NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
“DATA SCIENCE AND INFORMATION TECHNOLOGIES”
SPECIALIZATION
“BIOINFORMATICS – BIOMEDICAL DATA SCIENCE”

MASTER THESIS

"Automated workflow for Senolytic drug discovery using machine learning"

Batoul A. Khalil

Supervised by:

Athanasios Papakyriakou, Senior Researcher, National Center for Scientific Research “Demokritos”

Anastasia Krithara, Postdoctoral Researcher, head of the Biomedical and Health Informatics Team (BioHIT), National Center for Scientific Research “Demokritos”

Stavros Perantonis, Research Director, National Center for Scientific Research “Demokritos”

ATHENS
August 2024

MASTER THESIS

"Automated Workflow for Senolytic Drug Discovery Using Machine Learning"

Batoul A. Khalil
S.N :7115152200025

Supervised by:

Athanasios Papakyriakou, Senior Researcher, National Center for Scientific Research "Demokritos"

Anastasia Krithara, Postdoctoral Researcher, head of the Biomedical and Health Informatics Team (BioHIT), National Center for Scientific Research "Demokritos"

Stavros Perantonis, Research Director, National Center for Scientific Research "Demokritos"

EXAMINATION COMMITTEE:

Stavros Perantonis, Research Director, National Center for Scientific Research "Demokritos"

Anastasia Krithara, Postdoctoral Researcher, head of the Biomedical and Health Informatics Team (BioHIT), National Center for Scientific Research "Demokritos"

Dr Evangelia D. Chrysina, Research Director, ICB-NHRF

August 2024

Abstract

Cellular senescence, first identified in 1961, has become a major focus for biotech companies aiming to enhance human health. This phenomenon is characterized by cells ceasing to divide permanently in response to various internal and external stresses, such as telomere attrition, activation of oncogenes, and persistent DNA damage. The development of senolytic drugs has been hindered by the absence of well-characterized molecular targets.

In the recent study "Discovery of Senolytics Using Machine Learning," (14) researchers leveraged cost-effective machine learning algorithms trained on published data to address this challenge. The models, trained on both senolytic and non-senolytic compounds, were used to computationally screen a chemical library. Out of 21 experimentally tested compounds, three demonstrated senolytic activity, with Oleandrin showing superior potency compared to existing alternatives. This approach not only reduced drug screening costs but also showcased the potential of artificial intelligence in early-stage drug discovery.

Building on this work, our project aims to expand the methodology by exploring novel senolytic drug possibilities. Our approach involves three main steps: feature selection, model optimization, and performance comparison. We employed three feature selection techniques in parallel (random forest, mutual information, and Fiser score) and utilized cross-validation to optimize five machine learning models (SVM, XGB, NB, LR, RF). Our efforts included reproducing the results from the aforementioned study (14) and developing a test set to validate our models. Additionally, we incorporated new chemical descriptors called Mordred descriptors, in addition to RDkit descriptors used by the original authors. We also compared our models with a shallow neural network to explore the potential benefits of deep learning in this context.

The findings indicate that hyperparameter tuning and cross-validation positively influence model accuracy, and Mordred descriptors outperform RDkit descriptors. Additionally, feature selection based on mutual information leads to relatively higher precision than using Random Forest (RF) features, although it results in greater standard deviation.

SUBJECT AREA: Senolytics discovery, medical data science, Drug discovery, machine learning, artificial intelligence, Biochemistry.

KEYWORDS: Senolytics, Senescent, Machine learning, SVM, XGB, RF, MI, FS, Features Selection, RDkit descriptors, Mordred descriptors.

To every inspiring woman
To my husband and my family

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to Dr. Stavros Perantonis for agreeing to supervise me, believing in my abilities, and introducing me to Dr. Anastasia Krithara, and Dr. Athanasios Papakyriakou. I also extend my thanks to Dr. Anastasia Krithara for her overall supervision of my work and for creating a supportive and pleasant working environment. This thesis would not have been possible without the support of Dr. Athanasios Papakyriakou, who guided me through every step and provided daily supervision, even on weekends. Special thanks to Dr. Evangelia Chrysina for agreeing to participate in the examination committee of my thesis.

Furthermore, I would like to take this opportunity to express my gratitude to all professors and researchers of the postgraduate program “Data Science and Information Technologies-Specialization: Bioinformatics-Biomedical Data Science” of the Department of Information and Telecommunications, University of Athens in collaboration with the Biomedical Research Foundation, Academy of Athens and Athena Research Center, for sharing their valuable knowledge across a wide range of scientific topics. Furthermore, I want to thank Mrs. Konstantina Kanavou for her great support.

Last but certainly not least, I want to express my heartfelt thanks to my husband Omran for his different kinds of support, and the continually helping to overcome the language barrier, also for my family and my husband’s family for unconditional support. For my friends Thalassini, Daphne, and for my colleagues of the postgraduate program “Data Science and Information Technologies-Specialization: Bioinformatics-Biomedical Data Science”, finally for my childhood friends Raghad and Ali.

Contents

LIST OF FIGURES:

LIST OF TABLES:

Preface

This thesis was conducted at National and Kapodistrian University of Athens with collaboration of NCSR Demokritos, an environment known for fostering interdisciplinary research. Our work focuses on the discovery of senolytics—compounds that selectively target and eliminate senescent cells, which contribute to age-related diseases.

Based on recent study “Discovery of senolytic using machine learning” and supported by NCSR Demokritos, we employed machine learning techniques to identify new senolytic compounds. This approach not only accelerates drug discovery but also reduces associated costs.

The research resulted in improvements in the model's performance according to accuracy metrics, highlighting the potential for novel therapies to improve aging and health outcomes. This thesis is dedicated to my mentors, colleagues, and family, whose support has been instrumental in this endeavor. I hope it paves the way for future advancements in senolytic research.

1.INTRODUCTION:

“Cellular senescence is a permanent cell cycle arrest characterized by macromolecular damage and metabolic alteration, triggered by various stressors such as replicative exhaustion, oncogenic activation chemotherapy and radiation” (2). It exerts both advantageous and detrimental effects on the tissue microenvironment, aiding processes such as potent tumor suppression mechanism. Beyond their involvement in cancer and aging, the senescent program has been associated with adverse effects in conditions such as osteoporosis, osteoarthritis, pulmonary fibrosis, SARS-CoV-2 infection, hepatic steatosis, and neurodegeneration. Due to that discovering novel senolytics has gained more interest especially therapeutic agents designed to selectively target senescent cells for elimination. Senolytics have demonstrated significant promise in alleviating symptoms across various conditions in mice. (2) (3) However, the removal of senescent cells has also been linked to adverse effects, as it interferes with their beneficial roles in processes like wound healing and liver function. Despite encouraging results only, a limited number of compounds with proven senolytic action are known, with dasatinib and quercetin in combination therapy being the only two compounds showing efficacy in clinical trials.[14]

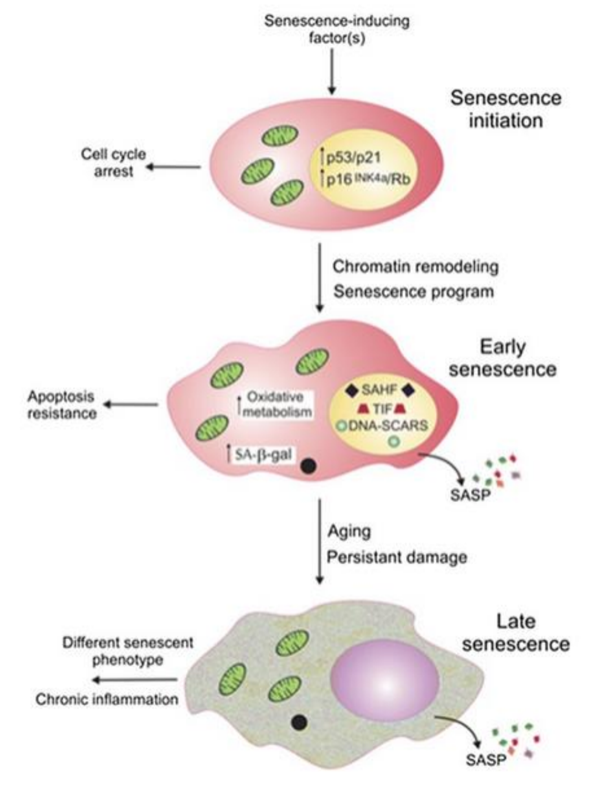


Figure 1.1 Three general phases of cellular senescence [55].

“The adoption of machine learning models, trained on molecular fingerprints and learned representation of chemical structures, has become prevalent for tasks such as bioactivity prediction” [4]. The study underscores the application of machine learning pipeline for the discovery of senolytics, leveraging a dataset compiled from diverse sources to train predictive models. The dataset comprising 58 known senolytics and a diverse set of negatives (compounds assumed to lack senolytic action) was assembled. The positive compounds were mined from literature and commercial patents, representing various chemical families such as flavonoids, cardiac glycoside, and antibiotics with senolytic action. This report focuses on comparing our methodology with that employed by the authors of "Discovery of senolytics using machine learning" [14]. While the authors utilized a pipeline involving Random Forest (RF) feature selection and XGBoost, we sought to replicate their approach and extend it by incorporating additional feature selection methods (such as Mutual Information (MI) and Fisher) and hyperparameter tuning. Furthermore, we aimed to enrich the pipeline by introducing hyperparameter tuning, a crucial step in optimizing the model's performance. This addition allowed us to explore various configurations and identify the most effective settings for our models. By expanding the methodology in this manner, we aimed to provide a comprehensive comparison of different approaches, enabling a more thorough evaluation of model performance and efficacy. Instead of solely relying on RDKit descriptors, we incorporated an additional set of descriptors known as Mordred descriptors. We then proceeded to compare the performance of these two sets of descriptors and analyzed their impact on the overall performance of the model. This approach allowed us to evaluate the effectiveness and complementary nature of each descriptor set, providing valuable insights into their respective contributions to the predictive accuracy of the model.

1.1 Motivation:

Cellular senescence, first identified in a lab setting in 1961, has become a major focus for biotech companies aiming to improve various human conditions. It is mainly characterized by cells stopping their growth permanently in response to both internal and external stresses, such as issues with telomeres, activation of cancer-causing genes, and ongoing DNA damage” (1). The lack of well characterized molecular targets has hindered the development of senolytic drugs. In the recent paper “Discovery of senolytic using machine learning” (14), the researchers

employed cost-effective machine learning algorithms trained on published data, The models were trained on senolytics and non-senolytics, the best model was used for computational screening of a chemical library, 21 compounds were tested experimentally, and 3 compounds found to have senolytic activity.
Oleandrin, in particular, exhibited improved potency compared to existing alternatives. This innovative approach significantly reduced drug screening costs and highlighted the potential of artificial intelligence in early-stage drug discovery. We aim to expand the paper's methodology and explore novel senolytic drug possibilities as part of our project's perspective improvement.

1.2 Thesis Contribution:

Several machine learning models have been evaluated for the purpose of finding new potential senolytics using 12 methods. The data used for the scope of this thesis are collected from two resources and are preprocessed in order to extract appropriate features and generate two datasets. The analysis of this project formulated using two targets positive and negative, positive refers to the compounds that are with senolytic activity and negative for the compounds that doesn't have a senolytic activity. The performance of various machine learning methods is assessed in terms of finding the model with the highest number of senolytic compounds while minimizing false positives compared to the proposed method in "Discovery of senolytic using machine learning" (14), we provide insights about the models' predictions, using explainability methods, and identify possible new senolytics compounds for testing in the lab.

In the scope of this thesis, we aim to answer the following questions:

1. Do the methods that we proposed outperform the one that presented in the recent paper (14).
2. Do Mordred descriptors enhance the model ability to predict the senolytic compounds.
3. Do the methods we proposed benefit from using other features selections.
4. What are the factors that played a role in improving the model performance.

1.3 Thesis Outline

The necessary background is presented in Chapter 2. More specifically, a brief overview of senolytics compounds is presented in Section 2.1. An overview of features selections and classical machine learning methods used in the scope of the current thesis is presented in Sections 2.2 and 2.3 respectively, and a summary of model explainability techniques is presented in Section 2.4. Recent and related works are presented in Section 2.5. Chapter 3 presents the methodology followed, including the problem statement in Section 3.1, an overview of the data used in Section 3.2, the data preprocessing and datasets generation in Section 3.3, model implementations for the different approaches in Section 3.4 as well as the overview of the evaluation metrics used for the comparison in Section 3.5. The comparison of the different methods is presented in Chapter 4, providing information regarding the experimental setup in Section 4.1 and the results in Section 4.2. A discussion over the obtained results is presented in Section 4.3. Finally, in Chapter 5, the conclusions of the current work and some perspectives on future directions for senolytics discovery are represented.

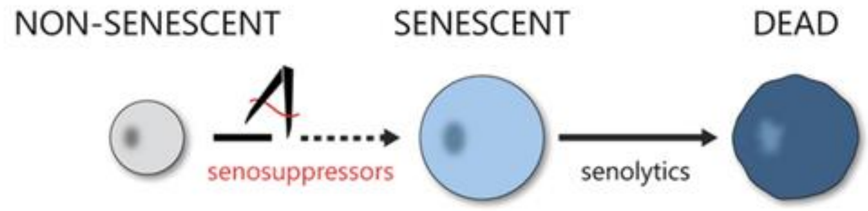
2. BACKGROUND AND RELATED WORK:

2.1 Senolytics compounds:

Senolytics are a category of drugs that reduce the impact of cellular senescence (SC), an effect associated with a range of chronic and age-related diseases (33).

