Received: 18 October 2024 | Revised: 28 February 2025 | Accepted: 13 March 2025 | Published online: 25 March 2025

RESEARCH ARTICLE

Medinformatics
2025, Vol. 00(00) 1–10
DOI: 10.47852/bonviewMEDIN52024571

BON VIEW PUBLISHING

Exploration of Various Supervised Machine Learning Algorithms for Predictive Modeling of *Caenorhabditis elegans* Lifespan-Extending Compound Dataset

Amisha Bisht[1], Disha Tewari[2], Kalpana Rawat[3], Priyanka Joshi[3], Sanjay Kumar[4] and Subhash Chandra[3,*]

[1]*Department of Botany, Soban Singh Jeena University (Pt. Badridutt Pandey Campus), India*

[2]*Department of Biotechnology, Kumaun University, India*

[3]*Department of Botany, Soban Singh Jeena University, India*

[4]*Department of Botany, Hukum Singh Bora Government Post Graduate College Someshwar, India*

Abstract: The discovery of compounds that extend lifespan is a key objective in aging research. The nematode *Caenorhabditis elegans* is an established model organism for studying aging due to its short lifespan and conserved molecular pathways related to longevity. This study aims to develop a predictive model for lifespan-extending compounds using a machine learning (ML) approach with mljar-supervised, an automated ML (AutoML) Python package. Various ML algorithms, including Decision Trees, Random Forest, Extra Trees, XGBoost, LightGBM, CatBoost, and Neural Network, were explored to analyze and predict the efficacy of compounds in extending *C. elegans* lifespan. In this work, we analyze data from the DrugAge database, which contains chemical compounds and their effect on the lifespan of model organisms. Predictive models were built using ML to predict whether a chemical compound will increase lifespan, using chemical descriptors calculated from each compound's chemical structure. The performance of the models was evaluated using metrics such as accuracy, precision, and recall. These evaluations demonstrated the exceptional predictive capability of the algorithms, achieving a remarkable accuracy rate. The results of this exploration provide insights into optimal ML models for predicting potential lifespan-extending compounds and highlight the importance of AutoML in accelerating research in aging and longevity.

Keywords: *Caenorhabditis elegans*, lifespan-extending compounds, predictive model, machine learning, automated machine learning (AutoML), mljar-supervised, aging research

1. Introduction

Aging is an intricate biological process that causes a slow decline in cellular capabilities, raising the risk of a variety of age-related disorders [1]. These mechanisms cause cumulative damage over time, gradually inhibiting the body's capacity to maintain homeostasis and adapt to stress. This contributes to the emergence of chronic diseases often associated with aging [2]. Identifying compounds that might extend life or reduce the effects of aging is an important field of research in biogerontology [3]. As a result, various model organisms are used in aging research, with the nematode *Caenorhabditis elegans* (*C. elegans*), being one of the most prominent [4]. *C. elegans*, a worm with well-defined genetics and a short lifespan, is a prominent model organism for aging research because of its simplicity and ease of manipulation in laboratory settings. Its short life cycle enables researchers to rapidly observe the impact of genetic and environmental influences on aging, making it an essential tool for studying the biological mechanisms that underlie longevity and age-related disorders [5, 6].

Over the last few decades, the DrugAge database, a curated collection of chemicals known to alter the lifespan of model organisms such as *C. elegans*, has evolved as an invaluable resource for drug development and aging studies [7]. This database offers a detailed list of compounds that may have antiaging effects and information on how they work and the biological pathways they affect. Scientists can use this data to pinpoint potential subjects for more in-depth study and to create tests that examine how these substances impact lifespan [8]. Moreover, the database helps uncover new targets for drugs and assists in confirming current antiaging treatments. As geroscience progresses, DrugAge remains essential in connecting basic research to clinical uses, stimulating innovation in therapies for aging [9].

Machine learning (ML), especially deep learning, has revolutionized drug discovery by analyzing large datasets to predict molecular properties, identify drug candidates, and optimize clinical trials. A review in Artificial Intelligence (AI) in Medicine highlights how AI, including ML and deep learning, has enabled extensive

*Corresponding author: Subhash Chandra, Department of Botany, Soban Singh Jeena University, India. Email: subhashchandra@ssju.ac.in

data analysis and molecular prediction in drug discovery [10]. In oncology, ML has helped predict cardiac risks from cancer treatments [11], and in cardiology, it has been applied to identify cardiovascular risks in cancer patients for early intervention [12]. While ML applications in aging are still emerging, AI's ability to analyze biological data offers promise for identifying biomarkers and therapeutic targets for age-related diseases.

ML algorithms significantly enhance the discovery process by analyzing vast and complex datasets to predict novel antiaging compounds and optimize therapeutic strategies [13]. ML bridges the gap between raw data and actionable insights, offering a transformative approach to predictive modeling in biological systems [14]. By evaluating molecular characteristics and interactions, ML algorithms can uncover hidden patterns and relationships that are challenging to identify using traditional methods, enabling researchers to prioritize potential compounds for experimental validation effectively [15]. This method speeds up the discovery process and improves the comprehension of intricate biological mechanisms [16]. Ultimately, leveraging ML in this field could revolutionize our approach to aging and facilitate the development of new treatments that enhance longevity [17, 18]. Through the examination of extensive datasets like DrugAge [https://genomics.senescence.info/drugs/], ML methods can identify trends and forecast the effectiveness of new compounds in prolonging lifespan. The emergence of automated ML (AutoML) platforms, like mljar-supervised, provides a convenient method to experiment with various ML algorithms without extensive knowledge of model optimization [19, 20]. AutoML can automatically assess various metrics, such as accuracy, recall, precision, and ROC-AUC, to compare and choose the most suitable ML models.

