Master Degree Project

UNIVERSITY OF SKÖVDE

Mathematical Modelling Simulation Data and Artificial Intelligence for the Study of Tumour - Macrophage Interaction

Master Degree Project in Bioinformatics
Second Cycle 30 credits
Autumn term 2023

Student: Jaysmita Khanindra Chaliha
a22jaych@student.his.se

Supervisor:
Dario Melguizo Sanchis, University of Skövde
Dario.melguizo.sanchis@his.se

Examiner: Sanja Jurcevic, University of Skövde
sanja.jurcevic@his.se

Abstract

The study explores the integration of mathematical modelling and machine learning to understand tumour-macrophage interactions in the tumour microenvironment. It details mathematical models based on biochemistry and physics for predicting tumour dynamics, highlighting the role of macrophages. Machine learning, particularly unsupervised and supervised techniques like K-means clustering, logistic regression, and support vector machines, are implemented to analyse simulation data. The thesis's integration of K-means clustering reveals distinct tumour behaviour patterns through the classification of tumour cells based on their microenvironmental interactions. This segmentation is crucial for understanding tumour heterogeneity and its implications for treatment. Additionally, the application of logistic regression provides insights into the probability of macrophage polarization states in the tumour microenvironment. This statistical model underscores the significant factors influencing macrophage behaviour and their consequent impact on tumour progression. These analytical approaches enhance the understanding of the complex dynamics within the tumour microenvironment, contributing to more effective tumour study strategies. The study presents a comprehensive analysis of tumour growth, macrophage polarization, and their impact on cancer treatment and prognosis. Ethical considerations and future directions focus on enhancing model accuracy and integrating experimental data for improved cancer diagnosis and treatment strategies. The thesis concludes with the potential of this hybrid approach in advancing cancer biology and therapeutic approaches.

Contents

List of Tables

List of Figures

Background

Mathematical Modelling

Mathematical models based on physics and biochemistry rules offer accurate descriptions of biological processes, enabling researchers to replicate experimental scenarios, develop novel hypotheses and allowing researchers to predict how biological systems behave and test these predictions using experimental data (Alden et al., 2020; Procopio et al., 2023; Salerno et al., 2013, 2015).

Such models have been long used to define biological functions. Previous research has utilised mathematical models to simulate biological assumptions, predict tumour dynamics, and evaluate treatment strategies (Frieboes et al., 2006; Hatzikirou, 2018; Macklin, 2017; Mascheroni et al., 2021a). These models have also contributed to understanding the impact of immune cells, including macrophages, in the tumour microenvironment (TME), while offering predictive capabilities and suggesting experimental directions, (de Pillis et al., 2005; Eftimie et al., 2011; Jansen et al., 2019; Mahlbacher et al., 2019; Makaryan et al., 2020; Shojaee et al., 2022).

Machine Learning

Machine learning, as a sub-branch of artificial intelligence (AI), relies on algorithms that learn from large datasets without explicit programming of rules or complex mechanisms and can predict outcomes or find patterns in input data, making them scalable and efficient tools (Baker et al., 2018; Benzekry, 2020). Machine learning techniques, particularly Deep Learning, have gained popularity in cancer research for diagnosis, prognosis, drug development, and personalised medicine (Benzekry, 2020; Gaw et al., 2019; L. S. Hu et al., 2015, 2017; Korfiatis et al., 2018; Kourou et al., 2015; Z. Li et al., 2017; Magazzù et al., 2022; Prasanna et al., 2017; Xi et al., 2018). AI algorithms have shown promise in various cancer-related tasks, including histology-based image analysis, molecular subtyping, and patient survival prediction (Bychkov et al., 2018; Ching et al., 2018; Courtiol et al., 2019; F. Hu et al., 2020; Huang et al., 2020; Jing et al., 2019; Katzman et al., 2018; Lai et al., 2020; Nir et al., 2019; Ryu et al., 2019; Saillard et al., 2020; Tabibu et al., 2019; K. Wang et al., 2018; Zadeh Shirazi et al., 2020).

Macrophages in Tumour Microenvironment

A complex interaction of several cell types, including cancer cells, immune cells, endothelial cells, and fibroblasts, occurs in the tumour microenvironment (TME). Macrophages, one of the non-tumour stromal cells within TME, play a crucial role in tumour growth, metastasis, as well as drug resistance (DeNardo & Ruffell, 2019; Ngambenjawong et al., 2017). They are traditionally classified into "M1" and "M2" subtypes based on pro-inflammatory and anti-inflammatory functions, but new data suggests a more complex polarisation system beyond the traditional "M1-M2" paradigm (Kerneur et al., 2022; Lantz et al., 2020; Lee et al., 2013; Yang et al., 2021; L. Zhang et al., 2020). Molecules within the tumour microenvironment influence the functional diversity of these particular macrophages, and their presence is often associated with poor prognosis in solid tumours (Gentles et al., 2015; Komohara et al., 2014; Pan et al., 2020; Q. Zhang et al., 2012). As a result, tumour-macrophage interactions have been extensively studied for their roles in tumour immunity and immunotherapy (Xiang et al., 2021).

Problem Definition

While machine learning isolates relevant inputs to predict outputs, mathematical modelling generates hypotheses for causal mechanisms based on observations (Baker et al., 2018). Both mathematical models and AI approaches have limitations that researchers have attempted to address. Mathematical models offer causality, but they rely on a comprehensive understanding of underlying biological mechanisms, fail to efficiently combine large datasets from different sources and different levels of resolution, oversimplifying complex biological processes while lacking the universal predictive capabilities of machine learning, which excels in providing patterns and predictions from data (Alber et al., 2019; Baker et al., 2018). On the other hand, AI algorithms have limitations like interpretability issues, the need for large datasets and ignoring the fundamental laws of physics that can result in ill-posed problems or non-physical solutions (Alber et al., 2019; Baker et al., 2018; Benzekry, 2020; Mascheroni et al., 2021a).

Advances in the field of medicine has enabled us to generate a vast variety of medical data from various different types of studies (Torkamani et al., 2017). Simultaneous, Artificial Intelligence has also made advances with the introduction of artificial neural networks and deep-learning (Y. Li et al., 2018). However, the recently developed machine learning models are dependent on relatively larger data sets (Nickel et al., 2016) and the field of medicine is yet to generate reliable data sufficient for the use of advance AI models (Camacho et al., 2018).

To overcome these limitations, hybrid approaches between mathematical modelling and machine learning have been explored to expand our understanding of cancer, with potential applications in clinical and industrial contexts (Alber et al., 2019; Baker et al., 2018; Peng et al., 2021; Procopio et al., 2023; von Rueden et al., 2020; Yauney & Shah, 2018).

Understanding macrophage subsets in cancer remains challenging, and further studies are needed to discover interactions with other immune cell indicators (Azizi et al., 2018; Chen et al., 2021; Müller et al., 2017; Q. Zhang et al., 2019), Since the most properties of such macrophages were mainly studied in mouse models, further investigation is needed to understand their role in humans due to species differences between rodents and humans (Yan & Wan, 2021; Zhu et al., 2017), imposing a challenges that can be further studied. While hybrid approaches that merge mathematical modelling with machine learning have been recently adopted in tumour biology, their application in understanding of tumour-macrophage interaction remains unexplored.

Problem Motivation

While both the computational techniques address important applications in the field of clinical oncology, it's evident that the research discussed in artificial intelligence and mathematical modeling has been undertaken by various communities (including AI/ML, pharmacometrics, statistics, and mathematicians). However, these communities have made separate advancements without significant collaboration. Uniting these disciplines could undoubtedly result in more robust techniques capable of enhancing our comprehension of cancer (Benzekry, 2020).

Despite their respective limitations, both mathematical modelling and machine learning show synergy where each technique seemingly compensates for the limitations imposed by the other, somehow complimenting the pros and cons of one another and opening exciting opportunities.

Where machine learning reveals correlation, multiscale modeling can probe whether the correlation is causal; where multiscale modeling identifies mechanisms, machine learning, coupled with Bayesian methods, can quantify uncertainty (Alber et al., 2019; Lytton et al., 2017; A. M. Tartakovsky et al., 2018; G. Tartakovsky et al., 2018). Incorporating biological knowledge into models can enhance interpretability, and mathematical approaches can develop physiologically relevant models that link input and output for simulating biological processes and treatments (Benzekry, 2020).

Various mathematical models exist that describe different aspects of tumour-macrophage interactions. In particular, cancer biology and immunology are research intensive fields, where a plethora of biological mechanisms remained to be discovered. As a result, developing a model at the best case involves assumptions of phenomenological terms that account for this lack of knowledge. Therefore, quantitative and personalised predictions are a daunting task. Meanwhile, in order to overcome these problems, data-driven techniques combined with mathematical modelling can improve the prediction accuracy of the models. Such methods have the potential to bring mathematical closer to the clinical practice (Benzekry, 2020; Mascheroni et al., 2021a; Shojaee et al., 2022).

Overview of Previous Research

Hybrid approaches that use mathematical modelling and machine learning together have recently been studied. One such research that studies the effects of radiation on cells and Boolean cancer modeling reveals that, when dealing with limited training data sets, simulation-based kernel methods that utilize approximate simulations to construct a kernel enhance the subsequent machine learning outcomes and outperform conventional machine learning approaches that lack prior knowledge (Alber et al., 2019; Deist et al., 2019).

Another such study conducted to anticipate the metastasis relapse in early-stage breast cancer. Rather than employing a biologically agnostic model for analyzing survival, Nicolò et al. developed a mathematical model to predict the specific timing of relapse. Furthermore, to address the relatively extensive set of covariates (21 in total), the authors turned to machine learning to perform feature selection (Benzekry, 2020; Nicolò et al., 2020).

Alternatively, a combined approach involving systems biology model and machine learning was developed using clinical data to forecast how patients would respond to anti-PD-1 immunotherapy with the goal of enhancing response rates. Through this methodology, the research pinpointed patient response biomarkers and uncovered potential mechanisms behind drug resistance. The model was employed to compute patient-specific kinetic parameters and make predictions about clinical outcomes, demonstrating the advantageous use of transfer learning with simulated clinical data to substantially enhance the accuracy of response predictions (Przedborski et al., 2021).

One specific study focused on clinical tumor predictions has introduced a Bayesian combination of machine learning and mathematical modeling, known as BaM3, designed to enhance clinically relevant predictions. This method leverages predictions from a mathematical model (C2) as intelligent priors, even when only partial knowledge of mechanisms and parameters is available. In addition, it rectifies model predictions by harnessing the predictive capabilities of infrequent non-modelable data (referred to as C1 and C3). The study showcased the potential of BaM3 using a synthetic dataset for glioma and two actual patient cohorts with leukemia and ovarian cancer. The

predictions generated by this approach closely align with real clinical data for individual patients, suggesting its potential utility in facilitating precise personalized clinical predictions (Mascheroni et al., 2021a).

In a recent study, a 3D simulation model was created, encompassing both angiogenesis and tumor growth. This model was employed to determine the concentration of vascular endothelial growth factor and to visually track the development of a microvascular network. Subsequently, the effectiveness of three distinct anti-angiogenic drugs at varying concentrations was assessed. Furthermore, a comprehensive understanding of tumor cell proliferation and endothelial cell angiogenesis mechanisms was put forth to offer precise forecasts for optimizing drug therapies. By employing machine learning techniques, the analysis of simulation output data also revealed additional characteristics, including tumor volume, tumor cell count, and the length of newly formed vessels. These parameters were investigated to gain insights into various stages of tumor growth and to assess the effectiveness of different pharmaceuticals (Mousavi et al., 2022).

Project Plan

Aim

To analyse dataset generated from mathematical model simulation, designed to stimulate tumour microenvironment with the implementation of various machine learning techniques.

Objectives

1. Simulate the chosen mathematical model (Ganguli & Sarkar, 2018) for tumour-macrophage interaction and obtain simulation results for machine learning analysis.
2. Apply unsupervised and supervised machine learning algorithms on simulation data to study the effects of tumour-macrophage interactions on overall tumour microenvironment.
3. Utilize K-means clustering algorithm to identify natural relationships and groupings within the simulation data.
4. Visualize clustering results, interpreting patterns in the dataset and label the data for supervised learning.
5. Implement logistic regression and SVM on the dataset obtained from K-means clustering for classification tasks.
6. Validate the integrated approach using experimental data, comparing predictions with observed outcomes.

Methods Description and Implementation

This study aimed to integrate machine learning techniques with mathematical modelling into a hybrid approach to study tumour-macrophage interaction in the tumour microenvironment. Figure 1. provides an outline of the proposed methodology.

Selection of Mathematical Model

This study aimed to integrate machine learning techniques with mathematical modelling into a hybrid approach to study tumour-macrophage interaction in the tumour microenvironment. Figure 1 provides an outline of the methodology implemented in this study. Appropriate mathematical models, fitting with the goal of this study, were explored to meet the expectations of this study. An extensive literature search was performed to find a suitable and publicly available mathematical model demonstrating the interaction between macrophages and tumour cells within the tumour microenvironment. Also, the BioModels Database (Glont et al., 2018; Malik-Sheriff et al., 2020) was searched to find manually curated and publicly available mathematical models to serve this study. The model was selected based on the parameters, variables, and factors used to build the respective model.

Based on all the above criteria, the mathematical model defined by (Ganguli & Sarkar, 2018) was used for this study. This model explored the interactions between the tumour microenvironment, composed of immune cells and cytokines, and the heterogeneous population of tumour cells originating from Cancer Stem Cells, giving the model a unique characteristic that expanded beyond studying tumour growth solely. The model attempted to encompass the interaction between the tumour and the immune system regarding cancer stem cell differentiation. Not only that, but the model also tested known treatment strategies and proposed improved protocols for potentially better cancer remission outcomes, along with addressing the processes behind cancer progression (Ganguli & Sarkar, 2018).

Mathematical Model Simulation

The purpose of this mathematical model simulation was to assess the data concerning each step of tumour-macrophage interaction, measuring how tumour-associated macrophages and their polarization could affect the tumour at different levels of growth while predicting the outcome in terms of overall tumour growth. The ordinary differential equation based mathematical model developed by (Ganguli & Sarkar, 2018), implemented in this study, mainly focused on tumour-macrophage interactions. The mathematical model obtained was tested and validated with experimental data. CellDesigner 4.4.2 (Funahashi et al., 2006, 2007), a specialized modelling platform, was used to simulate the mathematical model and to obtain results from the same. Validation method AUC, ROC and K-fold Cross validation were conducted on the obtained results.

Macrophages perform diverse functions, including pathogen phagocytosis, antigen presentation through major histocompatibility complex molecules, and cytokine production like IL-1, IL-6, and TNF-α (Sica et al., 2015; Wynn et al., 2013). Their identity and activities were influenced by factors such as developmental origin, tissue residence, and acute microenvironmental cues, making them a diverse collection of cell types with various functional roles in both homeostasis and pathological conditions (Bonnardel & Guilliams, 2018; DeNardo & Ruffell, 2019; Epelman et al., 2014; Ginhoux & Guilliams, 2016; Lavin et al., 2015). Although essential for defending against microorganisms, macrophages were also associated with autoimmune diseases and tumours, and they could contribute to tissue damage during infections and inflammatory diseases (Bashir et al., 2016; Benoit et al., 2008; Beschin et al., 2013; Dall'Asta et al., 2012; Das et al., 2015; Davies et al., 2013; Eguchi &

Manabe, 2013; Ginhoux & Jung, 2014; Kadomoto et al., 2021; Leopold Wager et al., 2015; Mantovani et al., 2013; Mège et al., 2011; Patel et al., 2017; Shapouri-Moghaddam et al., 2018; Shi & Pamer, 2011).

A complex interaction of several cell types, including cancer, immune, endothelial, and fibroblast cells, occurred in the tumour microenvironment. Tumour-associated macrophages, one of the non-tumour stromal cells within tumour microenvironment, played a crucial role in tumour growth, metastasis, and drug resistance (DeNardo & Ruffell, 2019; Ngambenjawong et al., 2017). Molecules within the tumour microenvironment influenced the functional diversity of t umour-associated macrophages, and their presence was often associated with poor prognosis in solid tumours (Gentles et al., 2015; Glont et al., 2018; Komohara et al., 2014; Q. Zhang et al., 2012). Tumour-associated macrophages were traditionally classified into "M1" and "M2" subtypes based on pro-inflammatory and anti-inflammatory functions (Kerneur et al., 2022; Lantz et al., 2020; Lee et al., 2013; Yang et al., 2021; L. Zhang et al., 2020). As a result, TAMs were extensively studied for their roles in tumour immunity and immunotherapy (Xiang et al., 2021).

The simulation in this study followed the initial parameter and variable values stated by (Ganguli & Sarkar, 2018). Simulation was carried out through ControlPanel of CellDesigner 4.4.2 software, where the Solver was set as SimulationCore and Error Tolerance was set as -6, which is the default value for the same. The model was simulated to replicate the temporal evolution and growth kinetics of tumour using the parameters discussed in their research. As such, the simulation was carried out for a time equivalent to 800 days until the system reached equilibrium, as stated in their study. The change in cellular concentrations of various cells was obtained, this included cancer cells, stem cells along with major immune cells and cytokines. Variables associated with M1 and M2 macrophages as well as cytokine IL10 were monitored closely to provide further information on tumour-associated macrophages and to study the regulatory behaviour of macrophages in the tumour microenvironment.

Data Collection

The results obtained from the mathematical model simulation were analysed and interpreted. The datasets obtained from the mathematical simulation represented various aspects of tumour growth and progression with the composition of the tumour microenvironment. For this study, data relevant to tumour-associated macrophages were given the main focus. The differential regulatory behaviour of type I and type II TAMs on the tumour cell data could be obtained by varying the specific parameters that governed the growth rate of the M1 and M2 macrophages. The M1:M2 ratio concerning tumour progression was the primary data in this study. Also, the production of cytokine IL10 and its concentration were taken into consideration due to its involvement with M2 macrophages.

The data collection method was based on the (Mousavi et al., 2022) protocol. For each time step, 1 minute in this simulation, a snapshot of tumour growth and changes in the microenvironment was collected in vital parameters of the growth process. To this end, three different simulation runs were performed, each for 800 days, to study the temporal evolution of the tumour and its microenvironment, from the early stages of tumour development to a supposedly steady state tumour growth stage. Each step (in seconds) provided the M1:M2 ratio and changes in IL10 concentration throughout the tumour development, growth, and progression.

This data collected from the mathematical model simulation was used as datasets for the machine learning algorithm later in the study. The data was reviewed to ensure the information's integrity and eliminate missing or incomplete data. This was carried out in R 4.3.2 and RStudio 2023.09.1 Build 494, where the data was checked for any missing values and data was analyse using elbow method

prior to labelling. However, the data was not preprocessed before clustering analysis, as it was obtained from a simulation and was considered to be complete and consistent.

The dataset size held significant importance for learning accuracy. In line with the requirements of machine learning, it was necessary to develop a simulation to generate an appropriate dataset for input into ML algorithms. Even though the trained algorithms contributed to reducing overall computation time, the duration of the simulation remained crucial in the dataset creation process (Mousavi et al., 2022). As a result, the simulation was configured with time steps set at 2 hours over 800 days. This configuration ensured the availability of sufficient tumour data at each time step.

Based on these assumptions, the dataset encompassed 9600-time steps. Each time step included information on the various interactions taking place within the tumour microenvironment, virtually, over the period of time stated above. The participating cells were deemed as variables for the given data.

Machine Learning Implementation

Implementing a machine learning algorithm aimed to address concerns about the interactions between tumours and macrophages in the tumour microenvironment without delving into the complexities of simulation models, and it was based on the (Mousavi et al., 2022) protocol. As previously mentioned, the primary input for the machine learning models was the mathematical simulation output, providing tumour data at various time steps during the simulation. To achieve the stated goals of the machine learning implementation, it was decided to employ two fundamental types of AI methods: supervised and unsupervised learning.

The primary motivation is to utilize machine learning for extracting meaningful insights from complex biological data. This is particularly crucial in understanding the dynamic tumor environment, which is characterized by a myriad of interactions and changes. The study aims to provide a nuanced understanding of these interactions, contributing to the broader field of oncology research. Also, machine learning is a suitable technique to analyze large datasets that can be generated by mathematical models' simulation, as these techniques are capable of handling large amount of data in a relatively short interval of time.

Unsupervised Learning: K-means Clustering

Unsupervised learning focuses on identifying natural relationships and groupings within the data without referencing specific outcomes (Bi et al., 2019). The approach aimed to discover patterns and similarities in the data. K-means is renowned for its straightforward implementation and computational efficiency, making it a practical choice for large datasets typically found in biological studies. This algorithm excels in identifying hidden patterns and natural groupings in unlabeled data, a crucial aspect when dealing with complex biological systems like tumor microenvironments. Given the potential for extensive datasets in tumor studies, K-means is ideal for its scalability and robustness in handling vast amounts of data.

K-Means is an unsupervised learning algorithm used for clustering. It partitions the dataset into K distinct, non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far apart) as possible. It does this by assigning data points to the cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum (Yadav & Sharma, 2012).

K-means clustering was used to perform feature selection and identify clusters in the simulation data, which was further used to label the dataset obtained from the mathematical model simulation before proceeding to supervised learning. The need to structure and simplify the simulation output for further analysis underpins the choice of K-means, as it effectively organizes data into coherent clusters that can be more easily interpreted and labeled for supervised learning.

The package cluster (Maechler et al., 2015) was implemented in this study to perform K-means clustering and to find the clusters. The number of cluster (K) was set to 2 after the dataset was analysed using elbow method and silhouette analysis. Principle component analysis was carried out to reduce the dimensionality of the dataset for efficient analysis of the same. The clustering results after analysis the simulation data was visualized through various plots and graphs for further interpretation of the results. The code for K-Means Clustering is provided in Appendix 1.

Supervised Learning: Support Vector Machines

Supervised learning is beneficial for estimating the machine learning model's proficiency, as it worked well when the outcome was known for each observation. This study implemented the commonly used supervised learning method, Logistic Regression and Support Vector Machines (SVM) (Bi et al., 2019). The results obtained from K-Means Clustering was used as the dataset to perform supervised learning. This was done as the dataset obtained from the mathematical model simulation was unlabelled and clustering this dataset provided this study with relevant data labels to carry out supervised learning.

SVM is a supervised learning model used for classification and regression analysis. The goal of the SVM algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly classifies the data points. To separate two classes of data points, there are many possible hyperplanes that could be chosen. The optimal hyperplane is the one that has the largest margin, i.e., the maximum distance between data points of both classes. SVM is effective in high-dimensional spaces and in situations where the number of dimensions exceeds the number of samples (Abaszade & Effati, 2018).

Support Vector Machine and its ability to perform both classification and regression makes SVM a versatile tool for various aspects of tumor analysis. For the binary classification task in the context of this thesis, the implementation of Support Vector Machines (SVM) is advocated. SVM is highly effective in high-dimensional spaces, typical in genetic and molecular data, ensuring robust performance even with a large number of features. The use of the Radial Basis Function (RBF) kernel allows for handling non-linear relationships in the data, which is a common scenario in complex biological interactions. Prudent preprocessing steps, including feature scaling and rigorous model validation using K-fold cross-validation (fold set at 5) was implemented to the data for the application of SVM. SVM kernel was set as Radial Basis Function and the packages used for the SVM were e1071, caret, pROC.The code for SVM is provided in Appendix 1.

Supervised Learning: Logistic Regression

Both SVM and logistic regression were used on the dataset to compare their performances and results. Logistic Regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes. It uses a logistic function to model a binary dependent variable, although it can be extended to model several classes of events. Logistic regression fits a special s-shaped curve by taking the linear regression and transforming the numeric estimate into a probability with the logistic function, which is an S-shaped curve (Fleiss et al., 1986).

Logistic Regression is particularly apt for binary classification tasks, which is common in medical studies where outcomes are often dichotomous (e.g., presence or absence of a tumor response), assuming a linear relationship between predictors and the log-odds of the outcome. This algorithm provides clear and interpretable results, crucial in medical research for understanding and communicating the relationship between variables. Assuming a linear relationship between variables, logistic regression could provide accurate and interpretable results in a computationally efficient and simple way. Logistic Regression not only classifies data but also provides probabilities, offering a nuanced view of the tumor-macrophage interactions, which is valuable in uncertain biological contexts. This was done to learn more about the nature of the data obtained from the mathematical model. Scaling and k-fold cross validation (fold set to 5) was also for logistic regression used to pre-process the data before analysis and result visualization. The packages used for logistic regression were tidyverse, caret, glmnet, pROC. The code for logistic regression is provided in Appendix 1.

