Available online at www.ijournalse.org

**Emerging Science Journal**

(ISSN: 2610-9182)

Vol. 7, No. 4, August, 2023

# A Binary Survivability Prediction Classification Model towards Understanding of Osteosarcoma Prognosis

Saravanan Muthaiyah [1*], Vivek Ajit Singh [2], Thein Oak Kyaw Zaw [1], Kalaiarasi S. M. Anbananthen [3], Byeonghwa Park [4], Myung Joon Kim [5]

[1] *Faculty of Management,Multimedia University, Selangor, Malaysia.*

[2] *Faculty of Medicine, Universiti Malaya, Selangor, Malaysia.*

[3] *Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia.*

[4] *Department of Management and Marketing, Valdosta State University, Georgia, United States.*

[5] *Department of Big Data Application, Hannam University, Daejeon, South Korea.*

**Abstract**

The objective of this study is to explore effective and innovative machine learning techniques that can assist medical professionals in developing more accurate prognoses that can enhance the survivability of osteosarcoma patients by investigating potential prognostic factors and identifying novel therapeutic approaches. A comprehensive analysis was conducted using a dataset of 128 osteosarcoma patients between 1997 to 2011. The dataset included 52 attributes in total that covered a wide range of demographics, together with information on clinical records, treatment protocols, and survival outcomes. Data was obtained from NOCERAL (National Orthopaedic Centre of Excellence in Research and Learning), Kuala Lumpur. Three distinct binary classification methods (i.e., random forest, support vector machine (SVM), and artificial neural network (ANN)) were employed to identify the prognostic factors that are associated with improved survival efficacy measures. The results of this study revealed that both SVM and ANN outperformed random forests in predicting survivability for both the 2-year and 5-year time frames. These findings indicate the potential of SVM and ANN as effective tools for predicting osteosarcoma survivability. The study signifies a significant step towards integrating machine learning techniques into the existing toolkit available to medical practitioners. This study contributes to the medical field by providing a comparative analysis of three prominent machine learning techniques for predicting osteosarcoma survivability. The superior performance of SVM and ANN over random forests highlights the potential of these methods in generating more accurate survivability predictions. Further development and refinement of these machine learning techniques hold promise for enhancing their effectiveness and instilling greater confidence among medical professionals and patients in the predictive capabilities of machine learning and artificial intelligence models for osteosarcoma survivability.

**Keywords:**
Osteosarcoma;
Survivability;
Prognosis;
Machine Learning.

**Article History:**

| | | | |
|---|---|---|---|
| **Received:** | 13 | February | 2023 |
| **Revised:** | 09 | June | 2023 |
| **Accepted:** | 02 | July | 2023 |
| **Available online:** | 12 | July | 2023 |

## 1- Introduction

Prognosis refers to the likelihood of recovering from a disease or the expected outcome of a treatment process [1]. It plays a crucial role in the healthcare industry, influencing medical decision-making and even affecting patients' decisions in other areas of their lives unrelated to medical care. A positive prognosis can instill optimism in patients, alleviating concerns about the disease and treatment. Conversely, a less favorable prognosis allows patients to make informed decisions about continuing treatment while preparing for the inevitable [2]. By equipping medical professionals with tools to enhance the accuracy and quality of prognoses, they can provide optimal care and treatment for their patients.

* **CONTACT**: saravanan.muthaiyah@mmu.edu.my

**DOI:** http://dx.doi.org/10.28991/ESJ-2023-07-04-018

These tools also contribute to achieving the United Nations' Sustainable Development Goal of ensuring healthy lives and promoting well-being [3]. While accurate prognoses of suboptimal outcomes may seem challenging, they empower patients to come to terms with their mortality and prepare for the future, benefiting the well-being of individuals in their social circle.

Therefore, the development of essential technological tools for more accurate prognoses is crucial to providing the best possible medical guidance to patients worldwide. As professionals who serve on the front lines, doctors are responsible for providing the best possible care and advice to their patients. This is especially important for osteosarcoma, a type of cancer that can also be detected in younger patients [4] who are only beginning to find their footing in life. On top of that, current treatments for osteosarcoma, such as surgery, radiation therapy, and chemotherapy, are done on a median approach as there are insufficient insights to determine the most accurate form of treatment needed for each individual. As such, significant progress needs to be made in developing the necessary tools that can assist medical professionals in providing more accurate prognoses to osteosarcoma patients. By providing a more accurate prognosis, doctors can provide the most optimal course of action to an osteosarcoma patient, and in the event of a poor prognosis, the patient can opt to forgo treatment and save themselves unnecessary time and money that would have been put towards an ineffective treatment process.

This research helps to achieve the goal of providing accurate prognoses by investigating machine learning techniques that can process vast amounts of patient data and construct an accurate prognostic model in a short period. From a theoretical point of view, there is currently a limited amount of research performed on the application of machine learning techniques to osteosarcoma prognoses, which may be due to osteosarcoma being less prevalent in the population than other types of cancers. However, there are many examples of the research being performed on other cancer types, such as breast cancer [5-7], which have been proven to demonstrate the effectiveness of machine learning techniques for developing prognostic models and can be drawn upon to supplement this research. Therefore, this study aims to discover a machine-learning technique that can assist medical professionals in the prognosis of osteosarcoma with high effectiveness.

### *1-1- Objective of the Study*

Based on the issues discussed earlier, this research was established with the objective of discovering the most effective machine learning technique that can assist medical professionals in developing accurate prognoses for osteosarcoma patients. To understand the context of the research problem and subsequently achieve this objective, the following research questions were formulated.

1) What is the current extent of osteosarcoma survivability among those who are afflicted with the disease?
2) What are the factors that influence the prediction of the survivability of osteosarcoma patients?
3) What is the effectiveness of existing machine learning techniques in assisting medical professionals with providing an accurate prognosis for osteosarcoma patients?

## 2- Literature Review

Osteosarcoma is a type of cancerous bone tumor that develops in areas of rapid bone turnover, occurring in the long bones of the limbs near the metaphyseal growth plate [8]. The most frequent sites for osteosarcoma occurrence are situated in the distal femur and proximal tibia of adolescent humans [9], with less common occurrences in the skull, jaw, or pelvis [8]. This is the most common form of primary bone malignant tumor found in humans [8] and often results in fatalities for adults and children alike [10]. However, while touted as the most common bone malignancy, the incidence of osteosarcoma in the human population is relatively sparse, with a reported worldwide incidence of 1 to 3.4 cases per million people per year [8, 9]. The typical curative process for osteosarcoma is surgery, but the survival of osteosarcoma patients who are treated with surgery alone is relatively low [8]. As with other cancers, osteosarcoma patients are also capable of succumbing to the disease after a certain period of time, which makes the ability to make survivability predictions for these patients much more valuable.

Osteosarcoma survival rates for 5-year periods have been found to have improved to approximately 70% in one study based in the United States as a result of improved clinical trials that began in the early 1980s [4]. However, the same study also finds that osteosarcoma survivability varies greatly in different age groups; younger patients ranging from 0 to 24 years of age had a relative 5-year survival rate of 61.6%, while adults in the 25–59-year range had relative survival rates of 58.7% for the same period. Meanwhile, osteosarcoma patients older than 60 years of age had a more dismal 5-year survival rate of 24.2%. Despite that, osteosarcoma survival rates rank higher than cancers that afflict other vital organs, such as the lungs and liver. For comparison, the World Health Organization cites a 10-15% 5-year overall survival rate for lung cancer [11]. Temperature-responsive hydrogels have gained significant attention in tissue engineering due to their ability to transition from a liquid or semi-solid state at ambient temperatures to a gel state at body temperature. This unique characteristic enables the loading of therapeutic compounds onto the hydrogel in its liquid form, which can then be easily solidified and administered when applied. These advancements have opened doors for

developing targeted hydrogels to combat osteosarcoma, aiming to induce tumor cell death and enhance survivability in patients [12]. Another study revealed a significant correlation between age and survival rate in elderly patients, demonstrating that advancing age is associated with reduced survivability. This finding is consistent with previous research that has also emphasized the significance of tumor stage as a substantial risk factor, with distantly metastatic tumors having a worse prognosis compared to localized tumors [13].

While data on the United States' Surveillance, Epidemiology, and End Results (SEER) program indicates a similar 5-year survival rate of 19% for lung cancer as well as 18% for liver cancer [14, 15]. However, osteosarcoma is not necessarily the highest in rank for survivability among cancers of the vital organs; American kidney cancer patients have been found to experience a 5-year relative survival rate of about 75% [12]. As for comparisons against cancers of non-vital organs, osteosarcoma is often outclassed in terms of survival rates as these non-vital organs can often be removed to eliminate the cancer from the patient's body. For instance, the same SEER data indicates that female breast cancer has a relative 5-year survival rate of 90% and prostate cancer has a survival rate of 98% [12]. Nevertheless, gaining the ability to make more accurate survivability predictions for osteosarcoma patients is indispensable in ensuring that the patients are given a clearer picture of what to expect from the disease. A prognosis is often made by medical professionals based on the factors surrounding a certain medical condition. The word "prognosis" is defined as the probability of recovering as per anticipation from the usual course of a disease or peculiarity of a case [1, 16]. Prognosis is done to predict the outcome of an ailment in order to devise a suitable treatment plan for a patient [16, 17]. As such, prognosis is an important stage in the healthcare process, and research should be further performed in this area to understand and improve future outcomes for patients afflicted with a certain health condition [18]. In the context of cancer prognoses, there are three key areas of prediction that are often discussed: cancer susceptibility, cancer recurrence, and cancer survivability [17].

