cancer.

Table 3.4: Hyper-Parameters for the Machine Learning Algorithms

Classifier	Hyperparameters		
Decision Tree	Criterion= entropy max_depth = 200 max_features = sqrt		
Random Forest	Criterion = gini max_depth = 200 max_features = log2 n_estimators = 20		
SVM	<table border="1"> <tr> <td>Kernel = rbf Gamma = 0.5 C = 0.1 Probability = True</td> <td>Kernel = 'poly' Degree = 3 C = 1 Probability = True</td> </tr> </table>	Kernel = rbf Gamma = 0.5 C = 0.1 Probability = True	Kernel = 'poly' Degree = 3 C = 1 Probability = True
Kernel = rbf Gamma = 0.5 C = 0.1 Probability = True	Kernel = 'poly' Degree = 3 C = 1 Probability = True		
AdaBoost	Learning_rate = 0.01, 0.05 When applied with Decision tree classifier as the base max_depth=1 param_grid, cv = 3		
LightGBM	random_state = 42, class_weight = balanced cv = 5 parameters = { 'n_estimators': [5, 10, 15, 20, 25, 50, 100], 'learning_rate': [0.01, 0.05, 0.1], 'num_leaves': [7, 15, 31], }		
K-mode Clustering	n_clusters = range (1-8) init = random n_init = 7 verbose = 1		
KNN	N_neighbours = 1- 21		
Gaussian Naïve Bayes	Default Parameters		

3.11.1 Survival Function

A survival function can be said as the probability of surviving beyond the time t . Where, survival function $S(t)$ and $0 < t < \infty$. As shown in the equation 3.11.1

$$S(t) = P(T \geq t) = 1 - F(t), t > 0 \quad (3.11.1)$$

where T is the experimental random variable (time to event), $F(t)$ is the cumulative distribution function of T and t is a fixed integer describing time. The survival function's graph $S(t)$ is referred as the survival curve, which starts at $S(t)=1$ and declines to 0 as t grows. When $t=0$, $S(t) = 1$ and when $t=\infty$ than $S(t) = 0$.

3.11.2 Kaplan Meier Method

Kaplan-Meier is one of the most prominent survival analysis technique. It includes computing probability of survival within a small interval of time. The Kaplan-Meier method uses the cumulative survival rate at the end of each year of follow-up and calculates the proportion of patients still alive at intervals as short as the accuracy of death recording allows.

As discussed, it is possible that the patient is still alive by the end of the study or has not followed-up, which means their survival time is unknown after the cancer detection for the given time-interval. This type of survival time and information is called as censoring. It can be due to many reasons and of many types.

Therefore, we are aware of the censored and truncation variables that may result in partial observations, but we cannot exclude them since each individual contributes information to the computation as long as they are event-free, and we do not want to reduce our sample size by removing those patients and the data information will be very less reliable. A skewed estimate will also come from excluding censored cases.

For both the cases censored and uncensored survival periods, Kaplan-Meier is regarded the simplest approach for assessing survival probability. It is determined by taking the probability of the number of patients who survived by the number of patients who were at danger at different periods. The expected probability of survival at a certain point in time can be given as: $1 - d/n$ where, d = number of subjects died / number of events occurred and n = number of subjects alive at the start of the time.

The cumulative survival probability over the follow-up time (t_j) for the subjects in study

is obtained by multiplying all the individual probabilities (p_j) for the specific time within the selected time interval. Such as the given equation depicts 3.11.2 :

$$\hat{S}(t_j) = \hat{P}(T > t_j) = p_1 \cdot p_2 \cdot \dots \cdot p_{t_j} \quad (3.11.2)$$

Therefore, the Kaplan-Meier is estimated as :

$$\hat{S}(t_k) = \prod_{t_k < t} S(t_{k-1}) \left(1 - \frac{d_k}{n_k}\right) \quad 1 < k < j \quad (3.11.3)$$

Here, n_k are the number of patients alive and are at risk before the time t_k , while d_k are the patients who died at time t_k . Whereas, $S(t_k - 1)$ represents the conditional probability of survival of the patient by the course of time, such that if the patient survived at time t_1 than it is multiplied as the input for the second instance, while $S(0)=1$.

3.12 Evaluation Metric

Evaluating the performance of the algorithm is one of the most important and trickiest in machine learning implementation. After running the model on the dataset, the outcome generated are evaluated and validated through different measures. The predicted results of Stages of cancer are so compared with the actual using different measures and model performance is estimated.

Model Validation: For this purpose we used k-fold cross validation, each model was validated through this iterative method. The detailed process is mentioned in the above section 3.7 .

Performance Evaluation: There are multiple evaluation metrics, out of which we used Confusion matrix, Accuracy score, Recall, Precision, F-measure and C-index.

A Confusion matrix is a table which consists of 4 combinations of predicted and actual values. This aids in determining whether or not the data is being sorted appropriately. Hence, the number of accurate and wrong predictions produced by the prediction models is presented in the confusion matrix. It consists of True positive, True negative, False positive and False negative values. Since the expected output is not binary but a multi-class problem, such that the output could be any of the possible thyroid cancer stage which means it can be Stage I, II, III, IVA or IVB. So the confusion matrix can not be formed as the usual binary matrix but it will consist of other possible outcome values

too. A general depiction for such multi-class confusion matrix is as shown in the figure 3.5 below :

	Predicted Values (A)	Predicted Values (B)	Predicted Values (C)
Actual Values (A)	True (AA)	False (AB)	False (AC)
Actual Values (B)	False (BA)	True (BB)	False (BC)
Actual Values (C)	False (CA)	False (CB)	True (CC)

Figure 3.5: Confusion Matrix Example for Multi-class Classification

Here, for A it will be further compressed into the binary format as given below in figure 3.6. In this way all the rest of the classes are determined and confusion matrix is formed. The values for True positive, True negative, False positive and False negative are obtained for each of the output class.

	Predicted	
Actual	True (AA)	False (AB, AC)
	False (BA, CA)	True (BB, BC,CB,CC)

Figure 3.6: Confusion Matrix for Multi-class Classification

True positive (TP) represents that a patient is rightly diagnosed for the particular cancer stage by the model, which our classifiers predicted correctly. True negative (TN) shows that the patient is not having the particular cancer stage and the model predicts it correctly too. Where, False positive (FP) is the opposite of TP such that the patient is of stage II but is predicted as not of this stage. While False negative (FN) shows that a patient is of the one specific cancer stage but is incorrectly predicted by the model.

Precision and Recall: Precision tells how much accurately the model has predicted the results. While Recall gives a bigger picture of the output, it tells that out of all the correct outcomes, how many of them are accurate and label them as true positives. For

the binary case both can be calculated as;

$$Precision = \frac{TP}{TP + FP} \quad (3.12.1)$$

$$Recall = \frac{TP}{TP + FN} \quad (3.12.2)$$

The simple binary case of precision and recall can be extended towards the multi-class concept. Such that if N is the number of confusion matrix for the multiple output classes than the matrix formed will be a k x k where k will be the number of classes.

$$Precision = \frac{N_{ii}}{\sum_j N_{ji}} \quad (3.12.3)$$

$$Recall = \frac{N_{ii}}{\sum_i N_{ij}} \quad (3.12.4)$$

Accuracy: Since our target variable consisted of multiple values, So the multi-class output accuracy is calculated as mentioned in equation 3.12.5;

$$Accuracy = \frac{\sum_{i=1}^l \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l} \times 100\% \quad (3.12.5)$$

where tpi are true positives, tni are the true negatives, fpi false positive ones and fni are false negatives, where l represents the number of classes.

It is understandable that a multi-class output could not give values with binary methods instead a classification report against each model is produced. A **Classification Report** for our dataset consists of Micro, Macro and Weighted results on the basis of number of classes in the target variable. Hence all the three evaluation measures precision, recall and f1 score were calculated against each label as micro, macro and weighted.

3.13 Summary

The Chapter discussed the whole methodology in steps by which the whole research proceeded further. The raw data is collected for Thyroid Cancer from the SEER repositories. The data gathering from SEER was not possible until SEER organization authorises the person to have its access. After data gathering, the main task related data pre-processing begins. The data is cleaned and important Features are selected, making up a new combination of features as TNMA. This combination is used in further Machine Learning model implementations. The models are trained on hyper-parameters.

CHAPTER 3: METHODOLOGY

The techniques for data balancing and models used are explained including the survival function. The Evaluation metrics used for the multi-class scenario are discussed in detail for the dataset.

The next chapters presents the results that are obtained after testing the models and concludes the research work with its future scope.

Implementation and Results

In this chapter, I have briefly explained the experiments and their outcomes. I used my findings on the dataset, to predict the stages for the cancer condition. Used the performance assessment measures to assess the models' performance.

4.1 Data Analysis

The data was analyzed prior to the models. It is very important to understand the data else it is hard to handle the machine learning models with the data. I took 5- year data of thyroid cancer, a time - to event data, but do not know exactly how many cases are there in each year. The figure 4.1 below shows thyroid cancer cases by year.

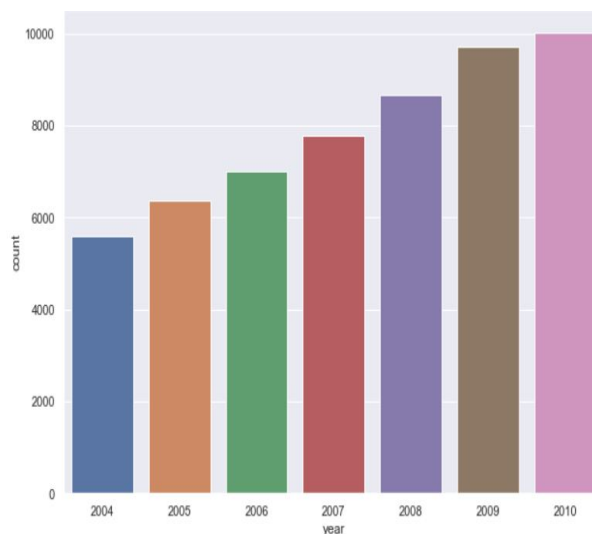


Figure 4.1: Thyroid Cancer cases over years

CHAPTER 4: IMPLEMENTATION AND RESULTS

We utilised the heat map to figure out the relationships between different data variables after pre-processing. The figure 4.2 shows different features and there correlation. The map shows how the TNMA values differ to each other, yet are closely related too, each TNMA value is unique. It is easy to interpret from this that each feature depends on the tumor type and the stage of the cancer.