The most harmful senescent cells (SC) resist apoptosis due to the upregulation of anti-apoptotic pathways, which protect them from their own inflammatory senescence-associated secretory phenotype (SASP). This allows them to survive while destroying nearby cells. Senolytics temporarily disable these anti-apoptotic pathways, inducing apoptosis in SC that exhibit a tissue-destructive SASP (34), by inhibiting or activating proteins that control resistance to apoptosis. The goal is to develop senolytics that can effectively and selectively eliminate senescent cells to treat age-related diseases while minimizing side effects.

Senolytics exert their effects through various molecular mechanisms aimed at disrupting the pro-survival pathways and enhancing cellular stress in senescent cells. They inhibit protective factors such as ephrins and PI3K with dasatinib and quercetin, respectively, and interfere with Bcl-2 family proteins by targeting upstream regulators like TrkB and direct inhibitors such as navitoclax. These agents also destabilize proteostasis and enhance proteotoxic stress by inhibiting HSP90 and other chaperones. Additionally, senolytics increase oxidative stress by exacerbating ROS production, often targeting mitochondrial function, as seen with procyanidin C1 and piperlongumine. Further mechanisms include inducing metabolic imbalances, altering lipid profiles, and disrupting intracellular pH regulation. Agents like cardiac glycosides exploit electrolyte imbalances, while others, such as SYK inhibitors, affect multiple signaling pathways to selectively induce apoptosis in senescent cells. The efficacy of senolytics and their specificity vary significantly across different cell types and senescence triggers. For instance, dasatinib is effective in killing senescent preadipocytes but not HUVECs, while quercetin shows the opposite pattern. Additionally, quercetin, effective in HUVECs, does not affect adult primary endothelial cells, indicating variability even within endothelial cells. Navitoclax, effective in various senescent cells, fails to selectively kill human primary preadipocytes. Similarly, fisetin's senolytic effect is limited to a narrow concentration range and is ineffective in some cell types. Cardiac glycosides like digoxin show cell-type-specific efficacy, influenced by mechanisms such as potassium ion regulation and autophagy inhibition, which vary with the senescence trigger and cell environment. This heterogeneity underscores the challenge of developing universal senolytics and suggests that combining senolytics or using sensitizers might extend their efficacy across different senescent cell types [35].



(56) Figure 2.1 Senolytics impact on the Sene 1

2.2 An overview of features selections:

Feature selection involves algorithms designed to reduce the dimensionality of data to enhance machine learning performance. Given a dataset with N features and M dimensions (or features, attributes), the goal of feature selection is to decrease M to M' such that $M' <= M$. This technique is a crucial and commonly employed method for dimensionality reduction. Feature selection offers several advantages: it enhances predictive accuracy, improves comprehensibility, increases learning efficiency, produces compact models, and facilitates effective data collection. By removing irrelevant and redundant features, it retains only the relevant ones, optimizing machine learning performance and efficiency. This technique is particularly useful in applications such as document categorization, medical diagnosis and prognosis, and gene-expression profiling (36).

“The feature selection methods that are routinely used in classification can be split into three methodological categories (Guyon et al., 2008; Bolón-Canedo et al., 2013): 1) filters; 2) wrappers; and 3) embedded methods. These methods differ in terms of 1) the feature selection aspect being separate or integrated as a part of the learning algorithm; 2) evaluation metrics; 3) computational complexities; 4) the potential to detect redundancies and interactions between features” (37). filters, which are independent of classification methods but lack sensitivity; embedded algorithms, which optimize attributes during classifier training and solve the minimal optimal problem; and wrappers, which provide deeper insights by integrating classification methods and are suited for all relevant problem (41).

Filter methods rank features based on their correlation with the class using statistical tests, selecting features above a threshold for input to classifiers. They are independent of the classifier, reducing overfitting but not considering classifier interaction, resulting in more general but less predictive models. Filter methods are computationally efficient and can handle high-dimensional data (42).

Wrapper methods, in contrast, use classifier performance to select feature subsets, accounting for feature dependencies and interactions, thus offering better predictive performance but at a higher computational cost. They perform heuristic searches to generate and evaluate feature subsets, choosing the one with the best classifier performance. Although they produce optimal subsets for specific classifiers and help avoid the need to set selection thresholds, they risk overfitting and lack generalizability to other classifiers. Overall, filter methods are preferred for their efficiency with high-dimensional data, while wrapper methods are favored for their superior predictive accuracy in specific classifier context (42). In this research we focus on five algorithms of features selection Random Forest, Fisher score, Mutual information, Principal component analysis, and t-distributed Stochastic Neighbor Embedding.

2.2.1 *Random Forest (RF):* are ensembles of tree predictors where each tree is built from a randomly sampled vector, ensuring independence and identical distribution across all trees. As the number of trees increases, the generalization error stabilizes. This error is influenced by the strength of the individual trees and their correlation. Randomly selecting features for node splits results in error rates that are competitive with Adaboost but more robust to noise. Internal estimates are used to monitor error, strength, correlation, and variable importance (52). RF is an evolution of Bagging which aims to reduce the variance of a statistical model, simulates the variability of data through the random extraction of bootstrap samples from a single training set and aggregates predictions on a new record (52) (51), with their ensemble approach and minimal tuning requirements, are promising for detecting weak and redundant attributes in all relevant feature selection. Challenges in discerning true relevance from random fluctuations can be mitigated using artificial contrast variables (41). This is one of the most widely used algorithms in ML, regardless of the type of problem to be solved and, although it is not possible to identify a model as the best for any type of problem, RF is undoubtedly one of the best in terms of performance, speed and generalizability (49)

2.2.2 *Mutual information (MI):* is a general measure of the relatedness between two random variables and has been actively used in the analysis of biomedical data. The mutual information between two discrete variables is conventionally calculated by their joint probabilities estimated from the frequency of observed samples in each combination of variable categories (5). In the context of feature selection, it can assess how much information about the target variable is contained in each feature (6). It is particularly useful for identifying associations where the interaction shape is unknown or varied [38], it does not assume a specific type of relationship between the variables. Unlike methods that require a predefined model of interaction (e.g., linear regression assumes a linear relationship), MI only relies on the entropies of the individual variables and their joint entropy. This flexibility allows MI to effectively detect and quantify the strength of associations regardless of their form, making it ideal for complex datasets where the nature of the interactions may be difficult to model or highly variable, [50] this making it ideal for the usage of features selection between the features of RDKit and Mordred descriptors.

2.2.3 *Fisher score (FS):* commonly known as Fisher's Score is a filter-based supervised feature selection method with feature weights" (17). This approach ranks features according to their ability to detect the labels of the dataset. FS provides a feature evaluation criterion and has been widely used (39). Since FS requires specifying the number of features by setting a threshold, we run the algorithm using a range of threshold values. This results in a variety of selected feature counts used in the method. The reason behind choosing FS is due to its stability in determining the ranks of the features.

2.2.4 *t-distributed Stochastic Neighbor Embedding (t-SNE):* is an unsupervised non-linear dimensionality reduction technique for data exploration and visualizing high-dimensional data". Parametric t-SNE aims to model the local structure of the data appropriately in the latent space, and it attempts to create separation between the natural clusters in the data (15).

t-SNE focus on preserve small distances by centering the gaussian distribution curve over a studied point then it measures the density of every other point in the high dimensional space under the gaussian distribution curve, then find the similarities between points, if two points are close to each other the similarity function gives 1, otherwise it gives 0. Student-t distribution as the heavy-tailed distribution to measure the pairwise similarities in the latent space. The weights of the parametric t-SNE network are now learned in such a way that the Kullback-Leibler divergence betheen the joint probability distributions P and Q is minimized. (40)

Kullback: is the distance matrix measures between the two-dimensional spaces, which is a very important parametric to determine the perplexity.

$$C = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

There are five parameters that control the optimization of t-SNE and therefore possibly the quality of the resulting embedding:

- Perplexity
- early exaggeration factor
- earning rate
- maximum number of iterations (40)

2.2.5 *Principal component analysis PCA:* is a mathematical technique that transforms data points from their original dimensions into new vectors, known as components. Each component is designed to capture as much variance from the original data as possible, with minimal information loss. PCA is a linear method that is most effective when the data exhibits a linear structure. It identifies the principal components by projecting the data onto lower dimensions, aiming to maximize variance while preserving significant pairwise distances. To determine the number of dimensions needed, we set the number of components to capture 99% of the variance in the dataset. This resulted in 111 dimensions. This indicates that the variance is distributed across many variables, requiring 111 principal components to account for 99% of the total variance in the dataset. (9)

2.3 Classical machine learning methods:

There is a great variety of ML methods that have been used in the pharmaceutical industry for the prediction of new molecular characteristics, biological activities, interactions and adverse effects of drugs. Some examples of these methods are Naive Bayes, Support Vector Machines, Random Forest and, more recently, Deep Neural Network (49)

Support Vector Machines (SVM): They are one of the most widely used techniques because of their superior performance and their ability to be generalized in high-dimensional domains, especially in bioinformatics (49). It minimizes the empirical classification error and maximizes the geometric margin (19). SVM finds a hyperplane in a space different from that of the input data x. It is a hyperplane in a feature space induced by a kernel K (the kernel defines a dot product in that space (Wahba, 1990) (18).

SVM widely used models in bioinformatics because of its ability to deal with complex, non-linear, high-dimensional and noisy problems (49).

Naive Bayes (NB): This model has been used in drug discovery for the prediction of possible drug targets (49). NB classifier simplifies learning by assuming that features are independent given class, despite this unrealistic assumption, the resulting classifier is remarkably successful in practice, often competing with much more sophisticated technique (26).

It is not the ideal algorithm for high dimensionality problems with a high number of attributes since it uses frequency tables to extract knowledge from the data and treats each variable as categorical and, in case of working with numerical variables, it must perform some kind of transformation (49).

Extreme Gradient Boosting (XGB): a scalable machine learning system for tree boosting, Ensembles are made up of decision tree models. These Trees are added one by one to the ensemble and trained to fix the mistakes made by the previous models. This method of ensemble learning is called boosting. To train the models, any loss function that can be differentiated is used along with a gradient descent optimization algorithm. This gives the technique its name" gradient boosting", as the loss gradient is minimized as the model is fit (20). XGBoost has two main benefits which are fast execution and superior model performance.

Logistic Regression (LR): is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X (21). LR is used to obtain odds ratio in the presence of more than one explanatory variable (22). LR is a probabilistic ML approach used for classification tasks (23), makes predictions based on probabilities obtained through maximum-likelihood estimations (24).

2.4 summary of model explainability techniques:

Nested cross validation: Cross validation is a k-fold cross-validation and partitions the variable data into k disjoint chunks of approximately equal size. In each iteration a training set is formed from different combination of k-1 chunks, with the remaining chunk used as the test set; a model is then fitted to the training set and its performance evaluated using the test set. The average of the performance metric on the validation set in each iteration is then used as an estimate of the generalization performance of a model fitted to all of the available data. There are two common procedures for selecting the best algorithm and tuning the hyperparameters via cross-validation (7). The reason behind using the nesting cross validation in comparison between the models in splitting a dataset for training and validation using scikit-learn, a common concern arises regarding the potential bias introduced by the random selection of the validation set. This randomness could lead to scenarios where the validation set is either too easy or too difficult to predict accurately. While traditional cross-validation addresses this by allowing multiple experiments with different validation sets, it becomes challenging when comparing the performances of various classifiers. This challenge is particularly evident when considering that a model's performance can fluctuate due to its configuration, notably its hyperparameters. Here's where nested cross-validation becomes indispensable. Nested cross-validation involves a loop of cross-validation processes, offering a more robust approach to comparing the performances of different models. By leveraging nested cross-validation, we can estimate a model's performance accurately while simultaneously optimizing its hyperparameters. This methodology is crucial, especially when dealing with datasets characterized by significant imbalances. In such cases, models tend to prioritize detecting the most frequent targets, potentially leading to overly optimistic performance evaluations. Nested cross-validation accounts for these complexities, providing a more reliable assessment of model performance. In our experiments, assessing whether the model had improved solely based on the training dataset proved challenging. This challenge prompted us to construct a test set with a distribution similar to that of the training set. To categorize compound labels, we relied on experimental results from recent papers and literature sources not included in our training data.

Optuna: a freely available optimization software, approaches hyperparameter optimization as the process of minimizing or maximizing an objective function, which evaluates a set of hyperparameters and returns its validation score. Optuna gradually constructs the objective function through interactions with the trial object. During the runtime of the objective function, search spaces are dynamically formed using methods of the trial object. Users are prompted to utilize the suggest API within the objective function to dynamically generate hyperparameters for each trial. When the suggested API is invoked a hyperparameter is statistically sampled based on the history of previously assessed trials (8).

CCN: is one of the neural network models adopted for drug response prediction. CNN has been actively used for image, video, text, and sound data due to its strong ability to preserve the local structure of data and learn hierarchies of features. In 2021, several methods had been developed for drug response prediction, each of which utilizes different input data for prediction (Baptista et al. 2021) (48).

2.5 Recent and related works:

Artificial intelligence (AI) has been transforming the practice of drug discovery in the past decade. Various AI techniques have been used in many drug discovery applications, such as virtual screening and drug design (45) (46) (14). Recently various factors were developed due to greater enthusiasm for utilizing machine learning approaches in the pharmaceutical industry (47).

