# Comparative Analysis of Machine Learning Models for Multiclass Classification: Application to Biological and Survey Data

Jinwen Lei
*z5435879*
*Jinwen Lei makes a submission on Edstem.

Jiacheng Ling
z5529848

*Abstract*—**This project aims to address two distinct classification problems using machine learning. The first task is to classify four age groups of abalone and the second one is focus on contraceptive method choice dataset. We present a series of data visualization and analysis to both datasets. Based on abalone dataset, we compare performance of different models. Then apply two top-performing models to the contraceptive method choice dataset.**

*Index Terms*—**Abalone, Contraceptive methods, Classification, Machine learning.**

## I. Introduccion

Text classification refers to train models to learn features from rows of data, and predicted the target category. Multiple models had been published in recent decades years. Models such as decision tree, random forest(RF) and eXtreme Gradient Boosting (XGBoost). Traditional methods performed better than deep learning models in the case of limited data and computational resources as traditional methods have simpler complexity than deep learning models.[1] In recent decades years, neural network had gained significant improvement on data classification. [2] Such methods were widely regarded as the most powerful tools, and most people used them in data processing projects.[3] Neural network can suffer from over-training and over-fitting, and also required larger size of dataset and better computational resources. In addition, neural network also need more complex hyperparameter tuning.[4][5]

During previous years, methods such as decision tree and random forest have shown advanced ability in multi-class classification tasks. Significant utilization has been implemented in biology and social sciences.[6][7] However, there still lack of studies in compare impact of hyperparameter optimization and regularization techniques across multiple model types, specifically in different datasets with distinct features.

In this project, we aim to use two datasets that are abalone and contraceptive methods as resources of multi-class classification to train different models. Then, compare performance of these models to discuss the result of neural network classifiers. Predict abalone age and contraceptive methods are time consuming and Labor intensive. Thus, explore more optimal methods such as neural network is essential. This project is focus on six different models that are decision tree, random forest, XGBoost, GradientBoost, simple neural network with Adam and simple neural network with SGD. Additionally, we examine regularization techniques in neural networks, testing combinations of dropout rates and L2 regularization to enhance model robustness. These methods are used to defined age groups for abalone and predicting contraceptive choices as categorical classes.

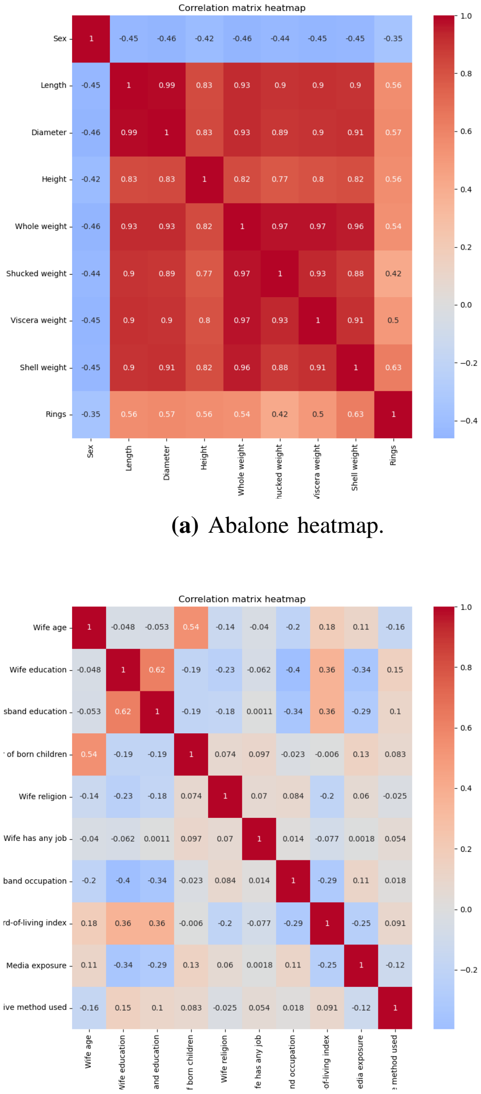
The rest report will discuss datasets, methodologies and results. Mainly focus on data features, models and parameter tuning, evaluation metrics and result comparation. Lastly, we will discuss current achievement, limitations and future work.