Relevant Methods Not Chosen for this Project

Baker et al. (2018) proposed two synergistic approaches, combining mathematical modeling and machine learning. The first involved using machine learning to create surrogate models for computationally demanding multiscale simulations, expediting future predictions. The second approach enriched machine learning-based pipelines by incorporating derived parameters from a mathematical approach, enhancing probabilistic models (Baker et al., 2018; Binder et al., 1997; Xu et al., 2012). This project implemented only the first approach due to challenges in establishing the second approach in tumor-related studies. Limited data availability in public databases and the need for a deeper understanding of macrophage polarization were key considerations (Baker et al., 2018). Reinforcement learning technique, specifically the Q-learning algorithm, was initially considered but deemed incompatible due to the static nature of the dataset implemented in this study (Bi et al., 2019). Mascheroni et al. (2021b) introduced the Bayesian combination, a novel method for personalized tumor growth prediction, integrating mathematical modeling and machine learning in a Bayesian framework. However, it was unsuitable for our project due to the need to replace the existing model and its computational intensity (Mascheroni et al., 2021b). Deep learning, based on artificial neural networks, was another compatible technique considered but faced challenges within our project's time constraints (Bi et al., 2019). Various mathematical models describe tumor-macrophage interactions in the tumor microenvironment (den Breems & Eftimie, 2016; Eftimie & Eftimie, 2019; X. Li et al., 2019). However, these were deemed unsuitable due to unavailability in databases like BioModels (Glont et al., 2018; Malik-Sheriff et al., 2020). The selected model met these criteria for inclusion in our project. The simulation data was also run through using linear regression (Appendix 4) and neural networks (Appendix 5). However, no further data analyses were carried out as the initial goal of the project was already achieved and also due to the time constrains regarding the project.

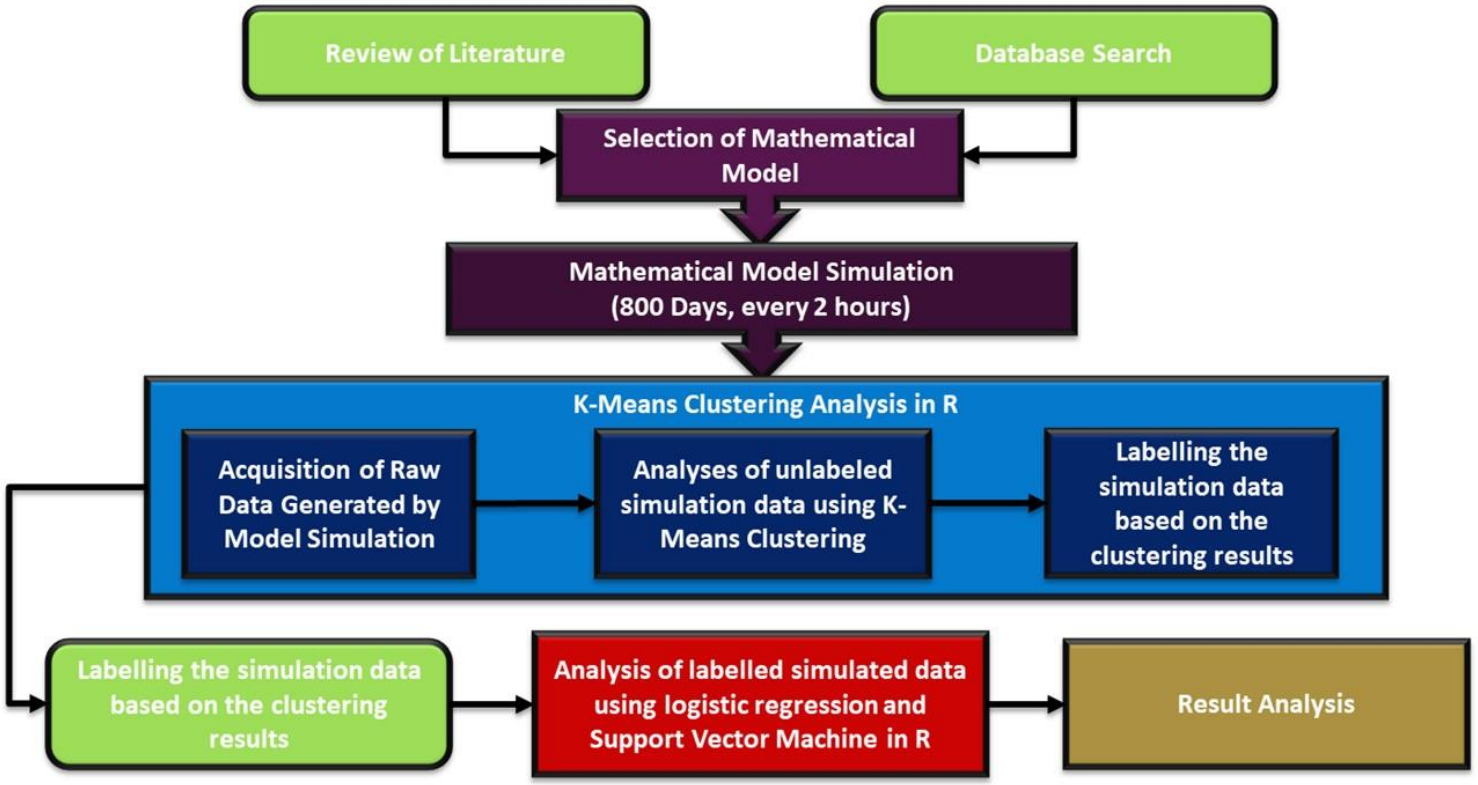

Figure 1: Outline of the methodology implemented in this study; Flowchart depicting data analysis progression: from literature review to database search, model selection, simulation, clustering, labelling, and final analysis with logistic regression and SVM in R.

Results and Discussion

Mathematical Model Simulation

The mathematical model chosen for this study had 14 variables, each corresponding to a cell type, as mentioned in Table 1. The variables/cell type and the initial cell concentration of each respective cell type at the beginning of the simulation for each variable are provided in Table 1. Various interactions between the variables are shown in Figure 2. The simulation lasted for a period that is equivalent to 800 days. Data for cellular concentration was recorded once every 2 hours. Thus, each variable consisted of data from 9600-time stamps. The results of this simulation are plotted in Figure 3 and Figure 4. The data obtained from this simulation was found to be linear and longitudinal in structure.

Upon examination of the simulation result, it was found that data represented a comprehensive array of variables relevant to cancer research and immune responses. The values of Cancer Stem Cells and Cancer Cells showcase a broad spectrum, indicating significant heterogeneity in cell populations. An exponential increase is predicted in cancer stem cells during the latter half of the simulation, reflecting a potential increase in cancer stem cells, which could indicate proliferation or treatment resistance.

In parallel, the variables Resistant Stem Cells and Resistant Cancer Cells shed light on the resistance distribution, offering insights into potential variations concerning treatment resistance within both stem and cancer cell populations. An increased concentration of these two cell types throughout simulation indicates potential challenges in treatment effectiveness or the development of resistance mechanisms.

Turning attention to the immune components, Tumor-Associated Macrophages (M1 and M2) exhibit substantial, underscoring their prominent role in the tumor microenvironment, providing nuanced insights into the inflammatory or anti-inflammatory states within the tumor throughout the simulation. The simulation predicts M2 concentration to increase with cancer stem cells, suggesting a potential shift towards an immunosuppressive environment, which could impact the immune system's ability to mount an effective anti-tumor response.

Similarly, the counts of T Helper Cells (T_H1 and T_H2) appear balanced, which also partake in inflammatory or anti-inflammatory states within the tumor. An increasing T_H2 is observed in the middle phase of the simulation, suggesting a potential shift towards an anti-inflammatory response, indicating a complex interplay between immune components that may influence cancer progression.

Cytotoxic T Cells (T_C) feature prominently in the dataset, suggesting a robust immune response against cancer cells. An increase in cytotoxic cells in the middle phase of the simulation is observed along with TH2, suggesting a rise in cytotoxic immune response, which is crucial for evaluating the immune system's capacity to target and eliminate cancer cells.

Concurrently, Regulatory T Cells (T-reg) were present, emphasizing the importance of maintaining a balance between an effective immune response and immune tolerance within the tumor microenvironment. A widening range of values is observed for this cell type in the simulation over time, suggesting variability in the degree of immunosuppression, emphasizing the complex interplay between regulatory T cells and the overall immune response.

Further examination delves into specific immune factors such as Interferon-gamma and Cytokines (IL10 and IL2), with their respective concentrations providing insights into the strength of immune responses and potential immunosuppressive conditions. However, it is crucial to interpret these findings in the broader biological context, considering the multifaceted roles these factors play in modulating immune responses. The concentration of interferon-gamma and IL2 are observed to be increasing in the middle phase of the simulation, suggesting immune activation, and indicating the potential for enhanced antitumor immune responses.

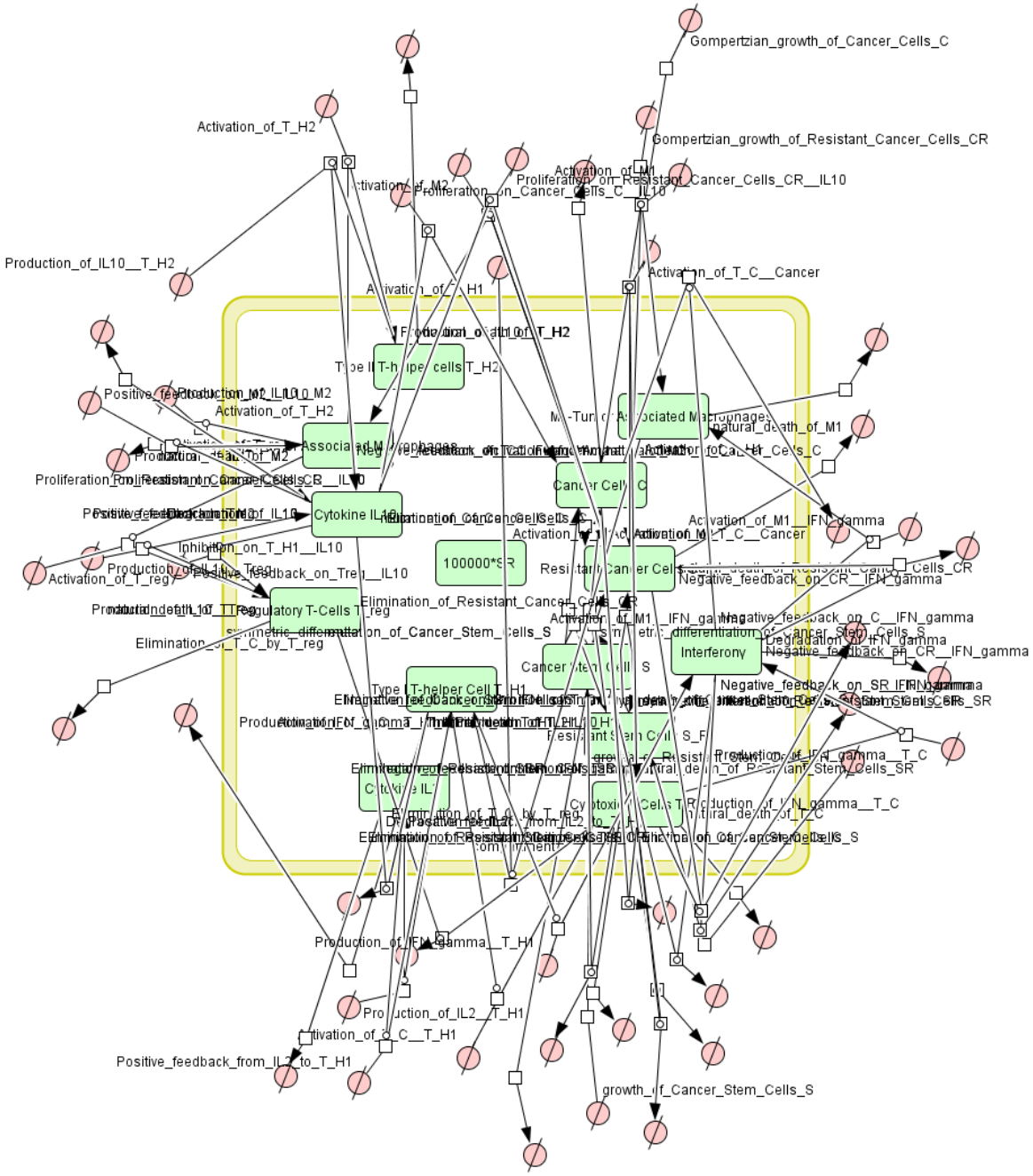

Figure 2. The interactions present in the mathematical model by (Ganguli & Sarkar, 2018). Obtained from Cell Designer software; Network diagram depicting the intricate web of interactions among cell populations and signalling molecules

within the tumor microenvironment model, highlighting the regulatory circuits, growth factors, and feedback loops that govern tumor progression and immune response over time.

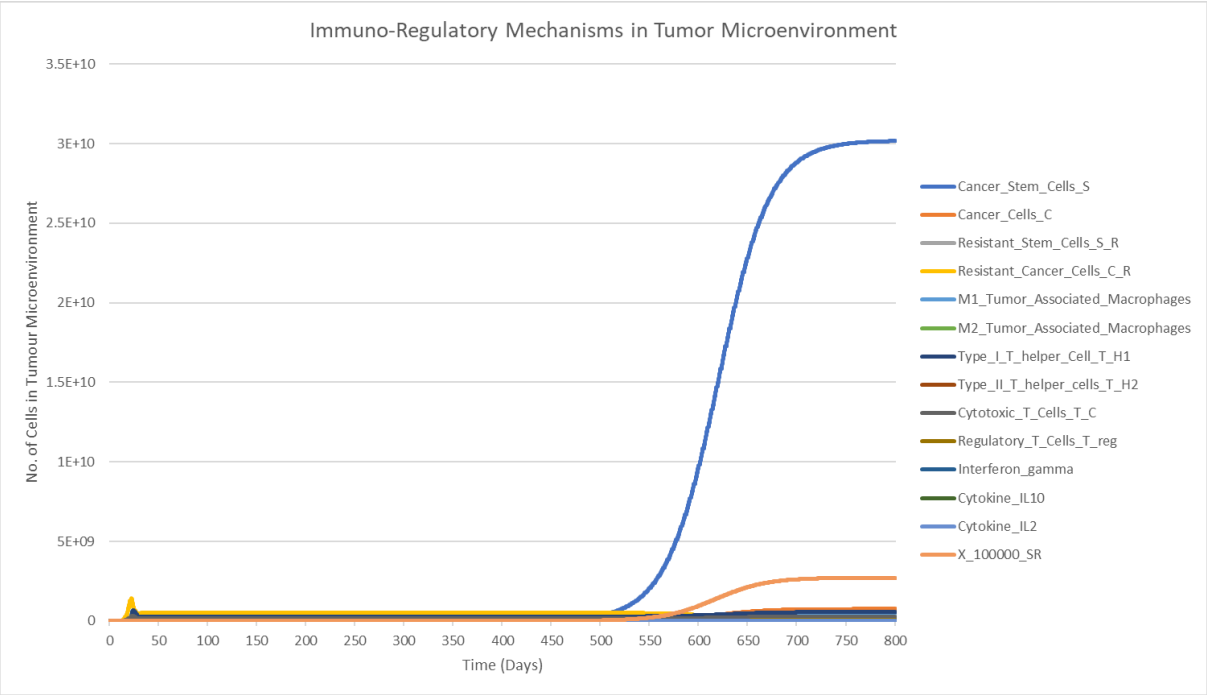

Figure 3. Change in concentration of different cells and cytokines in the tumour microenvironment with time (days); Graphical representation of the dynamic interplay between various cell populations and cytokines in the tumour microenvironment over an 800-day period, highlighting the exponential increase of cancer stem cells post-550 days and illustrating the complex immunological landscape within cancerous tissue.

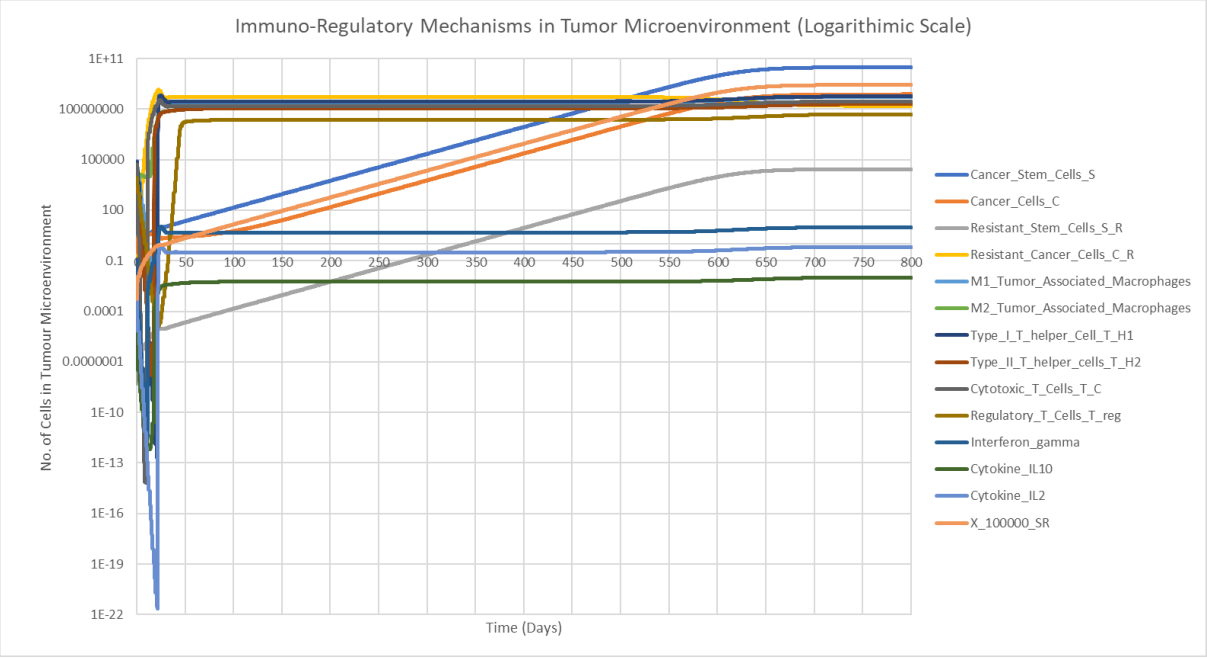

Figure 4. Change in concentration of different cells and cytokines in the tumour microenvironment with time (days) in logarithmic scale; Logarithmic scale visualization of cell and cytokine dynamics in the tumor microenvironment over 800 days, showcasing initial rapid changes and subsequent stabilization across various cell types, including cancer stem cells, resistant cells, and immune cells, emphasizing the nuanced evolution of tumor-immune interactions.

Table 1: Table of Model Variables and their respective initial values (Ganguli & Sarkar, 2018); Reference table for initial conditions in a mathematical model simulating tumor microenvironment, listing cell types and cytokines with corresponding symbols, as denoted in the graphical simulations, along with their respective initial quantities at the start of the 800-day period.

Variable	Symbol	Variable Name on Graph and Model	Initial Values
Stem Cell	S	Cancer_Stem_Cells_S	1
Stem Resistant Cell	S_R	Resistant_Stem_Cells_S_R	0
Cancer Cell	C	Cancer_Cells_C	0
Cancer Resistant Cell	C_R	Resistant_Cancer_Cells_C_R	0
Type-I Tumor Associated Macrophage	M1	M1_Tumor_Associated_Macrophages	85000
Type-II Tumor Associated Macrophage	M2	M2_Tumor_Associated_Macrophages	15000
Type-I Helper T Cell	T_{H1}	Type_I_T_helper_Cell_T_H1	71000
Type-II Helper T Cell	T_{H2}	Type_II_T_helper_cells_T_H2	12000
Cytotoxic T Cell	Tc	Cytotoxic_T_Cells_T_C	56000
Regulatory T Cell	Treg	Regulatory_T_Cells_T_reg	8000
Interleukin-10	IL10	Cytokine_IL10	0.0085
Interferon-γ	IFN-γ	Interferon_gamma	0.12
Interleukin-2	IL2	Cytokine_IL2	0.0094

Unsupervised Learning: K-Means Clustering

The k-means clustering results reveal two distinct clusters in the dataset: Cluster 1 and Cluster 2. Cluster 1 contains 2211 samples, while Cluster 2 is substantially more significant with 7390 samples, indicating an inherent imbalance in the cluster sizes and suggesting that the clustering algorithm is biased towards forming larger clusters. The clusters were formed on the basis of time.

Upon closer examination of the cluster, several noteworthy patterns emerge (Figure 7). In Cluster 1, the count of Cancer Stem Cells (Cancer_Stem_Cells_S) is significantly higher than in Cluster 2, suggesting a potential distinction in the stem cell population between the two clusters. Moreover, the counts of M1 and M2 Tumor-Associated Macrophages are notably elevated in Cluster 1, indicating a potential variation in the immune response. Likewise, Cluster 1 exhibits higher counts for both Type I and Type II T helper cells and cytotoxic T cells, suggesting a more active immune response in this cluster.

The overall analysis suggests that the clusters represent distinct biological profiles, with Cluster 1 potentially associated with a more active immune response and a higher population of cancer stem cells. The clusters may either indicate transition from healthy to disease state or early to late stages of cancer proliferation. The time-based clusters are diagrammatically represented in Figure 5.

The analysis of heatmap of clustering results (Figure 9) reveals diverse ranges and distributions among key features. Notably, features such as Cancer_Stem_Cells_S, and Resistant_Cancer_Cells_C_R exhibit wide-ranging values, indicating substantial variability. The increase in Cancer_Stem_Cells_S seems to correlate with the formation of cluster 1 as interpreted by the K-means clustering model. On analysis of the correlation matrix (Figure 6), we can immediately observe that the variable Resistant_Cancer_Cells_C_R has negative correlations with all the other variables suggesting an inverse relation between it and all the other variables, especially Resistan_Cancer_Stem_Cells_S_R, Cancer_Stem_Cells_S and Cancer_Cells_C.

The PCA results show the importance of each principal component. PC1 explains the most variance (89.94%), followed by PC2 (6.66%) (Figure 10). The cumulative proportion of variance reaches 99.02% by the third principal component. The low variance explained by the later components suggests that the data may be well-represented in lower dimensions.

However, the clustering results reveal notable imbalances, with one cluster containing 2211 data points and the other 7390, suggesting potential sensitivity to the initial conditions of the K-means algorithm or an inherent imbalance in the underlying data distribution (Figure 7). Analysis of clusters indicates significant differences in cell concentration, particularly in features related to cancer stem cells and resistant cell types. Upon inspection of the excel sheet generated by the algorithm, the clustering was found to occur at the time stamp that corresponds to 615 days and 20 hours, may suggest a varied state, possibly indicating phenotypic heterogeneity or a transitional phase in tumor evolution.

Interpretation of K-Means clustering results with respect to tumor-macrophage interaction

The k-means clustering results highlight intriguing patterns in the distribution of different cancer cell types and macrophages across the identified clusters. In Cluster 1, there is a pronounced elevation in

the count of Cancer Stem Cells. This finding suggests that Cluster 1 may be associated with a higher proportion of cells with stem cell properties, which could have implications for the aggressiveness and therapeutic response of the tumor. Cancer stem cells are often linked to tumor initiation, progression, and resistance to treatments, making their abundance a critical factor in understanding the nature of the identified clusters (Ahmed et al., 2018; Atashzar et al., 2020; Najafi et al., 2020).

Furthermore, the increased counts of M1 and M2 Tumor-Associated Macrophages (TAMs) in Cluster 1 introduce an intriguing aspect of the tumor microenvironment. Both M1 and M2 TAMs imply a complex interplay between pro-inflammatory and anti-inflammatory responses. This balance is crucial in shaping the immune landscape within the tumor, impacting tumor progression and therapeutic outcomes (Pan et al., 2020; Rakaee et al., 2019). The elevated levels of these macrophage subtypes in Cluster 1 suggest a more dynamic and interactive immune response within this cluster.

Moreover, the higher counts of Type I and Type II T helper cells and cytotoxic T cells in Cluster 1 point towards a potentially more robust antitumor immune response. These immune cell types play pivotal roles in recognizing and eliminating cancer cells (Cachot et al., 2021; Farhood et al., 2019; Jeong et al., 2023; Montfort et al., 2017). The increased presence of these cells in Cluster 1 may indicate a more favorable immunological environment for combating the tumor.