(Smer-Barreto et al.,2023) (14) proposed a machine learning pipeline for discovering senolytics, their pipeline included features selection using RF followed by manual comparison between SVM, XGB and RF, their results showed that XGB was the winner model for predicting the senolytics and the optimization of XGB has been done by optimizing the max-depth hyperparameter. The dataset was compiled from multiple sources, including academic publications and a commercial patent, to train machine learning models that predict senolytic activity. Utilizing this dataset, a library of over 4000 compounds was computationally screened, narrowing down to 21 candidate hits for further experimental validation. Experimental screenings in two model cell lines of oncogene- and therapy-induced senescence identified three compounds ginkgetin, oleandrin, and periplocin as having senolytic activity with potencies and dose-responses comparable to established senolytics. Notably, oleandrin exhibited greater potency and activity on its target, Na+/K+ ATPase, and its senolytic effector NOXA, outperforming known cardiac glycosides with senolytic properties (14).

(Wong et al.,2023) effort utilized graph neural networks to screen and predict senolytic activity from a vast chemical space of over 800,000 compounds. This approach achieved a positive predictive value of 11.6%, identifying structurally diverse senolytic compounds with favorable medicinal properties. Further validation confirmed the effectiveness of selected compounds, with some showing promising results in reducing senescent cell burden in aged mice. These findings highlight the significant advancements in using deep learning techniques for discovering effective senotherapeutics.

3. Methodology:

3.1 Problem statement:

As previously discussed in Section 1, the primary objective of our project is to compare the performance of the method described in the work of (Smer-Barreto et al.,2023) (14), which consists of three steps: feature selection using RF, classification using XGB, and hyperparameter tuning for the max_depth parameter, with our novel approaches aimed at improving the prediction of new senolytics.

Tasks:

1. Feature Selection Algorithms: Implement various feature selection algorithms (Mutual Information (MI), Fisher Score, Random Forest (RF)).
2. Hyperparameter Tuning and Model Comparison: Perform hyperparameter tuning and comparison across five models (SVM, XGB, NB, RF, LR).
3. Performance Evaluation: Evaluate performance using a test set.
4. Descriptor Comparison: Use Mordred descriptors in addition to RDKit descriptors and compare their performances.
5. Simple neural network.

Table 3.1 Explanatory tables of the methods

Method	Feature Selection	Hyperparameter Tuning	Classifiers	Descriptors
RF_XGB_RDKit	Random Forest	max_depth	XGBoost	RDKit
MI_XGB_RDKit	Mutual Information	max_depth	XGBoost	RDKit
FS_XGB_RDKit	Fisher Score	max_depth	XGBoost	RDKit
RF_Optuna_RDKit	Random Forest	Optuna for many hyperparameters	SVM, NB, RF, XGB, LR	RDKit
MI_Optuna_RDKit	Mutual Information	Optuna for many hyperparameters	SVM, NB, RF, XGB, LR	RDKit
FS_Optuna_RDKit	Fisher Score	Optuna for many hyperparameters	SVM, NB, RF, XGB, LR	RDKit
RF_XGB_Mord	Random Forest	max_depth	XGBoost	Mordred
MI_XGB_Mord	Mutual Information	max_depth	XGBoost	Mordred
FS_XGB_Mord	Fisher Score	max_depth	XGBoost	Mordred
RF_Optuna_Mord	Random Forest	Optuna for many hyperparameters	SVM, NB, RF, XGB, LR	Mordred
MI_Optuna_Mord	Mutual Information	Optuna for many hyperparameters	SVM, NB, RF, XGB, LR	Mordred
FS_Optuna_Mord	Fisher Score	Optuna for many hyperparameters	SVM, NB, RF, XGB, LR	Mordred

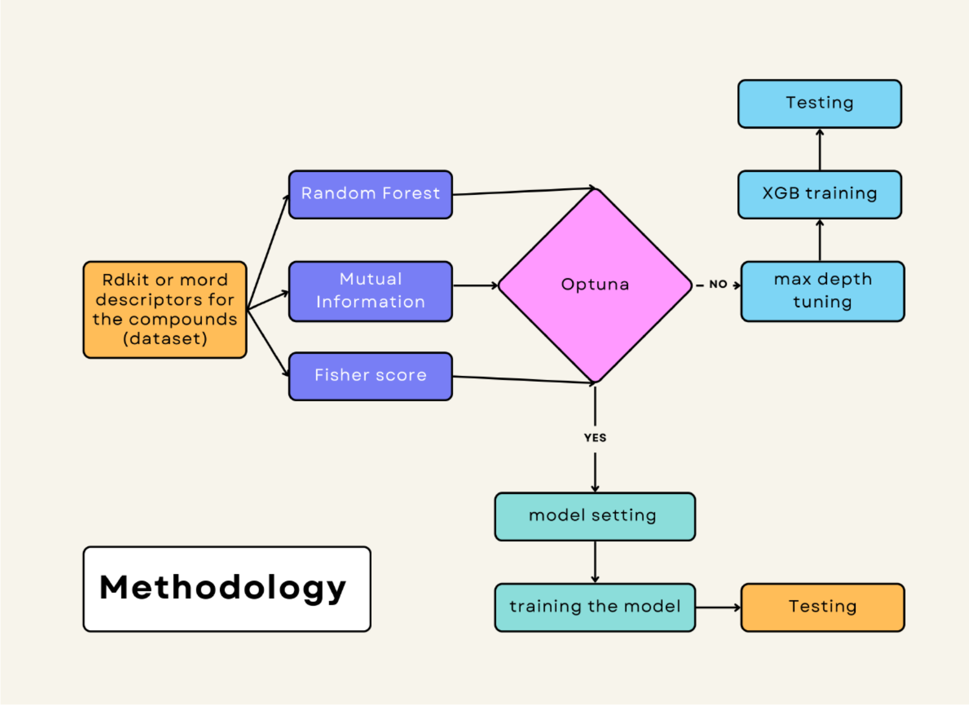


Figure 3.1 provides a comprehensive view of the methodology employed in our study. It serves as a roadmap for the steps taken, illustrating the interconnectedness of various processes involved in the feature selection, model optimization, and testing phases of our research

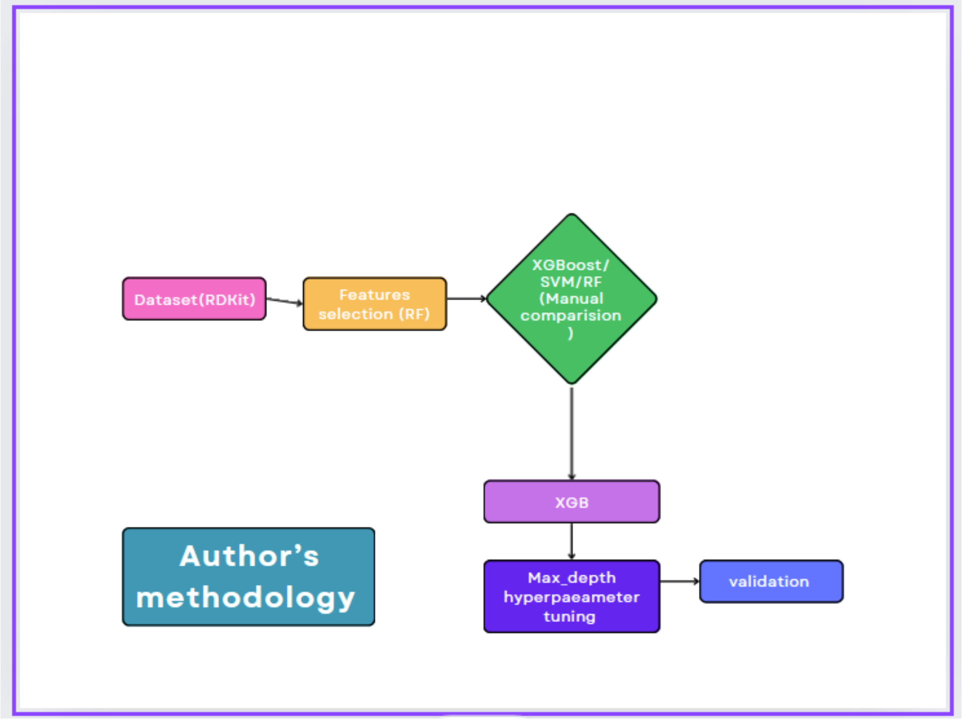


Figure 3.2 provides a comprehensive view of the methodology employed in the author's work *(14)*.

The project integrates these tasks to create twelve unique methods as follows:

- RF_XGB_RDKit: This method follows the original approach used by the authors, utilizing Random Forest for feature selection, XGB for classification, and RDKit descriptors.
- MI_XGB_RDKit and FS_XGB_RDKit: Similar to RF_XGB_RDKit but use Mutual Information and Fisher Score for feature selection, respectively.
- RF_Optuna_RDKit: Utilizes Random Forest for feature selection, Optuna for hyperparameter tuning with cross-validation, followed by model comparison through cross-validation. The best-performing model is trained and evaluated using the test set.
- MI_Optuna_RDKit and FS_Optuna_RDKit: Similar to RF_Optuna_RDKit but use Mutual Information and Fisher Score for feature selection, respectively. Additionally, methods using Mordred descriptors instead of RDKit descriptors were also developed.
- The method RF_XGB_RDKit is the method that the authors used in their work. MI_XGB_RDKit, FS_XGB_RDKit is the same method of the RF_XGB_RDKit the only difference is using different features selection method.
- RF_Optuna_RDKit in this method we use RF as features selection we use optuna with cross validation for hyperparameter tuning and another cross validation for models comparing. After finding the best model performance we train it and evaluate the performance using the test set.
- MI_Optuna _RDKit and FS_Optuna_RDKit are the same as RF_Optuna_RDKit the only difference the name of features selection we used. The rest of the method is similar to what we explained previously the one difference that we used Mordred descriptors instead of RDKit descriptors.

3.2 Overview of the data:

The datasets utilized in this project were based on two recent publications with a shared focus. The first source, previously mentioned in our work, will be referred to as ML (Smer-Barreto et al.,2023) (14). The second source, employing deep learning methods, will be referred to as DDN (Wong et al.,2023) (13).

A molecular descriptor MDs is defined as the “final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment” (30) (31)

MDs have become a very useful tool to carry out the search for similarities in molecular repositories, since they can find molecules with similar physicochemical properties according to their similarity to the values of the calculated descriptors (49).

Mordred descriptors: Mordred was developed to be a user-friendly software program with straightforward installation, extensive support for molecular descriptors, high-speed calculations, and built-in automated testing (32)

RDKit Descriptors: are numerical values that capture various chemical properties of a molecule. These descriptors are like fingerprints that represent the molecule in a way that can be used by computers to compare and analyze different molecules. There are many different types of RDKit descriptors, each encoding a specific aspect of a molecule's structure or properties. Some common examples include Molecular weight: The total mass of the molecule. Log P: A measure of a molecule's hydrophobicity (how well it dissolves in water). Hydrogen bond donors and acceptors: The number of atoms in a molecule that can form hydrogen bonds with other molecules. Fingerprint descriptors: These are binary vectors that encode the presence or absence of certain substructures in a molecule. RDKit descriptors are widely used in various cheminformatics applications, including Virtual screening: Identifying molecules that are likely to bind to a specific biological target. Quantitative structure-activity relationship (QSAR) modeling: Predicting the biological activity of a molecule based on its structure. Clustering and classification of molecules: Grouping molecules with similar properties together. By using RDKit descriptors, researchers can efficiently analyze large datasets of molecules and gain insights into their chemical properties and biological activities (53).

Our training data primarily originates from ML (14), but we have made modifications based on supplementary data from DDN (13), specifically the 'Data 10 μM' file (13) to enhance the reliability of our dataset and address discrepancies noted between the two datasets.

Certain compounds labeled as nonsenolytics in ML were tested in the lab by DDN and confirmed to be senolytics. These compounds were excluded from our training set and incorporated into our test set. Conversely, some compounds labeled as senolytics in ML were tested by DDN under specific conditions and concentrations, which suggested they were nonsenolytics. However, due to potential inefficiencies in the lab testing conditions, we retained the original ML labels, as these compounds have high potential to be senolytics according to the literature.

The test set was constructed by intersecting 'Data 10 μM' with 18 verified nonsenolytics (negatives) and including 3 positives from ML, 3 positives from DDN, and 45 positives from DDN's 'Data 10 μM' after removing any positives found among the verified nonsenolytics. The 3 compounds from ML and the 3 compounds from DDN represent the final results reported in their respective papers. The 45 positives from DDN's 'Data 10 μM' were gathered from supplementary files according to literature review. The 18 verified nonsenolytics were validated through lab testing as reported in the ML paper.s

Table 3.2 1 Sources and number of compounds in the training and test sets. Initial sets were processed to yield the final sets employed as described in the main text.

Compounds	**Initial Training**[a]	**Final Training**[b]	**Initial Test set**[c]	**Final Test set**[d]
Positives	58	58	45	38
Negatives	2,465	2,451	2,307	1,205
Total	2,523	2,509	2,352	1,243

[a] From the Supplementary File: "list_of_compounds_for_training.csv" of ML (14) taken from https://zenodo.org/records/7870357. Positives/Negatives as indicated in column "Senolytic".

[b] After removing 14 negatives that have been shown to exhibit senolytic activity in ref. (13) (negatives intersected with the initial DDN test set positives).

[c] from the tab "Data 10 μM" of the "Supplementary Data1" file (13) with positives highlighted.

[d] Added the 3 verified positives from ML (14), and the 3 verified positives from DDN (13) within positives. Added the remaining 18 tested non-senolytics from ML (14) within negatives. Then the test set was intersected with the corresponding positives and negatives of the training set so as to remove duplicates (13 positives and 1,120 negatives).

