International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6984
***Article Received**: 01 April 2023 **Revised**: 10 May 2023 **Accepted**: 25 May 2023*

Reinforcement of the Bank Loan Model using the Feature Selection Method of Machine Learning

Noopur Goel[1], Durgesh Kumar Singh[2]
[1]Head, Department of Computer Applications
VBS Purvanchal University
Jaunpur, India
noopurt11@gmail.com
[2]Research Scholar, Department of Computer Applications
VBS Purvanchal University
Jaunpur, India
durgeshsingh111@gmail.com

Abstract— Does feature selection and machine learning (ML) guarantee the effectiveness of the bank credit system model? This article aims to analyze this problem. In fact, in finance, expert-based credit risk models still dominate. In this study, we establish a new benchmark using consumer data and present machine learning methods. A risk prediction that is as accurate as possible is an important requirements for credit scoring models. In addition, regulators expect that the models should to be auditable and transparent. As a result, the superior predictive power of contemporary machine learning algorithms cannot be fully utilized in credit scoring because very simple predictive models, such as several ML classifiers, are still widely used. As a result, significant potential is missed, increasing reserves or the number of credit defaults. A framework for comparing scores before and after feature selection machine learning models that are transparent, auditable, and explainable is presented in this article, as well as the various dimensions that need to be taken into consideration in order to make credit scoring models understandable. In accordance with this framework, we give an overview of the models which demonstrate how it can be used in credit scoring, and compare the results to scorecards' interpretability. The model presented demonstrates that machine learning techniques can maintain their ability to enhance predictive power while still maintaining a comparable level of interpretability.

Keywords- Bank credit, Machine learning, Feature selection, Ensemble, Voting, Stacking, ROC (AUC) curve.

I. INTRODUCTION

Credit scoring systems aim to satisfy a minimum-loss principle for the sustainability of lending institutions by providing clients with a probability of default [1]. As a result, a credit scoring system aids in the decision-making process for credit applications, manages credit risks, and has an impact on the number of non-performing loans that are likely to result in bankruptcy, a financial crisis, or environmental sustainability. Although credit officers or expert-based credit scoring models have been determining whether borrowers can meet their requirements over the past ten years, this has changed over time due to technological advancements. In order to lessen each lending institution's potential loss, this modification necessitates the establishment of an automated credit decision-making system that can avoid opportunity losses or credit losses [2]. Because of this, the increasing number of financial services that do not involve a human being has made it increasingly important in recent years to use automated credit scoring. To put it another way, an accurate credit scoring model is needed for modern lending institutions to use technology and automation to cut down on operating costs. Although developing an effective model for determining a client's creditworthiness is extremely challenging, machine learning is now an essential component of credit scoring applications [3]. It is stated that the utilization of intricate algorithms in the context of the application of machine learning to financial services might lead to a lack of transparency for customers. Provide consumers, auditors, and supervisors with an explanation of a credit score and the resulting credit decision when challenged when using machine learning to assign credit scores and make credit decisions is typically more challenging. As a result, model developers face an increasing demand for tools to comprehend what their models have learned. Discriminant analysis, support vector machines, logistic regression, genetic algorithm, fuzzy logic, neural networks, Bayesian networks, decision trees, ensemble, and hybrid methods are just a few examples of ML algorithms that have been used in prior research. Chi-2 test, evolutionary feature selection with correlation, genetic algorithm, and hybrid feature-selection methods are just a few of the feature-selection approaches that numerous authors have proposed for credit scoring [4]. Predictions are based, in essence, on the characteristics of a phenomenon that are captured by machine learning. However, these characteristics may not only describe the intended

***Article Received**: 01 April 2023 **Revised**: 10 May 2023 **Accepted**: 25 May 2023*

phenomenon but may also be instructive in describing other phenomena, feature categories, or classes. This article compares the credit scoring before and after feature selection using a variety of machine learning techniques. We provide an extensive comparison of basic machine learning methods and feature selection-based models. The first method Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), Logistic Regression (LR), Random Forest (RF), AdaBoost, Decision Tree (DT), k-Nearest Neighbors (k-NN), Gradient Boosting Classifier (GBC) as base classifier and ensemble classifier with voting (hard) and stacking to measure performance on bank loan test data. Another six feature selection methods: Pearson, Chi-2, Recursive Feature Elimination (RFE), Logistic Regression (LR), Random Forest (RF) and LightGBM (LGBM) are used for the extraction of important features. These feature selection methods provide mixed features for performance measurement, and the second method is used to evaluate the performance of the above classifiers after feature selection (mixed features) [5]. For both methods, the most representative features are selected for effective modeling and they are compared with the results obtained from a test dataset with all features. Precision, recall, F1-score, confusion matrix and ROC (AUC) curve are used to validate the results obtained by these two methods.

The sections of this paper are as follows: Algorithms, tools, and techniques, as well as their significance, are discussed in section 2's background section. In section 3, the experimental diagram and its explanations were used. The information about the attributes and dataset has been described in section 4 of the experimental setup. In Section 5, the results of the experiments were discussed. Sections 6 and 7 respectively discuss the experiment's discussion and conclusion.

II. BACKGROUND

For a particular stage of the creditworthiness evaluation pattern's development, many of the previously proposed machines learning models are described here. However, due to the system's lack of adequate controls and the models' reliance on some short-term pattern emphasis, which may have an effect on the models' performance quality over time, certain risks are critical to these models.

A. Machine learning classifiers

- *Support vector machine (SVM)*

SVM is a supervised learning machine algorithm that can be used to solve regression and classification problems at the same time [6]. The SVM algorithm's objective is to find the most effective line or decision boundary for classifying n-dimensional space so that the new data point can be easily placed in the appropriate category in the future. A hyperplane is the name given to this best decision boundary. The extreme points or vectors that aid in the creation of the hyperplane are selected by SVM. The algorithm is referred to as a support vector machine because these extreme cases are referred to as support vectors. The number of features determines the hyperplane's dimension. The hyperplane is just a line if there are two input features. The hyperplane transforms into a two-dimensional plane when there are three input features. When the number of features is greater than three, it becomes difficult to imagine. The formula below can be used to locate an SVM classifier:

$$f(w,b) = \left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0,1-y_i(w^T x_i - b)\right)\right] + \lambda\ ||w||^2$$

Where, w and b are both convex functions of f.