Additionally, In Cluster 1, an elevated count of Cancer Stem Cells (Cancer_Stem_Cells_S) is accompanied with increase in IL10 expression introduces an intriguing aspect. The higher levels of IL10 may indicate an immunosuppressive microenvironment, as IL10 is known to suppress the activity of immune cells, particularly T cells and natural killer cells. This suggests a potential mechanism by which the tumor creates an immunosuppressive niche, allowing cancer stem cells to evade immune surveillance and promoting tumor progression (Mittal & Roche, 2015; Qiao et al., 2019).

The differential expression of IL10 is also closely tied to the M1 and M2 Tumor-Associated Macrophages (TAMs) observed in Cluster 1. M2 TAMs are generally associated with an immunosuppressive phenotype, and their presence and elevated IL10 could synergistically contribute to establishing an immunosuppressive microenvironment. This has implications for therapeutic strategies, as an immunosuppressed tumor microenvironment is often resistant to immunotherapies (F. Wang et al., 2018).

However, it is crucial to approach these findings cautiously due to the imbalanced cluster sizes, as Cluster 2 is substantially larger. The observed patterns may partly be influenced by the algorithm's bias towards forming larger clusters, and validation through alternative clustering methods or statistical techniques is warranted. Also, the increased count/concentration of the variable Cancer_Stem_Cells_S overshadowed the other variables with significantly lower count/concentration, which hindered the visualization of these variables.

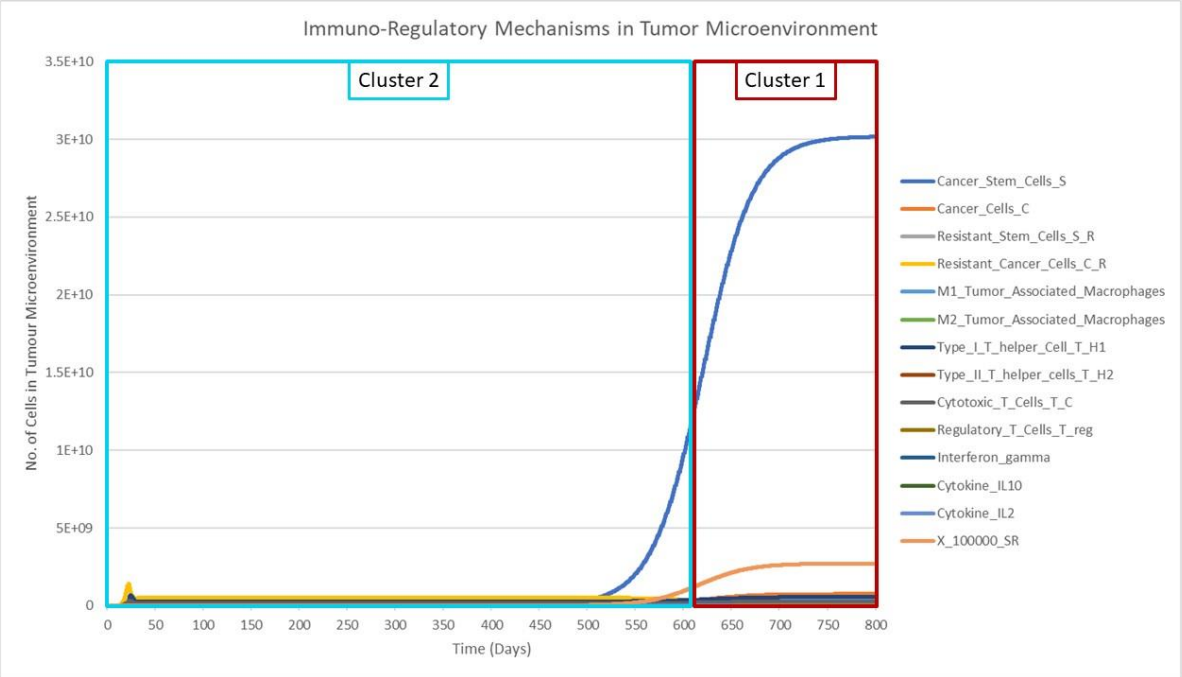

Figure 5: Dynamic Profiles of Cellular and Immune Factors Over Time in Two Clusters from Tumor Microenvironment Analysis; Graphical representation of the dynamic interplay between various cell populations and cytokines in the tumour microenvironment over an 800-day period, differentiated into two clusters. Cluster 2 (blue) exhibits a relatively stable profile, suggesting a controlled tumor environment or a dormant phase, whereas Cluster 1 (red) shows a marked escalation around day 616, indicative of a potential aggressive tumor growth phase or a significant shift in the microenvironmental dynamics.

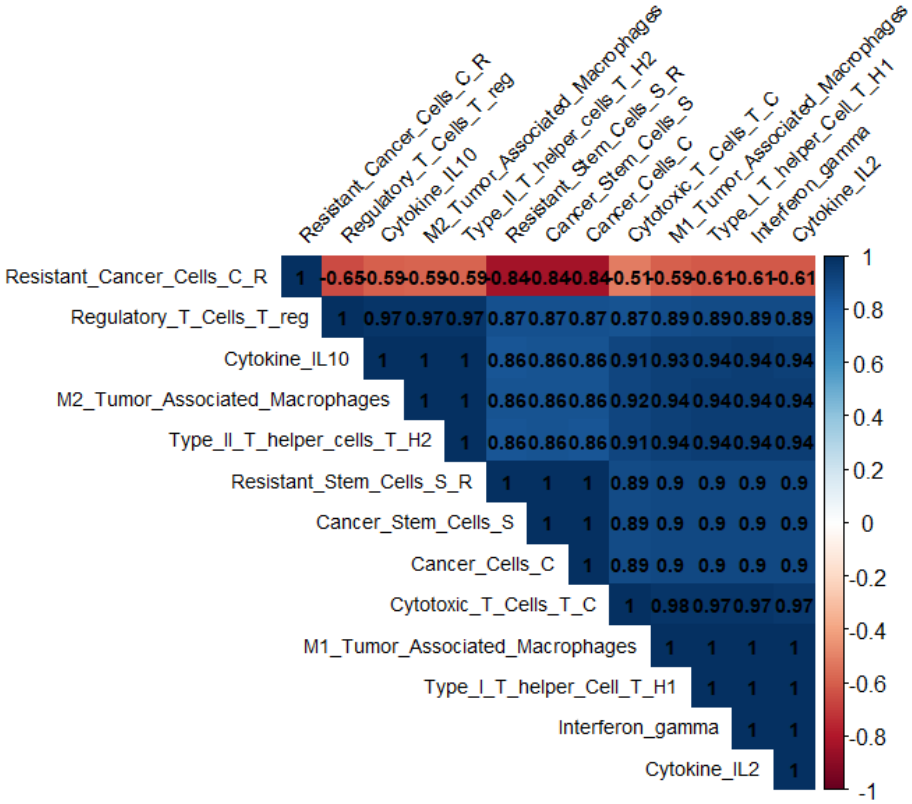

Figure 6. Correlation Matrix of Variables present in the simulation dataset. Correlation matrix showcasing the strength of associations between different cell types and cytokines within the tumour microenvironment, as revealed by k-means clustering, with values close to 1 indicating strong positive relationships, values near -1 indicating strong negative

relationships, and values around 0 suggesting no linear correlation. Highlighted anomalies in the matrix suggest unusual or unexpected associations, such as the negatively correlated clusters of Resistant_Cancer_Cells_C_R, especially with the variables Cancer_Stem_Cells_S, Resistant_Stem_Cells_S_R, Cancer_Cells_C that could reveal counterintuitive or regulatory interactions pivotal to understanding the complex biology of tumour immunology.

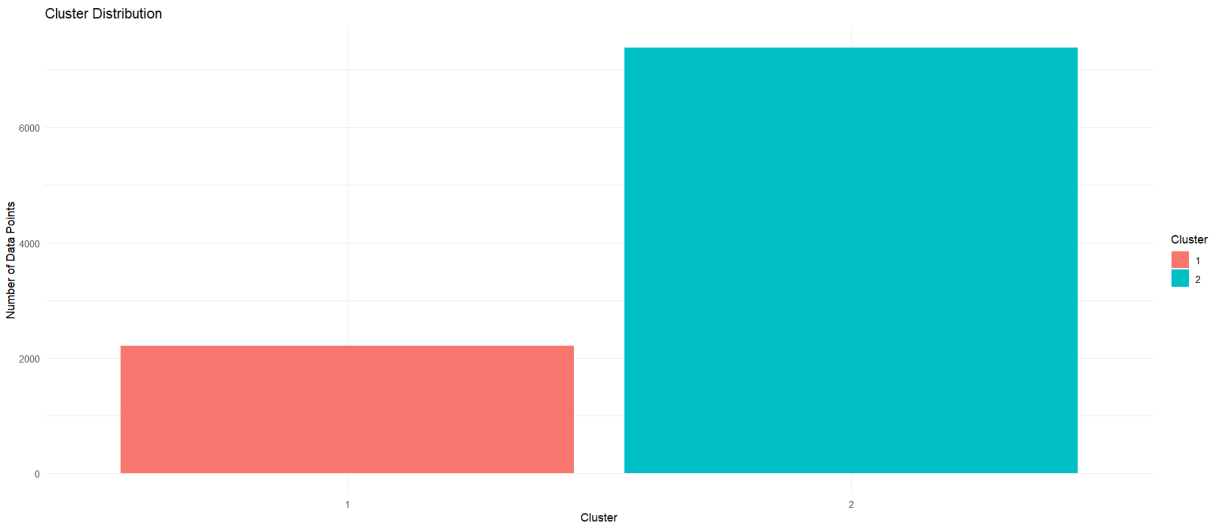

Figure 7. K-Means Clustering of Tumor Data: Visualizing Major and Minor Groups; This bar chart showcases the results of a K-Means clustering algorithm applied to tumor microenvironment data. It features two bars: a shorter red bar for Cluster 1 on the left, and a taller cyan bar for Cluster 2 on the right. The x-axis represents the two clusters, while the y-axis shows the number of data points in each. The noticeable size difference between the clusters (less than 3000 points in Cluster 1 and around 7000 in Cluster 2) suggests significant variance in data distribution, potentially indicating different tumor behaviors or patient groups. This disparity offers valuable insights into the tumor microenvironment, aiding in tailored therapeutic strategies and further biological research.

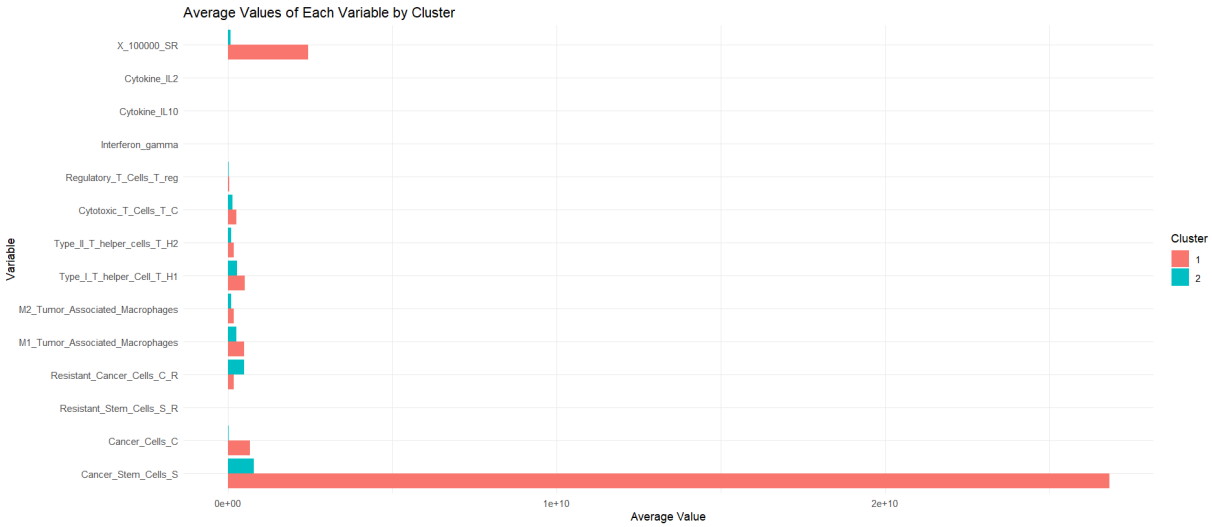

Figure 8. Comparison of Variable Means Between Two Clusters in K-Means Clustering of Tumor Microenvironment Data; This chart presents a comparison of variable means in two clusters identified by K-Means clustering in tumor microenvironment data. Each row on the y-axis signifies a distinct variable and the x-axis displays the average values of these variables in each cluster, with Cluster 1 (red) showing significantly higher averages for most variables compared to Cluster 2 (cyan). Notably, Cluster 1 is characterized by a high prevalence of cancer stem cells.

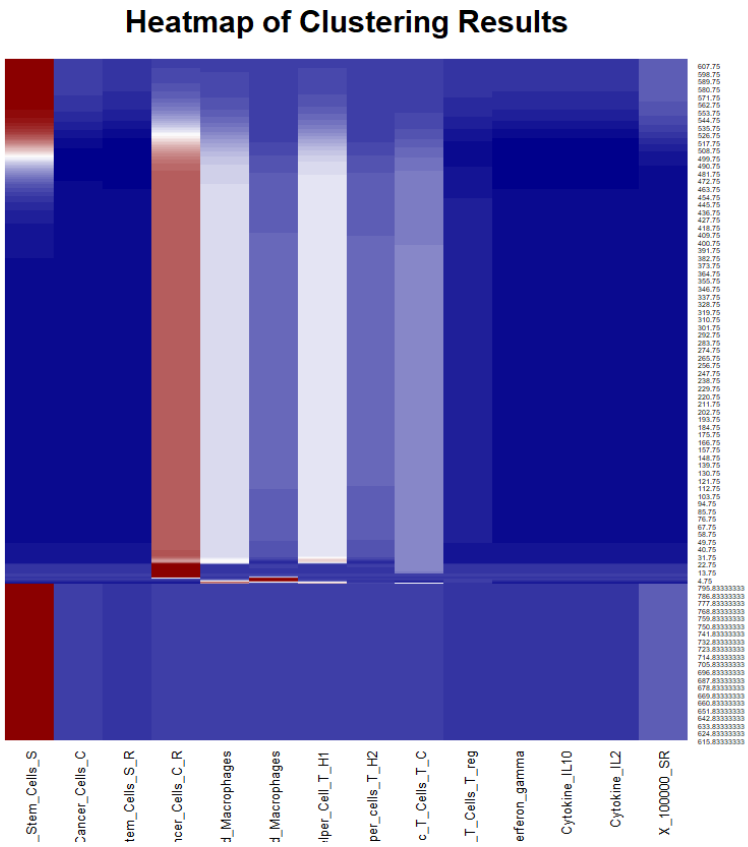

Figure 9: Heatmap of K-Means Clustering Results (x = variables, y = observation/data points) Highlighting Variability in Tumour Microenvironment; This heatmap visualizes the distribution of key cellular features across different samples, using a blue-white-red gradient to indicate low to high values, respectively. Notably, 'Cancer_Stem_Cells_S' and 'Resistant_Cancer_Cells_C_R' show significant variability, suggesting distinct biological behaviors. The clustering pattern, particularly the formation of cluster 1, demonstrates a correlation with increased levels of 'Cancer_Stem_Cells_S', which may have implications for understanding cancer progression and resistance mechanisms.

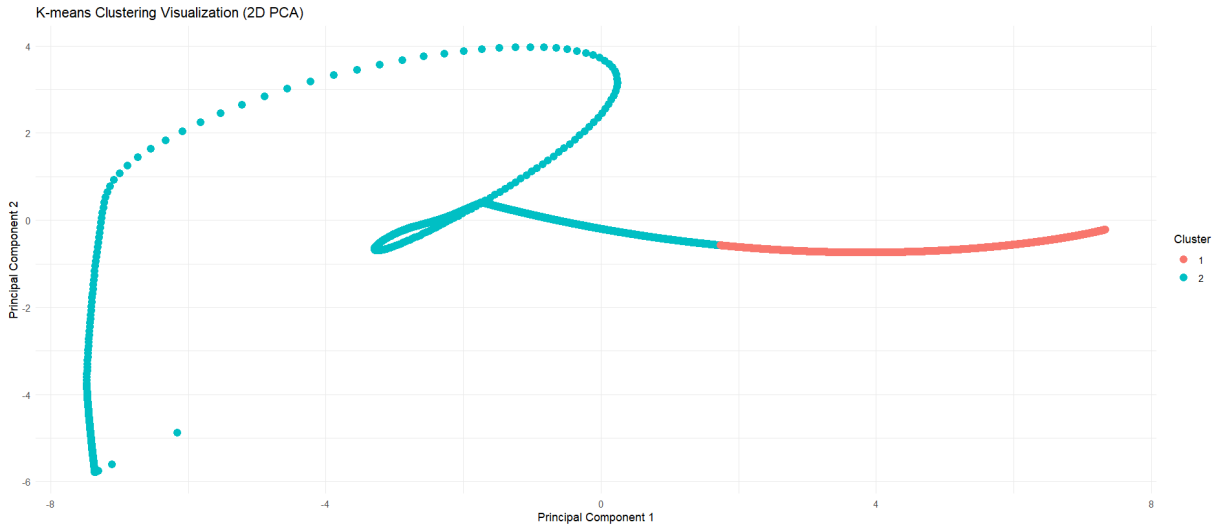

Figure 10: K-Means Clustering Results after Principal Component Analysis reduced visualisation (x = PC1, y = PC2); This scatter plot visualizes the segmentation of the tumour microenvironment into two distinct clusters using PCA for dimensionality reduction. The tightly grouped blue cluster may represent a homogenous state of cellular components, whereas the dispersed red cluster may suggest a varied state, possibly indicating phenotypic heterogeneity or a transitional phase in tumour evolution. Each point symbolizes the integrated profile of cancerous and immune cells at a given time point, highlighting the complex interactions within the tumour ecology.

Supervised Learning using Logistic Regression

The logistic regression analysis of the provided dataset revealed compelling insights. The model, trained on a comprehensive array of features encompassing cancer cells, stem cells, macrophages, T helper cells, and cytokines, exhibited statistical significance, as demonstrated by a substantial F-statistic (1.083e+04) with an exceptionally low p-value (< 2.2e-16). Coefficients for specific variables, such as "Cancer Cells" and "Resistant Stem Cells," were significantly different from zero ($p < 0.05$), indicating their importance in predicting the target variable "Cluster", whereas "Cancer Stem Cells" and "Resistant Cancer Cells" have high negative coefficients, suggesting a negative relationship with the variable "Cluster". The intercept was highly significant ($p < 2e-16$), reflecting the baseline level when all predictors are zero, implying their substantial influence on the target variable, "Cluster."

The model achieved an impressive multiple R-squared value of 0.9618, indicating that approximately 96.18% of the variance in the training data is explained. The Adjusted R-squared is also 0.9618, indicating that the model's goodness of fit is reliable even with multiple predictors. However, rank-deficient fit was observed during predictions, suggesting potential multicollinearity issues.

The Root Mean Squared Error (RMSE) on the testing set was 0.0795, signifying an average deviation of this magnitude between predicted and actual values. Validation on a separate dataset yielded a high R-squared value of 0.9706, reinforcing the model's predictive capability. The ROC curve highlighted the model's excellent discrimination between cancer and immune cell classes, with an AUC 1 (Figure 11).

Visualization tools, including residual and QQ plots, offered a nuanced understanding of the model's adequacy. The residuals appear centered around zero, indicating that the model captures the overall trend in the data (Figure 12). However, the result is sparsely spread, indicating that the model might only capture some variability. The histogram shows a roughly normal distribution of residuals (Figure 12), and the QQ plot indicates that the residuals are approximately normally distributed (Figure 14). Also, the confusion matrix indicates perfect performance (all diagonal elements are non-zero), which aligns with the high accuracy (Figure 15).

Interpretation of Logistic Regression results with respect to tumor-macrophage interaction

The coefficients calculated by the logistic regression model shed light on the relationship between cancer cells, macrophages, IL-10, and the assigned cluster. The coefficients for "Cancer Stem Cells" and "Resistant Stem Cells" emerge as highly significant (Figure 16), implying a significant influence on the assigned cluster. The negative coefficient for both the variables indicates an inverse relationship, hinting at the potential role of two in influencing a lower assigned cluster.

Surprisingly, the coefficients for M1 (pro-inflammatory) and M2 (anti-inflammatory) tumor-associated macrophages lack statistical significance. This finding is intriguing, suggesting that, in the context of this study, the specific subtypes of macrophages may not be primary determinants of the assigned cluster. This nuance underscores the complexity of the role played by macrophages in cancer and prompts further investigation.

Also, the coefficient for Cytokine IL10 is not statistically significant, implying that the abundance of IL-10 may not be a decisive factor in determining the assigned cluster. This unexpected result warrants

further exploration, especially given the well-established immunosuppressive role of IL-10. Understanding its potential interactions with other variables could unveil intricate relationships.

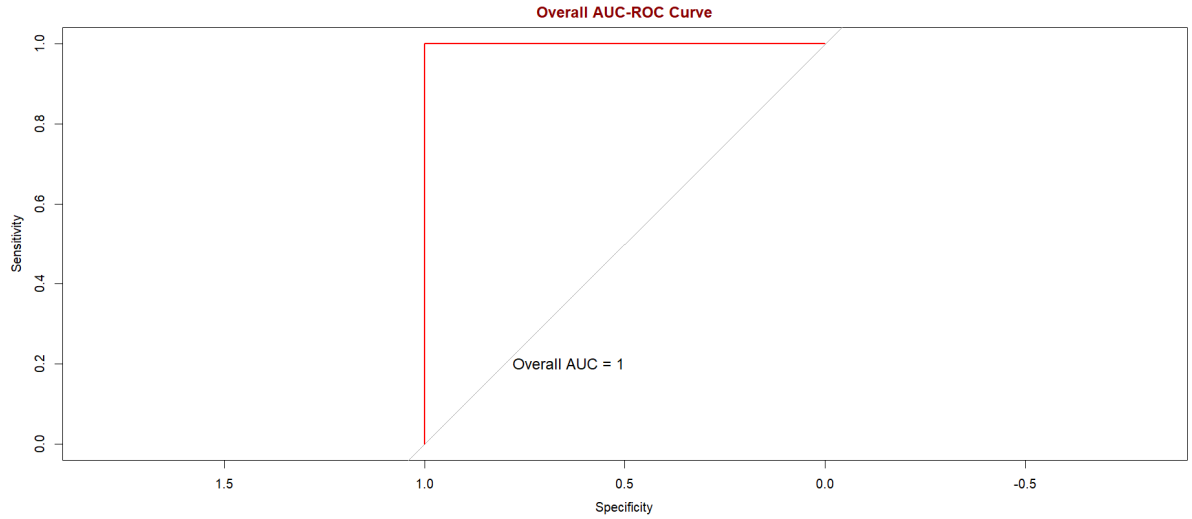

Figure 11. Logistic Regression ROC Curve Demonstrating Optimal Classifier Performance with an AUC of 1.0: This ROC curve illustrates an ideal classifier that perfectly distinguishes between the two classes with 100% sensitivity and 100% specificity, indicating no overlap between the positive and negative class distributions.