We utilized two types of physiochemical descriptors, which are among the most popular, namely the RDKit descriptors (53) and the Mordred descriptors (32). For our sets, 198 descriptors were generated by the RDKit library (2023.09.5) out of a total of 201 descriptors and 569 Mordred descriptors using Mordred library (1.2.0) out of a total number of 1826 descriptors calculated. Descriptors containing non-numeric values in at least one compound were discarded.

3.3 Data preprocessing and datasets generation:

Before the authors conducting their methodology, they checked various factors, they started checking the diversity of their training dataset, in order to achieve that, they performed clustering analysis using k_mean and Silhouette coefficients, the results of their analysis indicated a lack of clear clustering and low similarity among the senolytic compounds, which supports the dataset's diversity. The authors were interested in checking the diversity of the training dataset because Diverse dataset helps to capture this variability, ensuring that the developed models are more representative of the broader population, and avoiding bias and generalization issues, at the same time improves reliability. Furthermore, the Tanimoto distance graph analysis demonstrates the most senolytics are dissimilar in the chemical descriptor space.

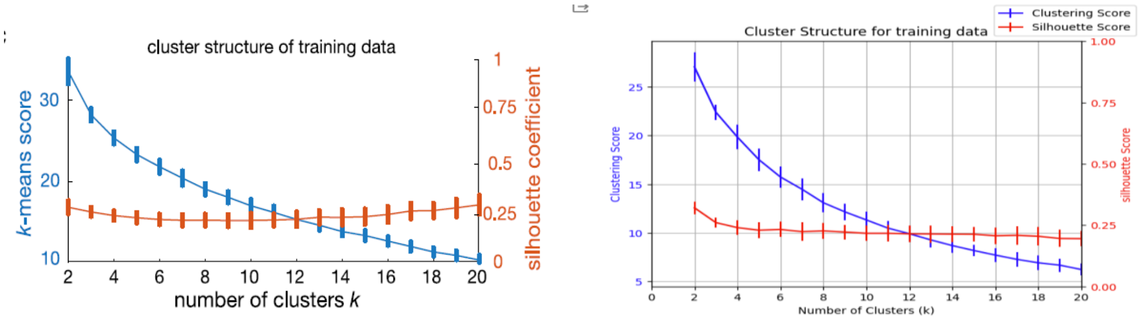


Figure 3.2 (a) Cluster structure of the senolytics employed for training using the RDKit descriptors as features. Plot shows the k-means clustering score and silhouette coefficient 58 averaged across compounds for an increasing number of clusters (k). Error bars denote one standard deviation over 100 repeats with different initial seeds. The lack of a clear "elbow" in the k-means score and low silhouette coefficients suggest poor clustering among the senolytics employed for training [Figure taken from ref. (14)].

Figure 3.2 (b) replicated the results reported by the authors by employing RDKit descriptors as features and conducting clustering on the structure of the senolytics used for training. Our analysis successfully reproduced the plot presented in the paper. However, upon closer examination, we discovered that the authors claimed to have used the standard scaler for normalization, whereas our findings suggest that they actually utilized the min-max scale

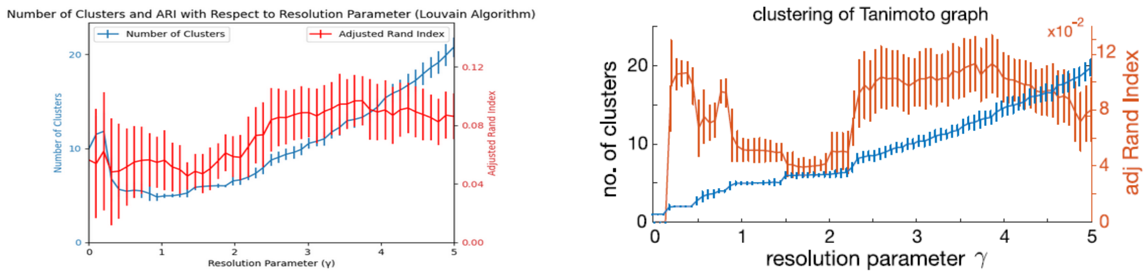


Figure 3.3 (a) authors result of clustering of the Tanimoto distance graph using the Louvain algorithm for community detection 60. Plot shows the average number of clusters with respect to the resolution parameter (γ) across 100 runs (error bars denote one standard deviation); increasing values of γ produce a larger number of clusters. We observe pronounced plateaus at 5 and 6 clusters, suggesting some degree of clustering in the data. We computed the

adjusted Rand index61 (ARI) averaged across all compounds to quantify the similarity between cluster labels and compound source labels. Low ARI values indicate that Louvain clusters are different from the literature source labels [Figure taken from ref. (14)]

Figure 3.3 (b) replicated results of clustering of the Tanimoto distance graph using the Louvain algorithm for community detection 60. increasing values of γ produce a larger number of clusters. We observe pronounced plateaus at 5 and 6 clusters, suggesting some degree of clustering in the data. We managed to produce similar results with the authors, but there are some differences, but the conclusion remains the same.

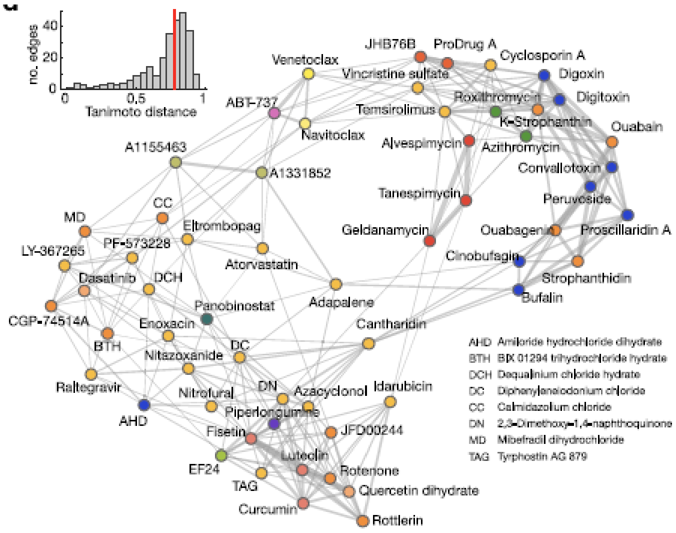
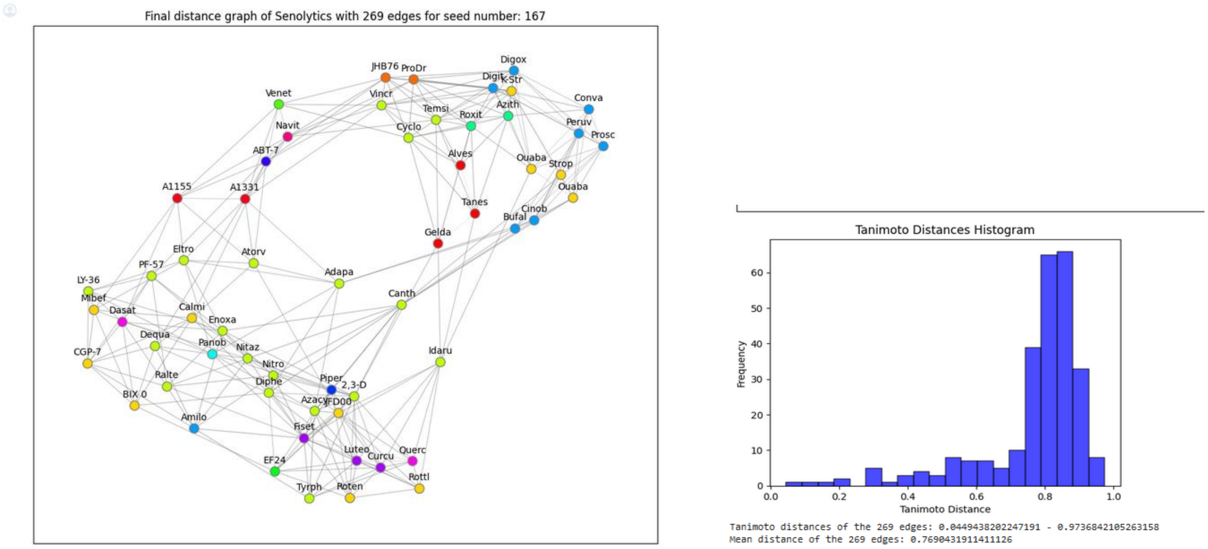


Figure 3.4 (a) Tanimoto distance graph of senolytics employed for training: nodes are compounds and edges represent compounds that are sufficiently close in the physicochemical feature space. Node color indicates the data source as in panel. To emphasize the overall dissimilarity between compounds, the authors set the edge thickness as the Tanimoto similarity (1-distance). Inset shows the distribution of Tanimoto distances across the 269 graph edges (median distance of 0.77) [Figure taken from ref. (14)].



```
Tanimoto distances of the 269 edges: 0.0449438202247191 - 0.9736842105263158
Mean distance of the 269 edges: 0.7690431911411126
```

Figure 3.4: (b) Tanimoto distance graph of senolytics employed for training; nodes are compounds and edges represent compounds that are sufficiently close in the physicochemical feature space. Node color corresponds to the data source. Our methodology involved initially computing the Tanimoto distance matrix of the structures, followed by

plotting the histogram of the distances. Subsequently, we identified k-nearest neighbors (k=7) based on cosine distances of scaled features. We then performed the intersection between the k-nearest neighbors' graph and the Tanimoto distances graph, with weights assigned based on Tanimoto distance. The distribution of the graph varies across trials. Through multiple iterations, we identified that the most consistent results, matching those presented in the paper, were achieved using a seed value of 168. This specific configuration resulted in 269 edges, with a mean Tanimoto distance of 0.7691

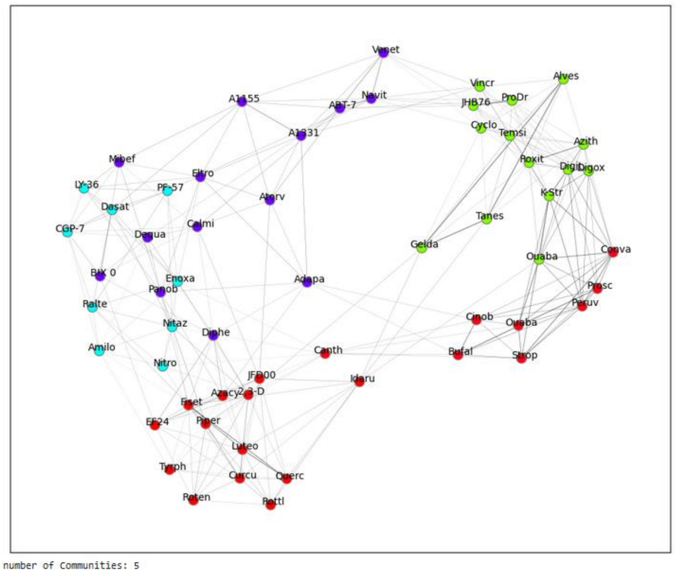


Figure 3.4 (c) The distance graph after the partition using louvain community and setting the resolution variable to 1.5 to visualize the result of this partition.

We used the Louvain algorithm to group compounds based on their similarities in the Tanimoto distance graph. The plot shows the average number of clusters across 100 runs, with error bars indicating variability. When we increase a parameter called γ, we get more clusters. Notably, the plot shows plateaus at 5 and 6 clusters, suggesting clear grouping in the data. the ARI scores showed pronounced troughs at the plateaus detected with the Louvain method (mean ARI < 0.05 for 100 runs of the clustering method), which we regarded as sufficient evidence that compounds do not cluster according to the source from which they were obtained. To check how well our clusters match the expected labels, we calculated the adjusted Rand index (ARI). Low ARI values mean our clusters differ a lot from the expected labels." the ARI scores showed pronounced troughs at the plateaus detected with the Louvain method (mean ARI < 0.05 for 100 runs of the clustering method), which we regarded as sufficient evidence that compounds do not cluster according to the source from which they were obtained".

Our plot differs from the paper due to various reasons like library versions and processing details. Despite these differences, our overall conclusion remains the same.

The author's Methodology consists of training machine learning models using the compiled dataset to computationally screen chemical libraries and identify potential candidates for experimental validation. To achieve this, the authors initiated a feature selection process on the complete dataset before any cross-validation or train -test split. Utilizing a random forest model and the average reduction of Gini index to measure impurity, they identified a reduced set of 165 normalized features. The authors claimed that the inherent high dimensionality of the training data was reflected in the fact that more than 100 dimensions were needed to accurately explain the data's variability (111 features for 99% explained variance) (14)

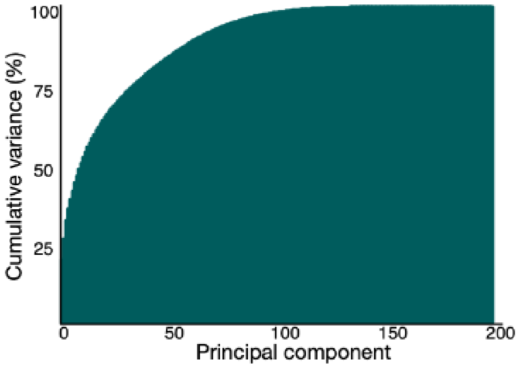
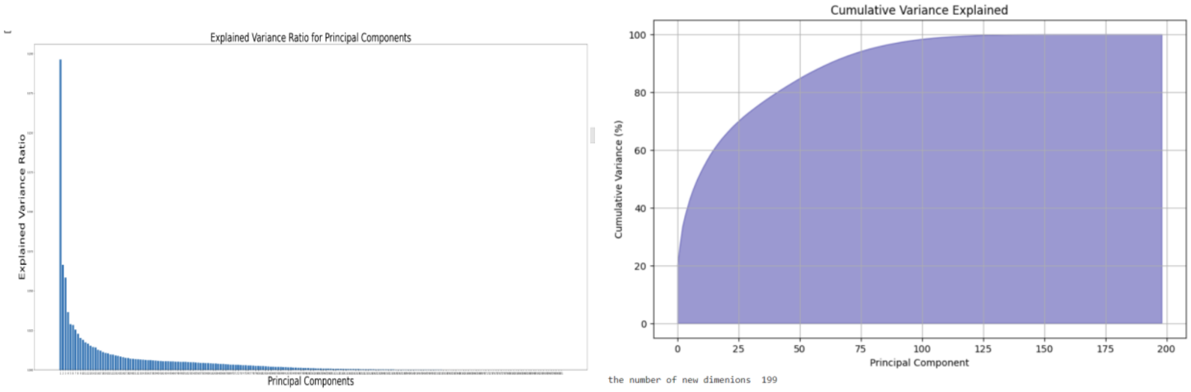


Figure 3.5 (a) explained variance plot displays the cumulative percentage of total variance explained by each successive principal component in a dataset, The plot helps to determine the number of principal components that should be retained by showing how much variance is captured by each component [Figure taken from ref. (14)]



```
the number of new dimenions  199
```

Figure 3.5 (b) The replication of the experiment showed that the dataset's variability can be accounted for by 99% using only 111 out of total 200 features.