In this study, we aim to develop a predictive model using various ML algorithms to assess potential lifespan-extending compounds in *C. elegans*. By leveraging mljar-supervised, an AutoML Python package, we explore and compare the performance of several ML algorithms, including Decision Trees, Random Forest, Extra Trees, XGBoost, LightGBM, CatBoost, and Neural Network (NN) classifiers. The selection of ML algorithms in this study is informed by their proven success in similar research focusing on lifespan extension and the identification of compounds with geroprotective properties. Algorithms such as Decision Trees and ensemble methods like Random Forests and Extra Trees were chosen for their proficiency in handling high-dimensional and intricate datasets, while effectively identifying nonlinear interactions between variables. Random Forests, for instance, have been successfully applied in predicting lifespan-extending chemical compounds, demonstrating their capability in such complex tasks [21]. XGBoost, LightGBM, and CatBoost were preferred for their exceptional performance in classification tasks, particularly with large datasets such as those in the DrugAge database, offering both speed and high accuracy. NNs were incorporated due to their ability to detect complex patterns and relationships that traditional models might miss. Together, these methods provide a robust and versatile approach to evaluating compounds that may extend lifespan, ensuring reliable and accurate predictions across diverse tasks. The DrugAge database provides a robust foundation for training and testing the predictive models, allowing us to identify promising candidates for further biological validation [22]. This exploration seeks to comprehensively evaluate different ML models for lifespan extension, highlighting the potential of AutoML-driven approaches to accelerating the discovery of geroprotective compounds.

This primary objective is to create a powerful predictive model ML technique to identify compounds that may help extend the lifespan of *C. elegans*. By integrating large-scale biological datasets and advanced computational approaches, the model is designed to evaluate and forecast how different compounds impact aging in this model organism. Given that *C. elegans* is widely used for aging research, the findings of this study could provide crucial insights into potential treatments for age-related diseases in humans. Ultimately, the goal is to accelerate the identification of lifespan-modulating compounds, facilitating more efficient screening processes and enhancing our understanding of the biological mechanisms of aging.

2. Materials and Methods

2.1. Dataset collection and preprocessing

The dataset utilized in this analysis includes positive and negative instances of chemical compounds according to their impact on the lifespan of *C. elegans*. Positive instances were obtained from the DrugAge database (Build 4, released on October 20, 2021) [23], which compiles substances with verified lifespan-extending effects (available at: genomics.senescence.info/drugs).

In contrast, the negative example compounds that did not demonstrate a beneficial impact on the lifespan of *C. elegans*—were primarily derived from supplementary material in an earlier study by Barardo *et al*. in 2017 [23]. This dataset was expanded in two key ways. First, we included compounds listed in DrugAge that showed lifespan-increasing effects in other organisms but were found to be detrimental to *C. elegans*. Second, some compounds originally categorized as negative in the Barardo *et al.,* study were reclassified as positive due to updated findings in DrugAge, reflecting advancements in research since the original dataset was compiled six years ago. The DrugAge database may have biases, including data imbalance, incomplete annotations, and publication bias favoring well-studied compounds. Addressing these requires dataset balancing, feature engineering, and unbiased metrics for accurate model evaluation.

Before modeling, the datasets underwent rigorous cleaning and preprocessing to address missing values, and normalization techniques were applied to maintain consistency throughout the data. The final dataset included 448 positive entries and 1,141 negative entries following this curation process, resulting in 1,589 compounds. These compounds were labeled for classification purposes: those associated with longevity (active) were marked as (1), while compounds lacking positive longevity effects were labeled as (0), indicating inactive. Consequently, from the pool of 1,589 compounds, 448 were designated as 1 and 1,141 as 0.

Chemical descriptors, essential attributes of small molecules that dictate their biological activity, were generated for each compound. All compounds were formatted in SMILES (Simplified Molecular Input Line Entry System) notation using Open Babel software [24], and the data were then processed using Mordred, a specialized software for descriptor calculation [25], within the Anaconda 3 environment (https://www.anaconda.com/). In total, this study yielded 1,613 molecular descriptors, encompassing both two-dimensional and three-dimensional features (see Supplementary file).

2.2. Exploration of various supervised machine learning algorithms for predictive modeling

For predictive modeling, we utilized the mljar-supervised AutoML Python package [https://github.com/mljar/mljar-supervised] to explore various supervised ML algorithms. The mljar-supervised AutoML Python package automates ML tasks such as data

preprocessing, model construction, and hyperparameter tuning for tabular data. Its computational efficiency is influenced by several factors, including the choice of algorithms (e.g., decision trees, ensemble methods), which affects training and inference times. Extensive hyperparameter tuning can increase computational demands, while automated preprocessing may add load, especially with large datasets. Additionally, the package supports parallel processing, utilizing multiple CPU cores to improve efficiency and reduce processing time.

The research used a Core i5 CPU, 8 GB RAM, and a 512 GB SSD. The software environment included Python 3 with Scikit-learn, TensorFlow, and other dependencies, utilizing local machines, cloud services (e.g., AWS, Google Cloud), or HPC clusters as needed. In this study, mljar-supervised package automatically tested multiple ML algorithms, optimized hyperparameters, and selected the best-performing model based on various metrics using the competing mode. Several ML algorithms, including Decision Trees, Random Forest, Extra Trees, XGBoost, LightGBM, CatBoost, and NNs, were applied to the dataset. The package also allows for stacking and ensembling models to improve predictive accuracy.

Hyperparameter optimization is essential for enhancing the predictive performance of ML models. In this study, the mljar-supervised AutoML package uses various strategies to fine-tune hyperparameters for each algorithm. For Decision Trees, Random Forests, and Extra Trees, optimization focuses on parameters like tree depth, minimum samples, and splitter, with methods such as grid and random search applied. For boosting algorithms like XGBoost, LightGBM, and CatBoost, important hyperparameters include learning rate, depth, and several estimators, optimized using grid search and Bayesian techniques to improve accuracy. NNs undergo tuning of layers, neurons, learning rate, and batch size, with optimization methods like Bayesian search ensuring the best architecture. The mljar-supervised package automates these processes, enhancing model generalization and minimizing overfitting, leading to improved performance across diverse datasets [26].

2.3. Data splitting

The Kennard-Stone method was employed to split the dataset into training and testing sets, ensuring both sets effectively represented the descriptive space of the dataset. The training set comprised 70% of the dataset, while the leftover 30% was allocated for testing purposes. A 70/30 train-test split ratio was adopted in this study to balance the amount of data available for model training and the robustness of the test set for performance evaluation. This ratio, widely used as a heuristic, ensures the model has sufficient data to learn effectively while retaining a test set large enough for meaningful validation. Additionally, this approach minimizes the risk of overly optimistic or biased performance results on the training data. This technique is well known for its efficiency in choosing representative samples from a heterogeneous dataset [27].