This research primarily focuses on predicting cancer survivability, which involves estimating outcomes such as life expectancy and disease progression after the cancer diagnosis [16, 17]. Early detection and treatment significantly impact long-term prognoses [19], emphasizing the need for more accurate tools to derive prognostic information. In the case of osteosarcoma, several factors influence the prognosis. The characteristics of the tumor itself, such as tumor necrosis, size, extension, and location, play a significant role in determining patient survivability [20]. Previous studies on osteosarcoma in the jaws have shown that larger and higher-grade tumors are associated with reduced survival rates [21]. Additionally, independent factors like the metastatic stage of the disease at presentation and gender have been identified as negative influences on survival probability [16, 22]. When developing prognostic models for osteosarcoma, these factors should be prioritized and given utmost attention.

### *2-1- Survivability Prediction*

Predicting a patient's probability of surviving osteosarcoma falls under the domain of survivability prediction. In a more generalized fashion, "survival" can be defined as a condition where the patient remains alive for a specific period of time after the diagnosis of a certain disease [23]. Also known as survivability analysis, the domain of survivability prediction falls under a subfield of statistics that aims to analyze and model the data that has the outcome of the time until the occurrence of a certain event of interest [24]. Survival analysis can also be defined as statistical methods that are used to examine changes over time related to a certain event [25]. In the case of cancer survivability, research in this field is primarily focused on predicting patient outcomes in terms of life expectancy, survivability, progression, or tumor-drug sensitivity after disease diagnosis [23]. Survival analysis methods can be broadly categorized into two categories, which are statistical methods and machine learning-based methods [26]. One statistical method that is commonly used to measure "time-to-event" data in survival analysis is the Kaplan-Meier method [26]. The Kaplan-Meier method measures the probability of an event happening in a certain time period by taking the number of patients that are affected by the event in that time period and dividing it by the total number of patients under study [26]. Similarly, the probability of an event not happening can be calculated by taking the number of patients that did not experience the event instead. In the case of osteosarcoma survivability, this method provides a simple way of calculating a patient's probability of surviving the disease over a certain period based on the historical data of other patients that have experienced the disease. This method is a univariate method, meaning that it can only calculate the time-to-event probability of a single variable at a time and thus cannot be used for multivariate analyses [25, 27]. One challenge that is particular to this field of study is the existence of censored instances, which are instances with event outcomes that become unobservable after a certain point in time [24].

These incomplete observations can occur due to various factors, such as a loss of contact with study members before the event happens, the intervention of external variables that affect the event, or an insufficient amount of time to observe the event [25]. For example, data sets for osteosarcoma survivability may only indicate patient survival for a certain number of years, such as 1 year or 5 years, after which the patients are no longer observed as part of the study. The Kaplan-Meier estimator is able to handle these censored instances by the nature of the statistical model itself, as it operates on the following three assumptions [27]: *i*. Participants who drop out or are censored from the study have the same survival capabilities as those who are still being followed. *ii*. The survival probabilities are the same for participants that are recruited early and late in the study. *iii*. The event being studied (e.g., patient death) occurs at the specified time. As a result, the Kaplan-Meier estimator provides a convenient method to study patient survivability for specific periods without requiring researchers to follow the progress of a certain patient for virtually limitless amounts of time until the

event under study happens. The capability of the estimator to account for censored instances allows researchers to work with data in specific time frames in order to establish a feasible idea of survivability for certain diseases. Another statistical model that is commonly used in the world of survival analysis is the Cox proportional hazards regression [28], often abbreviated as the Cox regression. This model is one of the most common regression modeling frameworks that enables the exploration of prognostic factors and the estimation of survival rates [29].

The Cox regression model is a semi-parametric model, meaning that the distribution of the outcome remains unknown even if it is based on a parametric regression model [24]. While its use in survival prediction for individuals is possible, the Cox regression typically places more focus on the differences in patient cohorts and is designed to gauge the effects of covariates on the changing hazard function [30–32]. As with other forms of regression models, the Cox regression is typically used to discover the weightage of each prognostic factor on the patient's survival, which differentiates itself from the Kaplan-Meier estimator, which only describes the probability of patient survival based on event data. As such, Cox regression is described as a multi-variate method because it can handle more than one variable in its analysis. This also makes the Cox regression more directly comparable to machine learning techniques for survivability prediction, which will be discussed later in this chapter. In summary, both the Kaplan-Meier estimator and Cox regression model are complementary tools in survival analysis and have their own ways of contributing towards the understanding of patient survivability.

## 3- Research Methodology

This research is designed to fulfill the primary objective of identifying the most effective machine learning technique for osteosarcoma survivability prediction. The primary work of this study is done in the modeling phase, where the cleaned data set is passed into machine learning algorithms to train a model that can predict the survivability of a patient with osteosarcoma. Before using machine learning techniques to develop prediction models, a Kaplan-Meier estimator is used to establish a baseline survival probability value for a better understanding of the patient survival rate. Extensive data cleaning and manipulation work was performed on attributes possessing these issues to ascertain the accuracy of the values provided. These are the following steps:

I. ***Removal of Missing Values****:* Some patients did not have values present for certain attributes. These missing values were rectified by inserting default values that are suitable for the attribute in question.

II. ***Values with Non-Standardized Wording****:* A single attribute may have different values that have the same intended meaning (e.g., "defaulted" and "defaulter" both meaning that the patient has defaulted from the study). These values were standardized using a common value, which is usually done by selecting one of the values in the domain.

III. ***Trailing Whitespaces****:* A single attribute may have values that look similar at first glance but in actuality have trailing whitespaces that cause the analysis program to interpret them as different values altogether (e.g., "Yes" and "Yes" with an additional space at the end). These values were standardized by removing the whitespaces.

IV. ***Multivalued Attributes****:* One attribute, which is metastasis location, contains multiple values in a single patient entry as the patient may experience metastatic growths in more than one location. In the initial data set, these values are combined into a single string with various non-standardized separator characters. These values were split up and assigned as multivalued lists for each patient entry to improve analysis capabilities.

The flowchart of the research methodology that was used to achieve the study's aims is shown in Figure 1.

As for the binary classification analysis, the random forest, support vector machine, and artificial neural network algorithms will be used for the purpose of developing prognostic models that are able to predict the potential survivability of a patient based on the factors present within the data set. These three techniques have been chosen for this study due to the presence of extensive literature that covers the effectiveness of each technique in providing accurate prognoses for patients afflicted with various forms of cancer, including osteosarcoma.

The random forest learner is provided with the cleaned data set containing both nominal and numerical attributes, while the support vector machine and artificial neural network learners are supplied with the data set that has undergone additional processing to be converted into all-numerical attributes. To avoid the over-fitting problem, 10-fold cross-validation is also used in the modeling process, where different parts of the data set are divided based on an 80:20 ratio of training data to test data over 10 iterations and later aggregated at the end to be evaluated for accuracy.

### *3-1- Machine Learning for Survivability Analysis*

In the modern era, where data is being generated at an unprecedented rate, there is a great amount of opportunity to extract valuable information that can be applied to improve processes. Such is also the case for the healthcare industry, where patient data is capable of being stored and retrieved regardless of location and time. However, data in its raw form alone does not provide much value without being processed further. To overcome this challenge, machine learning techniques have been introduced as a solution to harvest information that is hidden within the data [33]. The main outcome of the development of a certain machine learning technique is to produce a model that can be applied to tasks such as classification, prediction, or estimation [34]. With these techniques, researchers have been able to automatically select characteristics present within large amounts of structured data in order to increase risk classification accuracy [35].

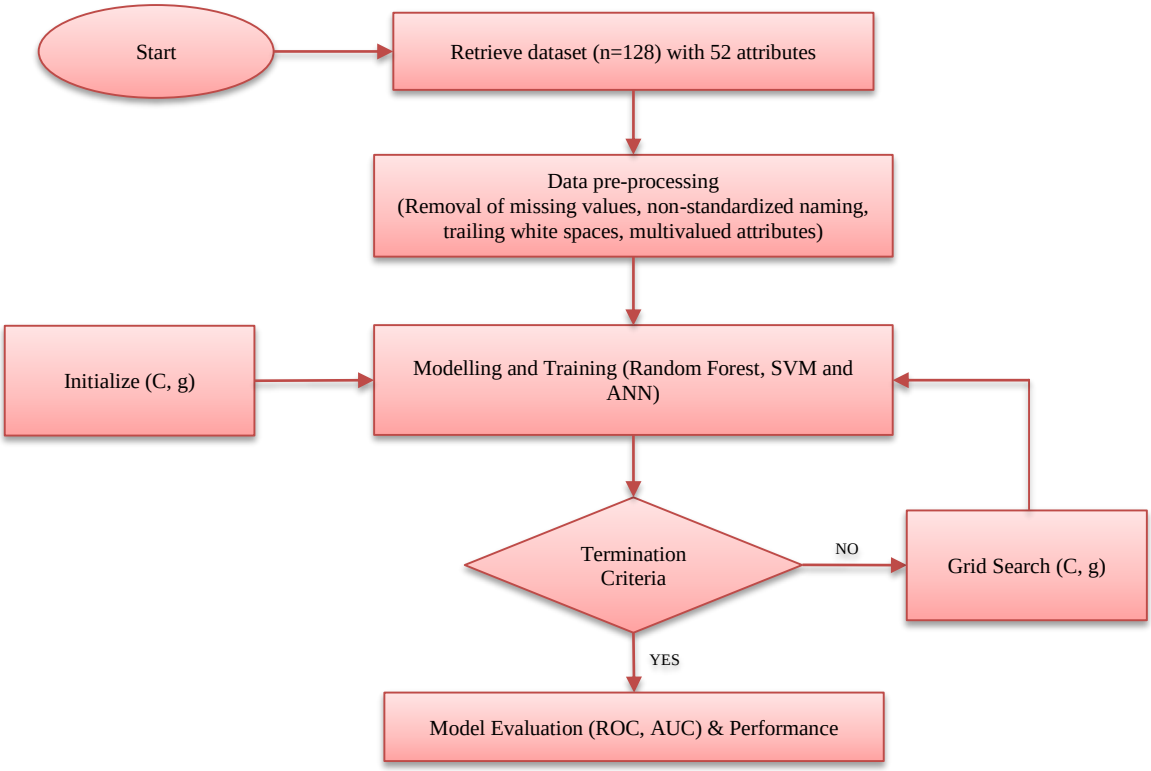


**Figure 1. Flowchart on research methodology**

The end goal of developing these models for use in the healthcare industry is so that the knowledge obtained from the data mining process can be used to assist healthcare providers in serving their patients better [36]. There are two commonly featured types of machine learning, which are supervised learning and unsupervised learning [33, 34]. While supervised learning requires labeled data with designated inputs and outputs for training, unsupervised learning does not require labeled data nor does it require a desired output [33]. In the field of cancer survivability research, many previous studies have utilized supervised learning techniques for developing new prognostic models. Several examples of supervised learning techniques that have been used for cancer prognosis are detailed in Table 1. For this research, the random forest, support vector machine, and artificial neural network techniques are selected for discussion and implementation due to the extensive literature present that describes the usage and prognostic accuracy of these algorithms in survivability prediction studies for various types of cancers, including osteosarcoma. The details of each algorithm, along with summaries of its various appearances in prior literature, are further described in later sections.