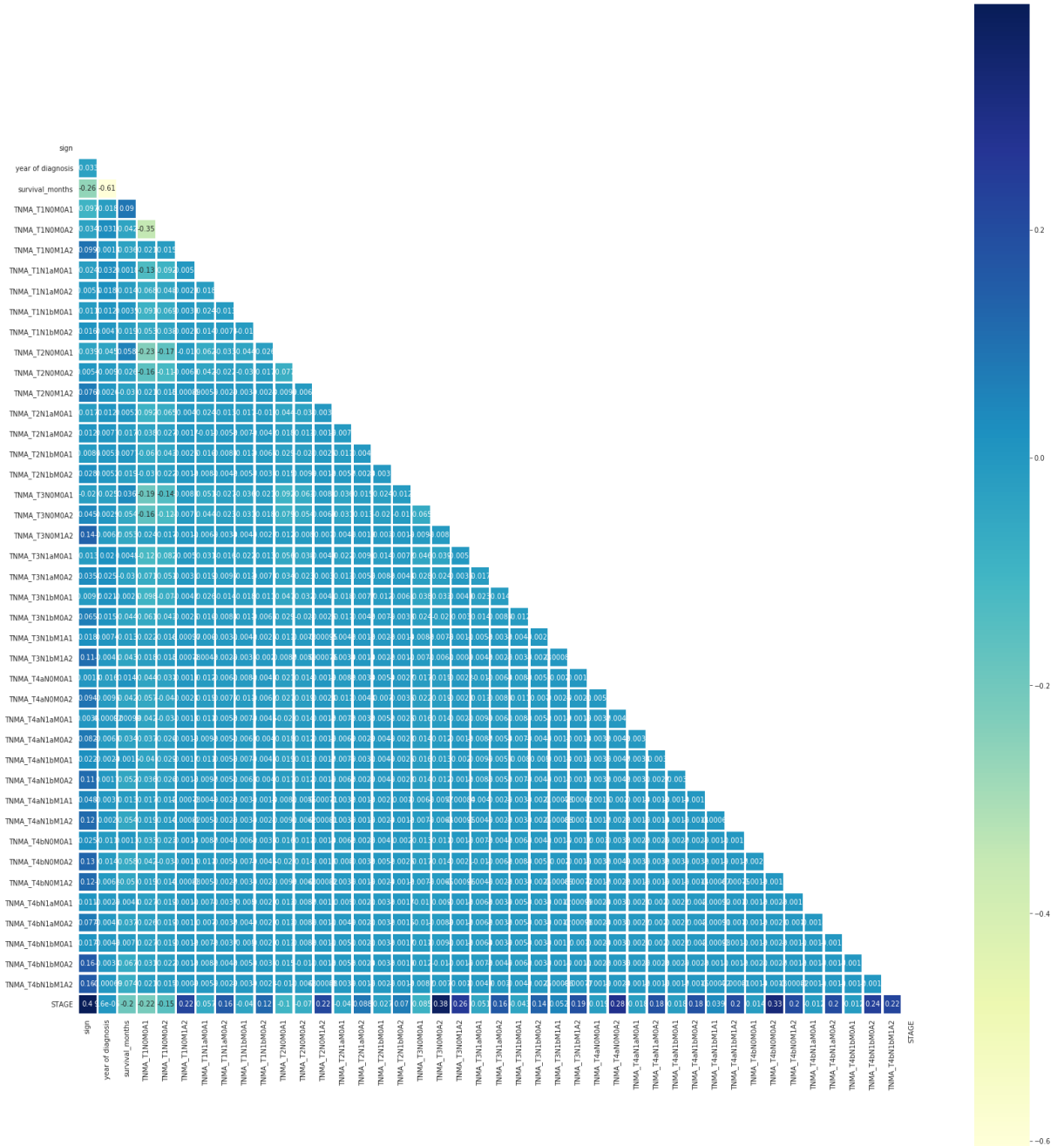


Figure 4.2: Thyroid Dataset Correlation Heat-map

It is clearly evident from the figure below that the amount of patients dead is less as compared to the patients alive. As mentioned in Chapter 2, the incidences for thyroid has drastically increased so has the mortality rate, but still not to the peaks of deaths as many other cancer types has a high death toll.

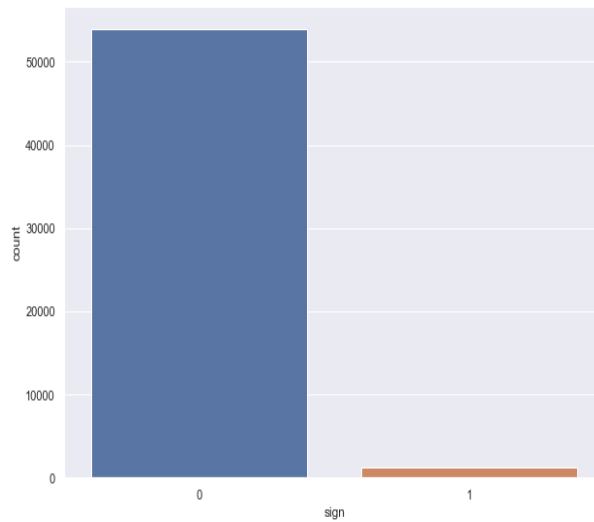


Figure 4.3: Thyroid Cancer Vital Status (Dead or Alive)

The patients in the critical situation are usually in the last stage of cancer. It is certain that a thyroid cancer patient may fall in this stage if not treated on time. The figure 4.4 shows that maximum of the patients are in the first stage as the stage increases the amount of patients is decreasing. But it does not mean that the patient remains in the same condition by next years.

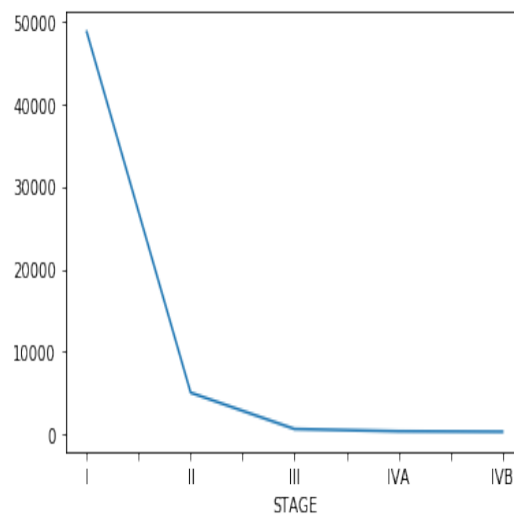


Figure 4.4: Patient count w.r.t Stages

4.2 Experiment Analysis

The Thyroid cancer dataset contained the selected features which includes Tumor, Lymph Nodes, Metastasis, Age as TNMA, survival time in months, year of diagnosis and vital sign as 0 or 1 stating the current state as dead or alive. The dataset was divided into two parts: training and validation. On the training dataset, several machine learning models are trained, and different hyper-parameters are optimized via cross-validation on the dataset. Multiple machine learning models and multiple sampling techniques which are implemented on the dataset produced quite promising results. The obtained results are in the form of confusion matrix from which a detailed multi-class classification report is generated. The confusion matrix give results in the form of TP, TF, FP, FN. In our case of multi-class classification, the metrics are generated against each target class. A classification report consists of micro, macro and weighted results on the basis of number of classes. Since the target variable in our research consisted of 5 possible classes hence all the three evaluation measures precision, recall and f1 score were calculated against each class as micro, macro and weighted.

Although, in such case it is not preferred to have an AUC-ROC score, since it will be calculated as either one class against rest or one vs one. In each case the averaged answer obtained is not really preferred and reliable, it is in the multi-class scenarios only. It is also not a good measure when there are imbalanced classes, in such conditions confusion matrix gives a better understanding.

It is well proven that using unbalanced data for machine learning models lead to very low sensitivity. In this research, its not only multi-class but also have the class imbalance. Using the cancer data to train Machine Learning models might result in very low sensitivity with such case. In order to avoid this and for the training datasets, we used SMOTE approaches to tackle the low imbalanced problem. The machine learning models were regularly tested and modified, and the ideal parameters for each model were found.