## II. methodology

### A. Dataset

In this section, two datasets are used. The firs one is abalone dataset which contains eight physical data such as shell weight, Length, Viscera weight et al. Moreover, the dataset also contains one column that is the actual age of abalone. In this project, the age of abalone are grouped into four classes: 0 to 7 years as Class 1, 8 to 10 years as Class 2, 11 to 15 years as Class 3 and Class 4 for elder than 15 years old. Contraceptive methods dataset use self-information of couples as features, including wife's age and education, husband's age and education and standard-of-living index et al. Contraceptive method used is the target of classification task, this section divide contraceptive method used into three groups, that are 1 for No-use, 2 for long-term and 3 for short-term.

*1) Abalone:* This dataset contains non-digital data which is in the 'Sex' column. The Sex column use F, M and I to present female, male and infant abalone. Thus, we need to convert M into 0, F into 1 and I into 2. In addition, the original dataset does not contains column names so we add the names manually. In order to predict age group, transfer the ages of abalone into corresponding class, and add this data into the additional column 'Class'. According to Fig. 1a, we can observe that almost all features have relatively low positive correlation with Ring age, while only Sex has negative correlation at -0.35. The highest positive correlation feature is Shell weight at 0.63. Other features like Diameter, Length and Height show weaker correlations with "Rings," suggesting they may not independently predict age well. We can also observe high correlations among features, this phenomenon



**(a)** Abalone heatmap.

**(b)** Contraceptive methods heatmap.

**Fig. 1.** Heat map for two datasets: (a) Heat map of Abalone dataset. (b) Heat map of Contraceptive methods dataset.

makes some features less informative in a model as it leads to redundancy in the data.

Fig. 2 illustrates the class distribution of the entire dataset. The distribution shows an obvious class imbalance problem between each age group. As Fig. 2 shows, Class 2 contains the most number of samples that can reach 1891, while Class 4 is the least one that have 261 samples. This indicates that the overall age distribution skews towards younger age groups. Thus, models probably be frustrated to classify abalone which over 15 years old.

*2) Contraceptive methods:* Due to Fig. 1b, all features have low correlation with contraceptive methods. The highest observed is with ”Media exposure” (around 0.12). This suggests that social characteristics may not be able to strongly predict comprehensive choice of people.