2.4. Machine learning algorithms

Several ML algorithms were implemented. Such as Decision Trees are a classification technique that divides a dataset into subgroups according to the most useful attributes. The model builds a tree-like structure, where each internal node represents a feature test, and each leaf node represents an outcome or class label. The model is renowned for its simplicity and interpretability [28]. Random Forest is a technique that involves merging multiple decision trees and averaging their forecasts to prevent overfitting and enhance generalization. The key feature of a Random Forest is that it introduces randomness into the process of building the trees by selecting a random subset of features at each split in the tree [21]. Extra Trees is similar to Random Forest, but with more randomization in feature selection and cut-point determination, making it more resilient and successful for classification and regression features [29]. XGBoost is a widely recognized gradient-boosting algorithm, praised for its scalability and high performance with huge datasets [30]. XGBoost optimizes both the training process and model performance using techniques such as regularization (to prevent overfitting), handling missing values, and parallelizing the computation [31]. LightGBM is a fast, scalable variation of the gradient boosting decision tree technique that excels at handling huge datasets. It uses histogram-based learning to group continuous features into discrete bins, reducing memory consumption and speeding up training [32, 33]. The CatBoost algorithm effectively mitigates overfitting and addresses the challenge of class imbalance. It performs well with default parameters, eliminating the need for extensive hyperparameter tuning [34]. NNs are strong models that can capture nonlinear interactions between input variables and outputs. They are made up of several layers, with each neuron giving weight to the inputs and using an activation function to calculate the output [35]. AutoML refers to the automation of the entire ML workflow for real-world problems. AutoML frameworks automatically select, optimize, and evaluate ML models, eliminating the need for manual intervention. This allows users to develop high-performing models with minimal ML expertise. These algorithms were used to create a robust screening framework, with AutoML automating the selection of the best-performing model [36, 37].

2.5. Model training validation

To validate the models, a fivefold cross-validation procedure was employed. Metrics such as accuracy, $F1$ score, precision, recall, and Matthews correlation coefficient (MCC) were calculated for each fold [38]. Receiver operating characteristic (ROC) curves and confusion matrices were also used to assess model performance, including true and false-positive rates [39, 40].

2.6. Performance metrics

Each ML method's performance was assessed using a range of important criteria. Accuracy refers to the ratio of accurate predictions generated by the model [41]. Precision measures the proportion of correct positive class identifications, while Recall measures the proportion of true positives that were correctly identified [42, 43]. The $F1$ Score balances precision and recall by calculating their harmonic mean. Furthermore, MCC provides a comprehensive statistic that accounts for both true and false positives and negatives in a balanced manner [44]. Finally, Log-Loss measures how closely predicted probabilities match actual values, which is an essential indicator of model success [45].

2.7. Model performance assessment

Model performance was analyzed using the above metrics, with particular emphasis on the area under the curve (AUC) of the ROC curve. This measure evaluates the model's ability to discriminate between lifespan-extending and non-extending compounds. The best-performing model was selected based on its balanced performance across these criteria.

3. Results and Discussion

3.1. Exploration of various supervised machine learning algorithms for predictive modeling

Exploring various supervised ML algorithms for predictive modeling, we utilized seven different ML algorithms—Decision Trees, Random Forest, Extra Trees, XGBoost, LightGBM, CatBoost, and NN classifiers—to develop an effective prediction model using the C. elegans lifespan-extending chemical database. The mljar-supervised AutoML method selected these classifiers based on their performance in the fivefold cross-validation of the training dataset [46]. The dataset was split into training and testing sets using the Kennard–Stone method to ensure a balanced distribution across the compound's descriptive space [47, 48].

Each ML model was trained on the dataset using fivefold cross-validation to develop cost-optimized models. The outcomes were assessed by generating confusion matrices and ROC curves for every algorithm.

3.2. Performance of machine learning models

The confusion matrix was the primary tool for assessing the performance of each model. The diagonal elements represent the true-positive (TP) and true-negative (TN) values, while the off-diagonal elements correspond to false positives (FP) and false negatives (FN) (see Table 1). By analyzing the confusion matrices of seven ML models (Figure 1A-G), we can assess their performance in terms of TP, TN, FP, and FN.

Among the models, the Decision Tree (Figure 1(A)) and Default_Random Forest (Figure 1(B)) performed the best, achieving perfect classification with no false positives or negatives. Both models recorded 45 true positives and 114 true negatives, making them the top performers. While still effective, the _Default_Extra Trees model (Figure 1(C)) had 10 false positives, slightly lowering its accuracy compared to the Decision Tree and Random Forest models. The _Default_XGBoost model (Figure 1(D)) performed better than Extra Trees with only 1 false positive, though it still fell short of the perfect scores achieved by the Decision Tree and Random Forest models. The _Default_LightGBM (Figure 1(E)) and _Default_CatBoost (Figure 1(F)) models each had 14 false positives, placing them behind Decision Tree, Random Forest, and XGBoost in terms of accuracy. Lastly, the _Default_Neural Network (Figure 1(G)) showed the weakest performance, with 20 false positives and 5 false negatives, making it the least accurate of the models evaluated.

The ROC curves, which evaluate classifier performance across different thresholds, are presented in Figure 2. The curves depict the true-positive rate (sensitivity) versus the false-positive rate, with the AUC representing the crucial performance metric. The Decision Tree (DT), _Default_Random Forest (RF), _Default_Extra Trees (ET), _Default_XGBoost (X), _Default_LightGBM (L), and _Default_CatBoost (C) models all achieved an AUC of 1.00, indicating perfect classification. In contrast, the _Default_NN model had an AUC of 0.94, which, while still strong, is slightly lower than the other models. Overall, the models with an AUC of 1.00 are considered optimal, while the NN model, despite its good performance, lags slightly behind due to its lower AUC score.

Table 1. Confusion matrix

Actual value	Predicted as 0 (Inactive)	Predicted as 1 (Active)
Labeled as 0 (Inactive)	True negative (TN)	False positive (FP)
Labeled as 1 (Active)	False negative (FN)	True positive (TP)

3.3. Predictive model development through machine learning algorithm comparison

The comparative analysis of ML algorithms presented in Table 2 highlights Decision Trees and _Default_Random Forest as the most effective predictive models. Both models achieved optimal performance across all evaluated metrics, including an exceptionally low log loss of 1e-06, an AUC of 1, an *F*1-score of 1, 100% accuracy, precision, and sensitivity/true-positive rate (TPR), along with an MCC of 1. These results indicate that both models are highly suitable for classification tasks in this study, exhibiting perfect predictive accuracy without any misclassifications.