4.3 Results

4.3.1 Experimental Result with Decision Tree

The figure below shows the classification report of the Decision tree classifier for the data. For the model, the sklearn library in python was used. Trained the model and tuned it on the hyper-parameters, shown in figure 4.5. The parameters were passed and out of them the best possible were selected by the model.

```
parameters = {'max_features': ['log2', 'sqrt'],
              'max_depth': [10, 200, 5000],
              'criterion': ['gini', 'entropy']}
```

Figure 4.5: Parameters passed for Decision Tree

After the model training and testing the classification report is generated. It is presented below in the figure 4.6:

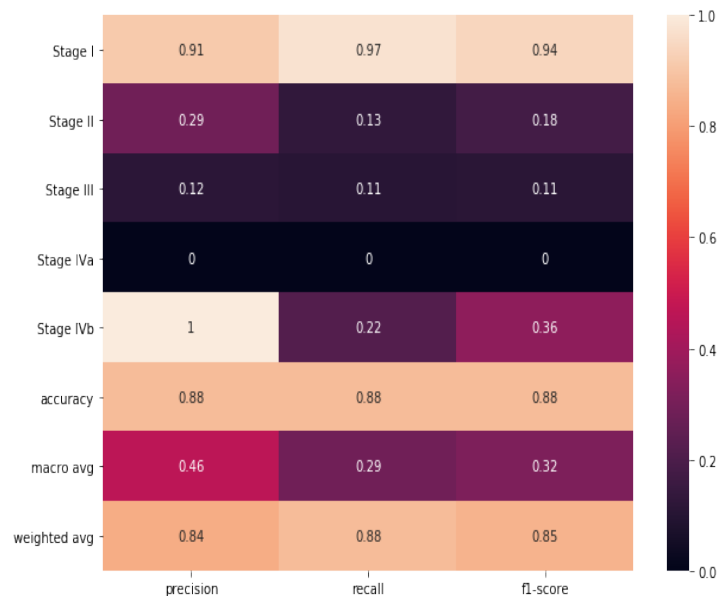


Figure 4.6: Classification-Report Decision Tree

As seen by the chart, the Decision Tree classifier gives the accuracy of 0.88 with the weighted average precision of 0.84, recall of 0.88 and F1 score of 0.85. It is seen that the figure 4.6 also shows the individual class precision, recall and F1 values such that for each Stage the values are calculated separately.

4.3.2 Experimental Result with Gaussian Naive Bayes

For this experiment, the dataset was trained on the default parameters of the Naive Bayes algorithm. The classifier worked quite well with the imbalance’s of the classes as seen through the given graph 4.7.

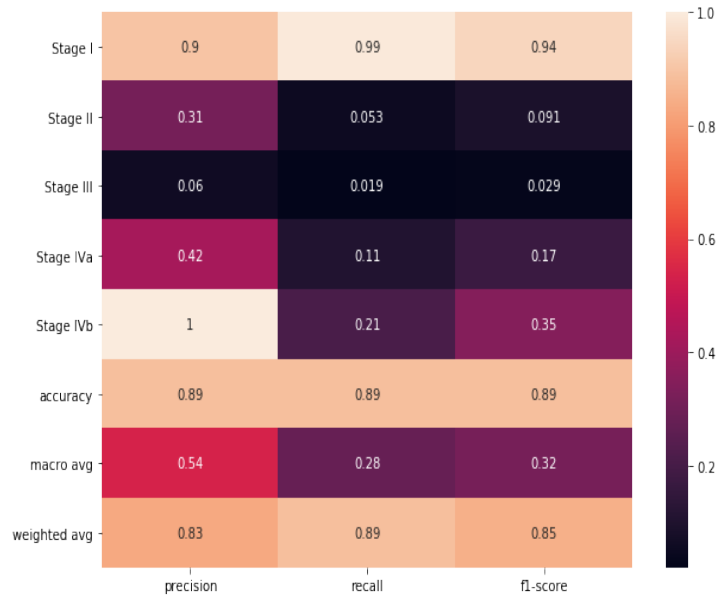


Figure 4.7: Classification-Report Naive Bayes

The Naive Bayes classifier gives the accuracy of 0.89 which means the classifier predicted stages are 89% accurate. With the weighted average precision of 0.81, recall of 0.89 and F1 score of 0.84. The individual stages values are also shown in the figure.

4.3.3 Experimental Result with Support Vector Machine

For Support Vector Machine, three different kernels for model training were used. That are linear, polynomial, radial basis function (rbf) kernels. Out of these RBF performed better than the rest of them.

For SVM linear kernel, the data was processed through the Principal component analysis (PCA). Although SVM can handle the imbalance of the classes but with PCA it produced much better results when the linear kernel is selected. The figure 4.8 below shows the graph for SVM linear kernel. The accuracy achieved is 0.89 while precision is 0.81, recall 0.89 and F1 score is 0.84 respectively.

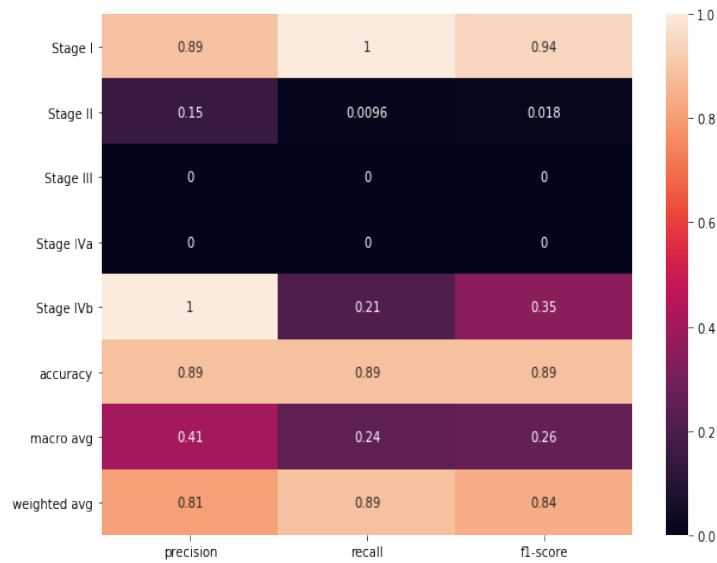


Figure 4.8: Classification-Report SVM (linear)

For SVM poly kernel, the hyper-parameters were tuned with the Grid Search. The accuracy obtained is 0.89. The figure 4.9 below shows the graph for SVM poly kernel.

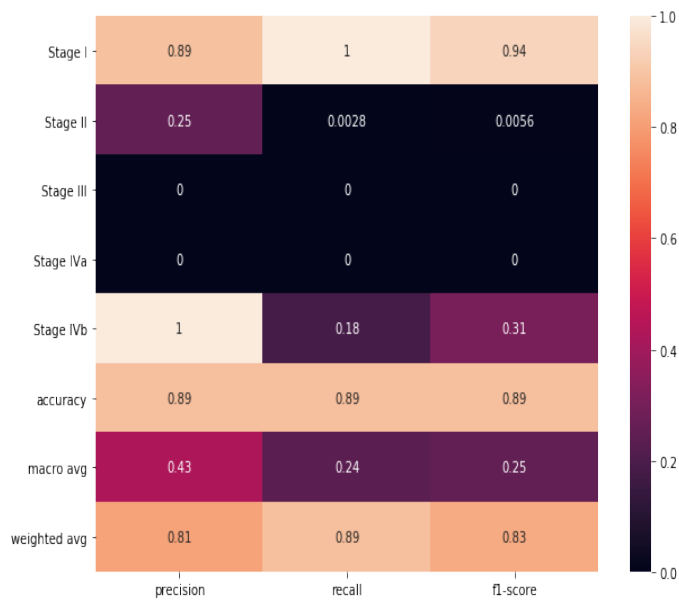


Figure 4.9: Classification-Report SVM (poly)

The figure 4.10 shows the classification report for SVM with RBF kernel. The model was tuned on the hyper-parameters such that the parameter used to avoid miss-classification or to have higher optimization is kept in range from min (1) to max (1000). The accuracy

attained is 0.87 but the precision value 0.83 is higher than the rest.

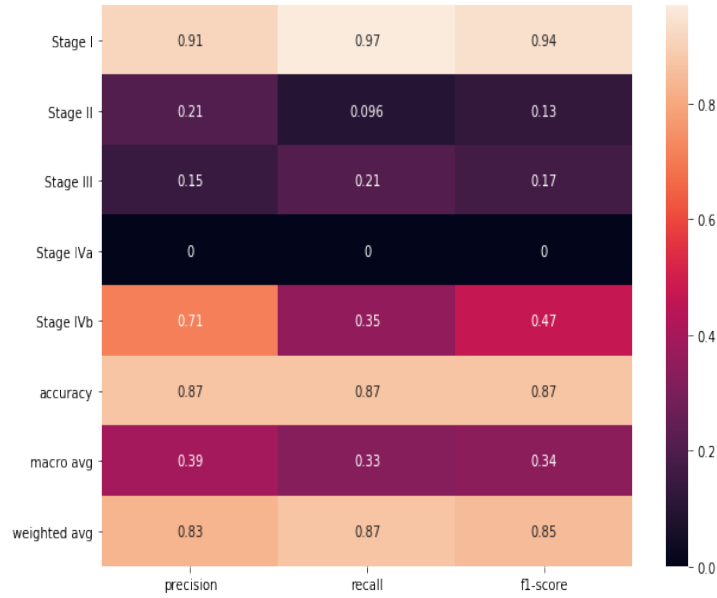


Figure 4.10: Classification-Report SVM (rbf)

4.3.4 Experimental Result with Random Forest

To implement Random Forest, we utilised Python’s sklearn library for random forest. The classifier’s overall performance was analyzed. The model was first tuned on hyper-parameters using Grid Search library in python. Figure 4.11 shows the classification report for RF, where data was processed with PCA.

The above figure shows the accuracy of 0.81. When considering an ensemble algorithm the estimated accuracy is high but the value obtained is low as per the estimations.

The figure 4.12 below presents the report when data was processed using the SMOTE. In order, to handle the class imbalance the SMOTE library for over-sampling is used with thyroid dataset. The model performed efficiently, the results obtained are higher such that the accuracy obtained is 91% with SMOTE oversampling measure. The precision, recall and F1-score are 0.94, 0.91 and 0.91 respectively. This higher precision score shows that max number of classes are rightly classified.