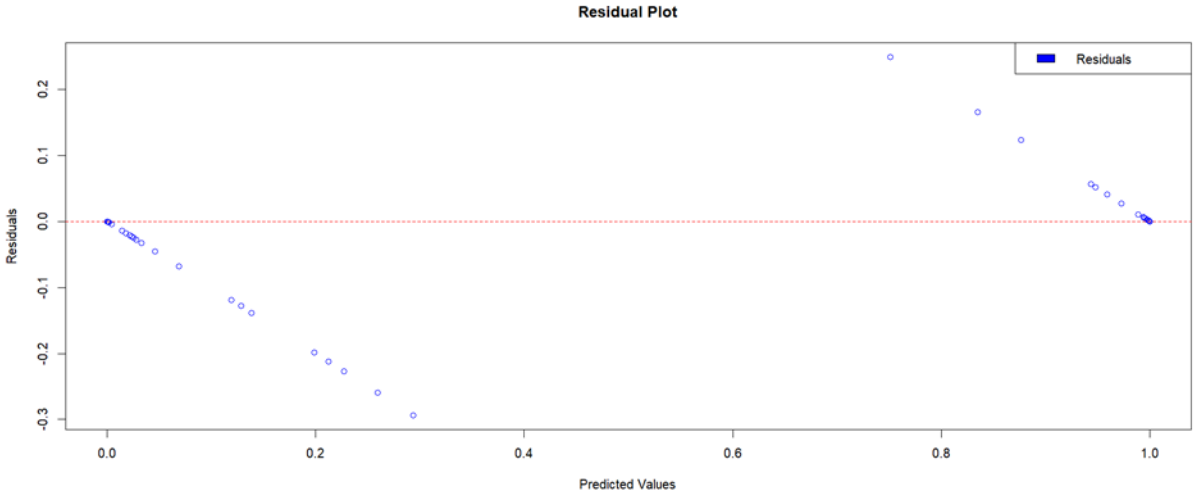

Figure 12: Evaluating Model Accuracy: Residuals in Logistic Regression Analysis; This chart analyses the accuracy of a logistic regression model, mapping predicted values (0 to 1) on the horizontal axis against residuals on the vertical axis. The residuals represent the differences between observed and predicted values, offering insight into the model's precision. Most residuals hover near the zero line, indicating a strong fit at lower predicted values. The red dotted line at zero is a benchmark for assessing fit deviations.

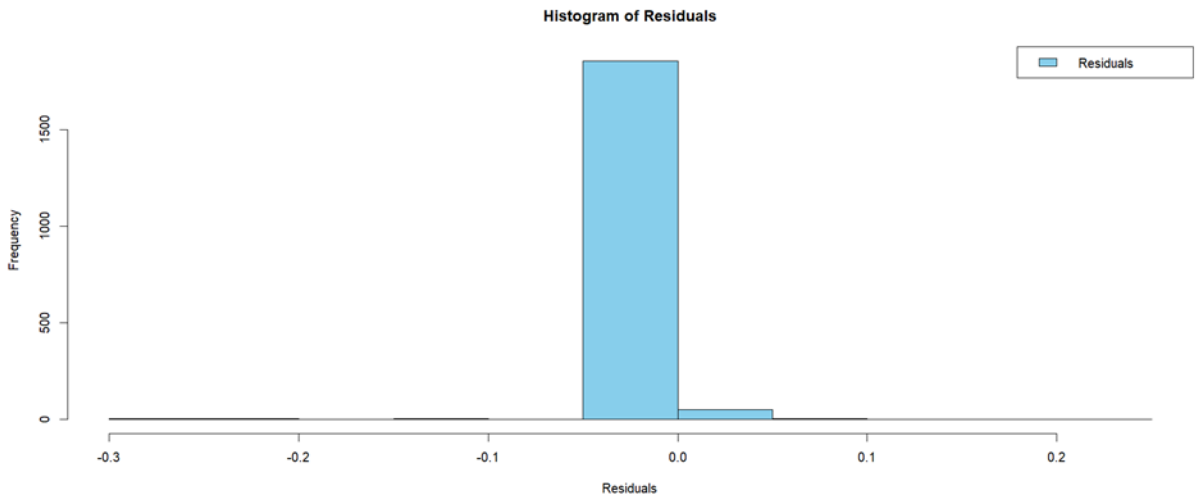

Figure 13: Assessing Model Precision: Histogram of Residuals from Logistic Regression; This histogram visually represents the residuals from a logistic regression model, plotted on the x-axis, ranging from approximately -0.3 to 0.2, against their frequency on the y-axis. The prominent central bar, closely hugging the zero mark, indicates that most residuals are minimal, suggesting a high accuracy of the model for the majority of data. The sparse bars at the extremes, particularly beyond -0.1 and 0.1, show that large prediction errors are uncommon, reinforcing the model's reliability.

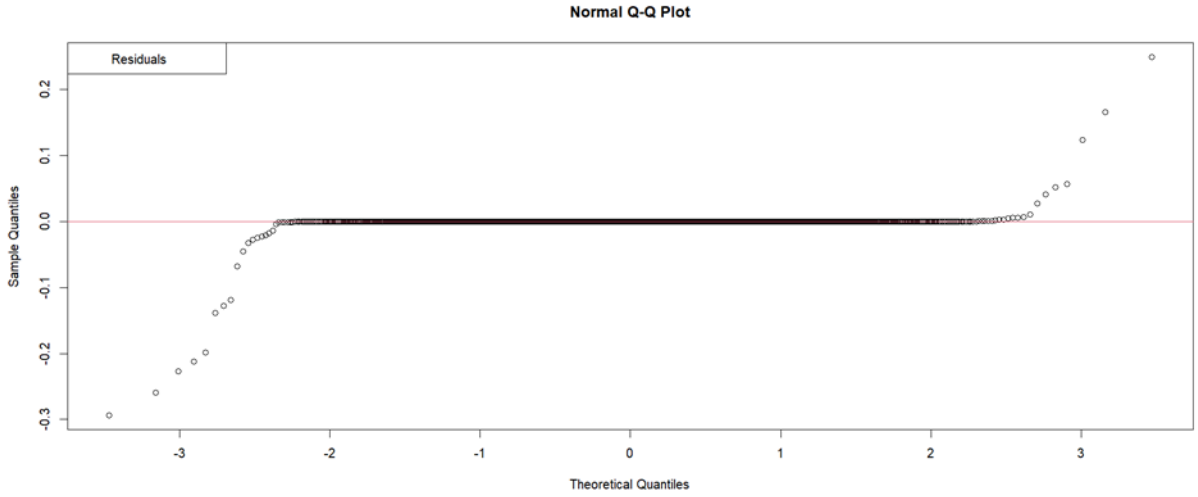

Figure 14: Normality of Residuals: Q-Q Plot Analysis in Logistic Regression; This Q-Q plot compares the theoretical quantiles of a standard normal distribution (horizontal axis) against the sample quantiles of residuals from a logistic regression model (vertical axis). The close alignment of points with the red line in the central part of the plot suggests that residuals largely conform to normality in this range, indicating random and unbiased errors in the middle range of predictions

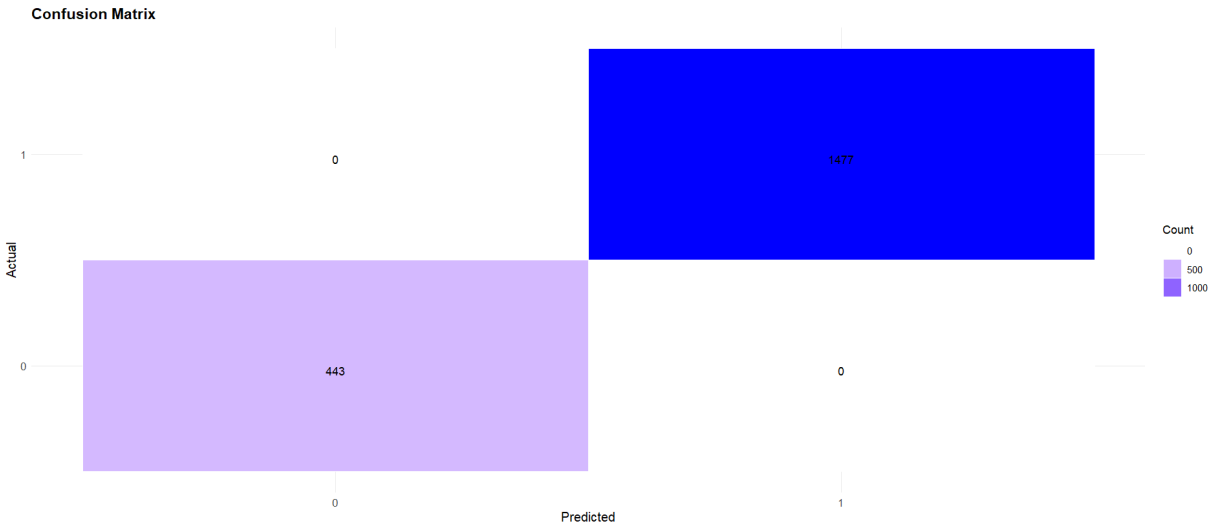

Figure 15. Evaluating Predictive Accuracy: Confusion Matrix in Logistic Regression; This confusion matrix illustrates the performance of a logistic regression model in predicting binary outcomes, such as the presence or absence of specific cell types or states in the tumour microenvironment. The matrix has two rows and columns, corresponding to the actual and predicted classes (0 for 'event did not occur', 1 for 'event occurred'). The intense dark blue in the true positive quadrant (top right) and the significant count in the true negative quadrant (bottom left), with the absence of false negative (top left) and false positive (bottom right), suggest a highly accurate model.

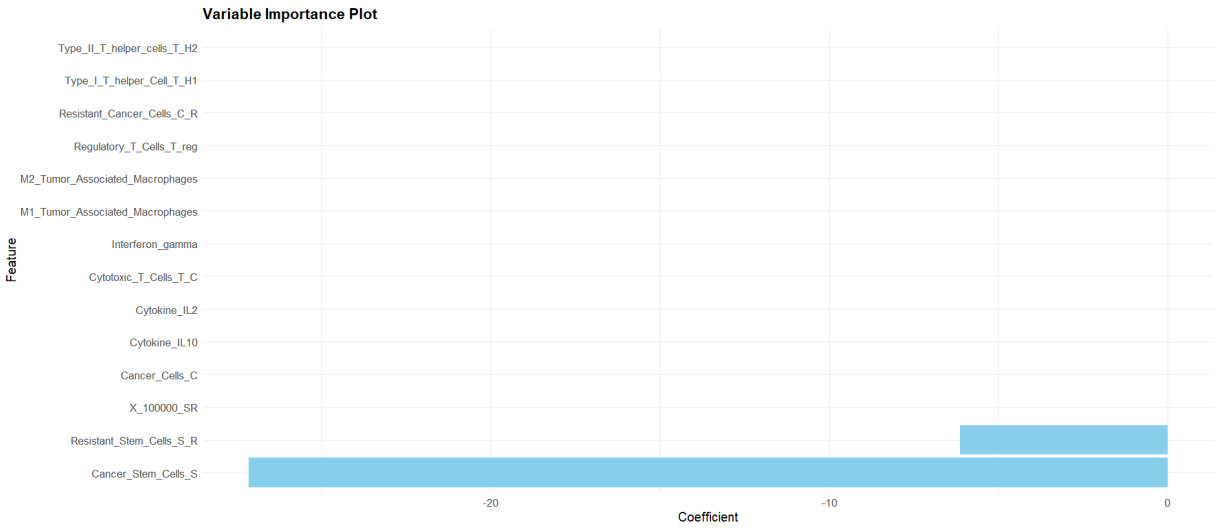

Figure 16. Interpreting Variable Impact: Coefficient Plot in Logistic Regression; This coefficient plot from a logistic regression model visualizes the influence of each predictor variable on the log-odds of the dependent variable. On the x-axis, the coefficients represent the effect of a one-unit change in each variable, keeping others constant. The y-axis enumerates predictor variables. Most variables show small coefficients near zero, implying limited individual impact. However, "Resistant_Stem_Cells_S_R" and "Cancer_Stem_Cells_S" stand out with longer leftward bars, indicating these are significant negative predictors; their increase correlates with decreased log-odds of the predicted outcome. This could be key in understanding tumour dynamics or treatment effectiveness.

Supervised Learning using Support Vector Machines (SVM)

The SVM analysis on the cancer data yielded compelling results, with hyperparameter tuning identifying optimal values of a cost of 0.01 and a gamma of 0.03125. The model achieved perfect classification during cross-validation on the training data, reflected in a confusion matrix with

accuracy, sensitivity, and specificity equal to 1 on the testing data. The ROC curve highlighted the model's excellent discrimination between cancer classes, with an AUC 1. While these results showcase the model's high predictive performance, SVM, as a black-box model, lacks biological interpretability (Camacho et al., 2018).

The visualization techniques in analyzing the cancer data provide valuable insights into the SVM model's performance. The ROC curve, a critical diagnostic tool, demonstrates the model's excellent discrimination with a curve close to the top-left corner and an AUC of 1 (Figure 17). Also, the confusion matrix indicates perfect performance (all diagonal elements are non-zero), which aligns with the high accuracy (Figure 18).

Interpretation of SVM results with respect to tumor-macrophage interaction

Due to its black box model, the results from the SVM model could not provide any significant data of biological interpretability. While the model can make accurate predictions, it's difficult to understand or interpret exactly how and why it's making these decisions based on the input data. This is a common challenge in many complex machine learning models. However, the accurate and perfect prediction results from SVM and logistic regression provide us insights on how simulation data might behave when implemented in machine learning algorithms. Simulation data is often considered 'idealistic' because it's usually cleaner and more controlled than real-world data. In real-world scenarios, data often contains noise (unwanted variations or irrelevant information), which can complicate analysis and model training. The success of the SVM model in this context might be partly due to the clean, noise-free nature of the simulation data. However, further studies need to be performed to learn more about the nature of simulation data for machine learning analysis.

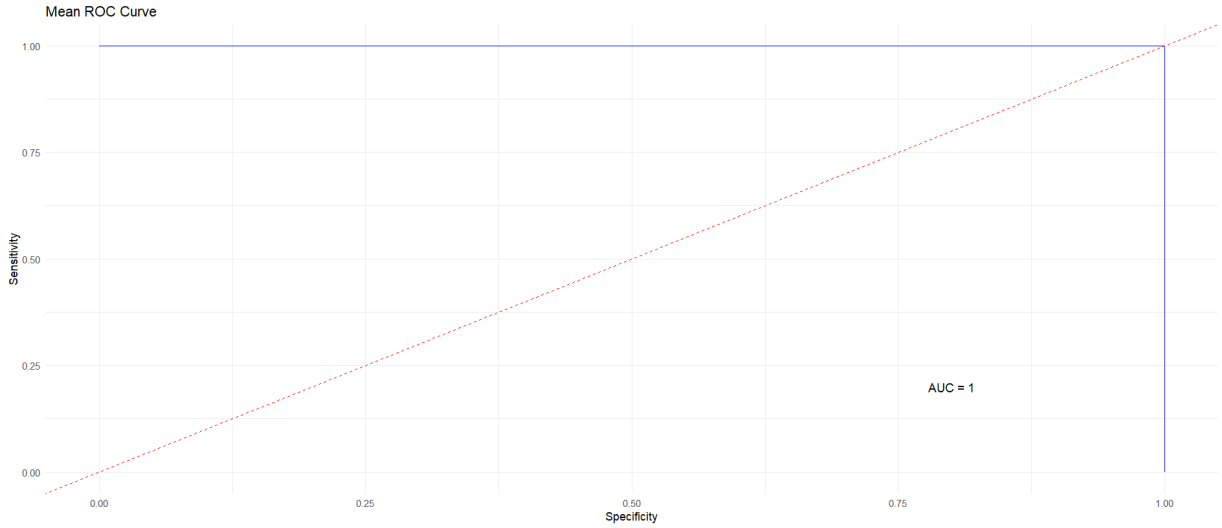

Figure 17: Ideal Classification: ROC Curve Analysis of SVM Classifier; This ROC (Receiver Operating Characteristic) curve illustrates the performance of a Support Vector Machine (SVM) classifier. The x-axis measures Specificity (False Positive Rate, FPR), and the y-axis measures Sensitivity (True Positive Rate, TPR). The curve, hugging the top and left plot borders, indicates an Area Under the Curve (AUC) of 1, symbolizing perfect classifier performance with exceptional sensitivity and specificity. It suggests the SVM classifier accurately identifies all positive and negative cases without any false positives or negatives.

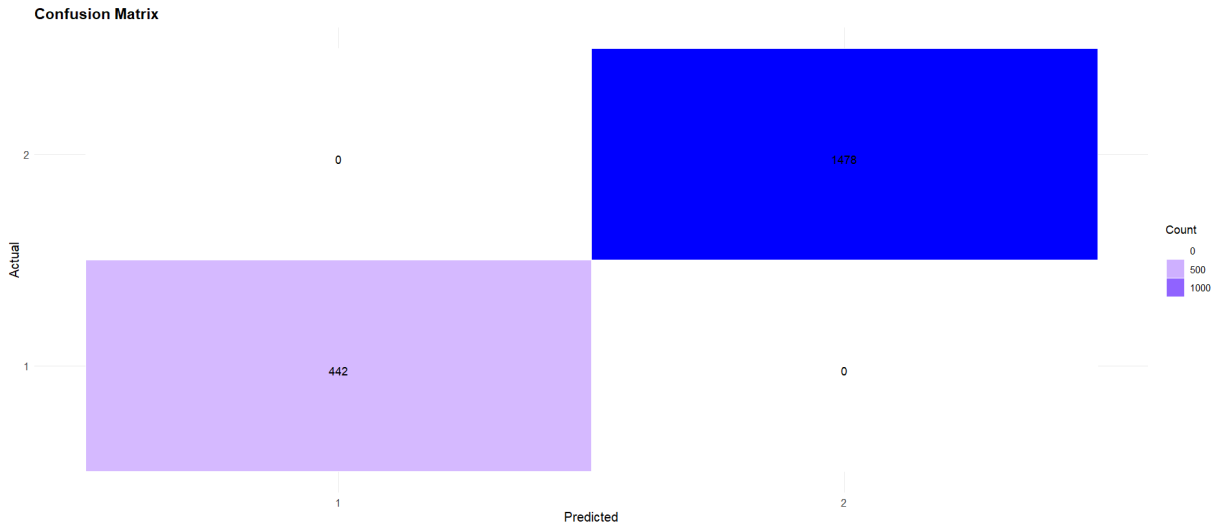

Figure 18: Optimal Prediction: Confusion Matrix of SVM Classifier; This confusion matrix visually represents the impeccable performance of a Support Vector Machine (SVM) classifier, with rows and columns correlating to actual and predicted classes, '1' and '2'. Deep colour shades reflect higher counts, and the matrix displays perfect classification accuracy, with a 100% success rate and no false predictions.

Machine Learning Analysis of a Larger Simulation Dataset

Another analysis was carried out where the time stamps in simulation dataset was increased by 4 times. The simulation period remained the same, i.e. 800 days, however the number of time stamps was increased to 38400, each corresponding to 30 mins within the 800 days period. This was carried out to validate and reciprocate the results obtained previously from the machine learning algorithms and to observe any differences due change in number of time stamps. The same method and protocol were followed as mentioned in the Methods Description and Implementation section, except changing the input file.

However, there were no significant changes observed after increasing the time stamps by 4 times, as seen in the Figures 16 – 32 in Appendix 3. The results obtained after running K-means Clustering, Logistic regression and SVM, were found to be consistent and identical with the results obtained previously (9600-time stamps). As the results were found to be identical for both, 9600-time stamps and 38400-time steps, the results for the latter are covered in Appendix 3. The calibration plot for logistic regression of 38400-time steps (Appendix 3, Figure 30) showed a relatively higher model confidence with its steep, almost vertical calibration curve, unlike the curve for 9600-time stamps (Appendix 2, Figure 19). A possible reason for this outcome can be the use of simulation data instead of real-world data, where the former would follow a stringent set of mathematical and physical principles for consistent outcome each time the model is simulated.

Ethical Issues

Although the project manages to circumnavigate through the ethical issues that arises from the use of medical or clinical data, it is to be noted that this study did not collect real-world data directly; instead, uses a computational model to generate the data. Mathematical models often make assumptions or simplifications to make the simulations more manageable. This can lead to data that may not fully capture the complexities and variations found in the real world. One notable characteristic of simulation data is the absence of noise. In real-world data, noise refers to random variations or errors that can be present in measurements. Simulation data, being idealistic, does not

incorporate such noise, making it cleaner and potentially easier for machine learning algorithms to learn patterns. Machine learning models often perform better when trained on clean and well-behaved data, as they can more easily discern patterns and relationships without interference from noise.

Another ethical issue that needs to be addressed is the potential data bias that was generated during K-Means Clustering. The algorithm generated two clusters, one of which was relatively larger than the other cluster, accounting to approximately 76% of the entire data set. Larger clusters may dominate the interpretation, overshadowing meaningful patterns in smaller clusters. This can lead to an incomplete or distorted understanding of the dataset. This may have been occurred due to certain variables being overrepresented and the rest being underrepresented in the dataset. Also, other clustering techniques maybe applied to further understand the dataset.

Scientific Contributions and Future direction

This study showcases the power of a data-driven approach in cancer research, where mathematical models are theoretical and validated and refined using machine learning techniques. This approach aligns with the current trend in biomedical research, emphasizing the importance of leveraging large datasets and computational methods for insights that can guide clinical decision-making. By examining cancer cells, stem cells, macrophages, T helper cells, and cytokines, the study provides a comprehensive overview of the intricate interplay within the tumor microenvironment, shedding light on potential factors influencing cancer progression.

Using unsupervised learning (K-Means Clustering) to identify distinct biological profiles is a significant step in understanding the heterogeneity within cancer populations. Identifying clusters with varying immune responses and cancer stem cell populations can provide nuanced insights into cancer's different stages and characteristics, offering potential avenues for targeted therapies and treatments. Using machine learning to validate mathematical models enhances the credibility of simulation results, and this cross-validation ensures that the mathematical models, often based on simplifications and assumptions, align with real-world data. The convergence of results from two different methodologies strengthens the reliability of the findings. The findings from machine learning can also be used to improve the pre-existing mathematical models by putting weightage on variables or simulation phases that are important through machine learning analysis.

This study is a step towards bridging the gap between simulations and real-world scenarios for future research to focus on validating mathematical models and machine learning predictions using clinical data. Future studies can be directed towards developing more sophisticated hybrid approaches between mathematical modelling and machine learning to improve the understanding of complex biological processes and enable more accurate predictions of outcomes. Another interesting aspect would be exploring the potential of combining mathematical models and AI approaches with experimental data to develop more accurate and reliable predictive models for cancer diagnosis, prognosis, and treatment. Likewise, new AI techniques incorporating the fundamental laws of physics and biology can be developed to generate more accurate and biologically plausible predictions.

Conclusion

In conclusion, this thesis has delved into the intricate dynamics of tumor-macrophage interactions, employing a dual approach of mathematical model simulation and various machine learning

techniques. The mathematical model, featuring 14 variables, revealed significant heterogeneity in cancer cell populations, with a noteworthy exponential increase in cancer stem cells suggesting potential implications for proliferation or treatment resistance. Unsupervised learning through K-means clustering identified distinct biological profiles, emphasizing an active immune response and higher cancer stem cell populations in Cluster 1. Supervised learning, particularly Logistic Regression, demonstrated a highly significant model with strong predictive capability, shedding light on the influence of specific variables like "Cancer Cells" and "Resistant Stem Cells" on assigned clusters. Support Vector Machines (SVM) showcased compelling results in classification, underscoring associations between different cancer cell types and tumor-associated macrophages.

The data used in this study is obtained from a mathematical model simulation and is bound to be idealistic. Unlike actual data, simulation data does not contain noise. This nature of the simulated data contributed to the machine learning algorithms to perform with high accuracy. One way to improve the machine learning models would be to combine simulated and real-world clinical data, thus providing higher-value results. Simultaneously, the mathematical simulation model used for this study encompassed a wide variety of variables contributing within the tumor microenvironment, instead of focusing solely on the tumor associated macrophages interactions. Contributing to results that do not heavily lean towards understanding impact of macrophages on the tumor cells.

The concept used in this thesis can be used to overcome the scarcity of data, especially in terms of deep learning analysis, which is a data-hungry model for precise analysis of biological functions such as cancer. The findings underscore the complexity of tumor-macrophage interactions, calling for further research to refine clustering methods and explore alternative machine-learning techniques. Furthermore, the observations from machine learning can help us understand and improve the existing mathematical models by understanding what factors and variables play significant roles in the model interaction. Integrating mathematical modelling and machine learning opens avenues for developing more sophisticated models capturing the nuances of these interactions, contributing to advancements in cancer biology and therapeutic strategies.