- *Gaussian Naïve Bayes (GNB)*

A variant of Naive Bayes that supports continuous data and follows the Gaussian normal distribution is known as GNB [7]. An assumption that is frequently made when working with continuous data is that the continuous values associated with each class are distributed in a normal (or Gaussian) manner. It is presumed that the features have a likelihood of:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i-\mu_y)^2}{\sqrt{2\sigma_y^2}}\right)$$

Continuous valued features and models are accepted by Gaussian Naive Bayes as belonging to a Gaussian (normal) distribution. To define such a distribution, all that is required to fit this model is the mean and standard deviation of the points within each label.

- *Logistic Regression (LR)*

Classification and predictive analytics frequently make use of the logistic regression model, also referred to as the logit model [8]. Based on a particular dataset of independent variables, LR estimates the probability that an event, such as voting or not voting, will take place. The dependent variable is limited to values between 0 and 1, as the outcome is a probability. A logit transformation is applied to the odds in logistic regression, which is the probability of success divided by the probability of failure. The following formulas show this logistic function, which is also known as the log odds or the natural logarithm of odds:

$$\mathrm{Logit}(p_i) = \frac{1}{(1+\exp(-p_i))}$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathrm{Beta\_0} + \mathrm{Beta\_1} * \mathrm{X\_1} + \cdots + \mathrm{B\_k} * \mathrm{K\_k}$$

- *Random Forest (RF)*

The supervised learning method includes the well-known machine learning algorithm RF. In ML, it can be utilized for both regression and classification issues [9]. It is based on the idea of ensemble learning, in which multiple classifiers are combined to solve a complex problem and boost the model's performance. RF is a classifier that takes the average of a number of decision trees on various subsets of the given dataset to increase that dataset's predictive accuracy. The RF predicts the final result based on the majority of votes cast for each prediction, rather than relying on a single decision tree. The problem of overfitting is avoided and accuracy is improved when there are more trees in the forest.

$$\mathrm{RFfi_i} = \frac{\sum_{\mathrm{j \in all\ trees}} \mathrm{normfi_{ij}}}{\mathrm{T}}$$

Where, RFfi sub(i) is the calculated importance of feature i from all of the Random Forest model's trees; normfi sub(ij) is the normalized importance of feature i in tree j; and T is the total number of trees.

- *AdaBoost*

The first and most effective method of boosting, AdaBoost aims to combine multiple weak classifiers into a single strong classifier [10]. An object's class might not be accurately predicted by a single classifier; however, by combining a number of weak classifiers and gradually learning from the incorrectly classified objects of the others, we can construct a strong model. The classifier described here could be implemented using any of your standard classifiers, such as LR or DT.

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

Where, each ft is a weak learner that takes an object x as input and returns a value indicating the object's class.

- *Decision Tree (DT)*

The DT algorithm is a member of the supervised learning algorithm family. The decision tree algorithm, in contrast to other supervised learning algorithms, can also be used to solve regression and classification problems [11]. Using a Decision Tree, a training model that can use simple decision rules inferred from previous data to predict the class or value of the target variable is the goal. When attempting to predict a record's class label using Decision Trees, we begin at the tree's base. The values of the record's attribute and the root attribute are compared. We jump to the next node based on comparison by following the branch that corresponds to that value.

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

Where S is the current state and Pi is the probability of an event i in state S or the percentage of class i in a state S node

- *K-Nearest Neighbor (K-NN)*

A non-parametric, supervised learning classifier known as the KNN or k-NN uses proximity to classify or predict the grouping of a single data point [12]. It can be used for either classification or regression problems, but most of the time it is used as a classification algorithm because it assumes that similar points can be found close to each other. In order to give a query point a class label, we use the k-NN algorithm to determine its closest neighbors. The distance between the query point and the other data points will need to be calculated in order to determine which data points are closest to the query point. These distance metrics aid in the formation of decision boundaries, which are calculated as follows and divide query points into various regions:

$$\text{Euclidean distance} = d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

$$\text{Manhattan distance} = d(x,y) = (\sum_{i=1}^{m} |x_i - y_i|)$$

$$\text{Minkowski distance} = (\sum_{i=1}^{n} |x_i - y_i|)^{\frac{1}{p}}$$

$$\text{Hamming distance} = D_H = \left(\sum_{i=1}^{k} |x_i - y_i|\right)$$

$$x = y \quad D = 0$$
$$x \neq y \quad D \neq 1$$

- *Gradient Boosting Classifier (GBC)*

Regression and classification, among other tasks, are examples of machine learning applications for gradient boosting [13]. Ensembles of weak prediction models, typically decision trees, are provided as a prediction model. The algorithm that is produced is referred to as gradient-boosted trees when a decision tree serves as the weak learner. It typically performs better than random forest. Similar to other boosting techniques, a gradient-boosted trees model is constructed stage-by-stage, but it extends these techniques by allowing for the optimization of any differentiable loss function.