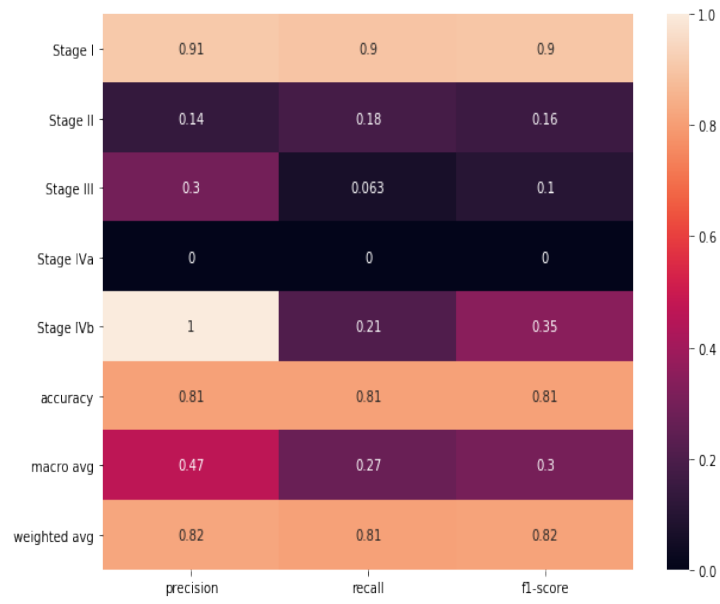


Figure 4.11: Classification-Report Random Forest using PCA

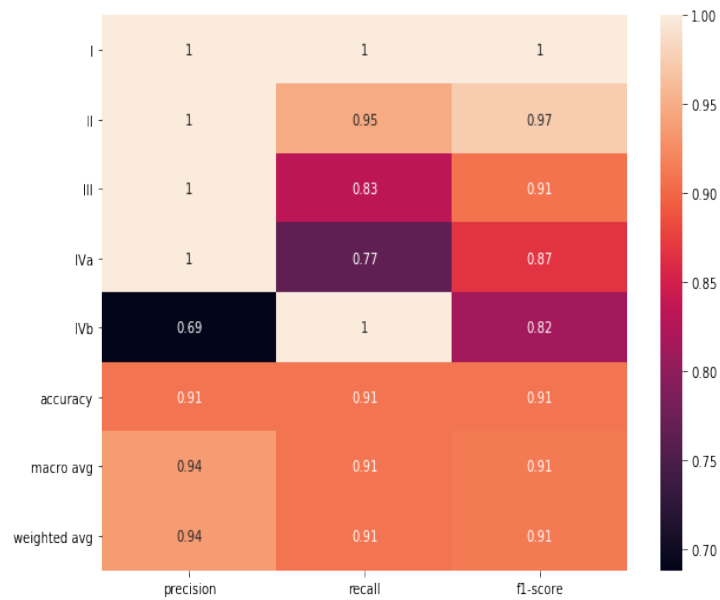


Figure 4.12: Classification-Report Random Forest using SMOTE

4.3.5 Experimental Result with AdaBoost

For thyroid predictions, AdaBoost was not just tuned on hyper-parameters but along with it a Decision-Tree model was used as the base algorithm for the ensemble method, figure 4.13.

```
param_grid = [{"n_estimators": [100, 115, 130, 145], "learning_rate": [.6, .7, .8, .9, 1]}]
ab_grid_search = GridSearchCV(AdaBoostClassifier(DecisionTreeClassifier(max_depth=1), algorithm="SAMME.R"),
    param_grid, cv=3, scoring=make_scorer(roc_auc_score),
    verbose=5
)
```

Figure 4.13: Parameters used with AdaBoost

The figure 4.14 shows the ensemble method AdaBoost classification report. Out of all the classifiers used, it performed best and consumed much less time in processing. With low time consumption, low GPU utilization and higher performance, this model outperformed rest of the ML models. The accuracy of 96% with an F1 score of 0.95, the classifier performed efficiently.

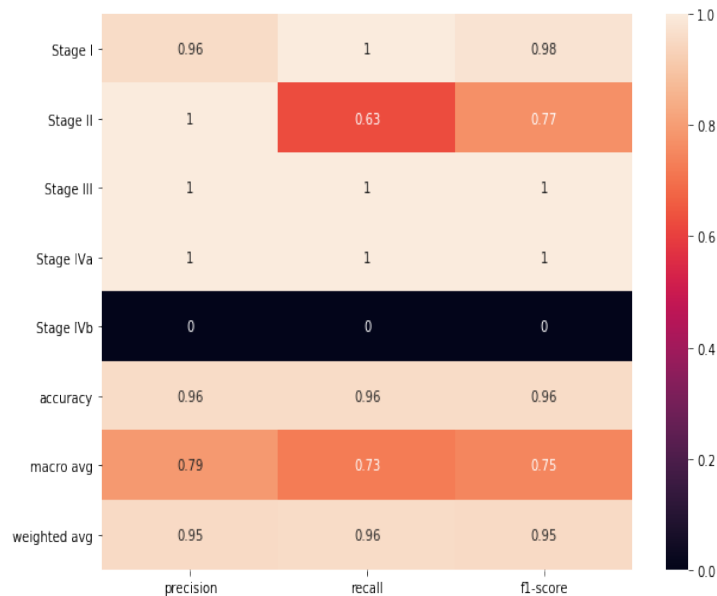


Figure 4.14: Classification-Report AdaBoost

4.3.6 Experimental Result with K-Mode Clustering

The unsupervised technique of clustering, called as K-mode clustering is used with thyroid data set. Since K-mode can handle the categorical data, so for this reason the encoded data was not used in K-mode clustering. In fact, the dataset in its original form without any encoding was used. The figure 4.15 presents the Elbow curve, which evaluates the clusters formed. As the elbow goes down, the point is reached where it has a stabilizes and goes down constantly.

The initial runs were set to 7 to find the best run, and as seen in the graph the optimal K value is at 5, since after k= 5 there is a drastic fall on k= 6. This can be interpreted as, 5 clusters formation would be most optimal for the thyroid dataset.

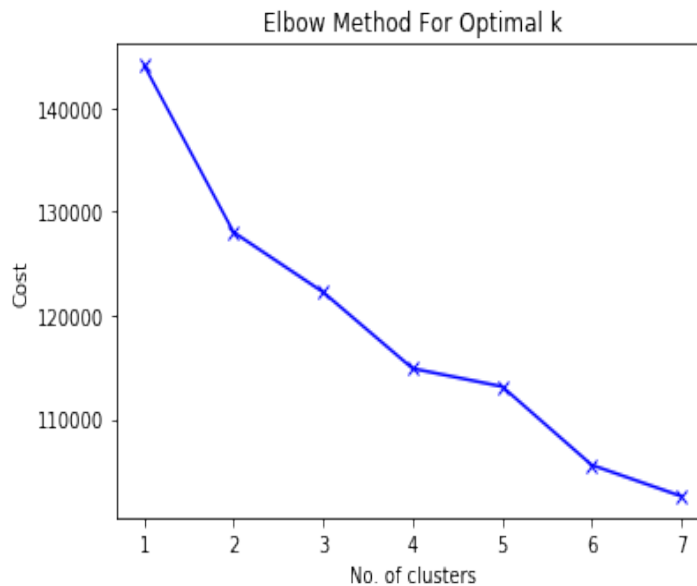


Figure 4.15: Optimal K values for Clusters

4.3.7 Experimental Result with KNN

K-nearest neighbours is used using the sklearn library. The Knn is trained iteratively such that the value is ranged from 1 to 21, such that 20 iterations occur and after each iteration the model test accuracy improves. The graph 4.16 below shows the Accuracy for different K values on each iteration.

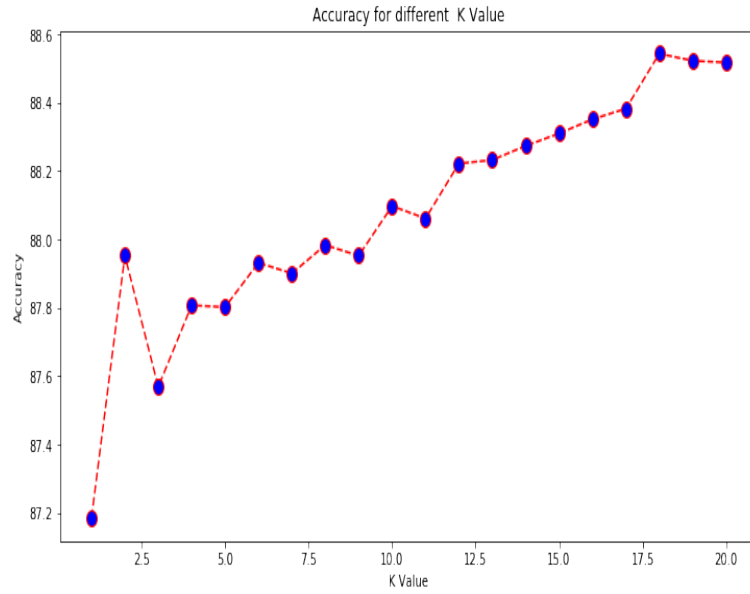


Figure 4.16: Accuracy plot for different values of K-NN

For K-NN, the train test split is set to 65 % and 35%. At every iteration the the training accuracy decreases and testing accuracy improves. By the last iteration, the KNN training accuracy obtained is 96% while KNN test accuracy is 88%. The classification report is given below 4.17 :