References

Abaszade, M., & Effati, S. (2018). Stochastic Support Vector Machine for Classifying and Regression of Random Variables. *Neural Processing Letters*, *48*(1), 1–29. https://doi.org/10.1007/s11063-017-9697-0

Ahmed, N., Escalona, R., Leung, D., Chan, E., & Kannourakis, G. (2018). Tumour microenvironment and metabolic plasticity in cancer and cancer stem cells: Perspectives on metabolic and immune regulatory signatures in chemoresistant ovarian cancer stem cells. *Seminars in Cancer Biology*, *53*, 265–281. https://doi.org/10.1016/j.semcancer.2018.10.002

Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W. W., Perdikaris, P., Petzold, L., & Kuhl, E. (2019). Integrating machine learning and multiscale modeling—Perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *Npj Digital Medicine*, *2*(1), Article 1. https://doi.org/10.1038/s41746-019-0193-y

Alden, K., Cosgrove, J., Coles, M., & Timmis, J. (2020). Using Emulation to Engineer and Understand Simulations of Biological Systems. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(1), 302–315. https://doi.org/10.1109/TCBB.2018.2843339

Atashzar, M. R., Baharlou, R., Karami, J., Abdollahi, H., Rezaei, R., Pourramezan, F., & Zoljalali Moghaddam, S. H. (2020). Cancer stem cells: A review from origin to therapeutic implications. *Journal of Cellular Physiology*, *235*(2), 790–803. https://doi.org/10.1002/jcp.29044

Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., Choi, K., Fromme, R. M., Dao, P., McKenney, P. T., Wasti, R. C., Kadaveru, K., Mazutis, L., Rudensky, A. Y., & Pe'er, D. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, *174*(5), 1293-1308.e36. https://doi.org/10.1016/j.cell.2018.05.060

Baker, R. E., Peña, J.-M., Jayamohan, J., & Jérusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, *14*(5), 20170660. https://doi.org/10.1098/rsbl.2017.0660

Bashir, S., Sharma, Y., Elahi, A., & Khan, F. (2016). Macrophage polarization: The link between inflammation and related diseases. *Inflammation Research: Official Journal of the European Histamine Research Society ... [et Al.]*, *65*(1), 1–11. https://doi.org/10.1007/s00011-015-0874-1

Benoit, M., Desnues, B., & Mege, J.-L. (2008). Macrophage polarization in bacterial infections. *Journal of Immunology (Baltimore, Md.: 1950)*, *181*(6), 3733–3739. https://doi.org/10.4049/jimmunol.181.6.3733

Benzekry, S. (2020). Artificial Intelligence and Mechanistic Modeling for Clinical Decision Making in Oncology. *Clinical Pharmacology and Therapeutics*, *108*(3), 471–486. https://doi.org/10.1002/cpt.1951

Beschin, A., De Baetselier, P., & Van Ginderachter, J. A. (2013). Contribution of myeloid cell subsets to liver fibrosis in parasite infection. *The Journal of Pathology*, *229*(2), 186–197. https://doi.org/10.1002/path.4112

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology*, *188*(12), 2222–2239. https://doi.org/10.1093/aje/kwz189

Binder, J., Koller, D., Russell, S., & Kanazawa, K. (1997). Adaptive Probabilistic Networks with Hidden Variables. *Machine Learning*, *29*(2), 213–244. https://doi.org/10.1023/A:1007421730016

Bonnardel, J., & Guilliams, M. (2018). Developmental control of macrophage function. *Current Opinion in Immunology*, *50*, 64–74. https://doi.org/10.1016/j.coi.2017.12.001

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., & Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, *8*, 3395. https://doi.org/10.1038/s41598-018-21758-3

Cachot, A., Bilous, M., Liu, Y.-C., Li, X., Saillard, M., Cenerenti, M., Rockinger, G. A., Wyss, T., Guillaume, P., Schmidt, J., Genolet, R., Ercolano, G., Protti, M. P., Reith, W., Ioannidou, K., De Leval, L., Trapani, J. A., Coukos, G., Harari, A., … Jandus, C. (2021). Tumor-specific cytolytic CD4 T cells mediate immunity against human cancer. *Science Advances*, *7*(9), eabe3348. https://doi.org/10.1126/sciadv.abe3348

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. *Cell*, *173*(7), 1581–1592. https://doi.org/10.1016/j.cell.2018.05.015

Chen, D., Zhang, X., Li, Z., & Zhu, B. (2021). Metabolic regulatory crosstalk between tumor microenvironment and tumor-associated macrophages. *Theranostics*, *11*(3), 1016–1030. https://doi.org/10.7150/thno.51777

Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, *14*(4), e1006076. https://doi.org/10.1371/journal.pcbi.1006076

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., Girard, N., Elemento, O., Nicholson, A. G., Blay, J.-Y., Galateau-Sallé, F., Wainrib, G., & Clozel, T. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, *25*(10), 1519–1525. https://doi.org/10.1038/s41591-019-0583-3

Dall'Asta, M., Derlindati, E., Ardigò, D., Zavaroni, I., Brighenti, F., & Del Rio, D. (2012). Macrophage polarization: The answer to the diet/inflammation conundrum? *Nutrition, Metabolism, and Cardiovascular Diseases: NMCD*, *22*(5), 387–392. https://doi.org/10.1016/j.numecd.2011.12.010

Das, A., Sinha, M., Datta, S., Abas, M., Chaffee, S., Sen, C. K., & Roy, S. (2015). Monocyte and macrophage plasticity in tissue repair and regeneration. *The American Journal of Pathology*, *185*(10), 2596–2606. https://doi.org/10.1016/j.ajpath.2015.06.001

Davies, L. C., Jenkins, S. J., Allen, J. E., & Taylor, P. R. (2013). Tissue-resident macrophages. *Nature Immunology*, *14*(10), 986–995. https://doi.org/10.1038/ni.2705

de Pillis, L. G., Radunskaya, A. E., & Wiseman, C. L. (2005). A validated mathematical model of cell-mediated immune response to tumor growth. *Cancer Research*, *65*(17), 7950–7958. https://doi.org/10.1158/0008-5472.CAN-05-0564

Deist, T. M., Patti, A., Wang, Z., Krane, D., Sorenson, T., & Craft, D. (2019). Simulation-assisted machine learning. *Bioinformatics*, *35*(20), 4072–4080. https://doi.org/10.1093/bioinformatics/btz199

den Breems, N. Y., & Eftimie, R. (2016). The re-polarisation of M2 and M1 macrophages and its role on cancer outcomes. *Journal of Theoretical Biology*, *390*, 23–39. https://doi.org/10.1016/j.jtbi.2015.10.034

DeNardo, D. G., & Ruffell, B. (2019). Macrophages as regulators of tumor immunity and immunotherapy. *Nature Reviews. Immunology*, *19*(6), 369–382. https://doi.org/10.1038/s41577-019-0127-6

Eftimie, R., Bramson, J. L., & Earn, D. J. D. (2011). Interactions between the immune system and cancer: A brief review of non-spatial mathematical models. *Bulletin of Mathematical Biology*, *73*(1), 2–32. https://doi.org/10.1007/s11538-010-9526-3

Eftimie, R., & Eftimie, G. (2019). Investigating Macrophages Plasticity Following Tumour-Immune Interactions During Oncolytic Therapies. *Acta Biotheoretica*, *67*(4), 321–359. https://doi.org/10.1007/s10441-019-09357-9

Eguchi, K., & Manabe, I. (2013). Macrophages and islet inflammation in type 2 diabetes. *Diabetes, Obesity & Metabolism*, *15 Suppl 3*, 152–158. https://doi.org/10.1111/dom.12168

Epelman, S., Lavine, K. J., & Randolph, G. J. (2014). Origin and Functions of Tissue Macrophages. *Immunity*, *41*(1), 21–35. https://doi.org/10.1016/j.immuni.2014.06.013

Farhood, B., Najafi, M., & Mortezaee, K. (2019). CD8 $^{+}$ cytotoxic T lymphocytes in cancer immunotherapy: A review. *Journal of Cellular Physiology*, *234*(6), 8509–8521. https://doi.org/10.1002/jcp.27782

Fleiss, J. L., Williams, J. B. W., & Dubro, A. F. (1986). The logistic regression analysis of psychiatric data. *Journal of Psychiatric Research*, *20*(3), 195–209. https://doi.org/10.1016/0022-3956(86)90003-8

Frieboes, H. B., Zheng, X., Sun, C.-H., Tromberg, B., Gatenby, R., & Cristini, V. (2006). An integrated computational/experimental model of tumor invasion. *Cancer Research*, *66*(3), 1597–1604. https://doi.org/10.1158/0008-5472.CAN-05-3166

Funahashi, A., Matsuoka, Y., Jouraku, A., Kitano, H., & Kikuchi, N. (2006). CellDesigner: A modeling tool for biochemical networks. *Proceedings of the 38th Conference on Winter Simulation*, 1707–1712.

Funahashi, A., Morohashi, M., Matsuoka, Y., Jouraku, A., & Kitano, H. (2007). CellDesigner: A Graphical Biological Network Editor and Workbench Interfacing Simulator. In S. Choi (Ed.), *Introduction to Systems Biology* (pp. 422–434). Humana Press. https://doi.org/10.1007/978-1-59745-531-2_21

Ganguli, P., & Sarkar, R. R. (2018). Exploring immuno-regulatory mechanisms in the tumor microenvironment: Model and design of protocols for cancer remission. *PLoS ONE*, *13*(9), e0203030. https://doi.org/10.1371/journal.pone.0203030

Gaw, N., Hawkins-Daarud, A., Hu, L. S., Yoon, H., Wang, L., Xu, Y., Jackson, P. R., Singleton, K. W., Baxter, L. C., Eschbacher, J., Gonzales, A., Nespodzany, A., Smith, K., Nakaji, P., Mitchell, J. R., Wu, T., Swanson, K. R., & Li, J. (2019). Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric MRI. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-46296-4

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., Nair, V. S., Xu, Y., Khuong, A., Hoang, C. D., Diehn, M., West, R. B., Plevritis, S. K., & Alizadeh, A. A. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature Medicine*, *21*(8), 938–945. https://doi.org/10.1038/nm.3909

Ginhoux, F., & Guilliams, M. (2016). Tissue-Resident Macrophage Ontogeny and Homeostasis. *Immunity*, *44*(3), 439–449. https://doi.org/10.1016/j.immuni.2016.02.024

Ginhoux, F., & Jung, S. (2014). Monocytes and macrophages: Developmental pathways and tissue homeostasis. *Nature Reviews. Immunology*, *14*(6), 392–404. https://doi.org/10.1038/nri3671

Glont, M., Nguyen, T. V. N., Graesslin, M., Hälke, R., Ali, R., Schramm, J., Wimalaratne, S. M., Kothamachu, V. B., Rodriguez, N., Swat, M. J., Eils, J., Eils, R., Laibe, C., Malik-Sheriff, R. S., Chelliah, V., Le Novère, N., & Hermjakob, H. (2018). BioModels: Expanding horizons to include more modelling approaches and formats. *Nucleic Acids Research*, *46*(D1), D1248–D1253. https://doi.org/10.1093/nar/gkx1023

Hatzikirou, H. (2018). Statistical mechanics of cell decision-making: The cell migration force distribution. *Journal of the Mechanical Behavior of Materials*, *27*(1–2). https://doi.org/10.1515/jmbm-2018-0001

Hu, F., Zhou, Y., Wang, Q., Yang, Z., Shi, Y., & Chi, Q. (2020). Gene Expression Classification of Lung Adenocarcinoma into Molecular Subtypes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(4), 1187–1197. https://doi.org/10.1109/TCBB.2019.2905553

Hu, L. S., Ning, S., Eschbacher, J. M., Baxter, L. C., Gaw, N., Ranjbar, S., Plasencia, J., Dueck, A. C., Peng, S., Smith, K. A., Nakaji, P., Karis, J. P., Quarles, C. C., Wu, T., Loftus, J. C., Jenkins, R. B., Sicotte, H., Kollmeyer, T. M., O'Neill, B. P., … Mitchell, J. R. (2017). Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro-Oncology*, *19*(1), 128–137. https://doi.org/10.1093/neuonc/now135

Hu, L. S., Ning, S., Eschbacher, J. M., Gaw, N., Dueck, A. C., Smith, K. A., Nakaji, P., Plasencia, J., Ranjbar, S., Price, S. J., Tran, N., Loftus, J., Jenkins, R., O'Neill, B. P., Elmquist, W., Baxter, L. C., Gao, F., Frakes, D., Karis, J. P., … Li, J. (2015). Multi-Parametric MRI and Texture Analysis to Visualize Spatial Histologic Heterogeneity and Tumor Extent in Glioblastoma. *PloS One*, *10*(11), e0141506. https://doi.org/10.1371/journal.pone.0141506

Huang, Z., Johnson, T. S., Han, Z., Helm, B., Cao, S., Zhang, C., Salama, P., Rizkalla, M., Yu, C. Y., Cheng, J., Xiang, S., Zhan, X., Zhang, J., & Huang, K. (2020). Deep learning-based cancer survival prognosis from RNA-seq data: Approaches and evaluations. *BMC Medical Genomics*, *13*(Suppl 5), 41. https://doi.org/10.1186/s12920-020-0686-1

Jansen, J. E., Gaffney, E. A., Wagg, J., & Coles, M. C. (2019). Combining Mathematical Models With Experimentation to Drive Novel Mechanistic Insights Into Macrophage Function. *Frontiers in Immunology*, *10*, 1283. https://doi.org/10.3389/fimmu.2019.01283

Jeong, S., Jang, N., Kim, M., & Choi, I.-K. (2023). CD4 $^{+}$ cytotoxic T cells: An emerging effector arm of anti-tumor immunity. *BMB Reports*, *56*(3), 140–144. https://doi.org/10.5483/BMBRep.2023-0014

Jing, B., Zhang, T., Wang, Z., Jin, Y., Liu, K., Qiu, W., Ke, L., Sun, Y., He, C., Hou, D., Tang, L., Lv, X., & Li, C. (2019). A deep survival analysis method based on ranking. *Artificial Intelligence in Medicine*, *98*, 1–9. https://doi.org/10.1016/j.artmed.2019.06.001

Kadomoto, S., Izumi, K., & Mizokami, A. (2021). Macrophage Polarity and Disease Control. *International Journal of Molecular Sciences*, *23*(1), 144. https://doi.org/10.3390/ijms23010144

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*, 24. https://doi.org/10.1186/s12874-018-0482-1

Kerneur, C., Cano, C. E., & Olive, D. (2022). Major pathways involved in macrophage polarization in cancer. *Frontiers in Immunology*, *13*, 1026954. https://doi.org/10.3389/fimmu.2022.1026954

Komohara, Y., Jinushi, M., & Takeya, M. (2014). Clinical significance of macrophage heterogeneity in human malignant tumors. *Cancer Science*, *105*(1), 1–8. https://doi.org/10.1111/cas.12314

Korfiatis, P., Lachance, D., Parney, I., Buckner, J., Eckel-Passow, J., Decker, P., Jenkins, R., Wrensch, M., Wiencke, J., Hansen, H., Rice, T., McCoy, L., Nelson, S., Clarke, J., Taylor, J., Luks, T., & Erickson, B. (2018). COMP-05. EVALUATION OF A DEEP LEARNING ARCHITECTURE FOR MRI PREDICTION OF IDH, 1p19q AND TERT IN GLIOMA PATIENTS. *Neuro-Oncology*, *20*(Suppl 6), vi64. https://doi.org/10.1093/neuonc/noy148.260

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. https://doi.org/10.1016/j.csbj.2014.11.005

Lai, Y.-H., Chen, W.-N., Hsu, T.-C., Lin, C., Tsao, Y., & Wu, S. (2020). Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific Reports*, *10*, 4679. https://doi.org/10.1038/s41598-020-61588-w

Lantz, C., Radmanesh, B., Liu, E., Thorp, E. B., & Lin, J. (2020). Single-cell RNA sequencing uncovers heterogenous transcriptional signatures in macrophages during efferocytosis. *Scientific Reports*, *10*(1), 14333. https://doi.org/10.1038/s41598-020-70353-y

Lavin, Y., Mortha, A., Rahman, A., & Merad, M. (2015). Regulation of macrophage development and function in peripheral tissues. *Nature Reviews. Immunology*, *15*(12), 731–744. https://doi.org/10.1038/nri3920

Lee, H.-W., Choi, H.-J., Ha, S.-J., Lee, K.-T., & Kwon, Y.-G. (2013). Recruitment of monocytes/macrophages in different tumor microenvironments. *Biochimica Et Biophysica Acta*, *1835*(2), 170–179. https://doi.org/10.1016/j.bbcan.2012.12.007

Leopold Wager, C. M., Hole, C. R., Wozniak, K. L., Olszewski, M. A., Mueller, M., & Wormley, F. L. (2015). STAT1 signaling within macrophages is required for antifungal activity against

Cryptococcus neoformans. *Infection and Immunity*, *83*(12), 4513–4527. https://doi.org/10.1128/IAI.00935-15

Li, X., Jolly, M. K., George, J. T., Pienta, K. J., & Levine, H. (2019). Computational Modeling of the Crosstalk Between Macrophage Polarization and Tumor Cell Plasticity in the Tumor Microenvironment. *Frontiers in Oncology*, *9*, 10. https://doi.org/10.3389/fonc.2019.00010

Li, Y., Jiang, W., Yang, L., & Wu, T. (2018). On neural networks and learning systems for business computing. *Neurocomputing*, *275*, 1150–1159. https://doi.org/10.1016/j.neucom.2017.09.054

Li, Z., Wang, Y., Yu, J., Guo, Y., & Cao, W. (2017). Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Scientific Reports*, *7*(1), 5467. https://doi.org/10.1038/s41598-017-05848-2

Lytton, W. W., Arle, J., Bobashev, G., Ji, S., Klassen, T. L., Marmarelis, V. Z., Schwaber, J., Sherif, M. A., & Sanger, T. D. (2017). Multiscale modeling in the clinic: Diseases of the brain and nervous system. *Brain Informatics*, *4*(4), 219–230. https://doi.org/10.1007/s40708-017-0067-5

Macklin, P. (2017). When Seeing Isn't Believing: How Math Can Guide Our Interpretation of Measurements and Experiments. *Cell Systems*, *5*(2), 92–94. https://doi.org/10.1016/j.cels.2017.08.005

Maechler, M., Rousseeuw, P., Struyf, A., & Hubert, M. (2015, July 21). *"Finding Groups in Data": Cluster Analysis Extended Rousseeuw etal.* https://www.semanticscholar.org/paper/%22Finding-Groups-in-Data%22%3A-Cluster-Analysis-Extended-Maechler-Rousseeuw/17a188e028a95d003144fcfd3a0839fe622deb79

Magazzù, G., Zampieri, G., & Angione, C. (2022). Clinical stratification improves the diagnostic accuracy of small omics datasets within machine learning and genome-scale metabolic modelling methods. *Computers in Biology and Medicine*, *151*, 106244. https://doi.org/10.1016/j.compbiomed.2022.106244

Mahlbacher, G. E., Reihmer, K. C., & Frieboes, H. B. (2019). Mathematical modeling of tumor-immune cell interactions. *Journal of Theoretical Biology*, *469*, 47–60. https://doi.org/10.1016/j.jtbi.2019.03.002

Makaryan, S. Z., Cess, C. G., & Finley, S. D. (2020). Modeling immune cell behavior across scales in cancer. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, *12*(4), e1484. https://doi.org/10.1002/wsbm.1484

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., Vu, M. T., Men, J., Maire, M., Kananathan, S., Fairbanks, E. L., Meyer, J. P., Arankalle, C., Varusai, T. M., Knight-Schrijver, V., Li, L., Dueñas-Roca, C., Dass, G., Keating, S. M., … Hermjakob, H. (2020). BioModels—15 years of sharing computational models in life science. *Nucleic Acids Research*, *48*(D1), D407–D415. https://doi.org/10.1093/nar/gkz1055

Mantovani, A., Biswas, S. K., Galdiero, M. R., Sica, A., & Locati, M. (2013). Macrophage plasticity and polarization in tissue repair and remodelling. *The Journal of Pathology*, *229*(2), 176–185. https://doi.org/10.1002/path.4133

Mascheroni, P., Savvopoulos, S., Alfonso, J. C. L., Meyer-Hermann, M., & Hatzikirou, H. (2021a). Improving personalized tumor growth predictions using a Bayesian combination of mechanistic modeling and machine learning. *Communications Medicine*, *1*(1), Article 1. https://doi.org/10.1038/s43856-021-00020-4

Mascheroni, P., Savvopoulos, S., Alfonso, J. C. L., Meyer-Hermann, M., & Hatzikirou, H. (2021b). Improving personalized tumor growth predictions using a Bayesian combination of mechanistic modeling and machine learning. *Communications Medicine*, *1*, 19. https://doi.org/10.1038/s43856-021-00020-4

Mège, J.-L., Mehraj, V., & Capo, C. (2011). Macrophage polarization and bacterial infections. *Current Opinion in Infectious Diseases*, *24*(3), 230–234. https://doi.org/10.1097/QCO.0b013e328344b73e

Mittal, S. K., & Roche, P. A. (2015). Suppression of antigen presentation by IL-10. *Current Opinion in Immunology*, *34*, 22–27. https://doi.org/10.1016/j.coi.2014.12.009

Montfort, A., Pearce, O., Maniati, E., Vincent, B. G., Bixby, L., Böhm, S., Dowe, T., Wilkes, E. H., Chakravarty, P., Thompson, R., Topping, J., Cutillas, P. R., Lockley, M., Serody, J. S., Capasso, M., & Balkwill, F. R. (2017). A Strong B-cell Response Is Part of the Immune Landscape in Human High-Grade Serous Ovarian Metastases. *Clinical Cancer Research*, *23*(1), 250–262. https://doi.org/10.1158/1078-0432.CCR-16-0081

Mousavi, M., Manshadi, M. D., Soltani, M., Kashkooli, F. M., Rahmim, A., Mosavi, A., Kvasnica, M., Atkinson, P. M., Kovács, L., Koltay, A., Kiss, N., & Adeli, H. (2022). Modeling the efficacy of different anti-angiogenic drugs on treatment of solid tumors using 3D computational modeling and machine learning. *Computers in Biology and Medicine*, *146*, 105511. https://doi.org/10.1016/j.compbiomed.2022.105511

Müller, S., Kohanbash, G., Liu, S. J., Alvarado, B., Carrera, D., Bhaduri, A., Watchmaker, P. B., Yagnik, G., Di Lullo, E., Malatesta, M., Amankulor, N. M., Kriegstein, A. R., Lim, D. A., Aghi, M., Okada, H., & Diaz, A. (2017). Single-cell profiling of human gliomas reveals macrophage ontogeny as a basis for regional differences in macrophage activation in the tumor microenvironment. *Genome Biology*, *18*(1), 234. https://doi.org/10.1186/s13059-017-1362-4

Najafi, M., Farhood, B., Mortezaee, K., Kharazinejad, E., Majidpoor, J., & Ahadi, R. (2020). Hypoxia in solid tumors: A key promoter of cancer stem cell (CSC) resistance. *Journal of Cancer Research and Clinical Oncology*, *146*(1), 19–31. https://doi.org/10.1007/s00432-019-03080-1

Ngambenjawong, C., Gustafson, H. H., & Pun, S. H. (2017). Progress in tumor-associated macrophage (TAM)-targeted therapeutics. *Advanced Drug Delivery Reviews*, *114*, 206–221. https://doi.org/10.1016/j.addr.2017.04.010

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, *104*(1), 11–33. https://doi.org/10.1109/JPROC.2015.2483592

Nicolò, C., Périer, C., Prague, M., Bellera, C., MacGrogan, G., Saut, O., & Benzekry, S. (2020). Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer. *JCO Clinical Cancer Informatics*, *4*, 259–274. https://doi.org/10.1200/CCI.19.00133

Nir, G., Karimi, D., Goldenberg, S. L., Fazli, L., Skinnider, B. F., Tavassoli, P., Turbin, D., Villamil, C. F., Wang, G., Thompson, D. J. S., Black, P. C., & Salcudean, S. E. (2019). Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images. *JAMA Network Open*, *2*(3), e190442. https://doi.org/10.1001/jamanetworkopen.2019.0442

Pan, Y., Yu, Y., Wang, X., & Zhang, T. (2020). Tumor-Associated Macrophages in Tumor Immunity. *Frontiers in Immunology*, *11*, 583084. https://doi.org/10.3389/fimmu.2020.583084

Patel, U., Rajasingh, S., Samanta, S., Cao, T., Dawn, B., & Rajasingh, J. (2017). Macrophage polarization in response to epigenetic modifiers during infection and inflammation. *Drug Discovery Today*, *22*(1), 186–193. https://doi.org/10.1016/j.drudis.2016.08.006

Peng, G. C. Y., Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W. W., Perdikaris, P., Petzold, L., & Kuhl, E. (2021). Multiscale Modeling Meets Machine Learning: What Can We Learn? *Archives of Computational Methods in Engineering*, *28*(3), 1017–1037. https://doi.org/10.1007/s11831-020-09405-5