.

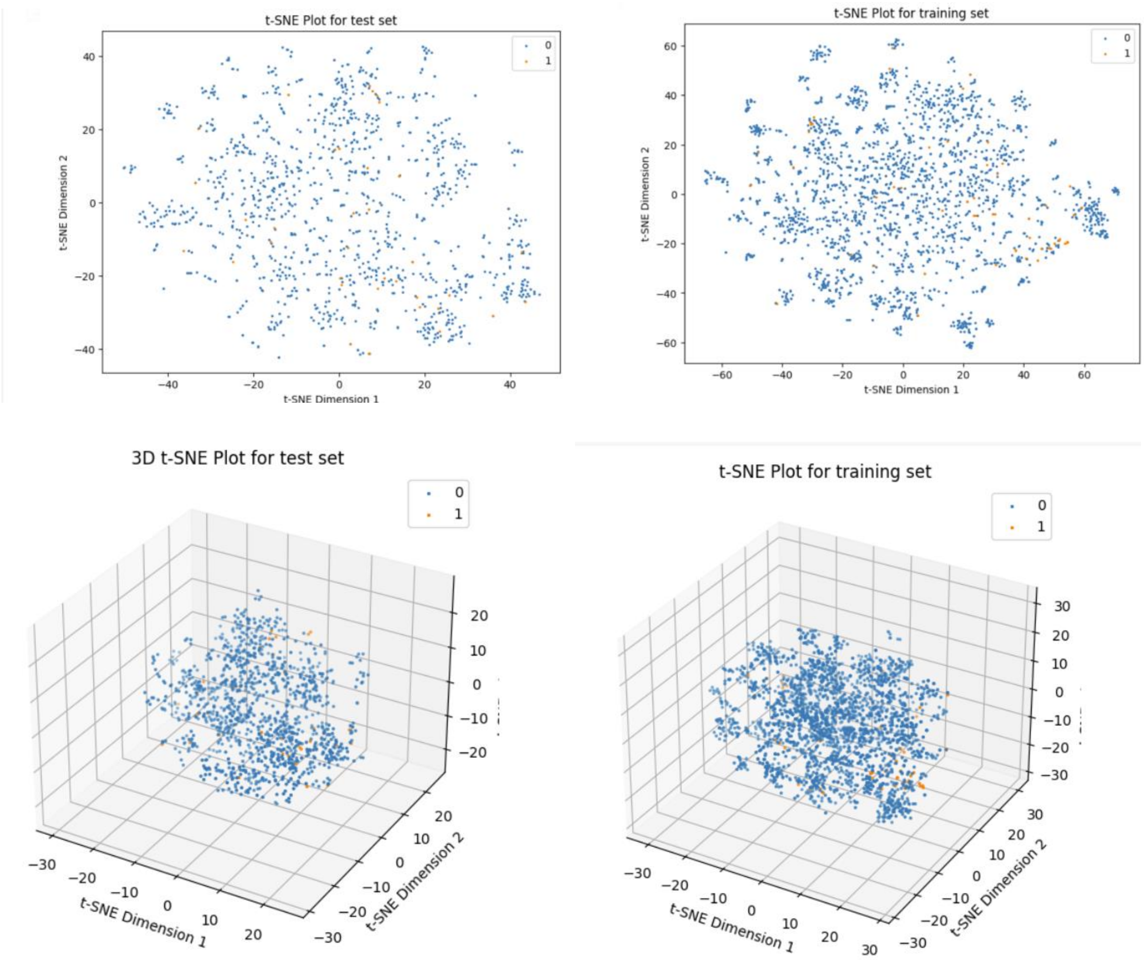
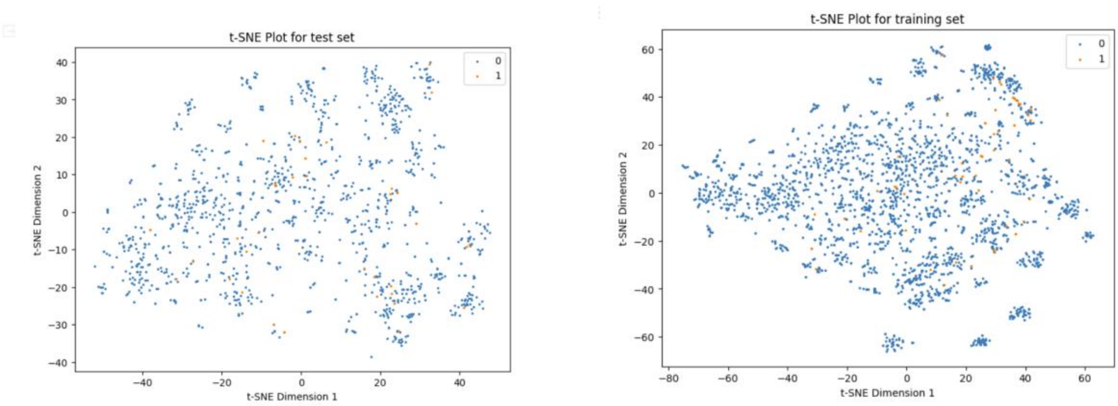


Figure 3.6: (a) The t-SNE plot, generated using RDKIt descriptors, illustrates the sparsity of compounds in the chemical space for both the test and training sets. Additionally, it provides a visual representation of the distribution of senolytics in two and three dimensions.



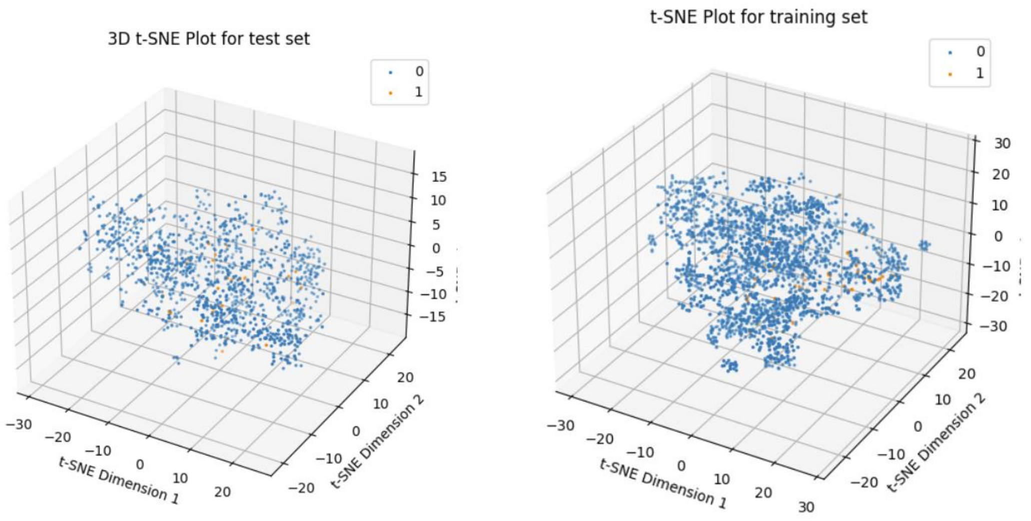
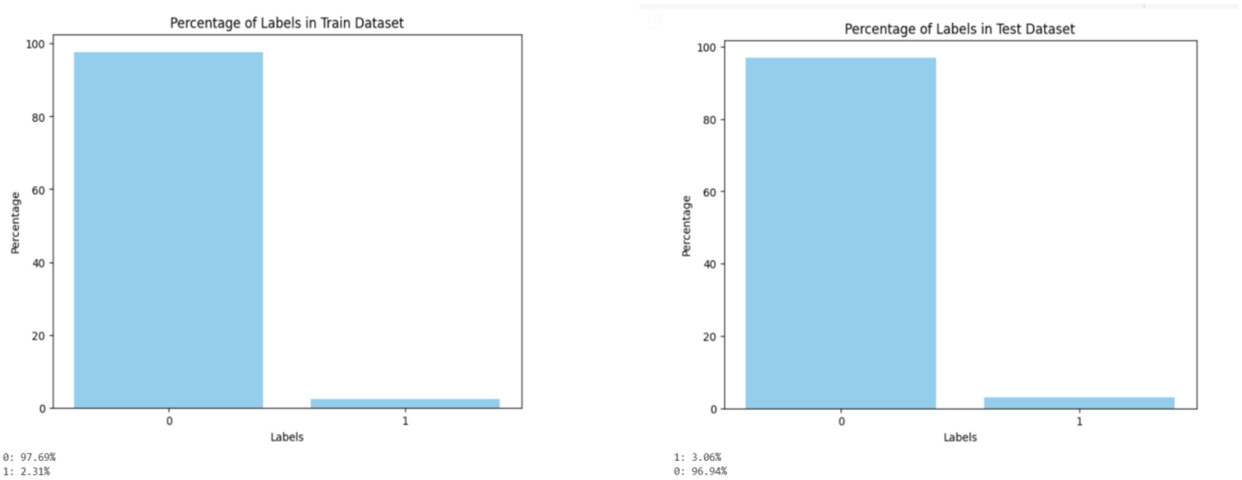


Figure 3.6 (b) The t-SNE plot, generated using Mordred descriptors, illustrates the sparsity of compounds in the chemical space for both the test and training sets. Additionally, it provides a visual representation of the distribution of senolytics in two and three dimensions.



```
0: 97.69%
1: 2.31%
```

```
1: 3.06%
0: 96.94%
```

Figure 3.7 plot shows the distribution of labels in the training and test set, we noticed the two datasets share a similar distribution.

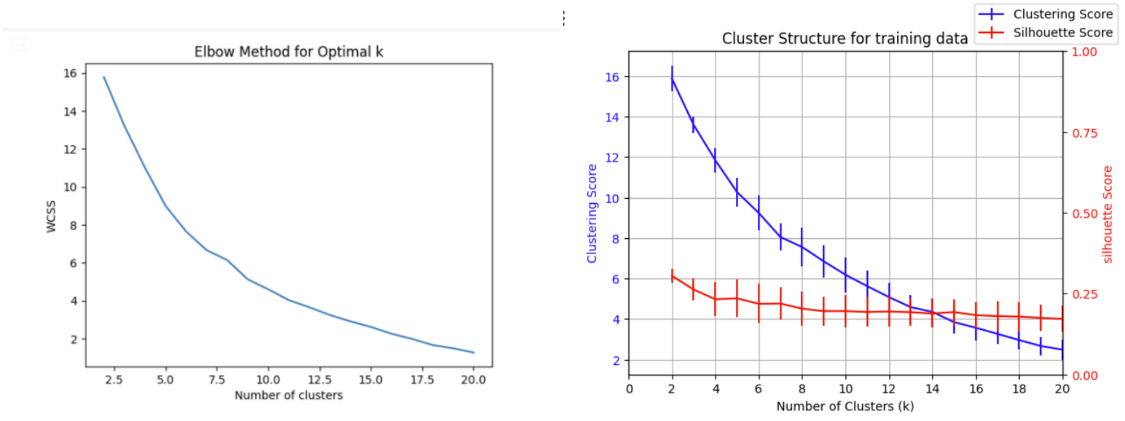


Figure 3.8 Cluster structure of the senolytics employed for testing using the Mordred descriptors. The plot displays the average k-means clustering score and silhouette coefficient across compounds as the number of clusters increases. The absence of a distinct "elbow" in the k-means score and the low silhouette coefficients indicate weak clustering among the senolytics tested.