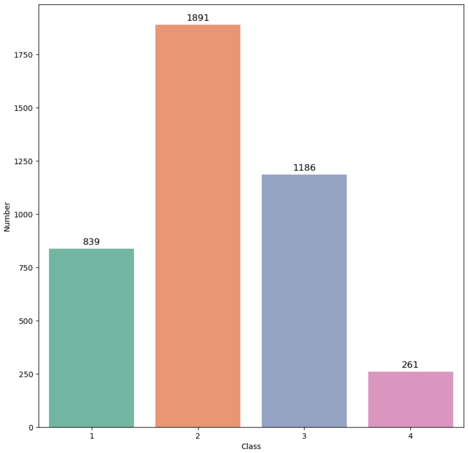


**Fig. 2.** Age distribution of Abalone.

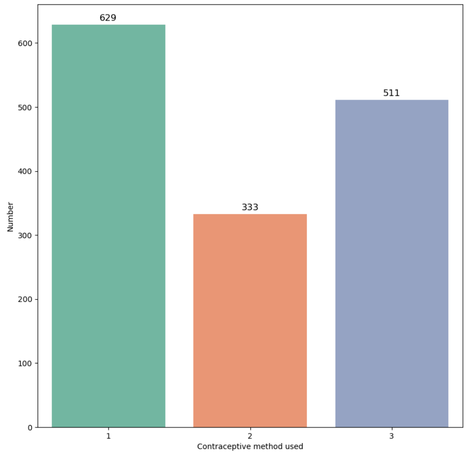


**Fig. 3.** Distribution of contraceptive methods used.

This dataset also shows relatively obvious imbalance problem between each class. As Fig. 3 shows, majority people prefer not to use contraceptive methods, which can reach 629 samples. The second popular one is short-term contraceptive method which is 511 samples in total. However, samples in Long-term contraceptive method shows the least distribution that is 333. This may makes classifiers tend to predict more common categories, which may affect the performance of the model. Specifically, models are probably frustrated to classify Long-term category as it has the least sample numbers to efficiently train the model.

| Age range | Corresponding class |
|---|---|
| 0 - 7 years | 1 |
| 8 - 10 years | 2 |
| 11 - 15 years | 3 |
| Greater than 15 years | 4 |

**Tab. I.** Caption

### *B. Materials*

This project is based on python environment, using sklearn to construct the trees and tensorflow to build neural network. Using UNSW terminal environment as running resources.

### *C. Models and experiental settings*

The models are used in this project are decision tree, random forest, XGBoost, Gradient boosting and Adam/SGD in Neural Network. Firstly, split data into two parts, use random split to contrast 60% of overall data as training set, and 40% as testing.

Fig. 4 presents the entire structure of decision tree in this project. Select four of important nodes, these nodes can be written in IF THEN rule as:

IF (Shell weight $\leq$ 0.11 AND Diameter $\leq$ 0.25 AND Shell weight $>$ 0.05) THEN Class = 1

IF (Shell weight $\leq$ 0.11 AND Diameter $\leq$ 0.25 AND Shell weight $\leq$ 0.05) THEN Class = 2

IF (Shell weight $>$ 0.44 AND Shucked weight $>$ 0.63) THEN Class = 3

IF (Shell weight $>$ 0.44 AND Shucked weight $\leq$ 0.63) THEN Class = 4

The features used by the tree are physical features of abalone. The decision tree model uses the binary split criterion and takes information gain as the splitting criterion. Final output is 4 age groups based on Tab. I. Combined with Fig. 4 analysis, we can infer that decision tree shows obvious imbalanced predicted results as it has more and more detailed predictions for Class 1 and Class 2. This indicates that decision tree has better performance on young abalone, and may has under-fitting problem when dealing with elderly abalone especially in Class 4. Such situation may happened because the dataset has more data of class 1 and class 2, which is possibly causes impacts on the model's performance.

Random forest uses randomization to build large number of decision tree, this algorithm is developed by Breiman. It use voting for classification problems or use averaging for regression problems. Also, it can achieve better performance as it can discover meaningful interactions and non-linear effect. [8]

Gradient boosting algorithm can suffer form over-fitting when the iterative process is not properly regulized. Several hyper-parameters are used considered to adjust the performance of Gradient boosting, including learning rate, maximum depth, minimum number of samples et al.[9] XGBoost is a powerful variants of tree boosting algorithms gives a novel solution to the real-world use-cases. [10][11] In order to search the better combination of parameters for each algorithm, we implement HO to models including Random Forest, XGBoost and Gradient Boosting and Decision Tree. We use hyper-parameter search(randomly set the depth, the number of trees .etc) to get the best model for each algorithm within 5 time search. Then, use the best model of each algorithm, train the dataset 5 times, calculating the mean accuracy of each algorithm(HO).

## III. RESULTS

### *A. PART A*

In this project, for Part A , we implement 5 different types of models to process the abalone dataset, respectively decision tree, random forest, XGBosst, Gradient Boosting and simple neural network.

First, we implement hyper-parameter optimization(HO) in these models except neural network. Find the model with the best performance in 5 times hyper-parameter search.

Second, we use the best model to train the dataset in 5 times, calculate the mean accuracy of training and test. As for the neural network, we implement two different types of optimization namely adam and sgd.