**Table 1. Supervised Machine Learning Techniques Used in Cancer Prognosis**

| Learning Technique | Publication(s) | Type of cancer data set |
|---|---|---|
| Random forest | Kaladhar et al. [6] | Stomach, Bronchus, Colon, Ovary, Breast |
| | Li et al. [37] | Osteosarcoma |
| | Montazeri et al. [7] | Breast |
| Support vector machine | Chao et al. [5] | Breast |
| | Li et al. [37] | Osteosarcoma |
| | Li et al [37] | Breast |
| Multilayer perceptron (artificial neural network) | Li et al [37] | Breast |
| Classification and regression tree | Kaladhar et al. [37] | Stomach, Bronchus, Colon, Ovary, Breast |
| Logistic model tree | Kaladhar et al. [37] | Stomach, Bronchus, Colon, Ovary, Breast |
| Naïve Bayesian | Kaladhar et al. [37] | Stomach, Bronchus, Colon, Ovary, Breast |
| | Li et al. [37] | Breast |
| 1-nearest neighbour | Li et al. [37] | Breast |
| AdaBoost | Li et al. [37] | Breast |
| RBF network | Li et al. [37] | Breast |
| C5.0 decision tree | Chao et al. [5] | Breast |

# 4- Data Overview

## *4-1- Data Set Characteristics*

The data set used as part of this research was obtained from NOCERAL, Faculty of Medicine, University of Malaya, Malaysia. The data contains information on patients diagnosed with osteosarcoma between the years of 1997 and 2011, including information on their survival status as well as various other factors that influence osteosarcoma prognosis, such as tumor location, metastasis, and tumor size. Extensive discussions were conducted with the research team in NOCERAL to ascertain the quality of the data set and to avoid ambiguity in understanding the values that are present within. The original data set contains data on a total of 128 patients with 52 attributes. The attributes present in the data set can be divided into the following categories.

*I.* Patient demographics (age, gender, ethnicity);

*II.* Osteosarcoma attributes (tumour characteristics, metastasis);

*III.* Treatment-related information (type of surgery/treatment, disease recurrence);

*IV.* Patient survivability.

The following subsections describe each of the patient attributes in the data set.

### *4-1-1- Patient Age*

Table 2 describes the age of the patients recorded in the data set. Patients in this study range from children as young as 5 years old to more elderly adults at 59 years old. The mean patient age is 16.225, while the median age is 15, indicating that osteosarcoma can be present in younger individuals.

**Table 2. Descriptive Analysis of Patient Age**

| | Patient age |
|---|---|
| **Minimum** | 5 |
| **Maximum** | 59 |
| **Mean** | 16.225 |
| **Standard deviation** | 7.914 |
| **Median** | 15 |

### *4-1-2- Gender of Patients*

Table 3 shows the frequency of patient genders recorded in the data set. A majority of the patients participating in this study are male (80 patients, or 62.5%), while the remaining 48 patients (37.5%) are female.

**Table 3. Frequency Table for Patients' Gender**

| Gender | Frequency | Percentage |
|---|---|---|
| Male | 80 | 62.5 |
| Female | 48 | 37.5 |
| **Total** | **128** | **100** |

### *4-1-3- Ethnicity of Patients*

Table 4 shows the frequency of patient ethnicities recorded in the data set. 86.7% of the patients participating in this study are either of Malay (59 patients, or 46.1%) or Chinese (52 patients, or 40.6%) descent. The remaining patients come from Indian (14 patients, or 10.9%) or other (3 patients, or 2.4%) ethnic backgrounds.

**Table 4. Frequency Table for Patient Ethnicity**

| Ethnicity | Frequency | Percentage |
|---|---|---|
| Malay | 59 | 46.1 |
| Chinese | 52 | 40.6 |
| Indian | 14 | 10.9 |
| Other | 3 | 2.4 |
| **Total** | **128** | **100** |

### *4-1-4- Primary Site of Tumor*

Table 5 shows the frequency of the primary tumor site diagnosed in the patients recorded in the data set. Almost all of the patients had tumors in extremity sites, with only six patients recorded to have tumors in axial sites. One patient was not applicable for this attribute.

**Table 5. Frequency Table for Primary Site of Tumour**

| Primary Site | Frequency | Percentage |
|---|---|---|
| Extremity | 121 | 94.5 |
| Axial | 6 | 4.7 |
| Not applicable | 1 | 0.8 |
| **Total** | **128** | **100** |

### *4-1-5- Tumour Location*

Table 6 shows the frequency of tumor locations detected for the patients recorded in the data set. Exactly half of the patients had tumors in their femurs, with another third having tumors in the tibia and humerus. The remaining patients have tumors detected in their fibulae, pelvises, radii, ribs, scapulae, and ulnae. Two patients are not applicable for this attribute.

**Table 6. Frequency Table for Tumour Location**

| Tumour location | Frequency | Percentage |
|---|---|---|
| Femur | 64 | 50.0 |
| Tibia | 31 | 24.2 |
| Humerus | 13 | 10.1 |
| Fibula | 8 | 6.2 |
| Pelvis | 4 | 3.1 |
| Radius | 2 | 1.6 |
| Rib | 2 | 1.6 |
| Scapula | 1 | 0.8 |
| Ulna | 1 | 0.8 |
| Not applicable | 2 | 1.6 |
| **Total** | **128** | **100** |

### *4-1-6-Metastasis*

Table 7 shows the frequency of metastasis presence for the patients recorded in the data set. More than half of the patients have developed metastatic growths, with only 34 patients not undergoing this experience. 16 patients did not apply for this condition or defaulted from the study before metastasis could be detected.

**Table 7. Frequency Table for Presence of Metastasis**

| Presence of metastasis | Frequency | Percentage |
|---|---|---|
| Yes | 78 | 60.9 |
| No | 34 | 26.6 |
| Not applicable/Defaulted | 16 | 12.5 |
| **Total** | **128** | **100** |

Table 8 shows the frequency of metastasis detection at diagnosis for the patients recorded in the data set. 54.7% of the patients did not have metastatic growths detected during their initial diagnosis with the hospital, while 39.1% of patients did. Eight patients did not apply for this attribute or defaulted from the study before metastasis could be ascertained at diagnosis.

**Table 8. Frequency Table for Metastasis at Diagnosis**

| Metastasis at diagnosis | Frequency | Percentage |
|---|---|---|
| No | 70 | 54.7 |
| Yes | 50 | 39.1 |
| Not applicable/Defaulted | 8 | 6.2 |
| **Total** | **128** | **100** |

Table 9 shows the frequency of metastasis detection during patient treatment as recorded in the data set. Under 10% of patients had their metastasis condition detected during treatment, with 47 others did not have such a condition detected. 69 patients are either not applicable for this attribute or have defaulted or passed away before metastasis could be detected in this phase.

**Table 9. Frequency Table for Metastasis during Treatment**

| Metastasis during treatment | Frequency | Percentage |
|---|---|---|
| No | 47 | 36.7 |
| Yes | 12 | 9.4 |
| Not applicable/Defaulted/Passed away | 69 | 53.9 |
| **Total** | **128** | **100** |

Table 10 shows the frequency of metastasis detection in patients after treatment has been done. 16 patients had their metastatic conditions detected at this stage, while 27 others did not. 85 patients are not applicable for this attribute or have defaulted from the study before metastasis could be detected at this stage.

**Table 10. Frequency Table for Metastasis after Treatment**

| Metastasis after treatment | Frequency | Percentage |
|---|---|---|
| No | 27 | 21.1 |
| Yes | 16 | 12.5 |
| Not applicable/Defaulted | 85 | 66.4 |
| **Total** | **128** | **100** |

Table 11 shows the frequency of metastasis presence in the first 12 months since the patient's case was first presented to the hospital. 9 patients had metastatic growths detected within the 12-month period, while 7 patients did not have metastases detected in the same period. 112 other patients are not applicable for this attribute or have passed away before metastasis could be detected.

**Table 11. Frequency Table for Presence of Metastasis in First 12 Months since Presentation**

| Presence of metastasis in first 12 months since presentation | Frequency | Percentage |
|---|---|---|
| Yes | 9 | 7.0 |
| No | 7 | 5.5 |
| Not applicable/Passed away | 112 | 87.5 |
| **Total** | **128** | **100** |

Table 12 shows the frequency of metastasis presence in the first 24 months since the patient's case was first presented to the hospital. 4 patients had metastases detected within this period, while 5 others did not. 119 patients are not applicable for this attribute.