Other models performed well but did not achieve the same level of perfection. The _Default_XGBoost model showed remarkable predictive capacity, with an AUC of 1 and an *F*1-score of 0.98, although its log loss (0.45) was slightly higher compared to the Decision Tree and Random Forest, suggesting that while its predictions were generally accurate, it was slightly less precise in minimizing error. The _Default_Extra Trees model showed good performance, with an *F*1-score of 0.90 and accuracy of 93%, but its lower precision (81%) and MCC (0.86) positioned it behind the leading models. Similarly, the _Default_LightGBM and _Default_CatBoost models yielded *F*1-scores of 0.86, accuracy of 91%, and precision of 76%, but their higher log loss values (0.30 and 0.07, respectively) and lower MCC scores (0.81), which suggest that although they performed reasonably well, they were not as reliable or consistent as the top models. These results point to areas where refinement may be needed, particularly in reducing error rates and enhancing model stability across different data distributions.

The _Default_Neural Network model exhibited the weakest performance, with the lowest *F*1-score (0.76), accuracy (84%), precision (66%), and TPR (88%), as well as the highest log loss (0.61), making it the least effective classifier in this context. This performance suggests that, although NNs are capable of modeling complex patterns, the model's architecture or training procedure may require further optimization to improve its effectiveness in this specific classification task.

Our results outperform similar studies, such as Ribeiro et al. [49], where the Random Forest model showed good performance but lacked consistency across all evaluation metrics [49]. The use of advanced validation methods, including fivefold cross-validation and the Kennard-Stone method, contributed to our enhanced results. Despite strong performance, the Default XGBoost model exhibited a slightly higher log loss, suggesting it could benefit from further optimization. Similarly, Extra Trees, LightGBM, and CatBoost showed higher log loss and lower MCC shown in Figures 3 and 4 respectively, pointing to areas for improvement. The Default NN model underperformed, requiring adjustments to its architecture.

3.4. Feature importance

In ML, feature importance involves evaluating the impact of each input feature (also called a predictor, attribute, or variable) on a model's ability to make predictions or classifications. It shows which features matter the most for the model's results, giving insights into how inputs relate to the outcome [50]. Even

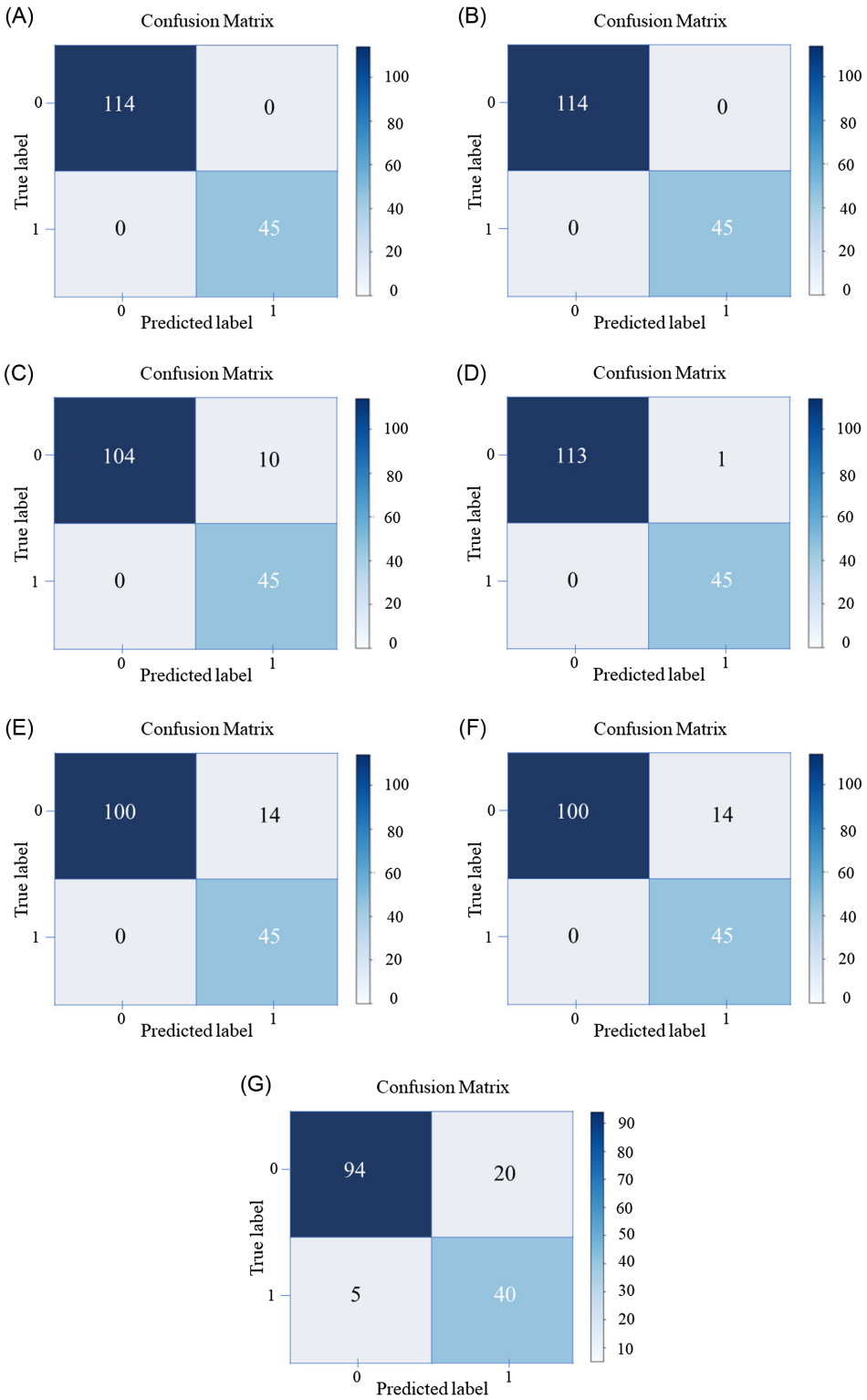

Figure 1. Confusion matrix of models. A. Decision Tree, B. _Default_Random Forest,_C. Default_Extra Trees, D. _Default_XGBoost, E. _Default_LightGBM, F. Default_CatBoost, and G. Default_Neural Network

though the provided dataset was perfectly balanced, conducting analysis using significantly weighted desired features aids in achieving results that are more accurate and precise [51]. Figure 5 presents a heatmap that illustrates the importance of features using feature engineering. This process involves employing data mining techniques to select a diverse array of attributes from a raw dataset. The selection of these attributes enhances accuracy and optimizes outcomes, thus boosting the performance of ML models.