Tab. II is the results which shows the performance of the 5 kinds of models in training the abalone dataset.

| | Training accuracy | Test accuracy |
|---|---|---|
| Decision Tree | 0.6424 | 0.6108 |
| Random Forest | 0.8206 | 0.6345 |
| XGBoost | 0.7509 | 0.6293 |
| Gradient Boosting | 0.8942 | 0.6114 |
| Neural network with sgd/adam | 0.6131/0.6342 | 0.6112/0.6333 |

**Tab. II.** Results of training

According to the data shown in Tab. II, random forest, XGBoost and Gradient boosting shows noticeable overfitting as there exist obvious gap between training and testing results, which can beyond 0.1. This is because the depth of each model is deep. These three models using over-fitting to increase the accuracy, however, due to the quality of the dataset, using the three models (HO) still can't reach a high test accuracy. In Fig. 1a, we can infer that the correlation between features is high. As integrated models will generate multiple decision trees or weak classifiers, features with high correlation will lead to increasion of models' complexity. Although models can effectively learn different combination of features, but they are frustrated to generalize what they have learnt to testing set. In contrast, simple structure of algorithm can avoid this problem as they are not able to learn all the features, so significant feature redundancy information is avoided. In this case, decision tree and neural network do not have over-fitting problems.

We compare the performance of neural network (with Adam optimizer) with or without L2 regularization and dropout in training the dataset and put results in Tab. III. The table shows the performance in different combinations of L2 regularization and dropout, 0 refers to the corresponding method does not use.

The overall performance of neural network does not change noticeably. Low L2 with a higher Dropout, over-fitting can be