Prasanna, P., Patel, J., Partovi, S., Madabhushi, A., & Tiwari, P. (2017). Radiomic features from the peritumoral brain parenchyma on treatment-naïve multi-parametric MR imaging predict long versus short-term survival in glioblastoma multiforme: Preliminary findings. *European Radiology*, *27*(10), 4188–4197. https://doi.org/10.1007/s00330-016-4637-3

Procopio, A., Cesarelli, G., Donisi, L., Merola, A., Amato, F., & Cosentino, C. (2023). Combined mechanistic modeling and machine-learning approaches in systems biology – A systematic literature review. *Computer Methods and Programs in Biomedicine*, *240*, 107681. https://doi.org/10.1016/j.cmpb.2023.107681

Przedborski, M., Smalley, M., Thiyagarajan, S., Goldman, A., & Kohandel, M. (2021). Systems biology informed neural networks (SBINN) predict response and novel combinations for PD-1 checkpoint blockade. *Communications Biology*, *4*(1), Article 1. https://doi.org/10.1038/s42003-021-02393-7

Qiao, J., Liu, Z., Dong, C., Luan, Y., Zhang, A., Moore, C., Fu, K., Peng, J., Wang, Y., Ren, Z., Han, C., Xu, T., & Fu, Y.-X. (2019). Targeting Tumors with IL-10 Prevents Dendritic Cell-Mediated CD8+ T Cell Apoptosis. *Cancer Cell*, *35*(6), 901-915.e4. https://doi.org/10.1016/j.ccell.2019.05.005

Rakaee, M., Busund, L.-T. R., Jamaly, S., Paulsen, E.-E., Richardsen, E., Andersen, S., Al-Saad, S., Bremnes, R. M., Donnem, T., & Kilvaer, T. K. (2019). Prognostic Value of Macrophage Phenotypes in Resectable Non–Small Cell Lung Cancer Assessed by Multiplex Immunohistochemistry. *Neoplasia*, *21*(3), 282–293. https://doi.org/10.1016/j.neo.2019.01.005

Ryu, H. S., Jin, M.-S., Park, J. H., Lee, S., Cho, J., Oh, S., Kwak, T.-Y., Woo, J. I., Mun, Y., Kim, S. W., Hwang, S., Shin, S.-J., & Chang, H. (2019). Automated Gleason Scoring and Tumor Quantification in Prostate Core Needle Biopsy Images Using Deep Neural Networks and Its Comparison with Pathologist-Based Assessment. *Cancers*, *11*(12), 1860. https://doi.org/10.3390/cancers11121860

Saillard, C., Schmauch, B., Laifa, O., Moarii, M., Toldo, S., Zaslavskiy, M., Pronier, E., Laurent, A., Amaddeo, G., Regnault, H., Sommacale, D., Ziol, M., Pawlotsky, J.-M., Mulé, S., Luciani, A., Wainrib, G., Clozel, T., Courtiol, P., & Calderaro, J. (2020). Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides. *Hepatology (Baltimore, Md.)*, *72*(6), 2000–2013. https://doi.org/10.1002/hep.31207

Salerno, L., Cosentino, C., Merola, A., Bates, D. G., & Amato, F. (2013). Validation of a model of the GAL regulatory system via robustness analysis of its bistability characteristics. *BMC Systems Biology*, *7*(1), 39. https://doi.org/10.1186/1752-0509-7-39

Salerno, L., Cosentino, C., Morrone, G., & Amato, F. (2015). Computational Modeling of a Transcriptional Switch Underlying B-Lymphocyte Lineage Commitment of Hematopoietic Multipotent Cells. *PLOS ONE*, *10*(7), e0132208. https://doi.org/10.1371/journal.pone.0132208

Shapouri-Moghaddam, A., Mohammadian, S., Vazini, H., Taghadosi, M., Esmaeili, S.-A., Mardani, F., Seifi, B., Mohammadi, A., Afshari, J. T., & Sahebkar, A. (2018). Macrophage plasticity, polarization, and function in health and disease. *Journal of Cellular Physiology*, *233*(9), 6425–6440. https://doi.org/10.1002/jcp.26429

Shi, C., & Pamer, E. G. (2011). Monocyte recruitment during infection and inflammation. *Nature Reviews. Immunology*, *11*(11), 762–774. https://doi.org/10.1038/nri3070

Shojaee, P., Mornata, F., Deutsch, A., Locati, M., & Hatzikirou, H. (2022). The impact of tumor associated macrophages on tumor biology under the lens of mathematical modelling: A review. *Frontiers in Immunology*, *13*, 1050067. https://doi.org/10.3389/fimmu.2022.1050067

Sica, A., Erreni, M., Allavena, P., & Porta, C. (2015). Macrophage polarization in pathology. *Cellular and Molecular Life Sciences: CMLS*, *72*(21), 4111–4126. https://doi.org/10.1007/s00018-015-1995-y

Tabibu, S., Vinod, P. K., & Jawahar, C. V. (2019). Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific Reports*, *9*, 10509. https://doi.org/10.1038/s41598-019-46718-3

Tartakovsky, A. M., Marrero, C. O., Perdikaris, P., Tartakovsky, G. D., & Barajas-Solano, D. (2018). *Learning Parameters and Constitutive Relationships with Physics Informed Deep Neural Networks* (arXiv:1808.03398). arXiv. https://doi.org/10.48550/arXiv.1808.03398

Tartakovsky, G., Tartakovsky, A. M., & Perdikaris, P. (2018). *Physics Informed Deep Neural Networks for learning parameters with non-Gaussian non-stationary statistics*. *2018*, H21J-1791.

Torkamani, A., Andersen, K. G., Steinhubl, S. R., & Topol, E. J. (2017). High-Definition Medicine. *Cell*, *170*(5), 828–843. https://doi.org/10.1016/j.cell.2017.08.007

von Rueden, L., Mayer, S., Sifa, R., Bauckhage, C., & Garcke, J. (2020). Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions. In M. R. Berthold, A. Feelders, & G. Krempl (Eds.), *Advances in Intelligent Data Analysis XVIII* (pp. 548–560). Springer International Publishing. https://doi.org/10.1007/978-3-030-44584-3_43

Wang, F., Li, B., Wei, Y., Zhao, Y., Wang, L., Zhang, P., Yang, J., He, W., Chen, H., Jiao, Z., & Li, Y. (2018). Tumor-derived exosomes induce PD1+ macrophage population in human gastric cancer that promotes disease progression. *Oncogenesis*, *7*(5), 41. https://doi.org/10.1038/s41389-018-0049-3

Wang, K., Duan, X., Gao, F., Wang, W., Liu, L., & Wang, X. (2018). Dissecting cancer heterogeneity based on dimension reduction of transcriptomic profiles using extreme learning machines. *PLoS ONE*, *13*(9), e0203824. https://doi.org/10.1371/journal.pone.0203824

Wynn, T. A., Chawla, A., & Pollard, J. W. (2013). Macrophage biology in development, homeostasis and disease. *Nature*, *496*(7446), 445–455. https://doi.org/10.1038/nature12034

Xi, Y., Guo, F., Xu, Z., Li, C., Wei, W., Tian, P., Liu, T., Liu, L., Chen, G., Ye, J., Cheng, G., Cui, L., Zhang, H., Qin, W., & Yin, H. (2018). Radiomics signature: A potential biomarker for the prediction of MGMT promoter methylation in glioblastoma. *Journal of Magnetic Resonance Imaging*, *47*(5), 1380–1387. https://doi.org/10.1002/jmri.25860

Xiang, X., Wang, J., Lu, D., & Xu, X. (2021). Targeting tumor-associated macrophages to synergize tumor immunotherapy. *Signal Transduction and Targeted Therapy*, *6*, 75. https://doi.org/10.1038/s41392-021-00484-9

Xu, C., Wang, C., Ji, F., & Yuan, X. (2012). Finite-Element Neural Network-Based Solving 3-D Differential Equations in MFL. *IEEE Transactions on Magnetics*, *48*(12), 4747–4756. https://doi.org/10.1109/TMAG.2012.2207732

Yadav, R., & Sharma, A. (2012). Advanced Methods to Improve Performance of K-Means Algorithm: A Review. *Global Journal of Computer Science and Technology*. https://www.semanticscholar.org/paper/Advanced-Methods-to-Improve-Performance-of-K-Means-Yadav-Sharma/e61b2d23108366d1780edd0873f26d3aa8099df9

Yan, S., & Wan, G. (2021). Tumor-associated macrophages in immunotherapy. *The FEBS Journal*, *288*(21), 6174–6186. https://doi.org/10.1111/febs.15726

Yang, Q., Zhang, H., Wei, T., Lin, A., Sun, Y., Luo, P., & Zhang, J. (2021). Single-Cell RNA Sequencing Reveals the Heterogeneity of Tumor-Associated Macrophage in Non-Small Cell Lung Cancer and Differences Between Sexes. *Frontiers in Immunology*, *12*, 756722. https://doi.org/10.3389/fimmu.2021.756722

Yauney, G., & Shah, P. (2018). Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection. *Proceedings of the 3rd Machine Learning for Healthcare Conference*, 161–226. https://proceedings.mlr.press/v85/yauney18a.html

Zadeh Shirazi, A., Fornaciari, E., Bagherian, N. S., Ebert, L. M., Koszyca, B., & Gomez, G. A. (2020). DeepSurvNet: Deep survival convolutional network for brain cancer survival rate classification based on histopathological images. *Medical & Biological Engineering & Computing*, *58*(5), 1031–1045. https://doi.org/10.1007/s11517-020-02147-3

Zhang, L., Li, Z., Skrzypczynska, K. M., Fang, Q., Zhang, W., O'Brien, S. A., He, Y., Wang, L., Zhang, Q., Kim, A., Gao, R., Orf, J., Wang, T., Sawant, D., Kang, J., Bhatt, D., Lu, D., Li, C.-M., Rapaport, A. S., … Yu, X. (2020). Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. *Cell*, *181*(2), 442-459.e29. https://doi.org/10.1016/j.cell.2020.03.048

Zhang, Q., He, Y., Luo, N., Patel, S. J., Han, Y., Gao, R., Modak, M., Carotta, S., Haslinger, C., Kind, D., Peet, G. W., Zhong, G., Lu, S., Zhu, W., Mao, Y., Xiao, M., Bergmann, M., Hu, X., Kerkar, S. P., … Zhang, Z. (2019). Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell*, *179*(4), 829-845.e20. https://doi.org/10.1016/j.cell.2019.10.003

Zhang, Q., Liu, L., Gong, C., Shi, H., Zeng, Y., Wang, X., Zhao, Y., & Wei, Y. (2012). Prognostic significance of tumor-associated macrophages in solid tumor: A meta-analysis of the literature. *PloS One*, *7*(12), e50946. https://doi.org/10.1371/journal.pone.0050946

Zhu, Y., Herndon, J. M., Sojka, D. K., Kim, K.-W., Knolhoff, B. L., Zuo, C., Cullinan, D. R., Luo, J., Bearden, A. R., Lavine, K. J., Yokoyama, W. M., Hawkins, W. G., Fields, R. C., Randolph, G. J., & DeNardo, D. G. (2017). Tissue-Resident Macrophages in Pancreatic Ductal Adenocarcinoma Originate from Embryonic Hematopoiesis and Promote Tumor Progression. *Immunity*, *47*(2), 323-338.e6. https://doi.org/10.1016/j.immuni.2017.07.014

Appendix/Appendices

Appendix 1: Code

The following is the code for K-means Clustering,

```
# Load necessary libraries
library(ggplot2)
library(cluster)
library(factoextra)
library(corrplot)
library(pheatmap)

# Set seed for reproducibility
set.seed(123)

# Read data from file
MMSimData <- read.table("./MMSimData.txt", header = TRUE, as.is = TRUE,
row.names = 1)

# Check and handle missing values
if (any(is.na(MMSimData))) {
  MMSimData <- na.omit(MMSimData)
}

# Choose the number of clusters (K) using the elbow method
wss <- numeric(10)
for (i in 1:10) {
  kmeans_temp <- kmeans(MMSimData, centers = i, iter.max = 1000000)
  wss[i] <- sum(kmeans_temp$withinss)
}
# Plot the elbow method graph
plot(1:10, wss, type = "b", main = "Elbow Method", xlab = "Number of
Clusters (K)", ylab = "Within-cluster Sum of Squares")

# Choose an optimal value for K based on the elbow method
optimal_k <- 2  # Set initial choice of K

# Apply K-means algorithm with the optimal K
kmeans_result <- kmeans(MMSimData, centers = optimal_k, iter.max =
1000000)

# Silhouette Analysis
silhouette_avg <- silhouette(kmeans_result$cluster, dist(MMSimData))
print(summary(silhouette_avg))

# Extract cluster assignments
cluster_assignments <- as.factor(kmeans_result$cluster)

# Perform PCA for 2D visualization
pca_result <- prcomp(MMSimData, scale. = TRUE)
pca_data <- as.data.frame(pca_result$x)
```

```r
# Combine PCA data with cluster assignments
pca_data <- cbind(pca_data, Cluster = cluster_assignments)

# Visualization
# 1. Plot PCA using ggplot
ggplot(pca_data, aes(x = PC1, y = PC2, color = Cluster)) +
  geom_point(size = 3) +
  labs(title = "K-means Clustering Visualization (2D PCA)", x = "Principal
Component 1", y = "Principal Component 2") +
  theme_minimal()

# 2. Plot PCA using factoextra
fviz_pca_ind(pca_result, geom.ind = "point", col.ind =
cluster_assignments, title = "K-means Clustering Visualization (2D PCA)")

# Plot clustering results using factoextra
fviz_cluster(kmeans_result, data = MMSimData, geom = "point",
             stand = FALSE, ellipse = TRUE,
             ellipse.type = "norm", main = "K-means Clustering (9600 time
steps)")

# 3. Heatmap
heatmap_data <- MMSimData
heatmap_data$Cluster <- cluster_assignments
# Order rows by cluster
heatmap_data <- heatmap_data[order(heatmap_data$Cluster), -
ncol(heatmap_data)]
# Data visualization through heatmap
heatmap_matrix <- as.matrix(heatmap_data[, 1:14])
heatmap(heatmap_matrix, Colv = NA, scale = "row", Rowv = NA,
        col = colorRampPalette(c("darkblue", "white", "darkred"))(50),
        main = "Heatmap of Clustering Results", cexCol = 0.8)

# 4. Correlation Plot
# Calculate the correlation matrix
corr_matrix <- cor(MMSimData[, -ncol(MMSimData)])
# Correlation Heatmap with adjustments
corrplot(corr_matrix, method = "color", type = "upper", order = "hclust",
         addCoef.col = "black", tl.col = "black", tl.srt = 45, tl.cex =
0.7,
         mar = c(0, 0, 2, 0), number.cex = 0.7, width = 10, height = 10,
         title = "Correlation Heatmap")

# Saving clustering results for supervised analysis
# Extract cluster assignments
cluster_assignments <- as.factor(kmeans_result$cluster)
# Assign cluster labels to MMSimData
MMSimData$Cluster <- as.factor(kmeans_result$cluster)
# Save the results to a CSV file
write.csv(MMSimData, file = "KMeansClusteringOutput.csv")
```

The following is the code for Logistic Regression,

```r
# Load necessary libraries
library(tidyverse)
library(caret)
library(glmnet)
library(Matrix)
library(pROC)
library(ggplot2)
library(RColorBrewer)

# Load your data
KMCdata <- read.csv("KmeansClusteringOutput.csv", header = TRUE, as.is =
TRUE, row.names = 1)

# Check for missing values
missing_values <- colSums(is.na(KMCdata))
print(missing_values)

# Convert cluster to a factor
KMCdata$Cluster <- as.factor(KMCdata$Cluster)

# Convert Cluster to a factor with consistent levels
level <- levels(KMCdata$Cluster)
KMCdata$Cluster <- factor(KMCdata$Cluster, levels = level)

# Split the data into training and testing sets
set.seed(123)  # Set seed for reproducibility

# Shuffle the data
set.seed(123)
shuffled_indices <- sample(nrow(KMCdata))
KMCdata <- KMCdata[shuffled_indices, ]


# Manual K-fold cross-validation
num_folds <- 5
fold_size <- nrow(KMCdata) %/% num_folds
cv_auc_values <- numeric(num_folds)
predictions_list <- list()

for (fold in 1:num_folds) {
  # Define the indices for the current fold
  start_index <- (fold - 1) * fold_size + 1
  end_index <- fold * fold_size

  # Extract the fold for validation
  validation_fold <- KMCdata[start_index:end_index, ]

  # Extract the remaining folds for training
  training_folds <- KMCdata[-(start_index:end_index), ]

  # Fit logistic regression model using glmnet
  x_train <- as.matrix(training_folds[, -ncol(training_folds)])
  y_train <- as.numeric(training_folds$Cluster) - 1
  scaled_x_train <- scale(x_train)
```

```
  model <- cv.glmnet(scaled_x_train, y_train, alpha = 1, family =
"binomial", nfolds = num_folds)

  # Make predictions on the validation set
  x_validation <- as.matrix(validation_fold[, -ncol(validation_fold)])
  scaled_x_validation <- scale(x_validation)
  predictions <- predict(model, newx = scaled_x_validation, s =
"lambda.min", type = "response")

  # Save predictions for later analysis
  predictions_list[[fold]] <- data.frame(Actual = validation_fold$Cluster,
Predicted = as.numeric(predictions[, 1]))

  # Calculate AUC for the current fold
  roc_curve <- roc(y_train, as.numeric(predict(model, newx =
scaled_x_train, s = "lambda.min", type = "response")[, 1]))
  cv_auc_values[fold] <- auc(roc_curve)
}


# Average AUC across folds
average_auc <- mean(cv_auc_values)
cat("Average AUC across folds:", round(average_auc, 3), "\n")

# Individual AUC-ROC curves
for (fold in 1:num_folds) {
  pred <- predictions_list[[fold]]
  roc_curve <- roc(as.factor(pred$Actual), pred$Predicted)
  auc_value <- auc(roc_curve)

  # Plot individual AUC-ROC curve
  plot(roc_curve, col = "blue", main = "AUC-ROC Curve", col.main =
"darkblue", lwd = 2,
       sub = paste("AUC =", round(auc_value, 3)))
}

legend("bottomright", legend = paste("Fold", 1:num_folds), col = "blue",
lty = 1, cex = 0.8)

# Calculate and plot overall AUC-ROC curve
# Use the predictions from the last fold for the overall AUC
overall_predictions <- predict(model, newx = scaled_x_validation, s =
"lambda.min", type = "response")
roc_curve <- roc(as.factor(validation_fold$Cluster),
as.numeric(overall_predictions[, 1]))

# Plot overall AUC-ROC curve
plot(roc_curve, col = "red", main = "Overall AUC-ROC Curve", col.main =
"darkred", lwd = 2)
text(0.8, 0.2, paste("Overall AUC =", round(auc(roc_curve), 3)), col =
"black", cex = 1.2, pos = 4)

# ... (rest of your existing code)
```

```r
# Calculate RMSE
rmse <- sqrt(mean((as.numeric(validation_fold$Cluster) - 1 -
as.numeric(overall_predictions[, 1]))^2))
cat("Root Mean Squared Error (RMSE):", round(rmse, 3), "\n")

# Histogram of Residuals
residuals <- as.numeric(validation_fold$Cluster) - 1 -
as.numeric(overall_predictions[, 1])
hist(residuals, main = "Histogram of Residuals", col = "skyblue", border =
"black", xlab = "Residuals")
legend("topright", legend = "Residuals", fill = "skyblue")

# Residual Plot
# Residual Plot
plot(as.numeric(overall_predictions[, 1]), residuals, main = "Residual
Plot", col = "blue", xlab = "Predicted Values", ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
legend("topright", legend = "Residuals", fill = "blue")

# QQ Plot
# QQ Plot
qqnorm(residuals)
qqline(residuals, col = 2)
legend("topleft", legend = "Residuals", col = "black")

# Actual vs. Predicted Plot
plot(as.numeric(overall_predictions[, 1]),
as.numeric(validation_fold$Cluster) - 1, main = "Actual vs. Predicted",
col = "green", xlab = "Predicted Values", ylab = "Actual Values")
abline(a = 0, b = 1, col = "red", lty = 2)


# 2. Confusion Matrix Heatmap using ggplot2
conf_matrix <- table(Actual = as.numeric(validation_fold$Cluster) - 1,
                     Predicted = as.numeric(overall_predictions > 0.5))

# Extract confusion matrix values
conf_matrix_values <- as.table(conf_matrix)

# Convert the confusion matrix to a data frame
conf_matrix_df <- as.data.frame.matrix(conf_matrix_values)

# Add row and column names to the data frame
conf_matrix_df <- cbind(Actual = rownames(conf_matrix_df), conf_matrix_df)
rownames(conf_matrix_df) <- NULL

# Reshape the data for ggplot
conf_matrix_long <- gather(conf_matrix_df, key = "Predicted", value =
"Count", -Actual)

# Confusion Matrix Heatmap
ggplot(conf_matrix_long, aes(x = Predicted, y = Actual, fill = Count)) +
  geom_tile(color = "white") +
  geom_text(aes(label = Count), vjust = 1, color = "black") +
  scale_fill_gradient(low = "white", high = "blue") +
```

```
  labs(title = "Confusion Matrix", x = "Predicted", y = "Actual") +
  theme_minimal() +
  theme(axis.text = element_text(size = 10), axis.title =
element_text(size = 12),
        plot.title = element_text(size = 14, face = "bold")) +
  guides(fill = guide_legend(title = "Count"))

# 3. Variable Importance Plot
coefficients <- as.vector(coef(model)[-1, ])
feature_names <- colnames(x_train)

# Check if lengths match
if (length(coefficients) == length(feature_names)) {
  # Create a data frame
  coef_df <- data.frame(Feature = feature_names, Coefficient =
coefficients)

  # Variable Importance Plot
  variable_importance_plot <- ggplot(coef_df, aes(x = reorder(Feature,
Coefficient), y = Coefficient)) +
    geom_bar(stat = "identity", fill = "skyblue") +
    coord_flip() +
    labs(title = "Variable Importance Plot", x = "Feature", y =
"Coefficient") +
    theme_minimal() +
    theme(axis.text = element_text(size = 10), axis.title =
element_text(size = 12),
          plot.title = element_text(size = 14, face = "bold")) +
    guides(fill = guide_legend(title = "Coefficient"))
  print(variable_importance_plot)

}

# 4. Calibration plot
# Fit the final model on the full training set
final_glmnet_model <- glmnet(scaled_x_train, y_train, alpha = 1, family =
"binomial", lambda = model$lambda.min)
final_model <- glm(as.numeric(validation_fold$Cluster) - 1 ~
predict(final_glmnet_model, newx = scaled_x_validation, type =
"response"), family = "binomial", maxit = 1000)

# Predict on the validation set
validation_probabilities <- predict(final_glmnet_model, newx =
scaled_x_validation, type = "response")

# Combine predictions and actual values
calibration_data <- data.frame(Predicted =
as.numeric(validation_probabilities), Actual =
as.numeric(validation_fold$Cluster) - 1)

# Calibration Plot with Custom Legend
calibration_plot <- ggplot(calibration_data, aes(x = Predicted, y =
Actual)) +
  geom_point(aes(color = "Actual vs. Predicted"), show.legend = TRUE) +
```

```
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se
= FALSE, aes(color = "Smoothed Line"), show.legend = TRUE) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color =
"red", aes(color = "Ideal Line"), show.legend = TRUE) +
  labs(title = "Calibration Plot", x = "Predicted Probability", y =
"Actual Probability") +
  theme_minimal() +
  theme(axis.text = element_text(size = 10), axis.title =
element_text(size = 12),
        plot.title = element_text(size = 14, face = "bold")) +
  scale_color_manual(values = c("black", "blue", "red"), guide = "legend",
                     labels = c("Actual vs. Predicted", "Smoothed Line",
"Ideal Line"))
```