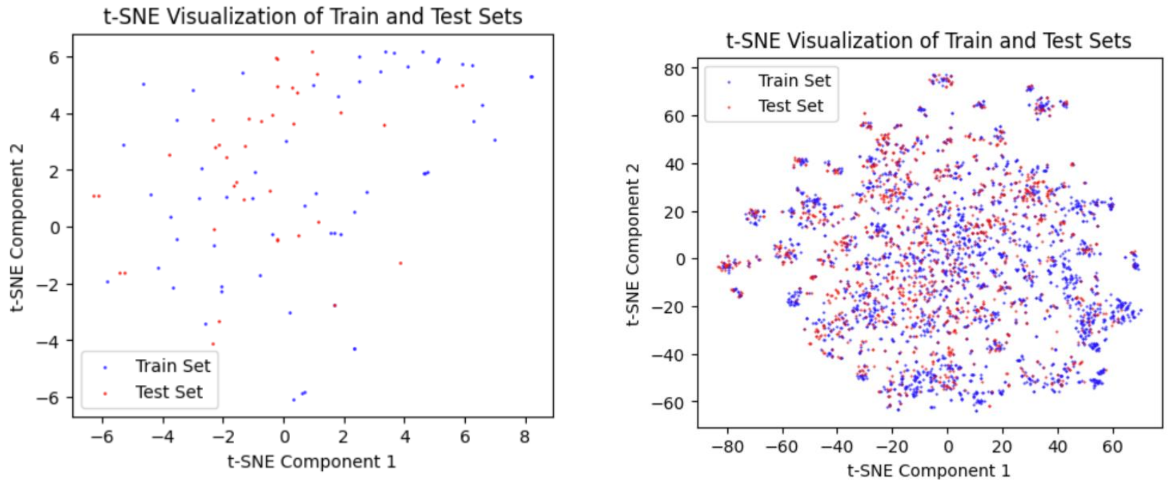


Figure 3.9 t-SNE visualization for the training and test sets using the Rdkit descriptors in the left is the 2-dimenstions visualization for the senolytics compounds for both test and train sets, on the right the visualization of all whole datasets of training and test, the blue dots represent the compounds from the training set and the red dots represent the compounds from the test set.

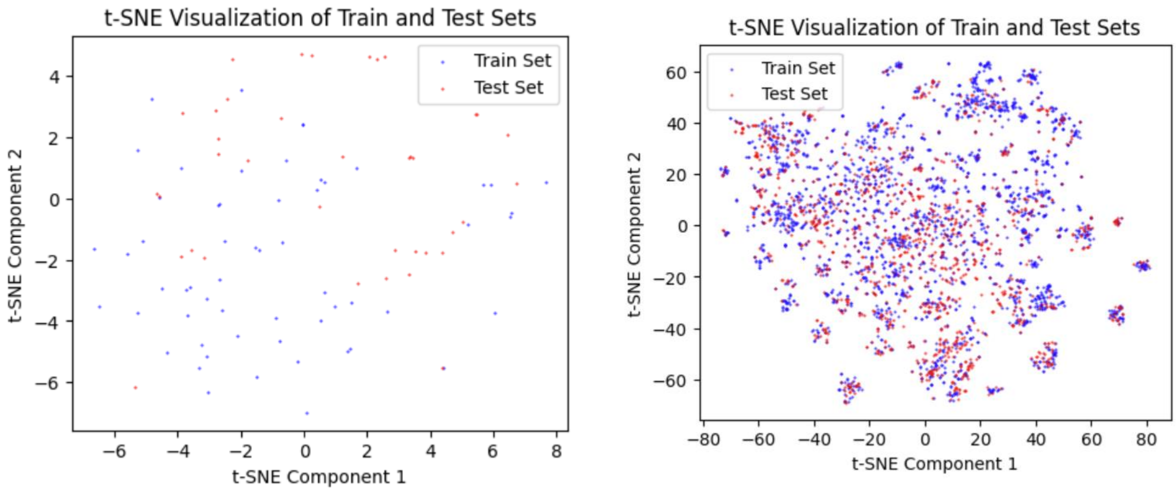


Figure 4.0 t-SNE visualization for the training and test sets using the Mordred descriptors in the left is the 2-dimenstions visualization for the senolytics compounds for both test and train sets, on the right the visualization of all whole datasets of training and test, the blue dots represent the compounds from the training set and the red dots represent the compounds from the test set.

The t-SNE visualizations indicate that the training and test sets have similar distributions, which is ideal for model training and testing. The clear overlap in the combined visualizations (for the datasets containing both negative and positive senolytics in RDKit and Mordred descriptors) suggests that the data was split in a balanced manner. The distinct clusters in the individual visualizations for senolytics show that while the training and test sets are distinguishable, they share common features, as evidenced by their overlap in the combined plots.

3.4 Model implementations:

We conducted 20 trials for all methods to address the variability in outcomes due to factors such as training data splits, model configuration after hyperparameter tuning, and the number of features selected in each trial.

Given the variability, averaging the results across trials was not effective for comparison as every trial represents a different model due to difference in model’s configuration. Instead, we ranked each trial based on precision, MCC, and recall in both validation and test sets. The best trial was identified as the one that ranked highly in both validation and test sets. This approach accounts for discrepancies where the top-ranked trial in validation might perform lower in the test set, and vice versa.

After identifying the best trial, we stabilized the model's performance by setting its configuration and trained it using the entire training set. The model's performance was then evaluated using the test set. Despite addressing two variability factors, we couldn't fix the number of features used, which is why we conducted 20 trials and averaged the test set performance. Our systematic approach of comparing different feature selection methods, hyperparameter tuning techniques, and descriptor types, followed by rigorous performance evaluation, provides a robust framework for improving senolytics prediction models.

Then we used CNN to compare its performance with the classical machine learning models, considering the computational expenses that CNN will require.

3.5 Evaluation metrics:

F1 score: harmonic average of precision and recall (28), F1 score is widely used to measure the success of a binary classifier when one class is rare (29)

$$F_1 \text{ score} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F1 score equation (27)

Recall: is the probability of detecting an item given that it is relevant (30)

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision: defined as the probability that an item is relevant given that it is detected by the algorithm (30)

$$\text{Precision} = \frac{TP}{TP + FP}$$

Matthews Correlation Coefficient (MCC): is a reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset (27).

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}}$$

(27)

Confusion Matrix: is a collection of all four types of results from a two-class prediction problem: true positives (TP), false positives (FP) also known as false discoveries or type-I errors, true negatives (TN), and false negatives (FN) also known as missed discoveries or type-II errors (54).

4. EXPERIMENTAL EVALUATION:

In this chapter, we present the results of our experimental evaluation of the algorithms under comparison. The purpose of this evaluation is to assess the effectiveness and performance of our methods relative to the baseline method described in reference paper (11). To facilitate this comparison, we designed an experimental pipeline that includes a set of well-defined tasks and metrics for performance evaluation. In Section 4.1, we describe the experimental setup pipeline, followed by a presentation of the evaluation results in Section 4.2. Here, we analyze the performance of the tested models in comparison to the baseline method. Finally, we discuss our findings, highlighting the strengths and weaknesses of the compared methods and providing insights for future research in this area.

4.1 Experimental setup:

The experimental pipeline for evaluating the model's ability to predict positive senolytics involved several key steps. After generating the final datasets as described in Chapter 3, we utilized nested cross-validation to conduct the experiments and obtain results. Nested k-fold cross-validation, which comprises two loops of cross-validation, was selected for its capability to evaluate multiple classifiers and determine the optimal configuration. Specifically, we performed nested cross-validation with 5 folds to find the best configuration of each classifier, followed by another 5-fold cross-validation for comparing the classifiers using their optimal configurations. All algorithms were trained with a range of hyperparameters, and the best values were chosen via cross-validation.

For Task 1, as outlined in Section 3.4, we determined the number of features to use, with each algorithm employing different approaches for feature selection. We compared the performance based on different feature selection techniques: for MI and RF, we used a threshold of zero, and for FS, we employed a range of thresholds. We conducted 20 trials for each method, yielding 20 results for both test and validation sets. One model per method was chosen based on its performance in both sets. Detailed results for each method are shown in Appendix A.

Deep learning:

The neural network architecture employed in this project is a convolutional neural network (CNN) tailored for binary classification tasks utilizing chemical descriptors. Beginning with convolutional layers to capture local patterns and spatial features, the network includes max-pooling layers for dimensionality reduction, followed by dense layers to extract higher-level abstractions. Specifically, it comprises two convolutional layers with 64 and 128 filters, respectively, each utilizing a ReLU activation function. Max-pooling layers with a pool size of 2 are interleaved between the convolutional layers. After flattening the output, two dense layers with 256 and 128 neurons, alongside ReLU activation and dropout regularization, are employed. The final layer, utilizing a sigmoid activation function, facilitates binary classification. Model optimization is performed using the 'adam' optimizer and 'binary_crossentropy' loss function, with accuracy, precision, recall, and MCC serving as the evaluation metrics during training and validation.

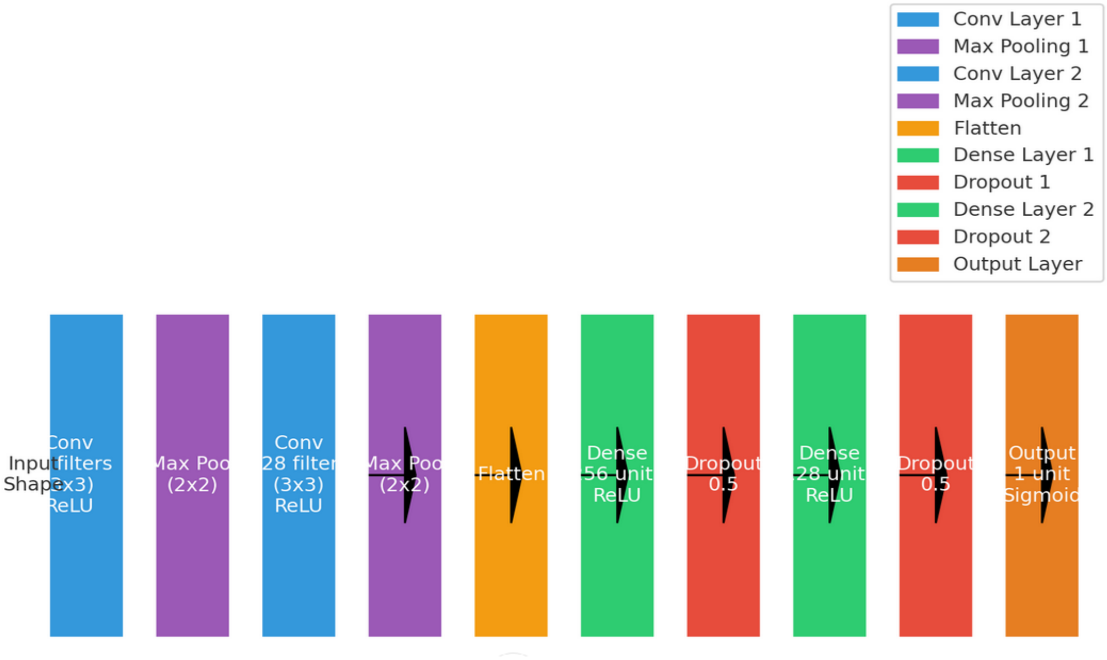


Figure 4.1 CNN architecture that employed in this research.

4.2 Results:

METRIC ±STD	RF_XGB_ RDKIT	MI_XGB_ RDKIT	FS_XGB_ RDKIT	RF_OPTUN A_RDKIT	MI_OPTUN A_RDKIT	FS_OPTU NA_RDKI T
PRECISION ±STD	0.05 ± 0.04	0.09 ± 0.04	0.10 ± 0.04	0.09 ± 0.01	0.08 ± 0.01	0.13 ± 0.02
RECALL ±STD	0.02 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.09 ± 0.01	0.03 ± 0.00	0.03 ± 0.00

METRIC ±STD	**RF_XGB_RDKIT**	**MI_XGB_RDKIT**	**FS_XGB_RDKIT**	**RF_OPTUNA_RDKIT**	**MI_OPTUNA_RDKIT**	**FS_OPTUNA_RDKIT**
F1 SCORE ±STD	0.02 ± 0.02	0.04 ± 0.02	0.04 ± 0.02	0.09 ± 0.01	0.04 ± 0.00	0.04 ± 0.01
MCC ±STD	0.01 ± 0.02	0.03 ± 0.02	0.03 ± 0.02	0.06 ± 0.01	0.03 ± 0.00	0.04 ± 0.01
TRUE POSITIVES (TP) ±STD	0.60 ± 0.50	1.00 ± 0.46	1.00 ± 0.46	3.45 ± 0.51	1.00 ± 0.00	1.00 ± 0.00
TRUE NEGATIVES (TN) ±STD	1195.35 ± 2.21	1194.90 ± 2.05	1195.20 ± 2.53	1170.90 ± 1.25	1194.4 ± 1.00	1198.05 ± 1.23
FALSE POSITIVES (FP) ±STD	9.65 ± 2.21	10.10 ± 2.05	9.80 ± 2.53	34.10 ± 1.25	10.06 ± 1.00	6.95 ± 1.23
FALSE NEGATIVES (FN) ±STD	37.40 ± 0.50	37.00 ± 0.46	37.00 ± 0.46	34.55 ± 0.51	37 ± 0.00	37.00 ± 0.00
NUMBER OF SELECTED FEATURES ±STD	170.90 ± 1.17	147.70 ± 5.88	138.40 ± 24.62	170.45 ± 1.76	148.05 ±4.70	138.40 ± 24.62

Table 4.1 The metric values for evaluation, obtained through RDKIt descriptors, were calculated after training the model on the entire training set and subsequently evaluating it on the test set.