***Article Received**: 01 April 2023 **Revised**: 10 May 2023 **Accepted**: 25 May 2023*

$$\gamma = \frac{\sum Residual}{\sum[previous\ Prob * (1 - Previous\ Prob)]}$$

Where the sign denotes the "sum of and PreviousProb denotes the probability that we have previously calculated

- *Voting Classifier (hard)*

A voting classifier is a type of machine learning model that predicts an output based on which model has the highest probability of selecting that class as the output and trains on a collection of other models [14]. It predicts the output class based on the class with the greatest number of votes by simply combining the results of each classifier that is fed into Voting Classifier. The idea is to create a single model that trains on these models and predicts output based on their combined majority of votes for each output class, as opposed to developing distinct models and determining their accuracy.

$$\tilde{y} = mode\{C_1(x), C_2(x), \ldots, C_m(x)\}$$

In this case, we use the majority vote of each classifier C_j to predict the class label ỹ

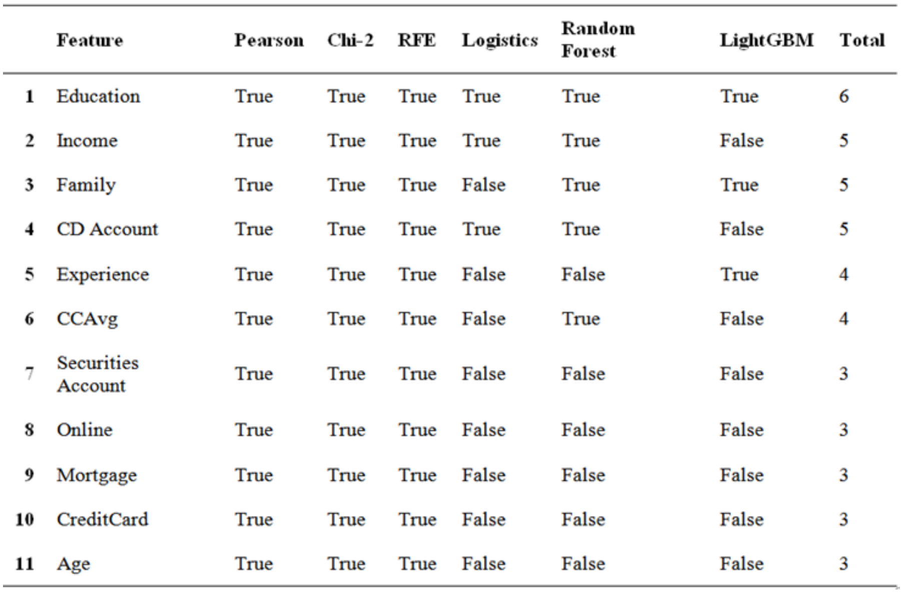
- *Stacking*

One of the most widely used ensemble machine learning methods is stacking, which is used to predict multiple nodes to construct a new model and boost model performance. We can train multiple models to solve similar problems with stacking, which then creates a new model with improved performance based on their combined output [15]. We can make better predictions for the future by combining several weak learners with Meta learners that have been parallel ensemble. An extended form of the model averaging ensemble technique, stacking is also known as a stacked generalization. In stacking, all sub-models participate equally based on their performance weights to create a new model with better predictions. The other models are stacking on top of this new one; Because of this, it is referred to as stacking.

B. *Feature Selection Techniques*

A more insightful model that considers a variety of factors in light of specific outcomes is made possible by machine learning. Phased regression and other feature selection techniques used to filter out irrelevant predictors are ignored by correlation matrices [16]. By ensuring that the words that are related to the worst feature are deleted at each round, it handles the conditions of best feature selection. In addition, the other strategic model that is followed when developing a machine learning model is to use cross-validation iteration to keep the predictor's subset as a current feature selection system. Following are some of the key features that are chosen as integral to the analysis models based on the inputs reviewed in the literature regarding potential machine learning models and the features that are chosen for analysis. These features can assist in improving the machine learning model's overall system. According to Table 1 and Figure 1, Pearson, chi-2, and RFE selected all 11 features as important, while LR selected Education, Income, Family, CD Account, and CCAvg, LGBM selected Education, Family, and Experience, and RF selected Education, Income, Family, CD Account, and CCAvg. Finally, a six-feature hybrid model selected Education, Income, Family, CD Account, CCAvg, and Experience for further analysis in the after feature selection model. ID and Zip code have been removed because they have no significance to bank loaning system.

TABLE I. SIX DIFFERENT METHODS FOR SELECTING FEATURES



	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	Total
1	Education	True	True	True	True	True	True	6
2	Income	True	True	True	True	True	False	5
3	Family	True	True	True	False	True	True	5
4	CD Account	True	True	True	True	True	False	5
5	Experience	True	True	True	False	False	True	4
6	CCAvg	True	True	True	False	True	False	4
7	Securities Account	True	True	True	False	False	False	3
8	Online	True	True	True	False	False	False	3
9	Mortgage	True	True	True	False	False	False	3
10	CreditCard	True	True	True	False	False	False	3
11	Age	True	True	True	False	False	False	3

***Article Received**: 01 April 2023 **Revised**: 10 May 2023 **Accepted**: 25 May 2023*

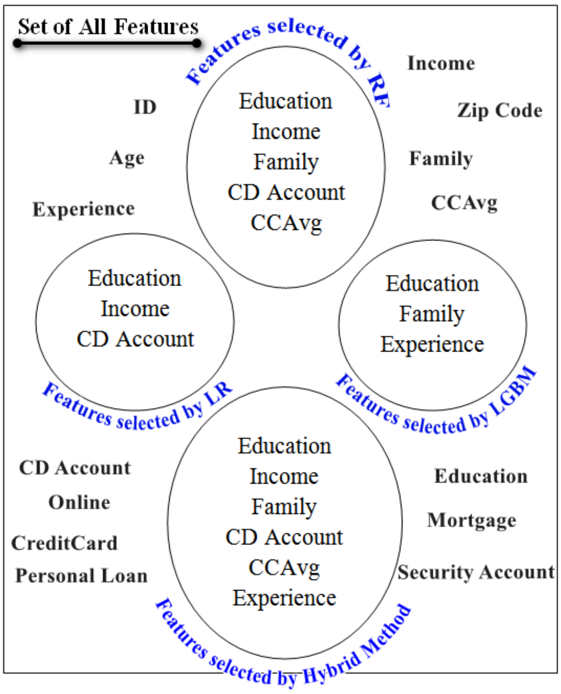


Figure 1. Feature chosen using a hybrid approach

III. EXPERIMENTAL METHODOLOGY

Experiments were carried out by the authors of this paper with the intention of developing a specific diagnostic system that would be able to use machine learning algorithms to objectively classify personal loans from the bank loaning system. A system of this kind would enable a loan officer or decision maker to initiate a personal loan and monitor model efficacy during a machine learning examination, providing them with technical support for their subjective evaluation. First, this paper shows how to use multiple base classifiers to find people in the customer dataset who need a loan. The main reason they are used is that base classifiers can solve many recognition problems without having to use experts to find relevant features. Without sufficient domain expertise, this strategy may be significant. However, in order to construct the hybrid set of features, a set of six feature selection techniques were used to identify the features of the dataset's most relevant and redundant features. In order to achieve accuracy, machine learning algorithms were used to analyze the selected features further. After comparing the accuracy of the classifiers (all features vs. hybrid features), this model is finally able to determine its appropriate objective. This article's comparison will help readers better understand each strategy's benefits and drawbacks. Figure 2 depicts the proposed model's experimental approach.