This heatmap visualizes the feature importance of the top 25 chemical descriptors used in predictive modeling, offering insights into their influence on model predictions. Each row corresponds to a specific chemical descriptor, such as AATS0p, BCUTm-1h,

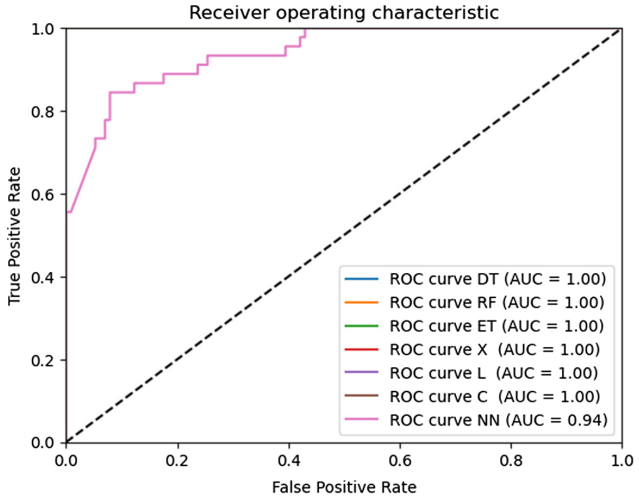

Figure 2. AUC curve values for seven classification algorithms as depicted in the ROC plot (X-axis: false-positive rate, Y-axis: true-positive rate)

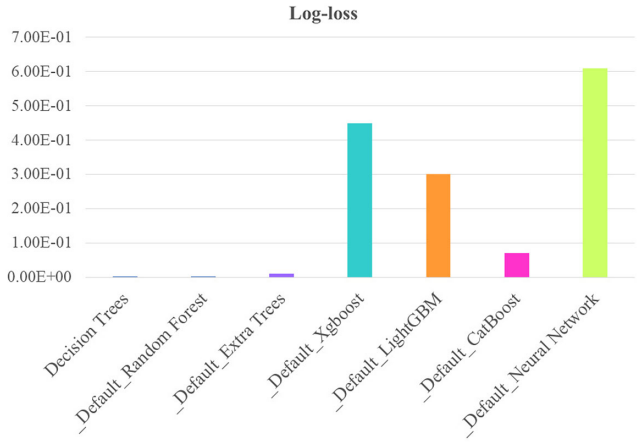

Figure 3. Comparative analysis of Log-Loss across models highlights superior performance of CatBoost, Decision Trees, Random Forest, and Extra Trees with minimal error

and SMR_VSA4, which represent various molecular properties, including atomic, structural, and functional group features. The columns on the x-axis denote different models or prediction sets, likely representing individual ML algorithms. The importance of descriptors varies across these models, providing a clearer understanding of their impact on specific predictions. The color intensity of the heatmap, ranging from lighter to darker blue, reflects the magnitude of feature importance, with darker shades indicating a higher influence on the model's predictions, while lighter or white areas signify lower or negligible importance. Among the descriptors, features such as name_diff_JGI2, AATS0p, and SMR_VSA4 demonstrate the highest influence, as indicated by the darker regions along their rows, suggesting their critical role in shaping the model outcomes. Overall, this visualization helps identify the most significant chemical descriptors for prediction accuracy, enabling researchers to refine models and prioritize these key features in future studies.

Overall, Decision Trees and Random Forest models exhibited flawless metrics, including near-zero Log-Loss (1e-06), perfect AUC (1), and an *F*1 score of 1, indicating exceptional performance. However, such perfection may signal overfitting, where the models rely heavily on memorizing the training data rather than identifying patterns that can extend to new datasets. Their notably short training times further suggest a potential over-reliance on specific dataset characteristics. In contrast, a

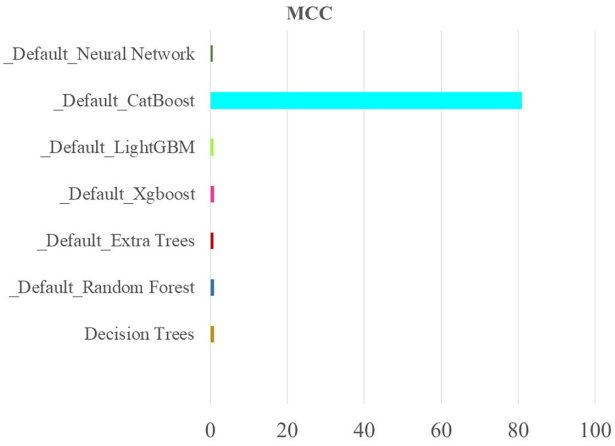

Figure 4. Matthews correlation coefficient (MCC) for seven machine learning models showcases CatBoost's exceptional predictive accuracy

sensitivity analysis through algorithm comparison revealed that models like Extra Trees and XGBoost demonstrated slightly less perfect performance (*F*1 Score: 0.9–0.98, MCC: 0.86–0.98), indicating a better balance between fitting the training data and generalizing to unseen data. NNs, LightGBM, and CatBoost with lower *F*1 Scores and MCC values further suggest a reduced risk

Table 2. Comparative evaluation of performance metrics employing different classifiers

Machine learning algorithm	logloss	AUC	*F*1 score	Accuracy	Precision (%)	Recall/TPR/ Sensitivity/ (%)	MCC	Train_time
Decision Trees	1e-06	1	1	1	1	100	1	33.23
_Default_Random Forest	1e-06	1	1	1	1	100	1	47.57
_Default_Extra Trees	0.01	1	0.9	0.93	0.81	100	0.86	45.36
_Default_Xgboost	0.45	1	0.98	0.99	0.97	100	0.98	33.48
_Default_LightGBM	0.30	1	0.86	0.91	0.76	100	0.81	28.7
_Default_CatBoost	0.07	1	0.86	0.91	0.76	100	081	119.9
_Default_Neural Network	0.61	0.94	0.76	0.84	0.66	88	0.66	34.81

Note: logloss: cross-entropy loss; AUC: area under the curve; TPR: true-positive rate; MCC: Matthew's correlation coefficient.