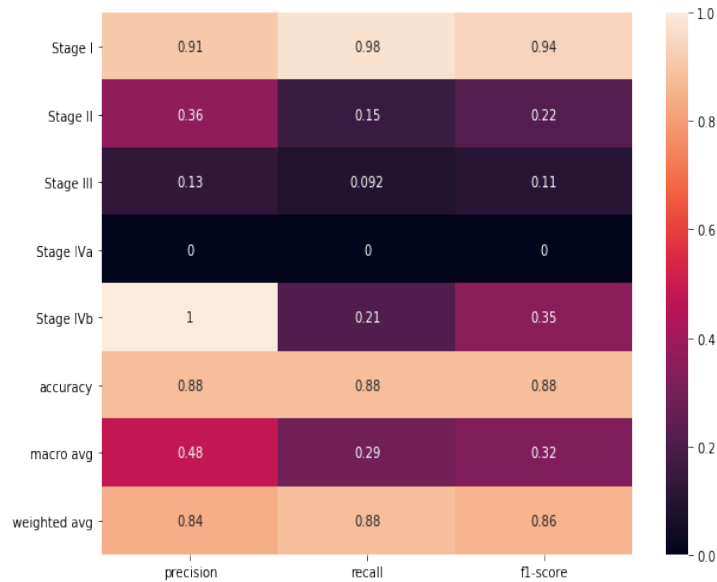


Figure 4.17: Classification-Report K-Nearest Neighbor

4.3.8 Experimental Result with Light Gradient Boosting Machine

LightGBM, was trained not with its own parameters but with the grid search cv shown in figure 4.18. It could handle the censored data easily. LightGBM stage predictions on the created prognostic combination of TNMA was positively co-related to the AJCC manual staging.

```

parameters = {
    'n_estimators': [5, 10, 15, 20, 25, 50, 100],
    'learning_rate': [0.01, 0.05, 0.1],
    'num_leaves': [7, 15, 31],
}

model_lgbm = lgbm.LGBMClassifier(
    random_state=42,
    class_weight='balanced',
)

model_lgbm = GridSearchCV(
    model_lgbm,
    parameters,
    cv=5)

```

Figure 4.18: Hyper-parameters for Light Gradient Boosting

The classification report for LightGBM is given below in figure 4.19. The model when tuned on hyper-parameters and balanced using Smote gave a higher accuracy of 0.91.

```

print(classification_report(model_lgbm_pred, y_test))

```

	precision	recall	f1-score	support
1	1.00	1.00	1.00	9630
2	0.91	1.00	0.95	8744
3	0.82	0.92	0.87	8675
4	0.83	0.86	0.84	9448
5	0.93	0.75	0.83	12196
accuracy			0.90	48693
macro avg	0.90	0.91	0.90	48693
weighted avg	0.90	0.90	0.90	48693

Figure 4.19: Classification-Report Light Gradient Boosting

4.3.9 Stage Predictions

After applying different Machine learning models, it became evident that not all of them are able to handle the class imbalance even if techniques to handle the imbalance data class are applied on the data-set. The models were tried one after the other, in order to analyze the performance of the specific model with dataset.

Since the dataset is challenging as not every patient is terminal and in critical stages or dead. This is the reason of large difference in the output Stage classes. As seen in the classification reports of the models, the individual classes precision recall values are varying prominently. Specifically in Stage IVA, the values are decreased to minimal which is happening because of the least number of patients in there terminal stage. Thyroid cancer already has a higher mortality, so having patients in terminal stage is a rare case.

This was a challenge during this research, handling it carefully and not skewing the results was the main concern since its a clinical data so no changes are acceptable. Out of all only LightGBM performed exceptionally, handling the imbalance in the classes and giving closely relatable results. It did not get effected by the minor classes with less patients. The precision, recall values for individual class are efficient and reliable outcomes. Even for Stage IVA it gave 0.83 and 0.86 precision and recall respectively.

For a better understanding, the predicted stages for the data are compared with the standard for cancer staging. This given table 4.1 represents the distribution of patients against the given TNMA values according to the manual AJCC staging, over the 5 prognostic groups by LightGBM. We can see that LightGBM assigned higher risk values in the last groups which means higher the risk higher the group stage, very similar to the AJCC stages. Although, exceptions exist, but those occurred due to the invalid survival rates estimations given in the data.

For example, a 5 year survival rate for T4bN1aM0A1 with approximately 100 patients can not be 94%, which means a higher risk group can not have a higher survival potency. But it is not a deniable fact that by using more large sample datasets with censorship handling techniques, with such robust model Cancer prognosis can be aided well.

Table 4.1: LightGBM and AJCC Grouping on thyroid cancer dataset generated from the SEER Database of thyroid cancer (papillary and follicular)

T	N	M	A	LightGBM Prognostic Groups	AJCC Manual Staging groups
T1	N0	M0	A1	1	I
T1	N0	M0	A2	1	I
T1	N0	M0	A2	1	I
T1	N1b	M0	A1	1	I
T2	N0	M0	A1	1	I
T2	N0	M0	A2	1	I
T2	N1a	M0	A1	1	I
T2	N1b	M0	A1	1	I
T3	N0	M0	A1	1	I
T3	N1a	M0	A1	1	I
T3	N1b	M0	A1	1	I
T4a	N0	M0	A1	1	I
T4a	N1a	M0	A1	1	I
T4a	N1b	M0	A1	1	I
T4b	N0	M0	A1	1	I
T4b	N1b	M0	A1	1	I
T1	N1a	M0	A2	2	II
T1	N1b	M0	A2	2	II
T2	N1a	M0	A2	2	II
T3	N0	M0	A2	2	II
T3	N1a	M0	A2	2	II
T4b	N1a	M0	A1	2	I
T2	N1b	M0	A2	3	II
T4a	N0	M0	A2	3	III
T2	N0	M1	A2	3	IVB
T4b	N0	M0	A2	3	IVA
T1	N0	M1	A2	3	IVB
T3	N0	M1	A2	3	IVB
T3	N1b	M0	A2	4	II
T3	N1b	M1	A1	4	II
T4a	N1a	M0	A2	4	III
T4b	N1a	M0	A2	4	IVA
T3	N1b	M1	A2	4	IVB
T4b	N0	M1	A2	4	IVB
T4a	N1b	M0	A2	5	III
T4a	N1b	M1	A1	5	II
T4a	N1b	M1	A2	5	IVB
T4b	N1b	M0	A2	5	IVA
T4b	N1b	M1	A2	5	IVB

4.3.10 Comparison of Machine Learning Classifiers

The table 4.2 presents the detail comparison of the models on the basis of the evaluation metrics. Though not a needed in multi-class scenarios, but AUC-ROC score was calculated with the one-to-rest and one-to-one multi-class parameter.

Table 4.2: Results of different Models on Thyroid Dataset for Stage Prediction

Classifier	Accuracy	Precision	Recall	F1 Measure	AUC_ROC
Decision Tree	88 %	0.84	0.88	0.85	0.68
Gaussian Naïve Bayes	89 %	0.83	0.89	0.85	0.85
SVM (poly)	88.60 %	0.81	0.88	0.83	0.71
SVM (rbf)	87 %	0.83	0.87	0.85	0.88
Random Forest (PCA)	81 %	0.82	0.81	0.82	0.69
Random Forest (SMOTE)	91 %	0.92	0.91	0.91	0.98
KNN	88.5 %	0.84	0.88	0.86	0.78
Adaboost (DT as base)	96 %	0.95	0.96	0.95	0.95
Adaboost	88.47 %	0.78	0.88	0.83	0.77
LightGBM	91 %	0.90	0.90	0.90	0.84

4.3.11 Survival Analysis

The chance of patients living to a given time point t following diagnosis is referred to as overall survival. The survival for the thyroid data for 5-year (months) with respect to the vital status of the patients i.e. alive or death is calculated using Kaplan Meier as given below in the figure 4.20. Where the survival time is given in months for 5-years. The figure shows the overall survival curve for the thyroid data. Due to the lower ratio of death, the curve maintains its drop. Thyroid cancer usually has the higher survival rate in 5-year or 10-year time span. Hence, the estimated survival is near the actual estimations.

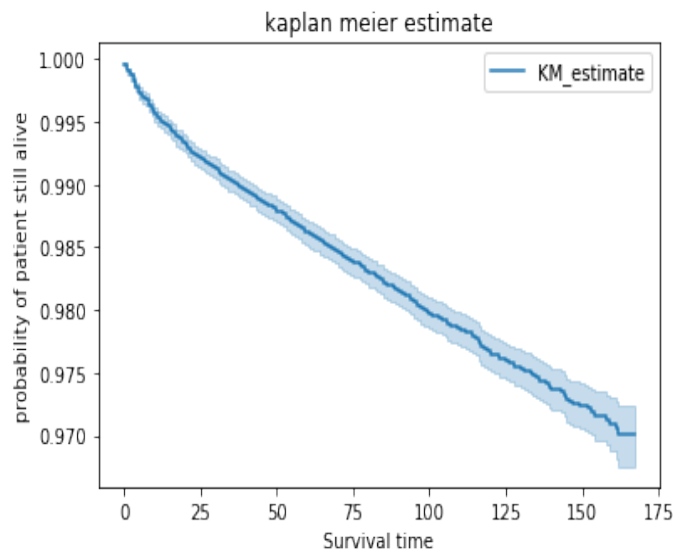


Figure 4.20: Kaplan Meier Estimate for 5-year on Thyroid Cancer Data

Cancer Stages:

Each Thyroid Cancer stage survival probability is generated and the survival rate is compared with the other stages respectively. The figure 4.21 below shows the comparison of **Stage I** and **Stage II** of Thyroid cancer. Since both the stages are having max incidences, it is important to see the difference in their survival rates. The survival time given on the x-axis is the time in months, showing the survival time in form of months for the period of 5-years, from 0 to 160 months.