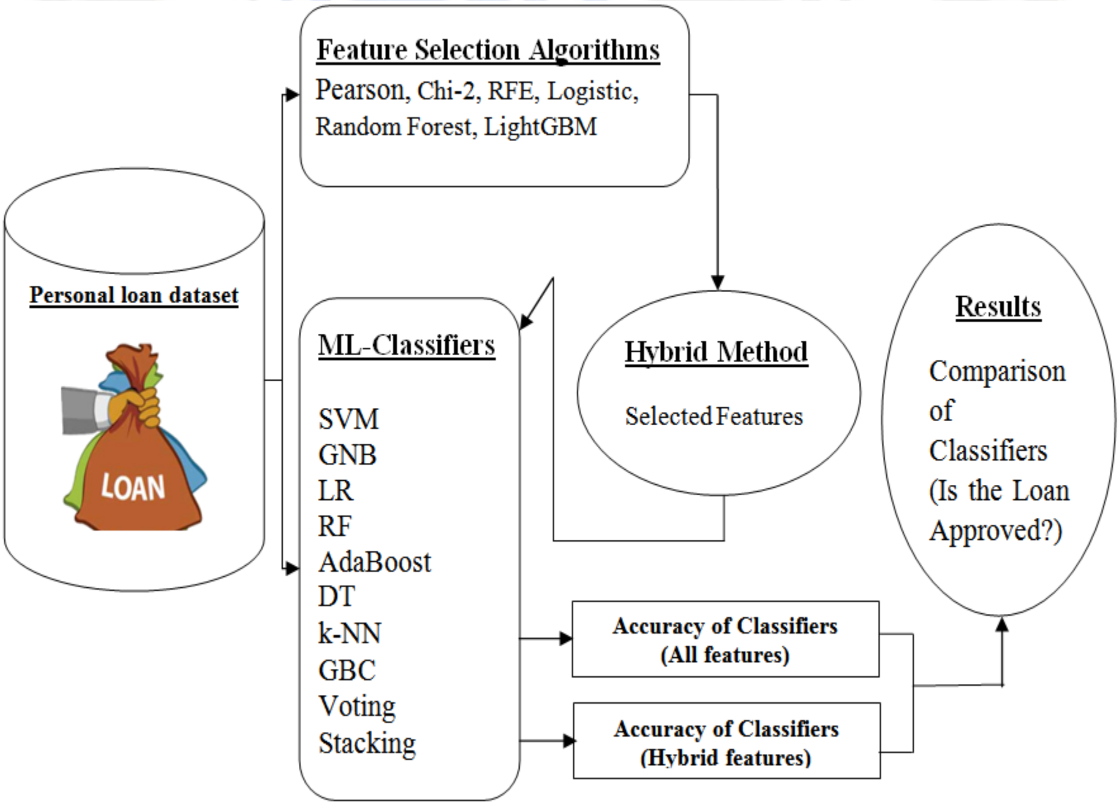
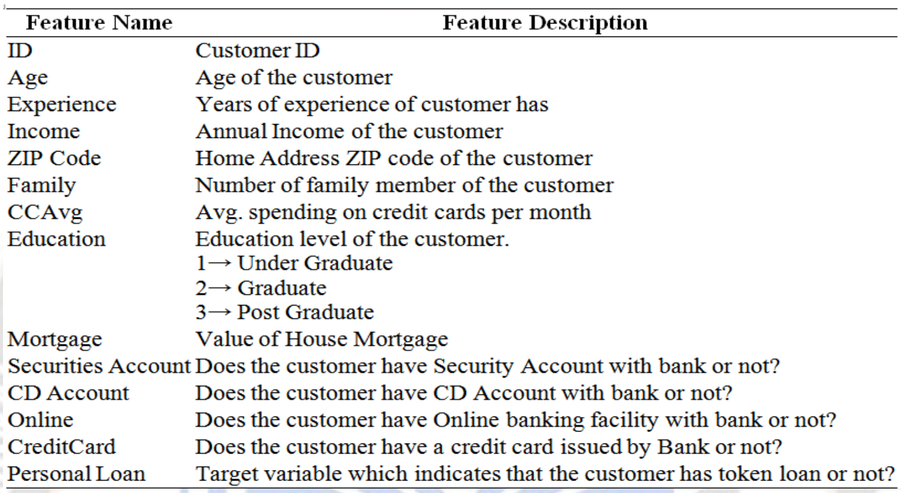

Feature Selection Algorithms
Pearson, Chi-2, RFE, Logistic,
Random Forest, LightGBM
Personal loan dataset
LOAN
ML-Classifiers
SVM
GNB
LR
RF
AdaBoost
DT
k-NN
GBC
Voting
Stacking
Hybrid Method
Selected Features
Results
Comparison of Classifiers (Is the Loan Approved?)
Accuracy of Classifiers (All features)
Accuracy of Classifiers (Hybrid features)

Figure 2. Diagram of the experiment used in the proposed work

IV. EXPERIMENTAL SETUP

Thera Bank generously donated the proprietary data sets used in this study for a personal loan campaign. A personal loan offer was sent to 5000 customers, of whom 480 gave a positive response with a value 1 and 4520 gave a negative response. The dataset "Bank_Personal_Loan_Modelling.xlsx" is obtained from the Kaggle data repository [17]. As shown in Table 2, there are fourteen attributes in this dataset. The attribute Personal Loan is a dependent feature of the dataset, while the name attribute ID and ZIP Code have no bearing on the bank loaning system. Therefore this two features have been removed for further analysis. There are no missing values in the dataset.