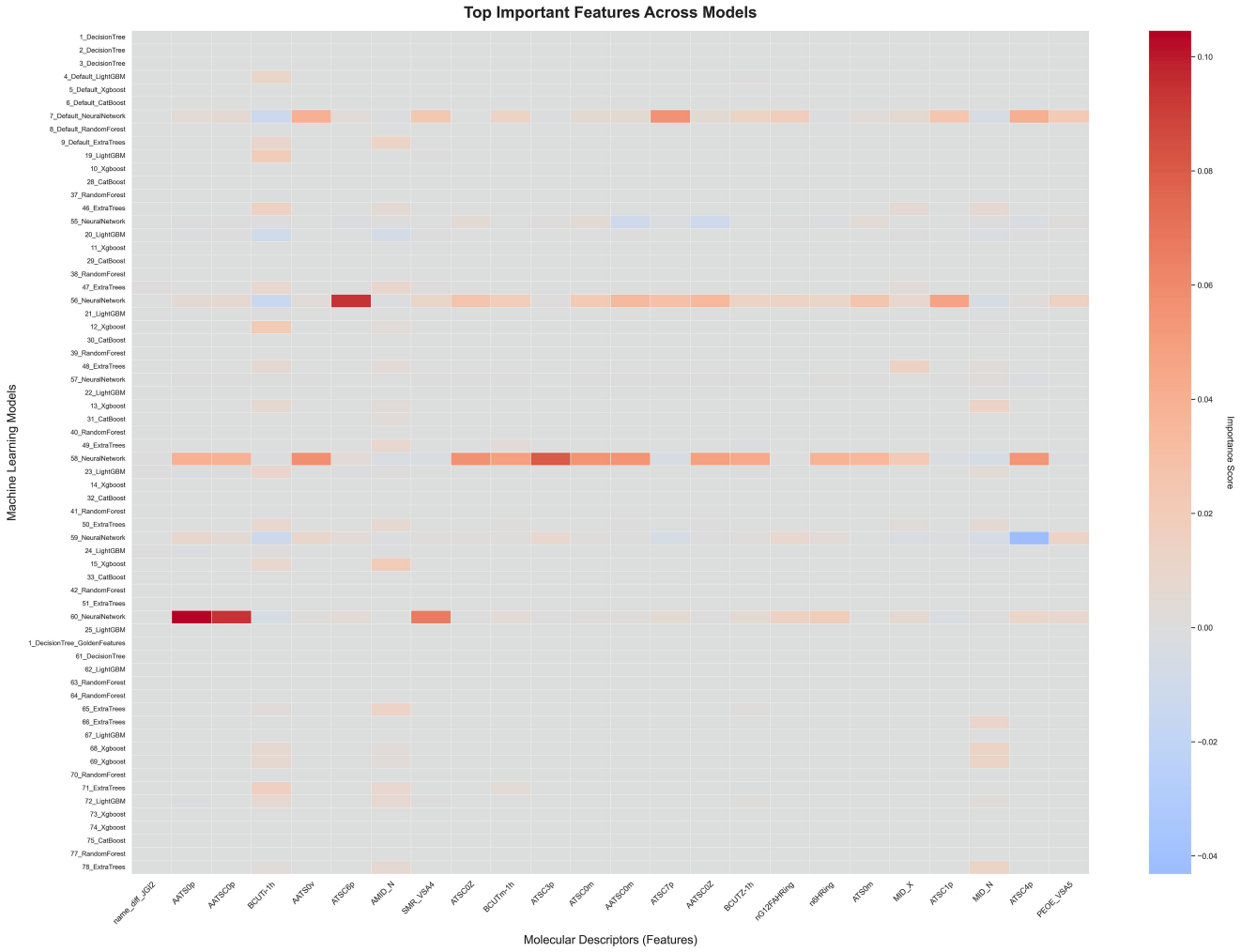

Figure 5. Feature importance heatmap: Top 25 chemical descriptors influencing model predictions

of overfitting. Furthermore, testing the models with noisy data could provide insights into their robustness and generalization capabilities, helping to assess how well they perform under different conditions and ensuring that the models are not overly sensitive to specific data characteristics.

The compounds predicted by our models may be involved in key biological processes related to lifespan extensions, such as cellular stress response, inflammation regulation, and autophagy, all of which have been shown to impact aging and longevity, consistent with findings from López-Otín *et al.* [52], which explores how modulating these pathways can extend lifespan and delay aging-related diseases [52]. The ability of our models to predict compounds that target these pathways is promising, suggesting that these predictions could uncover potential drug candidates with lifespan-extending effects. However, experimental validation is essential to confirm these findings. The predicted compounds should be tested in *in vitro* cell cultures or *in vivo C. elegans* lifespan assays, to assess their true efficacy in extending lifespan or promoting healthy aging. These experiments will provide essential data to substantiate the predicted effects of these compounds and clarify how they interact with biological systems at the molecular level.

Moreover, understanding the molecular mechanisms through which these compounds act on the identified pathways is essential for further development. Investigating how specific compounds regulate autophagy, reduce oxidative stress, or control inflammation can enhance our understanding of their therapeutic potential. Such insights would allow for the design of more targeted, effective geroprotective agents capable of delaying the onset of age-related diseases and extending healthspan.

In the broader context of drug discovery, the computational models developed in this study offer a robust tool for identifying promising compounds for aging-related diseases. By narrowing down the most likely candidates for lifespan extension, ML can help prioritize compounds for experimental testing, thus accelerating the drug development process. However, the accuracy of these *in silico* predictions is contingent on the quality of the input data and the assumptions made during model training. Further optimization of these models, combined with rigorous experimental validation, is essential to ensure that the predicted compounds can translate into viable therapies for aging-related diseases.

4. Conclusion

In conclusion, this study successfully developed a predictive model of lifespan-extending compounds for *C. elegans* from the DrugAge database using the AutoML platform, mljar-supervised. By exploring various ML algorithms, including Decision Trees, Random Forest, Extra Trees, XGBoost, LightGBM, CatBoost, and NNs. Decision Trees and Random Forest models showed perfect metrics, suggesting potential overfitting. In contrast, Extra Trees

and XGBoost demonstrated a better balance between fitting and generalizing, while NNs, LightGBM, and CatBoost with lower scores showed a reduced risk of overfitting. This robust performance highlights the capability of AutoML systems to accelerate aging research by efficiently predicting compounds with potential lifespan-extending properties. Using the DrugAge database and chemical descriptors further validated the model's ability to analyze compound efficacy, setting the stage for future research to discover therapeutic interventions in longevity science. However, experimental validation through biological testing is crucial to verify the predictions made in this study, especially to determine the actual efficacy of these compounds in extending lifespan and enhancing health in aging populations. This research lays the groundwork for integrating computational approaches into drug discovery and aging research, offering the potential to accelerate the identification of new therapeutic candidates that target pathways associated with aging.