It is seen in the graph that both stages have higher probability of survival over the

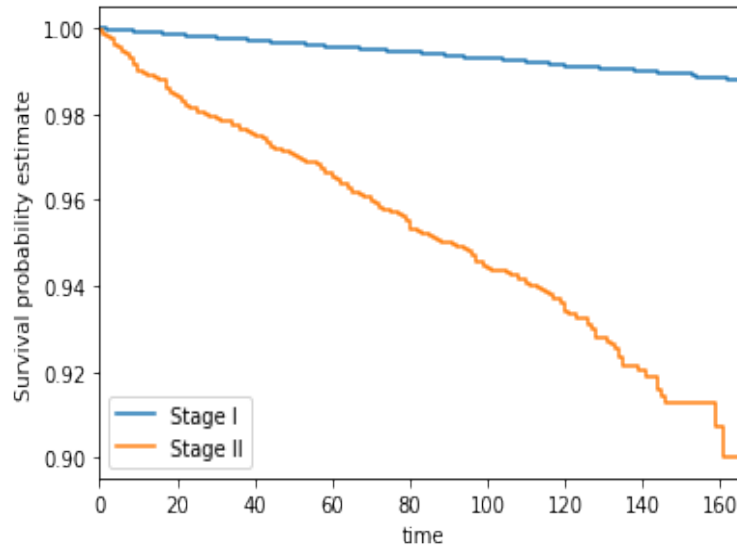


Figure 4.21: Kaplan Meier curve for Thyroid Cancer Stage I and II

years (months). That is why Thyroid cancer when in its early stage needs to be treated immediately else it may worsen even if the survival chance is higher. It is noticeable that Stage I survival probability lies between 1 to 0.98 which is a very high survival probability and explains a lot about the chances that might be available for the Stage I patient, by looking at the survival situation. Whereas the Stage II has a sort of straight curve downwards, although the survival probability is still high which lies between 1 to 0.90.

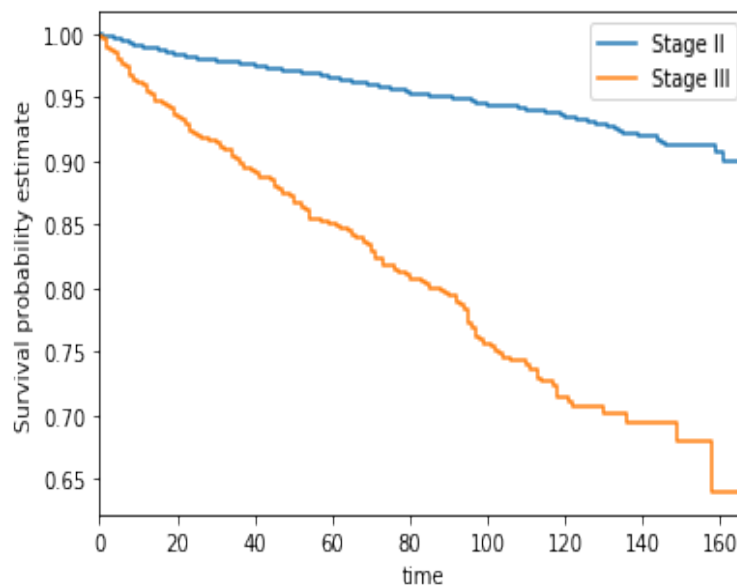


Figure 4.22: Kaplan Meier curve for Thyroid Cancer Stage II and III

The figure 4.22 presented above, represent **Stage II** and **Stage III** survival curves using Kaplan Meier. The graph depicts the particular stage patient survival over 5-years. The years represent time, which are given as months in the graph. As seen in the graph above, that stage III faces a sudden fall, and its survival probability experiences a lower survival rate which lies between 1 to 0.65. This drop of survival explains a lot about the condition of patient as the cancer reaches this stage. When we compare the these two Stages, the difference in the survival curve is much evident.

The figure 4.23 present the survival probability estimation and comparative survival curves for **Stage III** and **Stage IVA**. The two stages have a survival decrement but the Stage IVA (encoded as 4) has a very prominently low survival rate.

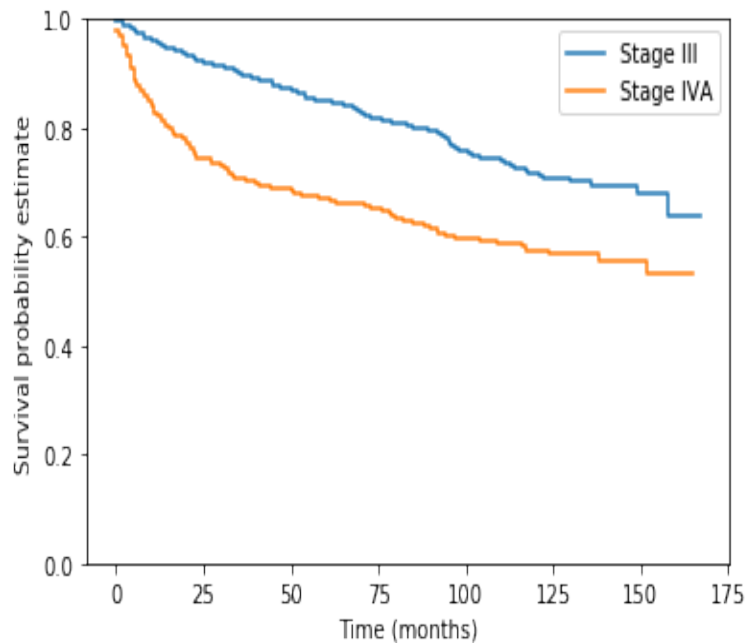


Figure 4.23: Kaplan Meier curve for Thyroid Cancer Stage III and IVA

There is a sudden down in the curve for Stage IVA. The curve is not a straight line decrement rather its more of an inward curve. This styled curve explains that there are a number of patients of thyroid cancer at much lower survival months. The survival probability lies between 0.97 to 0.45, which as compared to Stage III is less.

The next figure 4.24 gives the survival comparison curve and probability for the **Stage IVA** and **Stage IVB**. Those Cancer Stages which are considered as the last Stages for cancer, when the survival chances are narrowed down to 0.

Stage IVA has a significantly higher survival rate than Stage IVB. Although, both of

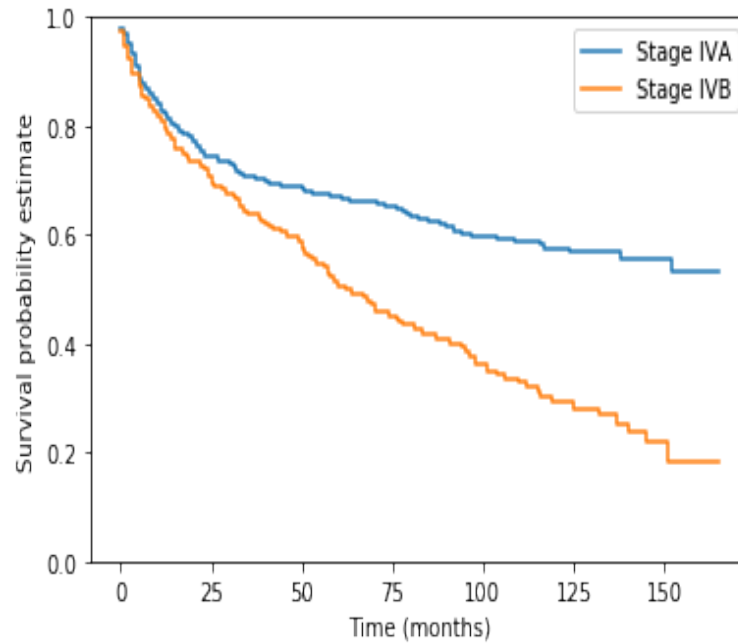


Figure 4.24: Kaplan Meier curve for Thyroid Cancer Stage IVA and IVB

them are critical stages but out of these two Stage IVB is the most deadly stage for any cancer patient. Stage IVA has a lower mortality rate than Stage IVB. Its least survival probability is lying near 0.65 while Stage IVB survival probability is below 0.2, which is not a survival rate by which a patient could survive over years.

Comparison:

The figure 4.25 shows all the stages in comparison with respect to others. The survival decrement in the last stages shows the situation of patients in those stages. Creating survival curves for the Stages allows us to develop a proactive plan for high-value patients for various survival risk segments along the timeline over months.

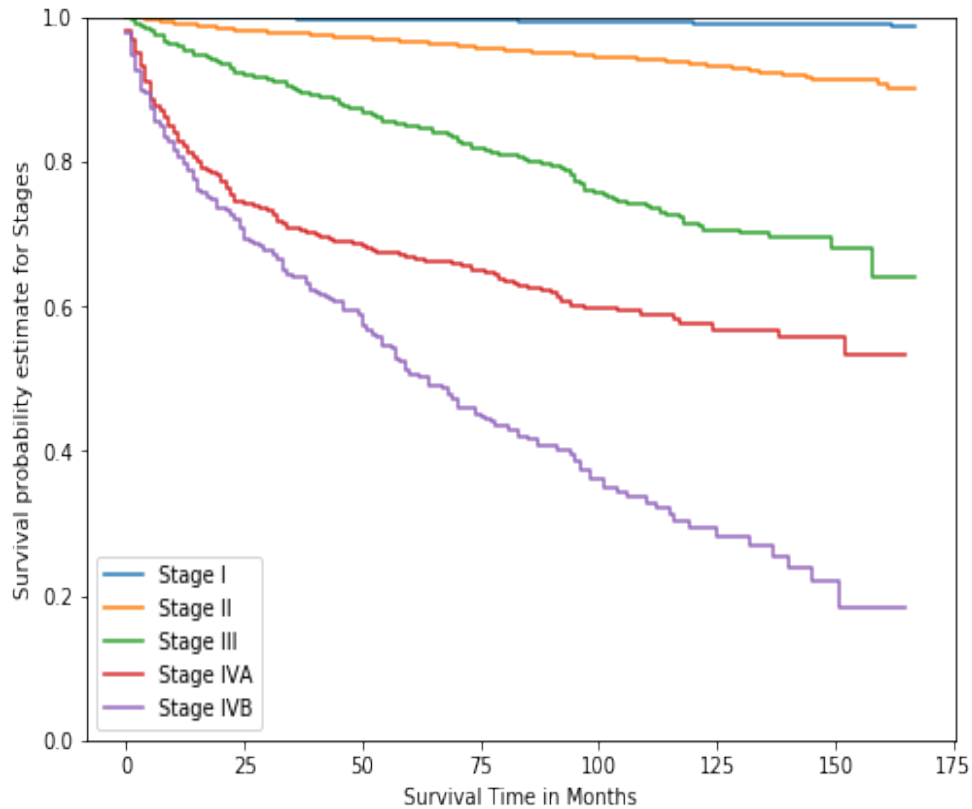


Figure 4.25: Kaplan Meier curves for Thyroid Cancer Stages

Impact of Age:

With respect to age the survival rate varies for cancer. The age was grouped into two groups as A1 (0-55) and A2 (55- 85+), for each group the survival rate is calculated through Kaplan Meier, shown in the figure below [4.26](#):

As the figure explains that lower is the age, higher is the survival probability over time. Similarly, higher is the age, lower will be the survival probability.