**Table 12. Frequency Table for Presence of Metastasis in First 24 Months since Presentation**

| Presence of metastasis in first 24 months since presentation | Frequency | Percentage |
|---|---|---|
| No | 5 | 3.9 |
| Yes | 4 | 3.1 |
| Not applicable | 119 | 93.0 |
| **Total** | **128** | **100** |

Table 13 shows the frequency of metastasis presence beyond the first 24 months since case presentation to the hospital. Beyond the first 2 years, 3 patients had metastatic growths detected, while 3 other patients did not. 122 patients were not applicable for this attribute.

**Table 13. Frequency Table for Presence of Metastasis beyond 24 Months since Presentation**

| Presence of metastasis beyond 24 months since presentation | Frequency | Percentage |
|---|---|---|
| No | 3 | 2.3 |
| Yes | 3 | 2.3 |
| Not applicable | 122 | 95.4 |
| **Total** | **128** | **100** |

Table 14 shows the frequency of metastatic growth locations separated by patients recorded in the data set. Over a third of the patients had metastases detected in the lungs only, while the remaining patients with metastases had growths in various combinations of locations. Meanwhile, Table 15 shows the same data but separates metastasis data by individual locations. Of the 55 patients with metastasis, almost all of them had metastatic growths in the lungs, followed by a small sum of patients having similar growths in the spine, ribs, and tibia. Five patients had unique growths in the humerus, iliac, lymph nodes, sternum, and scapula.

**Table 14. Frequency Table for Metastasis Location by Patient**

| Metastasis location by patient | Frequency | Percentage |
|---|---|---|
| Lungs | 46 | 35.9 |
| Lungs and spine | 2 | .61 |
| Lungs, ribs and tibia | 1 | 0.8 |
| Lungs and ribs | 1 | 0.8 |
| Humerus | 1 | 0.8 |
| Ribs, tibia, iliac and spine | 1 | 0.8 |
| Lymph nodes and lungs | 1 | 0.8 |
| Spine and sternum | 1 | 0.8 |
| Lungs, ribs, spine and scapula | 1 | 0.8 |
| Not applicable | 73 | 56.9 |
| **Total** | **128** | **100** |

**Table 15. Frequency Table for Metastasis Location by Location**

| Metastasis location by location | Frequency | Percentage of all patients | Percentage of 55 patients with metastasis |
|---|---|---|---|
| Lungs | 52 | 40.6 | 94.5 |
| Spine | 5 | 3.9 | 9.1 |
| Ribs | 4 | 3.1 | 7.3 |
| Tibia | 2 | 1.6 | 3.6 |
| Humerus | 1 | 0.8 | 1.8 |
| Iliac | 1 | 0.8 | 1.8 |
| Lymph nodes | 1 | 0.8 | 1.8 |
| Sternum | 1 | 0.8 | 1.8 |
| Scapula | 1 | 0.8 | 1.8 |

Table 16 shows the frequency of lung metastasis laterality found in the patients recorded in the data set. 26 patients have been found to have bilateral lung metastases, while 15 had unilateral metastases. 83 patients are not applicable for this attribute.

**Table 16. Frequency Table for Laterality of Lung Metastasis**

| Laterality of lung metastasis | Frequency | Percentage |
|---|---|---|
| Bilateral | 26 | 20.3 |
| Unilateral | 15 | 11.7 |
| Yes | 3 | 2.3 |
| No | 1 | 0.8 |
| Not applicable | 83 | 64.9 |
| **Total** | **128** | **100** |

Table 17 shows the frequency of the number of lung nodules detected in patients recorded in the data set. Almost one-fifth of the patients had less than four lung nodules detected, while 14.8% had four nodules or more. 84 patients are not applicable for this attribute.

**Table 17. Frequency Table for Number of Lung Nodules**

| Number of lung nodules | Frequency | Percentage |
|---|---|---|
| Less than 4 | 25 | 19.5 |
| Greater than or equal to 4 | 19 | 14.8 |
| Not applicable | 84 | 65.7 |
| **Total** | **128** | **100** |

Table 18 shows the frequency of distant metastasis presence in the patients recorded in the data set. Many patients did not exhibit the presence of distant metastatic growth, with only 11 patients having such a condition detected. 61 patients are not applicable for this attribute.

**Table 18. Frequency Table for Presence of Distant Metastasis**

| Presence of distance metastasis | Frequency | Percentage |
|---|---|---|
| No | 56 | 43.7 |
| Yes | 11 | 8.6 |
| Not applicable | 61 | 47.7 |
| **Total** | **128** | **100** |

### *4-1-7- Pathological Fracture*

Table 19 shows the frequency of pathological fracture presence in the patients recorded in the data set. Only six patients had such fractures detected, while the remaining 95.3% of patients did not.

**Table 19. Frequency Table for Presence of Pathological Fracture**

| Presence of pathological fracture | Frequency | Percentage |
|---|---|---|
| No | 122 | 95.3 |
| Yes | 6 | 4.7 |
| **Total** | **128** | **100** |

### *4-1-8- Histological Subtype*

Table 20 shows the frequency of the osteosarcoma histological subtype identified for the patients recorded in the data set. Almost two-thirds of patients had osteoblastic osteosarcoma, with another 11.7% suffering from chondroblastic osteosarcoma. The osteosarcomas of the remaining patients are separated into other histological subtypes.

**Table 20. Frequency Table for Histological Subtype**

| Histological subtype | Frequency | Percentage |
|---|---|---|
| Osteoblastic | 84 | 65.6 |
| Chondroblastic | 15 | 11.7 |
| Giant cell (rich) | 4 | 3.1 |
| Parosteal | 3 | 2.3 |
| Giant cell | 3 | 2.3 |
| Telangiectatic | 3 | 2.3 |
| Fibroblastic | 2 | 1.6 |
| Sarcomatoid | 1 | 0.8 |
| Not applicable | 13 | 10.3 |
| **Total** | **128** | **100** |

### *4-1-9- Histological Response*

Table 21 shows the frequency of patient histological response rates as recorded in the data set. 48 patients exhibited a response of less than 90%, while 43 other patients had responses that were 90% or greater. 37 patients were not applicable for this attribute.

**Table 21. Frequency Table for Histological Response**

| Histological response | Frequency | Percentage |
|---|---|---|
| Less than 90% | 48 | 37.5 |
| Greater than or equal to 90% | 43 | 33.6 |
| Not applicable | 37 | 28.9 |
| **Total** | **128** | **100** |

### *4-1-10- Tumor Size*

Table 22 shows the frequency of tumor sizes detected in the patients recorded in the data set. From the 102 patients that were eligible for this attribute, a large number of patients had been diagnosed with tumors that were greater than or equal to 10 centimeters in size.

**Table 22. Frequency Table for Tumour Size**

| Tumor size | Frequency | Percentage |
|---|---|---|
| Greater than or equal to 10 cm | 59 | 46.1 |
| Less than 10 cm | 43 | 33.6 |
| Not applicable | 26 | 20.3 |
| **Total** | **128** | **100** |

### *4-1-11- Enneking Stage*

Table 23 shows the frequency of Enneking stages for the patients recorded in the data set. Almost half of the patients recorded in the study are diagnosed with Stage 3 osteosarcoma, with 41 other patients diagnosed with Stage 2B and 1 patient diagnosed with Stage 1B. 24 patients were not applicable for this attribute.

**Table 23. Frequency Table for Tumor Size**

| Enneking stage | Frequency | Percentage |
|---|---|---|
| Stage 3 | 62 | 48.4 |
| Stage 2B | 41 | 32.0 |
| Stage 1B | 1 | 0.8 |
| Not applicable | 24 | 18.8 |
| **Total** | **128** | **100** |

### *4-1-12- Serum ALP*

Table 24 describes the serum alkaline phosphatase levels (ALP) of patients recorded in the data set. Patient serum ALP levels range from 56 to 3141m with a mean of 433.147 and a median value of 234.5.

**Table 24. Descriptive Analysis of serum ALP**

| | Serum ALP |
|---|---|
| **Minimum** | 56 |
| **Maximum** | 3141 |
| **Mean** | 443.147 |
| **Standard deviation** | 541.643 |
| **Median** | 234.5 |

### *4-1-13- Monocyte Count*

Table 25 describes the monocyte count for the patients recorded in the data set. The monocyte counts ranges from as low as $0.085\times10^9$ to as high as $3.021\times10^9$, with a mean of $0.674\times10^9$ and a median value of $0.575\times10^9$.

**Table 25. Descriptive Analysis of Monocyte Count**

| | Monocyte count ($\times10^9$) |
|---|---|
| **Minimum** | 0.085 |
| **Maximum** | 3.021 |
| **Mean** | 0.674 |
| **Standard deviation** | 0.448 |
| **Median** | 0.575 |

### *4-1-14- Lymphocyte Count*

Table 26 describes the lymphocyte count for the patients recorded in the data set. Patient lymphocyte counts range from $0.273\times10^9$ to $9.650\times10^9$, with a mean of $2.165\times10^9$ and a median value of $1.984\times10^9$.

**Table 26. Descriptive Analysis of Lymphocyte Count**

| | Lymphocyte count ($\times10^9$) |
|---|---|
| **Minimum** | 0.273 |
| **Maximum** | 9.650 |
| **Mean** | 2.165 |
| **Standard deviation** | 1.185 |
| **Median** | 1.984 |

### *4-1-15- Primary Surgery*

Table 27 shows the frequency of primary surgery types that the patients recorded in the data set have undergone. Almost two-thirds of patients have undergone salvage surgery, with the remaining patients being treated with amputation, conservative, or ablative salvage surgery. 15 patients were not applicable for this attribute. They had either defaulted from the study or passed away before primary surgery was performed.