METRIC	RF_XGB_MORD	MI_XGB_M ORD	FS_OPTUNA_MORD	RF_OPTUNA_MORD	MI_OPTUNA_MORD	FS_OPTUNA_MORD
NUMBER OF SELECTED FEATURES ±STD	399.20 ± 3.49	443.85 ± 4.67	415.5 ± 73.93	397.70 ± 4.27	442.10 ± 9.36	415.50 ± 73.93
MAX_DEPTH	9.00	7.00	9.00	Not Applicable	Not Applicable	Not Applicable
PRECISION ±STD	0.13 ± 0.06	0.14 ± 0.05	0.14 ± 0.06	0.14 ± 0.00	0.23 ± 0.16	0.09 ± 0.03
RECALL ±STD	0.03 ± 0.02	0.04 ± 0.01	0.03 ± 0.01	0.03 ± 0.00	0.02 ± 0.01	0.03 ± 0.00
F1 SCORE ±STD	0.05 ± 0.02	0.06 ± 0.02	0.05 ± 0.02	0.04 ± 0.00	0.03 ± 0.02	0.04 ± 0.00
MCC ±STD	0.05 ± 0.03	0.05 ± 0.03	0.05 ± 0.03	0.05 ± 0.00	0.06 ± 0.04	0.03 ± 0.01
TP (TRUE POSITIVES) ±STD	1.30 ± 0.57	1.35 ± 0.49	1.25 ± 0.44	1.00 ± 0.00	0.70 ± 0.47	1.00 ± 0.00
TN (TRUE NEGATIVES) ±STD	1195.80 ± 1.91	1196.05 ± 2.16	1196.95 ± 1.79	1199.00 ± 0.00	1203.00 ± 3.07	1194.15 ± 3.07
FP (FALSE POSITIVES) ±STD	9.20 ± 1.91	8.95 ± 2.16	8.05 ± 1.79	6.00 ± 0.00	2.00 ± 0.00	10.85 ± 3.07
FN (FALSE NEGATIVES) ±STD	36.70 ± 0.57	36.65 ± 0.49	36.75 ± 0.44	37.00 ± 0.00	37.30 ± 0.47	37.00 ± 0.0

Table 4.2 The metric values for evaluation, obtained through Mordred descriptors, were calculated after training the model on the entire training set and subsequently evaluating it on the test set.

4.3 Discussion over the obtained results:

4.3.1 discussion over the RDkit methods table1:

When comparing the methods based on their mean values along with their standard deviations (std), here's the conclusion: RF_Optuna_RDKit emerges as the best method overall, considering its high mean values across multiple metrics (Precision, F1 Score, MCC, TP, TN, FP, FN) and relatively low standard deviations, which signify consistent and reliable performance across different evaluations or datasets. For further explanation check appendix B.

4.3.2 Discussion over Mordred methods table 2:

When comparing the methods based on their mean values and standard deviations (std), the conclusion is as follows: RF_Optuna_Mord stands out as the best overall method. It exhibits high mean values across various metrics (Precision, F1 Score, MCC, TP, TN, FP, FN) and relatively low standard deviations, indicating consistent and reliable performance across different evaluations or datasets. On the other hand, MI_Optuna_Mord demonstrates higher precision but comes with a higher standard deviation. For further explanation check appendix C.

The process of identifying potential senolytics in the unlabeled dataset involved conducting 20 trials using different methods. Following this, we compiled a dataframe of compounds that exhibited a high probability of being senolytics. The unstable performance was anticipated due to the model's higher standard deviation. Among these, we identified compounds predicted by the model multiple times, as outlined below:

Name	Count
PD318088	20
O6-Benzylguanine	20
Spautin-1	20
Levulinic acid	20
CUDC-907	20
Thiazovivin	20
Peiminine	20
RN486	18
Resiquimod	13
Ethynodiol diacetate	10
Entacapone	4
Trimetazidine	4
Genistein	1

Table 4.3 The compounds predicted by the model of MI_Optuna_Mord to be senolytics using an unlabeled dataset have consistently appeared as such across 20 trials, with their frequency counts indicating repeated identification as potential senolytics

The compounds predicted by the RF_Optuna_Mord model to be senolytics using an unlabeled dataset have consistently appeared as such across 20 trials, with their frequency counts indicating repeated identification as potential senolytics. The compounds identified include: **Peiminine, Furosemide, Isoalantolactone, SB415286,**

PD318088, AZ5104, Halcinonide, Dihydrostreptomycin sulfate, and Thymoquinon. Each of these compounds was identified 20 times, demonstrating consistent prediction by the model.

Results of using CNN:

Validation Recall	Training Loss	Validation Loss	Validation Precision	Validation MCC	Test Precision	Test Recall	Test MCC
0.166667	0.000607	0.594402	0.285714	0.20382	0.043478	0.026316	0.010294

Table 4.4 The metric values for evaluation, obtained through RDKit descriptors, were calculated after using neural learning network.

Validation Recall	Training Loss	Validation Loss	Validation Precision	Validation MCC	Test Precision	Test Recall	Test MCC
0.583333	0.005655	0.132633	0.777778	0.666824	0.076923	0.0263	0.02768

Table 4.5 The metric values for evaluation, obtained through Mordred descriptors, were calculated after using neural learning network.

Comparison between methods that uses RDkit descriptors and the methods that uses Mordred descriptors:

Based on the comparison across these metrics, RF_Optuna_Mord generally performs better than RF_Optuna_RDKit in terms of precision, recall, F1 score, and MCC. It consistently reduces false positive. Therefore, RF_Optuna_Mord is considered the better method when using descriptors based on these evaluation criteria. For more explanation check appendix D.

5. Conclusions and Future scope:

In summary, our study aimed to refine the methodology proposed by the authors and investigators alternative model's effectiveness in identifying potential senolytics. We found that the original approach lacked efficiency in model comparison and feature selection, prompting us to enhance it by implementing cross-validation techniques. Our exploration also extended to utilizing various feature selection methods and assessing the performance of deep learning models with this context.

Our computational pipeline consisted of several crucial stages, including feature selection employing mutual information, random forest techniques, and Fisher scoring, hyperparameters tuning via Optuna, and nested cross-validation to pinpoint the optimal hyperparameters for each proposed model. This comprehensive methodology enabled us to assess model performance across multiple evaluation metrics and select the most suitable model for further analysis. We executed this pipeline by introducing one modification at a time and comparing it against the performance of the author's approach using a test set to evaluate efficacy.

After numerous trials utilizing RDKit descriptors, we observed that mutual information and Fisher score for features selection improved the model's performance. Additionally, implementing hyperparameter tuning as an additional change resulted in the model achieving higher scores in the metrics. Incorporating both hyperparameters tuning and features selection techniques yielded higher performance for Fisher with hyperparameters tuning, but it didn't add value for mutual information with hyperparameter tuning. Perhaps the similarity in performance between the model selected using Optuna and XGB.

Furthermore, our exploration extended to use of Mordred descriptors as an alternative to RDKit descriptors, providing a more comprehensive molecular representation for our models. This diversification allowed us to compare the performance of models trained on different sets of features and assess their suitability for identifying potential senolytics. We noted that the model effectively operated better with Mordred descriptors than RDKit descriptors. Following numerous trials utilizing Mordred descriptors, we observed improved performance for all methods, with the best method being RF_Optuna_Mord.

However, using a shallow neural network performed poorly compared to the classical methods, indicating that deep learning benefits from improvements to enhance performance, for instance, applying transfer learning could be advantageous due to the limited number of available samples.

It is worth mentioning that the lack of data samples and the significant imbalance greatly impacted the project's performance. We recommend conducting additional lab experiments to gather more samples for the dataset. Furthermore, Lab experiment for the compounds that we found.

Conducting a similar study for related problems, such as predicting new antibiotics, would be an efficient approach.

References

(1) Di Micco R, Krizhanovsky V, Baker D, Di Fagagna FD. Cellular senescence in ageing: from mechanisms to therapeutic opportunities. Nature Reviews Molecular Cell Biology. 2020;22(2):75-95. doi:10.1038/s41580-020-00314-w

(2) Fuhrmann-Stroissnigg H, Ling YY, Zhao J, et al. Identification of HSP90 inhibitors as a novel class of senolytics. Nature Communications. 2017;8(1). doi:10.1038/s41467-017-00314-z

(3) Chang J, Wang Y, Shao L, et al. Clearance of senescent cells by ABT263 rejuvenates aged hematopoietic stem cells in mice. Nature Medicine. 2015;22(1):78-83. doi:10.1038/nm.4010

(4) Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nature Communications. 2023;14(1). doi:10.1038/s41467-023-41948-6

(5) Seok J, Kang YS. Mutual Information between Discrete Variables with Many Categories using Recursive Adaptive Partitioning. Scientific Reports. 2015;5(1). doi:10.1038/srep10981

(6) Al-Ani A, Deriche M. Feature selection using a mutual information based measure. IEEE Xplore. June 2003. doi:10.1109/icpr.2002.1047405

(7) Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Systems With Applications. 2021;182:115222. doi:10.1016/j.eswa.2021.115222

(8) Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna. arXiv Preprint arXiv. July 2019. doi:10.1145/3292500.3330701

(9) Maćkiewicz A, Ratajczak W. Principal components analysis (PCA). Computers & Geosciences. 1993;19(3):303-342. doi:10.1016/0098-3004(93)90090-r

(10) Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences. 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202

(11) Ramsay J, Silverman BW. Functional Data analysis. Springer Science & Business Media; 2006.

(12) Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. Journal of Cheminformatics. 2018;10(1). doi:10.1186/s13321-018-0258-y

(13) Wong F, Omori S, Donghia NM, Zheng EJ, Collins JJ. Discovering small-molecule senolytics with deep neural networks. Nature Aging. 2023;3(6):734-750. doi:10.1038/s43587-023-00415-z

(14) Smer-Barreto V, Quintanilla A, Elliott RJR, et al. Discovery of senolytics using machine learning. Nature Communications. 2023;14(1). doi:10.1038/s41467-023-39120-1

(15) Van Der Maaten L. Learning a parametric embedding by preserving local structure. PMLR. https://proceedings.mlr.press/v5/maaten09a.html. Published April 15, 2009.

(16) Smer-Barreto V, Quintanilla A, Elliott RJR, et al. Discovery of senolytics using machine learning. Nature Communications. 2023;14(1). doi:10.1038/s41467-023-39120-1

(17) Sun L, Zhang XY, Qian YH, Xu JC, Zhang SG, Tian Y. Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. Applied Intelligence. 2018;49(4):1245-1259. doi:10.1007/s10489-018-1320-1

(18) Evgeniou T, Pontil M. Support Vector Machines: Theory and applications. In: Lecture Notes in Computer Science. ; 2001:249-257. doi:10.1007/3-540-44673-7_12

(19) Chau AL, Li X, Yu W. Convex and concave hulls for classification with support vector machine. Neurocomputing. 2013;122:198-209. doi:10.1016/j.neucom.2013.05.040

(20) Chen T, Guestrin C. XGBoost. arXiv Preprint arXiv:. August 2016. doi:10.1145/2939672.2939785

(21)DiGangi EA, Hefner JT. Ancestry estimation. In: Elsevier eBooks. ; 2013:117-149. doi:10.1016/b978-0-12-385189-5.00005-4

(22) Sperandei S. Understanding logistic regression analysis. Biochemia Medica. January 2014:12-18. doi:10.11613/bm.2014.003

(23) Reid T. Powell. "Computational precision therapeutics and drug repositioning." In Kenneth S. Ramos (Ed.), Comprehensive Precision Medicine (First Edition), Elsevier, 2024, Pages 57-74. ISBN 9780128242568. DOI: 10.1016/B978-0-12-824010-6.00063-0.

(24) Nima Rezaei, Parnian Jabbari. "Linear and logistic regressions in R." In Nima Rezaei, Parnian Jabbari (Eds.), Immunoinformatics of Cancers, Academic Press, 2022, Pages 87-125. ISBN 9780128224007. DOI: 10.1016/B978-0-12-822400-7.00004-X.

(25) Jiang L, Wang D, Cai Z, Yan X. Survey of Improving Naive Bayes for Classification. In: Lecture Notes in Computer Science. ; 2007:134-145. doi:10.1007/978-3-540-73871-8_14

(26) Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1). doi:10.1186/s12864-019-6413-7

(28) Goutte C, Gaussier E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: Lecture Notes in Computer Science. ; 2005:345-359. doi:10.1007/978-3-540-31865-1_25

(29) Chase Lipton Z, Elkan C, Narayanaswamy B. Thresholding classifiers to maximize F1 score. arXiv:1402.1892 [stat.ML]. https://arxiv.org/abs/1402.1892. Published February 8, 2014. Accessed May 14, 2014

(30) Zhu M. Recall, Precision and Average Precision. Research Gate; 2004. https://www.researchgate.net/publication/228874142_Recall_precision_and_average_precision.

(31) Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9783527628766. Published July 15, 2009

(32) Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. Journal of Cheminformatics. 2018;10(1). doi:10.1186/s13321-018-0258-y

(33) Power H, Valtchev P, Dehghani F, Schindeler A. Strategies for senolytic drug discovery. Aging Cell. 2023;22(10). doi:10.1111/acel.13948

(34) Kirkland JL, Tchkonia T. Senolytic drugs: from discovery to translation. Journal of Internal Medicine. 2020;288(5):518-536. doi:10.1111/joim.13141

(35) Rad AN, Grillari J. Current senolytics: Mode of action, efficacy and limitations, and their future. Mechanisms of Ageing and Development. 2024;217:111888. doi:10.1016/j.mad.2023.111888

(36) Liu H. Feature selection. In: Springer eBooks. ; 2011:402-406. doi:10.1007/978-0-387-30164-8_306

(37) Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics. 2022;2. doi:10.3389/fbinf.2022.927312

(38) Jeuken GS, Käll L. Pathway analysis through mutual information. Bioinformatics. 2024;40(1). doi:10.1093/bioinformatics/btad776

(39) Gu Q, Li Z, Han J. Generalized Fisher score for feature selection. arXiv (Cornell University). January 2012. doi:10.48550/arxiv.1202.3725

(40) Der Maaten LV, E. Hinton G. Visualizing data using T-SNE. Journal of Machine Learning Research. 2008;9:2579-2605. https://api.semanticscholar.org/CorpusID:5855042.