Acknowledgment

The authors would like to thank the Department of Botany at Soban Singh Jeena University, S.S.J. Campus, Almora (Uttarakhand), India, for providing the facilities and space needed for this research and recognize the Department of Botany at Pt. Badridutt Pandey Campus Bageshwar, Soban Singh Jeena University, Almora, for their support.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

Author Contribution Statement

Amisha Bisht: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Disha Tewari:** Validation, Visualization. **Kalpana Rawat:** Formal analysis, Data curation. **Priyanka Joshi:** Investigation, Visualization. **Sanjay Kumar:** Validation, Data curation, Supervision. **Subhash Chandra:** Conceptualization, Software, Validation, Data curation, Writing – review & editing, Supervision, Project administration.

References

[1] Bisht, A., Tewari, D., Kumar, S., & Chandra, S. (2024). Network pharmacology-based approach to investigate the molecular targets and molecular mechanisms of *Rosmarinus officinalis* L. for treating aging-related disorders. *Biogerontology*, *25*(5), 793–808. https://doi.org/10.1007/s10522-024-10122-w

[2] Jaul, E., & Barron, J. (2017). Age-related diseases and clinical and public health implications for the 85 years old and over population. *Frontiers in Public Health*, *5*, 335. https://doi.org/10.3389/fpubh.2017.00335

[3] Janssens, G. E., & Houtkooper, R. H. (2020). Identification of longevity compounds with minimized probabilities of side effects. *Biogerontology*, *21*(6), 709–719. https://doi.org/10.1007/s10522-020-09887-7

[4] Beydoun, S., Sridhar, A., Tuckowski, A. M., Wang, E., & Leiser, S. F. (2023). C22 disrupts embryogenesis and extends *C. elegans* lifespan. *Frontiers in Physiology*, *14*, 1241554. https://doi.org/10.3389/fphys.2023.1241554

[5] Jeayeng, S., Thongsroy, J., & Chuaijit, S. (2024). *Caenorhabditis elegans* as a model to study aging and photoaging. *Biomolecules*, *14*(10), 1235. https://doi.org/10.3390/biom14101235

[6] Kim, E. J. E., & Lee, S. J. V. (2019). Recent progresses on anti-aging compounds and their targets in *Caenorhabditis elegans*. *Translational Medicine of Aging*, *3*, 121–124. https://doi.org/10.1016/j.tma.2019.11.003

[7] de Magalhães, J. P., Abidi, Z., Dos Santos, G. A., Avelar, R. A., Barardo, D., Chatsirisupachai, K., . . . , & To, P. K. P. (2024). Human ageing genomic resources: Updates on key databases in ageing research. *Nucleic Acids Research*, *52*(D1), D900–D908. https://doi.org/10.1093/nar/gkad927

[8] Dönertaş, H. M., Fuentealba, M., Partridge, L., & Thornton, J. M. (2019). Identifying potential ageing-modulating drugs in silico. *Trends in Endocrinology & Metabolism*, *30*(2), 118–131. https://doi.org/10.1016/j.tem.2018.11.005

[9] Barardo, D., Thornton, D., Thoppil, H., Walsh, M., Sharifi, S., Ferreira, S., . . . , & de Magalhães, J. P. (2017). The DrugAge database of aging-related drugs. *Aging Cell*, *16*(3), 594–597. https://doi.org/10.1111/acel.12585

[10] Kokudeva, M., Vichev, M., Naseva, E., Miteva, D. G., & Velikova, T. (2024). Artificial intelligence as a tool in drug discovery and development. *World Journal of Experimental Medicine*, *14*(3), 96042. https://doi.org/10.5493/wjem.v14.i3.96042

[11] Zielinska, E. (2024). *Machine learning uses lung cancer scans to predict heart damage*. UK: Medical Press. Retrieved from: https://medicalxpress.com/news/2024-06-machine-lung-cancer-scans-heart.html

[12] Zhou, Y., Hou, Y., Hussain, M., Brown, S. A., Budd, T., Tang, W. W., . . . , & Cheng, F. (2020). Machine learning-based risk assessment for cancer therapy–related cardiac dysfunction in 4300 longitudinal oncology patients. *Journal of the American Heart Association*, *9*(23), e019628. https://doi.org/10.1161/JAHA.120.019628

[13] Hashemi, S., Vosough, P., Taghizadeh, S., & Savardashtaki, A. (2024). Therapeutic peptide development revolutionized: Harnessing the power of artificial intelligence for drug discovery. *Heliyon*, *10*(22), e40265. https://doi.org/10.1016/j.heliyon.2024.e40265

[14] Sharma, A., Lysenko, A., Jia, S., Boroevich, K. A., & Tsunoda, T. (2024). Advances in AI and machine learning for predictive medicine. *Journal of Human Genetics*, *69*(10), 487–497. https://doi.org/10.1038/s10038-024-01231-y

[15] 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

[16] Park, J., Beck, B. R., Kim, H. H., Lee, S., & Kang, K. (2022). A brief review of machine learning-based bioactive compound research. *Applied Sciences*, *12*(6), 2906. https://doi.org/10.3390/app12062906

[17] Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine learning in drug discovery: A review. *Artificial Intelligence Review*, *55*(3), 1947–1999. https://doi.org/10.1007/s10462-021-10058-4

[18] Zhavoronkov, A., Mamoshina, P., Vanhaelen, Q., Scheibye-Knudsen, M., Moskalev, A., & Aliper, A. (2019). Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Research Reviews*, *49*, 49–66. https://doi.org/10.1016/j.arr.2018.11.003

[19] Raj, R., Kannath, S. K., Mathew, J., & Sylaja, P. N. (2023). AutoML accurately predicts endovascular mechanical thrombectomy in acute large vessel ischemic stroke. *Frontiers in Neurology*, *14*, 1259958. https://doi.org/10.3389/fneur.2023.1259958

[20] Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, *104*, 101822. https://doi.org/10.1016/j.artmed.2020.101822