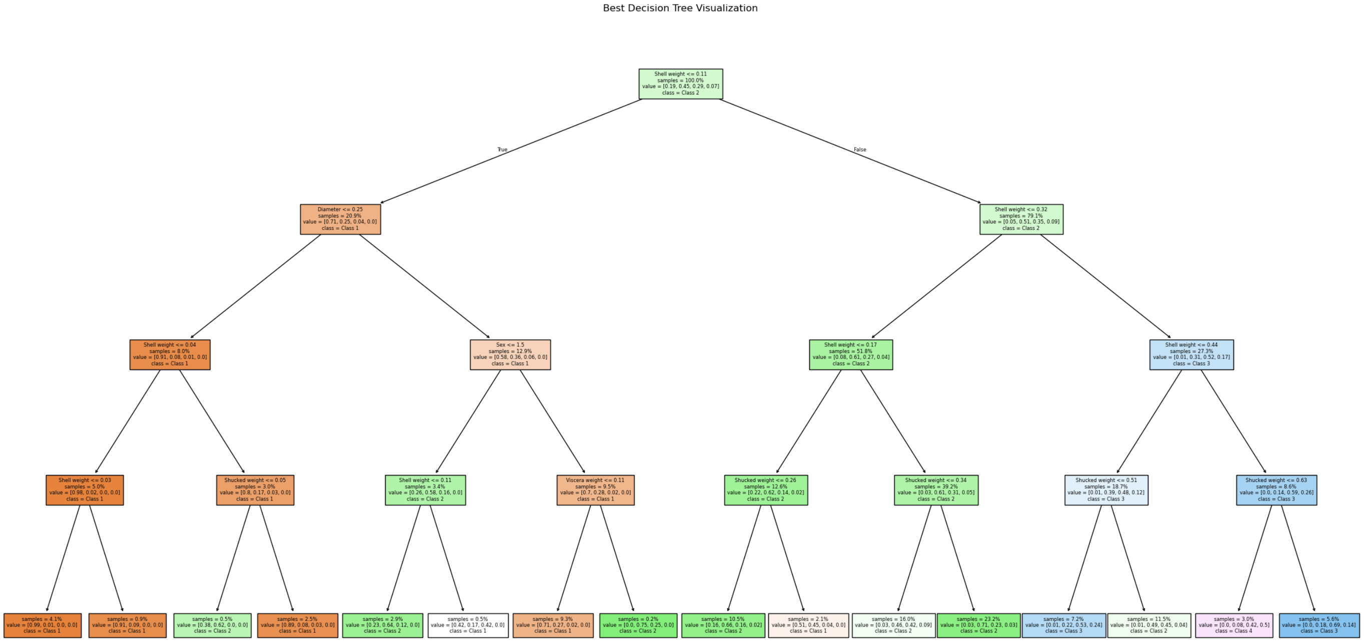


**Fig. 4.** Visualization of decision tree

| | L2 | Dropout | Training accuracy | Test accuracy |
|---|---|---|---|---|
| 1 | 0 | 0 | 0.6345 | 0.6421 |
| 2 | 0.01 | 0.3 | 0.5870 | 0.6188 |
| 3 | 0.0001 | 0.5 | 0.6553 | 0.6475 |

**Tab. III.** Using L2 and dropout

effectively reduced. In contrast, high L2 will lead to underfitting on neural network.

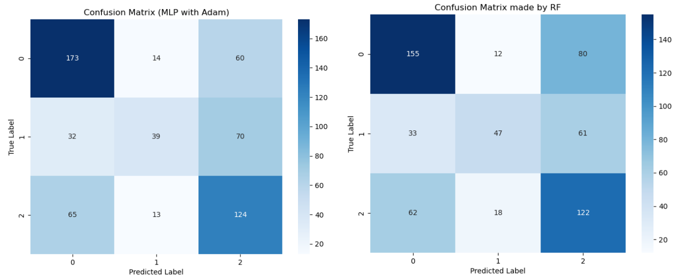
High correlation between multiple features is probably leads to this problem as the model will rely heavily on the redundancy between features caused by high correlation. L2 regulation is not able to reduce such redundancy, and regulation methods are usually used for solve over-fitting. Thus, L2 cannot optimize the performance of neural network. As Fig. I shows, the abalone dataset exists obvious imbalanced class problem. In this case, the generalization ability could be limited, and generalization methods cannot effectively balance the weight of each class. So the results of regularization model would not change significantly.

### B. PART B

It's clear to see that the Random Forest and the NN with adam optimizer have the best performance. Thus we use the two models to do the training in this part. Tab. IV shows the result of using NN:

| Model | F1 Score | ROC | Training Accuracy | Test Accuracy |
|---|---|---|---|---|
| NN | 0.5564 | 0.7577 | 0.6602 | 0.5492 |
| Random Forest | 0.5438 | 0.7452 | 0.5663 | 0.5695 |

**Tab. IV.** Comparison of NN and Random Forest results



**(a)** Confusion matrix of Neural Network

**(b)** Confusion matrix of Random Forest

**Fig. 5.** Comparison of Confusion Matrices for Neural Network and Random Forest. Class 0 represents method of 'No-use', Class 1 represents 'Long-term' method and Class 2 represents 'Short-term' method.

As Fig. 5a shows, we can know that the number of misclassified as class 2 is 60 in class 1 and 70 in class 1, and the number of misclassifications to class 2 is also higher than others. Due to Fig. 3, data imbalanced is significant in this dataset, which will lead to the model tends to predict class with more samples.

Use random forest instead of neural network, the confusion matrix is shown as Fig. 5b. Similar to neural network, he number of misclassifications to class 0 and class 2 is higher than class 1. Random forest also be impacted by data imbalance obviously. This shows that both models face similar challenges in distinguishing between the two categories. Compared with neural network, random forest relies more on splitting effect of features, so the improvement of model performance is limited when the relationship between features and categories is weak.