**Table 27. Descriptive Analysis of Lymphocyte Count**

| Primary surgery | Frequency | Percentage |
|---|---|---|
| Salvage | 83 | 64.8 |
| Amputation | 25 | 19.5 |
| Conservative | 4 | 3.1 |
| Salvage (ablation) | 1 | 0.8 |
| Not applicable/Defaulted/Passed away | 15 | 11.8 |
| **Total** | **128** | **100** |

### *4-1-16-Induction Chemotherapy*

Table 28 shows the frequency of patients who underwent induction chemotherapy as recorded in the data set. A majority of patients were treated with induction chemotherapy, with only 17 patients not doing so. Six patients are not applicable for this attribute or had defaulted from the study before the treatment was performed. The chemotherapy regimens used by patients for this treatment are shown in Table 29. Most patients are treated with the DCM regime, with 16 patients treated using the DC regime and 33 being treated with other regimes. 25 patients were either not applicable for this attribute, had defaulted, or had passed away before the regime could be completed.

**Table 28. Frequency Table for Induction Chemotherapy**

| Usage of induction chemotherapy | Frequency | Percentage |
|---|---|---|
| Yes | 105 | 82.0 |
| No | 17 | 13.3 |
| Not applicable/Defaulted | 6 | 4.7 |
| **Total** | **128** | **100** |

**Table 29. Frequency Table for Induction Chemotherapy Regime**

| Induction chemotherapy regime | Frequency | Percentage |
|---|---|---|
| DCM | 54 | 42.2 |
| DC | 16 | 12.5 |
| Other | 33 | 25.8 |
| Not applicable/Defaulted/Passed away | 25 | 19.5 |
| **Total** | **128** | **100** |

### *4-1-17- Adjuvant Chemotherapy*

Table 30 shows the frequency of patients who underwent adjuvant chemotherapy as recorded in the data set. Almost 80% of patients were treated with this chemotherapy, with only 2 patients not utilizing this treatment. Twenty-four patients were either not applicable for this attribute, had defaulted from the study before the treatment was performed, or had continued treatment at a different facility. The induction chemotherapy regimens undergone by patients are shown in Table 31. Like the induction chemotherapy regimens, most patients are treated with the DCM regime, with 10 patients treated using the DC regime and 36 being treated with other regimes. 28 patients are either not applicable for this attribute, had defaulted, or had passed away before the regime could be completed.

**Table 30. Frequency Table for Adjuvant Chemotherapy**

| Usage of adjuvant chemotherapy | Frequency | Percentage |
|---|---|---|
| Yes | 102 | 79.7 |
| No | 2 | 1.6 |
| Not applicable/Defaulted/Continued treatment elsewhere | 24 | 18.7 |
| **Total** | **128** | **100** |

**Table 31. Frequency Table for Adjuvant Chemotherapy Regime**

| Adjuvant chemotherapy regime | Frequency | Percentage |
|---|---|---|
| DCM | 54 | 42.2 |
| DC | 10 | 7.8 |
| Other | 36 | 28.1 |
| Not applicable/Defaulted/Passed away | 28 | 21.9 |
| **Total** | **128** | **100** |

### *4-1-18- Disease Recurrence*

Table 32 shows the frequency of patients experiencing recurrence of the osteosarcoma disease as recorded in the data set. Of the 75 applicable patients for this attribute, 46 did not experience recurrence, while 29 had recurrence detected.

**Table 32. Frequency Table for Disease Recurrence**

| Disease recurrence | Frequency | Percentage |
|---|---|---|
| No | 46 | 35.9 |
| Yes | 29 | 22.7 |
| Not applicable | 53 | 41.4 |
| **Total** | **128** | **100** |

### *4-1-19- Treatment Completion*

Table 33 shows the frequency of patients who have completed their treatments as recorded in the data set. Over 70% of patients are recorded as having successfully gone through their treatments, while 30 patients have not. Seven patients were not applicable for this attribute.

**Table 33. Frequency Table for Treatment Completion**

| Treatment completion | Frequency | Percentage |
|---|---|---|
| Yes | 91 | 71.1 |
| No | 30 | 23.4 |
| Not applicable | 7 | 5.5 |
| **Total** | **128** | **100** |

### *4-1-20- Patient Survival*

Table 34 shows the length of patient survival in rounded-down years as recorded in the data set. While almost half of the patients survived beyond 5 years, a similar number also succumbed to the disease within 3 years. Eight patients fell victim to the disease between 4 to 5 years after first being diagnosed with osteosarcoma.

**Table 34. Frequency Table for Length of Patient Survival in Years**

| Length of survival (years, rounded down) | Frequency | Percentage |
|---|---|---|
| 0 | 25 | 19.5 |
| 1 | 19 | 14.8 |
| 2 | 18 | 14.1 |
| 3 | 5 | 3.9 |
| 4 | 3 | 2.3 |
| 5 or greater | 58 | 45.4 |
| **Total** | **128** | **100** |

Based on the above data, the frequency of patients surviving after 2 years is shown in Table 35. Almost two-thirds of the recorded patients survived the disease beyond this period, with the remaining 44 patients succumbing to their illnesses.

**Table 35. Frequency Table for Patient Survival after 2 Years**

| Patient survival after 2 years | Frequency | Percentage |
|---|---|---|
| Yes | 84 | 65.6 |
| No | 44 | 34.4 |
| **Total** | **128** | **100** |

Similarly, the frequency of patients surviving after 5 years is shown in Table 36. While the number of patients succumbing to the disease increased by 14 to 58 patients from the first 2 years, over half of the patients still managed to survive beyond 5 years.

**Table 36. Frequency Table for Patient Survival after 5 Years**

| Patient survival after 5 years | Frequency | Percentage |
|---|---|---|
| Yes | 70 | 54.7 |
| No | 58 | 45.3 |
| **Total** | **128** | **100** |

## 5- Data Analysis Methods and Results

The data preparation process is done within the platform using various tools such as string manipulation and attribute normalizer nodes. For descriptive analysis purposes, statistical analysis was utilized to quickly obtain distribution values for each attribute that is presented in the dataset. The Kaplan-Meier survivability prediction value is obtained by supplying the data to the readily available Kaplan-Meier estimator node, while each of the machine learning models is trained and evaluated using built-in learner and predictor nodes as well as ROC curve generator tools.

### *5-1- Pilot Study*

A pilot study was conducted prior to this research with a subset of the same data set to test the research process and verify the capabilities of the data analysis platform. In the pilot study, the data on 40 randomly selected patients was used to test all three machine learning techniques as well as the Kaplan-Meier estimator for both 2-year and 5-year survival. Evaluation metrics remained the same in this pilot study, where the ROC curves for each machine learning model were observed and the AUCs were observed. In this pilot study, the machine learning prognostic models were able to produce AUCs of 0.705, 0.777, and 0.673 in terms of 2-year survival for the random forest, support vector machine, and artificial neural network algorithms, respectively, and AUCs of 0.508, 0.712, and 0.554 in terms of 5-year survival for the same algorithms. Meanwhile, the Kaplan-Meier estimator was able to produce survival estimates of 0.700 and 0.525 for both 2- and 5-year survival, respectively. This pilot study supports the validity of the research methodology and serves as a basis for the development of this research's complete analysis.

### *5-2- Kaplan-Meier Analysis*

Using the patient survival length data that is present within the data set, a Kaplan-Meier analysis is performed to discover the survivability rate of the osteosarcoma patients that are being studied as part of this research. The Kaplan-Meier curves for patient survival after 2 years and 5 years can be seen in Figures 2 and 3, respectively. Meanwhile, the exact Kaplan-Meier estimator values obtained are recorded in Table 37.

**Table 37. Frequency Table for Patient Survival after 5 Years**

| Patient survival length | Kaplan-Meier estimator value |
|---|---|
| 2 years | 0.656 |
| 5 years | 0.453 |