TABLE II. ATTRIBUTES DESCRIPTION



Feature Name	Feature Description
ID	Customer ID
Age	Age of the customer
Experience	Years of experience of customer has
Income	Annual Income of the customer
ZIP Code	Home Address ZIP code of the customer
Family	Number of family member of the customer
CCAvg	Avg. spending on credit cards per month
Education	Education level of the customer. 1→ Under Graduate 2→ Graduate 3→ Post Graduate
Mortgage	Value of House Mortgage
Securities Account	Does the customer have Security Account with bank or not?
CD Account	Does the customer have CD Account with bank or not?
Online	Does the customer have Online banking facility with bank or not?
CreditCard	Does the customer have a credit card issued by Bank or not?
Personal Loan	Target variable which indicates that the customer has token loan or not?

V. RESULTS

The analytical results of the various machine learning models used in this paper are presented in this section. In this paper, the validation/test set serves as the foundation for all model diagnostic metrics. Our experiment includes two steps for the purpose of comparison. In the first step, we look at our experiment before selecting features, which is the dataset with all features except ID and ZIP Code. In the second step, we look at the dataset with selected features, which are just six features chosen using a hybrid method.

A. Classifiers Accuracy

As a first step, we looked at each model's overall out-of-sample prediction accuracy on the test set to see which one performed best with our data [18]. Other metrics, such as precision, recall, the f1-score, and the confusion matrix, should also be evaluated in addition to the accuracy of the classifiers, which is the metric that is used to evaluate the performance of any classifier [19]. The classifier's accuracy is evaluated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B. Classification Reports

We talk about the f1-score, classifier precision, and recall in this section.

The precision is the ratio of all predicted values for a given risk class to all actual values. It is the ratio of the values on the leading diagonal to the total of all the values in that column in any of the confusion matrices above. In contrast, recall is a ratio of the actual values of a risk class to the actual values that were predicted to belong to that class. The f1-score is a metric that measures both precision and recall simultaneously, and the inverse relationship between precision and recall is common [20]. It conveys both precision and recall in one image. The two metrics' harmonic mean is it. The accuracy of each classifier is determined by dividing the sum of all elements in the confusion matrix by the sum of all elements on the principal diagonal. The weighted average calculates the weighted average of the precision, recall, and f1 scores, whereas the micro-average metric is the arithmetic mean of the three scores. The confusion matrix in Table 3 can be calculated using the following formula in addition to the f1-score, recall, and precision.

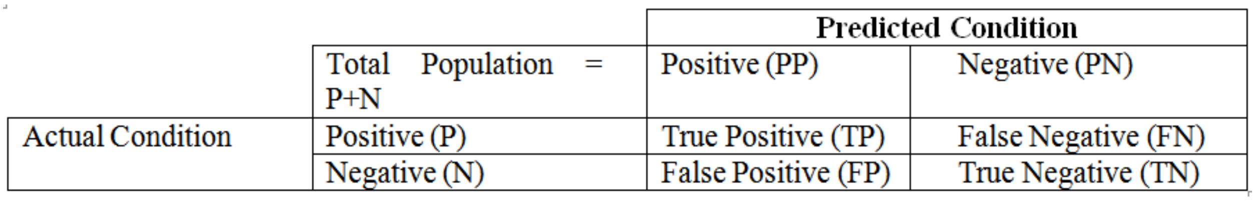
$$Precision = \frac{TP}{TP + FP}$$

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

$$F1\_score = \frac{2 \times precision \times recall}{precision + recall}$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6984
***Article Received**: 01 April 2023 **Revised**: 10 May 2023 **Accepted**: 25 May 2023*

TABLE III. CONFUSION MATRIX



		Predicted Condition	
	Total Population = P+N	Positive (PP)	Negative (PN)
Actual Condition	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

Where, FP (false positive) and FN (false negative) define inaccurate classification, while TP (true positive) and TN (true negative) describe accurate classification.

C. Sensitivity Analysis

Classifier validation metrics include the area under the curve (AUC) of receiver operating characteristic (ROC) curves [21]. A classifier's output quality is measured using ROC curves and AUCs; As a result, they assess how well a classifier has been tuned. The classifier's sensitivity TPR and specificity TNR typically make up for movement along the ROC curve, and the steeper the curve, the better. Sensitivity increases as we move up the ROC curve, while specificity decreases as we move right. The ROC curve at a 45-degree angle is comparable to flipping a coin. Additionally, the AUC performs better the closer it is to 1.

Now, we present each of the results of our experiments one by one.

D. Results based on all features

According to Table 4, the machine learning ensemble classifier RF had the highest accuracy, at 99.20%, followed by the stacking model, which had an accuracy of around 98.90%. In addition, the accuracy of ensemble classifiers like GBC and DT, as shown in Table 4, is superior to that of base classifiers in this dataset.

For the target variables 0 and 1, the classifier RF's precision, recall, and f1-score are superior to those of the other classifiers, which are 0.9913, 1.0000, 0.9956, and 1.0000, 0.9090, 0.9523, respectively. The scores for TP, FN, FP, and TN in the confusion matrix are 912, 0, 8, and 80, which is also better than any of the other performers in Table 4. Similar to the RF classifier, stacking model performs best after it.