But here a point of concern lies, as if we see the count of patients of Thyroid Cancer with respect to age, It is seen that maximum number of patients are from the age group of A1 (0-55). This and the Kaplan Meier curves for age explains the fact that why the mortality rate in Thyroid cancer is low. On the other hand it also explains the fact that majority of young adults are prey to this disease.

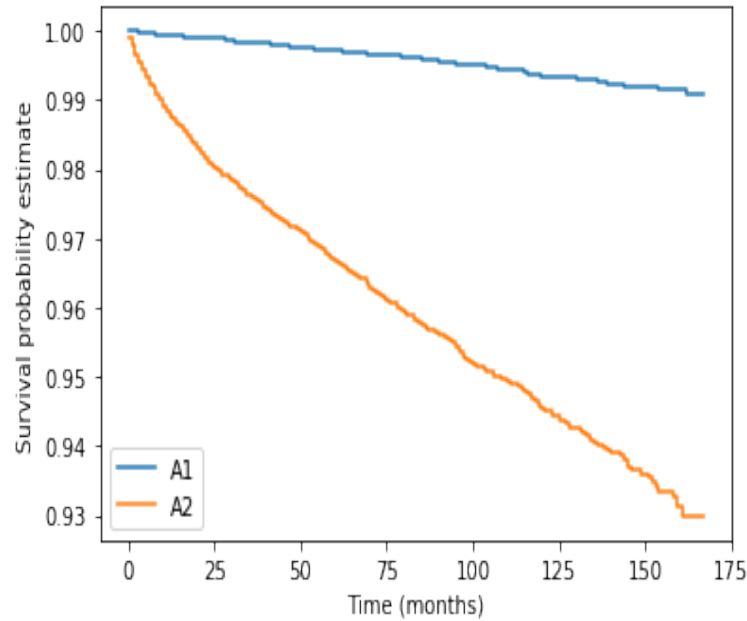


Figure 4.26: Kaplan Meier curve for Thyroid Cancer With Respect to Age

4.4 Summary

This chapter covers the implemented model's results and their evaluation. Data analysis is performed which is necessary to understand the data. Later the Experimental Analysis takes place that all the models implemented are evaluated and results are obtained. The ML models decision tree, naive bayes, SVM, random forest, adaboost, lightGBM, KNN and K-mode results are demonstrated in the form of a classification report. The survival analysis is conducted with thyroid stages and age. The analysis is discussed in detail. The next chapter concludes the thesis research work and presents the future goals.

Future Work and Conclusion

5.1 Conclusion

Thyroid cancer, the most frequent malignancy growing globally over the past few decades. Soon to be counted as the fourth most widely spread cancer. It is on the increase all over the world, particularly in developing countries.

In this thesis work, The main goal was to use the machine learning algorithm models that could predict thyroid cancer Stages and provide relevant information for clinical decision-making. And to determine the survival of the cancer over the specific period of time.

In order to identify distinct cancer stages, multiple machine learning algorithms are applied. The data of patients diagnosed with thyroid cancer was fetched from the SEER repository and passed through multiple pre-processing steps. Feature selection was performed such that only important features are included for the experiment rest were removed. Some of the main features included the tumour size (T), nodes (N), age, metastasis (M), survival time, cause of death and the year of diagnosis.

In this thesis, we have created a prognostic system using the T, N, M and A features by merging them and making a combination of TNMA. Using them, the work on the Stage prediction of the Thyroid Cancer is proceeded. Using the SEER database, it was made possible to create the prognostic data model of TNMA, which was built on AJCC staging tumor (T), nodes (N), metastasis (M) and age (A). The research's major purpose is to use idea of TNMA combinations and do a predictive analysis of dataset in order to determine the cancer stage and the impact on the performance of the models when this

combination data model is used.

There must be a target variable for every prediction outcome. In this case, cancer Stages were the target variable. When there are balanced classes, machine learning models perform better; otherwise, they tend to prefer the majority class. The class balancing approaches like oversampling and undersampling were used to eliminate the bias. The sample strategies have been demonstrated to not only balance the classes, but also to increase the model's performance. Various machine learning algorithms are utilised to classify the dataset and have a prediction algorithm. SVM, Decision tree, Random Forest, KNN, Naive bayes, AdaBoost, clustering and Gradient Boosting Algorithms are implemented to have predictive outcomes with best performance.

Results show that the LightGBM and Adaboost are the prominent ones with higher accuracy measure of 91% and 96% respectively. Though, Random Forest gave 91% accuracy when over-sampling technique was applied. Considering other evaluation measure of precision the values of LightGBM i.e 0.90 and AdaBoost i.e 0.95 are better than all other classifiers. LightGBM was able to handle such time-to-event data even with censorship of dataset. The imbalance of classes, representing the lesser number of patients in the terminal stages was well handled by the model and gave efficiently higher results of precision, recall and F1 measure for individual stage. The model was robust and generated results more similar to the AJCC manual grouping.

For Survival analysis, the Kaplan Meier was used and showed the survival rate for each stage. It is examined through the analysis that how much the Stages effect the survival rate of the patient. Specifically, it is to compare the overall survival of the whole dataset to that of the Stages of cancer. Since the thyroid mortality rate is low, so the overall survival of thyroid cancer is comparatively high, so the survival for 5 year time to event data was found to be within 1 to 0.9. While the survival probability for different stages showed that stage I has highest survival probability where Stage IVA and Stage IVB are with least values of 0.65 and 0.2 survival probability estimates.

This work not only helps in understanding the cancer current condition but also aids in treatment decision making and cancer prognosis.

5.2 Future Work

For future work, we are intended advance this research work to make it more understandable and effective. It will not only deal with the censoring of data but also aid in more accurate time-to-event predictions. Implementing the deep learning and Artificial Intelligence techniques for prediction of cancer survivability and extending this work to other cancer types. Since the results are promising, developing this project into an application so that it can be used in point of care at a hospital organization. The Future work in this project can be done in the following areas: Exploring the limitations of the project and develop new strategies. Exploring other methodologies and sampling techniques to improve the performance of the model and its behavior.

Bibliography

- [1] Munira Ferdous, Jui Debnath, and Narayan Ranjan Chakraborty. “Machine Learning Algorithms in Healthcare: A Literature Survey”. In: *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2020, pp. 1–6. DOI: [10.1109/ICCCNT49239.2020.9225642](https://doi.org/10.1109/ICCCNT49239.2020.9225642).
- [2] Ajay Kumar, Rama Sushil, and Arvind Kumar Tiwari. “Machine Learning Based Approaches for Cancer Prediction: A Survey”. In: *Digital Health eJournal* (2019).
- [3] Max Roser and Hannah Ritchie. “Cancer”. In: *Our World in Data* (2015). URL: <https://ourworldindata.org/cancer>.
- [4] Olayinka S Ilesanmi and S Ayanleke. “ep rin t n ot Pr pe er re vie we Pr ep ot t n pe er re vie we”. In: *Tropical Journal of Pharmaceutical Research* 18.12 (2019), pp. 2679–2686.
- [5] *GLOBOCAN*. URL: <https://gco.iarc.fr/>.
- [6] Rene Y. Choi et al. “Introduction to machine learning, neural networks, and deep learning”. In: *Translational Vision Science and Technology* 9.2 (2020), pp. 1–12. ISSN: 21642591. DOI: [10.1167/tvst.9.2.14](https://doi.org/10.1167/tvst.9.2.14).
- [7] Michael K. K. Leung et al. “Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets”. In: *Proceedings of the IEEE* 104.1 (2016), pp. 176–197. DOI: [10.1109/JPROC.2015.2494198](https://doi.org/10.1109/JPROC.2015.2494198).
- [8] Stefan Hegselmann et al. “Reproducible Survival Prediction with SEER Cancer Data”. In: *Proceedings of Machine Learning Research* 85.2017 (2018), pp. 49–66. ISSN: 2640-3498. URL: <https://proceedings.mlr.press/v85/hegselmann18a.html>.