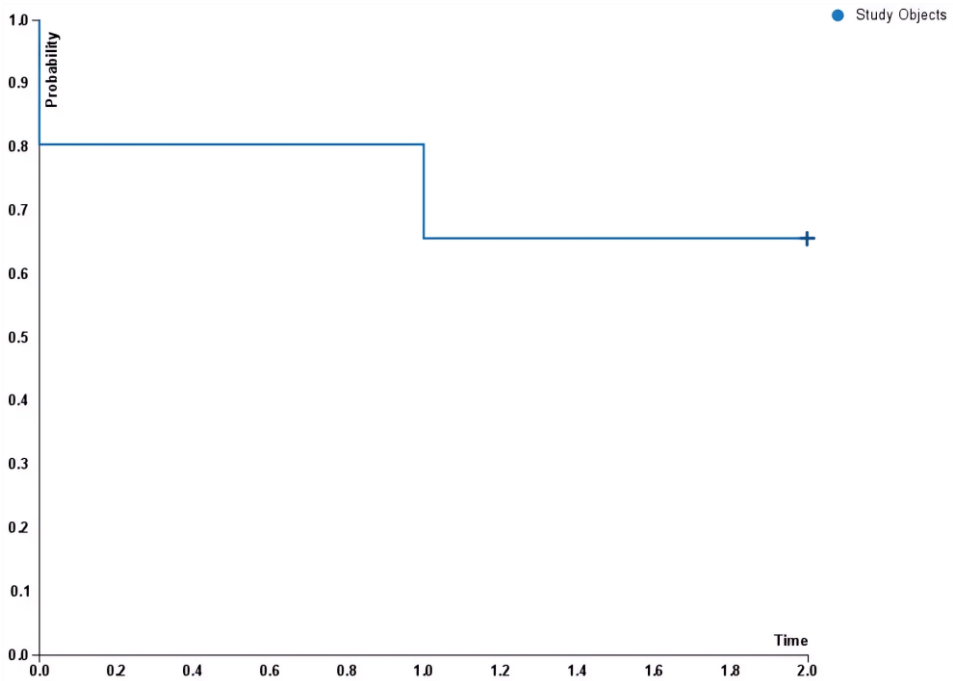


**Figure 2. Kaplan-Meier Curve for Patient Survival after 2 Years**

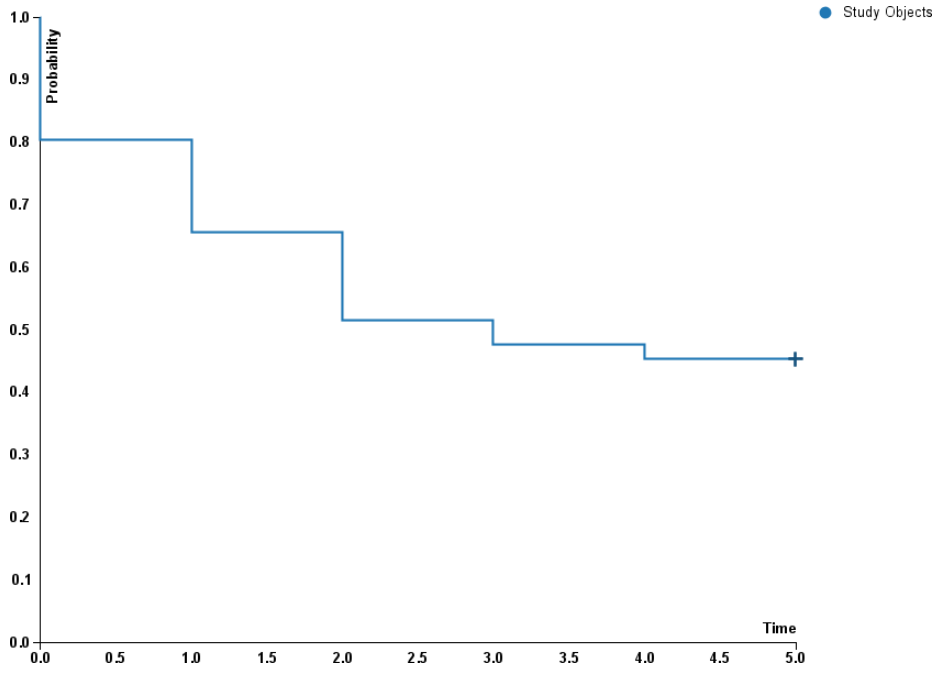


**Figure 3. Kaplan-Meier Curve for Patient Survival after 5 Years**

From the obtained results, the Kaplan-Meier estimator denotes that 65.6% of the studied patients are able to survive the osteosarcoma disease beyond 2 years, and only 45.3% are able to survive beyond 5 years. The 5-year curve also illustrates the fact that almost half of the patients did not manage to survive the disease past the first three years.

### *5-3- Binary Classification Analysis*

The following subsections describe the performance of each binary classification algorithm on the patient survival data. All three selected algorithms are used to develop survival prediction models for 2-year and 5-year survival, respectively, before making comparisons of the performance results to ascertain the best-performing model.

### *5-3-1- Random Forest*

The random forest (RF) learner is configured to use the information gain ratio as the tree splitting criterion and a limit of 100 models. The ROC curves for the 2-year and 5-year survival predictions using the RF algorithm are shown in Figures 4 and 5, respectively. From these curves, the AUC for 2-year survival is calculated to be 0.640, meaning that the model is able to correctly predict 64.0% of patients to survive osteosarcoma beyond 2 years. Similarly, the AUC for 5-year survival is calculated to have a value of 0.447, which means that the model can accurately predict that 44.7% of patients will survive the disease after 5 years.

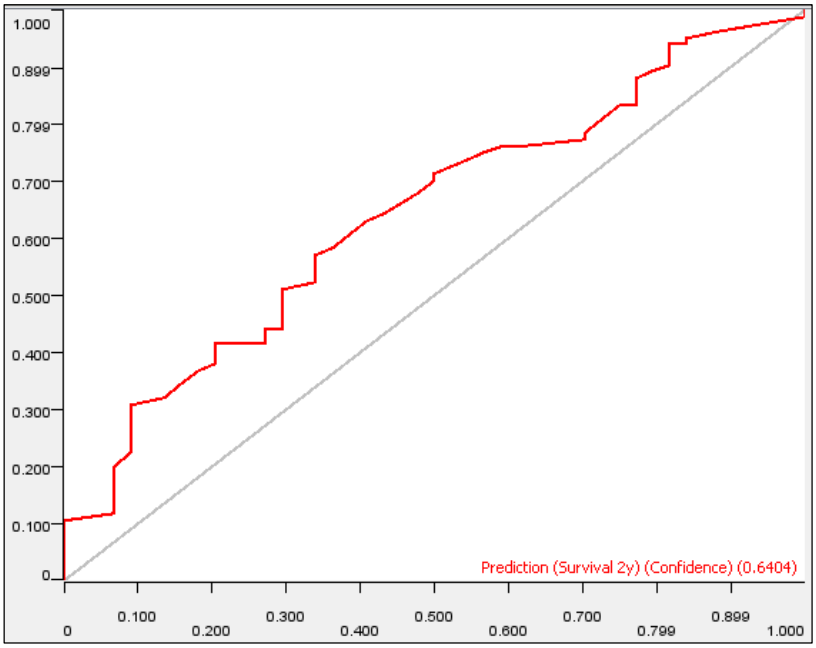


**Figure 4. ROC Curve for 2-Year Patient Survival with the Random Forest Algorithm**

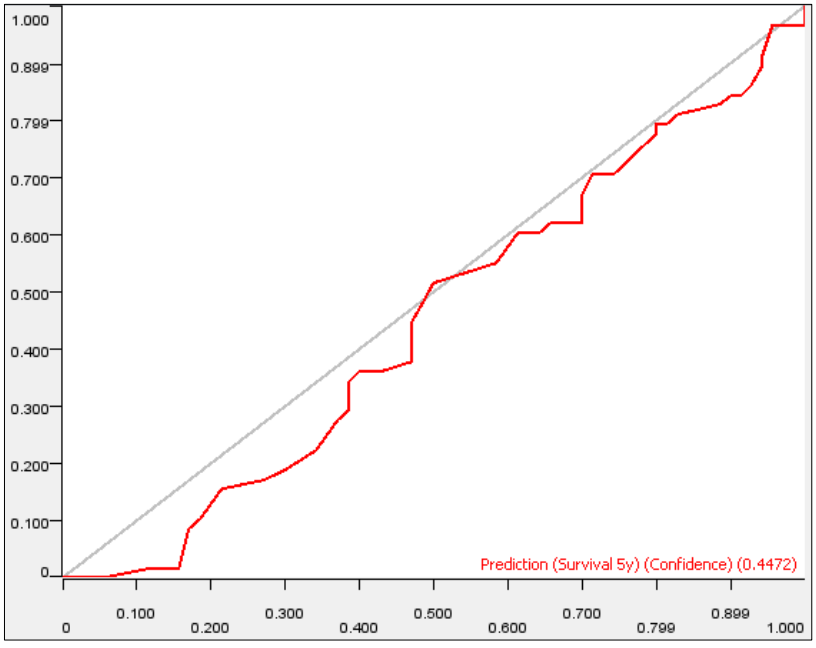


**Figure 5. ROC Curve for 5-Year Patient Survival with the Random Forest Algorithm**

### *5-3-2-Support Vector Machine*

The support vector machine (SVM) learner is configured to use a polynomial kernel with a bias, power, and gamma of 1.0 as well as an overlapping penalty of 1.0. The ROC curves for 2-year and 5-year survival generated by the SVM model can be seen in Figures 6 and 7, respectively. For the 2-year survival ROC curve, the AUC is calculated to be at 0.815, which indicates that the model is able to accurately predict that 81.5% of patients will survive the osteosarcoma disease after 2 years. As for 5-year survival, the AUC is calculated to have a value of 0.828, which means that the model is capable of correctly predicting that 82.8% of patients will survive after 5 years.