TABLE IV. CLASSIFICATION REPORT BEFORE FEATURE SELECTION

Classifiers	**0**			**1**			**Accuracy (%)**	**Confusion Matrix**
	Precision	Recall	F1-score	Precision	Recall	F1-score		
SVM	0.9637	0.9923	0.9778	0.8852	0.6136	0.7248	95.90	[[905 7] [34 54]]
GNB	0.9525	0.9254	0.9388	0.4035	0.5227	0.4554	89.00	[[844 68] [42 46]]
LR	0.9678	0.9890	0.9783	0.8529	0.6590	0.7435	96.00	[[902 10] [30 58]]
RF	0.9913	1.0000	0.9956	1.0000	0.9090	0.9523	99.20	[[912 0] [8 80]]
AdaBoost	0.9783	0.9912	0.9847	0.8947	0.7727	0.8292	97.20	[[904 8] [20 68]]
DT	0.9912	0.9901	0.9906	0.8988	0.9090	0.9039	98.30	[[903 9] [8 80]]
k-NN	0.9445	0.9703	0.9572	0.5714	0.4090	0.4768	92.10	[[885 27] [52 36]]
GBC	0.9891	0.9967	0.9929	0.9629	0.8863	0.9230	98.70	[[909 3] [10 78]]
Voting	0.9733	1.0000	0.9864	1.0000	0.7159	0.8344	97.50	[[912 0] [25 63]]
Stacking	0.9902	0.9978	0.9939	0.9753	0.8977	0.9349	98.90	[[910 2] [9 79]]

DOI: https://doi.org/10.17762/ijritcc.v11i7s.6984
***Article Received**: 01 April 2023 **Revised**: 10 May 2023 **Accepted**: 25 May 2023*

When determining whether two variables have a cause-and-effect relationship, correlation heat maps are frequently utilized [22]. The matrix data structure is utilized when there are multiple variables and the objective is to determine the correlation between all of them and store them using the appropriate data structure. In a correlation heat map, the correlation between the variables on each axis is shown in each square. The correlation is negative to positive. Values that are closer to zero indicate that there is not a linear trend between the two variables. The more positively correlated they are, the closer they are to one another; that is, the relationship between the two gets stronger the closer they are to each other. Similar is a correlation that is closer to -1, but rather than both increasing, one variable will decrease as the other does. Due to the fact that those squares are relating each variable to itself, the diagonals are all
one-dark black. For the remainder, the correlation between the two variables is greater when the number is larger and the color is darker. Because the same two variables are paired in those squares, the plot is also symmetrical about the diagonal. Figure 3 shows that the attributes Income, CCAvg, CD Account, Education, and Mortgage are highly correlated features because they all have high values when compared to the other attributes.

The ensemble classifiers GBC have excellent ROC curves and AUCs, with AUCs is 0.998 threshold and high ROC curves above the 45 degree in Figure 4. This demonstrates that our ensemble classifiers outperform random guessing in terms of predictive power. It should be noted that the presented ROC curves and AUCs for each classifier are superior to the test set result. However, with a score of 0.996, RF comes in second place.

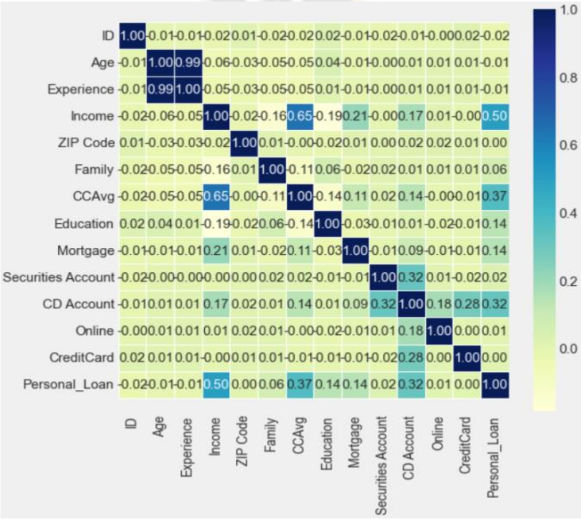


Figure 3. Correlation matrix of attributes

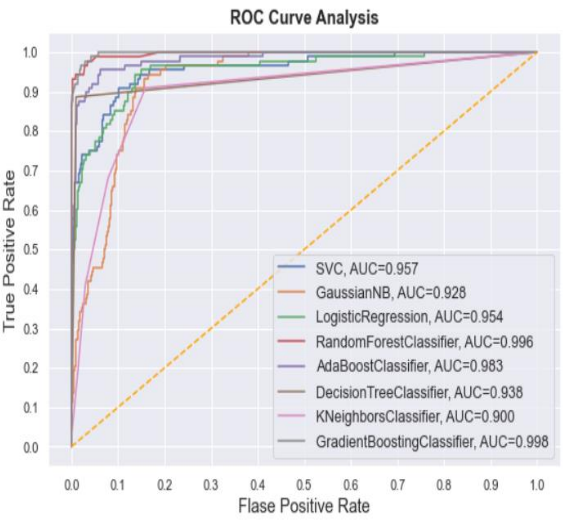


Figure 4. ROC (AUC) before feature selection

E. Results based on Hybrid features

The outcome based on hybrid features is described in this section. As can be seen in Table 5, the RF classifier once more outperformed the other classifiers with an accuracy of 99%. After the RF classifier, the stacking classifier has the highest accuracy of 98.90%**.**

Table 5 shows that the precision, recall, and f1-score as well as the confusion matrix performed better than the other classifiers we used for analysis. Figure 5's ROC (AUC) curve demonstrates that the RF classifier outperformed the other classifiers, with a score of 99.8%. With a ROC (AUC) score of 99.7%, the GBC classifier also performs better after the RF classifier in the validation of their scores**.**

TABLE V. CLASSIFICATION REPORT AFTER FEATURE SELECTION

Classifiers	**0**			**1**			**Accuracy (%)**	**Confusion Matrix**
	Precision	Recall	F1-score	Precision	Recall	F1-score		
SVM	0.9586	0.9923	0.9752	0.8750	0.5568	0.6805	95.40	[[905 7] [39 49]]
GNB	0.9541	0.9353	0.9446	0.4433	0.5340	0.4845	90.00	[[853 59] [41 47]]
LR	0.9634	0.9835	0.9734	0.7826	0.6136	0.6878	95.10	[[897 15] [34 54]]
RF	0.9912	0.9978	0.9945	0.9756	0.9090	0.9411	99.00	[[910 2] [8 80]]
AdaBoost	0.9752	0.9934	0.9842	0.9154	0.7386	0.8176	97.10	[[906 6] [23 65]]
DT	0.9901	0.9901	0.9901	0.8977	0.8977	0.8977	98.20	[[903 9] [9 79]]
k-NN	0.9439	0.9791	0.9612	0.6481	0.3977	0.4929	92.80	[[893 19] [53 35]]
GBC	0.9891	0.9967	0.9929	0.9629	0.8863	0.9230	98.70	[[909 3] [10 78]]
Voting	0.9712	1.0000	0.9854	1.0000	0.6931	0.8187	97.30	[[912 0] [27 61]]
Stacking	0.9902	0.9978	0.9939	0.9753	0.8977	0.9349	98.90	[[910 2] [9 79]]