[21] Kapsiani, S., & Howlin, B. J. (2021). Random forest classification for predicting lifespan-extending chemical compounds. *Scientific Reports*, *11*(1), 13812. https://doi.org/10.1038/s41598-021-93070-6

[22] Statzer, C., Jongsma, E., Liu, S. X., Dakhovnik, A., Wandrey, F., Mozharovskyi, P., . . . , & Ewald, C. Y. (2021). Youthful and age-related matreotypes predict drugs promoting longevity. *Aging Cell*, *20*(9), e13441. https://doi.org/10.1111/acel.13441

[23] Barardo, D. G., Newby, D., Thornton, D., Ghafourian, T., de Magalhães, J. P., & Freitas, A. A. (2017). Machine learning for predicting lifespan-extending chemical compounds. *Aging*, *9*(7), 1721–1737. https://doi.org/10.18632/aging.101264

[24] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*, 1–14. https://doi.org/10.1186/1758-2946-3-33

[25] Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, *10*(1), 4. https://doi.org/10.1186/s13321-018-0258-y

[26] Ilemobayo, J. A., Durodola, O., Alade, O., Awotunde, O. J., Olanrewaju, A. T., Falana, O., . . . , & Edu, O. E. (2024). Hyperparameter tuning in machine learning: A comprehensive review. *Journal of Engineering Research and Reports*, *26*(6), 388–395. https://doi.org/10.9734/jerr/2024/v26i61188

[27] Ferreira, R. D. A., Teixeira, G., & Peternelli, L. A. (2021). Kennard-Stone method outperforms the Random Sampling in the selection of calibration samples in SNPs and NIR data. *Ciência Rural*, *52*(5), e20201072. https://doi.org/10.1590/0103-8478cr20201072

[28] Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research*, *5*(4), 2094–2097.

[29] Chawla, A. (2024). Random forest vs. ExTra trees. *Daily Dose of Data Science*. Retrieved from: https://blog.dailydoseofds.com/p/random-forest-vs-extra-trees

[30] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[31] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[32] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . , & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.

[33] Ramalingam, K., Yadalam, P. K., Ramani, P., Krishna, M., Hafedh, S., Badnjević, A., . . . , & Minervini, G. (2024). Light gradient boosting-based prediction of quality of life among oral cancer-treated patients. *BMC Oral Health*, *24*(1), 349. https://doi.org/10.1186/s12903-024-04050-x

[34] Geeitha, S., Ravishankar, K., Cho, J., & Easwaramoorthy, S. V. (2024). Integrating cat boost algorithm with triangulating feature importance to predict survival outcome in recurrent cervical cancer. *Scientific Reports*, *14*(1), 19828. https://doi.org/10.1038/s41598-024-67562-0

[35] Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of artificial neural networks and deep learning. In O. A. Montesinos López, A. Montesinos López & J. Crossa (Eds.), *Multivariate statistical machine learning methods for genomic prediction* (pp. 379–425). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-89010-0_10

[36] Barbudo, R., Ventura, S., & Romero, J. R. (2023). Eight years of AutoML: Categorisation, review and trends. *Knowledge and Information Systems*, *65*(12), 5097–5149. https://doi.org/10.1007/s10115-023-01935-1

[37] Salehin, I., Islam, M. S., Saha, P., Noman, S. M., Tuni, A., Hasan, M. M., & Baten, M. A. (2024). AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, *2*(1), 52–81. https://doi.org/10.1016/j.jiixd.2023.10.002

[38] Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1). 37–63.

[39] Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4. https://doi.org/10.1186/s13040-023-00322-4

[40] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

[41] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, *16*(5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

[42] Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*(1), 5979. https://doi.org/10.1038/s41598-022-09954-8

[43] Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 5234–5243.

[44] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13. https://doi.org/10.1186/s12864-019-6413-7

[45] Bisht, A., Tewari, D., Rawat, K., Rawat, S., Almoyad, M. A. A., Wahab, S., . . . , & Chandra, S. (2024). Computational screening of matrix metalloproteinase 3 inhibitors to counteract skin aging from phytochemicals of *Nelumbo*

nucifera Gaertn. *Theoretical Chemistry Accounts*, *143*(6), 51. https://doi.org/10.1007/s00214-024-03125-w

[46] Wang, G., Li, C., Tang, F., Wang, Y., Wu, S., Zhi, H., . . . , & Zhang, J. (2023). A fully-automatic semi-supervised deep learning model for difficult airway assessment. *Heliyon*, *9*(5), e15629. https://doi.org/10.1016/j.heliyon.2023.e15629

[47] Li, T., Fong, S., Wu, Y., & Tallón-Ballesteros, A. J. (2020). Kennard-Stone balance algorithm for time-series big data stream mining. In *2020 International Conference on Data Mining Workshops*, 851–858. https://doi.org/10.1109/ICDMW51313.2020.00122

[48] Svitliishyi, M., Kovalishyn, V., Startseva, Y., & Hodyna, D. (2023). Application Kennard-Stone algorithm for QSAR studies. In *Proceedings of the13thInternational Scientific and Practical Conference: Scientific Horizon in the Context of Social Crises*, *144*, 534–539.

[49] Ribeiro, C., Farmer, C. K., de Magalhães, J. P., & Freitas, A. A. (2023). Predicting lifespan-extending chemical compounds for *C. elegans* with machine learning and biologically interpretable features. *Aging*, *15*(13), 6073–6099. https://doi.org/10.18632/aging.204866

[50] Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. Germany: Self-published.

[51] Jain, N., Jhunthra, S., Garg, H., Gupta, V., Mohan, S., Ahmadian, A., . . . , & Ferrara, M. (2021). Prediction modelling of COVID using machine learning methods from B-cell dataset. *Results in Physics*, *21*, 103813. https://doi.org/10.1016/j.rinp.2021.103813

[52] López-Otín, C., Pietrocola, F., Roiz-Valle, D., Galluzzi, L., & Kroemer, G. (2023). Meta-hallmarks of aging and cancer. *Cell Metabolism*, *35*(1), 12–35. https://doi.org/10.1016/j.cmet.2022.11.001

How to Cite: Bisht, A., Tewari, D., Rawat, K., Joshi, P., Kumar, S., & Chandra, S. (2025). Exploration of Various Supervised Machine Learning Algorithms for Predictive Modeling of *Caenorhabditis elegans* Lifespan-Extending Compound Dataset. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN52024571