BIBLIOGRAPHY

- [9] Xinyu Zhang et al. “Multi-channel convolutional neural network architectures for thyroid cancer detection”. In: *PLOS ONE* 17.1 (Jan. 2022), pp. 1–26. DOI: [10.1371/journal.pone.0262128](https://doi.org/10.1371/journal.pone.0262128). URL: <https://doi.org/10.1371/journal.pone.0262128>.
- [10] Dorota Dworakowska and Ashley B. Grossman. “Thyroid disease in the time of COVID-19”. In: *Endocrine* 68.3 (2020), pp. 471–474. ISSN: 15590100. DOI: [10.1007/s12020-020-02364-8](https://doi.org/10.1007/s12020-020-02364-8). URL: <http://dx.doi.org/10.1007/s12020-020-02364-8>.
- [11] Shikha Roy et al. “Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning”. In: *Scientific Reports* 10.1 (2020), pp. 1–15. ISSN: 20452322. DOI: [10.1038/s41598-020-60740-w](https://doi.org/10.1038/s41598-020-60740-w).
- [12] “How does the thyroid gland work?” In: *Institute for Quality and Efficiency in Health Care (IQWiG)*. Cologne, Germany: Institute for Quality and Efficiency in Health Care. URL: <https://www.ncbi.nlm.nih.gov/books/NBK279388/>.
- [13] Sharma S. Shahid MA Ashraf MA. “Physiology, Thyroid Hormone.” In: (2022). URL: <https://www.ncbi.nlm.nih.gov/books/NBK500006>.
- [14] Mark P.J. Vanderpump. “The epidemiology of thyroid disease”. In: *British Medical Bulletin* 99.1 (2011), pp. 39–51. ISSN: 00071420. DOI: [10.1093/bmb/ldr030](https://doi.org/10.1093/bmb/ldr030).
- [15] American Cancer Society. “What is Thyroid Cancer”. In: *Who* (2021), pp. 1–11. URL: <http://www.cancer.org/cancer/thyroidcancer/detailedguide/thyroid-cancer-what-is-thyroid-cancer>.
- [16] Quang T. Nguyen et al. “Diagnosis and treatment of patients with thyroid cancer”. In: *American Health and Drug Benefits* 8.1 (2015), pp. 30–38. ISSN: 19422970.
- [17] Yuejun Liu et al. “A Study on the Auxiliary Diagnosis of Thyroid Disease Images Based on Multiple Dimensional Deep Learning Algorithms”. In: *Current Medical Imaging Formerly Current Medical Imaging Reviews* 16.3 (Mar. 2020), pp. 199–205. ISSN: 15734056. DOI: [10.2174/1573405615666190115155223](https://doi.org/10.2174/1573405615666190115155223). URL: <http://www.eurekaselect.com/169020/article>.

BIBLIOGRAPHY

- [18] Shanu Verma, Rashmi Popli, and Harish Kumar. “Study of Thyroid Disease Using Machine Learning”. In: *Advanced Healthcare Systems*. Wiley, Feb. 2022, pp. 33–42. DOI: [10.1002/9781119769293.ch3](https://doi.org/10.1002/9781119769293.ch3). URL: <https://onlinelibrary.wiley.com/doi/10.1002/9781119769293.ch3>.
- [19] Nan Miles, Lin Wang, and Chuanjia Yang. “Improving The Diagnosis of Thyroid Cancer by Machine Learning and Clinical Data”. In: (2022). URL: <https://doi.org/10.48550/arXiv.2203.15804>.
- [20] American Cancer Society. “If You Have Cancer”. In: (2019), pp. 1–11. URL: <https://www.cancer.org/cancer/cancer-basics/if-you-have-cancer.html>.
- [21] Wenfei Liu et al. “A Proposed Heterogeneous Ensemble Algorithm Model for Predicting Central Lymph Node Metastasis in Papillary Thyroid Cancer”. In: *International Journal of General Medicine* Volume 15 (May 2022), pp. 4717–4732. ISSN: 1178-7074. DOI: [10.2147/IJGM.S365725](https://doi.org/10.2147/IJGM.S365725). URL: <https://www.dovepress.com/a-proposed-heterogeneous-ensemble-algorithm-model-for-predicting-centr-peer-reviewed-fulltext-article-IJGM>.
- [22] Uchechukwu C. Megwalu and Peter K. Moon. “Thyroid Cancer Incidence and Mortality Trends in the United States: 2000–2018”. In: *Thyroid* 32.5 (May 2022), pp. 560–570. ISSN: 1050-7256. DOI: [10.1089/thy.2021.0662](https://doi.org/10.1089/thy.2021.0662). URL: <https://www.liebertpub.com/doi/10.1089/thy.2021.0662>.
- [23] *Cancer Statistics Center*. URL: <https://cancerstatisticscenter.cancer.org/#/>.
- [24] D. Peraković A. Khan, K. T. Chui. “Future Scope of Machine Learning and AI in 2022”. In: *Insights2Techinfo* (2021). URL: <https://insights2techinfo.com/future-scope-of-machine-learning-and-ai-in-2022/>.
- [25] Konstantina Kourou et al. “Machine learning applications in cancer prognosis and prediction”. In: *CSBJ* 13 (2015), pp. 8–17. ISSN: 2001-0370. DOI: [10.1016/j.csbj.2014.11.005](https://doi.org/10.1016/j.csbj.2014.11.005). URL: <http://dx.doi.org/10.1016/j.csbj.2014.11.005>.
- [26] Leili Tapak et al. “Prediction of survival and metastasis in breast cancer patients using machine learning classifiers”. In: *Clinical Epidemiology and Global Health* 7.3 (2019), pp. 293–299. ISSN: 22133984. DOI: [10.1016/j.cegh.2018.10.003](https://doi.org/10.1016/j.cegh.2018.10.003). URL: <https://doi.org/10.1016/j.cegh.2018.10.003>.

BIBLIOGRAPHY

- [27] Pushpanjali Gupta, Sum-fu Chiang, and Prasan Kumar Sahoo. “Prediction of Colon Cancer Stages and Survival”. In: *Cancers* (2019), pp. 1–16.
- [28] Sung Mo Ryu, Sung Wook Seo, and Sun Ho Lee. “Novel prognostication of patients with spinal and pelvic chondrosarcoma using deep survival neural networks”. In: *BMC Medical Informatics and Decision Making* 20.1 (2020), pp. 1–10. ISSN: 14726947. DOI: [10.1186/s12911-019-1008-4](https://doi.org/10.1186/s12911-019-1008-4).
- [29] Mi Du et al. “Comparison of the tree-based machine learning algorithms to cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on seer database”. In: *Cancers* 12.10 (2020), pp. 1–16. ISSN: 20726694. DOI: [10.3390/cancers12102802](https://doi.org/10.3390/cancers12102802).
- [30] Pei Liu et al. “Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer”. In: *IEEE Transactions on Biomedical Engineering* 68.1 (2021), pp. 148–160. ISSN: 15582531. DOI: [10.1109/TBME.2020.2993278](https://doi.org/10.1109/TBME.2020.2993278).
- [31] Youness Khourdifi. “Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification”. In: *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)* (2018), pp. 1–5.
- [32] Farhad Imani et al. “Random Forest Modeling for Survival Analysis of Cancer Recurrences”. In: *IEEE International Conference on Automation Science and Engineering* 2019-August (2019), pp. 399–404. ISSN: 21618089. DOI: [10.1109/COASE.2019.8843271](https://doi.org/10.1109/COASE.2019.8843271).
- [33] Xinyu Zhang et al. “Multi-channel convolutional neural network architectures for thyroid cancer detection”. In: *PLOS ONE* 17.1 (Jan. 2022). Ed. by Le Hoang Son, e0262128. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0262128](https://doi.org/10.1371/journal.pone.0262128). URL: <https://dx.plos.org/10.1371/journal.pone.0262128>.
- [34] Claudio Casella et al. “The New TNM Staging System for Thyroid Cancer and the Risk of Disease Downstaging”. In: *Frontiers in Endocrinology* 9 (2018). ISSN: 1664-2392. DOI: [10.3389/fendo.2018.00541](https://doi.org/10.3389/fendo.2018.00541). URL: <https://www.frontiersin.org/article/10.3389/fendo.2018.00541>.
- [35] Wenfei Liu et al. “Prediction of lung metastases in thyroid cancer using machine learning based on SEER database”. In: *Cancer Medicine* December 2021 (2022), pp. 1–13. ISSN: 20457634. DOI: [10.1002/cam4.4617](https://doi.org/10.1002/cam4.4617).

BIBLIOGRAPHY

- [36] “Adjusted AJCC 6th ed. T, N, M, and Stage”. In: (2019). URL: <https://seer.cancer.gov/seerstat/variables/seer/ajcc-stage/6th/>.
- [37] “SEER Research Data Record Description”. In: (2019). URL: <https://seer.cancer.gov/data/seerstat/nov2017/TextData.FileDescription.%20pdf>.
- [38] Medical and The American Cancer Society editorial content Team. “Thyroid Cancer Stages”. In: (2019). URL: <https://www.cancer.org/cancer/thyroid-cancer/detection-diagnosis-staging/staging.html>.
- [39] T. L. Octaviani and Z. Rustam. “Random forest for breast cancer prediction”. In: *AIP Conference Proceedings* 2168.1 (2019), p. 020050. DOI: [10.1063/1.5132477](https://doi.org/10.1063/1.5132477). eprint: <https://aip.scitation.org/doi/pdf/10.1063/1.5132477>. URL: <https://aip.scitation.org/doi/abs/10.1063/1.5132477>.
- [40] Rezvan Ehsani and Finn Drabløs. “Robust Distance Measures for k NN Classification of Cancer Data”. In: *Cancer Informatics* 19 (Jan. 2020), p. 117693512096554. ISSN: 1176-9351. DOI: [10.1177/1176935120965542](https://doi.org/10.1177/1176935120965542). URL: <http://journals.sagepub.com/doi/10.1177/1176935120965542>.
- [41] Jin Hee Bae et al. “Feature Selection for Colon Cancer Detection Using K-Means Clustering and Modified Harmony Search Algorithm”. In: *Mathematics* 9.5 (2021). ISSN: 2227-7390. DOI: [10.3390/math9050570](https://doi.org/10.3390/math9050570). URL: <https://www.mdpi.com/2227-7390/9/5/570>.
- [42] Fei Li et al. “A Light Gradient Boosting Machine for Remaining Useful Life Estimation of Aircraft Engines”. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 2018, pp. 3562–3567. DOI: [10.1109/ITSC.2018.8569801](https://doi.org/10.1109/ITSC.2018.8569801).
- [43] Guolin Ke et al. “LightGBM: A highly efficient gradient boosting decision tree”. In: *Advances in Neural Information Processing Systems 2017-Decem.Nips* (2017), pp. 3147–3155. ISSN: 10495258.