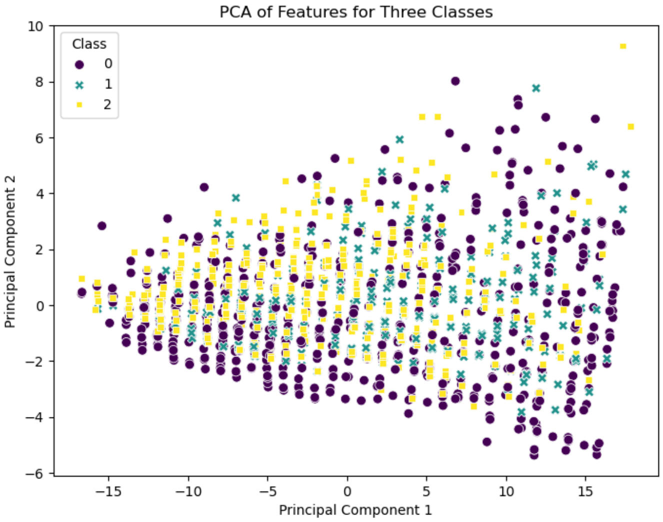


**Fig. 6.** Three classes distribution on feature map.

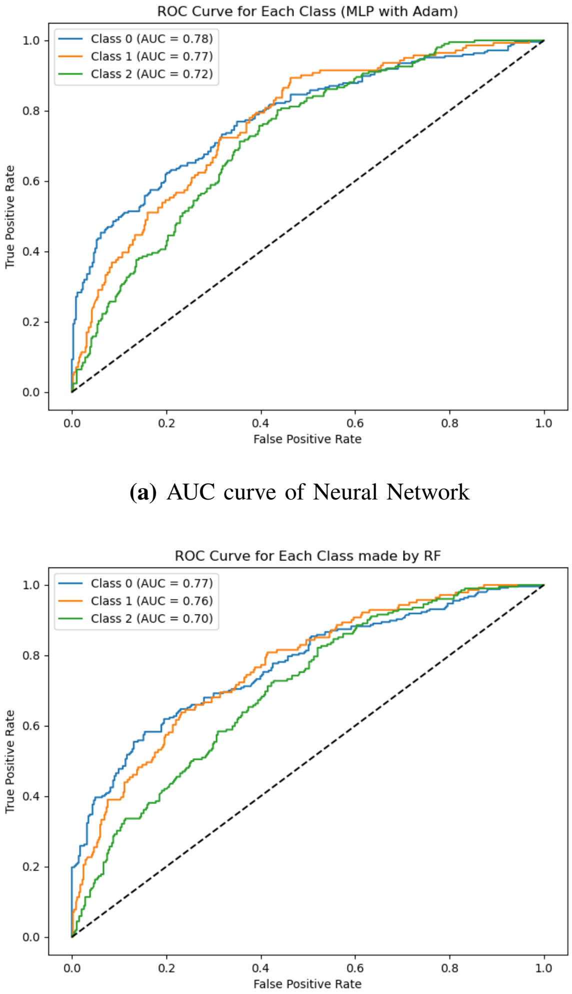
Fig. 6 illustrates the reason of this results, the overlap of the three features is significant high, especially for class 1 and class 2. Class 0 is briefly better than the other two classes, the distribution of class 1 and 2 are almost complete overlap while class 0 can have some outliers. So models can have better ability on classify class 0 than the others. In addition, models are more likely to classify samples into class 2 as the distribution of this class almost includes the other two classes.

As shown in Fig. 7a. The MLP model shows relatively good classification performance for Class 0 , as indicated by the higher AUC values and ROC curves for these classes.The model's ability to distinguish Class 2 and class 1 is weaker, as shown by its lower ROC curve, suggesting possible misclassifications for this class. And AUC-ROC curve is shown in Fig. 7b. The Random Forest model shows good discrimination for Class 0 and Class 1, as indicated by relatively high AUC values.The model's performance on Class 2 is weaker.

As Fig. 7 displayed, neural network and random forest shows very similar results on this dataset. Especially the AUC curve of class 1 and class 2 are very close, which indicates that both models have similar ability in classifying these two classes.