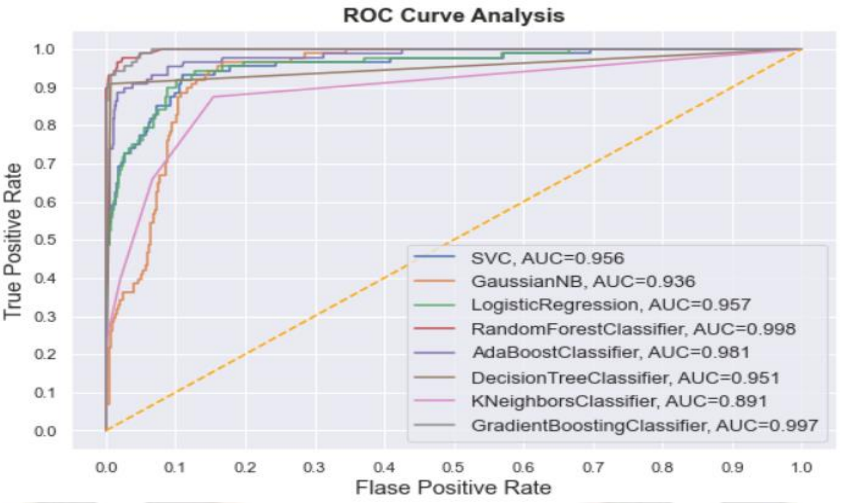
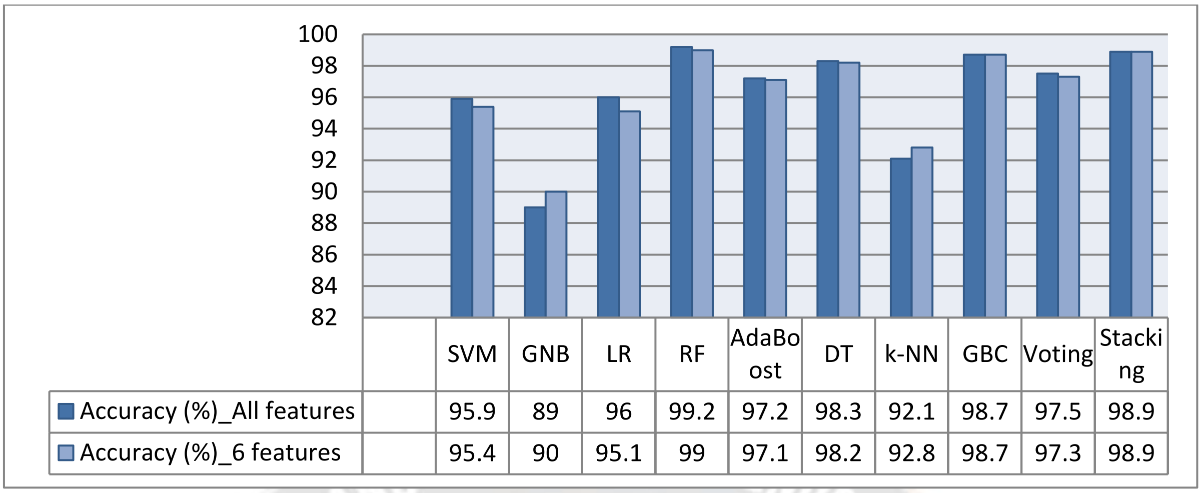


Figure 5. ROC (AUC) of Hybrid features

Last but not least, a graph plot comparing the accuracy of the classifier prior to feature selection (all features) and hybrid features is shown in Figure 6. When applied to all features or to reduced features (Hybrid features), it is clear that RF performs better than classifiers overall.



	SVM	GNB	LR	RF	AdaBoost	DT	k-NN	GBC	Voting	Stacking
■ Accuracy (%)_All features	95.9	89	96	99.2	97.2	98.3	92.1	98.7	97.5	98.9
□ Accuracy (%)_6 features	95.4	90	95.1	99	97.1	98.2	92.8	98.7	97.3	98.9

Figure 6. Comparison of classifiers with all features vs. 6-features

VI. DISCUSSION

The effectiveness of machine learning models in determining default in a credit environment was the subject of this study. In credit, there is typically no central customer credit database, and little to no information about a customer's credit history is available; this is the most common scenario. Lending institutions have a harder time deciding who to lend to because of this. This paper demonstrates that machine learning algorithms are effective at extracting hidden information from the data set, which aids in assessing credit defaults, to overcome the drawback. The test data set served as the basis for all performance metrics used in this paper. Several machine learning models were applied to the data set, but only those with an overall accuracy of 85 percent or higher on the test set were included in this paper.

The two best classifiers, RF and Stacking, are among the models discussed in this paper. All classifiers revealed a general exactness of somewhere around 90% on the test set. The stacking classifiers' ability to accurately predict bank-credit defaults was also demonstrated by other performance metrics (as shown in Sections 6). We used multiclass classification algorithms because they give us the additional advantage of having the average risk class, allowing us to further investigate customers who are predicted to be in that class before deciding whether to give them loans or not. As discussed in Section 3.2 of this manuscript, it is essential to note that feature selection using various methods plays a leading role in removing redundant features from the dataset. The most important thing is that our experiments have shown that the classifier's RF and Stacking are the best of several classifiers that were used in the bank loaning dataset. RF and stacking should be used because the classifier is most accurate both before and after feature selection. In order to predict defaults in a credit environment, subsequent studies will incorporate inflation and unemployment rate findings from this one.