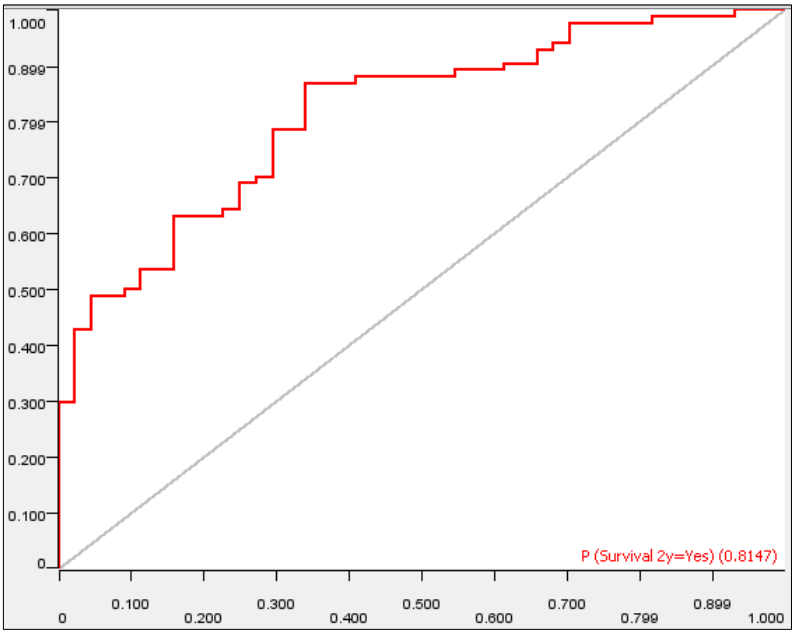


**Figure 6. ROC Curve for 2-Year Patient Survival with the Support Vector Machine Algorithm**

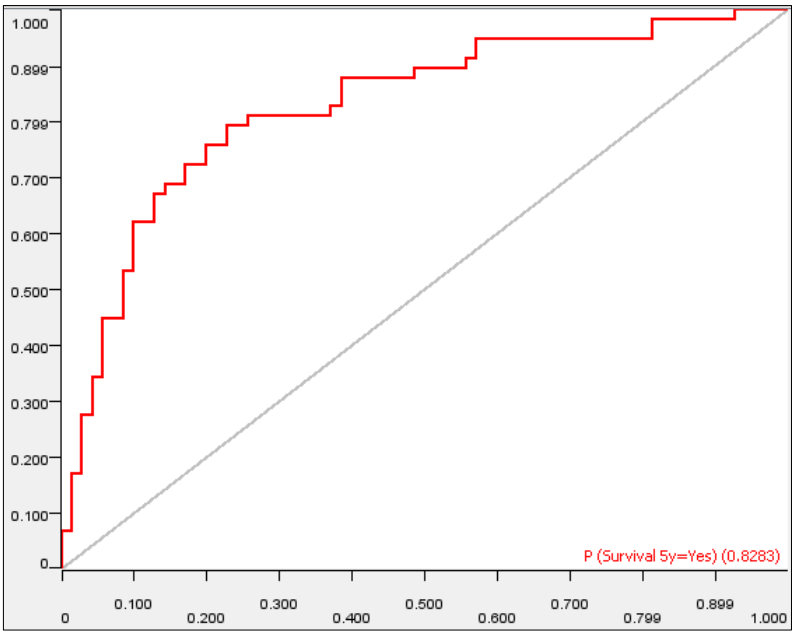


**Figure 7. ROC Curve for 5-Year Patient Survival with the Support Vector Machine Algorithm**

### *5-3-3- Artificial Neural Network*

For the artificial neural network (ANN) analysis, the multilayer perceptron learner was used with a maximum of 100 iterations, 1 hidden layer, and 10 hidden neurons per layer. The ROC curves obtained from the ANN model predictions for 2-year and 5-year survival are shown in Figures 8 and 9, respectively.

The 2-year survival model ROC curve presents an AUC of 0.833, which indicates that the model is capable of correctly predicting the survival of 83.3% of patients after 2 years. Meanwhile, the 5-year survival ROC curve has an AUC of 0.807, meaning that the model can successfully predict the survival of 80.7% of patients after 5 years.

### *5-4- Comparison of Binary Classification Model Performance*

Using the survival probability values obtained from the binary classification models, a bar chart comparing the values is constructed, as seen in Figure 10. From this chart, it can be seen that the performance of the random forest model is underwhelming, with both the 2-year survivability and the 5-year survivability prediction confidence probabilities underperforming against the support vector machine and artificial neural network models. The highest survivability prediction confidences can be observed in both the support vector machine and artificial neural network models, with each model edging out the other in the different studied time periods.

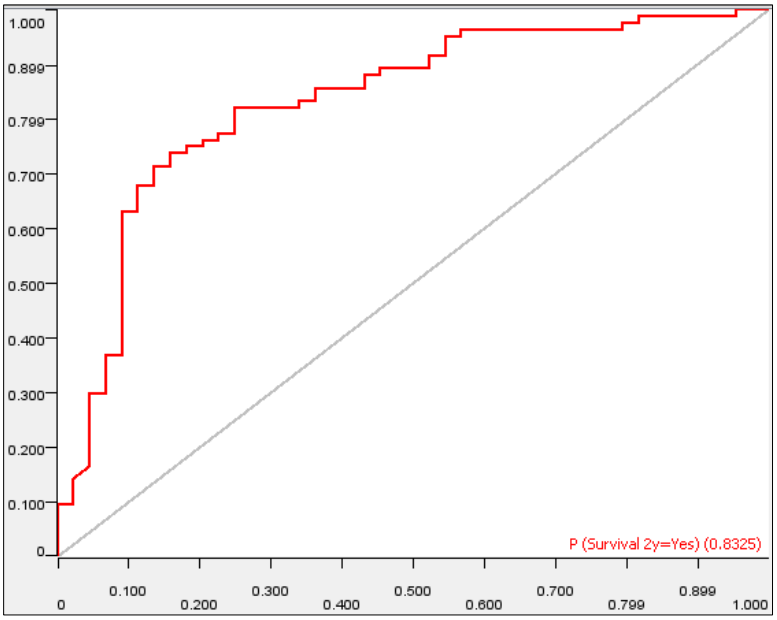


**Figure 8.** ROC Curve for 2-Year Patient Survival with the Multilayer Perceptron Artificial Neural Network Algorithm

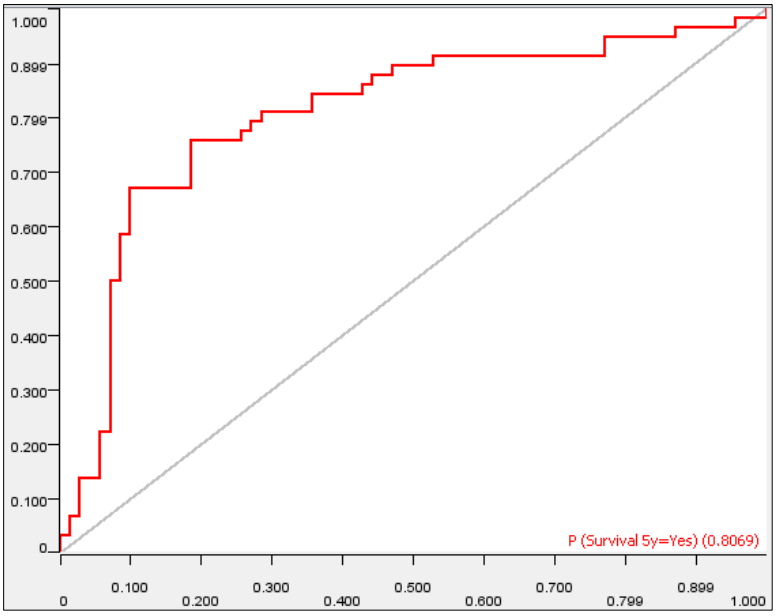


**Figure 9.** ROC Curve for 5-Year Patient Survival with the Multilayer Perceptron Artificial Neural Network Algorithm

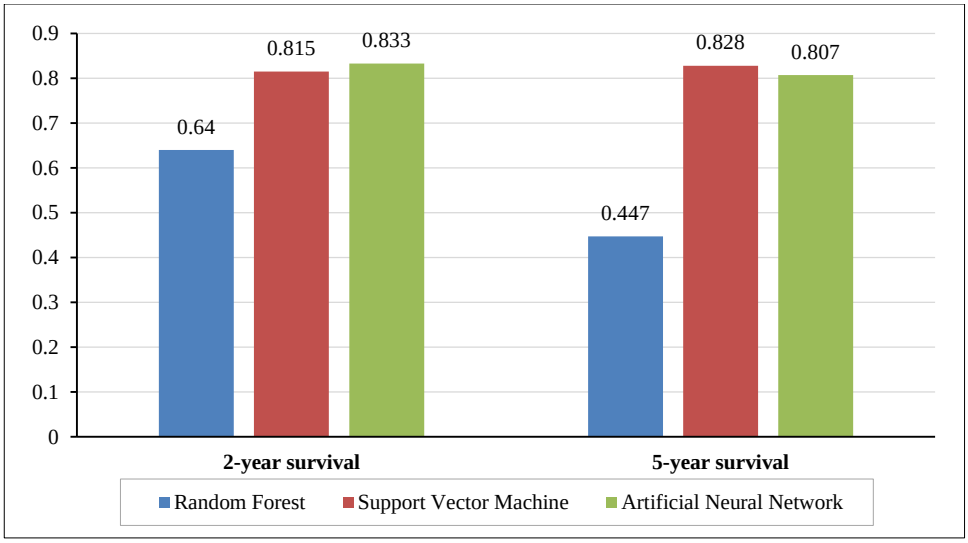


**Figure 10.** Bar Chart. Depicting Binary Classification Algorithm Performance

For patient survival after 2 years, the artificial neural network model is able to correctly predict surviving patients 1.8% better than the support vector machine model, while the support vector machine model is able to outperform the artificial neural network model by 2.1% for 5-year patient survivability prediction. From these results, it can be established that both the artificial neural network and support vector machine models are the most effective in predicting the survivability of osteosarcoma patients in this study, with each model topping the performance measures on the 2-year and the 5-year survivability predictions, respectively.

## 6- Conclusion

This study has discovered the current extent of survivability among patients that are affected by osteosarcoma as well as the effectiveness of binary classification machine learning techniques for predicting osteosarcoma survivability among patients. With survival rates being generally low for this disease, patients are often faced with doubts regarding their own mortality and the well-being of the loved ones around them should their affliction take a turn for the worse. As such, the ability to provide more accurate prognoses through machine learning-powered models is highly valuable for medical professionals to provide the best possible advice and treatment options to osteosarcoma patients and place more control of the patients fates in their own hands. With enough refinement, these machine learning techniques can become highly accurate and viable tools to assist doctors in making these prognoses with little doubt in their decision-making. Thus, this research provides yet another step forward towards achieving the goal of integrating machine learning techniques with tools that are currently available to medical practitioners. Nevertheless, there is potential for this study to be developed further to enhance the effectiveness of the applied machine learning techniques so that both medical professionals and patients can be more confident in the survivability predictions generated by machine learning- or artificial intelligence-based models. To gauge the effectiveness of machine learning techniques on predicting osteosarcoma survivability, this research has looked into three distinct binary classification techniques for comparison, which are the random forest, support vector machine, and artificial neural network techniques. The results obtained from this study indicate that the support vector machine and artificial neural network algorithms performed better than random forests for both 2-year and 5-year survivability. Thus, this research provides yet another step forward towards achieving the goal of integrating machine learning techniques with tools that are currently available to medical practitioners. Nevertheless, there is potential for this study to be developed further to enhance the effectiveness of the applied machine learning techniques in order for both medical professionals and patients to be more confident in the survivability predictions generated by machine learning- or artificial intelligence-based models.