One reason could be the high overlap between three classes. As mentioned before and in Fig. 6, models can be frustrated to process this dataset. Thus, the models cannot perform well on this dataset.

Random Forest perform slightly better than neural network due to its number of decision trees, even use the default random forest can get a good result. When it comes to neural network, though neural network depends on HO, just use optimizer and L2 with dropout can still reach a high accuracy.If spend more time on hyper-parameter search, the model can get much higher accuracy.Thus, neural network is not only a good model, but also a potential model.



**(a)** AUC curve of Neural Network

**(b)** AUC curve of Random Forest

**Fig. 7.** Comparison of AUC curves for Neural Network and Random Forest

## IV. DISCUSSION

In this project, we implemented different machine learning models to predict the age of abalone from physical measurements and to classify contraceptive methods. Evaluate performance of models by multiple parameters including accuracy, F1 score and AUC, also with variety kinds of visualization methods.

Our results indicates that all models cannot reach high accuracy in this multi-class classification project. But the performance of models still varies as over-fitting problem took place on random forest, XGBoost and Gradient boost.

One obvious limitation is the data quality. Imbalanced problem between classes and correlation could be the biggest hinder of achieving high results.

## V. CONCLUSION

The major contribution of our project was we explored a novel way to process biological and social survey data, which was based on machine learning models instead of manual analyze. In future studies, we expect to explore more advanced methods that can significant improve the performance in multi-class classification tasks. For example, improving the quality of data and improve a method to solve class imbalanced problem.

## REFERENCES

[1] Li, Q., Peng, H., Li, JX., *et al.*, 2022, 'A Survey on Text Classification: From Traditional to Deep Learning', *Association for Computing Machinery*, vol.13, no.2, ⟨10.1145/3495162⟩.

[2] Al-Saffar, AAM., Tao, H. & Talab, MA., 2017, 'Review of deep convolution neural network in image classification,' *International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pp. 26-31, ⟨10.1109/ICRAMET.2017.8253139⟩.

[3] Féraud, R., Clérot, F., 2002, 'A methodology to explain neural network classification,' *Neural Networks*, vol.15, no.22, pp. 237-246, ⟨https://doi.org/10.1016/S0893-6080(01)00127-7⟩.

[4] Livingstone, DJ., Manallack, DT. & Tetko, IV., 1997, 'Data modelling with neural networks: Advantages and limitations' *Journal of computer-aided molecular design*, vol.11, no.2, pp.135-142, ⟨10.1023/a:1008074223811⟩.

[5] Manallack, DT., Ellis, DD. & Livingstone, DJ., 1994, 'Analysis of linear and nonlinear QSAR data using neural networks' *Journal of Medicinal Chemistry*, vol.37, no.22, pp.3758-3767.

[6] Sahin, E.,Saul, CJ., Ozsarfati, E. *et al.*, 2018, 'Abalone Life Phase Classification with Deep Learning', *International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pp.163-167, ⟨10.1109/ISCMI.2018.8703232⟩.

[7] Wang, S., Gao, JZ., Lin, H. *et al.*, 2019, 'Dynamic Human Behavior Pattern Detection and Classification', *IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 159-166, ⟨10.1109/BigDataService.2019.00028⟩.

[8] Rigatti, SJ., 2017, 'Random Forest', *JOURNAL OF INSURANCE MEDICINE*, vol.41, no.1, pp. 31-39, ⟨https://doi.org/10.17849/insm-47-01-31-39.1⟩.

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

[10] Chen, T., Li, H., Yang, Q. & Yu, Y., 2013, 'General functional matrix factorization using gradient boosting', *Proceeding of 30th International Conference on Machine Learning(ICML'13)*, vol.28, no.1, pp. 436-444.

[11] Chen, T., Singh, S., Taskar, B. & Guestrin, C., 2015, 'Efficient second-order gradient boosting for conditional random fields', *Proceeding of 18th Artificial Intelligence and Statistics Conference (AISTATS'15)*, no.1.