The following is the code for Support Vector Machine,

```
# Install and load necessary packages
library(e1071)
library(pROC)
library(corrplot)
library(ggplot2)
library(reshape2)
library(caret)

# Load your data
KMCdata <- read.csv("KmeansClusteringOutput.csv", header = TRUE, as.is =
TRUE, row.names = 1)

# Convert response variable to a factor with two levels
KMCdata$cluster <- factor(KMCdata$Cluster)

# Set the number of folds for cross-validation
num_folds <- 5

# Create a data partition for cross-validation
set.seed(123)  # Set seed for reproducibility
folds <- createFolds(KMCdata$cluster, k = num_folds, list = TRUE)

# Initialize variables to store cross-validation results
cv_results <- NULL
all_true_labels <- all_predicted_probs <- NULL

# Initialize a list to store confusion matrices for each fold
confusion_matrices <- list()

# Define the parameter grid for tuning
param_grid <- expand.grid(
  cost = c(0.1, 1, 10),  # Adjust the range of cost values
  kernel = c("linear", "radial")  # Include other kernel types if needed
)

# Perform K-fold cross-validation with parameter tuning
for (i in 1:num_folds) {
  # Split the data into training and testing sets for the current fold
```

```r
  training_data <- KMCdata[-folds[[i]], ]
  testing_data <- KMCdata[folds[[i]], ]

  # Convert cluster variable to a factor with two levels
  training_data$cluster <- factor(training_data$cluster)
  testing_data$cluster <- factor(testing_data$cluster)

  # Scale the feature variables
  training_data[, -which(names(training_data) == "cluster")] <-
scale(training_data[, -which(names(training_data) == "cluster")])
  testing_data[, -which(names(testing_data) == "cluster")] <-
scale(testing_data[, -which(names(testing_data) == "cluster")])

  # Train the SVM model with parameter tuning
  svm_model <- svm(
    cluster ~ .,
    data = training_data,
    kernel = "radial",  # Adjust the kernel type
    cost = 1,           # Use a default cost value
    probability = TRUE
  )

  # Make predictions on the testing set
  predictions <- predict(svm_model, newdata = testing_data, probability =
TRUE)

  # Evaluate the performance (you can replace this with your own
evaluation metric)
  accuracy <- confusionMatrix(predictions,
testing_data$cluster)$overall["Accuracy"]

  # Store the results for this fold
  cv_results <- rbind(cv_results, data.frame(Fold = i, Accuracy =
accuracy))

  # Store true labels and predicted probabilities for ROC curve
  all_true_labels <- c(all_true_labels, as.numeric(testing_data$cluster) -
1)
  all_predicted_probs <- c(all_predicted_probs,
attributes(predictions)$probabilities[, 2])

  # Store confusion matrix for this fold
  confusion_matrices[[i]] <- as.table(confusionMatrix(predictions,
testing_data$cluster)$table)
}

# Display the cross-validation results
print(cv_results)

# Calculate the mean and standard deviation of the accuracy across folds
mean_accuracy <- mean(cv_results$Accuracy)
sd_accuracy <- sd(cv_results$Accuracy)

cat("Mean Accuracy:", mean_accuracy, "\n")
```

```r
cat("Standard Deviation of Accuracy:", sd_accuracy, "\n")

# Visualize the mean ROC curve
roc_curve <- roc(all_true_labels, all_predicted_probs)

# Convert roc_curve to a data frame
roc_data <- data.frame(
  specificity = 1 - roc_curve$specificities,
  sensitivity = roc_curve$sensitivities
)

# Plot ROC curve
roc_plot <- ggplot(roc_data, aes(x = 1 - specificity, y = sensitivity)) +
  geom_line(color = "blue") +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color =
"red") +
  labs(title = "Mean ROC Curve", x = "1 - Specificity", y = "Sensitivity")
+
  theme_minimal()

roc_plot + theme_minimal() + theme(axis.text.x = element_text(angle = 45,
hjust = 1), text = element_text(size = 12))

# Add AUC to the plot
auc_value <- auc(roc_curve)
cat("Mean AUC:", auc_value, "\n")

roc_plot +
  annotate("text", x = 0.8, y = 0.2, label = paste("AUC =",
round(auc_value, 2)), color = "black", size = 4) +
  theme_minimal()
```

```r
# Visualize the precision-recall curve using pROC package
pr_curve <- roc(all_true_labels, all_predicted_probs)

# Set a threshold (you can choose a threshold based on your preference)
threshold <- 0.5  # For example, you can use 0.5 as a threshold

# Extract precision-recall curve data for the specified threshold
pr_data <- coords(pr_curve, threshold = threshold, input = "threshold",
ret = c("sensitivity", "specificity"))

# Plot precision-recall curve
pr_plot <- ggplot(pr_data, aes(x = sensitivity, y = specificity)) +
  geom_line(color = "green") +
  labs(title = "Precision-Recall Curve", x = "Recall", y = "Precision") +
  theme_minimal()
pr_plot + theme_classic()

print(pr_plot)

# Plot confusion matrix for each fold
for (i in 1:num_folds) {
```

```
  # Convert confusion matrix to a data frame
  cm_df <- as.data.frame(as.table(confusion_matrices[[i]]))

  # Plot confusion matrix
  cm_plot <- ggplot(data = cm_df, aes(x = Prediction, y = Reference, fill
= Freq)) +
    geom_tile(color = "white") +
    scale_fill_gradient(low = "white", high = "steelblue") +
    labs(title = paste("Confusion Matrix - Fold", i), x = "Predicted", y =
"Actual") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(cm_plot)
}

# Visualize the calibration plot
calibration_data <- data.frame(
  observed = all_true_labels,
  predicted = all_predicted_probs
)

calibration_plot <- ggplot(calibration_data, aes(x = predicted, y =
observed)) +
  geom_smooth(method = "loess", se = FALSE, color = "orange") +
  geom_point() +
  labs(title = "Calibration Plot", x = "Predicted Probability", y =
"Observed Outcome") +
  theme_minimal()

print(calibration_plot)
```

Appendix 2: Miscellaneous Result Visualization (9600-time stamps)

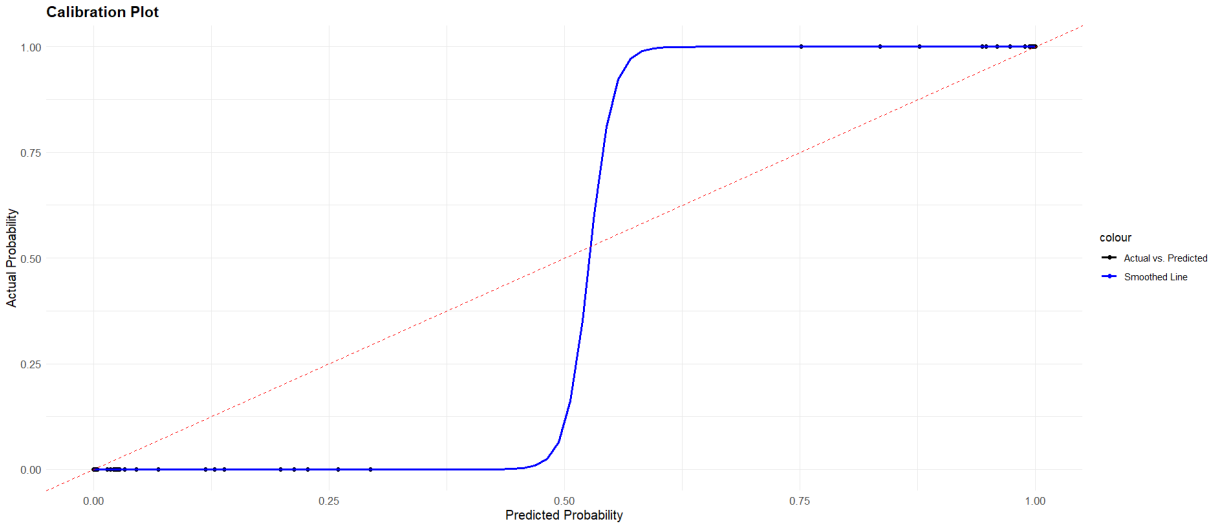

Figure 19. Calibration Plot for Logistic Regression

Appendix 3: Result Visualization (38400-time stamps)