VII. CONCLUSION

It is always the responsibility of financial institutions to ensure that all loan distributions to customers are highly secured. The establishment must devise an efficient creditworthiness evaluation system, which may result in a superior classification system, which is crucial. Patterns for evaluating credit make use of a lot of AI-based systems.

However, the alignment of the model to the current banking transaction and information systems, which can support the overall system enhancement of the systems, is one of the significant challenges that have been integral to the problem. When it comes to obtaining relevant data from the information system, the features chosen for analysis are typically ineffective, as is the case more frequently. However, more substantial systems that can analyze the credit profiles of individuals in a wider range of circumstances are required so that machine learning models can be enhanced to take into account a wider range of circumstances. Businesses may be able to track a more effective customer base if the features are able to support holistic creditworthiness analysis conditions.

When it comes to choosing non-conventional metrics for credit evaluation models, predictive modeling can provide a more in-depth analysis of the customer profile and the conditions, allowing for a more pragmatic evaluation of customer profiles. In the event that such a comprehensive system is developed, it may result in sustainable business practices when it comes to businesses receiving credit.

Competing interest: Authors have disclosed that there are no competing interests.

REFERENCES

[1] Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. Sustainability, 11(3), 699.

[2] Ricci, A., Jankowski, M., Pedersen, A., Sánchez, F., & Oliveira, F. Predicting Engineering Student Success using Machine Learning Algorithms. Kuwait Journal of Machine Learning, 1(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/118

[3] Van Thiel, D., & Van Raaij, W. F. F. (2019). Artificial intelligence credit risk prediction: An empirical study of analytical artificial intelligence tools for credit risk prediction in a digital era. Journal of Risk Management in Financial Institutions, 12(3), 268-286.

[4] Ubarhande, P., & Chandani, A. (2021). Elements of credit rating: a hybrid review and future research Agenda. Cogent Business & Management, 8(1), 1878977.

[5] Saidi, R., Bouaguel, W., & Essoussi, N. (2019). Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. Machine learning paradigms: theory and application, 3-24.

[6] Kabir, M. M., Shahjahan, M., & Murase, K. (2012). A new hybrid ant colony optimization algorithm for feature selection. Expert Systems with Applications, 39(3), 3747-3763.

[7] Mr. Vaishali Sarangpure. (2014). CUP and DISC OPTIC Segmentation Using Optimized Superpixel Classification for Glaucoma Screening. International Journal of New Practices in Management and Engineering, 3(03), 07 - 11. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/30

[8] Chaurasia, V., & Chaurasia, A. (2023). Detection of Parkinson's Disease by Using Machine Learning Stacking and Ensemble Method. Biomedical Materials & Devices, 1-13.

[9] Soria, D., Garibaldi, J. M., Ambrogi, F., Biganzoli, E. M., & Ellis, I. O. (2011). A ‘non-parametric’version of the naive Bayes classifier. Knowledge-Based Systems, 24(6), 775-784.

[10] Park, H. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing, 43(2), 154-164.

[11] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.

[12] Chaurasia, V., & Pal, S. (2022). Ensemble technique to predict breast cancer on multiple datasets. The Computer Journal, 65(10), 2730-2740.

[13] Feng, D. C., Liu, Z. T., Wang, X. D., Jiang, Z. M., & Liang, S. X. (2020). Failure mode classification and bearing capacity prediction for reinforced concrete columns based on ensemble machine learning algorithm. Advanced Engineering Informatics, 45, 101126.

[14] Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019, May). A brief review of nearest neighbor algorithm for learning and classification. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 1255-1260). IEEE.

[15] Costa, M. A., Wullt, B., Norrlöf, M., & Gunnarsson, S. (2019). Failure detection in robotic arms using statistical modeling, machine learning and hybrid gradient boosting. Measurement, 146, 425-436.

[16] Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. International journal of environmental research and public health, 18(6), 3317.

[17] Kumar, N. V. M. ., Raju, D. N. ., PV, G. ., & Subhashini, P. (2023). Real-Time User-Service Centric Historical Trust Model Based Access Restriction in Collaborative Systems with Blockchain Public Auditing in Cloud. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 69–75. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2509

[18] Chen, Y. L., Hsiao, C. H., & Wu, C. C. (2022). An ensemble model for link prediction based on graph embedding. Decision Support Systems, 157, 113753.

[19] Ahammad, D. S. H. ., & Yathiraju, D. . (2021). Maternity Risk Prediction Using IOT Module with Wearable Sensor and Deep Learning Based Feature Extraction and Classification Technique. Research Journal of Computer Systems and Engineering, 2(1), 40:45. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/19

[20] Krmar, J., Vukićević, M., Kovačević, A., Protić, A., Zečević, M., & Otašević, B. (2020). Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure-retention relationships modelling in micellar liquid chromatography. Journal of Chromatography A, 1623, 461146.

[21] Carmen Rodriguez, Predictive Analytics for Disease Outbreak Prediction and Prevention , Machine Learning Applications Conference Proceedings, Vol 3 2023.

[22] Bank_Personal_Loan_Modelling.xlsx. https://www.kaggle.com/code/pritech/bank-personal-loan-modelling/data. Accessed 25th January 2023.

[23] Li, H., Leung, K. S., Wong, M. H., & Ballester, P. J. (2015). Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. Molecular informatics, 34(2-3), 115-126.

[24] 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.

[25] Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., & Gottschlich, J. (2018). Precision and recall for time series. Advances in neural information processing systems, 31.

[26] Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. Physics in Medicine & Biology, 63(7), 07TR01.

[27] Besmer, M. D., Weissbrodt, D. G., Kratochvil, B. E., Sigrist, J. A., Weyland, M. S., & Hammes, F. (2014). The feasibility of automated online flow cytometry for in-situ monitoring of microbial dynamics in aquatic ecosystems. Frontiers in microbiology, 5, 265.