(41) Kursa MB, Rudnicki WR. The All Relevant Feature Selection using Random Forest. arXiv (Cornell University). January 2011. doi:10.48550/arxiv.1106.5112

(42) Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics. 2022;2. doi:10.3389/fbinf.2022.927312

(43) Zhu Y, Doornebal EJ, Pirtskhalava T, et al. New agents that target senescent cells: the flavone, fisetin, and the BCL-XL inhibitors, A1331852 and A1155463. Aging. 2017;9(3):955-963. doi:10.18632/aging.101202

(44) Li W, He Y, Zhang R, Zheng G, Zhou D. The curcumin analog EF24 is a novel senolytic agent. Aging. 2019;11(2):771-782. doi:10.18632/aging.101787

(45) Deng J, Yang Z, Ojima I, Samaras D, Wang F. Artificial intelligence in drug discovery: applications and techniques. Briefings in Bioinformatics. 2021;23(1). doi:10.1093/bib/bbab430

(46) Pérez-Sianes J, Pérez-Sánchez H, Díaz F. Virtual screening meets deep learning. Current Computer - Aided Drug Design. 2018;15(1):6-28. doi:10.2174/1573409914666181018141602

(47) Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine Learning in Drug Discovery: A review. Artificial Intelligence Review. 2021;55(3):1947-1999. doi:10.1007/s10462-021-10058-4

(48) Askr H, Elgeldawi E, Ella HA, Elshaier Y a. MM, Gomaa MM, Hassanien AE. Deep learning in drug discovery: an integrative review and future challenges. Artificial Intelligence Review. 2022;56(7):5975-6037. doi:10.1007/s10462-022-10306-1

(49) Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, et al. A review on machine learning approaches and trends in drug discovery. Computational and Structural Biotechnology Journal. 2021;19:4538-4558. doi:10.1016/j.csbj.2021.08.011

(50) Han M, Ren W, Liu X. Joint mutual information-based input variable selection for multivariate time series modeling. Engineering Applications of Artificial Intelligence. 2015;37:250-257. doi:10.1016/j.engappai.2014.08.011

(51) Aria M, Cuccurullo C, Gnasso A. A comparison among interpretative proposals for Random Forests. Machine Learning With Applications. 2021;6:100094. doi:10.1016/j.mlwa.2021.100094

(52) Breiman L. Random forests. In: Machine Learning. Vol 45. ; 2001:5-32. doi:10.1023/a:1010933404324

(53) rdkit.Chem.Descriptors module — The RDKit 2024.03.5 documentation. https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors.html.

(54) Brown JB. Classifiers and their Metrics Quantified. Molecular Informatics. 2018;37(1-2). doi:10.1002/minf.201700127

(55) Blasiak J. Senescence in the pathogenesis of age-related macular degeneration. Cellular and Molecular Life Sciences. 2020;77(5):789-805. doi:10.1007/s00018-019-03420-x

(56) Science & Pipeline | Atropos Therapeutics. Atropos Therapeutics. https://www.atroposthera.com/science-pipeline.

APPENDIX A

Method 1 (RF_XGB_RDKit) - Trial 1: Features: 171, Max Depth: 4 Test Set: Precision: 0.142857, Recall: 0.026316, F1: 0.044444, MCC: 0.049085, TP: 1, TN: 1199, FP: 6, FN: 37 Validation Set: Precision: 0.875, Recall: 0.411765, F1: 0.56, MCC: 0.59464, TP: 7, TN: 736, FP: 1, FN: 10 Features: 170, Hyperparameters: {'C': 89.6265, 'gamma': 9.965, 'kernel': 'poly'} Test Set: Precision: 0.166667, Recall: 0.026316, F1: 0.045455, MCC: 0.055058, TP: 1, TN: 1200, FP: 5, FN: 37 Validation Set: Precision: 1, Recall: 0.117647, F1: 0.210526, MCC: 0.339554, TP: 2, TN: 736, FP: 0, FN: 15

Method 2 (MI_XGB_RDKit) - Trial 6: Features: 142, Max Depth: 9 Test Set: Precision: 0.166667, Recall: 0.026316, F1: 0.045455, MCC: 0.055058, TP: 1, TN: 1200, FP: 5, FN: 37 Validation Set: Precision: 1, Recall: 0.294118, F1: 0.454545, MCC: 0.537958, TP: 5, TN: 736, FP: 0, FN: 12 Method 3 (FS_XGB_RDKit) - Trial 16: Features: 116, Max Depth: 9 Test Set: Precision: 0.230769, Recall: 0.078947, F1: 0.117647, MCC: 0.119554, TP: 3, TN: 1195, FP: 10, FN: 35 Validation Set: Precision: 0.8333, Recall: 0.294118, F1: 0.434783, MCC: 0.489145, TP: 5, TN: 735, FP: 0, FN: 16 Method 4 (RF_Optuna_RDKit) - Trial 10: Features: 170, Hyperparameters: {'C': 89.6265, 'gamma': 9.965, 'kernel': 'poly'} Test Set: Precision: 0.166667, Recall: 0.026316, F1: 0.045455, MCC: 0.055058, TP: 1, TN: 1200, FP: 5, FN: 37 Validation Set: Precision: 1, Recall: 0.117647, F1: 0.210526, MCC: 0.339554, TP: 2, TN: 736, FP: 0, FN: 15

Method 5 (MI_Optuna_RDKit) - Trial 8: Features: 146, Hyperparameters: {'C': 53.5967, 'gamma': 1.7274, 'kernel': 'rbf'} Test Set: Precision: 0.5, Recall: 0.026316, F1: 0.05, MCC: 0.109467, TP: 1, TN: 1204, FP: 1, FN: 37 Validation Set: Precision: 1, Recall: 0.117647, F1: 0.210526, MCC: 0.339554, TP: 2, TN: 736, FP: 0, FN: 15 Method 6 (RF_XGB_RDKit) - Trial 16: Features: 111, Hyperparameters: {'C': 42.783, 'gamma': 6.143, 'kernel': 'rbf'} Test Set: Precision: 0.1, Recall: 0.026316, F1: 0.041667, MCC: 0.03632, TP: 1, TN: 1196, FP: 9, FN: 37 Validation Set: Precision: 0.75, Recall: 0.176471, F1: 0.285714, MCC: 0.357856, TP: 3, TN: 735, FP: 1, FN: 14. Method 7 (RF_XGB_Mord) - Trial 5: Features: 503, Max Depth: 9 Test Set: Precision: 0.181818, Recall: 0.052632, F1: 0.081633, MCC: 0.083016, TP: 2, TN: 1196, FP: 9, FN: 36 Validation Set: Precision: 1, Recall: 0.117647, F1: 0.210526, MCC: 0.339554, TP: 2, TN: 736, FP: 0, FN: 15

Method 8 (MI_XGB_Mord) - Trial 0: Features: 533, Max Depth: 7 Test Set: Precision: 0.125, Recall: 0.026316, F1: 0.043478, MCC: 0.044147, TP: 1, TN: 1198, FP: 7, FN: 37

Validation Set: Precision: 0.8, Recall :0.235294, F1: 0.363636, MCC: 0.427882, TP: 4, TN: 735, FP: 1,FN: 13

Method 9 (FS_XGB_Mord) - Trial 20: Features: 212, Max Depth: 7 Test Set: Precision: 0.2, Recall: 0.026316, F1: 0.046512, MCC: 0.062546, TP: 1, TN: 1201, FP: 4, FN: 36 Validation Set: Precision: 1, Recall: 0.235294, F1: 0.380952, MCC: 0.480843, TP: 4, TN: 736, FP: 0, FN: 13. Method 10 (RF_Optuna_Mord) - Trial 5: Features: 506, Hyperparameters: {'C': 46.1245, 'gamma': 2.8592, 'kernel': 'rbf'} Test Set: Precision: 0.166667, Recall: 0.026316, F1: 0.045455, MCC: 0.055058, TP: 1, TN: 1200, FP: 5, FN: 37 Validation Set: Precision: 1, Recall: 0.117647, F1: 0.210526, MCC: 0.339554, TP: 2, TN: 736, FP: 0, FN: 15.

Method 11 (MI_Optuna_Mord) - Trial 13: Features: 346, Hyperparameters: {'C': 23.3017, 'gamma': 0.9808, 'kernel': 'rbf'} Test Set: Precision: 0.333333, Recall: 0.026316, F1: 0.04878, MCC: 0.086504, TP: 1, TN: 1203, FP: 2, FN: 37 Validation Set: Precision: 1, Recall: 0.058824, F1: 0.111111, MCC: 0.239942, TP: 1, TN: 736, FP: 0, FN: 16.

Method 12 (FS_Optuna_Mord) - Trial 16: Features: 279, Hyperparameters: {'C', 23.30174495312893), ('gamma', 0.9808224548231897), ('kernel', 'rbf')} Test Set: Precision: 0.2, Recall: 0.026316, F1: 0.046512, MCC: 0.062546, TP: 1, TN: 1201, FP: 4, FN: 37 Validation Set: Precision: 0.75, Recall: 0.176471, F1: 0.285714, MCC: 0.357856, TP: 3, TN: 735, FP: 1, FN: 14.

APPENDIX B:

Precision: FS_Optuna_RDKit has the highest mean precision of 0.13, with a relatively moderate standard deviation of ±0.02. This indicates that while FS_Optuna_RDKit shows the highest average precision, there is some variability in its performance across different runs or datasets.

Recall: RF_Optuna_RDKit has the highest mean recall of 0.09, with a standard deviation of ±0.01. This suggests that RF_Optuna_RDKit consistently performs well in terms of recall across different evaluations. F1 Score: RF_Optuna_RDKit has the highest mean F1 score of 0.09, with a standard deviation of ±0.01. This means RF_Optuna_RDKit achieves a good balance between precision and recall, although there is some variability in its F1 score performance.

MCC (Matthews Correlation Coefficient): RF_Optuna_RDKit has the highest mean MCC of 0.06, with a standard deviation of ±0.01. This metric indicates that RF_Optuna_RDKit is effective in capturing the balance between true positives and true negatives, with moderate variability in its MCC values.

True Positives (TP): RF_Optuna_RDKit has the highest mean TP values (3.45), with a standard deviation of ±0.51. This shows that this method consistently identifies a certain number of true positives.

False Positive (FP): RF_Optuna_RDKit has the highest mean FP values (34.1), with a standard deviation of ±1.25. This shows that this method consistently identifies a certain number of false positives, which consider a downside of the model performance.

Conclusion: RF_Optuna_RDKit emerges as the best method overall, considering its high mean values across multiple metrics (Precision, F1 Score, MCC, TP, TN, FP, FN) and relatively low standard deviations, which signify consistent and reliable performance across different evaluations or datasets, But with high false positive.

APPENDIX C:

Considering the standard deviation (std) values alongside the mean values provides additional insight into the consistency or variability of each method's performance across the 20 runs. Lower standard deviations generally indicate more consistent performance. Let's reconsider the assessment based on both mean and standard deviation values: Precision, Recall, F1 Score, MCC: Higher mean values are better. Lower standard deviations indicate more consistent performance. MI_Optuna_Mord has the highest mean Precision (0.23) and MCC (0.06), but its standard deviations are also relatively higher compared to other methods. RF_Optuna_Mord shows competitive mean scores with relatively lower standard deviations, indicating consistent performance across runs. TP (True Positives) and TN (True Negatives): Higher mean values are better. Lower standard deviations indicate more consistent performance. MI_XGB_Mord has the highest mean TP values (1.35), but its standard deviations is higher compared to RF_Optuna_Mord, which shows a slightly lower mean but potentially more consistent performance. FP (False Positives) and FN (False Negatives): Lower mean values are better. Lower standard deviations indicate more consistent performance. MI_Optuna_Mord has the lowest mean FP value (2.00) and competitive mean FN value (37.30) with relatively higher standard deviations. Conclusion with consideration of std values: While MI_Optuna_Mord has the highest mean values in key metrics like Precision and MCC, its higher standard deviations suggest more variability in performance across runs. RF_Optuna_Mord, on the other hand, shows competitive mean scores with generally lower standard deviations, indicating more stable performance. Therefore, RF_Optuna_Mord could be considered the best-performing method overall when accounting for both mean and standard deviation values across the metrics provided.

APPENDIX D:

Precision and F1 Score: RF_Optuna_Mord shows slightly higher precision (0.14 ± 0.00) compared to RF_Optuna_RDKit (Precision: 0.09 ± 0.01). Recall: RF_Optuna_Mord also has a lower recall (0.03 ± 0.00) compared to RF_Optuna_RDKit (Recall: 0.06 ± 0.01). MCC: Both RF_Optuna_RDKit and RF_Optuna_Mord have similar MCC values (0.06 ± 0.01 and 0.05 ± 0.00, respectively), with RF_Optuna_RDkit showing slightly higher average. True Positives (TP): RF_Optuna_Mord consistently identifies less true positives (1.00 ± 0.00) compared to RF_Optuna_RDKit (3.45 ± 0.51). True Negatives (TN): RF_Optuna_Mord has a slightly higher average of true negatives (1199.00 ± 0.00) compared to RF_Optuna_RDKit (1170.90 ± 1.25). False Positives (FP) and False Negatives (FN): RF_Optuna_RDKit has a higher average of false positives (34.55 ± 0.51) and lower of false negatives (34.55 ± 0.51) compared to RF_Optuna_Mord (6.00 ± 0.00 and 37.00 ± 0.00, respectively). Given our focus on reducing false positives and increasing precision, we can conclude that RF_Optuna_Mord outperforms RF_Optuna_RDKit.