## 7- Declarations

### *7-1- Author Contributions*

Conceptualization, S.M. and V.A.S.; methodology, S.M. and K.S.M.A.; investigation, V.A.S. and S.M.; validation: B.P. and M.J.K.; writing—original draft preparation, S.M.; writing—review and editing, T.O.K.Z.; project administration, T.O.K.Z. All authors have read and agreed to the published version of the manuscript.

### *7-2- Data Availability Statement*

The data presented in this study are available on request from the corresponding author.

### *7-3- Funding*

The authors received no financial support for the research, authorship, and/or publication of this article.

### *7-4- Acknowledgements*

We would like to acknowledge NOCERAL (National Orthopaedic Centre of Excellence in Research and Learning), Department of Orthopaedic Surgery, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia for the data provided to conduct this study. We would also like to acknowledge Multimedia University (MMU) for all the support provided.

### *7-5- Institutional Review Board Statement*

Ethical clearance number: EA0152022.

### *7-6- Informed Consent Statement*

Informed consent was obtained from the UM collaborator for all subjects involved in the study.

### *7-7- Conflicts of Interest*

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

## 8- References


[1] Merriam-Webster. (2023). Prognosis. "Merriam-Webster.Com" Dictionary. Available online: https://www.merriam-webster.com/dictionary/prognosis (accessed on April 2023).

[2] United Nations. (2014). Health-United Nations Sustainable Development. United Nations, New York, United States. Available online: https://www.un.org/sustainabledevelopment/health/ (accessed on April 2023).

[3] Mackillop, W. J. (2006). The Importance of Prognosis in Cancer Medicine. TNM Online. doi10.1002/0471463736.tnmp01.pub2.

[4] Mirabello, L., Troisi, R. J., & Savage, S. A. (2009). Osteosarcoma incidence and survival rates from 1973 to 2004: Data from the surveillance, epidemiology, and end results program. Cancer, 115(7), 1531–1543. doi:10.1002/cncr.24121.

[5] Chao, C. M., Yu, Y. W., Cheng, B. W., & Kuo, Y. L. (2014). Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree. Journal of Medical Systems, 38(10), 106. doi:10.1007/s10916-014-0106-1.

[6] Roshani, S., Coccia, M., & Mosleh, M. (2022). Sensor Technology for Opening New Pathways in Diagnosis and Therapeutics of Breast, Lung, Colorectal and Prostate Cancer. HighTech and Innovation Journal, 3(3), 356-375. doi:10.28991/HIJ-2022-03-03-010.

[7] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. Technology and Health Care, 24(1), 31–42. doi:10.3233/THC-151071.

[8] Kansara, M., Teng, M. W., Smyth, M. J., & Thomas, D. M. (2014). Translational biology of osteosarcoma. Nature Reviews Cancer, 14(11), 722–735. doi:10.1038/nrc3838.

[9] Misaghi, A., Goldin, A., Awad, M., & Kulidjian, A. A. (2018). Osteosarcoma: A comprehensive review. Sicot-J, 4, 12. doi:10.1051/sicotj/2017028.

[10] Lindsey, B. A., Markel, J. E., & Kleinerman, E. S. (2017). Osteosarcoma Overview. Rheumatology and Therapy, 4(1), 25–43. doi:10.1007/s40744-016-0050-2.

[11] Stewart, B. W., & Wild, C. P. (2014). World Cancer Report 2014. World Health Organization (WHO), Geneva, Switzerland.

[12] Tharakan, S., Raja, I., Pietraru, A., Sarecha, E., Gresita, A., Petcu, E., Ilyas, A., & Hadjiargyrou, M. (2023). The Use of Hydrogels for the Treatment of Bone Osteosarcoma via Localized Drug-Delivery and Tissue Regeneration: A Narrative Review. Gels, 9(4), 274. doi:10.3390/gels9040274.

[13] Tang, J., Wang, J. K., & Pan, X. (2022). A Web-Based Prediction Model for Overall Survival of Elderly Patients with Malignant Bone Tumors: A Population-Based Study. Frontiers in Public Health, 9, 1-12. doi:10.3389/fpubh.2021.812395.

[14] Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. CA: A Cancer Journal for Clinicians, 70(1), 7–30. doi:10.3322/caac.21590.

[15] Jiang, J., Pan, H., Li, M., Qian, B., Lin, X., & Fan, S. (2021). Predictive model for the 5-year survival status of osteosarcoma patients based on the SEER database and XGBoost algorithm. Scientific Reports, 11, 5542. doi:10.1038/s41598-021-85223-4.

[16] Muthaiyah, S., Singh, V.A. (2021). Bone Cancer Survivability Prognosis with KNN and Genetic Algorithms. Concepts and Real-Time Applications of Deep Learning. EAI/Springer Innovations in Communication and Computing. Springer, Cham, Switzerland. doi.org/10.1007/978-3-030-76167-7_8.

[17] Park, K., Ali, A., Kim, D., An, Y., Kim, M., & Shin, H. (2013). Robust predictive model for evaluating breast cancer survivability. Engineering Applications of Artificial Intelligence, 26(9), 2194–2205. doi:10.1016/j.engappai.2013.06.013.

[18] Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., & Altman, D. G. (2013). Prognosis Research Strategy (Progress) 3: Prognostic Model Research. PLoS Medicine, 10(2), e1001381. doi:10.1371/journal.pmed.1001381.

[19] Kyburz, D., & Finckh, A. (2013). The importance of early treatment for the prognosis of rheumatoid arthritis. Swiss Medical Weekly, 143. doi:10.4414/smw.2013.13865.

[20] Harting, M. T., Lally, K. P., Andrassy, R. J., Vaporciyan, A. A., Cox, C. S., Hayes-Jordan, A., & Blakely, M. L. (2010). Age as a prognostic factor for patients with osteosarcoma: An analysis of 438 patients. Journal of Cancer Research and Clinical Oncology, 136(4), 561–570. doi:10.1007/s00432-009-0690-5.

[21] LeCornu, M. G., Chuang, S. K., Kaban, L. B., & August, M. (2011). Osteosarcoma of the jaws: Factors influencing prognosis. Journal of Oral and Maxillofacial Surgery, 69(9), 2368–2375. doi:10.1016/j.joms.2010.10.023.

[22] Duchman, K. R., Gao, Y., & Miller, B. J. (2015). Prognostic factors for survival in patients with high-grade osteosarcoma using the Surveillance, Epidemiology, and End Results (SEER) Program database. Cancer Epidemiology, 39(4), 593–599. doi:10.1016/j.canep.2015.05.001.

[23] Kim, J., & Shin, H. (2013). Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. Journal of the American Medical Informatics Association, 20(4), 613–618. doi:10.1136/amiajnl-2012-001570.

[24] Wang, P., Li, Y., & Reddy, C. K. (2019). Machine Learning for Survival Analysis. ACM Computing Surveys, 51(6), 1–36. doi:10.1145/3214306.

[25] Dudley, W. N., Wickham, R., & Coombs, N. (2016). An Introduction to Survival Statistics: Kaplan-Meier Analysis. Journal of the Advanced Practitioner in Oncology, 7(1), 91-100. doi:10.6004/jadpro.2016.7.1.8.

[26] Stel, V. S., Dekker, F. W., Tripepi, G., Zoccali, C., & Jager, K. J. (2011). Survival analysis I: The Kaplan-Meier method. Nephron - Clinical Practice, 119 (1): c83–c88. doi:10.1159/000324758.

[27] Etikan, I. (2017). The Kaplan Meier Estimate in Survival Analysis. Biometrics & Biostatistics International Journal, 5(2), 55-59. doi:10.15406/bbij.2017.05.00128.

[28] Datema, F. R., Moya, A., Krause, P., Bäck, T., Willmes, L., Langeveld, T., Baatenburg De Jong, R. J., & Blom, H. M. (2012). Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression. Head and Neck, 34(1), 50–58. doi:10.1002/hed.21698.

[29] Li, L., Yang, Z., Hou, Y., & Chen, Z. (2020). Moving beyond the Cox proportional hazards model in survival data analysis: A cervical cancer study. BMJ Open, 10(7), e033965. doi:10.1136/bmjopen-2019-033965.

[30] Matsuo, K., Purushotham, S., Jiang, B., Mandelbaum, R. S., Takiuchi, T., Liu, Y., & Roman, L. D. (2019). Survival outcome prediction in cervical cancer: Cox models vs. deep-learning model. American Journal of Obstetrics and Gynecology, 220(4), 381.e1-381.e14. doi:10.1016/j.ajog.2018.12.030.

[31] Miladinovic, B., Kumar, A., Mhaskar, R., Kim, S., Schonwetter, R., & Djulbegovic, B. (2012). A Flexible Alternative to the Cox Proportional Hazards Model for Assessing the Prognostic Accuracy of Hospice Patient Survival. PLoS ONE, 7(10), e47804. doi:10.1371/journal.pone.0047804.

[32] Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217-222.

[33] Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. Eurasip Journal on Advances in Signal Processing, 2016(1), 2351–8014. doi:10.1186/s13634-016-0355-x.

[34] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. doi:10.1016/j.csbj.2014.11.005.

[35] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease Prediction by Machine Learning over Big Data from Healthcare Communities. IEEE Access, 5, 8869–8879. doi:10.1109/ACCESS.2017.2694446.

[36] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. Journal of Medical Systems, 36(4), 2431–2448. doi:10.1007/s10916-011-9710-5.

[37] Li, Z., Soroushmehr, S. M. R., Hua, Y., Mao, M., Qiu, Y., & Najarian, K. (2017). Classifying osteosarcoma patients using machine learning approaches. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea. doi:10.1109/embc.2017.8036768.