K-means Clustering

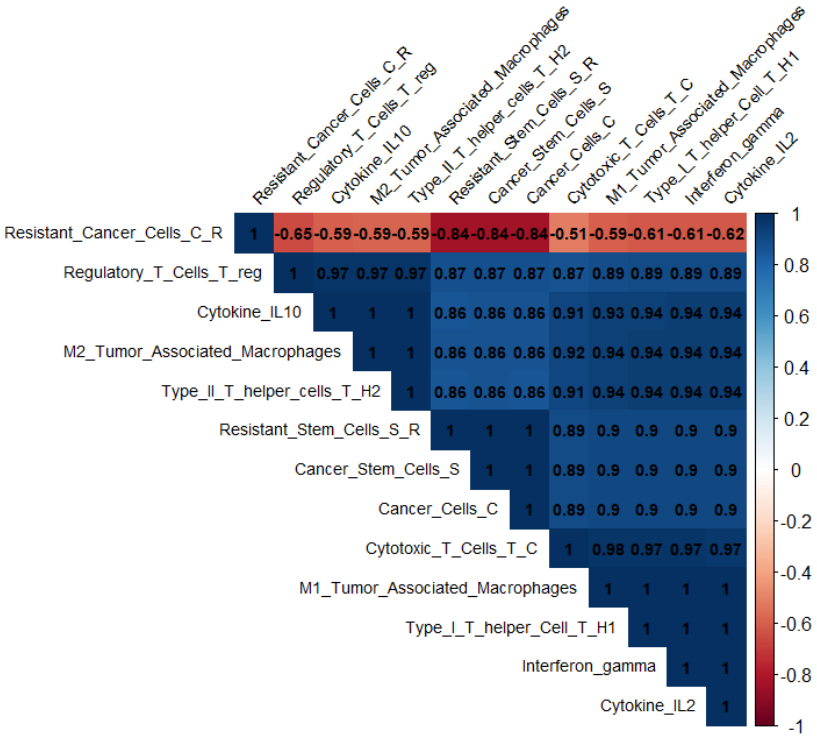

Figure 20. Correlation Matrix Heatmap

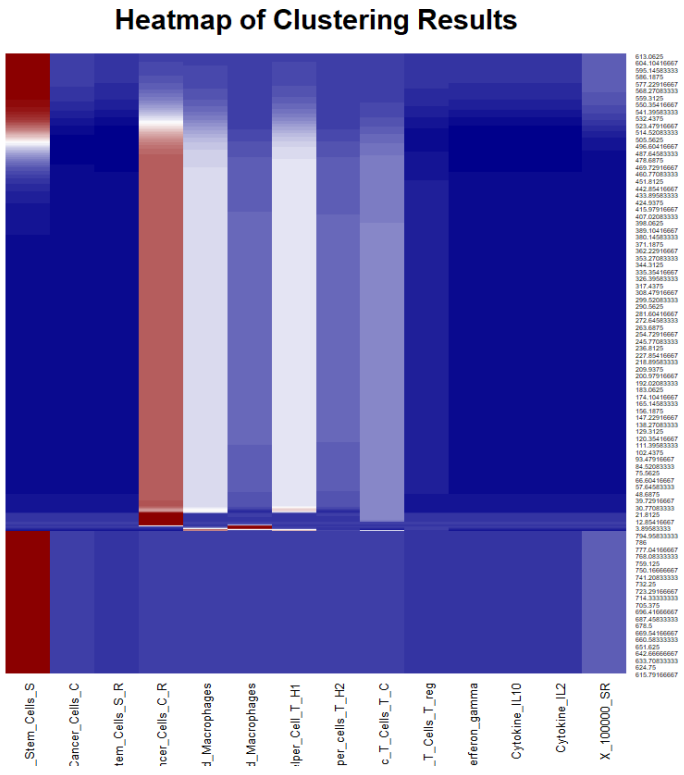

Figure 21. Heatmap of Clustering Results

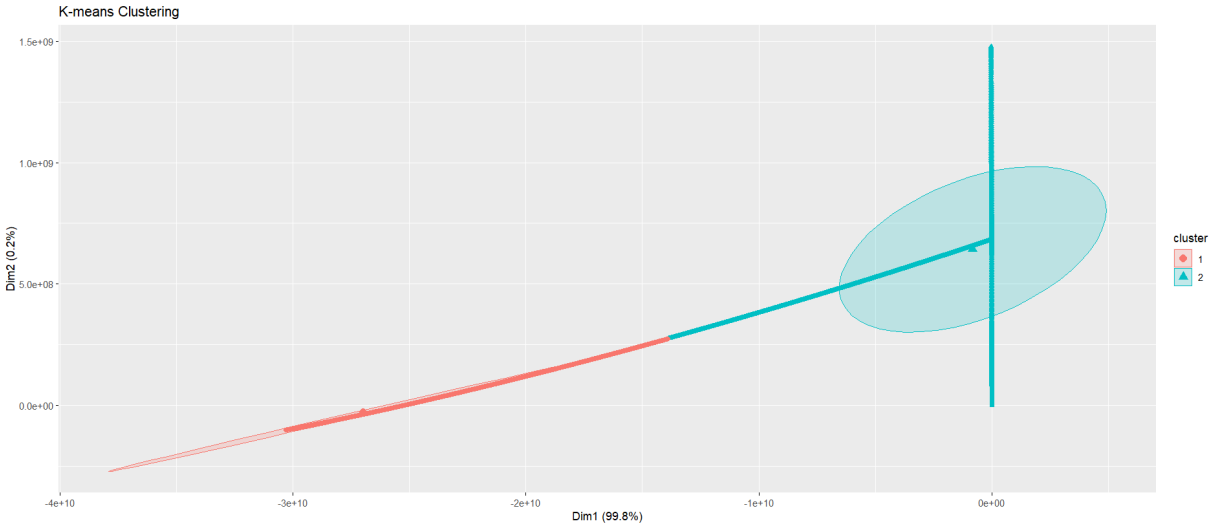

Figure 22. K-means Clustering results using factoextra package

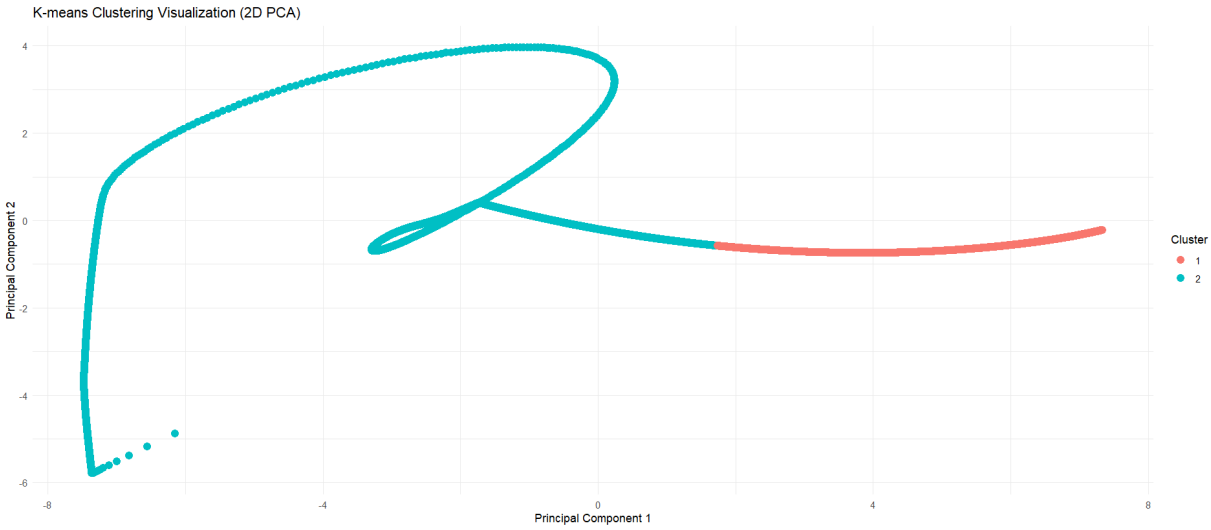

Figure 23. K-means Clustering results after 2D-PCA

Logistic Regressions

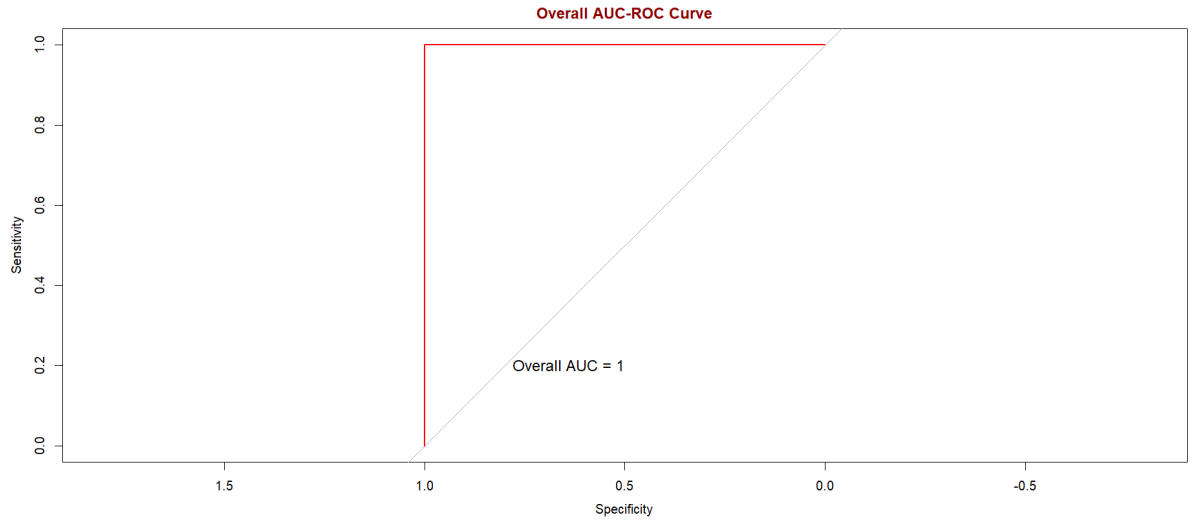

Figure 24. AUC-ROC for Logistic Regression

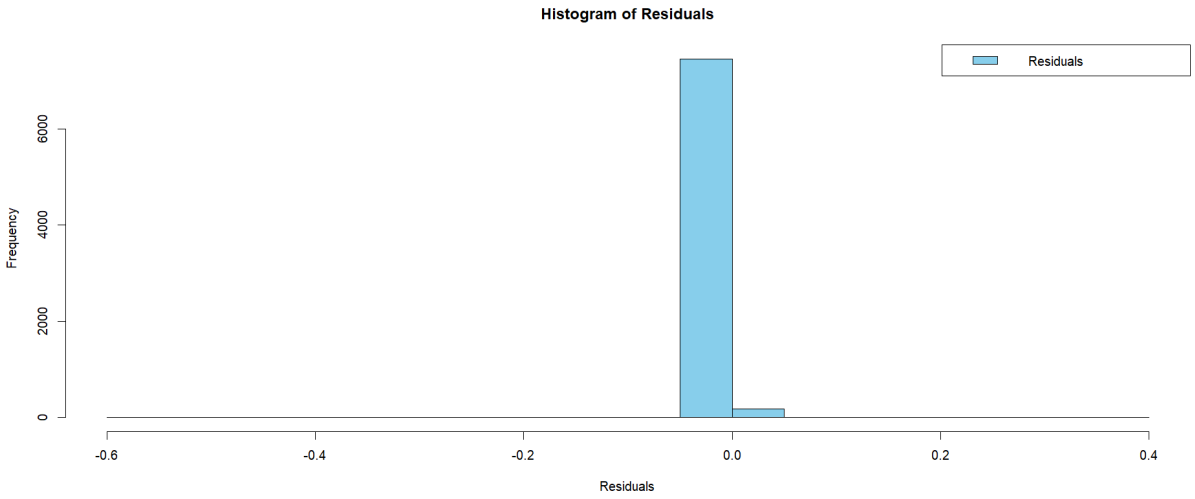

Figure 25. Histogram of Residuals for Logistic Regression

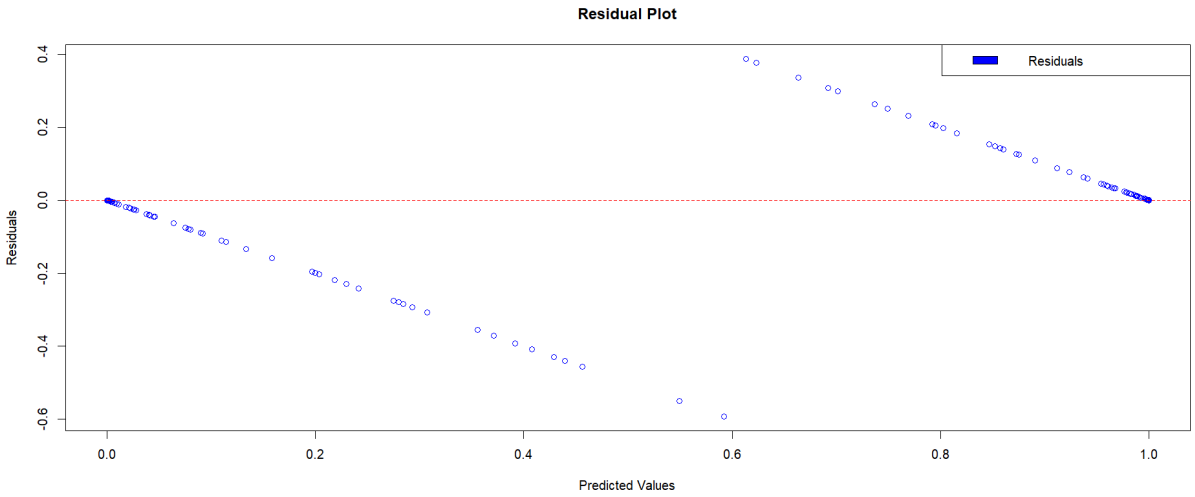

Figure 26. Residual Plot for Logistic Regression

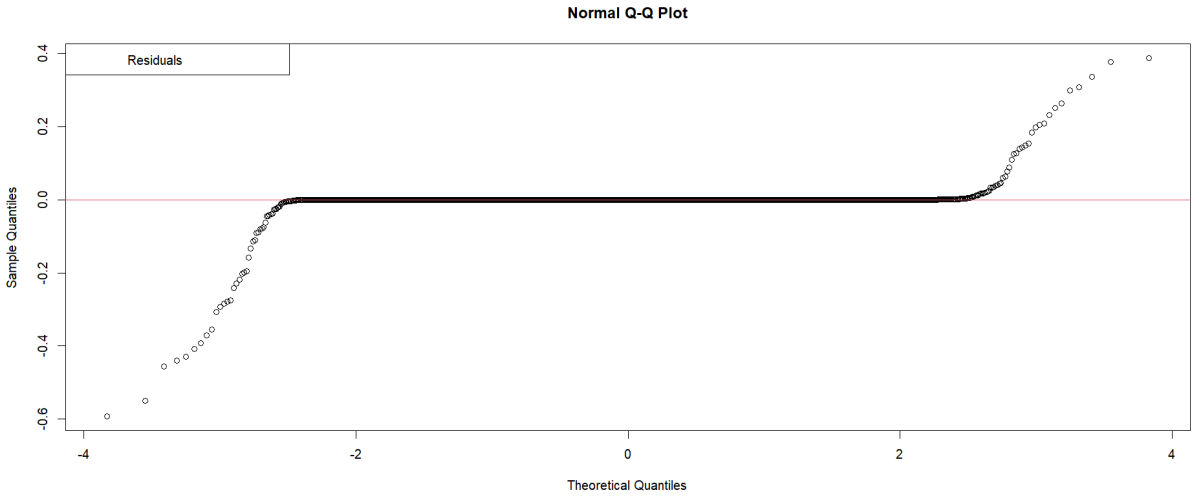

Figure 27. Normal QQ Plot for Logistic Regression

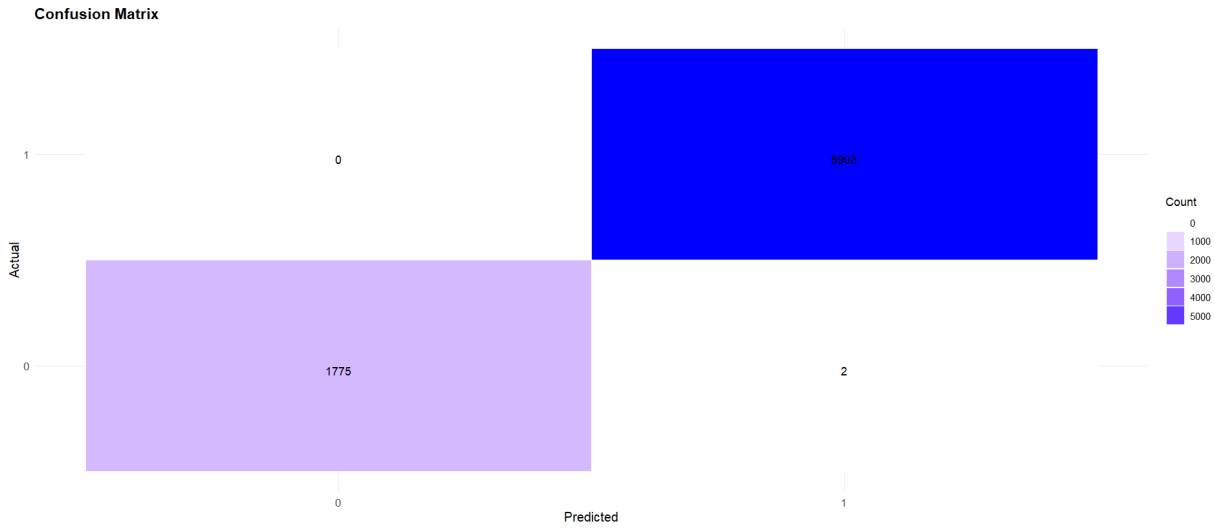

Figure 28. Confusion Matrix for Logistic Regression

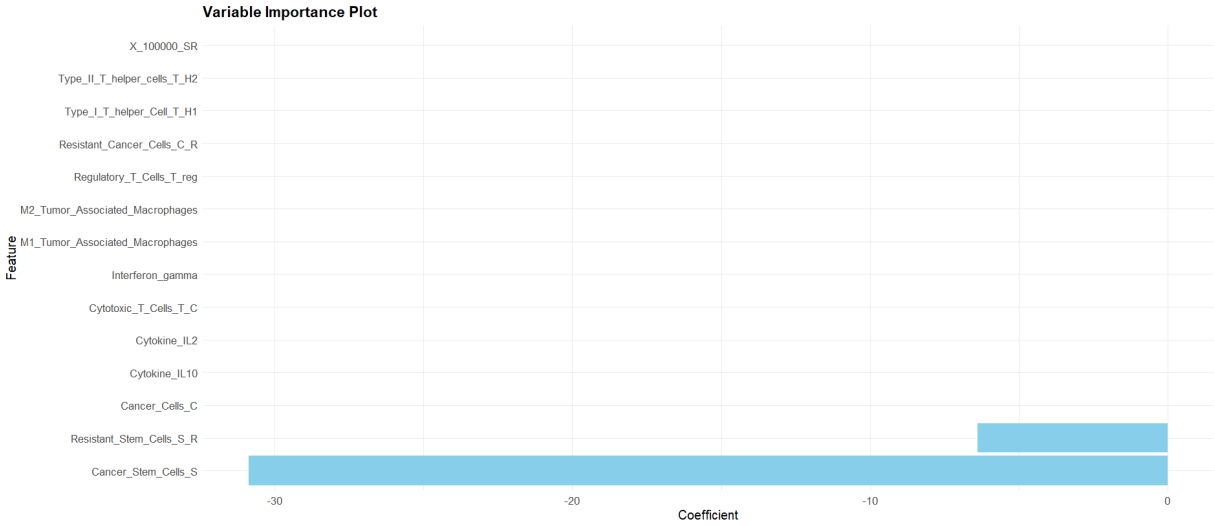

Figure 29. Variable Importance Plot for Logistic Regression

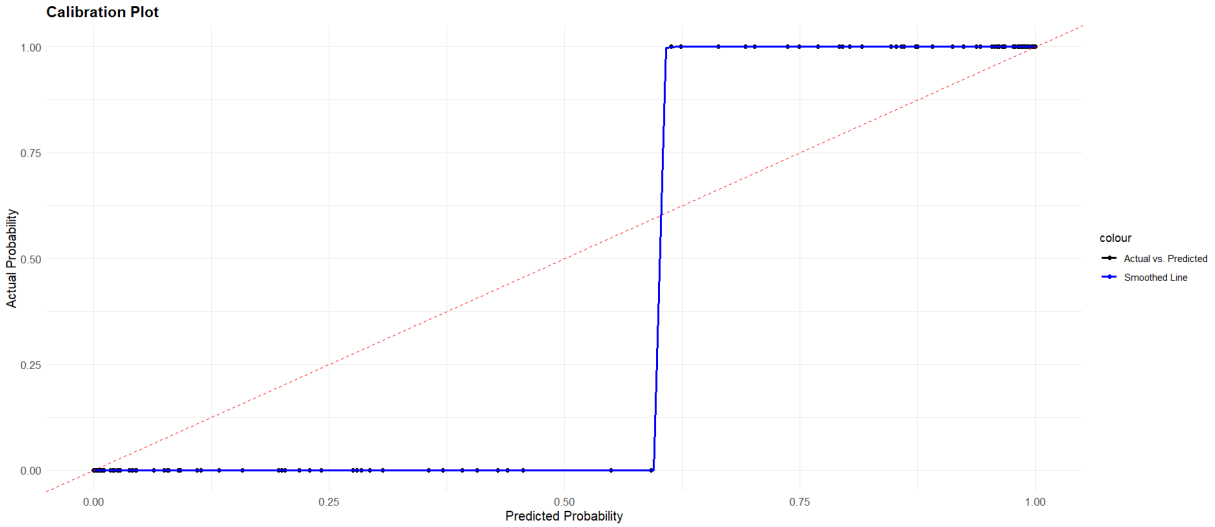

Figure 30. Calibration Plot for Logistic Regression

Support Machine Vector

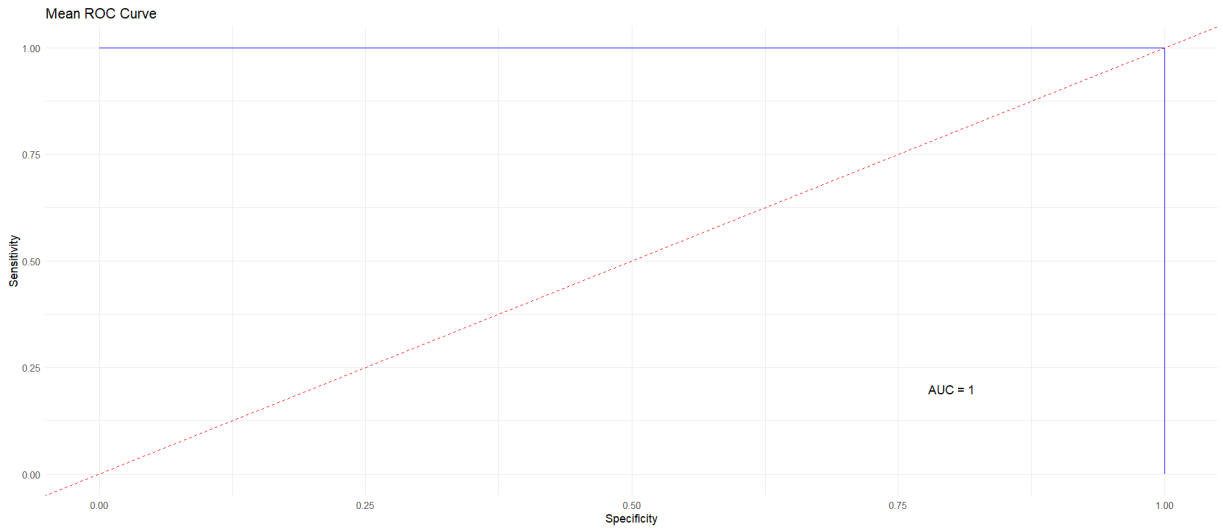

Figure 31. AUC-ROC for SVM

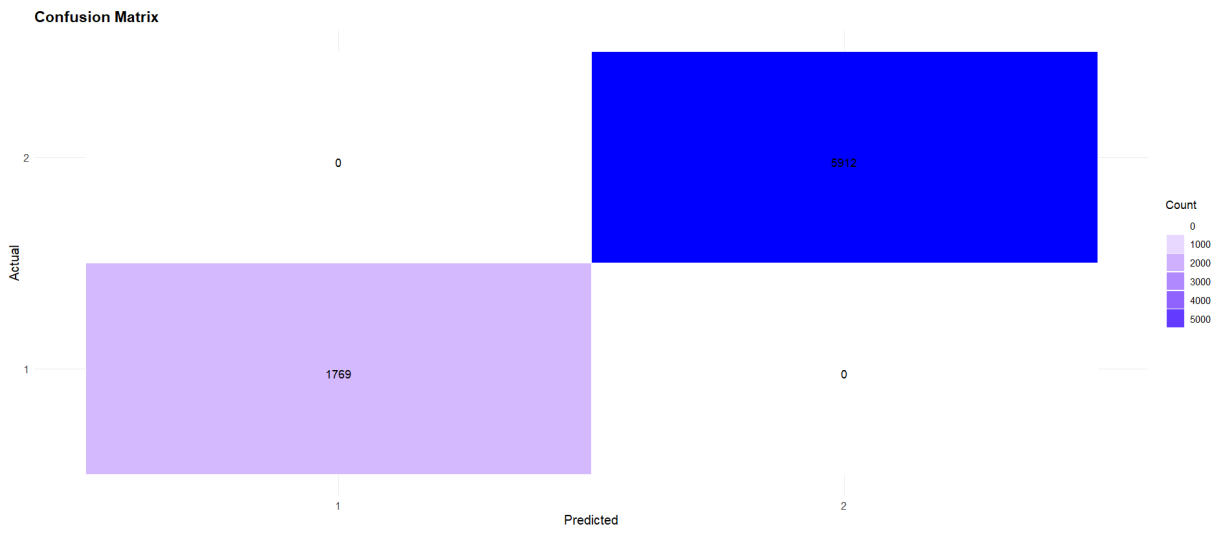

Figure 32. Confusion Matrix for SVM

Appendix 4: Linear Regression

```
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(ROCR)

# Load your data
KMCdata <- read.csv("KmeansClusteringOutput.csv", header = TRUE, as.is =
TRUE, row.names = 1)

# Display summary and head of the data
summary(KMCdata)

##  Cancer_Stem_Cells_S Cancer_Cells_C      Resistant_Stem_Cells_S_R
##  Min.   :1.000e+00   Min.   :        0   Min.   :    0.000
##  1st Qu.:5.581e+03   1st Qu.:      150   1st Qu.:    0.006
##  Median :8.727e+06   Median :   232626   Median :    8.852
##  Mean   :6.796e+09   Mean   :170380392   Mean   : 6236.114
##  3rd Qu.:9.581e+09   3rd Qu.:248804039   3rd Qu.: 9384.785
##  Max.   :3.016e+10   Max.   :749320004   Max.   :27143.956
##  Resistant_Cancer_Cells_C_R M1_Tumor_Associated_Macrophages
##  Min.   :0.000e+00          Min.   :        0
##  1st Qu.:3.137e+08          1st Qu.:261287298
##  Median :5.211e+08          Median :261379969
##  Mean   :4.279e+08          Mean   :312766740
##  3rd Qu.:5.215e+08          3rd Qu.:339875793
##  Max.   :1.416e+09          Max.   :640468287
##  M2_Tumor_Associated_Macrophages Type_I_T_helper_Cell_T_H1
##  Min.   :     9028               Min.   :        0
##  1st Qu.:111497154               1st Qu.:278474790
##  Median :111532479               Median :278572608
##  Mean   :123486405               Mean   :332998762
##  3rd Qu.:125184679               3rd Qu.:361279033
##  Max.   :199627612               Max.   :652741109
##  Type_II_T_helper_cells_T_H2 Cytotoxic_T_Cells_T_C
Regulatory_T_Cells_T_reg
##  Min.   :        0           Min.   :        0     Min.   :       0
##  1st Qu.:111396601           1st Qu.:151011165     1st Qu.:23448654
##  Median :111432289           Median :151059080     Median :23459587
##  Mean   :123261125           Mean   :175750560     Mean   :26874892
##  3rd Qu.:124878559           3rd Qu.:189930089     3rd Qu.:27389977
##  Max.   :199517068           Max.   :327354323     Max.   :49880509
##  Interferon_gamma Cytokine_IL10       Cytokine_IL2     X_100000_SR
##  Min.   : 0.000   Min.   :0.000000   Min.   :0.0000   Min.   :0.000e+00
##  1st Qu.: 4.796   1st Qu.:0.005757   1st Qu.:0.3214   1st Qu.:5.660e+02
##  Median : 4.798   Median :0.005759   Median :0.3215   Median :8.852e+05
##  Mean   : 5.725   Mean   :0.006375   Mean   :0.3842   Mean   :6.236e+08
##  3rd Qu.: 6.204   3rd Qu.:0.006460   3rd Qu.:0.4165   3rd Qu.:9.385e+08
##  Max.   :11.119   Max.   :0.010351   Max.   :0.7522   Max.   :2.714e+09
##     Cluster
##  Min.   :1.00
```

```
##  1st Qu.:2.00
##  Median :2.00
##  Mean   :1.77
##  3rd Qu.:2.00
##  Max.   :2.00

head(KMCdata)

##            Cancer_Stem_Cells_S Cancer_Cells_C Resistant_Stem_Cells_S_R
## 0                     1.000000      0.0000000             0.000000e+00
## 0.08333333            1.008655      0.2615118             4.286760e-09
## 0.16666667            1.017437      0.4926588             8.648148e-09
## 0.25                  1.026328      0.7040165             1.308557e-08
## 0.33333333            1.035317      0.8987405             1.760023e-08
## 0.41666667            1.044396      1.0787727             2.219322e-08
##            Resistant_Cancer_Cells_C_R M1_Tumor_Associated_Macrophages
## 0                        0.000000e+00                          85000.00
## 0.08333333               1.633207e-05                          78094.29
## 0.16666667               7.080665e-05                          71743.79
## 0.25                     1.751859e-04                          65905.64
## 0.33333333               3.456643e-04                          60539.97
## 0.41666667               6.040955e-04                          55609.54
##            M2_Tumor_Associated_Macrophages Type_I_T_helper_Cell_T_H1
## 0                                 15000.00                  71000.00
## 0.08333333                        14941.14                  65737.31
## 0.16666667                        14879.92                  60545.29
## 0.25                              14818.24                  55460.28
## 0.33333333                        14756.66                  50526.85
## 0.41666667                        14695.31                  45798.72
##            Type_II_T_helper_cells_T_H2 Cytotoxic_T_Cells_T_C
## 0                            12000.000             56000.000
## 0.08333333                   10388.337             37457.518
## 0.16666667                    8995.065             25084.708
## 0.25                          7790.032             16796.095
## 0.33333333                    6747.397             11231.999
## 0.41666667                    5844.973              7495.692
##            Regulatory_T_Cells_T_reg Interferon_gamma Cytokine_IL10
Cytokine_IL2
## 0                           8000.000       0.12000000  8.500000e-03
0.0094000000
## 0.08333333                 7360.631       0.07252665  1.638763e-03
0.0046065887
## 0.16666667                 6772.359       0.04397805  3.162704e-04
0.0022750107
## 0.25                       6231.103       0.02679639  6.131982e-05
0.0011393616
## 0.33333333                 5733.103       0.01644284  1.213379e-05
0.0005847317
## 0.41666667                 5274.903       0.01019175  2.612872e-06
0.0003124488
##             X_100000_SR Cluster
## 0          0.0000000000       2
## 0.08333333 0.0004286760       2
## 0.16666667 0.0008648148       2
```

```
## 0.25       0.0013085570       2
## 0.33333333 0.0017600228       2
## 0.41666667 0.0022193224       2

# Set seed for reproducibility
set.seed(123)

# Create training and testing sets
index_train <- createDataPartition(KMCdata$Cluster, p = 0.7, list = FALSE)
training_data <- KMCdata[index_train, ]
testing_data <- KMCdata[-index_train, ]

index_valid <- createDataPartition(testing_data$Cluster, p = 0.5, list =
FALSE)
validation_data <- testing_data[index_valid, ]
testing_data <- testing_data[-index_valid, ]

# Scale the features
preproc <- preProcess(training_data[, -ncol(training_data)], method =
c("center", "scale"))
training_data_scaled <- predict(preproc, training_data)
testing_data_scaled <- predict(preproc, testing_data)
validation_data_scaled <- predict(preproc, validation_data)

# Train the linear regression model
lm_model <- lm(formula = Cluster ~ ., data = training_data_scaled)
summary(lm_model)

##
## Call:
## lm(formula = Cluster ~ ., data = training_data_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54832 -0.00617  0.00257  0.00285  0.45001
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       1.768338   0.000968 1826.880  < 2e-16
***
## Cancer_Stem_Cells_S             -33.632214   0.448739  -74.948  < 2e-16
***
## Cancer_Cells_C                   66.021999   1.007314   65.543  < 2e-16
***
## Resistant_Stem_Cells_S_R        -31.908899   0.428318  -74.498  < 2e-16
***
## Resistant_Cancer_Cells_C_R        0.533919   0.202806    2.633  0.00849
**
## M1_Tumor_Associated_Macrophages   0.437615   0.381733    1.146  0.25167
## M2_Tumor_Associated_Macrophages   0.076917   1.070830    0.072  0.94274
## Type_I_T_helper_Cell_T_H1        -3.440746   3.317173   -1.037  0.29966
## Type_II_T_helper_cells_T_H2      -0.050667   1.050945   -0.048  0.96155
## Cytotoxic_T_Cells_T_C            -0.828937   0.382273   -2.168  0.03016
*
```

```
## Regulatory_T_Cells_T_reg          0.012587   0.007067    1.781  0.07494
.
## Interferon_gamma                 -3.033047   3.159067   -0.960  0.33704
## Cytokine_IL10                    -0.059239   0.055842   -1.061  0.28880
## Cytokine_IL2                      6.385987   5.967586    1.070  0.28461
## X_100000_SR                             NA         NA       NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07935 on 6707 degrees of freedom
## Multiple R-squared:  0.9647, Adjusted R-squared:  0.9646
## F-statistic: 1.41e+04 on 13 and 6707 DF,  p-value: < 2.2e-16

# Make predictions
predictions <- predict(lm_model, newdata = testing_data_scaled)

## Warning in predict.lm(lm_model, newdata = testing_data_scaled):
prediction from
## rank-deficient fit; attr(*, "non-estim") has doubtful cases

summary(predictions)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.886   1.976   1.997   1.788   1.999   2.039

# Evaluate the model
rmse <- sqrt(mean((testing_data_scaled$Cluster - predictions)^2))
cat("Root Mean Squared Error (RMSE):", rmse, "\n")

## Root Mean Squared Error (RMSE): 0.07746983

# Visualization Section

# 1. Actual vs. Predicted
plot(testing_data_scaled$Cluster, predictions,
     main = "Actual vs. Predicted",
     col = "blue",
     pch = 16,
     xlab = "Actual Values",
     ylab = "Predicted Values")
abline(0, 1, col = "red", lty = 2)
legend("topleft", legend = c("Observations", "Regression Line"), col =
c("blue", "red"), pch = c(16, NA), lty = c(NA, 2))

# 2. Residual Plot
residuals <- testing_data_scaled$Cluster - predictions
plot(predictions, residuals,
     main = "Residual Plot",
     xlab = "Predicted Values",
     ylab = "Residuals",
     col = "blue",
     pch = 16)
abline(h = 0, col = "red", lty = 2)
legend("topright", legend = "Residuals", col = "blue", pch = 16)
```

```r
# 3. Histogram of Residuals
hist(residuals,
     main = "Histogram of Residuals",
     xlab = "Residuals",
     col = "lightblue",
     border = "black")
legend("topright", legend = "Residuals", fill = "lightblue", border =
"black")

# 4. QQ Plot
par(mfrow = c(1, 2))  # Set up a 1x2 grid for plots
qqnorm(residuals, main = "QQ Plot")
qqline(residuals, col = 2)
legend("topleft", legend = "QQ Line", col = 2, lty = 1)

# Train the linear regression model on validation data
lm_model_validation <- lm(formula = Cluster ~ ., data =
validation_data_scaled)
summary(lm_model_validation)
```

```
##
## Call:
## lm(formula = Cluster ~ ., data = validation_data_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49083 -0.01169  0.00305  0.00365  0.49044
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.77175    0.01945  91.095  < 2e-16
***
## Cancer_Stem_Cells_S             -35.41326    1.18391 -29.912  < 2e-16
***
## Cancer_Cells_C                   61.98430    3.91705  15.824  < 2e-16
***
## Resistant_Stem_Cells_S_R        -35.91396    1.26920 -28.297  < 2e-16
***
## Resistant_Cancer_Cells_C_R       -5.18128    2.27089  -2.282 0.022660
*
## M1_Tumor_Associated_Macrophages  47.12238   13.99668   3.367 0.000781
***
## M2_Tumor_Associated_Macrophages  -1.82621    2.93317  -0.623 0.533643
## Type_I_T_helper_Cell_T_H1       -372.12606  111.82396  -3.328 0.000898
***
## Type_II_T_helper_cells_T_H2       8.10057   22.22790   0.364 0.715589
## Cytotoxic_T_Cells_T_C            20.54013    7.41116   2.772 0.005652
**
## Regulatory_T_Cells_T_reg          0.29050    0.64187   0.453 0.650916
## Interferon_gamma               -272.94778   83.67674  -3.262 0.001133
**
## Cytokine_IL10                    -6.73587   22.06576  -0.305 0.760210
## Cytokine_IL2                    583.24940  176.45407   3.305 0.000972
***
```

```
## X_100000_SR                              NA         NA       NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08603 on 1426 degrees of freedom
## Multiple R-squared:  0.9602, Adjusted R-squared:  0.9599
## F-statistic:  2649 on 13 and 1426 DF,  p-value: < 2.2e-16
```

```
# Make predictions on validation data
validation_predictions <- predict(lm_model_validation, newdata =
validation_data_scaled)
```

```
## Warning in predict.lm(lm_model_validation, newdata =
validation_data_scaled):
## prediction from rank-deficient fit; attr(*, "non-estim") has doubtful
cases
```

```
summary(validation_predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8636  1.6425  1.9964  1.7563  1.9992  2.0724
```

```
# 5. Validation: Actual vs. Predicted
plot(validation_data_scaled$Cluster, validation_predictions,
     main = "Validation: Actual vs. Predicted",
     col = "green",
     pch = 16,
     xlab = "Actual Values",
     ylab = "Predicted Values")
abline(0, 1, col = "red", lty = 2)
```

Appendix 5: Neural Network

```
# Load required libraries
library(neuralnet)
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

# Load your data
MMSimData <- read.csv("KmeansClusteringOutput8000.csv", header = TRUE,
as.is = TRUE, row.names = 1)

# Convert response variable to a factor with two levels
MMSimData$cluster <- factor(MMSimData$Cluster)

set.seed(123)
index_train <- createDataPartition(MMSimData$cluster, p = 0.7, list =
FALSE)
training_data <- MMSimData[index_train, ]
testing_data <- MMSimData[-index_train, ]

index_valid <- createDataPartition(testing_data$cluster, p = 0.5, list =
FALSE)
validation_data <- testing_data[index_valid, ]
testing_data <- testing_data[-index_valid, ]

# Convert cluster variable to a factor with two levels
training_data$cluster <- factor(training_data$cluster)
testing_data$cluster <- factor(testing_data$cluster)
validation_data$cluster <- factor(validation_data$cluster)

# Scale the features
preproc <- preProcess(training_data[, -ncol(training_data)], method =
c("center", "scale"))
training_data_scaled <- predict(preproc, training_data)
testing_data_scaled <- predict(preproc, testing_data)

# Train the neural network
nn_model <- neuralnet(cluster ~ ., data = training_data_scaled, hidden =
c(5, 2), linear.output = TRUE)

# Make predictions
predictions <- as.factor(round(predict(nn_model, newdata =
testing_data_scaled)))
predicted_levels <- levels(testing_data_scaled$cluster)

# Trim predictions to match the length of testing data
predictions <- factor(predictions[1:length(testing_data_scaled$cluster)],
levels = predicted_levels)

# Create confusion matrix
conf_matrix <- table(Reference = testing_data_scaled$cluster, Prediction =
```

```
predictions)
print(conf_matrix)
```

```
##          Prediction
## Reference   1   2
##         1 331   0
##         2   0   0
```

```
# Calculate other performance metrics
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
sensitivity <- diag(conf_matrix) / rowSums(conf_matrix)
specificity <- colSums(conf_matrix) - diag(conf_matrix) /
colSums(conf_matrix) - diag(conf_matrix)

print(paste("Accuracy:", accuracy))
```

```
## [1] "Accuracy: 1"
```

```
print(paste("Sensitivity:", sensitivity))
```

```
## [1] "Sensitivity: 1"   "Sensitivity: NaN"
```

```
print(paste("Specificity:", specificity))
```

```
## [1] "Specificity: -1"  "Specificity: NaN"
```